frontiers | Research Topics

Rapid, reproducible, and robust environmental modeling for decision support: Worked examples and open-source software tools

Edited by
Jeremy White, Michael Fienen, Catherine Moore and Anneli Guthke

Published in
Frontiers in Earth Science

ISSN 1664-8714
ISBN 978-2-8325-3581-3
DOI 10.3389/978-2-8325-3581-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Rapid, reproducible, and robust environmental modeling for decision support: Worked examples and open-source software tools

Topic editors

Jeremy White — Intera, Inc, United States

Michael Fienen — Upper Midwest Water Science Center, United States Geological Survey, United States

Catherine Moore — GNS Science, New Zealand

Anneli Guthke — University of Stuttgart, Germany

Citation

White, J., Fienen, M., Moore, C., Guthke, A., eds. (2023). *Rapid, reproducible, and robust environmental modeling for decision support: Worked examples and open-source software tools*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-3581-3

Table of contents

frontiers | Frontiers in Earth Science

TYPE Editorial
PUBLISHED 13 September 2023
DOI 10.3389/feart.2023.1260581

OPEN ACCESS

EDITED AND REVIEWED BY
Wouter Buytaert,
Imperial College London,
United Kingdom

*CORRESPONDENCE
Jeremy T. White,
jwhite@intera.com

RECEIVED 18 July 2023
ACCEPTED 06 September 2023
PUBLISHED 13 September 2023

CITATION
White JT, Fienen MN, Moore CR and Guthke A (2023), Editorial: Rapid, reproducible, and robust environmental modeling for decision support: worked examples and open-source software tools.
Front. Earth Sci. 11:1260581.
doi: 10.3389/feart.2023.1260581

Editorial: Rapid, reproducible, and robust environmental modeling for decision support: worked examples and open-source software tools

Jeremy T. White[1]*, Michael N. Fienen[2], Catherine R. Moore[3] and Anneli Guthke[4]

[1]Intera Geosciences, Perth, WA, Australia, [2]U.S. Geological Survey, Upper Midwest Water Science Center, Madison, WI, United States, [3]GNS Science, Wellington, New Zealand, [4]Stuttgart Center for Simulation Science, Cluster of Excellence EXC 2075, University of Stuttgart, Stuttgart, Germany

KEYWORDS

modeling, decision support, reproducability, uncertainty analysis (UA), open-source

Editorial on the Research Topic
Rapid, reproducible, and robust environmental modeling for decision support: worked examples and open-source software tools

To provide support for resource management decision making, computational modeling workflows in environmental simulations need to be efficient, reproducible, and robust with regard to informing assessments of the risk of unwanted outcomes. Each of these three attributes is difficult to achieve in practice; aspirations to simultaneously achieve all of them are truly lofty. Too often, modeling analyses are inefficient, the workflow is largely opaque and unknown, and the important simulated outcomes lack the context of uncertainty and/or risk. This Research Topic called for papers that demonstrate rapid, reproducible and/or robust modeling through worked examples and software tools (a preference for open source). The worked examples should demonstrate how the researcher aspired to be rapid, reproducible, and robust; we were interested in the process and approach as much as the results. We aim to stimulate discussion based on lessons learned and results presented, for other researchers and practitioners to build on. We particularly welcomed descriptions of trials and tribulations: What was difficult? What did not work? How were these issues overcome?[1]

Generally, we identified three categories of contributions:

- New open-source software tools designed to facilitate aspects of environmental, hydrological and geophysical modeling;
- New approaches to enable better decision support with modeling;
- Demonstrations/case studies of rapid, reproducible, and robust modeling.

1 Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

These contributions came from a wide range of author backgrounds and institutions. This diversity shows that there is broad interest from academia, industry, and government agencies in rapid, reproducible, and robust modeling workflows. We continue to help promote such methods through convening dedicated sessions at international conferences.

Open-source software to support modeling

Leaf and Fienen present Modflow-setup, a workflow toolset to automate the construction of numerical groundwater models for the MODFLOW platform from original geospatial and tabular datasets. The open-source, online code base is extensible through collaborative version control.

Moges et al. call for reproducible model benchmarking and diagnostics, which will find wide acceptance in modeling communities only through standardized methods and ready-to-use toolkits. Using the Jupyter platform, they have introduced HydroBench: an open-source toolset for objectively benchmarking hydrological models that can further be developed by the hydrological community.

Larsen et al. present pyGSFLOW, a Python toolset to transparently and reproducibly prepare input for and postprocess output of the integrated surface-water/groundwater model GSFLOW.

James et al. provide a new standard for geophysical data formats, termed GS Convention, to improve the interoperability, transferability, and long-term archival of such data. Their open-source toolset GSPy provides methods and workflows to build the respective standardized files.

Morvillo et al. present VisU-HydRA, a Python toolbox to compute exceedance probabilities and resilience measures as a basis for assessing the risk of groundwater contamination. It comes with a step-by-step tutorial to ensure reproducibility of the workflow.

Schorpp et al. introduce ArchPy, a python toolset for automating the construction of Quaternary geological models. This is an important step toward including these uncertainties in subsurface modeling workflows in a transparent and reproducible way, because the traditional approach required multiple manual steps using different software, which rendered updates with new data or automation almost intractable.

Pryet et al. present a scripted workflow that facilitates the use of reverse particle tracking in applied groundwater modeling as an efficient surrogate to more computationally demanding advection-diffusion transport modeling for well susceptibility evaluation.

Mudunuru et al. present an approach to improve the calibration of large-scale integrated hydrological models such as SWAT via deep-learning techniques. Compared to more traditional approaches, the proposed routine is more efficient and achieves higher skill scores in calibration.

New approaches to support model-based decision making

Hugman and Doherty discuss the challenge of choosing the right amount of model complexity for decision-making and propose a methodology that allows expert knowledge of system properties to inform the parameters of a structurally simple model. They demonstrate navigating the conflicting and competing objectives of simple and complex model designs on a case study of predictive modeling to support the management of a stressed coastal aquifer.

Elshall et al. present a method for prescreening-based subset selection with decision relevant metrics to exclude non-representative model runs from the prediction ensemble. Following the FAIR (Findability, Accessibility, Interoperability, and Reuse) Guiding Principles for scientific data management and stewardship, they developed and shared interactive Colab notebooks for data analysis.

Moore et al. present a sequential conditioning approach to account for geostatistical model uncertainty, which is shown to have a decisive impact on representing the connectivity of high permeability pathways in contaminant transport assessment.

Manewell et al. investigate spatial averaging functions to infer aquifer properties from aquifer test drawdowns under heterogeneity and feature boundaries. This helps to characterize and robustly estimate aquifer property heterogeneity in hydrogeological site investigation.

Case studies of rapid, reproducible, and robust workflows

Kitlasten et al. present a scripted, reproducible workflow to analyze the impact of ensemble size and vertical resolution on groundwater age predictions for New Zealand.

Standen et al. demonstrate a scripted and open-source application of decision-support modeling for managed aquifer recharge scenarios to mitigate aquifer contamination from saltwater intrusion in the Algarve region of Portugal.

Chambers et al. present a decision-support modeling analysis of the potential for increased groundwater flooding as a result of projected sea-level rise in the low-lying South Dunedin region of New Zealand. They incorporate risk into the analysis proving valuable new information to decision makers.

Brakenhoff et al. present a fully repeatable demonstration of large-scale transfer-function-noise modeling to differentiate contributions to observed groundwater level variations in a region of the Netherlands. Differentiating pumping and climate sources on water level impacts has important implications in how to manage water resources.

De Sousa et al. present a surface-water/groundwater modeling analysis of a semi-arid closed-basin in southwest Australia, and demonstrate efficient, at-scale application of several advanced analyses.

Höglund et al. report a fully-scripted decision-support modeling analysis within the context of contaminated groundwater discharging to surface-water. Innovative techniques are used to assimilate thermal measurements to better resolve patterns of surface-water/groundwater exchange, leading to improved modeling predictions, and ultimately decision support.

The editors are grateful to all of the authors for providing valuable contributions in the space of rapid, reproducible, and robust modeling. We hope that you enjoy reading these contributions. We also hope that in reading these contributions some of you may feel inspired to engage with the open-source community of your modeling field.

Author contributions

JW: Writing–original draft. MF: Writing–review and editing. CM: Writing–review and editing. AG: Writing–review and editing.

Conflict of interest

Author JW was employed by Intera, INC., a for-profit environmental consulting organization with no mechanism of direct financial benefit arising from JW's involvement in this Research Topic.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Author Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

frontiers in Earth Science

ORIGINAL RESEARCH
published: 25 January 2022
doi: 10.3389/feart.2022.786223

Prescreening-Based Subset Selection for Improving Predictions of Earth System Models With Application to Regional Prediction of Red Tide

Ahmed S. Elshall[1], Ming Ye[1], Sven A. Kranz[1], Julie Harrington[2], Xiaojuan Yang[3], Yongshan Wan[4] and Mathew Maltrud[5]*

[1]Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, United States, [2]Center for Economic Forecasting and Analysis, Florida State University, Tallahassee, FL, United States, [3]Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [4]Center for Environmental Measurement and Modeling, United States Environmental Protection Agency, Gulf Breeze, FL, United States, [5]Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, United States

We present the ensemble method of prescreening-based subset selection to improve ensemble predictions of Earth system models (ESMs). In the prescreening step, the independent ensemble members are categorized based on their ability to reproduce physically-interpretable features of interest that are regional and problem-specific. The ensemble size is then updated by selecting the subsets that improve the performance of the ensemble prediction using decision relevant metrics. We apply the method to improve the prediction of red tide along the West Florida Shelf in the Gulf of Mexico, which affects coastal water quality and has substantial environmental and socioeconomic impacts on the State of Florida. Red tide is a common name for harmful algal blooms that occur worldwide, which result from large concentrations of aquatic microorganisms, such as dinoflagellate *Karenia brevis*, a toxic single celled protist. We present ensemble method for improving red tide prediction using the high resolution ESMs of the Coupled Model Intercomparison Project Phase 6 (CMIP6) and reanalysis data. The study results highlight the importance of prescreening-based subset selection with decision relevant metrics in identifying non-representative models, understanding their impact on ensemble prediction, and improving the ensemble prediction. These findings are pertinent to other regional environmental management applications and climate services. Additionally, our analysis follows the FAIR Guiding Principles for scientific data management and stewardship such that data and analysis tools are findable, accessible, interoperable, and reusable. As such, the interactive Colab notebooks developed for data analysis are annotated in the paper. This allows for efficient and transparent testing of the results' sensitivity to different modeling assumptions. Moreover, this research serves as a starting point to build upon for red tide management, using the publicly available CMIP, Coordinated Regional Downscaling Experiment (CORDEX), and reanalysis data.

Keywords: regional environmental management, harmful algae blooms of red tide, climate models and Earth system models, HighResMIP of CMIP6, multi-model ensemble methods, sub-ensemble selection and subset selection, decision-relevant metrics

OPEN ACCESS

Edited by:
Anneli Guthke,
University of Stuttgart, Germany

Reviewed by:
Stefano Galelli,
Singapore University of Technology and Design, Singapore
Beate G. Liepert,
Bard College, United States

****Correspondence:***
Ming Ye
mye@fsu.edu

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *29 September 2021*
Accepted: *05 January 2022*
Published: *25 January 2022*

Citation:
Elshall AS, Ye M, Kranz SA, Harrington J, Yang X, Wan Y and Maltrud M (2022) Prescreening-Based Subset Selection for Improving Predictions of Earth System Models With Application to Regional Prediction of Red Tide. Front. Earth Sci. 10:786223. doi: 10.3389/feart.2022.786223

INTRODUCTION

To improve raw outputs directly given by Earth system models (ESMs) for providing useful services to societal decision making, a combination of multiple methods is often used such as bias-correction to account for systematic errors (Szabó-Takács et al., 2019; Wang et al., 2019), ensemble recalibration to improve ensemble characteristics (Manzanas et al., 2019), downscaling to improve the spatial and temporal resolution (Gutowski Jr. et al., 2016; Gutowski et al., 2020), and ensemble methods to select and combine different models. Ensemble methods are an active research area as multi-model ensemble can be more robust then a single-model ensemble (DelSole et al., 2014; Al Samouly et al., 2018; Wallach et al., 2018). Single model ensemble is a single Earth system model (ESM) with multiple realizations given perturbed parameters, initialization, physics, and forcings. Multi-model ensemble refers to an ensemble of multiple ESMs with single or multiple realizations of each ESM. Ensemble methods aim at selecting and combining multiple ESMs to form a robust and diverse ensemble of models. Ensemble methods include model weighting by assigning lower weights to less favorable models (Knutti, 2010; Weigel et al., 2010), bagging by using subsets of data or variables (Ahmed et al., 2019), subset-selection in which the best performing independent models are selected (Chandler, 2013; Herger et al., 2018; Ahmed et al., 2019; Hemri et al., 2020), and the combination of these methods (e.g., using subset selection prior to model weighting).

This study focuses on subset selection, which has not received adequate attention in climate and Earth system research (DelSole et al., 2013; Herger et al., 2018). In subset selection, a subset of models, which have better performance in a set of models, are selected as ensemble members. One model could perform better than other models due to more accurate parameterizations, higher spatial resolution, more tight calibration to relevant data sets, inclusion of more physical components, more accurate initialization, and imposition of more complete or more accurate external forcings (Haughton et al., 2015). In addition, one model could perform better than another model for a specific application as we show in this study. Accordingly, a question that often arises in multi-model combination is whether the original set of models should be screened such that "poor" models are excluded before model combination (DelSole et al., 2013). One argument is that combining all "robust" and "poor" models to form an ensemble (e.g., by assigning lower weights for poorly performing models than others) is an intuitive solution that has advantage over subset selection that uses the best performing model (Haughton et al., 2015). One justification is that, while the "poor" model can be useless by itself, it is useful when combined with other models due to error cancellation (Knutti et al., 2010; DelSole et al., 2013; Herger et al., 2018). Another justification is that no small set of models can represent the full range of possibilities for all variables, regions and seasons (Parding et al., 2020). On the other hand, it has been argued that the objective of subset selection is to create an ensemble of well-chosen, robust and diverse models, and thus if the subset contains a large enough number of the highest ranked and independent models, then it will have the characteristics that reflect the full ensemble (Evans et al., 2013).

Subset selection has several advantages and practical needs. First, a thorough evaluation is generally required to remove doubtful and potentially erroneous simulations (Sorland et al., 2020), and to avoid the least realistic models for a given region (McSweeney et al., 2015). Second, predictive performance can generally improve from model diversity rather than from larger ensemble (DelSole et al., 2014). A reason for this is that as more models are included in an ensemble, the amount of new information diminishes in proportion, which may lead to overly confident climate predictions (Pennell and Reichler, 2011). Accordingly, several studies (Herger et al., 2018; Ahmed et al., 2019; Hemri et al., 2020) developed evaluation frameworks in which subset selection is performed prior to model weighting. A third advantage of subset selection is to identify models based on physical relationships highlighting the importance of process-based model evaluation. For example, Knutti et al. (2017) defined the metric of September Arctic sea ice extent, showing that models that have more sea ice in 2100 than observed today and models that have almost no sea ice today are not suitable for the projection of future sea ice. There is no obvious reason to include these "poor model" that cannot simulate the main process of interest. Likewise, for our case study, we show that models that are unable to simulate the looping of a regional warm ocean current in the Gulf of Mexico (i.e., Loop Current) are unsuitable for our environmental management objective (i.e., prediction of the harmful algal blooms of red tide) as described later. Yun et al. (2017) indicate that incorporating such process-based information is important for highlighting key underlying mechanistic processes of the individual models of the ensemble. Fourth, subset selection allows for flexibility in terms of metrics and thresholds to tailor the multi-model ensemble for the needs of specific applications (Bartók et al., 2019). As noted by Jagannathan et al. (2020), model selection studies are often based on evaluations of broad physical climate metrics (e.g., temperature averages or extremes) at regional scales, without additional examination of local-scale decision-relevant climatic metrics, which can provide better insights on model credibility and choice. For example, Bartók et al. (2019) and Bartók et al. (2019) employ subset selection to tailor the ensemble for energy sector needs, and local agricultural need in California, respectively. Finally, another practical need for subset selection is that, due to high computational cost, it is common that only a small subset of models can be considered for downscaling (Ahmed et al., 2019; Parding et al., 2020; Sorland et al., 2020).

Although there is a need for an efficient and versatile method that finds a subset which maintains certain key properties of the ensemble, few work has been done in climate and Earth system research (Herger et al., 2018). Without a well-defined guideline on optimum subset selection (Herger et al., 2018; Ahmed et al., 2019; Bartók et al., 2019; Parding et al., 2020), it is unclear how to best utilize the information of multiple imperfect models with the aim of optimizing the ensemble performance and reducing the presence of duplicated information (Herger et al., 2018). It may

be difficult to predict exactly how many models are necessary to meet certain criteria, and subsets with good properties in one region are not guaranteed to maintain the same properties in other regions (Ross and Najjar, 2019). Typically, modelers make their own somewhat subjective subset choices, and use equal weighting for the models in the subset (Herger et al., 2018). A commonly used approach is model ranking, typically based on model performance to select the top models, which is generally the top three to five models (Jiang et al., 2015; Xuan et al., 2017; Hussain et al., 2018; Ahmed et al., 2019). For example, to derive an overall rank for each model, Ahmed et al. (2019) use comprehensive rating metric to combine information from multiple goodness-of-fit measures for multiple climate variables based on the ability to mimic the spatial or temporal characteristics of observations. Then to form the multi-model ensemble, Ahmed et al. (2019) select the four top-ranked models to evaluate the two cases of equal weighting and a bagging technique of random forest regression. A limitation of this approach is the arbitrary choice of the number of the top ranked model to include. For example, Ross and Najjar (2019) evaluate six subset-selection methods with respect to performance, and investigate the sensitivity of the results to the number of model chosen. They show that selection methods and models used should be carefully chosen. To aid this common approach of subset selection, Parding et al. (2020) present an interactive tool to compare subsets of CMIP5 and CMIP6 models based on their representation of the present climate, with user-determined weights indicating the importance of different regions, seasons, climate variables, and skill scores. This allows the users to understand the implications of their different subjective weights and ensemble member choices.

A less subjective approach for subset selection is to use a method that is designed to address specific key properties of the ensemble. In other words, a subset-selection method finds a subset which maintains certain key properties of the ensemble. Key properties include any combination of several criteria that are performance, ensemble range, ensemble spread, capture of extreme events, model independence, and decision relevant metrics. First, the performance criterion reflects the model's skills in representing past and present climate and Earth system states. Examples include subset-selection methods to favor skilled models (Bartók et al., 2019), and to eliminate models with poorest representation of the present system states (Parding et al., 2020). A second criterion is the range of projected climate and Earth system changes. For example, McSweeney et al. (2015) developed a subset-selection method that captures the maximum possible range of changes in surface temperature and precipitation for three continental-scale regions. Third, the model spread criterion ensures that the ensemble contains representative models that conserve as much as possible the original spread in climate sensitivity and climate future scenarios with respect to variables of interest (Mendlik and Gobiet, 2016; Bartók et al., 2019). Fourth, another subset selection criterion, which is related to model spread, is the captures extreme events (Cannon, 2015; Mendlik and Gobiet, 2016; Farjad et al., 2019). Although some sectors are affected by mean climate changes, the most acute impacts are related to extreme events (Eyring et al., 2019). Fifth, model independence is another important criterion, which can be accounted for using diverse approaches. Sanderson et al. (2015) propose a stepwise model elimination procedure that maximizes intermodel distances to find a diverse and robust subset of models. Similarly, Evans et al. (2013) and Herger et al. (2018) use an indicator method with binary weights to find a small subset of models that reproduces certain performance and independence characteristics of the full ensemble. Binary weights are either zero or one for models to be either discarded or retained, respectively. Sixth, an additional criterion that is particularly important from many climate services is to consider regional application and decision-relevant metrics (Bartók et al., 2019; Jagannathan et al., 2020). Since a primary goal of climate research is to identify how climate affects society and to inform decision making, a community generally needs rigorous regional-scale evaluation for different impacted sectors that include agriculture, forestry, water resources, infrastructure, energy production, land and marine ecosystems, and human health (Eyring et al., 2019). By considering this criterion, subset-selection is not based on general model evaluation irrespective of the application (e.g., Sanderson et al., 2017), but is rather based on regional model evaluation with sector-specific information (Elliott et al., 2015). This includes, for example, considering a combination of climate hazards at a specific region (Zscheischler et al., 2018), and the use of application-specific metrics as in this study.

This study complements an important aspect of subset selection by explicitly considering application specific metrics for subset selection based on a prescreening step. To find more skillful and realistic models for a specific process or application, we develop an indicator-based subset-selection method with a prescreening step. In a prescreening step, models are scored based on physical relationships and their ability to reproduce key features of interest, highlighting the importance of process-based and application specific evaluation of climate models. Our method extends the indicator method based on binary weights of Herger et al. (2018), by scoring each model based on evolving binary weights, which are either zero or one for models to be either discarded or selected, respectively, as explained in the method section. Thus, irrespective of the general predictive performance of the model for the variables of interest (e.g., temperature, sea surface height, wind speed, and precipitation), the model performance is evaluated based on suitability to specific applications for a given problem definition with key features of interest.

In this case study of red tide, models that cannot reproduce key features of interest are the models that cannot simulate the process of Loop Current penetration into the Gulf of Mexico, for example, along with other key features as explained in the method section. Red tide is a common name of harmful algae blooms that occur in coastal regions worldwide due to high concentrations of marine microorganisms such as dinoflagellates, diatoms, and protozoans. Along the West Florida Shelf in the Gulf of Mexico, red tide occurs by the increase of the concentration of *Karenia brevis*, a toxic mixotrophic dinoflagellate. This study focuses on Loop

Current (LC), which is one of the main drivers of red tide in the West Florida Shelf (Weisberg et al., 2014; Maze et al., 2015; Perkins, 2019). LC is a warm ocean current that penetrates and loops through the Gulf of Mexico until exiting the gulf to join the Gulf Stream. Several relations have been established between red tide and LC (Weisberg et al., 2014; Maze et al., 2015; Liu et al., 2016; Weisberg et al., 2019). The relation discussed in Maze et al. (2015) shows that the LC position, which can be inferred from sea surface heigh, can be a definitive predictor of a large red tide bloom possibility. Using CMIP6 and reanalysis data of sea surface heigh as described in the method section, we show that this prescreening-based subset-selection step can help reduce the ensemble size without degrading the predictive performance. We additionally illustrate the caveats of using non-representative models given the notation of error cancellation, showing that that a parsimonious ensemble can be more robust.

In the remainder of the manuscript, we present in *Methods* the red tide case study including the CMIP6 data, reanalysis data, and *Karenia brevis* data. *Methods* also presents the prescreening-based subset selection method. *Results* presents the results, which is following in *Discussion* by providing a discussion on subset selection, challenges of seasonal prediction, and the study limitations and outlook. Finally, we summarize our main findings, and draw conclusions in *Conclusion*.

METHODS

FAIR Guiding Principles

To better support transparency and reproducibility of scientific research, data and codes of scientific research should be part of the scholarly work, and must be considered and treated as a first-class research product (Horsburgh et al., 2020). We follow the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016). Accordingly, the data and codes that are used and developed for this study are Findable, Accessible, Interoperable, and Reusable (FAIR). With respect to the "findable" criterion, our data and codes for data analysis are presented in Jupyter notebooks (Elshall, 2021) to provide rich metadata about the used CMIP data, reanalysis data and *Karenia brevis* data (*Data*). With respect to the "Accessible" criterion, the notebooks are opensource and are available on GitHub (Elshall, 2021). Additionally, the notebooks are supported by Colab cloud computing to make the codes immediately accessible and reproducible by anyone with no software installation and download to the local machine. With respect to the "interoperable" criterion, which refers to the exchange and use of information, the notebooks provide rich metadata with additional analysis details not found in the manuscript. This allows users to make use of the presented information by rerunning the codes to reproduce the results, and to understand the sensitivity of the results to different assumptions and configurations as described in the manuscript. Also, the codes can be used to visualize additional data and results that are not shown in the manuscript as described below. With respect to the "reusable" criterion, all the used data are publicly available, and the codes have publicly data usage license. This allows the users to build additional components to the codes as discussed in the manuscript.

Data

The *Karenia brevis* cell count used in this study are from the harmful algal bloom database of the Fish and Wildlife Research Institute at the Florida Fish and the Wildlife Conservation Commission (FWRI, 2020). In the study area (**Figure 1**) and given the study period from 1993-01 to 2014-12, we identify 15 time intervals of large blooms, and 29 time intervals with no bloom; each time interval is six-month long. Following Maze et al. (2015), to identify a bloom/no-bloom event (z_t), a large bloom is defined as an event with the cell count exceeding $1{\times}10^5$ cells/L for ten or more successive days without a gap of more than five consecutive days, or 20% of the bloom length. Similar to Maze et al. (2015) we define no bloom as the absence of large bloom. The notebook "*Karenia_brevis*_data_processing" (Elshall, 2021) provides the data processing details.

We use global reanalysis data, which combine observations with shortrange weather forecast using weather forecasting models to fill the gaps in the observational records. We use the Copernicus Marine Environment Monitoring Service (CMEMS) monthly gridded observation reanalysis product. Th product identifier is Global_Reanalysis_PHY_001_030 (Drévillon et al., 2018; Fernandez and Lellouche, 2018), and can be download from Mercator Ocean International as part of the Copernicus Programme (https://resources.marine.copernicus.eu/products). The used CMEMS reanalysis product is a global ocean eddy-resolving reanalysis with approximatively 8 km horizontal resolution covering the altimetry from 1993 onward. Similar to CMIP6 data, we only focus on sea surface height above geoid, which is the variable name zos according to the Climate and Forecast Metadata Conventions (CF Conventions).

We use 41 CMIP6 model runs from 14 different models developed by eight institutes (Roberts et al., 2018, Roberts et al., 2019; Cherchi et al., 2019; Golaz et al., 2019; Held et al., 2019; Voldoire et al., 2019; Chang et al., 2020; Haarsma et al., 2020). CMIP6 data can be download from any node (e.g., https://esgf-data.dkrz.de/search/cmip6-dkrz) of the Earth System Grid Federation (ESGF) of World Climate Research Programme (WCRP). The study period is from 1993-01 to 2014-12. We select CMIP6 model runs from the historical experiment (Eyring et al., 2016) and the hist-1950 experiment (Haarsma et al., 2016), which are sibling experiments that use historical forcing of recent past until 2015. The historical simulation that starts from 1850 uses all-forcing simulation of the recent past (Eyring et al., 2016). The hist-1950 experiment that starts from 1950 uses forced global atmosphere-land simulations with daily 0.25° sea surface temperature and sea-ice forcings, and aerosol optical properties (Haarsma et al., 2016). For high-resolution models, our selection criteria are to select all model runs with gridded monthly "sea surface height above geoid," which is the variable name zos according to the Climate and Forecast Metadata Conventions (CF Conventions), with nominal resolution less than or equal to 25 km. For each model we only consider variable zos. Given the available CMIP6 data

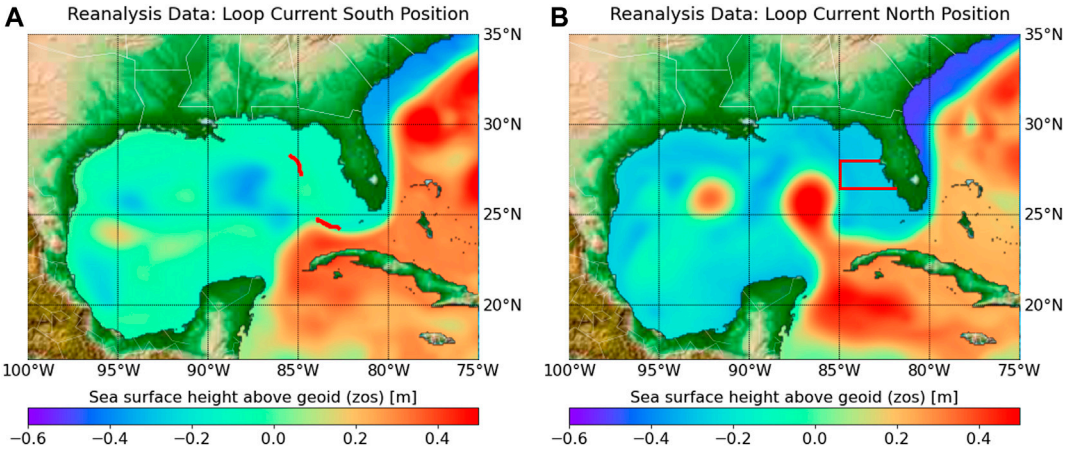

FIGURE 1 | Observation reanalysis data of sea surface height above geoid (zos) [m] showing **(A)** LC-S and **(B)** LC-N. Two red segments along the 300 m isobath in **(A)** are used to determine Loop Current position. The area where red tide blooms are considered by Maze et al. (2015) and this study is shown in the red box of **(B)**.

until September 2020 when this study started, this resulted in 33 model runs. We mainly focus on high-resolution models with eddy-rich ocean resolution, which is important for simulating Loop Current. For our analysis purpose, we include two models with standard resolution. One is EC-Earth3P with nominal ocean resolution of about 100 km given in the hist-1950 experiment with three model runs, and E3SM-1-0 with variable ocean resolution of 30–60 km given in the historical experiment with five model runs.

Model Independence

To account for model independence, we use institutional democracy (Leduc et al., 2016), which can be regarded as a first proxy to obtain an independent subset (Herger et al., 2018), reflecting a priori definition of dependence. For the same institution we created further subsets for different grids. This is the case for the standard- and medium-resolution models of EC-Earth-Consortium that use ORCA1 and ORCA025 grids, respectively. It is also the case for the high-resolution and medium-resolution model of MOHC-NERC that uses ORCA12 and ORC025 grids, respectively. The ORCA family is a series of global ocean configurations with tripolar grid of various resolutions. Thus, the considered 14 models that are listed aphetically by model name in **Table 1**, results in 11 independent model subsets.

For each independent model subset (IMS), multiple perturbed runs of (parameter) realizations (r), initializations (i), physics (p), and forcings (f) are considered. For example, IMS01 has only one model run r1i1p1f1, and IMS11 has seven model runs, three with perturbed initialization r1i (1-3)p1f1, and four with perturbed parameter realizations r (1-4)i1p1f3 as shown in **Table 1**. Note that this naming convention are relative given different modeling groups. For example, the coupled E3SM-1-0 simulations (Golaz et al., 2019) use five ensemble members that are r (1-5)i1p1f1 representing five model runs with different initialization. Each ensemble member (i.e., independent model subset, IMS) in **Table 1** contains one or more models, and each model has one or more model runs. These model runs of each ensemble member should not simply be included in a multi-model ensemble as they represent the same model, hence artificially increasing the weight of models with more model runs. On the other hand, using only one model run per ensemble member discards the additional information provided by these different runs (Brunner et al., 2019). Accordingly, the zos data of each ensemble member is averaged in the way described in *Loop Current Position and Karenia brevis Blooms*.

With the default model independence criteria of institutional democracy and ocean grid we identify 11 ensemble members listed in **Table 1**. The notebook "SubsetSelection" (Elshall, 2021) and its interactive Colab version (https://colab.research.google.com/github/aselshall/feart/blob/main/i/c2.ipynb) provide other model independence criteria that can be investigated by the users. For example, a second case is to use institutional democracy criterion as the first criterion, ocean grid as a second criterion and experiment as a third criterion, which results in 13 ensemble members. In this case historical experiment and hist-1950 experiment are assumed to be independent. A third case is to assume all models are independent, which results in 14 ensemble members. A fourth case is to assume all models are independent, and use experiment as a second criterion, which results in 16 ensemble members. A fifth case is to assume that all members are independent, which results is 41 ensemble members. The code additionally allows for any user defined criteria. While the presented results in this paper are all based on the default model independence criteria, the user can instantly use the above link to investigate the sensitivity of the prescreening and subset selection results and reproduce all figures and under different model independence criteria.

Loop Current Position and *Karenia brevis* Blooms

The mechanisms of initiation, growth, maintenance, and termination of red tides have not been fully understood. Yet

TABLE 1 | Independent model subsets based on institutional democracy and using ocean grid as a secondary criterion when applicable.

Independent model subset (IMS)	Institution	Country	Model (reference)	Experiment ID	Members	Ocean model resolution	Ocean model	Ocean grid	ESM nominal resolution (km)
IMS01	NCAR	United States	CESM1-CAM5-SE-HR (Chang et al., 2020)	hist-1950	r1i1p1f1	0.1° (11 km) nominal resolution	POP2	POP2-HR	25
IMS02	CMCC	Italy	CMCC-CM2-HR4 (Cherchi et al., 2019)	hist-1950	r1i1p1f1	0.25° from the Equator degrading at the poles	NEMO v3.6	ORCA025	25
			CMCC-CM2-VHR4 (Cherchi et al., 2019)	hist-1950	r1i1p1f1	0.25° from the Equator degrading at the poles	NEMO v3.6	ORCA025	25
IMS03	CNRM-CERFACS	France	CNRM-CM6-1-HR (Voldoire et al. (2019))	hist-1950	r (1-3) i1p1f2	0.25° (27–28 km) nominal resolution	NEMO v3.6	eORCA025	25
			CNRM-CM6-1-HR (Voldoire et al., 2019)	Historical	r1i1p1f2	0.25° (27–28 km) nominal resolution	NEMO v3.6	eORCA025	25
IMS04	DOE-E3SM-Project	United States	E3SM-1-0 (Golaz et al., 2019)	Historical	r (1-5) i1p1f1	60 km in mid-latitudes and 30 km at the equator and poles	MPAS-O	EC60to30	100
IMS05	EC-Earth-Consortium	Europe	EC-Earth3P (Haarsma et al., 2020)	hist-1950	r (1-3) i1p2f1	about 1° (110 km)	NEMO v3.6	ORCA1	100
IMS06	EC-Earth-Consortium	Europe	EC-Earth3P-HR (Haarsma et al., 2020)	hist-1950	r (1-3) i1p2f1	about 0.25° (27–28 km)	NEMO v3.6	ORCA025	25
IMS07	ECMWF	Europe	ECMWF-IFS-HR (Roberts et al., 2018)	hist-1950	r (1-6) i1p1f1	25 km nominal resolution	NEMO v3.4	ORCA025	25
IMS08			ECMWF-IFS-MR (Roberts et al., 2018)	hist-1950	r (1-3) i1p1f1	25 km nominal resolution	NEMO v3.4	ORCA025	25
IMS09	NOAA-GFDL	United States	GFDL-CM4 (Held et al., 2019)	Historical	r1i1p1f1	0.25° (27–28 km) nominal resolution	MOM6	tri-polar grid	50
			GFDL-ESM4 (Held et al., 2019)	Historical	r (2-3) i1p1f1	0.25° (27–28 km) nominal resolution	MOM6	tri-polar grid	50
IMS10	NERC	United Kingdom	HadGEM3-GC31-HH (Roberts et al., 2019)	hist-1950	r1i1p1f1	8 km nominal resolution	NEMO v3.6	ORCA12	10
	MOHC-NERC	United Kingdom	HadGEM3-GC31-HM (Roberts et al., 2019)	hist-1950	r1i (1-3) p1f1	25 km nominal resolution	NEMO v3.6	ORCA12	50
IMS11	MOHC	United Kingdom	HadGEM3-GC31-MM (Roberts et al., 2019)	hist-1950	r1i (1-3) p1f1	25 km nominal resolution	NEMO v3.6	ORCA025	100
			HadGEM3-GC31-MM (Roberts et al., 2019)	Historical	r (1-4) i1p1f3	25 km nominal resolution	NEMO v3.6	ORCA025	25

Loop Current, which is a warm ocean current that moves into the Gulf of Mexico, is an important factor that controls the occurrence of red tide (Weisberg et al., 2014; Maze et al., 2015; Perkins, 2019). Maze et al. (2015) shows that the difference between time intervals of large blooms and no blooms is statistically significant for the Loop Current's position. Maze et al. (2015) also show that the Loop current in a north position penetrating through the Gulf of Mexico is a necessarily condition for a large Karenia brevis bloom to occur. As such, when the Loop Current is in the south position shown in **Figure 1A**, which is hereinafter denoted as Loop Current-South (LC-S), then there is no large bloom (Maze et al., 2015). When the Loop Current is in the north position shown in **Figure 1B**, which hereinafter is denoted as Loop Current-North (LC-N), then there could be either large blooms or no blooms. This relationship between the loop current positions and Karenia brevis is based on retention time. With approximately 0.3 divisions per day, Karenia brevis is a slow growing dinoflagellate that requires an area with mixing slower than the growth rate to form a bloom (Magaña and Villareal, 2006). As such, LC-N increases the retention rate allowing bloom formation, if other conditions

are ideal (Maze et al., 2015). While there are several studies that establish different relationships between Loop Current and Karenia brevis (Weisberg et al., 2014; Maze et al., 2015; Liu et al., 2016; Weisberg et al., 2019), the aim of this study is not to support or refute any of these relationships, but to use the study of Maze et al. (2015) for the purpose of our subset selection analysis.

The LC and its eddies can be detected from sea surface height variability. When the difference between the average sea surface height of the north and south segments along the 300 m isobath (**Figure 1A**) is positive and negative, this is a good proxy for identify LC-N and LC-S, respectively (Maze et al., 2015). The zos data processing steps to determine the Loop Current positions (i.e., LC-N and LC-S) are as follows:

1) The zos data is preprocessed for the north and south segments (**Figure 1A**) for all model runs and observation analysis data. Model runs and observation reanalysis data are sampled using nearest neighborhood method along the line points (approximately spaced at 1 km interval between two neighboring points) of the north and south segments (**Figure 1A**). The nearest neighborhood sampling is performed using the python package of xarray project (http://xarray.pydata.org) that handles NetCDF (Network Common Data Form) data formats with file extension NC that is used typically for climate data (e.g., CMIP and reanalysis data). This has an additional practice advantage of reducing the size of the ESMs and reanalysis data. For example, in this case preprocessing CMIP6 and CMEMS data reduced that data size from more than 80 GB to about 11 MB interactive cloud computing feasible. Given data preprocessing, we have a zos datum $h_{(j,k,l,m,n,t)}$ for a model run with index j, an ensemble member with index k, a spatial point along the segment with index l, a segment (i.e., the north or south segment in **Figure 1A**) with index m, a model and reanalysis datasets temporal interval (i.e., 1 month) with index n, and a prediction interval with index t.
2) The expectation of zos data is taken for all model runs $j \in [1, J]$ of each ensemble member M_k

$$h_{k,l,m,n,t} = E_j\left(h_{j,k,l,m,n,t}\middle|M_k\right) \tag{1}$$

The size J of each ensemble member varies depending on the number of model runs in the ensemble member, with the minimum $J = 1$ for ensemble member IMS01 and the maximum $J = 7$ for ensemble member IMS11 (**Table 1**).

(3) The zos data is averaged for all ensemble members $k \in [1, K]$

$$h_{l,m,n,t} = E_k\left(E_j\left(h_{j,k,l,m,n,t}\middle|M_k\right)\right) \tag{2}$$

where k is the index of each ensemble member M_k. The size K of the multi-model ensemble varies based on subset selection (*Prescreening*), which determines the inclusion and exclusion of ensemble members. For example, using all available ensemble members without any subset selection results in $K = 11$ that is all the independent model subsets in **Table 1**. If we evaluate k for only one ensemble member for prescreening purpose (*Prescreening*), then $K = 1$.

4) For each of the north and south segments the expected zos is calculated for each segment

$$h_{m,n,t} = E_l\left[E_k\left(E_j\left(h_{j,k,l,m,n,t}\middle|M_k\right)\right)\right] \tag{3}$$

5) The zos data of the north segment is subtracted from the south segment

$$h_{n,t} = \Delta_m\left[E_l\left[E_k\left(E_j\left(h_{j,k,l,m,n,t}\middle|M_k\right)\right)\right]\right] \tag{4}$$

resulting in zos difference data $h_{n,t}$ with $n \in [1, N]$ and $t \in [1, T]$. As such, N represents the interval length such that $N = 3$ for a season interval, and $N = 6$ for a semiannual interval, and T represents the number of intervals. For example, given $N = 6$ as considered in this study and the 22-year study period, then $T = 44$.

6) The maximum $h_{n,t}$ in the 6-month interval is selected to obtain the zos anomaly per time interval

$$h_t = \max_{h_n}\left(\Delta_m\left[E_l\left[E_k\left(E_j\left(h_{j,k,l,m,n,t}\middle|M_k\right)\right)\right]\right]\right) \tag{5}$$

For each zos anomaly datum h_t, positive and negative values are used as an indicator of LC-N dominated interval and LC-S dominated interval, respectively. Selecting the maximum value $\max_{h_n}(.)$ is more robust than using the average value, which may dilute the signals since the Loop Current position is a cycling event, recalling that loop current has a random and chaotic cycle with the average period of 8–18 months per cycle (Sturges and Evans, 1983; Maze et al., 2015).

The objective of this analysis is not to model the LC cycle, but rather to use the relationship between Loop Current position and *Karenia brevis* bloom of Maze et al. (2015) to obtain a heuristic coarse-temporal-resolution relation between Loop Current position and Karenia brevis. Thus, the h_t values given by **Eq. 5** can be expressed as an indicator function for LC-N:

$$H_{LC-N}(h_t) = \begin{cases} 1, & h_t \geq 0 \\ 0, & h_t < 0 \end{cases} \tag{6}$$

and LC-S:

$$H_{LC-S}(h_t) = \begin{cases} 1, & h_t < 0 \\ 0, & h_t \geq 0 \end{cases} \tag{7}$$

such that $H_{LC-N}(h_t) = 1$ and $H_{LC-S}(h_t) = 1$ indicate a LC-N interval and LC-S interval, respectively. **Eqs 6** and **7** are convenient to use since we are not interested in the value of zos anomaly between the north and south segments per se, but rather in sign difference. Finally, **Eqs 5–7** are valid for both model simulation and observation reanalysis data, which hereinafter are donated as h_t and $h_{t,obs}$, respectively.

Model Performance Metrics

A model performance is based on its ability to reproduce the observed phenomena. We define three qualitative metrics to prescreen for physical relationships, and four quantitative metrics of the model performance. Based on this prescreening we can do subset selection. For prescreening, a process-based

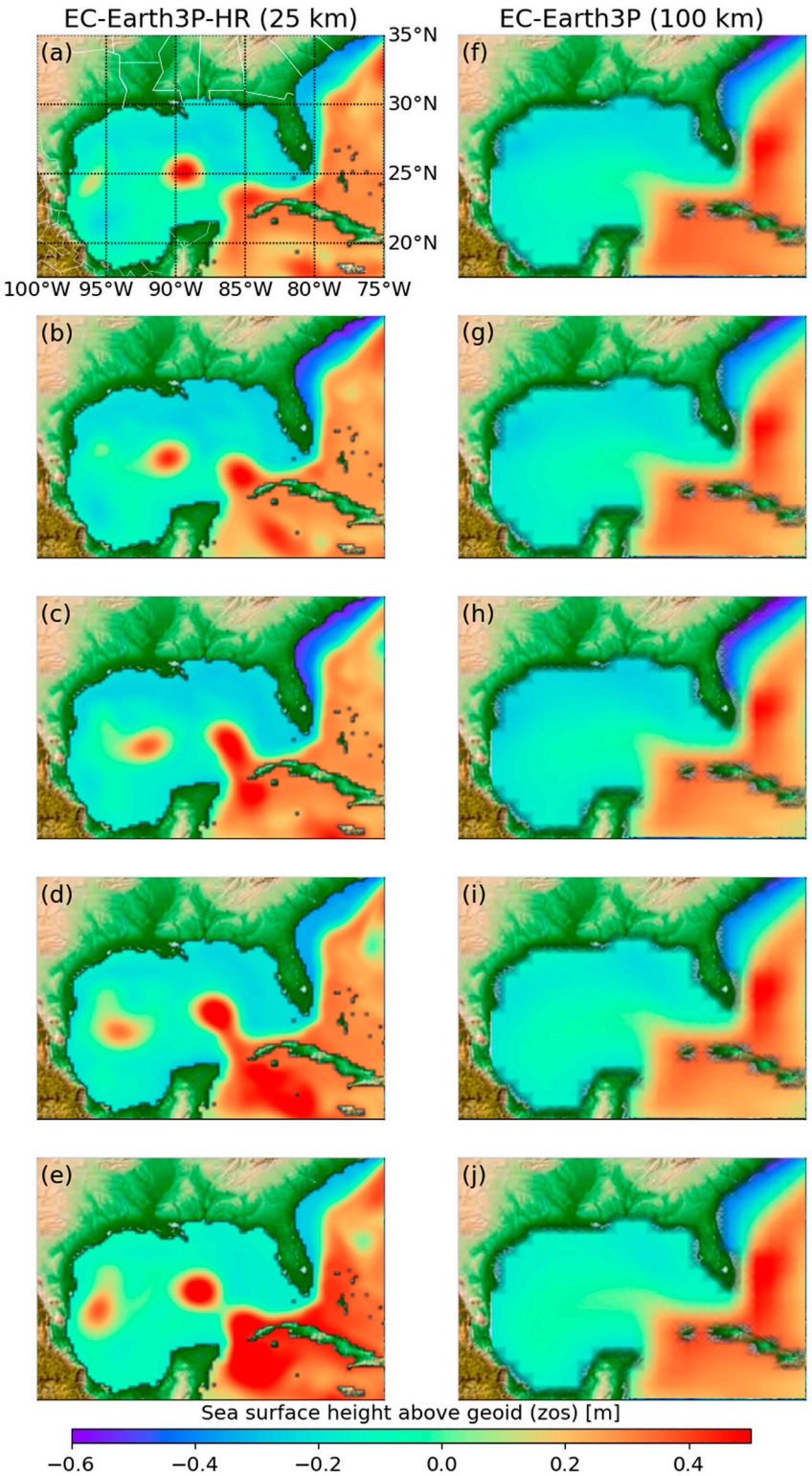

FIGURE 2 | Snapshots of sea surface height above geoid (zos) [m] from 1993-02 to 1993-06 simulated using **(A–E)** a high-resolution ESM, and **(F–J)** standard-resolution ESM with nominal resolution of 10 and 100 km, respectively.

metric is needed, for example, to understand if the model can simulate certain mechanistic aspects of the problem of interest. For example, Christensen et al. (2010) use metrics that capture aspects of model performance in reproducing large-scale circulation patterns and meso-scale signals. A qualitative metric reflects if the model is suitable or unsuitable for reproducing key features of the problem. In our case study, models that cannot reproduce key features of interest would be the models that cannot 1) simulate the penetration of LC into the Gulf of Mexico, 2) represent the alternation of LC in the North and South positions given the empirical method (**Eqs 5–7**), 3) reproduce the higher frequency of Loop Current in the northern and southern positions as described below. For example, with respect to (1), the Loop Current penetrates the Gulf of Mexico extending its northward reach with eddy shedding as shown by the high-resolution model EC-Earth3P-HR (**Figures 2A–C**). As such, intrusion of cooler water increases the stratification of the core of the Loop Current, and the Loop Current becomes unstable forming anticyclonic eddy that breaks from the parent Loop Current westward without reconnecting (Caldwell et al., 2019), as shown by the high-resolution model EC-Earth3P-HR (**Figures 2D,E**). On the other hand, the standard-resolution model EC-Earth3P (**Figures 2F–J**) cannot reproduce the observed physical phenomena, and thus unsuitable for this application. Models that are unable to simulate LC-N are unsuitable for this environmental management purpose. Justifications about selecting these three qualitative metrics and details about them are given below. Finally, for a further illustration of the models that are capable and incapable of reproducing the Loop Current, Elshall (2020) shows an animation of a Loop Current cycle of year 2010 given monthly zos data for all the 41 model runs in **Table 1** shown side-by-side with the reanalysis data. In addition, the reader can visualize the reanalysis data in **Figure 1** and the CMIP6 data in **Figure 2** for any month in the study period 1993–2015 using the Jupyter notebook "DataVisualization_zos" (Elshall, 2021), and its interactive Colab version (https://colab.research.google.com/github/aselshall/feart/blob/main/i/c1.ipynb).

The binary qualitative metrics (y_1-y_3) used for prescreening are as follows:

Physical phenomena simulation (y_1): Accurate simulation of Loop Current positions is generally a challenging task, yet the objective of this first metric is to determine if the model can simulate LC-N irrespective of the accuracy. Thus, the model receives a score one $y_1 = 1$ if it can simulate LC-N (e.g., **Figures 2A–E**), and zero $y_1 = 0$ otherwise (e.g., **Figures 2F–J**), i.e.,

$$y_1 = \begin{cases} 1, & \sum_{t=1}^{T} H_{LC-N}(h_t) > 0 \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) = 0 \end{cases} \tag{8}$$

such that $\sum_{t=1}^{T} H_{LC-N}(h_t)$ is the count on LC-N intervals given the total number of intervals $T = 44$ as explained before.

Oscillating event representation (y_2): This metric is specific to the method of Maze et al. (2015) for determining LC-N and LC-S. If the sea surface height is consistently higher at the north segment than at the south segment, then the model is unable to represent alternation of LC-N and LC-S according to the proxy method of Maze et al. (2015). In this case, the model receives a score zero $y_2 = 0$, and one $y_2 = 1$ otherwise, i.e.,

$$y_2 = \begin{cases} 1, & 0 < \sum_{t=1}^{T} H_{LC-N}(h_t) < T \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) = T \end{cases} \tag{9}$$

Oscillating event realism (y_3): If the frequency of LC-N is greater than that of LC-S for a model, the model receives the score of one $y_3 = 1$ and zero $y_3 = 0$ otherwise, i.e.,

$$y_3 = \begin{cases} 1, & \sum_{t=1}^{T} H_{LC-N}(h_t) \geq \sum_{t=1}^{T} H_{LC-S}(h_t) \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) < \sum_{t=1}^{T} H_{LC-S}(h_t) \end{cases} \tag{10}$$

It is more realistic that the frequency of LC-N is greater than that of LC-S. In the study of Maze et al. (2015), the ratio of the LC-S intervals $\sum_{t=1}^{T} H_{LC-N}(h_t)$ to the total number of intervals $T = 60$ is 0.267, given their altimetry data product with study period of 15 years and 3-month interval (i.e., $N = 3$). In this study the ratio of LC-S to total number of intervals is 0.273, given our reanalysis product with $T = 44$ and $N = 6$ as previously explained.

We define four quantitative metrics (y_4-y_7) to evaluate the predictive performance, and the scoring rules (y_8) to evaluate complexity. These performance criteria are as follows.

Oscillating event frequency (y_4): This is the ratio of the number of a LC position (LC-S or LC-N) to the total number of intervals. Hereinafter, we refer to the oscillating event frequency as the number of LC-S to the total number of intervals T,

$$y_4 = \frac{\sum_{t=1}^{T} H_{LC-S}(h_t)}{T} \tag{11}$$

which can be compared to reanalysis data that is 0.273 as presented in the results section. Additionally, we define the oscillating event frequency error as

$$y_{4,err} = \frac{\left|\sum_{t=1}^{T} H_{LC-S}(h_t) - \sum_{t=1}^{T} H_{LC-S}(h_{t,obs})\right|}{T} \tag{12}$$

which is the absolute difference of LC-S counts of ensemble prediction h_t and reanalysis data $h_{t,obs}$.

Temporal match error (y_5): This is a temporal match of model predictions and reanalysis data with respect to LC position for LC-N

$$y_{5,LC-N} = \frac{\sum_{t=1}^{T} H_{LC-N}(h_{t,obs}) - \sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0)}{\sum_{t=1}^{T} H_{LC-N}(h_{t,obs})} \tag{13}$$

for LC-S

$$y_{5,LC-S} = \frac{\sum_{t=1}^{T} H_{LC-S}(h_{t,obs}) - \sum_{t=1}^{T} (h_{t,obs} < 0 \wedge h_t < 0)}{\sum_{t=1}^{T} H_{LC-S}(h_{t,obs})} \tag{14}$$

and both positions

$$y_5 = \frac{T - \sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0) - \sum_{t=1}^{T} (h_{t,obs} < 0 \wedge h_t < 0)}{T} \tag{15}$$

such that $\sum_{t=1}^{T} H_{LC-N}(h_{t,obs})$ and $\sum_{t=1}^{T} H_{LC-S}(h_{t,obs})$ are the counts of the LC-N and LC-S intervals, respectively, given the observation reanalysis data $h_{t,obs}$; the terms $\sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0)$ and

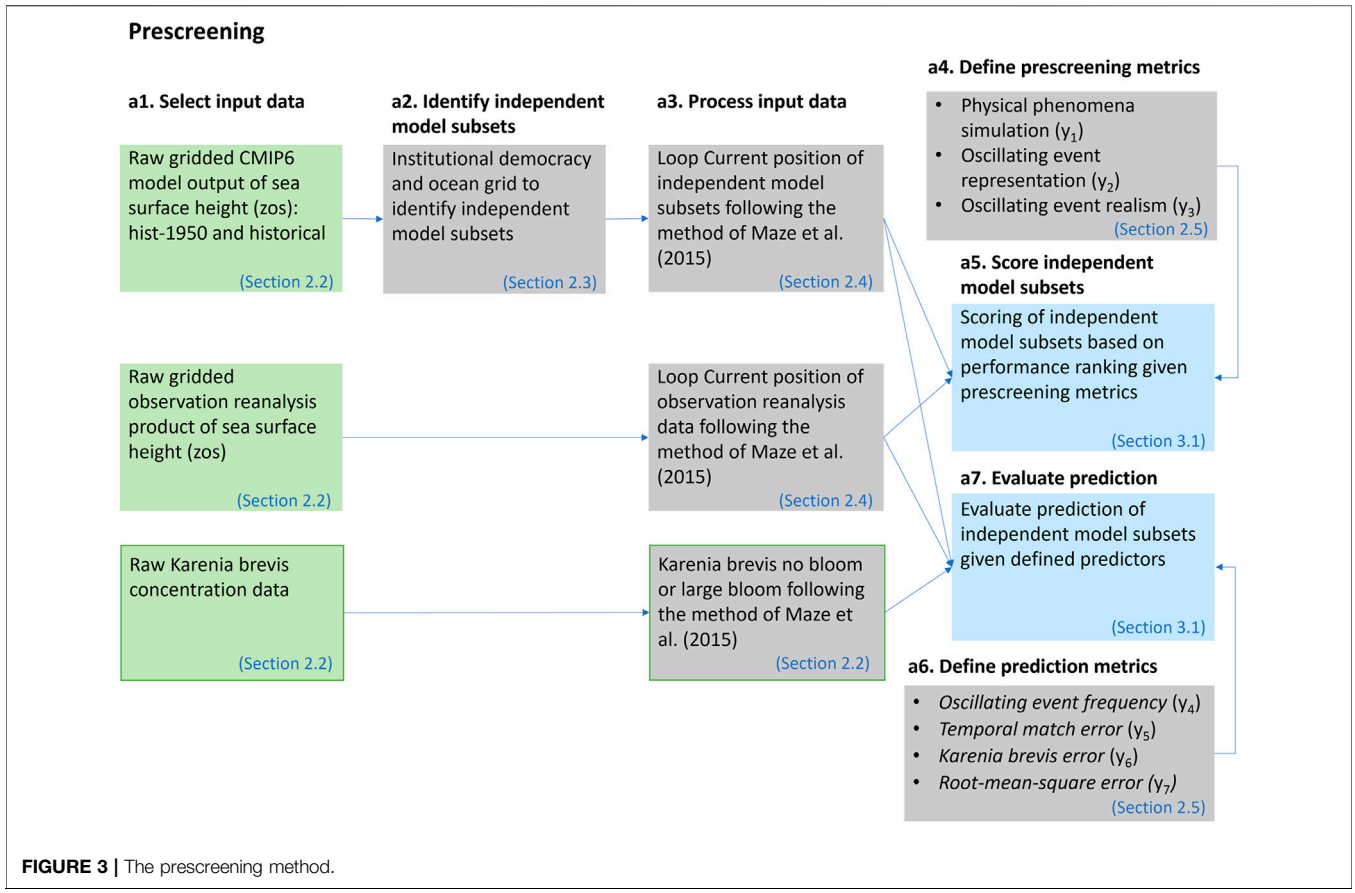

FIGURE 3 | The prescreening method.

$\sum_{t=1}^{T}(h_{t,obs}<0 \wedge h_t<0)$ are the temporal match counts of model simulation and reanalysis data for LC-N and LC-S, respectively. The logical conjunction $\wedge$ gives a value of one when the statement $(h_{t,obs}\geq 0 \wedge h_t \geq 0)$ is true if $h_{t,obs}\geq 0$ and $h_t \geq 0$ are both true, otherwise gives a value of zero if false. Temporal match is the most challenging task. While ESMs are well established on climate timescale, the temporal match at seasonal timescale can be challenging (Hewitt et al., 2017). Generally speaking, the hist-1950 and historical experiments are free-running, and accordingly are neither designed nor expected to have temporal coincide with real-world conditions, which is especially true for the historical experiment. However, one aim of this study is to investigate if any temporal match is possible given the used heuristic relation for determining Loop Current position with a coarse temporal resolution of 6-month interval.

Karenia brevis error (y_6): A false negative prediction of *Karenia brevis* bloom occurs when large bloom coincides with LC-S. For the study period, we define the *Karenia brevis* error as the ratio of the number of LC-S with large bloom to the number of large-bloom N_{bloom}

$$y_6 = \frac{\sum_{t=1}^{T}(h_t<0 \wedge H(z_t)=1)}{N_{bloom}} \tag{16}$$

where $H(z_t)$ is an indicator function with one and zero for large bloom and no bloom, respectively.

Root-mean-square error (y_7): It is the root-mean-square error (RMSE) between model simulation and reanalysis data

$$y_7 = \sqrt{\frac{\sum_{t=1}^{T}(h_t - h_{t,obs})^2}{T}} \tag{17}$$

The defined metrics (y_1- y_7) are specifically designed to judge the predictive performance of these ESMs with respect to the targets of a specific application, and are not meant to judge the predictive skill of these ESMs globally or regionally for general purposes. Judging the predictive skills of these models with respect to global or regional simulations of sea surface height above geoid (variable: zos) or any other variable, is beyond the scope of this work.

Prescreening

Evaluation of specific regional applications is another important criterion, which is the focus of this manuscript. We develop a subset-selection method that extends the binary method of Herger et al. (2018) based on a prescreening step as shown in **Figure 3**. Model independence is accounted for as described in *FAIR Guiding Principles*, and a score is obtained for each ensemble member using three binary qualitative metrics y_1- y_3 (*Model Independence*). Binary refers to a score of either zero or one if the ensemble member is unable or able to produce the metric target. The three binary metrics (**Eqs**

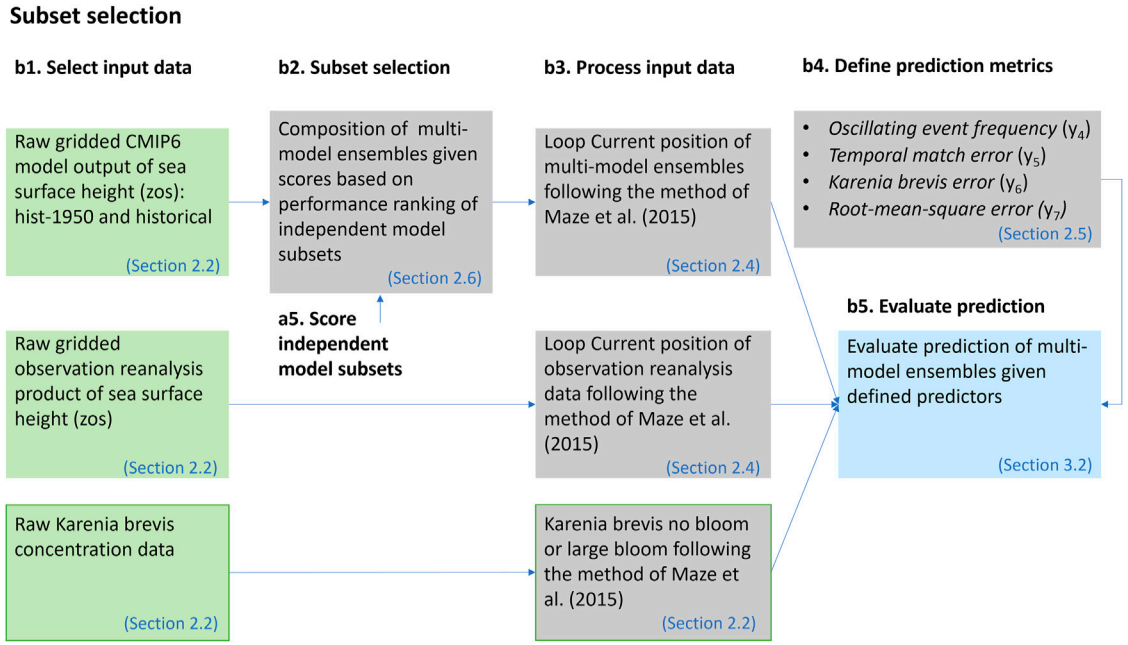

FIGURE 4 | The subset-selection method.

8–10) are evolving such that if the ensemble member fails the first metric, then it will consequently fail in the other two, and will accordingly receive a score of zero. For example, given score (y_1, y_2, y_3), the model receives a score from zero to three for score (0,0,0), (1,0,0), (1,1,0), and (1,1,1), respectively. In other words, if a model score is one for y_3 (**Eq. 10**) it will by default score ones for y_1 (**Eq.11**) and y_2 (**Eq.9**).

Subset Selection

The subset selection step is shown in **Figure 4**. In this step we compose five multi-model ensembles using simple-average multi-model ensemble (SME). Each SME is composed of ensemble members based on prescreening score. The notation SME3210 means that members with prescreening score from zero to three are included in the ensemble. The notation SM321X means that members with prescreening score from one to three are included in the ensemble and members with prescreening score of zero are excluded, and so on. Ensemble SME321X, SME32XX, and SME3XXX exclude ensemble members based on the three binary qualitative metrics (y_1- y_3), respectively. These are evolving metrics such that if an ensemble member scores zero in y_1, it will score zero in y_2 and y_3, and have an overall score of zero. If a model has a score $y_3 = 1$, it will by default score one in y_1 and y_2, and have an overall score of three. As such, SME3210 contains all ensemble members with scores from zero to three, which is all the 11 ensemble members listed in **Table1**. On the other hand, SME3XXX contains the best ensemble members, which are the ones with a score of three. Ensemble SME32XX contains ensemble members with scores of three and two, and so on. On the other hand, ensemble SMEXXX0 contains only the least performing ensemble members with a score of zero. More discussion on the model scores is given in the next section. We evaluate the predictive performance of these five multi-model ensembles using the quantitative metrics (y_4-y_7). The evaluation of these five multi-model ensembles serves multiple purposes as described in the results section.

RESULTS

Prescreening

We plot the oscillation of the Loop Current position for each ensemble member (**Figure 5**), following the zos data processing steps described in *Loop Current Position and Karenia brevis Blooms*. This is to conduct qualitative comparison between the reanalysis data (**Figure 5A**) and the prediction of each ensemble member (**Figures 5B–L**). Accordingly, we score the ensemble member given its performance with respect to three binary evolving metrics (y_1-y_3). The score is zero if the ensemble member fails to pass all the three metrics. This is the case for E3SM-1-0 of DOE-E3SM-Project (**Figure 5E**) and the EC-Earth3P of EC-Earth-Consortium (**Figure 5F**). As these two ensemble members do not pass the first metric of physical phenomena simulation (y_1) that is the simulation of the LC-N, then accordingly they score zero in the next two metrics of oscillating event representation (y_2) and oscillating event realism (y_3). This is not unexpected as these two ensemble members are standard-resolution ESMs, which do not have improved process description as the high-resolution ESMs do. The standard-resolution grids EC60to30 of E3SM-1-0 and ORCA1 of EC-Earth3P do not explicitly resolve the mesoscale eddies and boundary currents, but rather require global parametrization

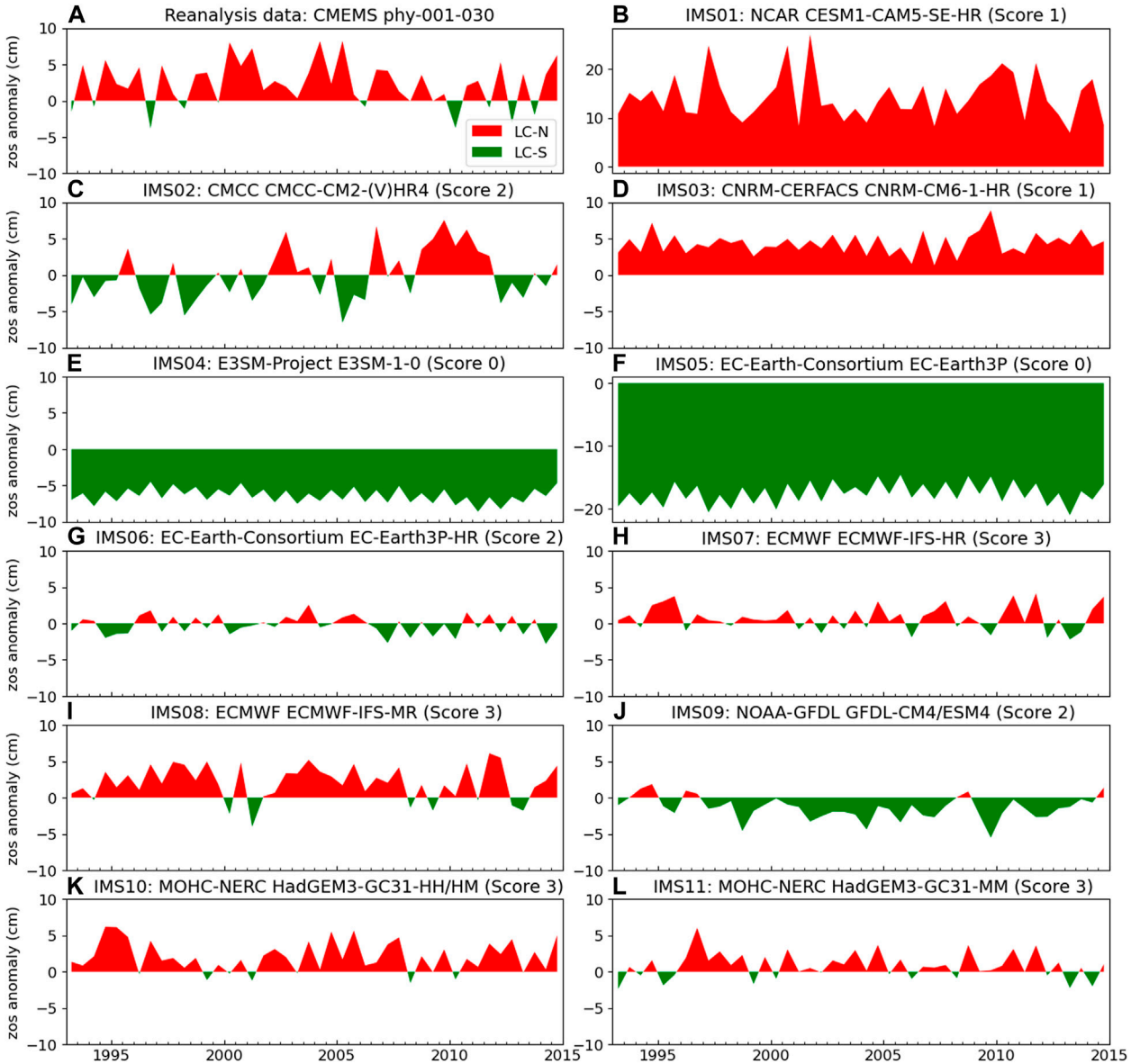

FIGURE 5 | The surface height above geoid (zos) anomaly (**Eq. 5**) of **(A)** reanalysis data, and **(B–L)** enesmble members (i.e, independent model subsets). The title of the reanalysis data shows the data provider name, and product ID. The title of ensemble member shows ensemble member number, modeling group name, model name(s), and ensmble member score.

of mesoscale eddies. For example, EC60to30 is an eddy closure (EC) grid with global parameterization that is not designed to resolve regional spatial phenomena. On the other hand, with a high horizontal resolution, the eddy-permitting grids such as eORCA12, ORCA12, eORCA025, and ORCA025 (**Table 1**) can resolve mesoscale eddies, and do not require ocean eddy flux parameterization. For comparison of high- and standard-resolution grid see also **Figure 2**. On the other hand, the model runs of CESM1-CAM5-SE-HR of NCAR (**Figure 5B**) and CNRM-CM6-1-HR of CNRM-CERFACS (**Figure 5D**) can simulate LC-N, but without a sign difference of zos at the two segments (**Figure 1A**), and accordingly fail in the second metric of oscillating event representation (y_2). These two ensemble members receive a score of one. This score does not indicate that the sea surface height simulation of these models is poor in general, but rather that these models are unsuitable for this target given the problem definition. The ensemble members of CMCC-CM2-(V)HR4 of CMCC (**Figure 5C**), EC-Earth3P-HR of EC-Earth-Consortium (**Figure 5G**), and GFDL-CM4/ESM4 of NOAA-GFDL (**Figure 5J**), pass the second metric, but fail on the oscillating event realism (y_3). These ensemble members show a higher LC-S frequency than LC-N, which is not consistent with the reanalysis data (**Figure 5A**). Accordingly, these three ensemble members receive a score of two. Finally, the

TABLE 2 | Raw data of Loop Current at North (LC-N) and South (LC-S) positions, and their relation to the occurrence of large blooms for reanalysis data, and each ensemble member (i.e., independent model subset, IMS). The ensemble size is the number of model runs per ensemble member, and the reanalysis data has only one realization. Note given Score (y_1, y_2, y_3) the model receives a score from 0 to 3 for Score (0, 0, 0), Score (1,0,0), Score (1, 1, 0), and Score (1, 1, 1), respectively.

IMS	Ensemble	Count		Count LC-N		Count LC-S		Temporal match			RMSE	Score
	Size	LC-N	LC-S	No-Bloom	Large-Bloom	No-Bloom	Large-Bloom	LC-N	LC-S	Total		
Reanalysis data	1	32	12	17	15	12	0	32	12	44	0	3
IMS01	1	44	0	29	15	0	0	32	0	32	13.16	1
IMS02	2	20	24	14	6	15	9	15	7	22	5.48	2
IMS03	4	44	0	29	15	0	0	32	0	32	4.02	1
IMS04	5	0	44	0	0	29	15	0	12	12	9.27	0
IMS05	3	0	44	0	0	29	15	0	12	12	20.16	0
IMS06	3	20	24	13	7	16	8	13	5	18	4.34	2
IMS07	6	31	13	21	10	8	5	24	5	29	3.77	3
IMS08	3	36	8	22	14	7	1	28	4	32	3.87	3
IMS09	3	8	36	6	2	23	13	5	9	14	5.06	2
IMS10	4	35	9	24	11	5	4	26	3	29	3.88	3
IMS11	7	30	14	20	10	9	5	22	4	26	4.08	3

ensemble members that pass the three evolving binary metrics and receive a score of three are ECMWF-IFS-HR of ECMWF (**Figure 5H**), ECMWF-IFS-MR of ECMWF (**Figure 5I**), HadGEM3-GC31-HH/HM of MOHC-NERC (**Figure 5K**), and HadGEM3-GC31-MM of MOHC-NERC (**Figure 5L**). Visual inspection shows that these four ensemble members are qualitatively similar to the reanalysis data (**Figure 5A**) with respect to Loop Current position oscillation.

Using metrics y_4–y_7, we evaluate the predictive performance of these 11 ensemble members with respect to reanalysis data as shown in **Table 2**. According to Maze et al. (2015) there are no red tide blooms for LC-S, and there are either large blooms or no blooms for LC-N. The results of our reanalysis data shown in **Table 2** are consistent with Maze et al. (2015) such that none of the 12 intervals of LC-S has large blooms for the study period. Out of the 32 intervals of LC-N, 15 intervals have large blooms. This indicates that LC-N is a necessarily condition for the large bloom to occur and be sustained. Given the reanalysis data, the LC-S frequency is 0.273 for our 22-year study period, which is comparable to Maze et al. (2015), which is 0.267 for their 15-year study period. The ensemble members IMS07, IMS10, IMS11, and IMS08 have the best agreement with the reanalysis data showing LC-S frequencies (y_4) of 0.295, 0.318, 0.205, and 0.182, respectively. These correspond to the oscillating event frequency errors ($y_{4,err}$) of 0.022, 0.045, −0.068, and −0.091, respectively. Ensemble members that can simulate the oscillation of LC-N and LC-S and have the best temporal match are IMS08, IMS07, IMS10, and IMS11 with temporal match error (y_5) of 27, 34, 34, and 41%, respectively. Given the high-resolution model runs, IMS08, IMS07, IMS10, and IMS11 have the lowest *Karenia brevis* error (y_6) of 0.1, 0.3, 0.3, and 0.3, respectively. IMS09, IMS08, IMS10, IMS03 have the lowest RMSE (y_7) of 3.77, 3.87, 3.88, and 4.02, respectively. While no ensemble member is consistently ranked as the top ensemble member given the four metrics, IMS08 is ranked twice as the top ensemble member given the two metrics y_5 and y_6. Thus, this analysis shows that there is no single ensemble member that consistently perform better with respect to all metrics, and that different ensemble members show both over and underestimation of zos anomaly. These two remarks indicate the importance of using a multi-model ensemble.

Subset Selection

There is generally no specific guideline on the composition of multi-model ensemble of ESMs. While composing information from multiple imperfect ensemble members can be an arbitrarily task, the prescreening step can help find subsets that maintain key features of the problem of interest. We first discuss the two ensembles of SME3210 and SME321X. The ensemble SME3210, which includes both high- and standard-resolution model runs, is generally a flawed ensemble composition, since we know from prior existing knowledge of other studies (Caldwell et al., 2019; Hoch et al., 2020) that standard-resolution ESMs are generally incapable of simulating Loop Current. On the other hand, SME321X is the most straightforward ensemble composition that acknowledges prior information, and includes all high-resolution runs that are capable of simulating Loop Current. We consider SME321X as our reference ensemble. **Figure 6** shows the predictive performance of the four multi-model ensembles. Large red tide blooms do not occur for LC-S given reanalysis data (**Figure 6A**). Comparing reanalysis data (**Figure 6A**) and the multi-model ensembles (**Figures 6B–E**) shows that ensembles based on prior information (i.e., SME321X, SME32XX, and SME3XXX) correspond better to reanalysis data than without accounting for prior information (i.e., SME3210).

Visual examination in **Figure 6** is insufficient to understand the impact of prescreening information (i.e., SME32XX and SME3XXX) in comparison to the reference ensemble SME321X without prescreening information, and qualitative metrics are needed. **Table 3** quantitatively shows that including standard-resolution model runs (i.e., SME3210) results in prediction degradation with respect to the four qualitative metrics (y_4-y_7). As can be calculated from raw data in **Table 3**, SME321X shows relatively good agreement with the reanalysis data with a LC-S frequency (y_4) of 0.227, temporal match error (y_5) of 36%, *Karenia brevis* bloom error (y_6) of 20%, and RMSE (y_7) of 3.71.

FIGURE 6 | Temporal match of large bloom/no bloom with Loop Current positions given the surface height above geoid (zos) anomaly (**Eq. 5**) of **(A)** reanalysis data, and **(B–E)** simulations of four multi-model ensembles. Positive and negative bars indicate Loop Current North (LC-N) and Loop Current South (LC-S), respectively.

Another approach for ensemble composition is to use information from the prescreening step. These are ensembles SME32XX and SME3XXX that exclude the models that cannot represent the oscillation of LC-N and LC-S (y_2). Ensemble SME3XXX only includes model runs with realistic presentation of LC-N and LC-S (y_3). SME32XX shows degraded predictions with respect to the reference ensemble SME321X for all the four quantitative metrics (y_4-y_7). This is not unexpected since members of SME321X show both under and overestimation. For simple model average of model runs with over and underestimation the errors are expected to cancel out (Herger et al., 2018). However, this is not the case for SME3XXX that leverages on most information gained from the prescreening step (i.e., by only including the best members that meet the targets

TABLE 3 | Raw data of Loop Current at North (LC-N) and South (LC-S) positions, and their relation to the occurrence of large blooms simple-average multi-model ensemble (SME). The ensemble size refers to the number of model runs per multi-model ensemble.

SME	Ensemble size	Count		Count LC-N		Count LC-S		Temporal match			RMSE
		LC-N	LC-S	No-Bloom	Large-Bloom	No-Bloom	Large-Bloom	LC-N	LC-S	Total	
Reanalysis data	1	32	12	17	15	12	0	32	12	44	0
SME3210	41	3	41	2	1	27	14	2	11	13	5.13
SME321X	33	34	10	22	12	7	3	25	3	28	3.71
SME32XX	28	23	21	17	6	12	9	17	6	23	3.92
SME3XXX	20	35	9	22	13	7	2	28	5	33	3.68
SMEXXX0	8	0	44	0	0	29	15	0	12	12	13.52

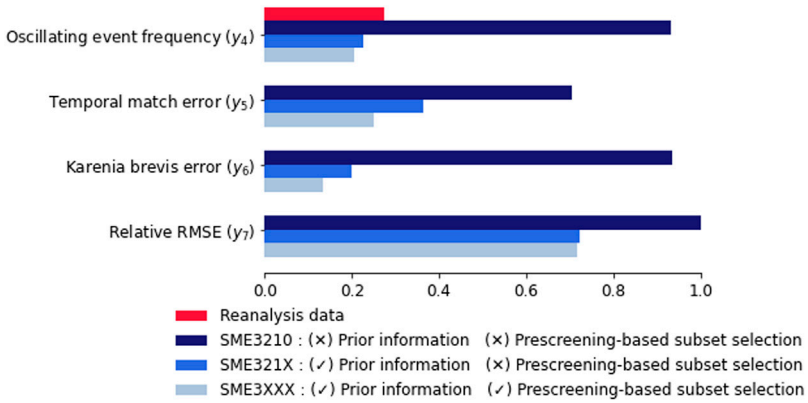

FIGURE 7 | Predictive performance (y_4–y_7) given different ensemble composition criteria.

of interest). SME3XXX shows mixed predictive performance with respect to the reference ensemble showing better performance with respect to temporal match error (y_5) of 25% (versus 36% for the reference ensemble), *Karenia brevis* error (y_6) of 13% (versus 20% for the reference ensemble), and RMSE (y_7) of 3.68 (versus 3.71 for the reference ensemble), but inferior performance with respect to LC-S frequency (y_4) of 0.205 (versus 0.273 and 0.227 for the reanalysis data and reference ensemble, respectively). Yet temporal coverage error is not important for future predictions as discussed in *Discussion*. The relatively good performance of SME3XXX is expected, because this ensemble ensures that members with good performance are only included.

Table 3 additionally shows the case of SMEXXX0, which only considers standard-resolution runs. SMEXXX0 shows a poor predictive performance with respect to all metrics. We present the SMEXXX0 ensemble to illustrate the breakthrough of the HighResMIP of CMIP6. With respect to sea surface height simulation and regional phenomena, our results clearly show the significant improvement of the high-resolution runs of CMIP6 in comparison to the standard-resolution models that are typical to CMIP5.

Ensemble Composition

Our results show that using prior information is important for ensemble composition, and prescreening- based subset selection can be helpful. **Figure 7** summarizes the effect of different ensemble composition criteria. Prior information appears as an important criterion that should be considered as SME3210 has the worst predictive performance with respect to the other ensembles given y_4–y_7. Prescreening-based subset selection seems to relatively improve the predictive performance given y_5–y_7, and slightly degraded performance with respect to y_4. However, pre-screening-based subsect selection has a second conceptual advantage. Given prior information, the first approach of using all the available ensemble members (i.e., SME321X) is a straightforward choice that can result in error cancellation. The second approach of using information from prescreening results in a reduced size ensemble (i.e., SME3XXX), which maintains the most important ensemble characteristics with respect to the problem of interest. While in the first approach we attempt to maintain a more conservative ensemble, with the second approach we create an ensemble with robust ensemble members. Our results suggest that pre-screening based subset section used to substitute or prior to model weighting, which is a subject of a future research.

DISCUSSION

Subset Selection

To find a robust ensemble that improves the predictive performance of ESMs, this article shows the importance of subset selection based on prior information, prescreening, and

process-based evaluation. By evaluating the prescreening-based subset-selection method we deduce two key points as follows. First, we present additional advantages to subset selection that are not well recognized in the literature, which is the importance of subset selection based on process-based evaluation similar to Yun et al. (2017). Eliminating models from an ensemble can be justified if they are known to lack key mechanisms that are indispensable for meaningful climate projections (Weigel et al., 2010). As shown in this study, models that cannot simulate the processes of interest based on a prescreening step can be excluded from the ensemble without degrading the ensemble prediction. Second, the selection of subset-selection method depends on the criteria that are relevant for the application in question (Herger et al., 2018). For example, the process-based evolving binary weights developed in this study is particularly important to eliminate non-representative models. Unlike other subset-selection methods in literature that can be technically challenging to implement, we present a subset-selection method that can be frequently used, as it is intuitive and straightforward to apply. This approach is an addition to subset-selection literature, and is not meant to supersede any of the existing approaches in the literature.

Seasonal Prediction Limitations

Improving seasonal prediction of ESMs to provide useful services for societal decision making is an active research area. Techniques to improve temporal correspondence between predictions and observations at the regional scale is needed for climate services in many sectors such as energy, water resources, agriculture, and health (Manzanas et al., 2019). In this study we used raw outputs without using a postprocessing method to improve temporal correspondence of seasonal prediction. Our results show that the temporal correspondence is not poor, which could be just coincident. Alternatively, this could be attributed to the chosen Loop Current position heuristic with a coarse-temporal-resolution. Accordingly, given a long 6-month period, this is not a month-by-month or season-by-season temporal match, but rather a pseudo-temporal correspondence that captures the general pattern of a dynamic process. Accordingly, using this heuristic relationship, a form of temporal relationship might be possible as long as there is no large drift. If such a temporal correspondence cannot be established for ESMs for Loop Current or other factors that drives the red tide, this would limit the use of the ESMs in terms of providing an early warning system. However, this will not affect the main purpose of the intended model, which is to understand the frequency and trend of red tide under different climate scenarios and estimating the socioeconomic impacts accordingly. If temporal correspondence is required, seasonal prediction of ESMs has generally been possible through statistical and dynamical downscaling methods, and other similar techniques such as pattern scaling and use of analogue (van den Hurk et al., 2018). Alternatives to more complex statistical downscaling techniques to improve temporal correspondence include bias correction (Rozante et al., 2014; Oh and Suh, 2017; Wang et al., 2019), ensemble recalibration (Sansom et al., 2016; Manzanas et al., 2019), and postprocessing techniques such as copula-based postprocessing (Li et al., 2020). For example, to improve temporal correspondence of seasonal prediction, Manzanas (2020) use bias correction and recalibration methods to remove mean prediction bias, and intraseasonal biases from drift (i.e., lead-time dependent bias).

Limitations and Outlook

In this study we present the advantages of subset selection using Loop Current prediction as an example. We show these advantages for the simplest case of using a deterministic analysis, and by considering only historical data. For red tide management purpose, which is to understand the frequency of red tide and the corresponding socioeconomic impacts under different climate scenarios, further steps are needed. First, using CMIP6 model projection data is important to understand the frequency and future trends of red tide under different Shared Socioeconomic Pathways (SSPs) of CMIP6 in which socio-economic scenarios are used to derive emission scenarios without mitigation (i.e., baseline scenario) and with mitigation (i.e., climate polices). Additionally, CMIP6 data can be readily replaced by high resolution data of Coordinated Regional Downscaling Experiment (CORDEX) as soon as they become available. CORDEX which is driven by the CMIP outputs, provides dynamically downscaled climate change experiments for selected regions (Gutowski Jr. et al., 2016; Gutowski et al., 2020). Second, we need to extend our method to a probabilistic framework that considers both historical and future simulations. As historical assessment criteria are not necessarily informative in terms of the quality of model projections of future climate change, identifying the performance metrics that are most relevant to climate projections is one of the biggest challenges in ESM evaluation (Eyring et al., 2019). As the choice of model is a tradeoff between good performance in the past and projected climate change, selecting only the best performing models may limit the spread of projected climate change (Parding et al., 2020). Exploring such trade-off is warranted in a future study in which a probabilistic framework (e.g., Brunner et al., 2019) is needed to account for model performance, model independence, and the representation of future climate projections. Third, it is imperative to consider not only Loop Current, but also other factors that control red tide such as alongshore and offshore wind speed, African Sahara dust, and atmospheric CO2 concentration need to be considered. To account for these different factors simultaneously to predict red tide, machine learning is needed similar to the study of Tonelli et al. (2021) that uses CMIP6 data and machine learning to study marine microbial communities under different climate scenarios. In summary, there are still many further steps needed to develop a probabilistic machine learning framework for regional environmental management of red tide using ESMs of CMIP6 and CORDEX when available. This study is merely a showcase for the potential of using ESMs for red tide management.

CONCLUSION

To improve ensemble performance and to avoid prediction artifacts from including non-representative models, which are models that cannot simulate the process(es) of interest, we introduce a prescreening based subset-selection method. Including non-representative models with both over and underestimation can

result in error cancellation. Whether to include or exclude these non-representative models from the ensemble is a point that requires further investigation through studying model projection. We present a generic subset-selection method to exclude non-representative models based on process-based evolving binary weights. This prescreening step screens each model with respect to its ability to reproduce certain key features. This research emphasizes the importance of ensemble prescreening, which is a topic that is rarely discussed. The presented subset-selection method is flexible as it scores each model given multiple binary criteria. This allows the user to systematically evaluate the sensitivity of the results to different choices of ensemble members. Such flexibility is generally needed to allow the user to understand the implication of ensemble subset selection under different cases (e.g., historic versus historic and future simulations, etc.). Our prescreening-based subset selection method is not meant to replace any of the existing approaches in the literature, but to provide a straightforward and easy-to-implement approach that can be used for many climate services in different sectors as needed.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Elshall, A.S. (2021). Codes for the article of prescreening-based subset selection for improving predictions of Earth system models for regional environmental management of red tide (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5534931.

AUTHOR CONTRIBUTIONS

MY, SK, JH, XY, and YW: motivation and framing for the project. AE, MY, and SK: method development and execution. AE: manuscript development and writing. MY, SK, JH, XY, YW, and MM: manuscript editing and improvements. All authors read and approved the submitted version.

FUNDING

This work is funded by NSF Award #1939994.

ACKNOWLEDGMENTS

We thank two reviewers for their constructive comments that helped to improve the manuscript. We thank Emily Lizotte in the Department of Earth, Ocean, and Atmospheric Science (EOAS) at Florida State University (FSU) for contacting the Florida Fish and Wildlife Conservation Commission (FWC) to obtain the *Karenia brevis* data. We thank FWC for data provision. We are grateful to Maria J. Olascoaga in the Department of Ocean Sciences at University of Miami for our communication regarding *Karenia brevis* data analysis. We thank Sally Gorrie, Emily Lizotte, Mike Stukel, and Jing Yang in EOAS at FSU for their fruitful discussion and suggestions on the project. We dedicate this paper to the memory of Stephen Kish the former professor in EOAS at FSU, who assisted with the motivation and framing for the project.

REFERENCES

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., and Chung, E.-S. (2019). Selection of Multi-Model Ensemble of General Circulation Models for the Simulation of Precipitation and Maximum and Minimum Temperature Based on Spatial Assessment Metrics. *Hydrol. Earth Syst. Sci.* 23, 4803–4824. doi:10.5194/hess-23-4803-2019

Bartók, B., Tobin, I., Vautard, R., Vrac, M., Jin, X., Levavasseur, G., et al. (2019). A Climate Projection Dataset Tailored for the European Energy Sector. *Clim. Serv.* 16, 100138. doi:10.1016/j.cliser.2019.100138

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R. (2019). Quantifying Uncertainty in European Climate Projections Using Combined Performance-independence Weighting. *Environ. Res. Lett.* 14, 124010. doi:10.1088/1748-9326/ab492f

Caldwell, P. M., Mametjanov, A., Tang, Q., Van Roekel, L. P., Golaz, J. C., Lin, W., et al. (2019). The DOE E3SM Coupled Model Version 1: Description and Results at High Resolution. *J. Adv. Model. Earth Syst.* 11, 4095–4146. doi:10.1029/2019MS001870

Cannon, A. J. (2015). Selecting GCM Scenarios that Span the Range of Changes in a Multimodel Ensemble: Application to CMIP5 Climate Extremes Indices*. *J. Clim.* 28, 1260–1267. doi:10.1175/JCLI-D-14-00636.1

Chandler, R. E. (2013). Exploiting Strength, Discounting Weakness: Combining Information from Multiple Climate Simulators. *Phil. Trans. R. Soc. A.* 371, 20120388. doi:10.1098/rsta.2012.0388

Chang, P., Zhang, S., Danabasoglu, G., Yeager, S. G., Fu, H., Wang, H., et al. (2020). An Unprecedented Set of High-Resolution Earth System Simulations for Understanding Multiscale Interactions in Climate Variability and Change. *J. Adv. Model. Earth Syst.* 12, e2020MS002298. doi:10.1029/2020MS002298

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., et al. (2019). Global Mean Climate and Main Patterns of Variability in the CMCC-CM2 Coupled Model. *J. Adv. Model. Earth Syst.* 11, 185–209. doi:10.1029/2018MS001369

Christensen, J., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M. (2010). Weight Assignment in Regional Climate Models. *Clim. Res.* 44, 179–194. doi:10.3354/cr00916

DelSole, T., Nattala, J., and Tippett, M. K. (2014). Skill Improvement from Increased Ensemble Size and Model Diversity. *Geophys. Res. Lett.* 41, 7331–7342. doi:10.1002/2014GL060133

DelSole, T., Yang, X., and Tippett, M. K. (2013). Is Unequal Weighting Significantly Better Than Equal Weighting for Multi-Model Forecasting? *Q.J.R. Meteorol. Soc.* 139, 176–183. doi:10.1002/qj.1961

Drévillon, M., Régnier, C., Lellouche, J.-M., Garric, G., and Bricaud, C. (2018). Quality Information Document For Global Ocean Reanalysis Products Global-Reanalysis-Phy-001-030. 48.

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., et al. (2015). The Global Gridded Crop Model Intercomparison: Data and Modeling Protocols for Phase 1 (v1.0). *Geosci. Model. Dev.* 8, 261–277. doi:10.5194/gmd-8-261-2015

Elshall, A. S. (2021). Codes for the Manuscript of Prescreening-Based Subset Selection for Improving Predictions of Earth System Models for Regional Environmental Management of Red Tide. *Zenodo*. doi:10.5281/zenodo.5534931

Elshall, A. S. (2020). Sea Surface Height above Geoid: AVISO Altimetry Data versus ESM Simulations of Loop Current. Available at: https://youtu.be/9Guohel814w (Accessed May 19, 2021).

Evans, J. P., Ji, F., Abramowitz, G., and Ekström, M. (2013). Optimally Choosing Small Ensemble Members to Produce Robust Climate Simulations. *Environ. Res. Lett.* 8, 044050. doi:10.1088/1748-9326/8/4/044050

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., et al. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization. *Geosci. Model. Dev.* 9, 1937–1958. doi:10.5194/gmd-9-1937-2016

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking Climate Model Evaluation to the Next Level. *Nat. Clim Change* 9, 102–110. doi:10.1038/s41558-018-0355-y

Farjad, B., Gupta, A., Sartipizadeh, H., and Cannon, A. J. (2019). A Novel Approach for Selecting Extreme Climate Change Scenarios for Climate Change Impact Studies. *Sci. Total Environ.* 678, 476–485. doi:10.1016/j.scitotenv.2019.04.218

Fernandez, E., and Lellouche, J. M. (2018). Product User Manual For The Global Ocean Physical Reanalysis Product Global_Reanalysis_ Phy_001_030. 15.

FWRI (2020). HAB Monitoring Database. *Fla. Fish Wildl. Conservation Comm.* Available at: http://myfwc.com/research/redtide/monitoring/database/ (Accessed December 23, 2020).

Golaz, J.-C., Caldwell, P. M., Roekel, L. P. V., Petersen, M. R., Tang, Q., Wolfe, J. D., et al. (2019). The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution. *J. Adv. Model. Earth Syst.* 11, 2089–2129. doi:10.1029/2018MS001603

Gutowski Jr., W. J., Jr., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., et al. (2016). WCRP COordinated Regional Downscaling EXperiment (CORDEX): a Diagnostic MIP for CMIP6. *Geosci. Model. Dev.* 9, 4087–4095. doi:10.5194/gmd-9-4087-2016

Gutowski, W. J., Ullrich, P. A., Hall, A., Leung, L. R., O'Brien, T. A., Patricola, C. M., et al. (2020). The Ongoing Need for High-Resolution Regional Climate Models: Process Understanding and Stakeholder Information. *Bull. Am. Meteorol. Soc.* 101, E664–E683. doi:10.1175/BAMS-D-19-0113.1

Haarsma, R., Acosta, M., Bakhshi, R., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., et al. (2020). HighResMIP Versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR - Description, Model Computational Performance and Basic Validation. *Geosci. Model. Dev.* 13, 3507–3527. doi:10.5194/gmd-13-3507-2020

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., et al. (2016). High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model. Dev.* 9, 4185–4208. doi:10.5194/gmd-9-4185-2016

Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J. (2015). Weighting Climate Model Ensembles for Mean and Variance Estimates. *Clim. Dyn.* 45, 3169–3181. doi:10.1007/s00382-015-2531-3

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and Performance of GFDL's CM4.0 Climate Model. *J. Adv. Model. Earth Syst.* 11, 3691–3727. doi:10.1029/2019MS001829

Hemri, S., Bhend, J., Liniger, M. A., Manzanas, R., Siegert, S., Stephenson, D. B., et al. (2020). How to Create an Operational Multi-Model of Seasonal Forecasts? *Clim. Dyn.* 55, 1141–1157. doi:10.1007/s00382-020-05314-2

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M. (2018). Selecting a Climate Model Subset to Optimise Key Ensemble Properties. *Earth Syst. Dynam.* 9, 135–151. doi:10.5194/esd-9-135-2018

Hewitt, H. T., Bell, M. J., Chassignet, E. P., Czaja, A., Ferreira, D., Griffies, S. M., et al. (2017). Will High-Resolution Global Ocean Models Benefit Coupled Predictions on Short-Range to Climate Timescales? *Ocean Model.* 120, 120–136. doi:10.1016/j.ocemod.2017.11.002

Hoch, K. E., Petersen, M. R., Brus, S. R., Engwirda, D., Roberts, A. F., Rosa, K. L., et al. (2020). MPAS-Ocean Simulation Quality for Variable-Resolution North American Coastal Meshes. *J. Adv. Model. Earth Syst.* 12, e2019MS001848. doi:10.1029/2019MS001848

Horsburgh, J. S., Hooper, R. P., Bales, J., Hedstrom, M., Imker, H. J., Lehnert, K. A., et al. (2020). Assessing the State of Research Data Publication in Hydrology: A Perspective from the Consortium of Universities for the Advancement of Hydrologic Science, Incorporated. *WIREs Water* 7, e1422. doi:10.1002/wat2.1422

Hussain, M., Yusof, K. W., Mustafa, M. R. U., Mahmood, R., and Jia, S. (2018). Evaluation of CMIP5 Models for Projection of Future Precipitation Change in Bornean Tropical Rainforests. *Theor. Appl. Climatol* 134, 423–440. doi:10.1007/s00704-017-2284-5

Jagannathan, K., Jones, A. D., and Kerr, A. C. (2020). Implications of Climate Model Selection for Projections of Decision-Relevant Metrics: A Case Study of Chill Hours in California. *Clim. Serv.* 18, 100154. doi:10.1016/j.cliser.2020.100154

Jiang, Z., Li, W., Xu, J., and Li, L. (2015). Extreme Precipitation Indices over China in CMIP5 Models. Part I: Model Evaluation. *J. Clim.* 28, 8603–8619. doi:10.1175/JCLI-D-15-0099.1

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in Combining Projections from Multiple Climate Models. *J. Clim.* 23, 2739–2758. doi:10.1175/2009JCLI3361.1

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V. (2017). A Climate Model Projection Weighting Scheme Accounting for Performance and Interdependence. *Geophys. Res. Lett.* 44, 1909–1918. doi:10.1002/2016GL072012

Knutti, R. (2010). The End of Model Democracy? *Climatic Change* 102, 395–404. doi:10.1007/s10584-010-9800-2

Leduc, M., Laprise, R., de Elía, R., and Šeparović, L. (2016). Is Institutional Democracy a Good Proxy for Model Independence? *J. Clim.* 29, 8301–8316. doi:10.1175/JCLI-D-15-0761.1

Li, M., Jin, H., and Brown, J. N. (2020). Making the Output of Seasonal Climate Models More Palatable to Agriculture: A Copula-Based Postprocessing Method. *J. Appl. Meteorology Climatology* 59, 497–515. doi:10.1175/JAMC-D-19-0093.1

Liu, Y., Weisberg, R. H., Lenes, J. M., Zheng, L., Hubbard, K., and Walsh, J. J. (2016). Offshore Forcing on the "Pressure point" of the West Florida Shelf: Anomalous Upwelling and its Influence on Harmful Algal Blooms. *J. Geophys. Res. Oceans* 121, 5501–5515. doi:10.1002/2016JC011938

Magaña, H. A., and Villareal, T. A. (2006). The Effect of Environmental Factors on the Growth Rate of Karenia Brevis (Davis) G. Hansen and Moestrup. *Harmful Algae* 5, 192–198. doi:10.1016/j.hal.2005.07.003

Manzanas, R. (2020). Assessment of Model Drifts in Seasonal Forecasting: Sensitivity to Ensemble Size and Implications for Bias Correction. *J. Adv. Model. Earth Syst.* 12, e2019MS001751. doi:10.1029/2019MS001751

Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., et al. (2019). Bias Adjustment and Ensemble Recalibration Methods for Seasonal Forecasting: a Comprehensive Intercomparison Using the C3S Dataset. *Clim. Dyn.* 53, 1287–1305. doi:10.1007/s00382-019-04640-4

Maze, G., Olascoaga, M. J., and Brand, L. (2015). Historical Analysis of Environmental Conditions during Florida Red Tide. *Harmful Algae* 50, 1–7. doi:10.1016/j.hal.2015.10.003

McSweeney, C. F., Jones, R. G., Lee, R. W., and Rowell, D. P. (2015). Selecting CMIP5 GCMs for Downscaling over Multiple Regions. *Clim. Dyn.* 44, 3237–3260. doi:10.1007/s00382-014-2418-8

Mendlik, T., and Gobiet, A. (2016). Selecting Climate Simulations for Impact Studies Based on Multivariate Patterns of Climate Change. *Climatic Change* 135, 381–393. doi:10.1007/s10584-015-1582-0

Oh, S.-G., and Suh, M.-S. (2017). Comparison of Projection Skills of Deterministic Ensemble Methods Using Pseudo-simulation Data Generated from Multivariate Gaussian Distribution. *Theor. Appl. Climatol* 129, 243–262. doi:10.1007/s00704-016-1782-1

Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., et al. (2020). GCMeval - an Interactive Tool for Evaluation and Selection of Climate Model Ensembles. *Clim. Serv.* 18, 100167. doi:10.1016/j.cliser.2020.100167

Pennell, C., and Reichler, T. (2011). On the Effective Number of Climate Models. *J. Clim.* 24, 2358–2367. doi:10.1175/2010JCLI3814.1

Perkins, S. (2019). Inner Workings: Ramping up the Fight against Florida's Red Tides. *Proc. Natl. Acad. Sci. USA* 116, 6510–6512. doi:10.1073/pnas.1902219116

Roberts, C. D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., and Keeley, S. P. E. (2018). Climate Model Configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS Cycle 43r1) for HighResMIP. *Geosci. Model. Dev.* 11, 3681–3712. doi:10.5194/gmd-11-3681-2018

Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., et al. (2019). Description of the Resolution Hierarchy of the Global Coupled HadGEM3-GC3.1 Model as Used in CMIP6 HighResMIP Experiments. *Geosci. Model. Dev.* 12, 4999–5028. doi:10.5194/gmd-12-4999-2019

Ross, A. C., and Najjar, R. G. (2019). Evaluation of Methods for Selecting Climate Models to Simulate Future Hydrological Change. *Climatic Change* 157, 407–428. doi:10.1007/s10584-019-02512-8

Rozante, J. R., Moreira, D. S., Godoy, R. C. M., and Fernandes, A. A. (2014). Multi-model Ensemble: Technique and Validation. *Geosci. Model. Dev.* 7, 2333–2343. doi:10.5194/gmd-7-2333-2014

Samouly, A. A., Luong, C. N., Li, Z., Smith, S., Baetz, B., and Ghaith, M. (2018). Performance of Multi-Model Ensembles for the Simulation of Temperature Variability over Ontario, Canada. *Environ. Earth Sci.* 77, 524. doi:10.1007/s12665-018-7701-2

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *J. Clim.* 28, 5171–5194. doi:10.1175/JCLI-D-14-00362.1

Sanderson, B. M., Wehner, M., and Knutti, R. (2017). Skill and independence Weighting for Multi-Model Assessments. *Geosci. Model. Dev.* 10, 2379–2395. doi:10.5194/gmd-10-2379-2017

Sansom, P. G., Ferro, C. A. T., Stephenson, D. B., Goddard, L., and Mason, S. J. (2016). Best Practices for Postprocessing Ensemble Climate Forecasts. Part I: Selecting Appropriate Recalibration Methods. *J. Clim.* 29, 7247–7264. doi:10.1175/JCLI-D-15-0868.1

Sørland, S. L., Fischer, A. M., Kotlarski, S., Künsch, H. R., Liniger, M. A., Rajczak, J., et al. (2020). CH2018 - National Climate Scenarios for Switzerland: How to Construct Consistent Multi-Model Projections from Ensembles of Opportunity. *Clim. Serv.* 20, 100196. doi:10.1016/j.cliser.2020.100196

Sturges, W., and Evans, J. C. (1983). On the Variability of the Loop Current in the Gulf of Mexico. *J. Mar. Res.* 41, 639–653. doi:10.1357/002224083788520487

Szabó-Takács, B., Farda, A., Skalák, P., and Meitner, J. (2019). Influence of Bias Correction Methods on Simulated Köppen–Geiger Climate Zones in Europe. *Climate* 7, 18. doi:10.3390/cli7020018

Tonelli, M., Signori, C. N., Bendia, A., Neiva, J., Ferrero, B., Pellizari, V., et al. (2021). Climate Projections for the Southern Ocean Reveal Impacts in the Marine Microbial Communities Following Increases in Sea Surface Temperature. *Front. Mar. Sci.* 8, 636226. doi:10.3389/fmars.2021.636226

van den Hurk, B., Hewitt, C., Jacob, D., Bessembinder, J., Doblas-Reyes, F., and Döscher, R. (2018). The Match between Climate Services Demands and Earth System Models Supplies. *Clim. Serv.* 12, 59–63. doi:10.1016/j.cliser.2018.11.002

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 DECK Experiments with CNRM-CM6-1. *J. Adv. Model. Earth Syst.* 11, 2177–2213. doi:10.1029/2019MS001683

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P. J., et al. (2018). Multimodel Ensembles Improve Predictions of Crop-Environment-Management Interactions. *Glob. Change Biol.* 24, 5072–5083. doi:10.1111/gcb.14411

Wang, H.-M., Chen, J., Xu, C.-Y., Chen, H., Guo, S., Xie, P., et al. (2019). Does the Weighting of Climate Simulations Result in a Better Quantification of Hydrological Impacts? *Hydrol. Earth Syst. Sci.* 23, 4033–4050. doi:10.5194/hess-23-4033-2019

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C. (2010). Risks of Model Weighting in Multimodel Climate Projections. *J. Clim.* 23, 4175–4191. doi:10.1175/2010JCLI3594.1

Weisberg, R. H., Liu, Y., Lembke, C., Hu, C., Hubbard, K., and Garrett, M. (2019). The Coastal Ocean Circulation Influence on the 2018 West Florida Shelf K . Brevis Red Tide Bloom. *J. Geophys. Res. Oceans* 124, 2501–2512. doi:10.1029/2018JC014887

Weisberg, R. H., Zheng, L., Liu, Y., Lembke, C., Lenes, J. M., and Walsh, J. J. (2014). Why No Red Tide Was Observed on the West Florida Continental Shelf in 2010. *Harmful Algae* 38, 119–126. doi:10.1016/j.hal.2014.04.010

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Xuan, W., Ma, C., Kang, L., Gu, H., Pan, S., and Xu, Y.-P. (2017). Evaluating Historical Simulations of CMIP5 GCMs for Key Climatic Variables in Zhejiang Province, China. *Theor. Appl. Climatol* 128, 207–222. doi:10.1007/s00704-015-1704-7

Yun, K., Hsiao, J., Jung, M.-P., Choi, I.-T., Glenn, D. M., Shim, K.-M., et al. (2017). Can a Multi-Model Ensemble Improve Phenology Predictions for Climate Change Studies? *Ecol. Model.* 362, 54–64. doi:10.1016/j.ecolmodel.2017.08.003

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., et al. (2018). Future Climate Risk from Compound Events. *Nat. Clim Change* 8, 469–477. doi:10.1038/s41558-018-0156-3

Author Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Earth Science

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/feart.2022.867379

Complex or Simple—Does a Model Have to be One or the Other?

Rui Hugman[1]* *and John Doherty*[1,2]

[1]National Centre for Groundwater Research and Training, College of Science and Engineering, Flinders University, Adelaide, SA, Australia, [2]Watermark Numerical Computing, Brisbane, QLD, Australia

The primary tasks of decision-support modelling are to quantify and reduce the uncertainties of decision-critical model predictions. Reduction of predictive uncertainty requires assimilation of information. Generally, this information resides in two places: 1) expert knowledge emerging from site characterization and 2) field measurements of present and historical system behavior. The former is uncertain and should therefore be expressed stochastically in a model. The range of parameter and predictive possibilities can then be constrained through history-matching. Implementation of these Bayesian principles places conflicting demands on the level of model structural complexity. A high level of structural complexity can facilitate expression of expert knowledge by establishing model details that are recognizable by site experts, and through supporting model parameters that bear a close relationship to real-world hydraulic properties. However, such models often run slowly and are numerically delicate; history-matching therefore becomes difficult or impossible. In contrast, if endowed with enough parameters, structurally simple models facilitate the achievement of a good fit between model outputs and field measurements. However, the values with which parameters are endowed may bear a looser relationship with real-world properties and are therefore less receptive to information born of expert knowledge. The model design process is therefore one of compromise. In this paper we describe a methodology that reduces the cost of compromise by allowing expert knowledge of system properties to inform the parameters of a structurally simple model. The methodology requires the use of a complementary model of strategic, but not excessive, structural complexity that is stochastic, fast-running and requires no history-matching. We demonstrate the approach using a real-world case in which modelling is used to support management of a stressed coastal aquifer. We empirically validate the approach using a synthetic model.

Keywords: optimisation, expert knowledge, modelling, groundwater, decision-support, complexity, data-assimilation, uncertainty

OPEN ACCESS

Edited by:
Anneli Guthke,
University of Stuttgart, Germany

Reviewed by:
Laura Foglia,
University of California, Davis, United States
Ty Ferre,
University of Arizona, United States

***Correspondence:**
Rui Hugman
rui.hugman@flinders.edu.au

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *01 February 2022*
Accepted: *12 April 2022*
Published: *09 May 2022*

Citation:
Hugman R and Doherty J (2022) Complex or Simple—Does a Model Have to be One or the Other? Front. Earth Sci. 10:867379. doi: 10.3389/feart.2022.867379

1 INTRODUCTION

Groundwater systems are complex. Our ability to characterise this complexity is limited. It is not possible to calculate the exact outcomes of a proposed groundwater management action, as they depend on too many unknown system details. However, it is often possible to characterize them probabilistically. Hence, forecasts of future system behaviour can be accompanied by estimates of their uncertainties. This is essential to risk-based decision-making (Freeze et al., 1990; Doherty and Moore, 2020).

In principle, a model that represents a high level of system and process detail (referred to as a "complex" model herein) supports 1) quantification of the uncertainties of management-critical predictions through inclusion of all facets of system behaviour on which those predictions depend, and 2) reduction of predictive uncertainty by employing parameter sets that allow it to also replicate the historical behaviour of the system. Complex models are generally physically based; their numerical details attempt to mimic what is known of reality. Ideally, field or laboratory measurements of system properties can therefore directly inform the values of their parameters and the prior probability distributions of these parameters. The link between expert knowledge and model parameterization is therefore direct.

A major problem with complex models, however, is that they are generally characterized by long run times, and often react badly to parameters sets that embody stochastic expression of hydraulic property heterogeneity. They make poor partners for software such as PEST and PEST++ which can implement the Bayesian imperative of constraining model parameters so that the model's outputs can replicate historical system states, thereby reducing uncertainties associated with its evaluation of future system states.

Assimilation of historical information can be rendered more tractable by use of a fast-running, numerically-stable, structurally simple model. Such a model can be physically-based; however, its design philosophy is to represent the repercussions of system detail on model outputs, rather than explicitly representing the detail itself. The link between parameters of a structurally simple model and real-world hydraulic properties may therefore become more abstract. However, if a strategically-designed, structurally simple model is sufficiently highly parameterized, programs such as PEST and PEST++ can easily find sets of parameters that allow it to replicate historical system behaviour well. This has the potential to reduce the uncertainties of some decision-critical predictions. However, this reduction in uncertainty must be balanced against the need to inflate prior parameter uncertainties because of their looser links with recognizable hydraulic properties, and for the need for any bias that is induced through history-matching of a simplified model to be included in the posterior uncertainties of its predictions (Doherty and Simmons, 2013).

Predictive bias can be reduced by strategic design of a simplified model; it can also be reduced through strategic design of the history-matching process. See White et al. (2014). The problem of assigning prior probability distributions to parameters of a structurally simple model, especially those that may be somewhat abstract because they are used to characterise its boundary conditions, has (to the authors' knowledge) received little, if any, attention in the modelling literature. A simple response to this problem is the assignment of generous prior uncertainties to pertinent parameters, and/or the use of noninformative priors and/or uniform prior probability distributions. While this strategy avoids the under-estimation of predictive uncertainty, it may erode the decision-support potential of numerical modelling by excluding information born of expert knowledge from the modelling process.

This paper demonstrates a methodology that can be used to characterise the prior probability distributions of some parameters that are assigned to a structurally simple model, in particular those that characterise one of its boundary conditions. This methodology relies on the use of a complementary model of greater structural complexity that embodies explicit (yet still simplified) representation of geometry and processes that are replaced by the simple model's boundary condition. Running of the complementary physically-based model is computationally expensive; however relatively few runs of this model are required.

We demonstrate this approach to assessment of boundary parameter prior stochasticity using a model that was built to support the management of a highly stressed coastal aquifer at Vale do Lobo, Portugal. A structurally simple, but parametrically complex, constant-density model represents the onshore portion of the aquifer. A Cauchy (i.e., "general head") boundary condition dispenses with the need for this model to simulate groundwater process in that part of the aquifer that extends a considerable distance offshore. Heads and flows computed by a structurally and parametrically stochastic, variable-density model are used to assign heads and conductances to the boundary condition of the onshore model.

This paper is organised as follows. **Section 2** describes the hydrogeology of the Vale do Lobo area, the problems that beset it, and a model that was built to support improved management of the area. However, the discussion is brief as a more complete description can be found elsewhere (Hugman et al., 2021). **Section 3** describes the methodology that was developed for prior stochastic parameterisation of the coastal boundary condition of the Vale do Lobo management model; implementation details of this methodology are also provided. Reference is made to **Supplementary Material** wherein the methodology is verified using a synthetic case whose details are inspired by Vale do Lobo hydrology. **Section 4** describes stochastic history-matching and deployment of the Vale do Lobo management model. Discussion and conclusions follow in **Section 5**.

2 VALE DO LOBO

This section describes modelling undertaken for an aquifer system in southern Portugal which faces the threat of sea water intrusion because of excessive groundwater extraction for irrigation. The study area is located to the west of Faro, capital of the Algarve province of Portugal (**Figure 1**).

2.1 Conceptual Model

The Vale do Lobo (VL) subsystem of the Campina de Faro aquifer occupies an area of about 32 km^2. The boundary which separates it from the Campina de Faro subsystem to its east is a management rather than a hydrogeological boundary; it runs roughly perpendicular to topographic contours. To the northwest, the VL subsystem boundary is defined by the Carcavai fault zone. Low permeability marls outcrop along the northern boundary; this creates a barrier between the VL system and a highly permeable karstic aquifer further to the north.

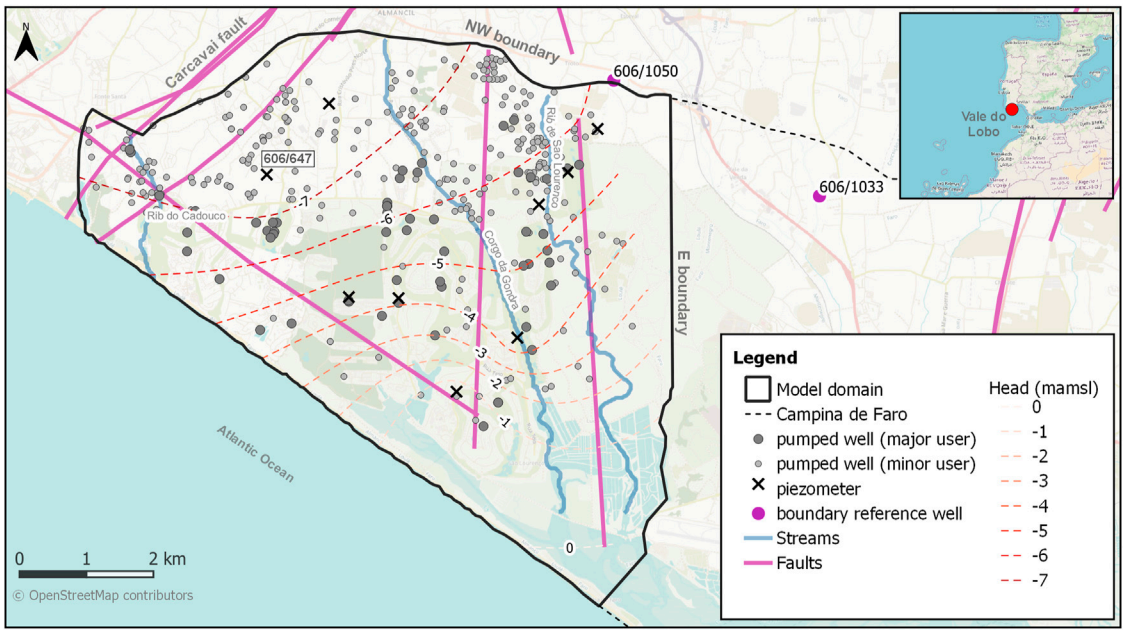

FIGURE 1 | Location and main hydrogeological features of the Vale do Lobo aquifer system. The locations of groundwater abstraction wells, piezometers, and contoured average measured hydraulic heads during 2020 are also depicted.

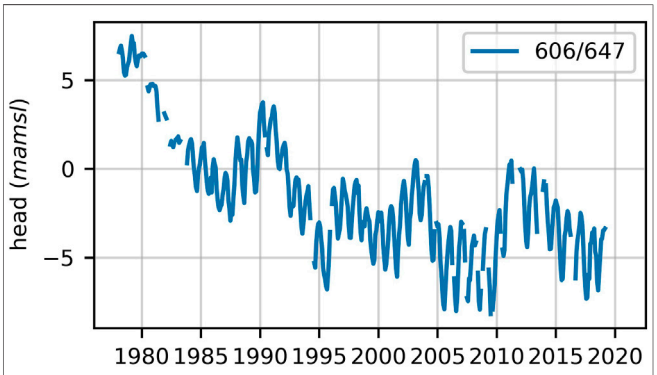

FIGURE 2 | Time-series of hydraulic heads measured in piezometer 606/647. The location of this piezometer is labelled in **Figure 1**.

The VL subsystem is comprised of two aquifers. An upper, phreatic aquifer and a lower, semi-confined aquifer formed of calcareous sandstones and limestones. Nearby and offshore drilling suggests that the base of the semi-confined aquifer reaches a depth of 350 m below mean sea level at the coast. A clay aquitard, typically about 10 m thick, separates the two aquifer formations. The aquifer formations dip towards the coast at about 4°. Fresh water that flows seaward through the deep aquifer emerges in the sea at an unknown distance offshore. However, the offshore perseverance of the overlying aquitard, together with all other details of offshore freshwater flow and the geology which controls it, are unknown.

Figure 2 displays heads measured in a representative well (see the labelled location in **Figure 1**). The well is open to the deep aquifer. This is the longest record of piezometric heads available in the VL area. It exhibits a gradual decline from the late 1970s to the late 1990s, at which time groundwater levels appear to reach a new equilibrium. Groundwater heads are, on average, below sea level in the deep aquifer. The lowest levels are in the central and north-western corner of the system. Groundwater appears to flow towards this area of depressed heads from all system boundaries, including from adjacent aquifers and from the coast.

The local regulatory agency estimates that total groundwater use during 2019 was around 6.45 Mm^3. Most of this water was extracted from the deep aquifer. Diffuse recharge to the phreatic aquifer is estimated to be about 3.46 Mm^3/yr (i.e., 108 mm/yr). The proportion of this water which reaches the deep aquifer is unknown. The deep aquifer probably receives most of its recharge laterally through its northern and/or eastern boundaries.

Chloride concentrations measured at observation wells within the deep aquifer are mostly below 300 mg/L. However, these measurements, as well as anecdotal evidence pertaining to the quality of water extracted from production wells, suggest increasing salinities over time at some locations. There is some evidence that dissolution of evaporites that are disseminated through sediments which comprise the deep aquifer may be partly responsible for increased chloride concentrations. It appears, therefore, that VL groundwaters have not yet suffered a serious decline in quality because of seawater intrusion. Nevertheless, hydraulic head data makes it clear that its occurrence is inevitable.

2.2 The Problem

Continued extraction of water from the VL system at current rates is unsustainable. In the long term, it will result in serious degradation of

aquifer water quality through seawater intrusion. In the short term, it will occasion noncompliance with an EU Water Framework Directive (WFD) which specifies that average abstraction from a groundwater system must be lower than 90% of average annual recharge. Local authorities are looking at a number of ways to comply with this directive. These include the impositions of limits on existing water use licenses and/or implementation of managed aquifer recharge (MAR). However, technical and legal obstacles presently impede both of these options.

Modelling that is described herein serves to:

1) Estimate the historical components of the VL water budget, along with their uncertainties. These include calculation of groundwater withdrawals where records are unavailable, and quantification of lateral inflows from neighbouring systems. Outcomes of these analyses provide an estimate of the minimum change in water budget required to comply with WFD requirements.
2) Explore management strategies that maximize groundwater use while mitigating the potential for seawater intrusion. Outcomes of these analyses are intended to identify maximum allocation limits for existing groundwater users.

2.3 The Numerical Model

2.3.1 Model Structure

A numerical model was developed in order to explore management options for the VL subsystem. The model is a composite of a constant-density MODFLOW 6 (Langevin et al., 2021) model and eight LUMPREM (Doherty, 2020) models. LUMPREM stands for "lumped parameter recharge model"; LUMPREM models are used to compute irrigation demand, and hence groundwater withdrawal rates employed by the MODFLOW 6 model.

History-matching is undertaken over the period October 2000–October 2020. Unfortunately, records of historical abstraction over this period are incomplete. However, they comprise the only direct measurement of any aspect of the VL subsystem water balance. Other aspects of the water balance must be inferred from the system's response to this extraction. A single LUMPREM model is used to calculate irrigation demand for each groundwater user group. LUMPREM models are calibrated against measured extraction rates where these are available.

The groundwater flow model employs a single layer. This represents the deeper semi-confined aquifer. Its landward boundaries coincide with that of the VL system described above. To its southwest, the coastline bounds the model domain. Aquifer top and bottom elevations are interpolated from onshore and offshore borehole logs and topographic elevations of outcrops.

The bottom boundary of the VL groundwater model is a no-flow boundary. A general-head boundary (GHB) is ascribed to its top. This simulates the connection of the lower aquifer with the overlying phreatic aquifer. Temporal and spatial variation in GHB heads at each cell are calculated by applying a linear function of surface elevation to measured time series of heads in the phreatic aquifer. Constants of this function are adjustable during history-matching.

All lateral boundaries of the VL model are also simulated using GHBs. For the north-western and eastern boundaries of the VL model, GHB heads vary with time while conductance is time-invariant. Time-varying heads are obtained from time-series of measured heads at representative piezometers; these are spatially interpolated along the boundaries.

2.3.2 Model Parameters

An array of 565 pilot points, distributed throughout the model domain, is employed to represent spatial variation of each of hydraulic conductivity, specific storage and conductance of the aquitard which separates the deep aquifer from the upper aquifer.

Pilot points are also placed along all model boundaries. These are used to represent spatial variation of conductance along these boundaries. They are also used to parameterise spatial variation of heads along these boundaries. However, as stated above, time-varying heads at certain locations along the north-western and eastern model boundaries are informed from heads measured in nearby wells. Hugman et al. (2021) for details.

As stated above, the coastal boundary of the VL model is represented by a continuous, coast-coincident GHB. Heads and conductances along this boundary are parameterised using pilot points. Statistical characterisation of these heads and conductances is discussed next.

Statistical characterisation of model parameters is required for two reasons. It is used in formulation of a regularisation scheme which seeks parameter uniqueness through model calibration. It is also used in calculation of an ensemble of parameter realisations which comprise samples of a parameter probability distribution. Prior means and probability density functions for hydraulic conductivity, specific storage and aquitard conductance can be assigned on the basis of expert knowledge. Expert knowledge can also be employed in assigning prior means and standard deviations to pilot points which represent heads along the north-western and eastern boundaries. Conductances along these boundaries are not well informed by expert knowledge; hence large prior uncertainties must be assumed. Fortunately, they are moderately well constrained through history matching.

The assignment of prior means and variances/covariances to coastal GHB boundary heads and conductances is a different matter. These parameters are somewhat abstract, as they are numerical surrogates for complex off-shore processes whose details are poorly known. However, there are physical and geometric constraints on these offshore processes that are set by local geology and the mechanics of density-dependent groundwater flow. Conceptually, these can be used to place constraints on the values of coastal GHB boundary parameters. It is to this subject that we now turn.

3 THE COASTAL BOUNDARY CONDITION

3.1 General

Confined and semi-confined coastal aquifers can convey fresh groundwater 10 s or even 100 s of kilometres beyond the coast

(Post et al., 2013; Knight et al., 2018). In confined systems, fresh groundwater discharges where the permeable formation outcrops on the seabed. In semi-confined systems, freshwater percolates through the overlying low permeability formation, reaching the sea as diffuse discharge along its bed. The offshore extent of freshwater in the latter case is determined by the hydraulic characteristics of the system and the stresses to which it has been subjected.

Data on offshore and under-sea conditions is rarely available. Stresses that determine the nature of present-day freshwater outflow can date back thousands of years. Explicit (necessarily stochastic) representation of these conditions in the same model as that which is used for aquifer management may require extension of the model grid tens of kilometres offshore; it may also require a long "wind-up" time in which the system is subjected to uncertain stresses. In conducting this modelling, the effects of density differences cannot be ignored.

For exploration of VL aquifer management options, we represent offshore conditions implicitly using a GHB. Meanwhile, constant density, freshwater conditions are assumed to prevail under land. This is in accordance with current assessments of the VL subsystem. A single-layer, fast-running model can then be used for assimilation of onshore data and probabilistic exploration of management options.

3.2 Conditions at the Boundary

Water enters or leaves a GHB in proportion to the difference between the head assigned to the boundary and that calculated by the model for the cell which the boundary occupies. The constant of proportionality is the boundary's conductance. For a coastal boundary condition, the head ascribed to the GHB represents a head at some point offshore. GHB conductance represents the resistance to flow between the model cell (i.e., the coast) and that point. Conceptually, the coastal GHB can provide simplified numerical representation of the hydraulic linkage between the model and an offshore portion of the same system. However, this simplification ignores the effects of changes in offshore storage while assuming that the dynamics of offshore flow do not change drastically throughout the simulated period.

Use of a GHB to represent a coastal system which is at equilibrium resembles an approach recommended by Lu et al. (2015). However real-world, coastal aquifer systems are rarely at equilibrium. Semi-confined coastal aquifers in particular can take a very long time to reach a new equilibrium because of the relatively slow movement of the fresh-saltwater interface, and the potentially large volume of water stored offshore (Knight et al., 2018). Matters are further complicated when the purpose of modelling is to assess the long-term consequences to a system of a change in onshore pressures.

A GHB that attempts to alleviate the need to simulate offshore conditions must be capable of representing the range of conditions to which the onshore part of the aquifer is subjected. In the current case these conditions are 1) those which prevailed prior to development, 2) those which prevailed in the last few decades when water extraction induced a landward hydraulic gradient, and 3) those which will prevail in the future when groundwater withdrawals are reduced in order to promote aquifer sustainability.

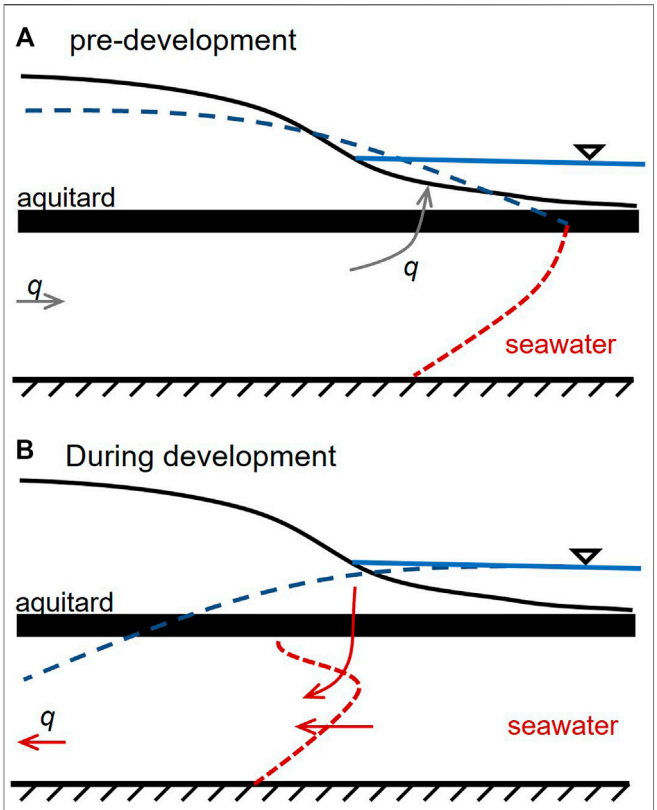

FIGURE 3 | Schematic representation of **(A)** emergence of fresh water through the aquitard overlying a semi-confined aquifer under pre-development conditions, and **(B)** landward flow of water under conditions that result in active seawater intrusion (adapted from Hugman et al., 2021).

Figure 3 illustrates the first two of these conditions. Conditions in the future lie between these two extremes. The methodology that we now describe provides a statistical characterisation of GHB parameters that can emulate these conditions. Its use is based on the premise that offshore processes can be represented by time-invariant boundary conditions over the simulation time of the VL model. We demonstrate below that this is actually the case.

3.3 Stochasticity of GHB Parameters

Before describing the methodology through which a prior probability distribution can be ascribed to GHB parameters, we describe the manner in which the stochasticity of these parameters is represented. We then describe how values of variables which define this representation can be derived through strategic use of a density-dependent model to represent the range of possible offshore conditions.

Let the vector **h** represent heads ascribed to pilot points that are used to parameterize the coastal model boundary of the VL management model; let the vector **c** represent pilot point conductances. Collectively, boundary parameters are therefore

represented by the composite vector $\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}$. Stochastic representation of boundary parameters requires that mean values be provided for these parameters at the locations of all pilot points; these mean values are represented by the vector $\begin{bmatrix}\underline{h}\\ \underline{c}\end{bmatrix}$. It also requires that a covariance matrix $C\left(\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}\right)$ be ascribed to these parameters. This covariance matrix is used to represent spatial correlation between parameters of the same type at each pilot point along the boundary, as well as correlation between parameters of different types. We assume a Gaussian probability distribution; note, however, that this assumption is not invoked until realisations of boundary parameters are generated for use during stochastic history-matching. Meanwhile prior means, together with the prior covariance matrix, form the basis of regularized inversion through which model calibration is achieved.

Characterization of Stochasticity using a Density-Dependent Model.

The prior mean vector and the prior mean covariance matrix of the $\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}$ parameter set which characterize the coastal boundary of the VL model are obtained through a two-step process. The first of these steps samples a one-dimensional counterpart of $\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}$. We refer to this vector as $\begin{bmatrix}h\\ c\end{bmatrix}$; it possesses just two elements (each of them random), namely a single head h and a single conductance c. Once enough samples of $\begin{bmatrix}h\\ c\end{bmatrix}$ have been obtained, its mean $\begin{bmatrix}\underline{h}\\ \underline{c}\end{bmatrix}$ and covariance matrix $C\left(\begin{bmatrix}h\\ c\end{bmatrix}\right)$ can be estimated. Samples of $\begin{bmatrix}h\\ c\end{bmatrix}$ are obtained by running a two-dimensional, cross-section, variable density model many times, using random samples of hydraulic properties and geometry which are representative of the offshore VL aquifer system.

In the second step, the stochastic description of $\begin{bmatrix}h\\ c\end{bmatrix}$ is modified to provide a stochastic description of $\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}$. That is, a mean $\begin{bmatrix}\underline{h}\\ \underline{c}\end{bmatrix}$ vector and a covariance matrix $C\left(\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}\right)$ are determined. Once these are available, random realizations of $\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}$ can be generated through standard statistical sampling. Details of the manner in which $C\left(\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}\right)$ is obtained are provided in **Supplementary Material**. Briefly, the steps are as follows:

1) Assume a correlation length for h and c along the boundary. This is a heuristic decision that supports representation of spatial variability of these variables along the boundary, while suppressing its expression on too short or too long a spatial scale.

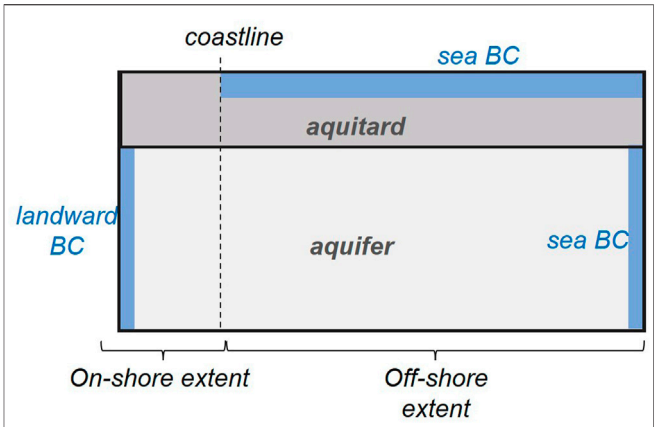

FIGURE 4 | Schematic representation of a two-dimensional, density-dependent, cross-sectional model (adapted from Hugman et al., 2021).

2) Determine the variance of h and c and the correlation between h and c from strategic deployment of a suite of random, one-dimensional, sectional models in the manner described below.
3) Use linear algebra (see **Supplementary Material**) to derive expressions for $\begin{bmatrix}\underline{h}\\ \underline{c}\end{bmatrix}$ and $C\left(\begin{bmatrix}\mathbf{h}\\ \mathbf{c}\end{bmatrix}\right)$ based on the above information

Details are provided in **Supplementary Material**.

3.4 Details of the Density-dependent Model

A stochastic, physically based model of the offshore system is used to obtain samples of $\begin{bmatrix}h\\ c\end{bmatrix}$. The model is coarse, but reflects the properties, geometry and dynamics of the real-world VL system.

Figure 4 depicts the domain of a two-dimensional, cross-sectional, density dependent SEAWAT model (Langevin et al 2008). This model simulates conditions which are illustrated schematically in **Figure 3**. Part of its domain lies beneath land, and part of its domain lies beneath the sea. The model simulates head-driven flow of water through a confined aquifer and under-sea emergence of water through a semi-confining aquitard.

An ensemble of model realizations is constructed. Each realization of the model is endowed with random hydraulic properties, a random landward boundary head, and randomized aspects of its geometry sampled from uniform or log-uniform distributions representative of those that characterize the VL subsystem. **Table 1** lists aspects of model design which vary between realizations.

For each model realization, aquifer and aquitard properties are homogeneous within their respective subdomains. However, they are different between realizations. On the seaward side of the coast, a constant-head boundary is introduced to all top-layer cells and along the vertical model boundary; the head is equivalent to 0 m of salt water. Cells comprising the vertical

TABLE 1 | Aspects of the design of the density-dependent model which vary between realizations.

Design variable	Lower bound	Upper bound	Units	Distribution type
Aquifer horizontal hydraulic conductivity	1	100	m/day	Log-uniform
Aquifer thickness	20	500	m	Uniform
Aquifer porosity	0.1	0.3	—	—
Aquifer specific storage	1×10^{-6}	1×10^{-3}	m^{-1}	Log-uniform
Aquitard vertical hydraulic conductivity	1×10^{-4}	1×10^{-3}	m/day	Log-uniform
Depth to top of aquitard below sea level at landward edge of model domain	0	100	m	Uniform
Seaward dip of all model layers	0	6	Degrees	Uniform
Pre-development landward head	5	20	m	Uniform
Distance from coast to landward model boundary	2.5	5.0	km	Uniform
Post-development landward head	−10.0	0.0	m	Uniform

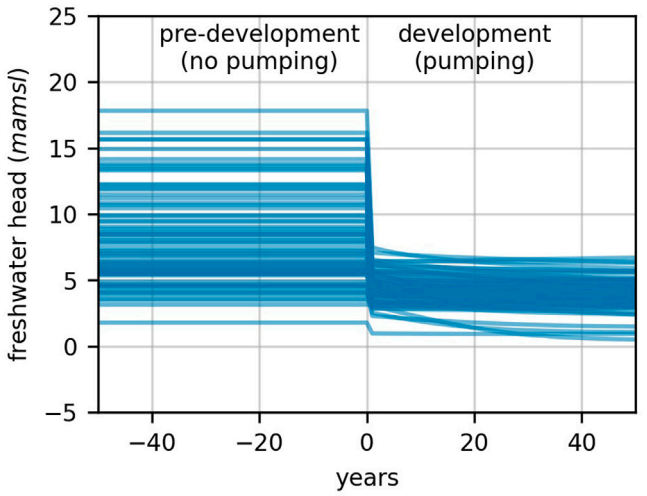

FIGURE 5 | Simulated freshwater heads at the coast during pre-development and development conditions simulated by the cross-section model.

landward model boundary are assigned a uniform freshwater head. Depending on the model stress period, this is either positive (i.e., above sea level) to simulate pre-development conditions, or negative (i.e., below sea level) to simulate present-day, extractive conditions. The value of landward heads is randomly selected from realization to realization. The distance from the coast to the landward model boundary also varies from realization to realization. Collectively, the range of post-development landward heads, and onshore distances to these heads, encompass those which presently prevail in the VL subsystem.

For each realization, the model is first run under pre-development conditions until equilibrium is established. It is then run for a 50-year period in which landward head is set in accordance with present-day conditions. Fifty years corresponds to the time over which groundwater has been extracted from the VL subsystem. The last 20 years of the last period are thus representative of conditions which prevailed during history-matching of the VL management model described above. For each realization, time series of simulated concentrations and heads at the coastline, together with budget components, are recorded.

Over all realizations the location of the toe of the interface varies from 30 km offshore to 500 m onshore. **Figure 5** shows the freshwater head at the coast plotted against time for the 100 realizations that were employed for this study. For most realizations, a sudden change in head occurs shortly after landward extraction of water commences. The head remains reasonably constant thereafter.

The difference in head across the coastline between pre-development and development conditions is used to determine values of h and c for a single realization of the VL model GHB. For any one realization of the density-dependent model, let the freshwater head at the coastline be H_o when water flows toward the sea (i.e., under pre-development conditions), and H_i when water flows toward the land (i.e. under post-development conditions). Values of H_o and H_i are easily obtained from outputs of this model; a value for H_i is established by averaging model-calculated coastal heads over the development period.

Let q_o and q_i denote flow of water (fresh and saline) under the coastline during pre- and post-development conditions. Note also that q_i and q_o have opposite signs. We wish to describe flow across the coastal boundary using a GHB with head h and conductance c. Under outflow conditions:

$$q_o = (H_o - h)c \tag{1}$$

while under inflow conditions:

$$q_i = (H_i - h)c \tag{2}$$

These two equations can be solved for the two unknowns h and c. The solutions are:

$$c = \frac{q_o - q_i}{H_o - H_i} \tag{3}$$

$$h = \frac{q_o H_i - q_i H_o}{q_o - q_i} \tag{4}$$

By running the two-dimensional, sectional, density-dependent model many times, many different values of $\begin{bmatrix} h \\ c \end{bmatrix}$ can be obtained using **Equations (3)**, **(4)**. Values of $\begin{bmatrix} \underline{h} \\ \underline{c} \end{bmatrix}$ and $C\left(\begin{bmatrix} \mathbf{h} \\ \mathbf{c} \end{bmatrix}\right)$ can then calculated in the manner described above and in **Supplementary Material**.

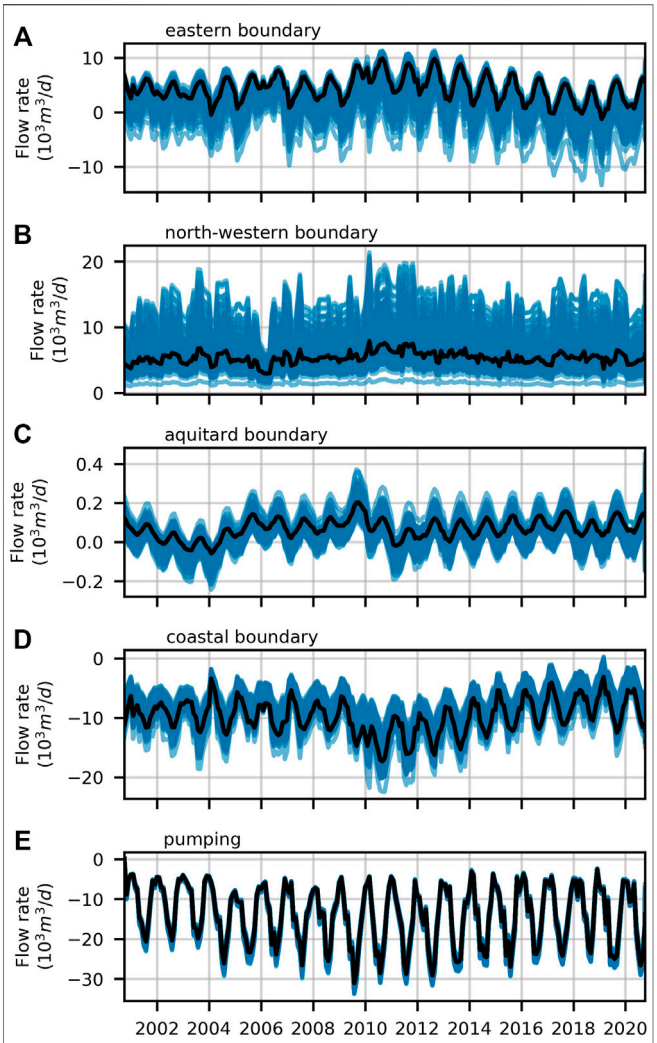

FIGURE 6 | Time-series for each of the VL water balance components simulated with the post-history-matching parameter ensemble (blue lines); net flow through the **(A)** eastern, **(B)** north-western **(C)** aquitard and **(D)** coastal boundaries, and **(E)** extracted through pumping. The black line highlights time series calculated using the parameter set achieved through model calibration (adapted from Hugman et al., 2021).

4 HISTORY-MATCHING AND DEPLOYMENT OF THE VL MANAGEMENT MODEL

Only a brief description of history-matching and deployment of the VL model is provided here. For more details, see Hugman et al. (2021).

4.1 History Matching

History-matching of the VL model was a two-step process. First, parameters were calibrated to obtain a parameter field of minimum error variance using PEST_HP (Doherty, 2021). During this process, parameter uniqueness was attained using preferred-value Tikhonov regularization supplemented with prior parameter covariance matrices to promote parameter field smoothness. An ensemble of 200 parameter realizations was then sampled from a linear approximation to the posterior parameter probability distribution. These were adjusted in order for model outputs to match field measurements using the PESTPP-IES ensemble smoother (White, 2018).

The inversion process featured a total of 2,241 adjustable parameters. These include 211 parameters for the set of LUMPREM models and three sets of 565 pilot point parameters representing hydraulic conductivity, specific storage and aquitard conductance within the VL model domain. The remaining parameters pertain to lateral GHB's. Of these 58 were used to characterize head and conductance along the coastal GHB.

The history-matching observation dataset included both hard and soft data. LUMPREM-calculated water demands were compared to measured extraction rates recorded by groundwater users where these are available. Simulated heads were compared to borehole heads measured in the piezometric monitoring network. Temporal head differences were also compared; this encourages the history-matching process to replicate seasonal head variations. The calibration dataset also included constraints of minimum heads in abstraction wells and upper limits on lateral recharge.

4.2 Model Deployment

4.2.1 Components of the Water Budget

Compliance with the European Water Directive (WFD) requires evaluation of components of the VL subsystem water balance. Time series of water budget components were computed using 200 samples of the posterior parameter probability distribution calculated by PESTPP-IES. These include flows through the coastal model boundary, through each of the landward model boundaries, and from the overlying unconfined aquifer. It also includes LUMPREM-calculated extraction rates. **Figure 6**. Time-averaged components of the water balance, together with respective uncertainty standard deviations, are listed in **Table 2**.

Features of interest in **Table 2** include the following:

- Inflow to the deep (exploited) aquifer from the overlying unconfined aquifer comprises only a small component of the overall water balance.
- There is a considerable influx of water to the system from the seaward side of the coastal boundary.
- Inflow to the system from the north-western boundary is large.
- Uncertainties associated with all lateral boundary inflows are large.

TABLE 2 | Water balance of VL aquifer over the history-matching period.

Inflow from	Mean (Mm^3/yr)	Standard deviation (Mm^3/yr)
Eastern boundary	0.23	0.59
North-western boundary	2.30	0.89
Aquitard	1.4×10^{-3}	0.01
Coastal boundary	2.64	0.42
Pumping	−5.17	0.17

The WFD requires that groundwater extraction be less than 90% of average annual (freshwater) recharge in managed groundwater systems such as Vale do Lobo by 2027. This criterion is simplistic and difficult to apply to the VL subsystem, as extraction from the system induces recharge. Nevertheless, if **Table 2** is viewed as a static ledger, it can be established that 2.9 ± 1.07 Mm^3/yr of extra recharge is required if all non-coastal inflow is to exceed pumping. This value represents the lower limit of additional recharge (or reduced abstraction) that is required to meet WFD requirements. However, it probably underestimates alterations to the current water balance that are required to preclude the threat of saltwater intrusion.

4.3 Optimal Distribution of Extraction

VL irrigation is almost entirely dependent on groundwater extraction. Reduction of extraction is a prerequisite for reducing the risk of seawater intrusion. However, a strategic distribution of extraction locations may mitigate the amount by which extraction must be reduced in order to ensure sustainability.

The VL management model was used to calculate how much water can be extracted from the VL subsystem under the constraint that extraction remains sustainable. Decision variables in this optimisation problem are extraction rates employed by the nine major VL groundwater user groups; it is assumed that the wells from which these groups extract water is unchanged from that which they presently employ. As extraction rates attributed to these user groups change, so too does the geographical distribution of groundwater extraction from the VL subsystem.

Maximization of extraction is subject to a single constraint. This is that there be zero net flow of water landward from the VL coastal boundary into the model domain. We recognise that this constraint does not preclude seawater intrusion. It only ensures net freshwater discharge to the sea. Extraction rates which satisfy this constraint may exceed those that preclude seawater intrusion. Thus, outcomes of this optimisation problem provide an upper estimate of the sustainable rate of extraction from the VL subsystem. More detailed constraints can be imposed if this value is large enough to warrant further investigation. Results presented below suggest that it is not likely to be worth the effort.

4.3.1 Difficulties

History-matching ascribes high conductances to the eastern GHB of the VL model. This indicates that the VL system cannot be managed in isolation. Pumping of the deep aquifer to the east of this boundary lowers groundwater levels within the VL subsystem. (Recall the artificial nature of the VL management boundary.)

To accommodate this issue the optimization problem is solved with five different head distributions ascribed to the eastern GHB, these simulating five different intensities of water extraction from the neighbouring subsystem. The highest of these five head distributions mimic pre-development conditions in which extraction from the neighbouring, easterly subsystem is minimal. The lowest of these five head distributions are slightly above those that characterise present-day conditions. The optimization problem is solved five times, once for each of these boundary head distributions.

To accommodate parameter uncertainty, these optimization problems are solved in two different ways. First the VL model is endowed with a single parameter set, namely the parameter set emerging from model calibration. We refer to these five solutions as "risk neutral", for they take no account of posterior parameter uncertainty. Ideally the calibrated parameter field, and model predictions that are made using this field, lie somewhere near the centre of their respective posterior probability distributions. These optimization problems are solved using the PESTPP-OPT optimizer that is supplied with the PEST++ suite.

We then solve the optimization problem using the ensemble of parameter sets that were calculated by the PESTPP-IES ensemble smoother. In this case, definition of the optimization constraint is varied to accommodate parameter uncertainty. Groundwater extraction is now optimized under the constraint that inland flow from the coastal boundary is zero or less (i.e., flow is outward) for all 200 parameter fields that comprise samples of the posterior parameter distribution. This simulates the operation of a risk-averse management strategy. This optimization-under-uncertainty problem is solved using the CMAES_HP global optimizer (supplied with PEST_HP).

4.3.2 Results

Risk neutral and risk averse solutions to the optimization problem are depicted in **Figures 7A–C**. The vertical axis of **Figure 7A** expresses optimal total extraction as a fraction of current extraction; the horizontal axis expresses conditions at the eastern boundary using an observation well that is close to this boundary. **Figure 7B, C** depict the distribution of optimized extraction between user groups. The major groundwater users in the area are several golf courses, an area of intensive agriculture and distributed small users. They are intentionally not named in **Figure 7** for privacy.

As stated above, the lowest of the five head distributions attributed to the eastern GHB corresponds to heads that are slightly above those that are currently being experienced. **Figure 7A** demonstrates that it is not possible to obtain a feasible solution to the constrained optimization problem under head conditions that are any lower than this, including those which characterise present-day conditions. It follows that if water use on the other side of the eastern boundary continues at its current rate, any extraction from the VL aquifer is unsustainable regardless of the risk stance. This is because eastern groundwater extraction induces landward flow of water through the VL subsystem. The VL subsystem cannot be managed in isolation.

5 DISCUSSION AND CONCLUSION

5.1 The "Simple" Model Setup

It is apparent from the above description that construction, calibration and deployment of the simple model that is described herein requires a rather complicated workflow. The model is "simple" in that it runs fast and is numerically stable; this is achieved by representing processes and structure through simple functions and boundary conditions. All of these are customized, this requiring that software be written for their implementation and parameterisation. Fine-tunning this setup required some trial-and-

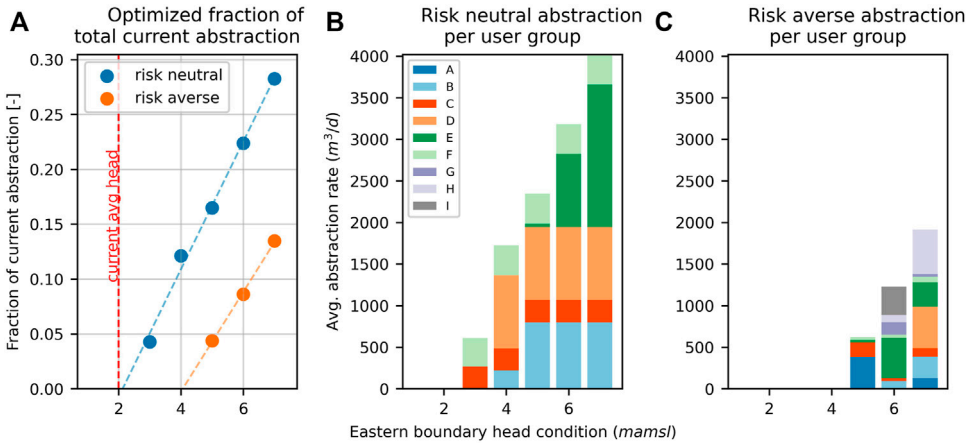

FIGURE 7 | (A) Solutions to constrained extraction optimization problems and, relative extraction rates for different user groups for **(B)** the five solutions of the risk neutral optimization problem and **(C)** the three solutions of the risk averse optimization problem (adapted from Hugman et al., 2021).

error; the workflow was revised many times as modelling proceeded. This is not an uncommon occurrence in real-world, decision-support modelling. The advent of new information, and lessons that are painfully learned during model development, frequently require that modelling plans be considerably revised. Fortunately, all components of all of the workflows that are outlined in this paper are scripted. Furthermore, the run time of the management model is small. This allowed rapid experimentation and testing of different ideas in ways that were flexible and reproducible.

All real-world modelling is difficult. Innovation is a necessity. "Back to the drawing board" moments are common. The advantages of reproducibility and reduced mental overhead that are enabled by a scripted workflow and a structurally simple model cannot be overemphasized.

5.2 Complementary Models

This paper demonstrates how abstract, non-physical parameters that characterise the boundary conditions of a fast-running, decision-support model can be informed by expert knowledge. This enables derivation of the prior expected values of boundary condition parameters, as well as characterisation of the prior uncertainties of these parameters. This obviates the need for non-informative priors whose use may unnecessarily inflate the posterior uncertainties of decision-critical model predictions.

The methodology described herein is based on the conjunctive use of two models. One of these models is complex. Its role is to simulate system structures and processes that are represented in an abstract manner in a complementary, simple, fast-running model that is designed with data assimilation and uncertainty quantification in mind. Simulation of hydrogeological complexities by the complex model cannot be exact, for many of the structures and processes that it includes are only vaguely known. Its task is to represent the ramifications of this vague knowledge for parameters that are assigned to part of the simpler management model—one of its boundary conditions in this case. Representation of this boundary must be stochastic. To the extent that expert-knowledge-informed stochasticity of the complex model can limit the stochasticity of management model boundary parameters, the latter's role in supporting environmental decision-making is enhanced, for the uncertainties of some of its predictions may thereby be reduced.

It is evident from the above that such an approach is of value in cases where 1) prediction uncertainty is affected by abstract parameter uncertainty and 2) this uncertainty is not reduced through history matching. In such cases, as discussed, expert knowledge can be an important (or only) source of information to constrain decision-pertinent uncertainties.

Figure 8 displays a generalisation of our workflow. It can be summarized as follows:

- A simple model is constructed in which numerically problematic processes and structures are replaced with approximate representations that are populated by non-physically based parameters. Forecasts of interest simulated by the simple model are affected by the uncertainties of these parameters where these cannot be adequately constrained by field observations of system states and fluxes.
- A complementary complex model is constructed which explicitly represents some of the processes and structures that prevail at a study site. Prior probability distributions for parameters of the complex model are characterized using expert knowledge derived from site characterisation. The complex model is populated using an ensemble of parameter fields sampled from this expert-knowledge-informed prior probability distribution. From the outcomes of many complex model runs, equivalent values for the more abstract parameters of the simple model are calculated.
- Samples of abstract parameters which are thus obtained are used to characterize the prior probability distribution of simple model parameters.
- The simple model is then subjected to stochastic history-matching. Posterior parameter probabilities obtained through this process therefore reflect information gained from both measured data and expert knowledge. These are used to explore predictive outcomes of management interest.

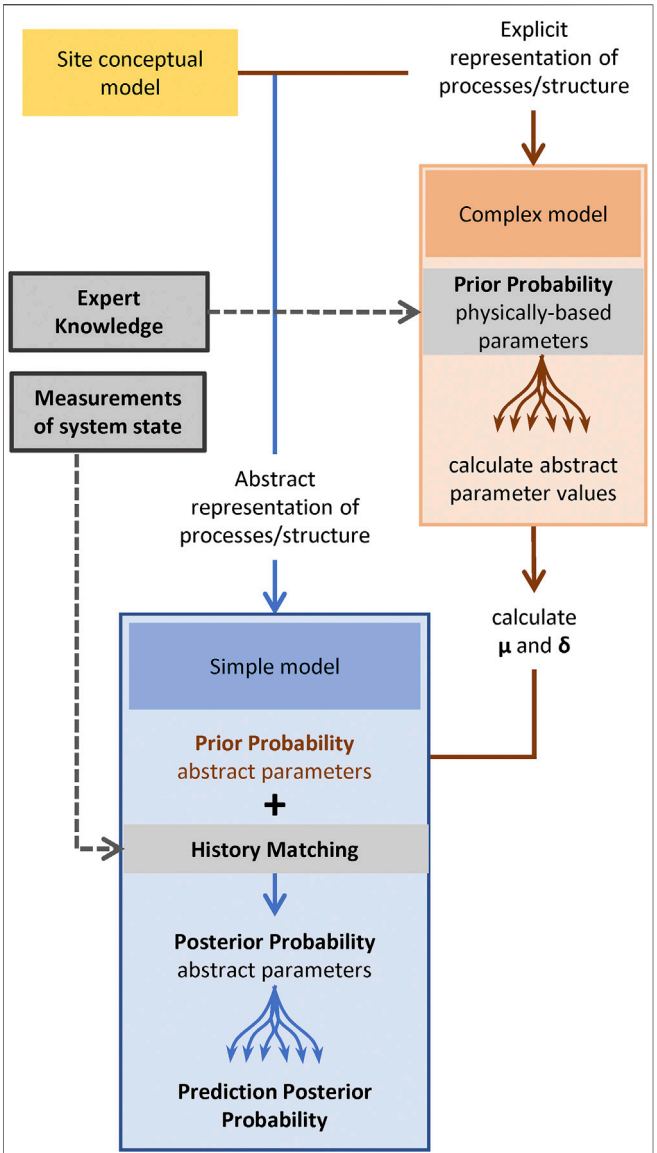

FIGURE 8 | Diagram of a workflow that uses complementary models to extract information from both expert knowledge and measurements of system state.

It is safe to say that few, if any, modelling alternatives to the approach that we have documented herein would have enabled data assimilation, uncertainty quantification, and exploration of optimized management strategies for the VL subsystem in ways that were achieved with the present modelling strategy. The option of building a single, complex, density dependent model (an option that is often taken in coastal contexts) would have made high-end data assimilation and uncertainty quantification almost impossible because of long run times and likely numerical instability. These problems would have been exacerbated if the domain of this model was made to extend for a considerable distance under the sea in order to simulate freshwater outflow. This could not have been accommodated in a properly stochastic fashion. Simplifications in representing outflow conditions and testing of only one or a small number of outflow options, may have led to unquantifiable predictive bias.

Decision-support modelling always requires compromises. These compromises do not result in loss of "simulation integrity", for integrity of simulation is an unachievable goal. If the task of decision-support modelling is properly defined (see, for example Doherty and Moore, 2021), the search for a compromise requires that reduction of uncertainty accrued through assimilation of expert knowledge be traded off against reduction of uncertainty accrued through history-matching. If a model is structurally simple, if may be capable of fitting a history-matching dataset very well. However, if this dataset is not rich in prediction-specific information, the parametric and predictive bias that may be incurred through model structural simplicity, may outweigh the benefits of assimilation of these data. Meanwhile, use of an abstract model design may have precluded parameter value insights gained from expert knowledge of real-world hydraulic properties. The optimal compromise will always be context and prediction specific.

This study demonstrates that selection of an appropriate level of model structural complexity does not always have to be an "either/or" choice. This is because the model that performs data assimilation does not need to be the same model as that which is used to extract information from expert knowledge. Two different models can be used conjunctively for these two different purposes. They can "meet" at a boundary condition of the data assimilation model. The parameters of this boundary condition are thereby provided with prior probability distributions that reflect process and properties that are missing from the model itself, but that may have profound effects on the uncertainties of management-critical predictions.

By separating the tasks of assimilating information that is resident in measurements of system state on the one hand, and assimilating information that is resident in expert knowledge on the other hand, each model can be tuned to its primary task. Fast execution speed and numerical stability are primary requirements for the former task. An ability to represent the range of possible conditions that are compatible with expert knowledge is a primary requirement for the latter task. The "complex" model that was used for the latter task in work that is documented herein, was actually not very complex; it is a two-dimensional sectional model. However, its two-dimensional design enhanced, rather than impeded, its ability to explore a wide range of hydraulic and geometric possibilities for undersea emergence of fresh water off the coast of Vale do Lobo. These possibilities were then made available to the management model through stochastic characterisation of its coastal boundary condition.

5.3 Final Remarks

Decision-support modelling in coastal areas is notoriously difficult. Uncertainties are often high. The potential of available data to reduce these uncertainties may be limited.

This may be a blessing rather than a curse. Compromises must be made. These compromises often require that a modeller identify sources of information that may be most effective in reducing the uncertainties of predictions that he/she cares about. Modelling must then be tuned to extraction of information from these sources without inducing bias or inflating uncertainties by impeding or distorting the flow of information from other sources to too great an extent. However, where uncertainties are already high, the costs in

extra uncertainty that are incurred by designing a model, and a modelling workflow, that are optimal for one task but not necessarily another, may not be great in comparison to these inherent, background uncertainties that are born of information insufficiency. This gives a modeller the freedom to test different workflows, and then tailor them to suit his/her needs.

We have demonstrated in this paper that strategic use of complementary models may make modelling choices less painful. A "win/lose" choice can be transformed into a "win/win" choice. We have also attempted to demonstrate that where innovation is required, the use of automated workflows allows a modeller to free him/herself from the yoke of having to adhere to one particular workflow. Different options can be rapidly tested. Once a suitable option is discovered, it can be rapidly improved. The cost of exploration becomes remarkably low. The journey of discovery becomes remarkably rewarding.

DATA AVAILABILITY STATEMENT

Data pertaining to modelling of the real-world case study is proprietary and cannot be shared. Data for the synthetic case can be provided on request. Requests to access the datasets should be directed to rui.hugman@flinders.edu.au.

AUTHOR CONTRIBUTIONS

RH and JD conceived the idea. RH analysed the data, set up the modelling workflows and undertook all simulations and their postprocessing. JD derived the solution to obtain the stochastic characterization of spatially distributed parameters. JD implemented adaptations to MODFLOW6 and prepared PEST_CMAES setups. Both authors discussed the results and contributed to the final manuscript.

FUNDING

This work was undertaken through the "Groundwater Modelling Decision Support Initiative" (GMDSI), conducted under the auspice of the National Centre for Groundwater Research and Training (NCGRT), Flinders University, South Australia with funding from BHP and Rio Tinto.

ACKNOWLEDGMENTS

We thank Kathleen Standen (University of the Algarve) for her contributions to gathering and processing data for the Vale do Lobo case study reported in this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.867379/full#supplementary-material

REFERENCES

Doherty, J., and Moore, C. (2020). Decision Support Modeling: Data Assimilation, Uncertainty Quantification, and Strategic Abstraction. *Groundwater* 58, 327–337. doi:10.1111/gwat.12969

Doherty, J., and Moore, C. (2021). Decision Support Modelling Viewed through the Lens of Model Complexity. *A GMDSI Monogr.* South Australia: National Centre for Groundwater Research and Training, Flinders University. doi:10.25957/p25g-0f58

Doherty, J., and Simmons, C. T. (2013). Groundwater Modelling in Decision Support: Reflections on a Unified Conceptual Framework. *Hydrogeol. J.* 21, 1531–1537. doi:10.1007/s10040-013-1027-7

Doherty, J. (2021a). *PEST_HP PEST for Highly Parallelized Computing Environments*. Brisbane: Watermark Numerical Computing.

Doherty, J. (2021b). *Version 2 of the LUMPREM Groundwater Recharge Model*. Brisbane: Watermark Numerical Computing.

Freeze, R. A., Massmann, J., Smith, L., Sperling, T., and James, B. (1990). Hydrogeological Decision Analysis: 1, A Framework. *Groundwater.* 28, 738–766.

Hugman, R., Doherty, J., and Standen, K. (2021). Model-Based Assessment of Coastal Aquifer Management Options. *A GMDSI Worked Example Report*. South Australia: National Centre for Groundwater Research and Training, Flinders University. doi:10.25957/a476-x588

Knight, A. C., Werner, A. D., and Morgan, L. K. (2018). The Onshore Influence of Offshore Fresh Groundwater. *J. Hydrol.* 561, 724–736. doi:10.1016/j.jhydrol.2018.03.028

Langevin, C. D., Thorne, D. T., Jr., Dausman, A. M., Sukop, M. C., and Guo, Weixing. (2008). SEAWAT Version 4: A Computer Program for Simulation of Multi-Species Solute and Heat Transport: U.S. Geological Survey Techniques and Methods Book 6, 39 p. U.S. Geological Survey, Reston, Virginia.

Langevin, C. D., J. D, Hughes, E. R, Banta, A. M, Provost, R. G, Niswonger, and S, Panday. (2021). MODFLOW 6 Modular Hydrologic Model Version 6.2.2: U.S. Geological Survey 652 Software Release, 30 July 2021. doi:10.5066/F76Q1VQV

Lu, C., Werner, A. D., Simmons, C. T., and Luo, J. (2015). A Correction on Coastal Heads for Groundwater Flow Models. *Groundwater* 53, 164–170. doi:10.1111/gwat.12172

Post, V. E. A., Groen, J., Kooi, H., Person, M., Ge, S., and Edmunds, W. M. (2013). Offshore Fresh Groundwater Reserves as a Global Phenomenon. *Nature* 504, 71–78. doi:10.1038/nature12858

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the Predictive Consequences of Model Error with Linear Subspace Analysis. *Water Resour. Res.* 50, 1152–1173. doi:10.1002/2013WR014767

White, J. T. (2018). A Model-independent Iterative Ensemble Smoother for Efficient History-Matching and Uncertainty Quantification in Very High Dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Earth Science

ORIGINAL RESEARCH
published: 19 May 2022
doi: 10.3389/feart.2022.884075

Automated Hierarchical 3D Modeling of Quaternary Aquifers: The ArchPy Approach

Ludovic Schorpp[1]*, *Julien Straubhaar*[1] and *Philippe Renard*[1,2]

[1]Centre for Hydrogeology and Geothermics, University of Neuchâtel, Neuchâtel, Switzerland, [2]Department of Geosciences, University of Oslo, Oslo, Norway

When modeling groundwater systems in Quaternary formations, one of the first steps is to construct a geological and petrophysical model. This is often cumbersome because it requires multiple manual steps which include geophysical interpretation, construction of a structural model, and identification of geostatistical model parameters, facies, and property simulations. Those steps are often carried out using different software, which makes the automation intractable or very difficult. A non-automated approach is time-consuming and makes the model updating difficult when new data are available or when some geological interpretations are modified. Furthermore, conducting a cross-validation procedure to assess the overall quality of the models and quantifying the joint structural and parametric uncertainty are tedious. To address these issues, we propose a new approach and a Python module, ArchPy, to automatically generate realistic geological and parameter models. One of its main features is that the modeling operates in a hierarchical manner. The input data consist of a set of borehole data and a stratigraphic pile. The stratigraphic pile describes how the model should be constructed formally and in a compact manner. It contains the list of the different stratigraphic units and their order in the pile, their conformability (eroded or onlap), the surface interpolation method (e.g., kriging, sequential Gaussian simulation (SGS), and multiple-point statistics (MPS)), the filling method for the lithologies (e.g., MPS and sequential indicator simulation (SIS)), and the petrophysical properties (e.g., MPS and SGS). Then, the procedure is automatic. In a first step, the stratigraphic unit boundaries are simulated. Second, they are filled with lithologies, and finally, the petrophysical properties are simulated inside the lithologies. All these steps are straightforward and automated once the stratigraphic pile and its related parameters have been defined. Hence, this approach is extremely flexible. The automation provides a framework to generate end-to-end stochastic models and then the proposed method allows for uncertainty quantification at any level and may be used for full inversion. In this work, ArchPy is illustrated using data from an alpine Quaternary aquifer in the upper Aare plain (southeast of Bern, Switzerland).

Keywords: automated modeling, geological modeling, stochastic, hierarchy, Quaternary, Python, open-source, multiplepoint statistics

OPEN ACCESS

Edited by:
Jeremy White,
Intera, Inc., United States

Reviewed by:
Shawgar Karami,
Amirkabir University of Technology, Iran
Emmanouil Varouchakis,
Technical University of Crete, Greece

***Correspondence:**
Ludovic Schorpp
ludovic.schorpp@unine.ch

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *25 February 2022*
Accepted: *15 April 2022*
Published: *19 May 2022*

Citation:
Schorpp L, Straubhaar J and Renard P (2022) Automated Hierarchical 3D Modeling of Quaternary Aquifers: The ArchPy Approach. Front. Earth Sci. 10:884075. doi: 10.3389/feart.2022.884075

1 INTRODUCTION

When constructing a 3D groundwater flow model, one of the first steps is to build a geological model. This includes defining the geometry of the stratigraphic units, filling them with a spatial distribution of lithofacies, and finally filling the lithofacies with petrophysical parameter values. The construction of these models is often complex and involves multiple assumptions and computing tools (Pyrcz and Deutsch, 2014; Ringrose and Bentley, 2016; Wellmann and Caumon, 2018). It is necessary to evaluate the uncertainties related to the parameter values, and indeed, geological structures of the aquifer or inversion using hydrogeological or geophysical data are also important sources of uncertainty. It is therefore extremely important to be able to construct all the components of these geological models in a manner that is fully automated, well documented, and repeatable. In this study, our aim is to introduce a new tool that can be used for this purpose for Quaternary aquifers.

The history of geological modeling techniques is rich and diverse (Matheron, 1963; Mallet, 1989; Koltermann and Gorelick, 1996; Ringrose and Bentley, 2016). However, some geological features such as the Quaternary formations are still difficult to model. These sediments were deposited during various sedimentological events, acting at different scales, both temporally and spatially, leading to complex relations and hierarchical structures. Larger and bigger units are the results of the aggregation of subunits of smaller hierarchical order that can themselves be the results of the aggregation of sub-subunits of even smaller hierarchical order, and so on (Miall et al., 1991; Heinz and Aigner, 2003; Bridge, 2009). The definition of this stratigraphic hierarchy is very important when analyzing field data (Aigner et al., 1996; Ford and Pyles, 2014) but also to develop stochastic modeling techniques.

However, one difficulty is that the concept of hierarchy is used differently depending on the modeling techniques, and the hierarchical modeling does not necessarily match exactly what is meant by stratigraphic hierarchy. For example, Neuman (1990) approached the question of the hierarchy by showing that it is likely that the hydraulic conductivity of hierarchical sedimentary deposits should have a truncated-power law variogram, while Ritzi et al. (2004) used the same type of tools but derived different types of variograms. In these approaches the sedimentological heterogeneity is not represented explicitly but represented by multi-Gaussian fields having specific correlation structures. On the other hand, Scheibe and Freyberg (1995) and Ramanathan et al. (2010) constructed highly detailed simulations of fluvial deposits using the concept of hierarchical deposits to investigate the effective properties of these types of sediment.

When modeling aquifers, the word *hierarchy* is often used with a slightly different meaning. It generally means that the modeling of the hydraulic conductivity field includes several steps such as the modeling of stratigraphic units using a given technique, followed by the modeling of the lithofacies within the stratigraphic units, and, finally, followed by the modeling of the hydraulic conductivities within the facies. This hierarchical modeling approach may include only two or all of these steps. It was used in many case studies (Weissmann and Fogg, 1999; Feyen and Caers, 2006; Comunian et al., 2011; Bennett et al., 2019). We note that this approach can be refined by using categorical geostatistical modeling methods to define stratigraphic units in which subunits can be modeled again using categorical simulations tools and so on to obtain multiple levels of hierarchy (Zappa et al., 2006; Comunian et al., 2016). But this last method does not account for the fact that sedimentary units are usually deposited as subhorizontal layers and that, in general, their geometry is controlled by a set of stratigraphic rules. Other methods account for that information, such as the implicit interpolation method implemented in Geomodeller 3D or in Gempy software (Calcagno et al., 2008; de la Varga et al., 2019). In this approach, the user defines the order of the stratigraphic units and the relations between them, allowing us to automatically model the volumes. This is convenient for building complex models in an efficient manner, but Zuffetti et al. (2020) showed that these tools cannot properly handle the concept of subunits within a stratigraphic unit. These authors therefore propose to formalize the stratigraphic hierarchy and define rules allowing us to automatically construct 3D models based on that concept, and they show promising results.

All these observations show that there is still a need for a tool that would facilitate stochastic 3D geological modeling. The tool must allow the user to conduct a complete and proper uncertainty quantification including uncertainty at the level of the unit geometry as well as their lithologies and properties. The tool must allow for carrying out cross-validation efficiently, meaning that it must be possible to remove any type of data and reconstruct an ensemble of models automatically. The model must also be easy to update when new data are acquired or if the geological interpretation is modified.

The aim of this study is to present the ArchPy approach. The method includes the two types of hierarchy discussed earlier. ArchPy defines the stratigraphic units from borehole data while accounting for as many hierarchical levels as needed from a stratigraphic point of view. But the method is also hierarchic in the sense that the same method will include the modeling of the stratigraphic units and then the modeling of the lithofacies within the units and, finally, the modeling of the properties within the lithofacies. ArchPy is a methodology but it is also a Python module allowing the automatic generation of stochastic, reproducible, and hierarchical models from borehole data and a geological concept.

To minimize the user interventions during the simulations, we rely on a formal description of the geological concept that ArchPy uses to construct the hierarchical model. This type of formal description is not new; it was described, for example, by Renard and Courrioux (1994) for fracture networks, and is at the core of Geomodeller 3D or Gempy in which a geological pile is defined to express the relations between the geological events and must be given explicitly (Calcagno et al., 2008). However, as revealed by Zuffetti et al. (2020), the manner in which the pile is defined and used in these codes does not allow us to describe the stratigraphic hierarchy. It is therefore necessary to find more general ways to describe the pile. One approach, using a tree, was proposed in the study by Zuffetti et al. (2020). Here, we adjust this initial data

structure, and we extend it to include additional information allowing us to encapsulate all the knowledge required to automatically build the 3D model. We call this formal description the stratigraphic pile (SP). The SP contains the description of the interpolation methods for all surfaces bounding the stratigraphic units, as well as the description of the simulation methods and parameters for filling the different units with lithofacies and properties.

It is important to note that the geological data in boreholes or outcrops are not always representing the actual position of the boundary between two units; instead, they indicate that this boundary should be lower or above these data. Such situations arise in the presence of erosion or hiatus in the deposition sequence or because a borehole is too shallow and does not reach the base of a given unit. These data are frequent and can be treated as inequalities (Dubrule and Kostov, 1986; Mallet, 1989; Freulon and de Fouquet, 1993; Straubhaar and Renard, 2021). The use of such data can significantly increase the quality of the simulations as shown, for example, by Freulon and de Fouquet (1993). In the ArchPy methodology, we include not only the possibility of interpolating the boundaries between stratigraphic units using such inequalities but we also propose a method to automatically identify the inequalities in the borehole data to facilitate the automatic updating of models with large borehole data sets.

ArchPy is also an object-oriented Python package allowing us to illustrate the applicability and the benefits of the proposed approach. While describing the methodology, we will also discuss the key objects that are used to implement the concepts underlying the approach. Its Python interface and open-source nature facilitate its use for a large number of users.

The main novelty of the proposed approach and this software is to allow fast and reproducible simulations of Quaternary aquifers as well as their related uncertainties to any desired hierarchical level (unit, subunits, facies, and properties).

This article first describes the different components of the ArchPy approach and then illustrates its main features using a synthetic and a real case example.

2 ARCHPY APPROACH

In this section, we first present a brief overview of the main components of the proposed methodology. We then describe in detail the concept of stratigraphic pile (SP) and the way we deal with the stratigraphic rules (erosion and hiatus). All the simulation steps and modeling guidelines are explained using a synthetic case. In the following sections, for the sake of brevity, the word *unit* will refer to the stratigraphic unit as defined in the SP and *lithology* or *facies* will refer to the different lithofacies that can be found inside these units.

2.1 General Overview

The final aim of ArchPy is to generate an ensemble of petrophysical models (or property models) that describe the spatial distribution of specified properties consistent with the location of the units and facies. To achieve these results, ArchPy proceeds in several steps (**Figure 1**). The input data are a stratigraphic pile (SP) and a set of hard data (HD). The HD can be either borehole data or punctual information (e.g., from outcrops). First, the HD are processed to extract the contact points (equalities and inequalities). Then, a whole simulation takes place hierarchically in three main steps:

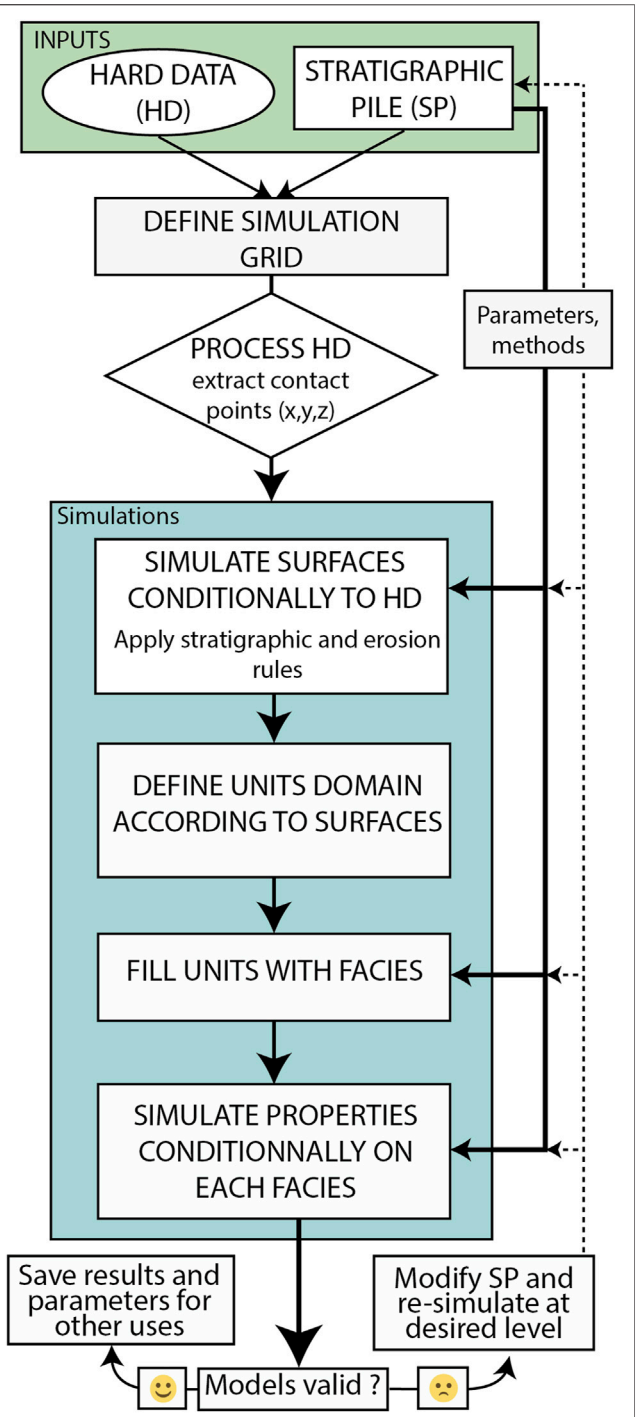

FIGURE 1 | Overview of the ArchPy approach.

1) simulate the surfaces delimiting the unit boundaries and thus allowing the definition of the stratigraphic unit domain;
2) simulate the facies to fill each unit using various geostatistical methods according to prior geological knowledge; and
3) simulate the properties inside each facies independently.

All these steps are done conditionally to the HD. In the end, the final models are validated by the user. If they are not satisfying (based on expert knowledge or on some criterion), the SP can be modified (e.g., on the simulation methods), and previous steps are re-executed, depending on the modified parameters.

ArchPy can be used in several manners. It can be used to generate one or an ensemble of models to quantify uncertainty. It can be used to facilitate the update of geological models when new data are collected in the field. It can also be coupled with an inversion technique to express the prior distribution of the geological and petrophysical parameter values.

It is important to note that each step depends on the results of the previous ones. For example, after a first complete simulation, if the only parameters changed are those of the filling step, it will only be required to simulate the facies in the units and, subsequently, the properties inside them. Similarly, if only one surface has been recomputed and the others have been kept, the only unit domains that will need to be re-simulated (as well as their filling) will be those impacted by a modification of this surface. This flexibility is important for dealing with large inverse problems as the number of unknown parameters can make the problem tedious and difficult (Biegler et al., 2011). It allows us to focus on particular units of interest and only simulate parts of the domain at any desired level, without being forced to simulate the whole system each time a modification is decided.

2.2 Stratigraphic Pile

The concept of stratigraphic pile (SP) is the backbone of the ArchPy methodology. Indeed, it contains almost all the information needed for the simulations, including the stratigraphic relations between units and the description of the simulation methods of the surfaces, the filling, and the properties as well as the parameters controlling them (**Figure 2**). The SP is made of different components (coded as Python objects): units, surfaces, facies, properties, and possible other piles. Following an object-oriented programming logic, all these objects have different attributes (name, color, interpolation method, etc.) that define and differentiate them. These object attributes can also be composed of other different objects. A practical example is the unit object which has a *list_lithofacies* attribute, containing the lithofacies objects that populate this unit.

1) **Surfaces.** They delimit the top of the units and are defined using an interpolation (or simulation) method and by a contact type to indicate if the surface is conformable (onlap) or erosional (erode). The difference is that an erode surface erodes older units where it is simulated below their top surface while an onlap surface is simply ignored in such locations, meaning there is no deposition (see **Section 2.3**).
2) **Units.** They correspond to stratigraphic units that can be observed in HD or on outcrops. Each unit belongs to an SP, and has a specific order index to determine its position in the pile. A unit also needs a surface object to determine its upper boundary, and its lower boundary is defined by the surface at the top of the underlying unit. Finally, the unit contains a description of the filling method and a list of facies objects to simulate inside the unit domain.
3) **Facies.** The facies describe the hydro- or lithofacies that are contained inside the units. They can be very different depending on the modeling purposes and data available (e.g., stratigraphic facies, lithologies, or USCS codes). A list of facies is given for each unit to indicate which ones to simulate during the facies simulation. Furthermore, each facies can be composed by one or more properties that are simulated inside the facies domain (where a specific facies is present).
4) **Properties.** They are independent ArchPy objects that composed the different facies. For each property inside a facies, some parameters can be set, such as the mean value of the property (inside a specific facies), the covariance model, and the method of simulation.
5) **Stratigraphic Pile.** An SP is the combination of all the objects described before that synthesize the geological concept. Note that an SP can be inserted as a filling method within a unit. This allows us to construct a stratigraphic hierarchy. For example, in **Figure 2**, the sub-pile PB is used as input for filling unit B.

Thus, the stratigraphic pile is an object that can easily be manipulated and modified. Using such an approach, the user can focus more on the conceptual aspects of the modeling (unit relations, erosional events, spatial distribution of the lithologies or facies, etc.). In clear terms, the whole modeling process can be easily reproduced because all the steps are documented in the SP.

2.3 Erosion Rules

Erosion (stratigraphic) rules (ERs) describe how the surfaces influence each other after having been simulated. They are similar to those denominated "geological rules" by Calcagno et al. (2008). **Figure 3** shows a simple example of the difference between onlap and erode behaviors. Here, two surfaces are simulated: S1 (old) and S2 (young), while the gray surface is simply set to the topography or the digital elevation model (DEM). S1 is defined onlap and S2 has both behaviors. If S2 is specified onlap, unit 2 (young, in black in the figure) is not deposited where S2 is below S1, whereas if S2 is specified erode, the effective top of unit 1 (young, in yellow in the figure) is set to S2. This approach allows incorporating geological time directly by choosing the appropriate truncation operation to remain consistent with the sedimentological history (Wellmann and Caumon, 2018). Another rule is that a surface cannot be above (or below) the DEM (resp. bottom of the domain); if this case arises, the surface will be automatically set to the DEM (resp. bottom). These ER are applied each time a new surface is simulated during the surface simulation.

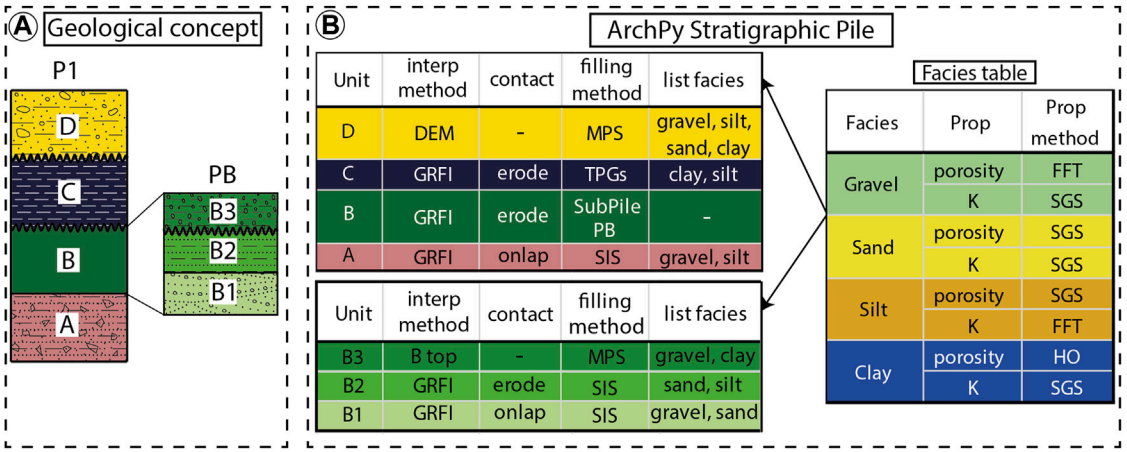

FIGURE 2 | Schematic representation of an ArchPy stratigraphic pile **(B)** given a geological concept **(A)**; interp. method, interpolation method; GRF, Gaussian random functions; DEM, digital elevation model; Prop method, property simulation method; MPS, multiple-points statistics; SIS, sequential indicator simulation; TPG, truncated pluri-Gaussians; HO, homogeneous; K, hydraulic conductivity; FFT, fast Fourier transform; SGS, sequential Gaussian simulation.

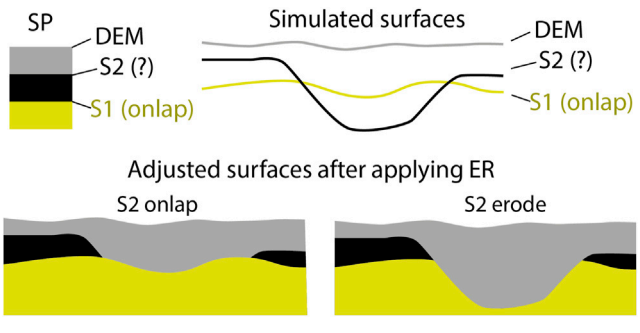

FIGURE 3 | Schematic example of the erosion rules (ER) used in ArchPy. The surfaces are first simulated from oldest to youngest and then adjusted following the ER.

2.4 Synthetic Example

To illustrate the ArchPy capabilities, **Figure 4** shows one stochastic realization of hydraulic conductivity and porosity based on the SP of **Figure 2**. The domain dimensions (x, y, z) are $3 \times 1 \times 0.2$ km^3 with a spatial resolution of $15 \times 15 \times 4$ m^3. The model is purely synthetic, but it mimics a valley filled by a series of sedimentary episodes. For this example, we assume that unit A is a moraine deposit only filled with gravel and silt. The B formation is deposited during three sub-steps (sub-pile PB, **Figure 2**). It represents a fluvio-glacial environment including three stages (B1, B2, and B3). On top of that, unit C is an important glacio-lacustrine phase where only fine particles are deposited (clay and silt). Unit D ends the process by setting up a fluvial environment that was more active in the southern part of the area (toward the -y main axis). The different steps to obtain this result will be explained in detail in the following sections.

2.5 Data Pre-Processing

The main hard data to describe the geology in a Quaternary environment are the borehole information. For each borehole, ArchPy requires the following: a borehole ID, the depth, and location (x, y, and z) of the borehole as well as a stratigraphic unit log and/or a lithology log. Both logs contain the elevation of the top of the interpreted units and lithologies in the simulation grid reference system. These locations will be used as the conditioning point for simulations. An important thing to note is that the logs of borehole objects in the ArchPy interface only need top elevation for each unit/lithology encountered. To facilitate the geostatistical simulation of the properties, we consider only regular Cartesian grids for the moment in ArchPy (to avoid support effects). The extension and parameters of the grid are provided by the user. The SP and eventual sub-piles are also given as input. We then pre-process the borehole data (HD) to check that they are consistent with the SP and to extract the contact points between the units. The difficulty here (which is not intuitive) is that a contact between two units in a borehole does not necessarily correspond to the top of the unit located below this contact because of possible erosion or sedimentation hiatus. It is therefore necessary to analyze the borehole logs to identify the information that the contacts bring about the possible positions of the surfaces. To code this information, we will define contact points where the position of the surface is perfectly defined (equality contact point) and those which provide only indirect information (inequality contact point). Formally, an inequality contact point consists of a lower or upper bound for the actual surface.

Figure 5 shows how HD are interpreted automatically in the ArchPy pre-processing step given four different examples. We do not fully detail each example for brevity, but the main points are covered.

Considering the example showed in **Figure 5A**, three surfaces need to be simulated: blue, green, and red tops (the yellow top is defined by the DEM). Equality points can be safely attributed in B1 between all contacts as there are no hiatus and no erosion layer in the Pile. However, in B2 and B3 boreholes, as the blue outcrops directly, its top must go above the surface, assuming erosion at the

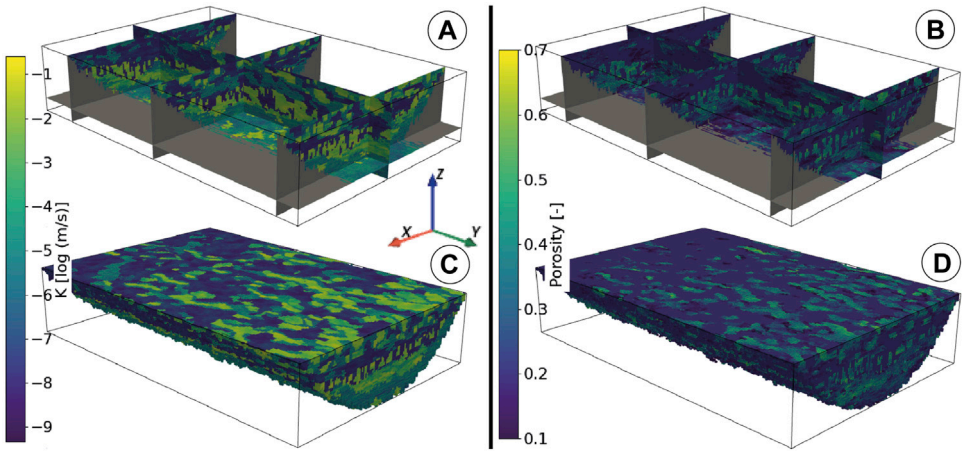

FIGURE 4 | Realizations of two different properties (**(A)** hydraulic conductivity and **(B)** porosity) for the synthetic case. **(C)** and **(D)** are 3D blocs of the **(A)** and **(B)** realizations, respectively. The simulations used respective corresponding facies realizations of **Figure 7** as simulation domains. Vertical exaggeration = 3x.

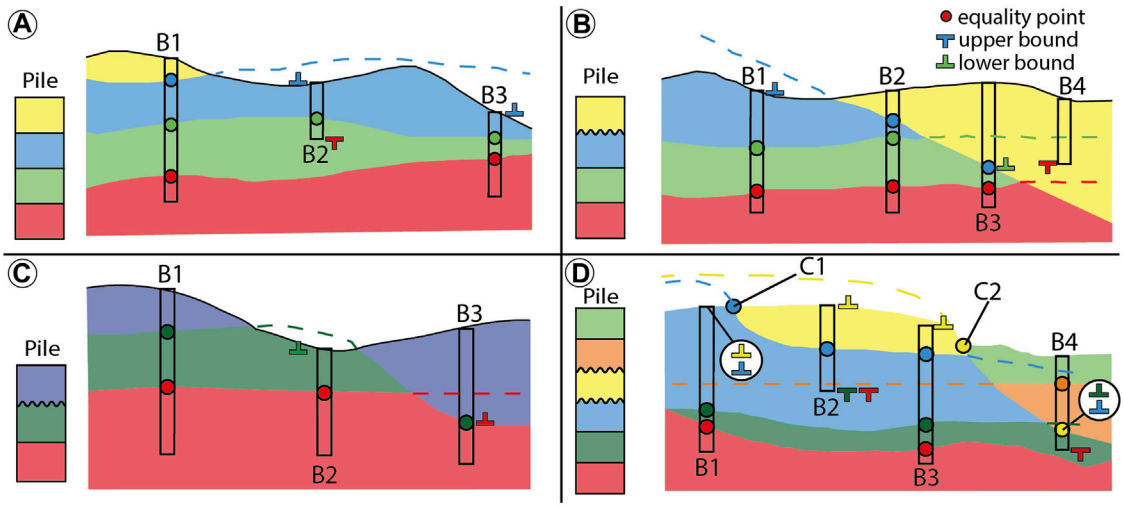

FIGURE 5 | Four examples **(A–D)** of how inequalities and equalities are extracted from boreholes and field data. **(B)** indicates borehole information and **(C)** a unit contact (e.g., observed on the field). For each example, a stratigraphic pile is defined to indicate the relationships between the units and the nature of the surface contact (straight line is onlap and corrugated is erode). Dashed lines represent simulated surfaces before applying erosion rules (see **Section 2.3**).

surface. This means that the tops of the borehole B2 and B3 are then lower bounds for the top of the blue unit. Also, the red unit is not encountered by B2 which indicates that it must go below the bottom of the borehole which is an upper bound for the top of the red unit. The **Figure 5B** example adds an erosional event (the blue top) that cross-cuts the green and red top layers. This implies that we must add an inequality contact point (lower bound) in B3 for the green top and equality contacts along the erosion surface (B2 and B3) for the blue top because of the ER. Indeed, the green top in B3 cannot be considered as the actual green top since the green unit has been eroded. The same event occurs in **Figure 5C** in B3 where the red unit has been eroded by an erosional event (the green top).

The more complex example (**Figure 5D**) shows that the number of extracted data can become important, especially when the number of layers increases. Here, additional outcrop information has been added with unit contacts (C1 and C2) that inform about a transition between two units. Yellow goes above as it is the unit reaching the topography. The contact between yellow and blue is an equality for blue (the only erode layer above blue is yellow, but it has been deposited and thus cannot have eroded blue). The bottom of B2 ends in the blue unit which indicates that all layers below it (dark green and red) must go below the bottom of the borehole. When two units in a borehole are separated by two (or more) erosion surfaces, ArchPy assumes that the contact belongs to the younger erosion surface. This is shown in

Figure 5D at B4 between the orange and dark green units. In the pile, these two units are separated by two other units (blue and yellow) characterized by their erosional top. As yellow is younger than blue, the equality point is attributed to yellow.

All boreholes are then processed this way, and HD are extracted and assigned to respective surfaces. It is relevant to note that this step is completely automated in ArchPy and the only required inputs are the SP, the boreholes, and the simulation grid.

2.6 Simulation of the Surfaces and Units

Using the HD and the information provided by the SP, ArchPy performs a 2D simulation (interpolation) for each surface of the SP over the complete domain. Simulations are generally performed conditionally on HD, but unconditional simulations are also possible by defining a mean altitude.

The surfaces are simulated successively from the oldest to the most recent (for a hierarchic level). After having simulated a surface, we apply the ER (see **Section 2.3**). The surfaces are also simulated hierarchically, which implies that surfaces of higher order (main units) are simulated before those of lower order (subunits). For example, in the case of the pile in **Figure 2**, ArchPy first computes the surfaces of the top of A, B, and C, and only after, the surfaces of the top of B1 and B2 are computed inside the unit B. It means that no surface of the top of these subunits can go above or below the limits of B unit, even if it is an erode surface. The other lower hierarchical units (if present) are simulated following the same strategy. The top unit surface is not simulated since it must be equal to the digital elevation model (DEM) for consistency; it is then simply set as equal to the DEM. Equivalently, as lower limits of all units are defined by the top of the underlying unit, the bottom last unit must be defined. In ArchPy, this is done by setting it to the bottom of the simulation domain.

In the SP, the user must indicate an interpolation method for each layer among the following choices: simple and ordinary kriging (SK and OK, resp., Chilès and Delfiner, 2009), multi-Gaussian random functions with or without inequalities (GRF, Chilès and Delfiner, 2009), basic 2D interpolation methods (linear, nearest neighbors, and cubic) using *SciPy* (Virtanen et al., 2020), and, finally, direct sampling multiple-points statistics algorithm with or without inequalities (MPS, Mariethoz et al., 2010; Straubhaar and Renard, 2021). For the multi-Gaussian simulation methods, a normal-score transform can be applied automatically to the HD if they do not follow a Gaussian distribution. Most of the methods are taken from the Python module *Geone* that provides a set of geostatistical and MPS modeling tools. For each method, the user has to provide the set of required parameters. For example, for the kriging or multi-Gaussian simulations, a variogram model must be provided. The inference of the parameters can be done manually or automatically if sufficient data are available using the *Geone* toolbox. For the MPS approach, a training image and the relevant parameters also need to be provided. Anisotropy can be easily modeled by choosing appropriate variogram or MPS parameters.

Once all the surfaces of the SP are defined, it is straightforward to define the volumes representing the units (unit domains), knowing the top and bottom surfaces for each unit according to the ER. These volumes are discretized by defining which cells are intersected by the surfaces for each (x, y) location. All the cells lying vertically between these 2 cells are assigned to the unit domain. Concerning the intersected cells, they belong to the unit domain only if they go above (or below) the middle of the cell for the top surface (or bot resp.).

Figure 6 shows two realizations of the unit domains. The realizations are conditioned to the borehole data and the stratigraphic pile. The effect of using subunits is clearly visible: the B2 (middle green) top surface (which is set as erode) does not cross-cut unit A (as expected) in the front cross-section of both realizations. This approach allows for representing the uncertainty of the position and extension of the units. By running a large number of realizations, the uncertainty can be quantified: for example, probability maps can be produced for each unit by post-processing those results.

2.7 Simulation of the Facies

Once the stratigraphic unit volumes are defined, it is possible to fill these volumes with different facies or lithologies using different geostatistical methods. The simulation takes place in each unit independently, even if the same facies is present in different units, as in the example in **Figure 2** where the sand appears in B and D units. This means that only the sand HD located inside the unit D will be taken into account when simulating the facies (sand) inside unit D. If a certain facies HD, which does not belong to a specific unit, is present inside its domain (e.g., a sand HD in unit C, **Figure 2**), these HD will be ignored. In such cases, warnings are issued by the software. This situation leads to inconsistencies with facies HD that principally reflect a probable mismatch in the HD (false geological interpretation) or in the geological concept.

As facies (resp. property) simulations are dependent on unit (resp. facies) simulation results, at least one facies (resp. property) simulation must be done for each unit (resp. facies) realization. It means that if the modeler chooses 100 unit realizations and 100 facies simulations, a total of 100 × 100 simulations will be generated. The same logic applies for the property simulations.

In the SP, the user must define one simulation method for each unit. The choices available in the current version of the code are as follows: homogeneous (one unique facies for the whole unit), sequential indicator simulations (SIS, Journel, 1983; Journel and Isaaks, 1984), Truncated pluri-Gaussians (TPG, Loc'h et al., 1994; Mariethoz et al., 2009; Armstrong et al., 2011), multiple-points statistics (MPS, Mariethoz et al., 2010; Straubhaar and Renard, 2021), and sub-pile which indicates that the unit will be populated by another pile containing subunits (e.g., PB in **Figure 2**).

Thus, multiple facies simulation techniques can be used to assess the uncertainty. Note that it is rather straightforward to switch from one method to another. This capability allows us to cover a broad uncertainty space by providing the user with different simulation methods within the same framework. Hence, if little geological knowledge is available for the spatial distribution of the facies within a unit, SIS can be used since little

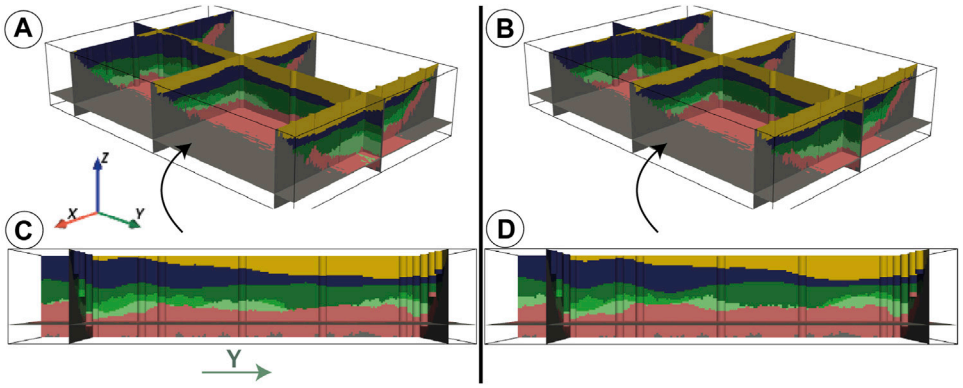

FIGURE 6 | Two realizations **(A,B)** of the units (1st step of the ArchPy simulations) for the synthetic case. **(C)** and **(D)** are lateral view of the realizations **(A)** and **(B)**, respectively. The colors of the units are those defined in the stratigraphic pile of **Figure 2**. They are presented with cross-sections for visualization purposes. Vertical exaggeration = 3x.

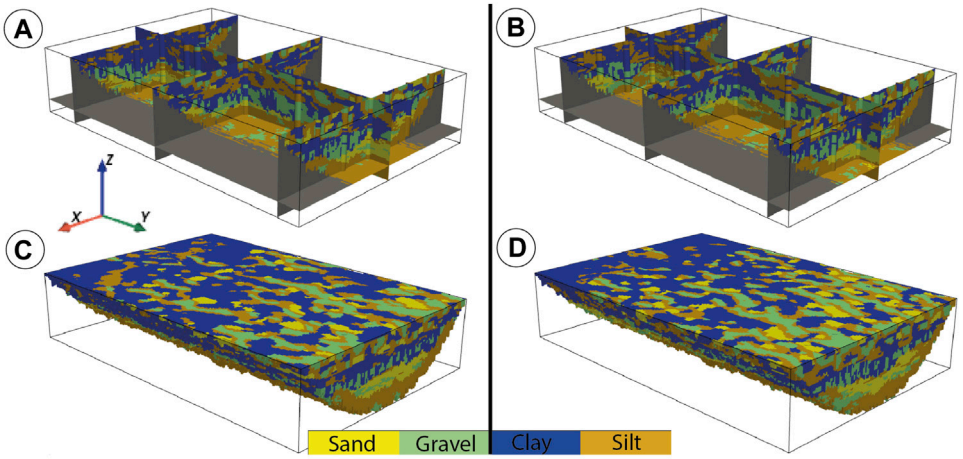

FIGURE 7 | Two realizations **(A,B)** of the lithofacies (2nd step of the ArchPy simulations) for the synthetic case. **(C,D)** are 3D bloc realizations of **(A)** and **(B)**, respectively. The simulations used the respective corresponding units of **Figure 6** as simulation domains. Vertical exaggeration = 3x.

user inputs are required, while if there is more detailed geological knowledge available, other methods can be used, such as TPG or MPS, if an analog geological concept can be defined (e.g., training image). All the geostatistical methods used for this step are included inside *Geone* except for the TPG that are directly included inside ArchPy with various tools to define the truncated flag or estimate the variogram parameters of the underlying multi-Gaussian random fields.

Figure 7 shows two realizations of lithologies according to the unit realizations (**Figure 6**). The spatial variability of the lithologies is significant despite the fact that only four lithologies have been defined. This is mainly due to the combination of structural heterogeneity coming from the stratigraphic units and the lithology distribution within these units. This allows the exploration of many different plausible realities that are consistent with the HD and the concept. We can also observe the non-stationarity in unit D where the channels are sparser in the back than in the front (along the *y* axis). Indeed, it is important to mention that most of the facies simulation methods available in ArchPy can be non-stationary, allowing a much better representation of geological trends and exploration of the uncertainty.

2.8 Simulation of the Properties

Once the facies are simulated, the simulation of the properties is straightforward and requires little input. Indeed, there are only two requirements: define the properties that must be simulated and define how to simulate them (method and parameters). The available methods for the moment are multi-Gaussian Random Fields (GRF) or homogeneous. If a GRF is used, two methods can be used to generate them: fast Fourier transform moving average (FFT, Ravalec et al., 2000) and sequential Gaussian simulation (SGS, Deutsch and Journel 1992). As the FFT method needs to perform the simulations on an entire grid, it can be less effective than SGS, especially if there are many lithologies and units. It is important to note that properties are simulated for a given

lithofacies, independently and sequentially for each unit. For example, if we have a facies which is present in multiple units (consider the sand in **Figure 2**), the properties will be simulated first in the sand occurrences of the top unit and sequentially in the other occurrences in underlying units. This allows us to avoid spurious correlations that can arise if we consider the whole sand domains at once. Indeed, if sands of different units are in contact, they should not be considered as part of the same entity and thus should not be simulated together.

Conditional simulations are available with punctual data (x, y, z and property value). As some methods require at most one HD per cell, values that lie in the same cell are averaged if necessary.

2.9 Implementation of ArchPy

The ArchPy methodology is coded in an open-source Python code[1]. The code is designed using an object-oriented approach. All the concepts described earlier are implemented using classes of objects designed to match the concepts and to facilitate their use. Most of the data imports and exports are based on simple text files. The geostatistical kernel of ArchPy is based on the Geone[2] Python library. For the visualization, ArchPy integrates some functionalities to produce various figures (e.g., stratigraphic units at a specific hierarchical level, only specified units, and cross-sections). These plots are generated mainly using PyVista[3]. If needed, data can also be exported in a vtk format for further use. Post-processing tools are also provided, to estimate, for example, the probability of encountering a specific unit (facies) or estimate the excepted value of a property over a part of the simulation domain. The structure and the principles underlying the ArchPy code are designed to allow the user to script the construction of the geological model in a very flexible manner. This also facilitates the coupling of ArchPy with any forward or inverse simulator. Some example Python notebooks are given on the online repository of the code.

3 A FIRST FIELD APPLICATION

3.1 The Upper Aare Valley

The upper Aare Valley (**Figure 8B**) is a Quaternary alpine valley located between the cities of Thun and Bern in Switzerland with a complex and rich geological history due to its proximity to the Alps (Kellerhals et al., 1981; Haeuselmann et al., 2007). Previous studies on Quaternary deposits, on this particular site (Schlüchter, 1989; Preusser and Schlüchter, 2004) or at a regional scale (Preusser et al., 2011; Graf and Burkhalter, 2016), have shown the complex relations occurring between multiple depositional and erosional processes (mainly glacial, glacio-fluvial, and glacio-lacustrine). This led to the valley being incised and filled with a wide variety of sediments and facies (tills/moraines, fluvial gravels, glacio-lacustrine deposits, lake deposits, alluvial cones, etc.) explaining the great heterogeneity of this type of deposit. Two main aquifers have been identified in the valley: a superficial one which is actively used for drinking water supply, shallow geothermal energy, and some local industries and a deep one that is poorly known due to its higher depth (only few boreholes have reached it). **Figure 8B** shows that the superficial aquifer is mainly composed of the Aare gravels, the Late Glacial alluvial deposits, alluvial cones, and the Münsingen gravels.

A major hydrogeological synthesis of the valley was undertaken at the end of the 1970s and at the beginning of the 1980s (Kellerhals et al., 1981). Since then, additional data have been collected (Schlüchter (1989); Preusser and Schlüchter (2004)) for various projects in different parts of the valley, but no new hydrogeological synthesis has been assembled and published. Among the new data, the Swiss Geological Survey has systematically gathered the borehole data for Quaternary sediments in several pilot sites and homogenized the data and terminology (Volken et al., 2016). This data set includes around 800 digitized boreholes in the upper Aare Valley. A geological model of the valley filling has also been produced by the Swiss Geological Survey to illustrate how those data can be used. In addition, a valley scale towed Transient Electromagnetic Survey has been acquired and published in 2021 (Neven et al., 2021) by the University of Neuchâtel to better constrain and characterize the aquifer dimensions and its internal heterogeneity.

3.2 Modeling Area and Borehole Dataset

To illustrate the ArchPy approach, we chose an area with a high density of boreholes located in the south of the Valley (**Figure 8B**). The extent of this area is given by its coordinates in the CH 1903+ - LV95 system: lower left corner, 2′611′000 m/1′178′000 m, and upper-right corner, 2′613′000 m/1′182′000 m.

The depth of the boreholes rarely exceeds 50 m; they do not reach the deepest aquifer and stay in the shallow one that is 50–60 m thick in this part of the valley (Kellerhals et al., 1981). The data set contains a large part of the boreholes that have been drilled over decades in the area (Volken et al., 2016). Each borehole is described in terms of intervals with information about the granulometry (lithofacies), the units encountered, the quality of the interpretations, etc. However, unit data can be missing, contrary to granulometry data, meaning that for many boreholes, only lithofacies data are available. Granulometry information is described with one, two, or up to three different grain sizes, each defined with a USCS classification code (Casagrande, 1948). No hydraulic conductivity data have been taken into account.

The boreholes (133 in total) intercept a total of four major stratigraphic units: Aare young gravel (YG, Holocene), Late Glacial alluvial deposits (LGA, Holocene), late glacio-lacustrine deposits (LGL, Holocene), and Late Glacial Till (LGT, late Pleistocene). The LGT appears only on two boreholes on the southern part of this section of the aquifer and will therefore be difficult to model. YG and LGA are the most present units and constitute the largest part of the shallow aquifer in this area while LGL and LGT are more scattered and can be seen as its bottom.

Note that this stratigraphic pile is simplified given that 23 major stratigraphic units can be distinguished on the entire valley

[1]http://www.github.com/randlab/ArchPy.
[2]http://www.github.com/randlab/geone.
[3]https://www.pyvista.org/.

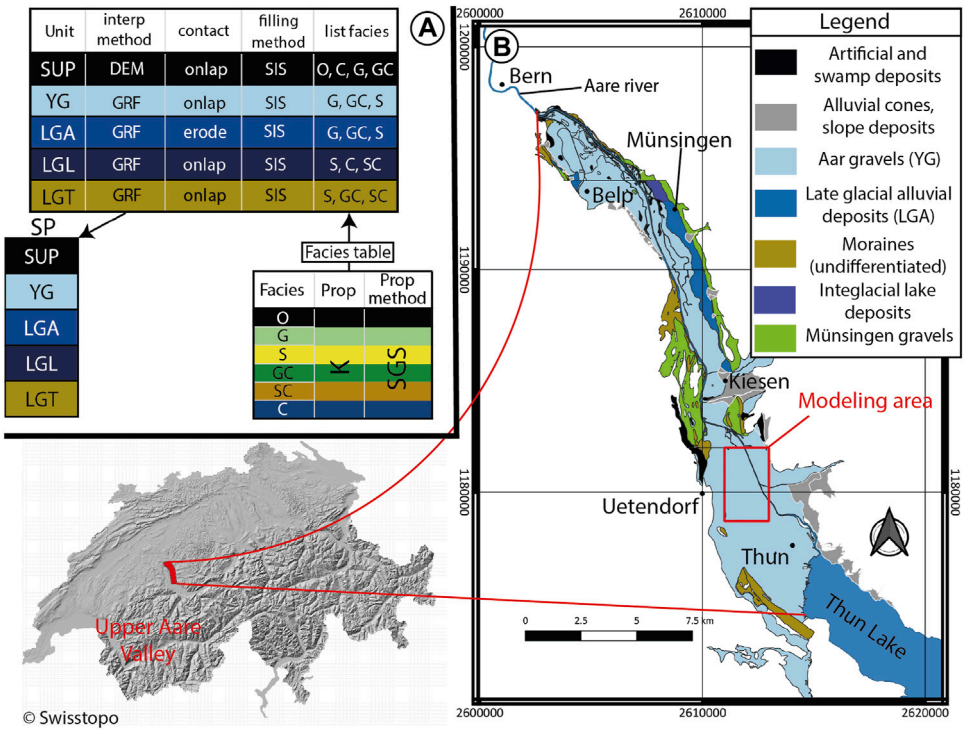

FIGURE 8 | (A) Schematic stratigraphic pile used for the ArchPy simulations of the local geological models in the upper Aare Valley. It shows the different units that have been taken into account, their stratigraphic relations (which is above/below (i.e., younger/older)), the nature of its contact (erode or onlap), the interpolation methods used or the lithofacies (see **Table 2**), and the property to simulate (hydraulic conductivity (*K*) using sequential Gaussian simulations (SGS)). See **Section 2.2** for more details. **(B)** Situation and simplified geological map of the upper Aare Valley. Only unconsolidated deposits that are near the aquifer are shown. The moraines are represented undifferentiated. The coordinates are presented in CH 1903+ - LV95 (epsg: 2056). All the data used come from the Swiss Geological Survey (Swisstopo).

(Volken et al., 2016). However, most of these units are absent in the modeled area.

3.3 Modeling Settings

The extension of the simulation domain is 2 km in the W-E direction and 3.3 km in the N-S direction. The elevation ranges from 520 to 570 m a.s.l and the resolution is 15 × 15 × 1 m, which implies a number of cells of 134 × 220 × 60 ($nx \times ny \times nz$). The top of the domain is defined according to the DEM of Switzerland with a resolution of 25 × 25 m (DHM25, Swisstopo) and the bottom is defined by a raster map of the bedrock elevation (TopFels25, Swisstopo), also with a resolution of 25 × 25 m, both freely distributed by the Swiss Federal Office of Topography.

The units were defined mainly on the basis of the HD and the geological knowledge of this area (Kellerhals et al., 1981). Five units were recognized (**Figure 8A**). A superior unit (SUP) was added which includes superficial (soil and peat) and artificial (anthropogenic) deposits. No subunit has been defined as such data are not available in this actual dataset. The top surfaces of the units have been modeled with GRF to take the effect of inequalities into account, except for the SUP top surface which was set to the DEM as it is the most superficial unit. The associated covariance models (variograms) were estimated using an automatic fitting method (least squares optimization) on the HD. The optimized parameters are shown in **Table 1**. Most

TABLE 1 | Covariance model parameters (C: contribution and r: range) used for the surface interpolation of each surface. All models are isotropic, except the LGT one with an orientation of N-S for the major axis). No covariance model was fitted for SUP as its surface is defined by the DEM. Subscripts *exp* and *sph* indicate exponential and spherical covariance models.

Unit	r_{sph} [m]	C_{sph} [m^2]	r_{exp} [m]	C_{exp} [m^2]	Nugget [m^2]
YG	2,986	8.9	5,000	17.8	0
LGA	2,854	24.8	4,846	49.5	1.0
LGL	2,531	19.0	3,942	38.1	1.0
LGT	(2000, 4,000)[a]	200	-	-	-

[a](*Ranges in x and y directions, resp.*).

TABLE 2 | Grouped USCS codes. It indicates in which group code the classical USCS groups are rearranged.

Grouped code	USCS classical groups
O	(OH, OL, Pt)
G	(G, G-GM, GW-GM, GP-GM, GP, GW, GP-GC, G-GC)
S	(S-SM, S, SP-SM, SP, SW, S-SC, SP-SC, SW-SM)
GC	(GM, GC, GC-GM)
SC	(SM, SC, SC-SM)
C	(ML, CL-ML, CL, CM, CH)

TABLE 3 | Covariance model parameters (C: contribution and r: ranges in *x*, *y*, and *z* directions) manually adjusted and used for the SIS of each unit. For units where the number of data points was too low (LGL and LGT), a default model was taken ("Default" row). The ranges are given in the three main axis directions without any rotation (*x* axis goes toward E and *y* axis toward N). Subscripts *exp* and *sph* indicate exponential and spherical covariance models.

Unit (lithofacies)	r_{sph} [m]	C_{sph} [m^2]	r_{exp} [m]	C_{exp} [m^2]	Nugget [m^2]
SUP (O)	(200, 400, 5)	0.15	-	-	0.1
SUP (C)	(200, 200, 5)	0.11	-	-	0
SUP (G)	(200, 200, 5)	0.25	-	-	0
SUP (GC)	(300, 100, 4)	0.20	-	-	0
YG (G)	-	-	(50, 50, 15)	0.22	0
YG (GC)	(200, 200, 1)	0.06	(300, 200, 6)	0.09	0
YG (S)	(200, 200, 15)	0.03	(200, 200, 1)	0.03	0
LGA (G)	(100, 200, 8)	0.12	(100, 200, 20)	0.12	0
LGA (GC)	(100, 150, 15)	0.10	(100, 150, 15)	0.11	0
LGA (S)	(50, 100, 15)	0.13	-	-	0.01
Default	(100, 100, 10)	Variable[a]	-	-	0

[a]*Variance was adjusted according to lithofacies proportions.*

TABLE 4 | Covariance model parameters (C: contribution and r: ranges in *x*, *y*, and *z* directions) used for the property simulations. Subscript *exp* indicates an exponential covariance model.

Lithofacies	r_{exp} [m]	C_{exp} [m^2]
O	(50, 50, 2)	0.1
G	(50, 50, 2)	0.25
S	(50, 50, 2)	0.16
GC	(50, 50, 2)	0.2
SC	(50, 50, 2)	0.2
C	(50, 50, 2)	0.2

of the surfaces were defined as onlap except the LGA top surface which represents a former terrace of the Aare river, generally deposited at a slightly higher altitude than YG (Kellerhals et al., 1981).

In the HD, the lithofacies are described by up to three different grain sizes. We chose to only take the most present one for each layer and we also grouped certain similar USCS codes (**Table 2**) to reduce the number of lithofacies to 7: others (O), gravel (G), sand (S), clayey gravel (GC), clayey sand (SC), and clay (C). The other facies regroup superficial codes such as OH or Pt. Facies were then considered within a unit if their proportions exceed 5% (inside that specific unit).

For the sake of simplicity in that example, all the units were filled using SIS. Prior variography analysis on the lithofacies HD shows significant variability which required the SIS variograms to be fitted manually; the chosen parameters are given in **Table 3**. Only the hydraulic conductivity *K* property has been simulated for that example using the covariance models given in **Table 4**. Note that adding other properties is possible and very simple since only the interpolation method and the covariance models (for each facies) are required.

The ArchPy Aare model was run several times to illustrate its applicability for uncertainty estimation. In that example, we generated 10 simulations of the stratigraphic units. For each stratigraphic unit simulation, we generated 10 facies simulations. Finally, for each combined realization, we generated 1 unconditional simulation for *K*. This procedure resulted in a total of 100 simulations ($10 \times 10 \times 1$). The code allows us to proceed in this manner, but it also permits us to simulate all the components successively for each realization (units, facies, and properties). These different modes of simulation can be used for quantifying the impact of these different sources of uncertainty on the distribution of the properties but also on their groundwater flow or geophysical responses.

Figures 9–12 show the results of ArchPy simulations conditioned to the borehole data. The figures illustrate the type of heterogeneity and complexity that can be modeled rather simply using the ArchPy approach. For example, the two unit realizations (**Figures 9A,B**) differ significantly while being consistent and honoring both the borehole data. This variability is important for quantifying the uncertainty. To visualize that part of the uncertainty, ArchPy allows the user to compute the probability of observing a specific unit. **Figure 10** shows in yellow the locations where it is quite certain that a given unit is present and with which thickness. For example, **Figure 10A** shows that the unit YG is well constrained in the eastern and northern part of the domain due to the important number of boreholes that reach it. The unit LGA seems to be more present in the southern part of the area (**Figure 10B**), thinner than the unit YG and almost absent (or very thin) in the north. The unit LGT (**Figure 10C**) does not display such trends and has a more uncertain distribution, probably mainly due to a lack of data (shallow boreholes).

Lithofacies simulations are shown in **Figures 9C**, **11** and are the results of the filling of the simulations shown in **Figure 9A**. These simulations honor both the borehole data and geometry of the stratigraphic units. As for the stratigraphic units, it is possible to compute and produce figures showing the probability of occurrence of each facies.

Finally, two simulated hydraulic conductivity fields are shown in **Figure 12**. They display a broad range of values that is expected for this geology and that honor all the borehole data. It also shows the complex relations between the property values, the stratigraphic units, and the lithofacies. The variability between the realizations suggests a strong heterogeneity that would have been extremely difficult to model properly without the hierarchical approach (e.g., Feyen and Caers 2006; Zappa et al., 2006; Zech et al., 2021). The mean of the logarithm of *K*

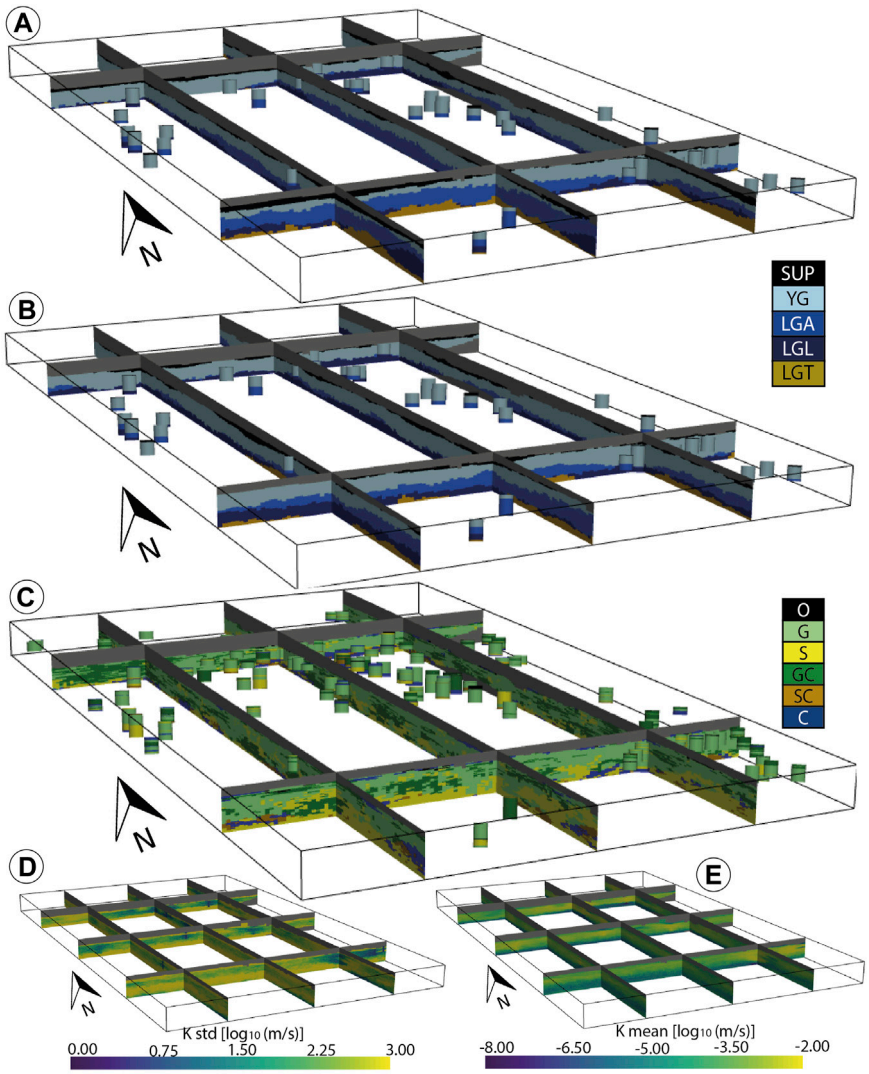

FIGURE 9 | Aare aquifer results obtained for **(A,B)** two unit realizations, **(C)** one facies realization (within model **(A)**), and **(D,E)** K standard deviation and mean simulated along the 100 models; units are in log10 (m/s). Vertical exaggeration = 3x.

(**Figure 9E**) highlights the location of the aquifer where the values are likely to be especially high. These locations also coincide with those where the standard deviation (**Figure 9D**) is low, indicating that the property values are better defined inside the aquifer than outside. Elsewhere, the standard deviation values may be quite high, easily reaching 2 $\log_{10}$ (m/s), indicating an uncertainty of up to two orders of magnitude. The variance is low around the boreholes as expected.

4 DISCUSSION

One of the most important novel features of the ArchPy approach is the extended concept of stratigraphic pile (SP) as compared to the findings of Calcagno et al. (2008), for example. This concept has been shown to be an effective way to formulate all the geological knowledge into one entity (practically, a Python object). Thanks to this representation, it is easy to embed multiple SPs inside other SPs and to simulate the units to any level of hierarchy and do this without any particular restrictions. By including various interpolation and simulation methods which can be applied independently for each unit, lithofacies, and property, the ArchPy approach offers a high flexibility to the user who can adapt the methods to the quantity of available data and the complexity that he needs to represent for a specific site. In addition, the use of inequality data that are automatically derived from the SP and the borehole data allows ArchPy to extract a larger amount of information from boreholes than what is usually done in alternative geo-modeling tools.

The results obtained for the upper Aare Valley illustrate the type of stochastic models that can be easily and rapidly constructed for Quaternary deposits using the ArchPy

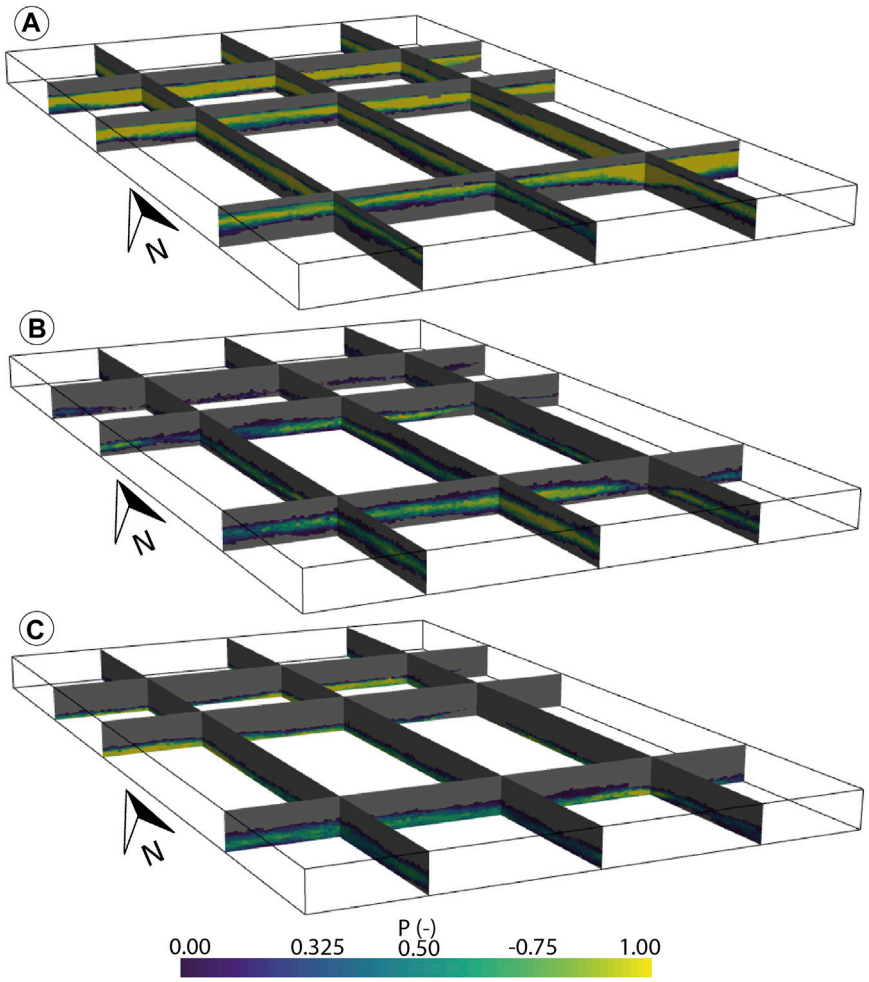

FIGURE 10 | Probability of occurrence along the 100 models for units YG **(A)**, LGA **(B)**, and LGL **(C)**. Vertical exaggeration = 3x.

approach. Due to the simple assumptions made about the geology and the concepts (simplified SP and use of SIS to fill the units), some aspects of the proposed method could not be illustrated in this example. Several limitations in the data set were also identified. For example, the LGT unit is not well constrained because only a few boreholes reach it. Indeed, most boreholes in this area are drilled for hydrogeological purposes (Kellerhals et al., 1981), and local communities are generally not interested in reaching the LGT unit because of its lower hydraulic conductivity as compared to the YG or LGA unit. A sampling bias is also expected in the lithofacies inside the LGL and LGT units. Indeed, we observe that the simulations of these units tend to have more sand (S code) than expected in glacio-lacustrine or till deposits. This can be due to a sampling bias in the borehole database because the areas of high permeability are preferentially drilled while clay and silt areas are generally avoided. Since the simulated proportions of lithofacies are conditioned on the HD, the lithofacies simulations can reflect this bias. It is, however, possible to correct it by imposing proportions that differ from those of the HD, but further secondary information should then be used to guide the simulations. One possible method to correct that bias could also be to use geophysical data, as we will discuss more in detail below.

Another important feature of the ArchPy approach is that it allows quantifying the uncertainty by generating an ensemble of models. The uncertainty can be evaluated at any desired hierarchical level among the units, subunits, sub-subunits, lithofacies, or properties. The uncertainty on the geological concept and stratigraphic pile (SP) itself can also be evaluated. This type of uncertainty was not covered in the example of the upper Aare valley, where the concept was simple. But there are situations in which different geological concepts or different geostatistical models for the different components of the SP are plausible. Using the ArchPy approach and its scripting possibilities, it is straightforward to automatically explore all these possibilities and generate an ensemble of models that covers this uncertainty.

Due to its recent existence, ArchPy still lacks some interesting features and has several limitations. First of all, it is only usable through scripts in Python, which may prevent a certain number

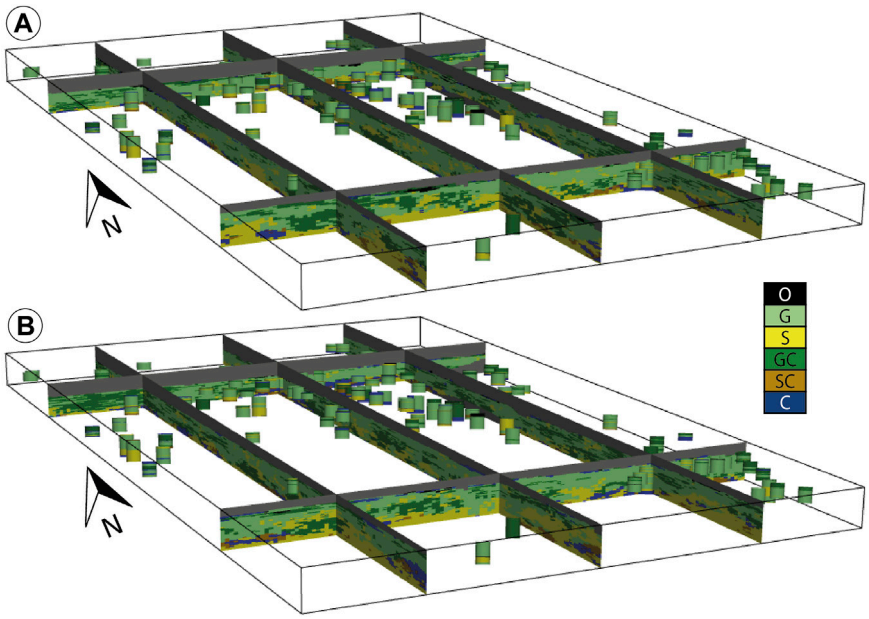

FIGURE 11 | (A, B) Two facies realizations that are the results of the filling of the unit realization in **Figure 9A**. Boreholes show the spatial distribution of the HD.

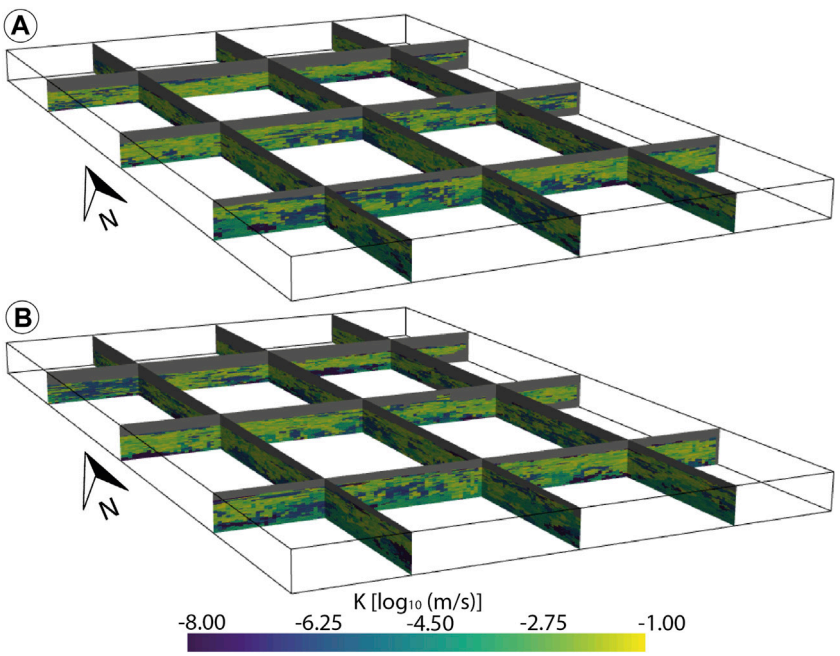

FIGURE 12 | (A, B) Two property realizations made on the two facies realizations in **Figure 11**.

of people from using it. However, examples are provided and can easily be edited; therefore, it is not necessary to be an expert in Python to use ArchPy. This approach has many advantages such as ensuring efficient model update when new data are acquired or accurately documenting the model construction steps. In future, one could construct a graphical user interface (GUI). The main limitation of ArchPy for the moment is that it assumes that the boundaries of the stratigraphic units can be modeled using functions that can be represented on a 2D grid. Therefore, ArchPy cannot, at the moment, represent overturned folds. It also does not include faults. We consider that these situations are extremely rare in Quaternary environments; adapting ArchPy to

account for these structures would be feasible but is not currently a priority. ArchPy is also limited by the set of geostatistical methods that are proposed. It implies, for example, that the code may be slow if the number of borehole data and inequalities is important. We will continue to optimize the methods as much as possible. Concerning the simulation of non-Gaussian data, a normal-score transform should always be considered before using any GRF method. But the user must be aware that this kind of transformation is only suitable when the number of data is sufficient (a Cumulative Function Distribution can be built). When data are sparse, it is simply possible to assume that the data follow a normal distribution or the use of other available methods can be considered (MPS). Moreover, GRF simulations, performed on data normally transformed, do not guarantee that the covariance will be preserved in the original data space. The use of more advanced simulation methods such as Direct Sequential Simulation (Soares, 2001) could be a solution. The final note is about the trends in the data (surface elevations or facies proportions). Such behaviors can be modeled but not in a fully automated manner as it requires user-inputs (e.g., local facies proportions over the domain). These must be computed or derived externally. However, implementing such routines in future updates is straightforward.

Because it is simple and easy to run ArchPy automatically, it is straightforward to conduct parameter sensitivity analysis. We even suggest that the parameters should be tested as well as the stratigraphic pile, including all its various geostatistical components, using cross-validation. This approach should be used to test and compare different alternative SPs. The procedure consists in splitting the borehole data and applying a K-fold cross-validation approach as we discussed in a previous study (Juda et al., 2020). An ensemble of models is generated to predict the units, lithologies, and properties at the location of a subset of the boreholes (removed from the HD). A score can then be computed to compare the quality of the stochastic predictions with the actual data. We plan to incorporate cross-validation frameworks inside the ArchPy architecture.

To go a step further, ArchPy is already coupled with several geophysical and groundwater flow simulation tools Cockett et al. (2015); Bakker et al. (2016). Property models generated using ArchPy (e.g., resistivity, gravity, storativity, and hydraulic conductivity) can be passed to forward models. The outputs are retrieved and compared with real field measurements which are then used to adapt the ArchPy models to reduce the misfit between both actual and simulated data. For example, this adaptation could be done in a Monte Carlo scheme (Tokdar and Kass 2010) or with ensemble methods (Chen and Oliver, 2012). This approach opens the way toward geologically constrained joint inversion involving different forward models.

5 CONCLUSION

The ArchPy approach that is proposed in this study combines many techniques that are well known (geostatistical simulation techniques for continuous or categorical variables). One important novelty is to formally separate the description of the list of tasks that are required to construct the model and the construction of the model itself. This is done by embedding all the geological and geostatistical knowledge in an object called "stratigraphic pile." Based on this formalism, a piece of software can be constructed that can automate all the tedious tasks of the model construction. The Python module that implements the ArchPy approach allows the fast and reproducible creation of an ensemble of stochastic models respecting both the conditional data and the user inputs (geological concepts). The only inputs required are the digital elevation model, the borehole data, and how to interpolate and simulate the different components of the model (surfaces, lithofacies, and properties); the rest is up to ArchPy. The simulations take place during three main phases: simulation of the units, of the lithofacies, and then of the properties. Each step depends on the previous ones. A major novelty is that the stratigraphic pile allows defining a hierarchical stratigraphy and therefore allows modeling automatically consistent subunits of any hierarchical level within units of higher levels. The code allows quantifying uncertainty using a sound geostatistical model. It also allows updating the model easily when new data are available or embedding the model construction into an inverse procedure. The code is open-source and freely distributed. Due to its open-source nature, the coupling with other software is facilitated. It opens the doors to an easier and more accessible geological modeling of Quaternary aquifers.

DATA AVAILABILITY STATEMENT

The github repository of the code as well as the synthetical dataset used can be found on http://www.github.com/randlab/ArchPy. The data concerning the Aar model are available on request at the Swiss geological survey, Swisstopo.

AUTHOR CONTRIBUTIONS

LS: ArchPy software development and data preparation. JS: Geone library development and support. LS and PR: conceptualization and methodology. LS: original draft preparation and editing. PR and JS: review and editing. PR: funding acquisition, supervision, and project administration. All authors contributed to the article and approved the submitted version.

FUNDING

This research is funded by the Swiss National Science Foundation under the contract 200020_182600/1 (PheniX project).

ACKNOWLEDGMENTS

The authors would like to greatly thank Alexis Neven for the discussions about the design of ArchPy, his active use of the code,

and, thus, for his help in finding most of the bugs and issues. The authors also thank all the partners of the project for their inputs and discussions and, in particular, the Swiss Geological Survey (Swisstopo) for having prepared and shared the dataset, the Canton of Bern, and GEOTEST AG as well as KELLERHALS + HAEFELI AG for their support. The authors would also like to thank two reviewers who helped to improve the quality of this article.

REFERENCES

Aigner, T., Asprion, U., Hornung, J., Junghans, W.-D., and Kostrewa, R. (1996). Integrated Outcrop Analogue Studies for Triassic Alluvial Reservoirs: Examples from Southern germany. *J. Pet. Geol.* 19, 393–406. doi:10.1111/j.1747-5457.1996.tb00446.x

Armstrong, M., Galli, A., Beucher, H., Loc'h, G., Renard, D., Doligez, B., et al. (2011). *Plurigaussian Simulations in Geosciences*. Springer Science & Business Media.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting Modflow Model Development Using python and Flopy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Bennett, J. P., Haslauer, C. P., Ross, M., and Cirpka, O. A. (2019). An Open, Object-Based Framework for Generating Anisotropy in Sedimentary Subsurface Models. *Groundwater* 57, 420–429. doi:10.1111/gwat.12803

Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., et al. (2011). *Large-Scale Inverse Problems and Quantification of Uncertainty*. John Wiley & Sons.

Bridge, J. S. (2009). *Rivers and Floodplains: Forms, Processes, and Sedimentary Record*. John Wiley & Sons.

Calcagno, P., Chilès, J. P., Courrioux, G., and Guillen, A. (2008). Geological Modelling from Field Data and Geological Knowledge. *Phys. Earth Planet. Interiors* 171, 147–157. doi:10.1016/j.pepi.2008.06.013

Casagrande, A. (1948). Classification and Identification of Soils. *T. Am. Soc. Civ. Eng.* 113, 901–930. doi:10.1061/TACEAT.0006109

Chen, Y., and Oliver, D. S. (2012). Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother. *Math. Geosci.* 44, 1–26. doi:10.1007/s11004-011-9376-z

Chilès, J.-P., and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons.

Cockett, R., Kang, S., Heagy, L. J., Pidlisecky, A., and Oldenburg, D. W. (2015). Simpeg: An Open Source Framework for Simulation and Gradient Based Parameter Estimation in Geophysical Applications. *Comput. Geosciences* 84, 142–154. doi:10.1016/j.cageo.2015.09.015

Comunian, A., De Micheli, L., Lazzati, C., Felletti, F., Giacobbo, F., Giudici, M., et al. (2016). Hierarchical Simulation of Aquifer Heterogeneity: Implications of Different Simulation Settings on Solute-Transport Modeling. *Hydrogeol. J.* 24, 319–334. doi:10.1007/s10040-015-1343-1

Comunian, A., Renard, P., Straubhaar, J., and Bayer, P. (2011). Three-dimensional High Resolution Fluvio-Glacial Aquifer Analog - Part 2: Geostatistical Modeling. *J. Hydrology* 405, 10–23. doi:10.1016/j.jhydrol.2011.03.037

de la Varga, M., Schaaf, A., and Wellmann, F. (2019). GemPy 1.0: Open-Source Stochastic Geological Modeling and Inversion. *Geosci. Model Dev.* 12, 1–32. doi:10.5194/gmd-12-1-2019

Deutsch, C. V., and Journel, A. G. (1992). *GSLIB. Geostatistical Software Library and User's Guide*. New York: Oxford University Press.

Dubrule, O., and Kostov, C. (1986). An Interpolation Method Taking into Account Inequality Constraints: I. Methodology. *Math. Geol.* 18, 33–51. doi:10.1007/BF00897654

Feyen, L., and Caers, J. (2006). Quantifying Geological Uncertainty for Flow and Transport Modeling in Multi-Modal Heterogeneous Formations. *Adv. Water Resour.* 29, 912–929. doi:10.1016/j.advwatres.2005.08.002

Ford, G. L., and Pyles, D. R. (2014). A Hierarchical Approach for Evaluating Fluvial Systems: Architectural Analysis and Sequential Evolution of the High Net-Sand Content, Middle Wasatch Formation, Uinta Basin, Utah. *Bulletin* 98, 1273–1303. doi:10.1306/12171313052

Freulon, X., and de Fouquet, C. (1993). "Conditioning a Gaussian Model with Inequalities," in *Geostatistics Tróia '92*. Editor A. Soares (Dordrecht: Springer Netherlands), Quantitative Geology and Geostatistics), 1, 201–212. doi:10.1007/978-94-011-1739-5_17

Graf, H. R., and Burkhalter, R. (2016). Quaternary Deposits: Concept for a Stratigraphic Classification and Nomenclature-An Example from Northern Switzerland. *Swiss J. Geosci.* 109, 137–147. doi:10.1007/s00015-016-0222-7

Haeuselmann, P., Granger, D. E., Jeannin, P.-Y., and Lauritzen, S.-E. (2007). Abrupt Glacial Valley Incision at 0.8 Ma Dated from Cave Deposits in Switzerland. *Geol* 35, 143–146. doi:10.1130/G23094A

Heinz, J. r., and Aigner, T. (2003). Hierarchical Dynamic Stratigraphy in Various Quaternary Gravel Deposits, Rhine Glacier Area (SW Germany): Implications for Hydrostratigraphy. *Int. J. Earth Sci.* 92, 923–938. doi:10.1007/s00531-003-0359-2

Journel, A. G., and Isaaks, E. H. (1984). Conditional Indicator Simulation: Application to a Saskatchewan Uranium Deposit. *Math. Geol.* 16, 685–718. doi:10.1007/BF01033030

Journel, A. G. (1983). Nonparametric Estimation of Spatial Distributions. *Math. Geol.* 15, 445–468. doi:10.1007/BF01031292

Juda, P., Renard, P., and Straubhaar, J. (2020). A Framework for the Cross-Validation of Categorical Geostatistical Simulations. *Earth Space Sci.* 7, e2020EA001152. doi:10.1029/2020EA001152

Kellerhals, P., Haefeli, C., and Tröhler, B. (1981). *Grundlagen für Schutz und Bewirtschaftung der Grundwasser des Kantons Bern Hydrogeologie Aaretal, zwischen Thun und Bern*. Tech. rep. Bern: Wasser- u. Energiewirtschaftsamt des Kantons Bern.

Koltermann, C. E., and Gorelick, S. M. (1996). Heterogeneity in Sedimentary Deposits: A Review of Structure-Imitating, Process-Imitating, and Descriptive Approaches. *Water Resour. Res.* 32, 2617–2658. doi:10.1029/96wr00025

Le Loc'h, G., Beucher, H., Galli, A., and Doligez, B. (1994). Improvement in the Truncated Gaussian Method: Combining Several Gaussian Functions. ECMOR IV - 4th European Conference on the Mathematics of Oil Recovery, Roros, Norway. European Association of Geoscientists & Engineers. doi:10.3997/2214-4609.201411149

Mallet, J.-L. (1989). Discrete Smooth Interpolation. *ACM Trans. Graph.* 8, 121–144. doi:10.1145/62054.62057

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009). Truncated Plurigaussian Simulations to Characterize Aquifer Heterogeneity. *Groundwater* 47, 13–24. doi:10.1111/j.1745-6584.2008.00489.x

Mariethoz, G., Renard, P., and Straubhaar, J. (2010). The Direct Sampling Method to Perform Multiple-point Geostatistical Simulations. *Water Resour. Res.* 46. doi:10.1029/2008WR007621

Matheron, G. (1963). Principles of Geostatistics. *Econ. Geol.* 58, 1246–1266. doi:10.2113/gsecongeo.58.8.1246

Miall, A. (1991). "Hierarchies of Architectural Units in Terrigenous Clastic Rocks and Their Relationship to Sedimentation Rate," in *The Three-Dimensional Facies Architecture of Terrigenous Clastic Sediments, and its Implications for Hydrocarbon Discovery and Recovery*. Editors A. D. Miall and N. Tyler (tulsa, oklahoma: Society for sedimentary geology edn), 3. doi:10.2110/csp.91.03.0006

Neuman, S. P. (1990). Universal Scaling of Hydraulic Conductivities and Dispersivities in Geologic Media. *Water Resour. Res.* 26, 1749–1758. doi:10.1029/wr026i008p01749

Neven, A., Maurya, P. K., Christiansen, A. V., and Renard, P. (2021). tTEM20AAR: A Benchmark Geophysical Data Set for Unconsolidated Fluvioglacial Sediments. *Earth Syst. Sci. Data* 13, 2743–2752. doi:10.5194/essd-13-2743-2021

Preusser, F., Graf, H. R., Keller, O., Krayss, E., and Schlüchter, C. (2011). Quaternary Glaciation History of Northern Switzerland. *E&G Quat. Sci. J.* 60, 282–305. doi:10.3285/eg.60.2-3.06

Preusser, F., and Schlüchter, C. (2004). Dates from an Important Early Late Pleistocene Ice Advance in the Aare Valley, Switzerland. *Eclogae Geol. Helv.* 97, 245–253. doi:10.1007/s00015-004-1119-4

Pyrcz, M. J., and Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling*. Oxford University Press.

Ramanathan, R., Guin, A., Ritzi, R. W., Jr, Dominic, D. F., Freedman, V. L., Scheibe, T. D., et al. (2010). Simulating the Heterogeneity in Braided Channel Belt Deposits: 1. A Geometric-Based Methodology and Code. *Water Resour. Res.* 46. doi:10.1029/2009wr008111

Ravalec, M. L., Noetinger, B., and Hu, L. Y. (2000). The FFT Moving Average (FFT-MA) Generator: An Efficient Numerical Method for Generating and Conditioning Gaussian Simulations. *Math. Geol.* 32, 701–723. doi:10.1023/a:1007542406333

Renard, P., and Courrioux, G. (1994). Three-dimensional Geometric Modeling of a Faulted Domain: The Soultz Horst Example (Alsace, france). *Comput. Geosciences* 20, 1379–1390. doi:10.1016/0098-3004(94)90061-2

Ringrose, P., and Bentley, M. (2016). *Reservoir Model Design*. Springer.

Ritzi, R. W., Dai, Z., Dominic, D. F., and Rubin, Y. N. (2004). Spatial Correlation of Permeability in Cross-Stratified Sediment with Hierarchical Architecture. *Water Resour. Res.* 40. doi:10.1029/2003wr002420

Scheibe, T. D., and Freyberg, D. L. (1995). Use of Sedimentological Information for Geometric Simulation of Natural Porous Media Structure. *Water Resour. Res.* 31, 3259–3270. doi:10.1029/95wr02570

Schlüchter, C. (1989). The Most Complete Quaternary Record of the Swiss Alpine Foreland. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 72, 141–146. doi:10.1016/0031-0182(89)90138-7

Soares, A. (2001). Direct Sequential Simulation and Cosimulation. *Math. Geol.* 33, 911–926. doi:10.1023/a:1012246006212

Straubhaar, J., and Renard, P. (2021). Conditioning Multiple-Point Statistics Simulation to Inequality Data. *Earth Space Sci.* 8, e2020EA001515. doi:10.1029/2020EA001515

Tokdar, S. T., and Kass, R. E. (2010). Importance Sampling: A Review. *WIREs Comp. Stat.* 2, 54–60. doi:10.1002/wics.56

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

Volken, S., Preisig, G., and Gaehwiler, M. (2016). GeoQuat: Developing a System for the Sustainable Management, 3D Modelling and Application of Quaternary Deposit Data. *Swiss Bull. Appl. Geol.* 21, 3–16.

Weissmann, G. S., and Fogg, G. E. (1999). Multi-scale Alluvial Fan Heterogeneity Modeled with Transition Probability Geostatistics in a Sequence Stratigraphic Framework. *J. Hydrology* 226, 48–65. doi:10.1016/s0022-1694(99)00160-2

Wellmann, F., and Caumon, G. (2018). 3-D Structural Geological Models: Concepts, Methods, and Uncertainties. *Adv. Geophys.* 59, 1–121. doi:10.1016/bs.agph.2018.09.001

Zappa, G., Bersezio, R., Felletti, F., and Giudici, M. (2006). Modeling Heterogeneity of Gravel-Sand, Braided Stream, Alluvial Aquifers at the Facies Scale. *J. Hydrology* 325, 134–153. doi:10.1016/j.jhydrol.2005.10.016

Zech, A., Dietrich, P., Attinger, S., and Teutsch, G. (2021). A Field Evidence Model: How to Predict Transport in Heterogeneous Aquifers at Low Investigation Level. *Hydrol. Earth Syst. Sci.* 25, 1–15. doi:10.5194/hess-25-1-2021

Zuffetti, C., Comunian, A., Bersezio, R., and Renard, P. (2020). A New Perspective to Model Subsurface Stratigraphy in Alluvial Hydrogeological Basins, Introducing Geological Hierarchy and Relative Chronology. *Comput. Geosciences* 140, 104506. doi:10.1016/j.cageo.2020.104506

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Earth Science

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/feart.2022.904271

Decision-Support Groundwater Modelling of Managed Aquifer Recharge in a Coastal Aquifer in South Portugal

Kath Standen[1,2]*, *Rui Hugman*[3] *and José Paulo Monteiro*[1,2]

[1]Centro de Ciências e Tecnologias da Água (CTA), Universidade do Algarve, Faro, Portugal, [2]CERIS, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, [3]National Centre for Groundwater Research and Training, Flinders University, Groundwater Modelling Decision Support Initiative (GMDSI), Adelaide, SA, Australia

OPEN ACCESS

Edited by:
Michael Fienen,
United States Geological Survey,
United States

Reviewed by:
Ali Al-Maktoumi,
Sultan Qaboos University, Oman
John Masterson,
United States Geological Survey
(USGS), United States

***Correspondence:**
Kath Standen
k.e.standen@ualg.pt

Specialty section:
This article was submitted to
Hydrosphere,
a section of the journal
Frontiers in Earth Science

Received: 25 March 2022
Accepted: 11 May 2022
Published: 31 May 2022

Citation:
Standen K, Hugman R and
Monteiro JP (2022) Decision-Support
Groundwater Modelling of Managed
Aquifer Recharge in a Coastal Aquifer
in South Portugal.
Front. Earth Sci. 10:904271.
doi: 10.3389/feart.2022.904271

The Vale do Lobo sector of the Campina de Faro aquifer system in the Algarve (Portugal) is at risk of seawater intrusion. Managed Aquifer Recharge (MAR) is being considered to avoid groundwater quality deterioration. Numerical modelling was undertaken to assess the feasibility of several proposed MAR schemes. Although some data is available, many aspects of system behaviour are not well understood or measured. We demonstrate the use of a structurally simple but parametrically complex model for decision-making in a coastal aquifer. Modelling was designed to facilitate uncertainty reduction through data assimilation where possible, whilst acknowledging that which remains unknown elsewhere. Open-source software was employed throughout, and the workflow was scripted (reproducible). The model was designed to be fast-running (rapid) and numerically stable to facilitate data assimilation and represent prediction-pertinent uncertainty (robust). Omitting physical processes and structural detail constrains the type of predictions that can be made. This was addressed by assessing the effectiveness of MAR at maintaining the fresh-seawater interface (approximated using the Ghyben-Herzberg relationship) below specified thresholds. This enabled the use of a constant-density model, rather than attempting to explicitly simulating the interaction between fresh and seawater. Although predictive uncertainty may be increased, it is outweighed by the ability to extract information from the available data. Results show that, due to the limit on water availability and the continued groundwater extraction at unsustainable rates, only limited improvements in hydraulic heads can be achieved with the proposed MAR schemes. This is an important finding for decision-makers, as it indicates that a considerable reduction in extraction in addition to MAR will be required. Our approach identified these limitations, avoiding the need for further data collection, and demonstrating the value of purposeful model design.

Keywords: uncertainty, data-assimilation, complexity, MAR (managed aquifer recharge), numerical modelling

1 INTRODUCTION

Seawater intrusion is a global issue exacerbated by increasing dependence on coastal groundwater resources, sea level rise and climate change. Most severe cases of salinization occur where groundwater levels fall below mean sea level and the groundwater flow direction turns landward (Werner, 2017). The interactions between fresh and saline groundwater involve complex density-dependent and hydrochemical processes, and are therefore inherently difficult and expensive to monitor, investigate and manage (Werner et al., 2013).

Corrective measures for aquifers already impacted by, or at risk of, seawater intrusion essentially comprise two options: reducing extraction rates, or artificially increasing recharge (Abarca et al., 2006). Both are often prevented or limited due to regulatory issues. It is often not possible to revoke existing groundwater abstraction licences, and decisions to reduce groundwater use have far-reaching economic, political, and social consequences. Managed Aquifer Recharge (MAR) includes a suite of methods to enhance aquifer recharge that are increasingly used to maintain, enhance, and secure groundwater systems under stress (Dillon et al., 2019). However, many countries lack detailed regulations making implementation challenging (Yuan et al., 2016). Although global implementation of MAR is increasing, it is not keeping pace with increasing groundwater extraction (Dillon et al., 2019).

MAR is expensive, particularly for seawater intrusion barriers, or where deep recharge boreholes are needed (Vanderzalm et al., 2022). Further pre-treatment of water prior to discharge is often necessary, particularly where urban wastewater or storm water is used (Dillon et al., 2019). Such schemes have high capital and operational costs. However, in comparison to desalination of seawater as an alternative water source, additional treatment of wastewater incurs lower energy costs, and fewer environmental impacts (Koussis et al., 2010).

Given the costs associated with MAR, and the challenges in predicting seawater intrusion, identifying the appropriate course of action is difficult. It is hard to demonstrate to stakeholders why, and when, action is necessary. However, decision-making under uncertainty is the norm for most decisions of consequence in groundwater management (Caers, 2011). Notwithstanding, decision-makers need to be informed of the risks surrounding their decisions. This requires quantifying the uncertainty of decision outcomes. Modelling supports decision making by providing the means to consolidate available data and information to both quantify, and reduce, the uncertainty surrounding outcomes of a management action (Doherty and Moore, 2021).

The subsurface is complex, and data on aquifer properties and boundary conditions are typically very limited. Expressing their uncertainty requires the use of many parameters to allow spatial variability to emerge through history-matching of the model to the historical behaviour of the system. The methods employed by industry standard tools for history-matching, PEST (Doherty, 2020) and PESTPP (White et al., 2020) software suites, need to run models many times to calculate parameter sensitivities. Therefore, incorporating existing information on system properties and past system behaviour into a model that is capable of quantifying and reducing uncertainty, requires a model that is fast and stable. To accomplish this, model development and deployment must be purposefully designed to achieve these two goals (Caers, 2011; Doherty and Moore, 2021). Models that simulate the effects of density changes between fresh and seawater require fine spatial and time discretization, have long run times, and are susceptible to numerical instability (Dausman et al., 2010). As a result, model-based data assimilation and uncertainty quantification become difficult, if not impossible (Carrera et al., 2010). Therefore, an alternative approach is needed.

Using a simpler model of a coastal aquifer introduces limitations in the types of questions that the model can answer. These limitations can be overcome by purposeful design of the modelling workflow and the prediction. We present a case study of decision-support groundwater modelling to assess MAR as a solution to seawater intrusion in the Vale do Lobo aquifer, Portugal. We demonstrate the development and use of a constant-density, highly-parameterized model, building upon the methods described in Hugman and Doherty, 2022. This approach enables the assimilation of information from both expert knowledge and field measurements to quantify and reduce predictive uncertainty. The effectiveness of MAR is assessed in terms of whether MAR can raise hydraulic heads sufficiently to levels preventative of seawater intrusion; a simple metric enabling this feasibility stage assessment of MAR.

Details of the study area and conceptual model are provided in **Section 2**, an outline of the problem, modelling rationale and design are described in **Section 3**. The numerical model configuration is described in **Section 4**, the data assimilation and uncertainty quantification process in **Section 5**, with discussion and conclusions presented in **Section 6**.

2 STUDY AREA

The study area is located to the west of Faro, capital of the Algarve province of Portugal, and comprises the western part of the Campina de Faro aquifer system, known as the Vale do Lobo (VL) sector as shown in **Figure 1**. The aquifer covers an area of 32 km^2. Groundwater from this coastal aquifer has been used extensively for irrigation over the last 50 years, for golf, tourism, and agricultural purposes. Long term annual average rainfall is approximately 600 mm/yr largely falling between November and April, whilst potential evapotranspiration is approximately 1,600 mm/yr with a substantial excess over rainfall during the summer months (DRAP-ALGARVE, 2021). Most irrigation is applied between the months of March and October and is almost entirely supplied from groundwater. Consequently, hydraulic heads are now below sea level across much of the aquifer (0 to -9 m above sea level (asl)), and several boreholes can no longer be used due to chloride concentrations of 927–2,242 mg/l measured in 2019 (Fernandes et al., 2020)). Currently the VL sector does not meet the regulatory requirement of "good" quantitative status under the EU Water Framework Directive (WFD), where

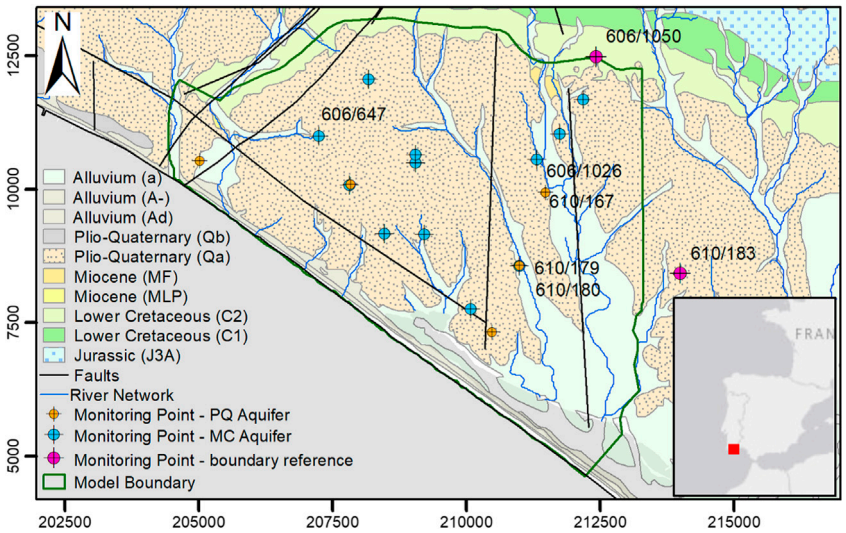

FIGURE 1 | Location and main hydrogeological features of the Vale do Lobo aquifer system, including piezometer locations.

groundwater extraction is required to be less than 90% of average annual recharge. The aquifer is at risk of further deterioration and losing "good" status based on water quality considerations considering that the Portuguese threshold value for chloride is 250 mg/l (APA, 2016). Stakeholders are interested in understanding to what extent Managed Aquifer Recharge (MAR) can be part of the solution to reverse the decline in hydraulic heads and prevent further seawater intrusion, recognising that achieving an aquifer-scale solution to SWI is a very ambitious aim, likely to require a combination of methods.

The VL sector is bounded to the east by an administrative boundary which divides the western sector of the aquifer (at risk of seawater intrusion), from the Faro sector to the east, where the problems affecting groundwater status are related with excess nitrates due to agriculture (Stigter et al., 2011). These sectors were defined to enable appropriate independent measures for each sector to be defined in the River Basin Management Plans (RBMPs) (APA, 2016) to meet WFD requirements. To the northwest, the VL boundary is defined by the Carcavai fault zone. An outcrop of Lower Cretaceous strata form the northern boundary of the VL sector, with Jurassic sediments forming a karstic aquifer further to the north.

The aquifer is formed of a thick sedimentary sequence of superimposed sedimentary basins of Mesozoic and Cenozoic age, underlain by Palaeozoic basement. The VL sector is comprised of two aquifers, an upper phreatic sand to sandy clay aquifer of Plio-Quaternary (PQ) age, and a lower semi-confined aquifer of calcareous sandstones and limestones of mainly Miocene (MC) age. A clay aquitard, with an average thickness of 10 m, separates these two aquifers. The PQ is absent at the northern boundary of the VL and increases to a maximum thickness of around 70 m in the south-east, where it is postulated that PQ sediments infilled a karstic depression in the MC surface (Carvalho et al., 2012). The PQ is highly heterogenous with 5 distinct layers mapped (Manuppella et al., 2007).

Although deep borehole records are limited, correlation with offshore and onshore borehole logs suggest that the MC aquifer reaches a depth of 350 m below mean sea level at the coast. It is underlain by the same low permeability marls of Lower Cretaceous age that form the northern boundary of the aquifer (Lopes et al., 2006). In addition to the faulted north-western boundary of the aquifer, two NNW-SSE-oriented faults transect the eastern part of the VL area (Manuppella et al., 2007). Their locations are somewhat uncertain; it is possible that their alignment is closer to that of the streams than depicted in **Figure 1**. A strike-slip fault is also located parallel to the coast approximately 1 km inland.

Most groundwater extraction is now from the MC aquifer, although the PQ aquifer was exploited historically by shallow, large diameter wells (Almeida et al., 2000). Current groundwater extraction is estimated at 6.45 Mm^3/yr (APA, 2016), based on measured extraction for the major groundwater users, and estimated extraction based on land cover and crop type for the smaller users who are not required to submit extraction returns. Detailed borehole construction records are limited, and it is often unclear if, and where, extraction is occurring from the phreatic aquifer.

The environmental regulator, the Agência Portuguesa do Ambiente (APA), estimates long term annual diffuse recharge to the VL sector is 3.46 Mm^3/yr. However, diffuse recharge is limited by the weathered red clays found at the surface, and it is recognized that a major, but unquantified, source of water to the aquifer is likely to be groundwater flowing laterally from the northern boundary from Cretaceous and Jurassic strata (Almeida et al., 2000; Hugman, 2016).

Hydraulic heads are regularly monitored by APA and are available for boreholes in the long-term monitoring network

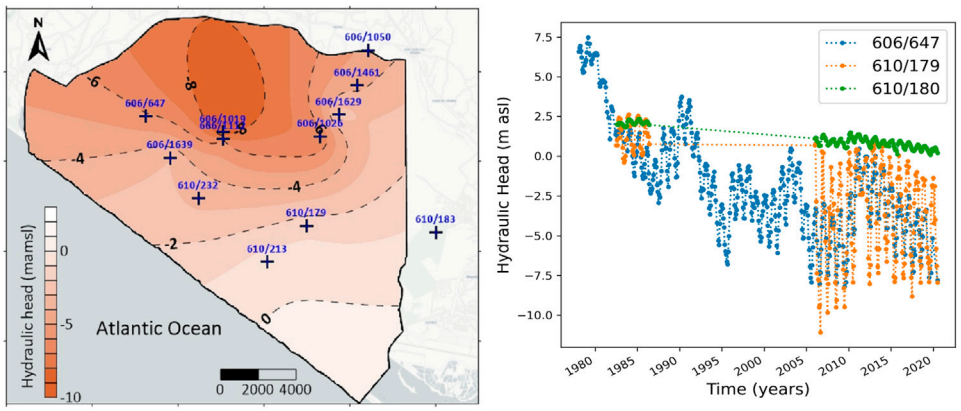

FIGURE 2 | Hydraulic head contours from semi-confined aquifer, October 2018 **(A)**, Selected hydraulic head time series at piezometer locations 606/647 (semi-confined), 610/179 (semi-confined), and 610/180 (phreatic) **(B)**.

(SNIRH, 2021). Additional heads measured monthly by the groundwater users in piezometers and extraction boreholes were also made available for use in this study. The location of hydraulic head time series used for history matching are shown in **Figure 1**.

Piezometric contours of the MC aquifer, along with selected hydraulic head time series, are shown on **Figure 2A**. The contours show that hydraulic heads are below sea level (between 0 and −9 m asl) across most of the aquifer, with the lowest values in the centre and north of the aquifer, resulting in radial flow towards this depression. Time series from three boreholes with the longest period of record are shown in **Figure 2B**, indicating that hydraulic heads were already declining during the 1980s, possibly reaching a new equilibrium since the late 1990s with higher seasonal variation, in both the PQ and MC aquifers. Hydraulic heads in the PQ are only measured in 5 locations, but these generally show slightly higher heads with reduced seasonal fluctuations compared to heads in the MC. Piezometers 610/179 (MC) and 610/180 (PQ) are adjacent to one another and represent the only location where heads are measured simultaneously in both aquifers.

Time series measurements of chloride concentrations over time are only available at four locations in the VL sector, with 2 of these exhibiting increasing trends (SNIRH, 2021). A monitoring program during 2019/2020 encountered chloride concentrations up to 2,200 mg/l in extraction boreholes, with land managers reporting that several boreholes are no longer used as their chloride concentrations are too high for irrigation (Fernandes et al., 2020).

Previous numerical modelling studies covering this area have included density-driven flow (DDF) models to investigate seawater intrusion (Hugman, 2016), assessment of nitrate contamination in the eastern sector of the Campina de Faro (Costa et al., 2021), and to assess the potential of using greenhouse runoff as water source for MAR (Costa et al., 2020). More recently, (Hugman et al., 2021) investigated sustainable extraction rates to avoid seawater intrusion.

3 THE PROBLEM

It is clear the current rates of extraction from the VL sector are unsustainable. Meeting the water balance requirement of the WFD will not prevent seawater intrusion. Without other mitigation measures, groundwater extraction would need to be reduced to 30% of current rates in VL, possibly even less (Hugman and Doherty, 2022). This would be exceedingly difficult to achieve in practice. There are few viable alternatives, and these are expensive, i.e., replacing groundwater use with desalinated seawater.

MAR has been identified as a potential mitigation measure. However, additional treatment is likely to make it an expensive option, the water available for MAR is limited, and legal issues would need to be overcome. Before committing to further investment in investigating MAR options, decision-makers need to understand whether it is likely to prevent seawater intrusion in this aquifer.

3.1 MAR Design and Water Availability

Two types of water are potentially available for MAR in this area: 1) ephemeral river flow, and 2) treated wastewater. The Ribeira da São Lourenço 1) flows from north to south close to the eastern boundary of the aquifer, with average annual flow of 1.25 Mm^3/yr between 1996 and 2008. Flow occurs on average 77 days per year. No flow is recorded in some years. Preliminary pre-settlement basin designs limit the average MAR recharge from this source to 0.5 Mm^3/yr (Standen et al., 2021).

Treated wastewater 2) is available from three treatment works in the area: Quinta do Lago, Vale do Lobo and Faro Noroeste. In 2020, available volumes were 0.76, 0.16 and 1.50 Mm^3/yr respectively (written communication, Águas do Algarve, S.A.).

The preferred MAR design would use surface infiltration basins recharging into the PQ, thereby avoiding direct injection into the MC, and allowing soil-aquifer treatment in the unsaturated zone. However, the current understanding

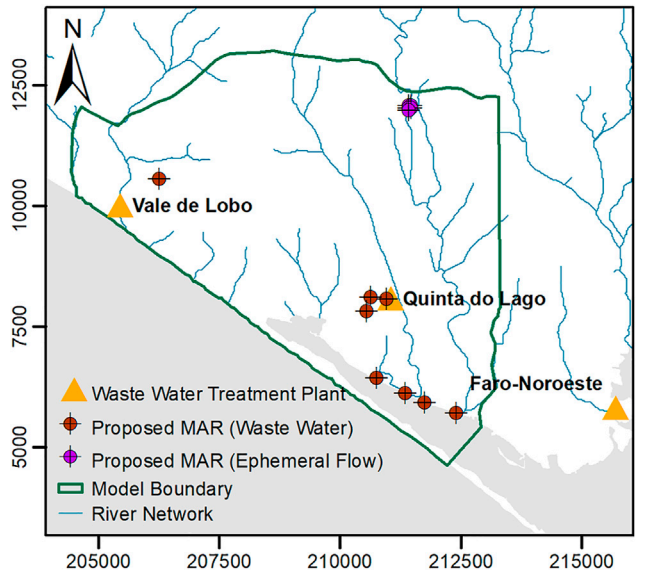

FIGURE 3 | Locations of wastewater treatment plants and proposed MAR borehole locations.

of the permeability of the PQ and the presence of the aquitard suggests this option is unlikely to be feasible. Therefore, recharge is proposed by boreholes into the MC, at locations close to the water sources, as shown on **Figure 3**.

3.2 Modelling Rationale and Prediction

3.2.1 Rationale

To model the physical coastal aquifer processes requires density-coupled flow and transport models. These require fine spatial and time discretization, with typically very long run times, and are susceptible to numerical instability. They also require the offshore part of the system to be characterized and included in the model, yet these aspects of the system are often poorly known. Sharp-interface codes offer an alternative, but simulated outcomes can be quite sensitive to initial conditions, definition of the coastal boundary condition, and they still require the offshore portion to be modelled explicitly (Bakker and Schaars, 2013; Coulon et al., 2021).

To achieve a fast and stable numerical model, process complexity is reduced by using a constant-density model. We assume that the changes in density do not play a large role in the aquifer response during the simulation period. Although this simplification will introduce some error, this will likely be small in comparison to other sources of uncertainty in the model (Caers, 2011; Doherty and Moore, 2021). The prediction cannot be based on chloride concentrations; therefore, an alternative prediction is described in **Section 3.2.2** below.

The model structure is simplified in terms of reducing the model layers and extent. The rigid structure of the offshore portion of the model is replaced by flexible parameters which represent the offshore extent. This allows us to stochastically represent the uncertainty in aquifer structure and properties offshore through physically abstract parameters. It removes the need for an offshore extent entirely, reducing the number of grid cells significantly, whilst also avoiding hard wiring assumed (but unknown) offshore structure and properties into the model. The number of layers is then limited to main hydro-stratigraphic units that are likely to control the hydraulic head response to the current pressures, and the proposed artificial recharge.

Initial assessment of the water volumes available compared to the estimated aquifer water balance indicates that these volumes may be insufficient to achieve the aquifer-scale improvements in hydraulic heads necessary to prevent SWI. However, given the uncertainties in the water balance, and the interest from regulators and stakeholders in MAR, the modelling presented herein investigates the feasibility of MAR in more detail.

3.2.2 The Prediction

Modelling undertaken herein aims to determine whether MAR can prevent seawater intrusion, an ambitious but important aim. The depth of the fresh-seawater interface as a function of hydraulic head can be obtained using the Ghyben-Herzberg relationship (Bear and Verruijt, 1987), based on the assumptions of static equilibrium, stationary seawater, and assuming that a sharp interface exists between fresh and salt water:

$$z = \alpha \times h \tag{1}$$

where z is the position of the interface below sea level [m], α [-] is defined as $\rho_f/(\rho_s - \rho_f)$, where ρ_f [M/L^3] *and* ρ_s [M/L^3] are the fresh water and sea water densities respectively, and h is the hydraulic head [m]. The minimum value of hydraulic head that ensures the fresh-seawater interface does not rise above a specified depth can be calculated using **Eq. 1**.

The effectiveness of MAR is assessed on its ability to maintain hydraulic heads at levels that ensure that the interface remains deeper than a critical value at specified locations (e.g., deeper than the base of existing extraction boreholes). This is admittedly a coarse metric. It ignores the effects of dispersion, the (potentially wide) transition zone between fresh and seawater, and up-coning in response to individual extractions. However, it is a metric that allows preliminary assessment at the aquifer scale of the feasibility of the scheme. Modelling in this context cannot ensure that MAR will be successful; however, it can determine if MAR will not be successful. As the purpose of this exercise is to assess whether it is worth exploring these schemes further, such a prediction is sufficient, and more robust, than attempting to simulate the full complexity of processes and structure.

4 NUMERICAL MODEL DEVELOPMENT

The groundwater model was constructed using MODFLOW6 (MF6) (Langevin et al., 2021), using the open source Flopy environment (v.3.3.4) (Bakker et al., 2016). The model has three stress periods: an initial steady state period to obtain representative heads and extraction rates for the start of the second stress period; a transient period from October 2000 to October 2020. A third stress period extends the model for 20 years

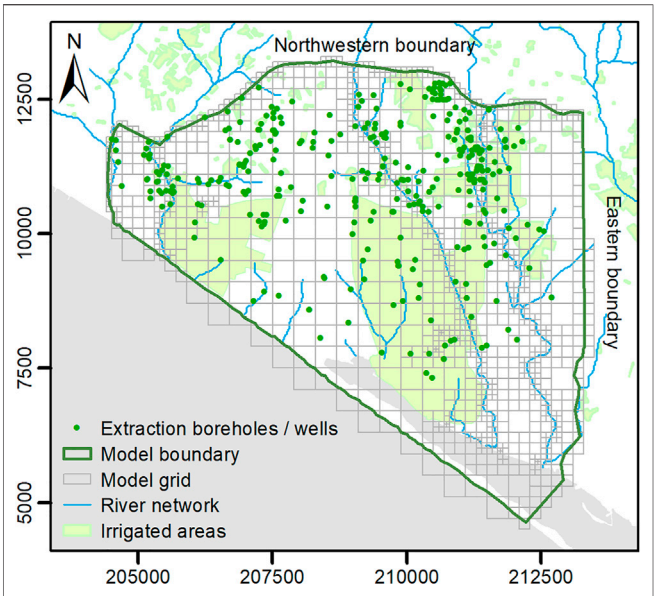

FIGURE 4 | Irrigated areas and extraction boreholes /wells, with model domain, grid and boundaries.

incorporating MAR, with the same hydrological inputs and extraction rates from the calibration period. It was not possible to start from a pre-development scenario, due to a lack of head data from this time.

The model was discretized with 400 × 400 m cell size, with quadtree mesh refinement applied using the open-source software, GRIDGEN (Lien et al., 2015). Cell sizes were reduced adjacent to potential drains and the MAR borehole locations. The model has three layers representing the phreatic aquifer, the aquitard, and the semi-confined aquifer (all layers were necessary to capture the hydraulic head response under MAR). To avoid discontinuous layers, the upper layer was assigned a minimum thickness where necessary.

The lumped parameter recharge model, LUMPREM (Doherty J., 2021), was used to estimate both recharge and groundwater withdrawal for irrigation, based on daily rainfall and potential evapotranspiration from the Faro-Patacão meteorological station (DRAP-ALGARVE, 2021). Recharge is applied to layer 1 (the phreatic aquifer), at rates depending on rainfall, irrigation, evapotranspiration, and the capacity and current volume of the soil-moisture store. Recharge occurs up to the potential evapotranspiration rate until the soil moisture store is empty, with rates decreasing as the volume of the soil moisture store decreases (the shape of this function is controlled by the gamma parameter). Transfer of rainfall-recharge to layer 3 (the semi-confined aquifer) is limited by the presence of the clay aquitard separating the two aquifers.

Using LUMPREM allowed estimation of groundwater extraction based on irrigation demand, thereby accounting for missing extraction data. It can be integrated with MF6 and PEST by the python package Lumpyrem (Hugman, 2021). The combined model (MF6 + LUMPREM) includes LUMPREM models for each of the major groundwater users, the extensive agriculture (non-metered) group, and non-irrigated land. Recharge was applied to areas defined by grid intersection with the respective land uses. Total groundwater withdrawal for irrigation was applied as time-varying total extraction rates for each group, these are then sub-divided between individual extraction wells. The locations of irrigated areas and extraction boreholes/wells are shown on **Figure 4** in relation to the model grid and boundaries.

The inland boundaries are represented by Cauchy (i.e., general head) boundary conditions applied to the semi-confined aquifer. These represent the inflows to the MC aquifer from the Jurassic aquifer to the north, and the eastern sector of the Campina de Faro aquifer to the east). The heads vary according to time series measured at 606/1050 and 610/183 for the northwestern, and eastern boundaries respectively (at locations shown on **Figure 1**). Definition of the coastal boundary condition for the semi-confined aquifer is described in **Section 4.1**, whilst for the phreatic aquifer, a head correction of 1.0124 was applied, based on the method of Lu et al. (2015). For all the boundaries, conductance is time-invariant, as are heads for the coastal boundary.

4.1 Coastal Boundary

As previously described, there is little to no data on hydraulic properties or system behaviour in the offshore portion of the aquifer system. Rather than attempt to simulate it explicitly, we represent the offshore conditions implicitly with a general head boundary, using the approach described in Hugman and Doherty (2022). This enables us to limit the model domain to the onshore portion of the system, where freshwater conditions are assumed to prevail. In turn, this allows us to ignore the effects of density differences and use a fast-running model that enables data assimilation and uncertainty analysis.

General head boundaries require specification of head and conductance parameters. Conceptually, these parameters represent the linkage between the model and the offshore portion of the system. However, they omit the effects of changes in offshore storage and assume that the dynamics of offshore flow do not change significantly during the simulated period. As such, these head and conductance parameters take on a somewhat “abstract” nature. As they are no longer physically-based, these parameters are no longer useful recipients for expert knowledge, And as they are not informed by measured data, uncertainty can be large.

The approach described in Hugman and Doherty (2022) enables the transfer of expert knowledge to these abstract parameters through the use of a simple-complex model pair. The “complex” model simulates physical process which are omitted from the “simple” model. The complex model is simulated for an ensemble representative of stresses and hydraulic properties. Values for the abstract parameters in a corresponding “simple” model are calculated for each realization. This allows the statistical distribution of abstract parameters to be characterized.

For the VL, this is achieved with use of a complementary two-dimensional DDF model (using SEAWAT) of the VL semi-confined aquifer. It was run for a long pre-

development period (during which flow is towards the sea), followed by a post-development (land-ward flow) period. A total of 100 stochastic realisations were created, sampling from the prior probability distribution of aquifer properties and inland heads based on the aquifer conceptualization (expert knowledge). By recording head and flow for both pre- and post- development conditions for each realization, values of head and conductance at the coastal boundary were obtained through the following equations:

$$q_o = (H_o - h)c \quad (2)$$

$$q_i = (H_i - h)c \quad (3)$$

For sea-ward flow and land-ward flow conditions respectively. These two equations can be solved for the two unknowns h and c. The solutions are:

$$c = \frac{q_o - q_i}{H_o - H_i} \quad (4)$$

$$h = \frac{q_o H_i - q_i H_o}{q_o - q_i} \quad (5)$$

Where H is the freshwater head at the coastline [m], q is groundwater flow under the coastline [m^3/d], and for a general head boundary along the coastline, h represents the head [m], and c the conductance [m^2/d]. The subscripts o and i represent outflow (pre-development) and inflow (post-development) conditions respectively. Values of H_o and H_i and values of q_o and q_i are obtained from the complex model for each realization. The mean and covariance of the heads and conductance can be calculated to form the combined covariance matrix:

$$C\left(\begin{bmatrix} h \\ c \end{bmatrix}\right) = \begin{bmatrix} \sigma_h^2 & \sigma_{hc} \\ \sigma_{ch} & \sigma_c^2 \end{bmatrix} \quad (6)$$

Where c is the value of $\log_{10}$ conductance [$\log_{10}$ m^2/d] at the coastal boundary, σ_h^2 is the variance of heads, σ_c^2 is the variance of $\log_{10}$ conductance, and σ_{hc} and σ_{ch} are the variance of head with $\log_{10}$ conductance, and the variance of $\log_{10}$ conductance with head respectively.

The values of h and c for a single point are used to characterize the full length of the coastal boundary by pilot points. However, values of h and c are expected to show some degree of spatial correlation along the boundary. Therefore, a joint probability distribution is required. Values were selected from a probability distribution that has the mean of $\begin{bmatrix} \underline{h} \\ \underline{c} \end{bmatrix}$ and whose covariance matrix is $C\left(\begin{bmatrix} h \\ c \end{bmatrix}\right)$ based on a maximum distance over which spatial correlation could be expected, by specifying an exponential decay of h correlation with distance, i.e., an exponential variogram, from which a covariance matrix can be obtained using the PPCOV utilities in PEST. The mean values of this prior probability distribution, together with the covariance matrix, form the basis of regularized inversion through which model calibration is achieved. The coastal boundary condition parameters characterised in this manner are thus informed by expert knowledge. This enables representation of uncertainty, whilst constraining it as much as is reasonable.

5 DATA ASSIMILATION AND UNCERTAINTY QUANTIFICATION

5.1 Methods

For the combined model a solution of minimum error variance (MEV) was sought using PEST_HP (Doherty, 2020), employing a highly-parameterized approach. A unique solution was obtained using Tikhonov (preferred value) regularization. This was followed by history-matching and uncertainty quantification (and reduction) using PESTPP-IES (White, 2018).

5.1.1 Parameterisation and Prior Information

An array of 962 pilot points distributed across the model domain, layers and boundaries allowed spatial variation of parameters. For aquifer properties these included horizontal hydraulic conductivity (K) for all layers (and thus for ratio-linked vertical hydraulic conductivity), specific yield (layer 1), and storativity (layer 3). Pilot points were placed manually, located between observation points and extraction well/borehole locations, and between these features and the model boundaries. Pilot points were also included along model boundaries and drains to allow spatial variation in boundary condition parameters.

Recharge and groundwater withdrawal for irrigation vary by land use zones linked to LUMPREM models, the parameters of which are adjustable. Prior to coupling LUMPREM and MF6, LUMPREM model parameters were first calibrated against measured extraction rates. Obtained values were subsequently used as initial parameter values when calibrating the combined model. LUMPREM provided a time series of groundwater extraction totals for each major groundwater user, and these were subdivided into groundwater extraction rates at each extraction point with a multiplier. As the extraction rates at each well were unknown, the multipliers were allowed to vary if needed during the calibration process.

The prior estimates of parameters, including the LUMPREM parameters, are shown in **Supplementary Table S1** of the Supplementary Material, with the mean of the prior probability distribution representing preferred values in the regularization. The model is parameterized with a total of 1,437 adjustable parameters. Parameter field uniqueness is achieved through numerical regularization which seeks minimum departure of each parameter from a user-specified "preferred value." For spatially varying parameters, covariance matrices are used instead of regularization weights to ensure smoothness of emergent parameter fields.

5.1.2 Observations and Weighting

In total, 5,103 observations were included as history matching targets. A total of 12 hydraulic head time series from the semi-confined, and 5 from the phreatic aquifer were used as history-

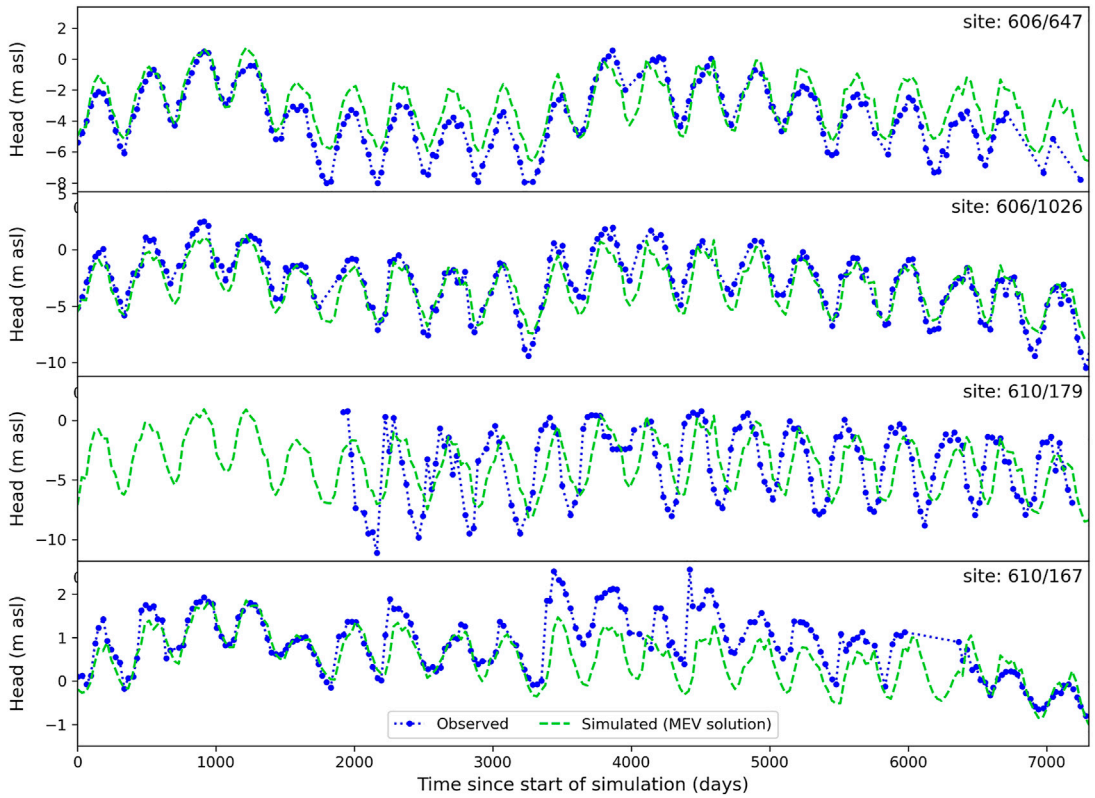

FIGURE 5 | Measured and simulated hydraulic heads for 606/647, 606/1026 and 610/179 from the semi-confined aquifer, and 610/167 from the phreatic aquifer.

matching targets. At one location, head differences between the two aquifers were also included as observations (610/179 and 610/180 in **Figure 2B**). Metered quantities of groundwater extraction reported to APA from 2010 onwards were also included as observations.

First-order temporal variations were calculated by subtracting each observation from the previous observation, giving equal importance to the temporal changes in the observation borehole time series as the actual measurement value (White et al., 2014; Foster et al., 2021; Hugman et al., 2021).

Soft data was also incorporated, with drains set at ground level across the entire model domain, and observations of zero flow included, where appropriate.

The weighting scheme aimed to give equal importance to matching of heads and extraction rates in the history-matching process. Heads were sub-divided into several observation groups to increase the weight of boreholes in different layers, and those that exhibited different responses. Groundwater extraction observations also were sub-divided to account for large difference in the temporal resolution of observations between the groups.

5.1.3 History Matching and Uncertainty Quantification

The PESTPP-IES iterative ensemble smoother generates alternative, calibration-constrained, parameter realizations, by sampling from a selected probability distribution (White, 2018). The parameter realizations are then iteratively adjusted until the model outputs attain a better fit to observations. In this case, the linear approximation to the posterior probability distribution was used as the starting point for PESTPP-IES, as often this can provide a better starting point for the process (Gallagher and Doherty, 2020).

Noise was added to the non-zero weighted observations by replacing the observation weights used during the history matching process with the inverse of the standard deviation of measurement noise. These were applied to heads (0.1 m) and pumping rates (0.5–2.5 m^3/d), with larger uncertainty applied to the non-metered groundwater users. The PEST utility RANDOBS was used to generate realisations containing noise-enhanced observations. The number of realisations (200) was selected to be more than double the number of uniquely identifiable pieces of information in the calibration dataset (90) identified by the PEST utility SUPCALC (Doherty J. E., 2021) following other recent studies (Hayley et al., 2019).

5.2 Results

5.2.1 Calibration

The resulting MEV parameter set achieved a good fit to measured observations of both hydraulic heads and groundwater extraction. In general, a better fit was obtained for heads in the semi-confined aquifer compared to the phreatic (as shown in **Figure 5**). This is not surprising, as there are fewer head observation points in the phreatic aquifer. The PQ formation

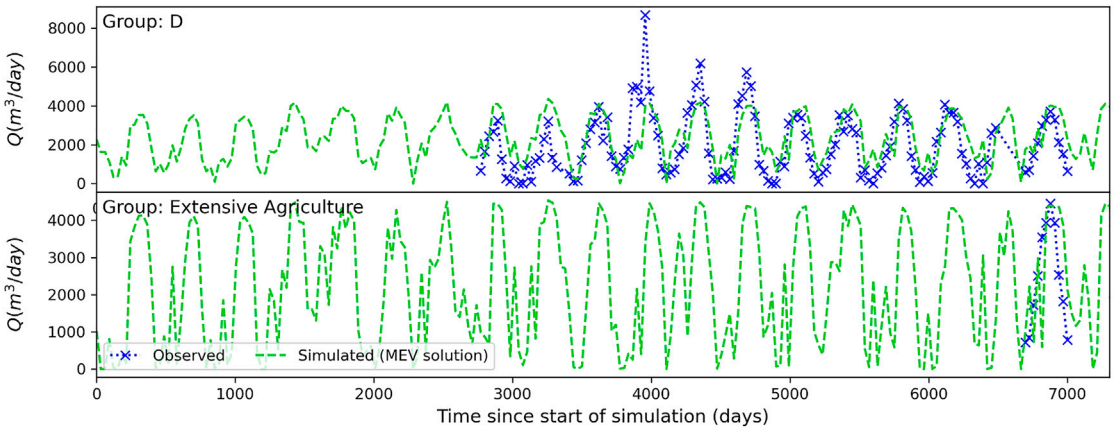

FIGURE 6 | Measured and simulated extraction rates for user group D, and the extensive agriculture group.

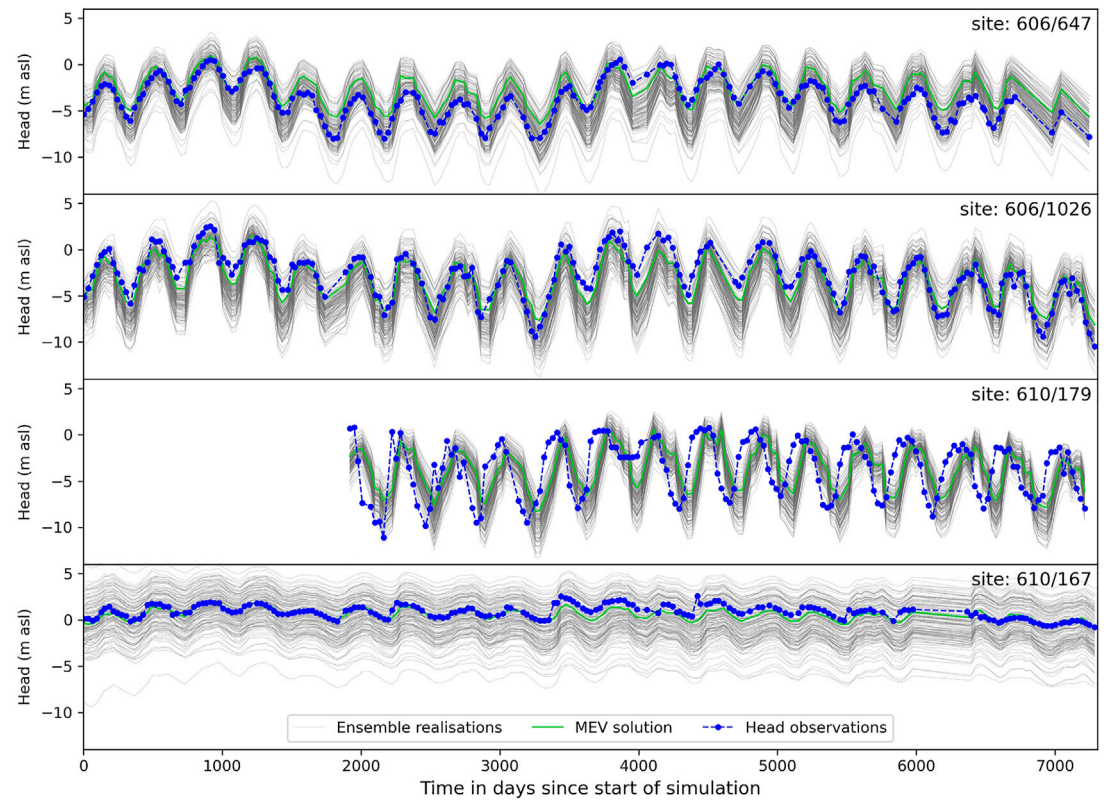

FIGURE 7 | Measured and ensemble of simulated hydraulic heads for 606/647, 606/1026 and 610/179 from the semi-confined aquifer, and 610/167 from the phreatic aquifer.

is known to be highly heterogeneous, and it is difficult to determine if, and where, extraction is occurring from this phreatic aquifer. As the fit of 610/167 only improved once extraction was permitted from both aquifers, this suggests that extraction is occurring from the PQ in this area.

Simulated and observed extraction rates are shown in **Figure 6**. In general, simulated extractions match measured extractions well, particularly in the central and eastern parts of the model.

Calibrated total annual average recharge values of 0.33–0.59 Mm^3/yr, with an average of 0.44 Mm^3/yr, were obtained. These values are an order of magnitude lower than the APA estimate (3.46 Mm^3/yr), which has been recognised as an over-estimate by several authors (Almeida

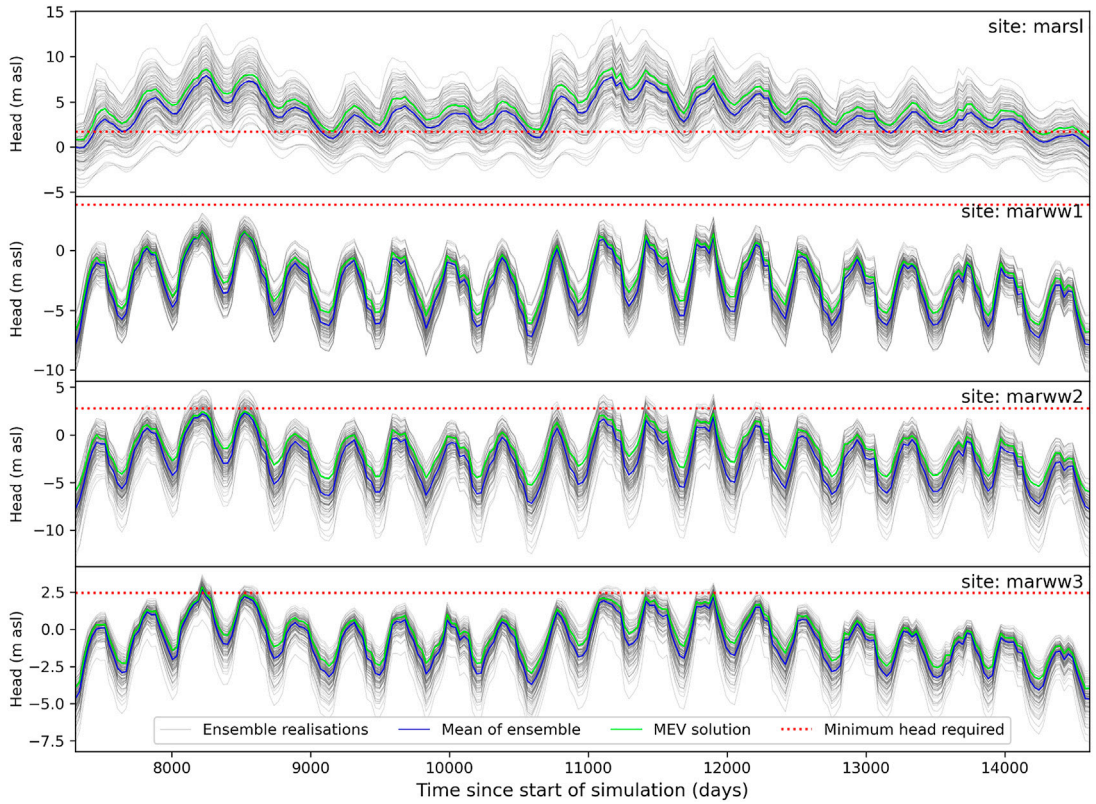

FIGURE 8 | Predicted hydraulic heads at MAR locations, showing MEV model results (green), each ensemble member (grey), mean of ensemble (blue), and the minimum head requirement at that location (red dashed).

et al., 2000; Hugman, 2016). The calibrated recharge values reflect the conceptual understanding that weathered red clays at the ground surface are of low permeability, limiting diffuse rainfall-recharge to the phreatic aquifer. The lowest recharge rates occur under the non-irrigated land (2 mm/yr), which accounts for 25 km^2 of the total 32 km^2. The other land uses have higher recharge rates (4–295 mm/yr) and include irrigation return. Diffuse recharge is largely prevented from reaching the semi-confined aquifer by the presence of the aquitard, with the majority of inflow occurring at depth from the adjacent aquifer systems.

5.2.2 History-Matching

Of the 200 realizations, 138 resulted in model convergence. The remaining model runs generally failed due to convergence issues related to drying of the upper layers. History matching results are shown in **Figure 7** for the same piezometers as **Figure 5**, along with the MEV results. The ensemble encompasses almost all the observations, apart from piezometer 610/179 where heads recover earlier in the year than the model predicts, indicating that extraction in this location perhaps ceases earlier in the year than expected by the soil-moisture balance. The ensemble resulted in a wider distribution of heads in the phreatic aquifer, as shown by 610/167, where although the temporal variation in heads matches the measured data well, there is a large range of predicted groundwater levels in this location. This occurred despite increasing the weight of the phreatic aquifer observations.

TABLE 1 | Average head differences (m) at MAR locations during 20 years simulation period (MAR scenario minus no-MAR scenario).

Location	5th Percentile	Mean	95th Percentile
Marww1	0.78	2.03	3.36
Marww2	0.28	2.01	3.80
Marww3	0.77	1.62	2.80

5.2.3 MAR Scenario Results

The impact of MAR at the locations denoted Marsl (Ribeira da São Lourenço), Marww1 (Quinta do Lago), Marww2 (Vale do Lobo) and Marww3 (Faro Noroeste) is shown in **Figure 8**, where the ensemble of predicted heads is plotted against the minimum head required at each location. Results at extraction boreholes are not shown, as the impact of MAR is negligible.

At Marsl, the heads are highly dependent on the variability of ephemeral flow, with large increases occurring during recharge periods. However, these are short-lived, falling rapidly to levels similar to the minimum head requirement when additional recharge is not occurring. This indicates that MAR is probably

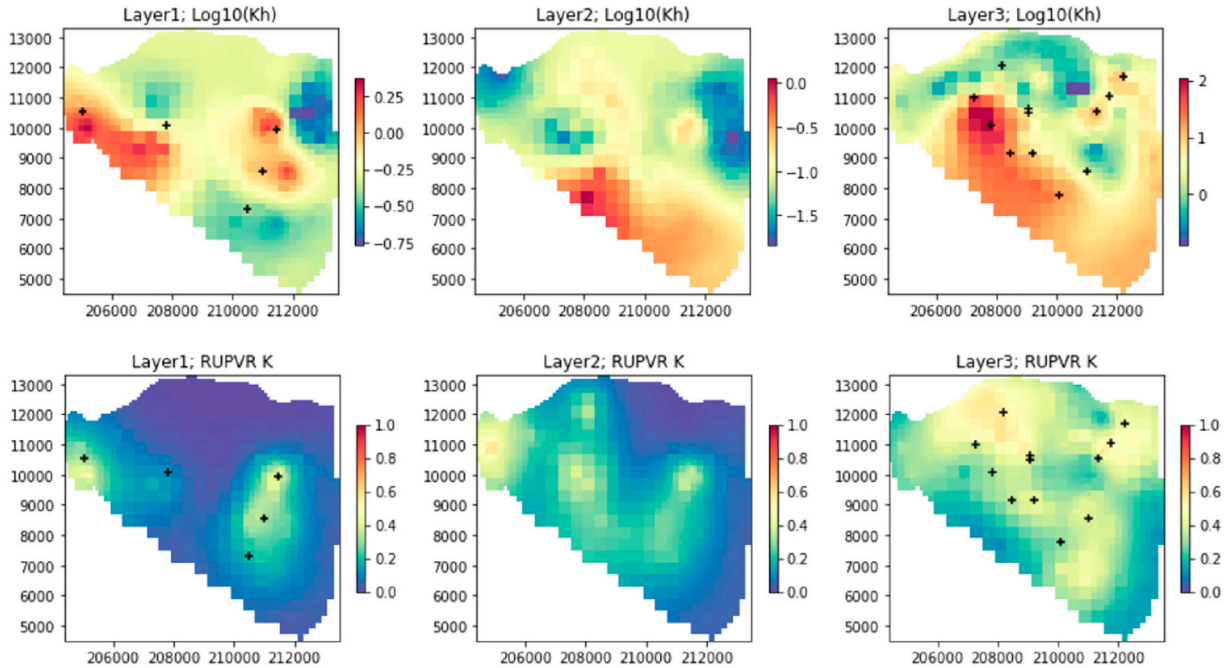

FIGURE 9 | Spatial distribution of (log) hydraulic conductivity in layers 1 (phreatic), 2 (aquitard) and 3 (semi-confined) (top, left to right), and RPUVR of hydraulic conductivity in layers 1, 2 and 3 (bottom, left to right), location of hydraulic head observations for each layer indicated by black crosses.

not necessary at this location; a location further downstream would be more beneficial.

At the other MAR locations, the minimum head requirement is only met during limited times and for some realisations. A no-MAR scenario was run to identify the head improvements resulting from MAR. The time-averaged head differences show only limited improvement in hydraulic heads as shown in **Table 1** (averages are not appropriate for Marsl due to the ephemeral flow variability and are not presented). Time series of head differences, in the **Supplementary Figure S1**, indicate that head improvements occur rapidly after the implementation of MAR.

These results raise the question whether it is possible to reach the minimum heads under any scenario, remembering that heads for the pre-development period are unknown. This was examined by undertaking an additional scenario with no extraction (and no MAR). The minimum heads required at each extraction borehole were compared to the predicted heads (5^{th} percentile of the ensemble) at those locations, confirming that the minimum heads could be met with a small number of exceptions (see **Supplementary Table S2**). These occurred where extraction boreholes are deep (up to 200 m), or where the boreholes were located close to the eastern boundary. Here, the average heads are already low (1.1 m asl), preventing the minimum head requirement from being met close to the boundary. This provides confidence that somewhere between no-extraction and current extraction plus MAR, a management solution to protect the aquifer exists.

5.2.4 Insights From Linear Analysis

The spatial distributions of hydraulic conductivity for each layer from the MEV parameter set are plotted in **Figure 9**, along with the corresponding values of the relative parameter uncertainty variance reduction (RUPVR) (Doherty J. E., 2021). This ratio varies from 0 to 1, with higher values indicating the locations where posterior parameter uncertainty has been reduced in comparison to the prior during history-matching. Of particular interest is the area in the centre of the model, which appears to have relatively higher K in both layer 1 and layer 2, where the RPUVR shows that the uncertainty has been reduced to a greater extent than the surrounding area. This is an important insight, which could justify further site investigation for a potential infiltration basin MAR scheme in this location.

Values of RUPVR were low for pilot points along the boundary conditions, with mean values of 3×10^{-2} to 6×10^{-6} obtained for conductance and head values, indicating that history matching was not effective in reducing uncertainty in the boundary condition parameters, outlining the importance of constraining the prior probability distributions by the method described in **Section 4.1**.

6 DISCUSSION AND CONCLUSIONS

6.1 The Workflow

The combined model was scripted, and therefore reproducible. Furthermore, conceptual changes identified during the model construction and calibration process could easily be altered in the

base model. Whilst the model was designed to have a fast run time, the scripting involved significant time investment and cognitive effort for the modeller in moving away from GUI based methods to the open-source tools described herein, but it was considered to be worth it for the resulting flexibility and reproducibility.

The model development and deployment were considered simultaneously, with reduced process complexity (constant-density) and structural complexity (modelling the offshore extent and processes as described in **Section 4.1**). The resulting model was capable of uncertainty quantification and reduction, but with limitations in terms of the predictions it can make. For an initial first-order assessment, evaluating the effectiveness of MAR against minimum heads was an acceptable compromise. This relatively simple metric quickly identified that MAR was not likely to be successful and thus no further effort was put into a more comprehensive analysis. If this had not been the case, further efforts into designing adequate metrics would have been warranted. An alternative (or complementary) analysis could use the results from the complementary DDF model to determine the relation between fresh-saltwater interface response to changes in flux across the GHB coastal boundary. If a defensible relation between change in flux and gradient reversal could be established, this would allow the magnitude of change in GHB flux to be used as a metric for effectiveness.

Calibrating with PEST_HP was time-consuming. Balancing the weights requiring subjective expert knowledge about the important features of the system. To obtain an acceptable fit across all observation groups required testing of multiple weighting strategies. However, calibration allowed the use of linear analysis. This identified (with the RPUVR statistic) that the uncertainty of coastal boundary parameters was not reduced by history-matching. This provided further justification for the method used to stochastically characterise the coastal boundary, which constrained the prior probability distribution. It also enabled the linearized posterior probability distribution to be used as the starting point for PESTPP-IES, reducing the number of model convergence failures during this process (one-third of realizations failed to converge even with this workflow).

Where decisions need to be made relatively quickly to protect the aquifer, the use of a simpler model is beneficial. If building a complex model takes too long, decisions are likely to be taken before such a model is available (Caers, 2011). Furthermore, if a complex model cannot quantify and reduce uncertainty, a likely outcome given the nature of DDF models, then the decision-support such a model can provide is limited.

6.2 The VL Sector

This case study demonstrates the development of a decision-support groundwater model to assess the effectiveness of MAR to prevent seawater intrusion in a coastal aquifer system, whilst allowing reduction of prediction uncertainty through data assimilation in a highly-parameterized framework. Process complexity was reduced using a constant-density model, along with a complementary 2D DDF model, to allow stochastic characterisation of the head and conductance along the boundary. This allowed us to achieve the fast run times necessary to undertake history-matching and reduce predictive uncertainty.

Evaluating MAR by the ability to achieve minimum heads that prevent the seawater interface encroaching above the base of the current extraction boreholes is pragmatic. It permits a preliminary, aquifer-wide assessment, and allows regulators and stakeholders to understand the benefits and limitations of MAR with a simple metric. The results demonstrate that MAR cannot increase the hydraulic heads sufficiently to attain the minimum heads required, even locally. Therefore, the proposed MAR schemes cannot prevent the interface from reaching the base of the existing extraction boreholes, and seawater intrusion in the VL cannot be mitigated by MAR alone.

The minimum heads can be met for the majority of locations in a "no-extraction" scenario, the exception being deep boreholes close to the eastern boundary. Here heads are not sufficiently high enough to prevent seawater intrusion, indicating that the VL sector cannot be entirely protected from seawater intrusion even under this scenario without concurrent management action in the eastern part of the Campina de Faro.

This modelling, in conjunction with that of Hugman and Doherty (2022), identifies for the first time, the true scale of the problem in this area, and how difficult it will be to resolve. A significant reduction in extraction will be needed in addition to, or as an alternative to MAR. Hugman and Doherty (2022) have shown that extraction rates would need to be reduced at least to 30% of current rates in VL, possibly even less. Required reduction in extraction would be less in conjunction with MAR. An integrated approach to water management in the VL sector could use the available treated waste-water directly for irrigation as an alternative to MAR. Although this has not been explicitly modelled, the implication of our model results is that the waste-water volumes remain insufficient, and further reductions in extraction would still be required.

Predicted climate change impacts on rainfall indicate that for the RCP4.5 scenario, rainfall is expected to decrease by 10% in the south of Portugal, with an associated reduction in wet days of 10–20%, which will lead to associated reductions in recharge (Soares et al., 2017). River flows in the Mediterranean region are likely to be even more intermittent in the future due to climate change, with an increasing number of zero flow events (Schneider et al., 2013), reducing the availability of water for MAR from this source. Meanwhile, socio-economic and agricultural development in the region will result in increased water demand for irrigation (Stigter et al., 1998; Hugman et al., 2017). These compounding factors will result in higher demand at a time when less water is available. Without action, the aquifer will face even more severe pressures in the future.

Collecting further information on the aquifer properties and state of seawater intrusion, such as geophysics and further water quality studies, adds to the available body of knowledge, but it is time-consuming and expensive. Meanwhile, decisions are not taken. The existing data is already rich in prediction-specific information, as measured water levels are available close to where

water level predictions are required. We have demonstrated an approach and associated model to support decision-making with the data currently available. This modelling has limitations, but we are still able to state with a relative degree of confidence that investing in MAR on its own is not going to solve the problem. In conjunction with Hugman & Doherty (2022), we have demonstrated that substantial further actions are needed to protect groundwater quality in the VL sector.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The raw data is available from the websites referenced in the text (SNIRH and DRAP-ALGARVE). The remainder of the data was provided directly by the Agência Portuguesa do Ambiente, elements of which are confidential.

AUTHOR CONTRIBUTIONS

KS and RH conceived the idea, KS analysed the data. KS (with support and guidance from RH) set up the modelling workflows and undertook all simulations and their postprocessing. RH completed the linear algebra to characterise the coastal boundary. KS and RH discussed the results and contributed to the final manuscript. JPM provided supervision, reviewed the manuscript and obtained the MARSoluT research funding.

FUNDING

KS received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 814066 (Managed Aquifer Recharge Solutions Training Network—MARSoluT). RH received funding through the "Groundwater Modelling Decision Support Initiative" (GMDSI), conducted under the auspice of the National Centre for Groundwater Research and Training (NCGRT), Flinders University, South Australia, with funding from BHP and Rio Tinto.

ACKNOWLEDGMENTS

We thank Edite Reis of the Agência Portuguesa do Ambiente for providing additional data used in this study, and John Doherty for this advice during the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.904271/full#supplementary-material

REFERENCES

Abarca, E., Vázquez-Suñé, E., Carrera, J., Capino, B., Gámez, D., and Batlle, F. (2006). Optimal Design of Measures to Correct Seawater Intrusion. *Water Resour. Res.* 42, 1–14. doi:10.1029/2005WR004524

Almeida, C., Mendonça, J. L., Jesus, M. R., and Gomes, A. J. (2000). *Sistemas Aquíferos de Portugal Continental*. Lisboa: INAG, Instituto da Água.

APA (2016). Plano de Gestão de Região Hidrográfica Das Ribeiras Do Algarve (Rh8) Parte 2 - Caracterização e Diagnóstico. Available at: https://apambiente.pt/sites/default/files/_SNIAMB_Agua/DRH/PlaneamentoOrdenamento/PGRH/2016-2021/PTRH8/PGRH_2_RH8_Parte2_Anexos.pdf.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., and Starn, J. J. (2016). Scripting MODFLOW Model Development Using Python and FloPy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Bakker, M., and Schaars, F. (2013). Modeling Steady Sea Water Intrusion with Single-Density Groundwater Codes. *GroundWater* 51, 135–144. doi:10.1111/j.1745-6584.2012.00955.x

Bear, J., and Verruijt, A. (1987). *Modeling Groundwater Flow and Pollution - Theory and Application of Transport in Porous Media*. Dordrecht, Holland: D. Reidel Publishing Company.

Caers, Jef. (2011). *Modeling Uncertainty in the Earth Sciences*. Chichester, UK: John Wiley & Sons.

Carrera, J., Hidalgo, J. J., Slooten, L. J., and Vázquez-Suñé, E. (2010). Computational and Conceptual Issues in the Calibration of Seawater Intrusion Models. *Hydrogeology J.* 18, 131–145. doi:10.1007/s10040-009-0524-1

Carvalho, J., Ramalho, E., Dias, R., Pinto, C., and Ressurreição, R. (2012). A Geophysical Study of the Carcavai Fault Zone, Portugal. *Pure Appl. Geophys.* 169, 183–200. doi:10.1007/s00024-011-0318-y

Costa, L. R. D., Hugman, R. T., Stigter, T. Y., and Monteiro, J. P. (2021). Predicting the Impact of Management and Climate Scenarios on Groundwater Nitrate Concentration Trends in Southern Portugal. *Hydrogeology J.*, 2501–2516. doi:10.1007/s10040-021-02374-4

Costa, L. R. D., Monteiro, J. P. P. G., and Hugman, R. T. (2020). Assessing the Use of Harvested Greenhouse Runoff for Managed Aquifer Recharge to Improve Groundwater Status in South Portugal. *Environ. Earth Sci.* 79, 1–15. doi:10.1007/s12665-020-09003-5

Coulon, C., Pryet, A., Lemieux, J. M., Yrro, B. J. F., Bouchedda, A., and Gloaguen, E. (2021). A Framework for Parameter Estimation Using Sharp-Interface Seawater Intrusion Models. *J. Hydrology* 600, 126509. doi:10.1016/j.jhydrol.2021.126509

Dausman, A. M., Langevin, C., Bakker, M., and Schaars, F. (2010). *A Comparison between SWI and SEAWAT – the Importance of Dispersion , Inversion and Vertical Anisotropy*. São Miguel, Azores, Portugal: 21st Salt Water Intrusion Meeting, 271–274.

Dillon, P., Stuyfzand, P., Grischek, T., Lluria, M., Pyne, R. D. G., and Jain, R. C. (2019). Sixty Years of Global Progress in Managed Aquifer Recharge. *Hydrogeology J.* 27, 1–30. doi:10.1007/s10040-018-1841-z

Doherty, J. E. (2021b). PEST Model-independent Parameter Estimation User Manual Part II: PEST Utility Support Software, Watermark Numerical Computing. Brisbane.

Doherty, J. E. (2020). *PEST_HP - PEST for Highly Parallelized Computing Environments*. Brisbane: Watermark Numerical Computing.

Doherty, J., and Moore, C. (2021). in *Decision Support Modelling Viewed through the Lens of Model Complexity* (Adelaide: National Centre for Groundwater Research and Training, Flinders University). doi:10.25957/p25g-0f58*A GMDSI Monograph. South Australia*

Doherty, J. (2021a). *Version 2 of the LUMPREM Groundwater Recharge Model*. Brisbane: Watermark Numerical Computing.

DRAP-ALGARVE (2021). Weather Station Data for Patação-Faro. Available at: https://www.drapalgarve.gov.pt/ema/pat.htm (Accessed December 22, 2021).

Fernandes, J., Midões, C., Ferreira, A., Castanheira, A., Monteiro, F., and Pereira, A. (2020). *GeoEra TACTIC project, Pilot description and assessment: Campina de Faro Aquifer System*. Portugal.

Foster, L. K., White, J. T., Leaf, A. T., Houston, N. A., and Teague, A. (2021). Risk-Based Decision-Support Groundwater Modeling for the Lower San Antonio River Basin. *Groundwater* 59, 581–596. doi:10.1111/gwat.13107

Gallagher, M., and Doherty, J. (2020). *Water Supply Security for the Township of Biggenden: A GMDSI Worked Example Report*. South Australia: Flinders University. doi:10.25957/5f4c5815fc0d1

Hayley, K., Valenza, A., Whilte, E., Hutchinson, B., and Schumacher, J. (2019). Application of the Iterative Ensemble Smoother Modeling Case Study. *Water* 11, 1649. doi:10.3390/w11081649

Hugman, R., Doherty, J., and Standen, K. (2021). *Model-Based Assessment of Coastal Aquifer Management Options. A GMDSI Worked Example Report*. South Australia: Flinders University. doi:10.25957/a476-x588

Hugman, R., and Doherty, J. (2022). Complex or Simple-Does a Model Have to be One or the Other?. *Frontiers in Earth Sci.* 10, 1–12. doi:10.3389/feart.2022.867379

Hugman, R. (2021). Lumpyrem. Available at: https://github.com/rhugman/lumpyrem (Accessed March 15, 2021).

Hugman, R. (2016). *Numerical Approaches to Simulate Groundwater Flow and Transport in Coastal Aquifers – from Regional Scale Management to Submarine Groundwater Discharge*. Portugal: Universidade do Algarve.

Hugman, R., Stigter, T., Costa, L., and Monteiro, J. P. (2017). Numerical Modelling Assessment of Climate-Change Impacts and Mitigation Measures on the Querença-Silves Coastal Aquifer (Algarve, Portugal). *Hydrogeology J.* 25, 2105–2121. doi:10.1007/s10040-017-1594-0

Koussis, A. D., Georgopoulou, E., Kotronarou, A., Lalas, D. P., Restrepo, P., and Destouni, G. (2010). Cost-efficient Management of Coastal Aquifers via Recharge with Treated Wastewater and Desalination of Brackish Groundwater: General Framework. *Hydrological Sci. J.* 55, 1217–1233. doi:10.1080/02626667.2010.512467

Langevin, C. D., Hughes, J. D., Banta, E. R., Provost, A. M., and Panday, S. (2021). *MODFLOW 6 Modular Hydrologic Model*. doi:10.5066/F76Q1VQV

Lien, J.-M., Liu, G., and Langevin, C. D. (2015). *GRIDGEN Version 1 . 0 : A Computer Program for Generating Unstructured Finite-Volume Grids*. doi:10.3133/ofr20141109

Lu, C., Xin, P., Li, L., and Luo, J. (2015). Seawater Intrusion in Response to Sea-Level Rise in a Coastal Aquifer with a General-Head Inland Boundary. *J. Hydrology* 522, 135–140. doi:10.1016/j.jhydrol.2014.12.053

Manuppella, G., Ramalho, M., Antunes, M. T., and Pais, J. (2007). *Carta Geologia de Portugal na escala de 1:50,000: Notícia Explicativa da folha 53-A Faro*. Lisboa.

Schneider, C., Laizé, C. L. R., Acreman, M. C., and Flörke, M. (2013). How Will Climate Change Modify River Flow Regimes in Europe? *Hydrology Earth Syst. Sci.* 17, 325–339. doi:10.5194/hess-17-325-2013

SNIRH (2021). Sistema Nacional de Informação de Recursos Hidricos, Agência Portuguesa Do Ambiente. Available at: https://snirh.apambiente.pt/index.php?idMain=1&idItem=1.4&uh=M&sa=M12 - CAMPINA DE FARO (Accessed December 20, 2020).

Soares, P. M. M. M., Cardoso, R. M., Lima, D. C. A., Miranda, P. M. A. A., and Daniela, C. A. L. (2017). Future Precipitation in Portugal: High-Resolution Projections Using WRF Model and EURO-CORDEX Multi-Model Ensembles. *Clim. Dyn.* 49, 2503–2530. doi:10.1007/s00382-016-3455-2

Standen, K., Hugman, R., and Monteiro, J. P. (2021). "Managed Aquifer Recharge as a Solution for an Over-exploited Aquifer in South Portugal: Development of a Decision-Support Groundwater Model (Abstract)," in 47th International Association of Hydrogeologists Conference, Brussels, September 6th – 10th, 2021.

Stigter, T. Y., Carvalho Dill, A. M. M., and Ribeiro, L. (2011). Major Issues Regarding the Efficiency of Monitoring Programs for Nitrate Contaminated Groundwater. *Environ. Sci. Technol.* 45, 8674–8682. doi:10.1021/es201798g

Stigter, T. Y., van Ooijen, S. P. J., Post, V. E. A., Appelo, C. A. J., and Carvalho Dill, A. M. M. (1998). A Hydrogeological and Hydrochemical Explanation of the Groundwater Composition under Irrigated Land in a Mediterranean Environment, Algarve, Portugal. *J. Hydrology* 208, 262–279. doi:10.1016/S0022-1694(98)00168-1

Vanderzalm, J., Page, D., Dillon, P., Gonzalez, D., and Petheram, C. (2022). Assessing the Costs of Managed Aquifer Recharge Options to Support Agricultural Development. *Agric. Water Manag.* 263, 107437. doi:10.1016/j.agwat.2021.107437

Werner, A. D., Bakker, M., Post, V. E. A., Vandenbohede, A., Lu, C., and Ataie-Ashtiani, B. (2013). Seawater Intrusion Processes, Investigation and Management: Recent Advances and Future Challenges. *Adv. Water Resour.* 51, 3–26. doi:10.1016/j.advwatres.2012.03.004

Werner, A. D. (2017). On the Classification of Seawater Intrusion. *J. Hydrology* 551, 619–631. doi:10.1016/j.jhydrol.2016.12.012

White, J. T. (2018). A Model-independent Iterative Ensemble Smoother for Efficient History-Matching and Uncertainty Quantification in Very High Dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the Predictive Consequences of Model Error with Linear Subspace Analysis. *Water Resour. Res.* 50, 1152–1173. doi:10.1002/2013WR014767

White, J. T., Hunt, R. J., Fienen, M. N., Doherty, J. E., and Survey, U. S. G. (2020). Approaches to Highly Parameterized Inversion: PEST++ Version 5, a Software Suite for Parameter Estimation, Uncertainty Analysis, Management Optimization and Sensitivity Analysis. *Tech. Methods* 64. doi:10.3133/tm7c26

Yuan, J., van Dyke, M. I., and Huck, P. M. (2016). Water Reuse through Managed Aquifer Recharge (MAR): Assessment of Regulations/guidelines and Case Studies. *Water Qual. Res. J. Can.* 51, 357–376. doi:10.2166/wqrjc.2016.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Earth Science

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/feart.2022.916198

VisU-HydRA: A Computational Toolbox for Groundwater Contaminant Transport to Support Risk-Based Decision Making

Maria Morvillo, Jinwoo Im and Felipe P. J. de Barros*

Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA, United States

Obtaining accurate and deterministic predictions of the risks associated with the presence of contaminants in aquifers is an illusive goal given the presence of heterogeneity in hydrological properties and limited site characterization data. For such reasons, a probabilistic framework is needed to quantify the risks in groundwater systems. In this work, we present a computational toolbox `VisU-HydRA` that aims to statistically characterize and visualize metrics that are relevant in risk analysis with the ultimate goal of supporting decision making. The `VisU-HydRA` computational toolbox is an open-source Python package that can be linked to a series of existing codes such as MODFLOW and PAR^2, a GPU-accelerated transport simulator. To illustrate the capabilities of the computational toolbox, we simulate flow and transport in a heterogeneous aquifer within a Monte Carlo framework. The computational toolbox allows to compute the probability of a contaminant's concentration exceeding a safe threshold value as well as the uncertainty associated with the loss of resilience of the aquifer. To ensure consistency and a reproducible workflow, a step-by-step tutorial is provided and available on a GitHub repository.

Keywords: uncertainty quantification (UQ), stochastic hydrogeology, reproducible, decision making, probabilistic risk analysis

OPEN ACCESS

Edited by:
Anneli Guthke,
University of Stuttgart, Germany

Reviewed by:
Md Nazmul Azim Beg,
Tulane University, United States
Mariaines Di Dato,
Helmholtz-Zentrum für Umweltforschung UFZ, Germany

****Correspondence:***
Maria Morvillo
morvillo@usc.edu

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *08 April 2022*
Accepted: *23 May 2022*
Published: *14 June 2022*

Citation:
Morvillo M, Im J and de Barros FPJ (2022) VisU-HydRA: A Computational Toolbox for Groundwater Contaminant Transport to Support Risk-Based Decision Making. Front. Earth Sci. 10:916198. doi: 10.3389/feart.2022.916198

1 INTRODUCTION

Assessing the risks associated with the presence of pollutants in groundwater often times relies on the use of mathematical models. The key challenge is that model predictions in the subsurface environment are subject to significant amount of uncertainty. These uncertainties arise due to insufficient site characterization and our inability to fully resolve the spatial fluctuations of hydrological properties at multiple scales. The combined effect of these factors leads to uncertainty in model input parameters which render model outputs to be uncertain (Rubin, 2003). Quantifying this uncertainty and understanding how it propagates to quantities of interest, such as environmental performance metrics (de Barros et al., 2012), are critical in risk analysis as well as for decision makers to better allocate resources toward uncertainty reduction (de Barros and Rubin, 2008) and define optimal aquifer remediation strategies (Cardiff et al., 2010).

A number of computational approaches have been proposed to quantify the effects of aquifer heterogeneity on the spatiotemporal dynamics of a solute plume (see the following review articles, Dentz et al., 2011; Neuman and Tartakovsky, 2009; Fiori et al., 2015) and the associated uncertainty

(e.g., Kapoor and Kitanidis, 1998; Fiori and Dagan, 2000; Dentz and Tartakovsky, 2010; Meyer et al., 2013; de Barros and Fiori, 2014; Boso and Tartakovsky, 2016; Ciriello and de Barros, 2020). The computational stochastic frameworks provided in most of the above-mentioned works are of analytical or semi-analytical nature (i.e. based on perturbation theory). Many of these analytical approaches have been applied to human health probabilistic risk analysis (Andričević and Cvetković, 1996; de Barros and Rubin, 2008) and successfully compared with concentration field data (i.e., de Barros and Fiori, 2021). Fully numerical approaches allow to relax on simplifying assumptions, and uncertainty is typically quantified through the Monte Carlo framework (e.g., Maxwell et al., 2008; Siirila et al., 2012; Henri et al., 2016; Im et al., 2020). An in-depth analysis of the reliability assessment of the computed statistical moments obtained from Monte Carlo simulations within the context of subsurface hydrology is provided in Ballio and Guadagnini (2004). In addition to Monte Carlo methods, other approaches for uncertainty quantification are reported in the literature (see Oladyshkin and Nowak, 2012; Ciriello et al., 2017, and references therein). A review comparing the advantages of different numerical stochastic methodologies used for uncertainty quantification can be found in Zhang et al. (2010).

Despite significant contributions in computational methods in the field of stochastic hydrogeology (see Rubin, 2003), many of the developed computational tools are not easily accessible to the hydrological community. With the exception of few tools (e.g. Li and Liu, 2006; Maxwell et al., 2015; Hammond et al., 2014, amongst others), most existing computational tools are difficult to access and are not open source. There is an ever-increasing need within the hydrological community for models' transparency and reproducibility. Being able to reproduce numerical results is, nowadays, an essential feature that needs to be present in tools used for environmental modeling, risk analysis and data management (e.g., Fienen and Bakker, 2016; Fienen et al., 2022). The usage of collaborative coding environments (such as GitHub, Dabbish et al., 2012), where scripts are written in open-source languages (e.g., Python, vanRossum, 1995), has the potential of creating clear, shareable and reproducible knowledge. As stated by White et al. (2020), the absence of those characteristics can reduce the credibility of the model as a decision support tool and hamper resource management efforts.

In this work, we present a computational toolbox that links the various components relevant for the estimation of the pollutant concentration at an environmentally sensitive target and its associated uncertainty. The computational framework builds upon existing computational tools such as HYDRO_GEN (Bellin and Rubin, 1996), FloPy (Bakker et al., 2016), and a GPU-based random walk particle tracking code (Rizzo et al., 2019). The key features of this computational toolbox are as follows:

- The proposed computational toolbox is fully *open source*, including coding language and utilized software, to enhance accessibility of the modeling workflow.
- All the utilized software are run through Python scripts, to provide a complete, transparent and repeatable record of the modeling process following the ideas put forth in Bakker et al. (2016).
- All the steps necessary to construct the model, compute risks and uncertainty, are smoothly connected in a unified script, to ensure efficiency in computing the Monte Carlo iterations, for uncertainty quantification.
- All the software have been selected for their precision, robustness, compatibility among each other, user-friendliness and transparency.
- All files are shared on a GitHub repository, to be constantly accessible, editable and expandable as a communal effort in creating a consistent and efficient modeling framework.

As mentioned in Bakker et al. (2016), models are commonly constructed with a graphical user interface (GUI), due to their interactive environment and guided structure in populating the model and post processing the results. However, when GUIs are utilized for constructing and post processing numerical groundwater flow and transport models, no records of the modeling process are available, limiting repeatability and accessibility of the employed modeling framework. On the other hand, Python scripts can be seen as a complete, clear, easy and readable structure of the modeling approach. It serves as a documentation of the model input data and provides the hydrological community a cooperative script-based workflow that is easy to access (Peñuela et al., 2021). The visual and interactive nature of our workflow and software package enhance the accessibility and understanding of model predictions, overcoming the communication limits of static documentation, with the final goal of assisting decision makers (Woodruff et al., 2013). For such reasons, we provide a step-by-step tutorial on how to utilize the proposed computational toolbox and illustrate how this toolbox can be employed to 1) perform risk assessment and 2) improve our fundamental understanding of the role of aquifer heterogeneity in the physics of contaminant transport.

2 PROBLEM STATEMENT

We start by considering a scenario where an hazardous substance is released into an aquifer with ambient base flow rate Q_b. The contaminant plume originating from the source zone will undergo a series of physical and (bio)chemical processes until it reaches a receptor, e.g. a compliance plane or pumping wells. Due to limited site characterization of the subsurface environment, the spatiotemporal dynamics of the solute plume is subject to uncertainty. In this work, the main source of uncertainty stems from the randomness of the hydraulic conductivity field, denoted by K. Under these conditions, decision makers are interested in determining the *probability* that the contaminant concentration C at an environmentally sensitive location will exceed a threshold value C^* established by a regulatory agency, namely $\text{Prob}[C > C^*]$. The concentration estimate C at a given location $\mathbf{x} = [x_1, \ldots, x_d]$, with d denoting the

dimensionality of the flow domain, and time t is typically obtained by solving the partial differential equations governing the physics of transport.

The concept of reliability, traditionally defined as the probability of non-failure, can be employed to formulate this problem. Let $\xi(t)$ denote the aquifer reliability [-] evaluated at an environmentally sensitive target exposed to contamination for a given realization of the hydraulic conductivity field. In this work, we define that aquifer reliability function at the environmentally sensitive target as follows:

$$\xi(t;\mathcal{V}_T) = \begin{cases} 1, & C(\mathbf{x} \in \mathcal{V}_T, t) \leq C^*, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where t is time and $\mathcal{V}_T$ is the geometric configuration (i.e. volume, area, line or point) that characterizes the dimensions of the environmentally sensitive location. In order to measure the capability of the aquifer to recover its reliability as a potable water source, we will rely on the concept of loss of resilience. The aquifer resilience loss, denoted here by R_L, over a given time period $t_0 \leq t \leq t_0 + t_f$ is computed at the target location $\mathcal{V}_T$ as follows:

$$R_L(t;\mathcal{V}_T) = \int_{t_0}^{t_0+t_f} \Psi(t;\mathcal{V}_T)\mathrm{d}t, \tag{2}$$

where

$$\Psi(t;\mathcal{V}_T) = 1 - \xi(t;\mathcal{V}_T). \tag{3}$$

Due to the inherent uncertainty in the parameters characterizing the aquifer system, such as the K-field, the functions C, ξ and R_L are regarded as random functions. As a consequence, both Ψ and R_L are characterized in terms of their statistical moments and their probability density functions (or cumulative density functions). For this work, we will compute uncertainty in Ψ and R_L through Monte Carlo framework. With the goal of ensuring transparency and reproducibility of the results presented in this work, we provide all the codes necessary to compute **Eqs 2, 3** for a fully worked illustration. We include a detailed description of the files and scripts for the illustrations, following the mindset presented in some works within the hydrological community (e.g., Fienen and Bakker, 2016; White et al., 2020).

3 METHODS AND IMPLEMENTATION

Our approach for modeling the resilience loss of an aquifer consists of following four components: the generation of the hydraulic conductivity fields, a groundwater flow and contaminant transport model accompanied by an initially known contaminant injection zone, the computation of **Eq. 2** and a Monte Carlo framework to evaluate the uncertainty associated with the resilience loss. For our computational illustrations, we will assume that groundwater flow and contaminant transport take place in a hypothetical two-dimensional (2D) aquifer. The computational domain has dimensions ℓ_i along the ith direction where $i = 1, 2$. Below we provide details regarding each step.

1. *Geology and Geostatistics*: First, a description of the site geology is needed. This description is based on a grid of hydraulic conductivity values that will serve as input for the groundwater flow model. The log-conductivity field is assumed to be isotropic and spatially heterogeneous $Y(\mathbf{x}) = \log K(\mathbf{x})$. As previously mentioned, Y is uncertain given the incomplete information of the hydrogeological system. Therefore, Y is regarded as a random space function (RSF) (Kitanidis, 1997; Rubin, 2003) and considered here as statistically stationary and multi-variate Gaussian. The RSF model for Y is therefore characterized by the mean (μ_Y), variance (σ_Y^2) and spatial covariance $C_Y(\mathbf{r})$ of Y with $\mathbf{r}$ denoting the lag-distance. We adopt an exponential model for C_Y with isotropic correlation length λ. The ensemble of random Y fields used in the simulations is generated using the robust HYDRO_GEN tool (Bellin and Rubin, 1996).
2. *Groundwater Flow Field*: Groundwater flow is assumed to be at steady-state and far from the presence of sinks and sources. The governing equation for the flow field is provided in **Appendix A**, see **Eq. A1**. Permeameter-like boundary conditions are assumed, i.e. prescribed hydraulic heads at the inlet (h_{in}) and outlet (h_{out}) along the longitudinal direction of the computational domain and no-flux conditions at the remaining boundaries. The groundwater balance equations are solved numerically and the solution of hydraulic head in the computational domain is obtained. Groundwater fluxes are computed using MODFLOW (Harbaugh, 2005) together with the Python library FloPy (Bakker et al., 2016). The dimensions of the numerical grid block are Δx_i, with $i = 1, 2$. The velocity field can be calculated through Darcy's law by combining the computed hydraulic head with the hydraulic conductivity and with the knowledge of the porosity.
3. *Solute Transport*: A contaminant is instantaneously released along a source zone with area $\mathcal{A}_o = \Delta s_1 \times \Delta s_2$. The initial concentration of the contaminant is given by C_0. We assume that transport is non-reactive and governed by the advection-dispersion equation (see **Appendix A**, **Eq. A2**). Transport is solved through the use of a Lagrangian-based simulator, i.e., Random Walk Particle Tracking (RWPT). The transport simulator is a parallelized GPU-based RWPT dubbed PAR^2 (Rizzo et al., 2019). The transport simulations will allow to compute the concentration breakthrough curve at a given target location (i.e. a protection zone or an observation well). The concentration values at the target location are then compared to a given regulatory threshold value, C^*, for the pollutant of interest. Based on the solute breakthrough curves and C^*, we can then calculate the resilience loss, **Eq. 2**.
4. *Uncertainty Quantification*: To estimate the uncertainty associated with the quantities of interest, we employ a

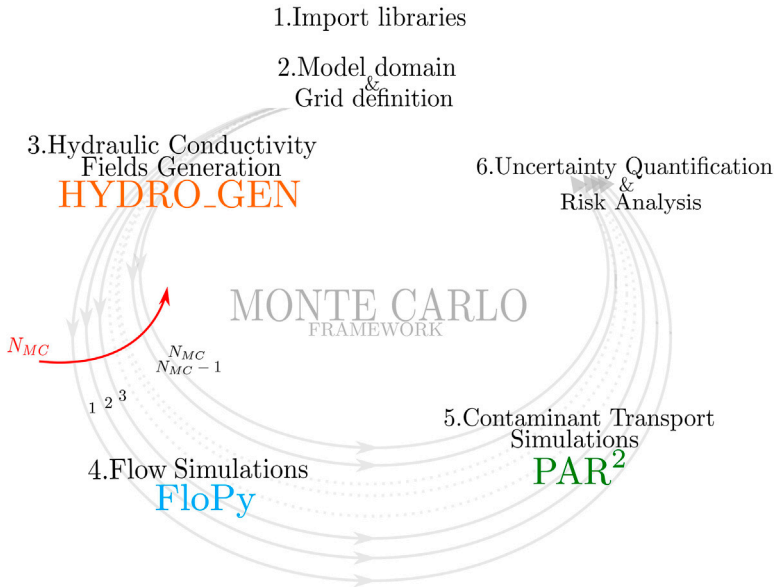

FIGURE 1 | A schematic representation of VisU-HydRA. As a computational toolbox for groundwater contaminant transport and its uncertainty, VisU-HydRA integrates all necessary numerical software and post-processing codes in a single Jupyter Notebook. Each component (1-6) is executed by each code cell in the Jupyter Notebook in order. Colored are the names of the employed software, and N_{MC} stands for the number of Monte Carlo realizations.

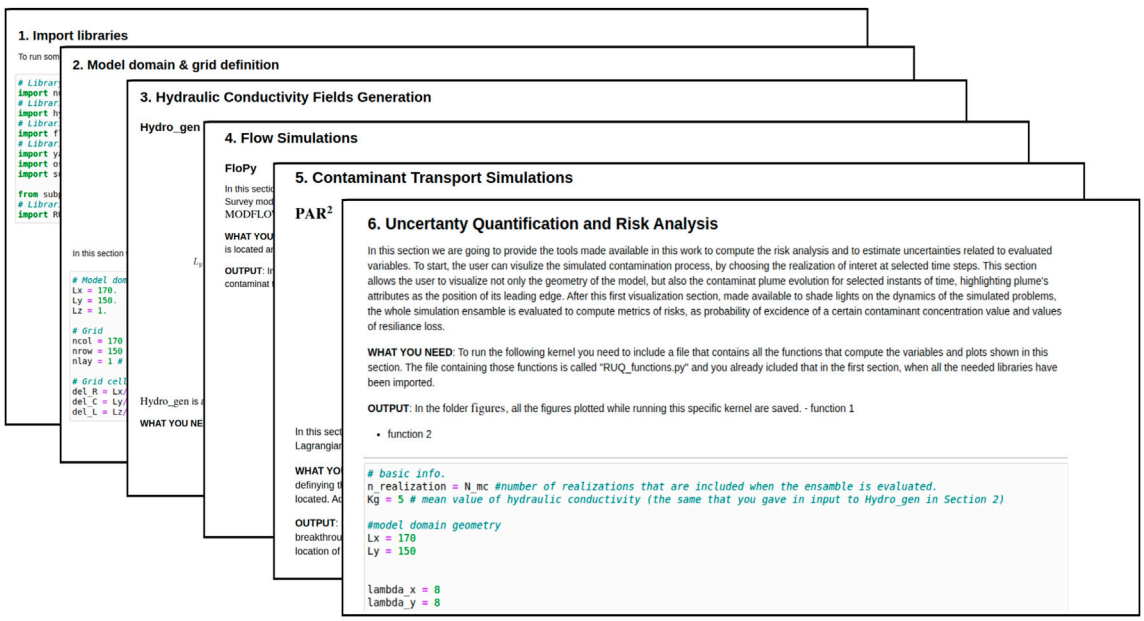

FIGURE 2 | Examples of Jupyter Notebook code cells corresponding to the numbered processes (see **Figure 1**). Detailed descriptions are included in each cell so that users can easily reproduce the results of the tutorial presented in this study.

Monte Carlo (MC) framework. In this approach, a series of hydraulic conductivity field realizations are generated, based upon some geostatistical representation of the subsurface environment (see step 1). Then, the groundwater flow and contaminant transport equations are solved for each realization of the K field. This results in a statistical description of the concentration breakthrough curves at a given location and consequently, the resilience loss, **Eq. 2**. In our fully worked out example, we evaluate an ensemble consisting of five hundred realizations of the conductivity field (N_{MC} = 500).

TABLE 1 | Input parameters used in the proposed tutorial for hydraulic conductivity field generation and flow and transport simulations.

Random space function model for $Y = \log K$		
Symbol	**Value**	**Units**
$\ell_1 \times \ell_2$	170 × 150	[m]
μ_Y	1.6	[m/day]
$K_G = \exp[\mu_Y]$	5	[m/day]
σ_Y^2	3	[-]
λ_1, λ_2	8, 8	[m]
Flow Simulations		
h_{in}, h_{out}	1, 0	[m]
$\Delta x_1 \times \Delta x_2$	1 × 1	[m]
t_{TOT}	1,000	[days]
Δt	4	[days]
Transport Simulations		
α_1, α_2	0.01, 0.001	[m]
D_m	8.6×10^{-5}	[m^2/day]
s_1^0, s_2^0	25, 65	[m]
Δs_1, Δs_2	12, 20	[m]
ν_1^0, ν_2^0	117, 55	[m]
$\Delta \nu_1$, $\Delta \nu_2$	12, 40	[m]
C_0	1	[mg/L]
C^*	0.001	[mg/L]
N_p	10^5	[-]

4 TUTORIAL

Here, we describe the structure and the contents of the computational code adopted in our study and a tutorial. For our case study, we provide an open-source code to ensure the reproducibility and re-usability of its model outputs (Peñuela et al., 2021). We will make use of the appealing features of Jupyter Notebooks (i.e., web-based applications) to write a simple and transparent code. The proposed package is aimed at Visualizing Uncertainty for Hydrological Risk Analysis and it is called `VisU-HydRA`.

The `VisU-HydRA` package is available on GitHub repository (https://github.com/mariamorvillo/VisU-HydRA). A Jupyter Notebook, named "Tutorial_MC_F&T.ipynb", contains all the scripts necessary to produce the results and analysis associated with the case study presented in this work (see further details in **Section 3**). All other functions and files are also made available to support the user in better understanding the features of the tutorial and to eventually apply all (or some) of the available tools to scenarios of their interest. As shown in **Figure 1**, the tutorial consist of six components. These components are subdivided into code cells (see **Figure 2**) in the Jupyter Notebook as follow:

1. *Import Libraries*: The first tutorial code cell imports libraries which are needed in order to run some of the tools offered by Python. A Python library is simply a collection of codes, or modules of codes, that can be used in a program for specific operations. Among the most commonly used libraries, `NumPy` supports large matrices, multi-dimensional data and consists of in-built mathematical *functions* to facilitate the computations (Oliphant, 2006). Further libraries are needed in order to generate spatially random K fields, to assist in the execution of flow and contaminant transport simulations and in post-processing the output data to compute relevant quantities such as the reliability and resilience loss of the aquifer.
2. *Model domain & Grid Definition*: In the second code cell, the dimensions of the computational domain and related contaminant source and target zone areas as well as the characteristics of the numerical mesh (i.e. the discretization scheme) are declared. The hydraulic properties of the aquifer, such as the hydraulic conductivity, are specified for each cell which will serve as input for the flow simulator (such as MODFLOW). In any finite-difference based flow simulation, such as the one employed in this analysis through MODFLOW, the hydraulic heads are calculated at discrete points in space; those points are termed the nodes of the Finite Difference or Finite Volume grid with dimensions $n_{col} \times n_{row} \times n_{lay}$ (e.g., Zheng and Bennett, 2002).
3. *Hydraulic Conductivity Fields Generation*: This code cell contains the first step of the MC loop (see **Figure 1**). As stated in **Section 3**, this analysis considers 500 realizations of the hydraulic conductivity field, namely $N_{MC} = 500$. To start, spatially heterogeneous log-conductivity fields $Y \equiv \ln[K]$ are generated, with the characteristics indicated in **Table 1**. To produce the ensemble of spatially correlated Y field realizations, we use HYDRO_GEN (Bellin and Rubin, 1996). The HYDRO_GEN executable is available for Linux and Mac platforms, meaning that if a Windows platform is used, the user needs to generate a file containing the K fields realizations ("Kfileds_Hydrogen.npy") on a different platform. The output of the code cell is a ".npy" file containing the K values related to all the generated fields. Each of these K fields are distributed along the file columns. The user has to declare only the number of MC realizations and the used operating system in the code cell. All the information, such as the geostatistical parameters used to generate the random K fields, have to be indicated on the "hydrogen_info.txt" file. This ".txt" file can be easily compiled following the instructions included in the "manual_hydrogen.pdf" file made available by the software creators (Bellin and Rubin, 1996) and included in the GitHub repository, as all the documents discussed in this work (see **Table 2**).
4. *Flow Simulations*: The fourth code cell computes the second step of the MC framework, by processing the flow simulations using the randomly generated heterogeneous K fields. The flow field is computed using numerical simulator MODFLOW (Harbaugh, 2005). To create, run, and post-process MODFLOW-based models, the Python package FloPy (Bakker et al., 2016) is employed. The FloPy library is imported in the first step on this tutorial. The MODFLOW executable, "mf2005dbl.exe," is needed in order to run this code cell (in the GitHub repository, the one for Windows platform), while all the variables related to the flow simulation (see Harbaugh, 2005, for details regarding MODFLOW and its packages),

TABLE 2 | A list of software, Python libraries, configuration files for each step and the corresponding output data. Users need to install and prepare them to run the tutorial that reproduces the data and the analysis shown in this work.

	What you need	What to install	Output
K Fields Generation	• hydrogen_linux or hydrogen_mac • hydrogen_input.txt • hydrogen.py	• numpy library	• Kfield_Hydrogen.npy
Flow Simulations	• mf2005dbl.exe	• flopy library	• N_{MC} model-*.ftl files in the folder tmp
Contaminant Transport Simulations	• NVIDIA GPU • config.yaml • config-tmp.yaml • par2.exe	• yaml library • os library • subprocess library	• N_{MC} result-*.csv files in the output folder • snap-{}-*.cvs files for each chosen simulation time step ({})
Uncertainty Quantification & Risk Analysis	• RAUQ_function.py	• matplotlib library	• data_output folder

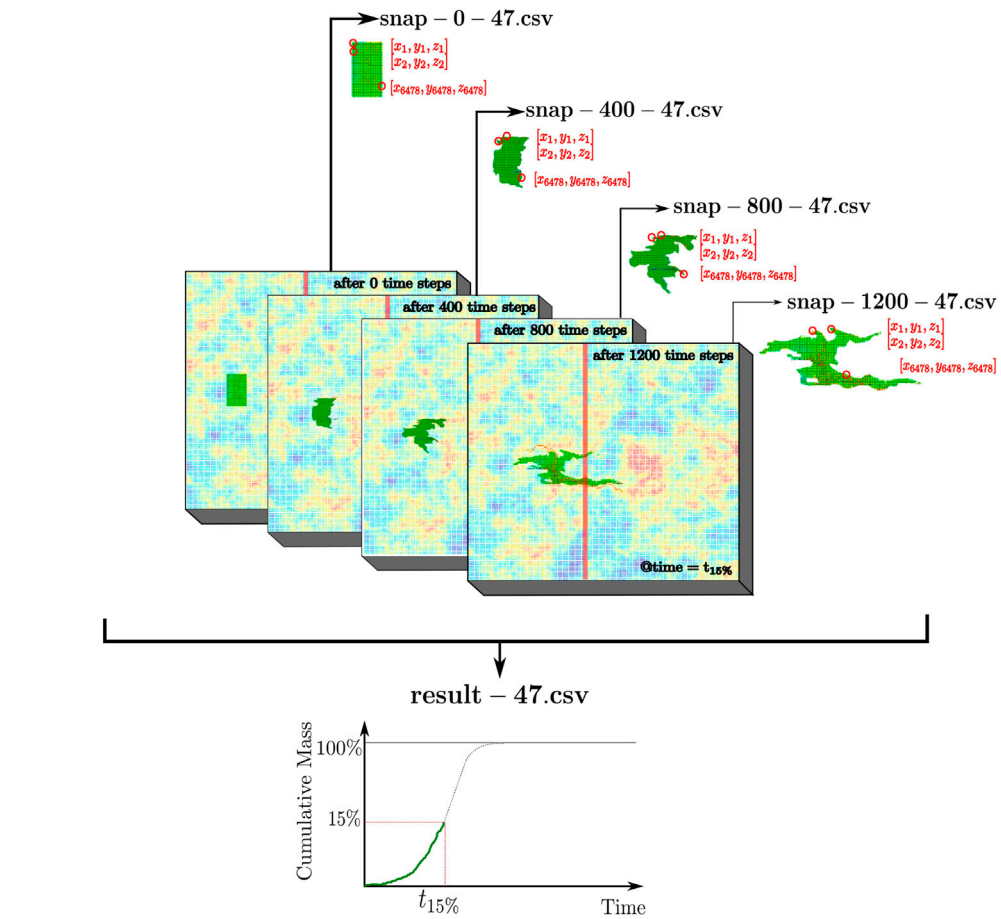

FIGURE 3 | Schematic representation of the contaminant transport simulations code cell output. The image schematizes a possible output (depending on the declared parameters in the YAML file) of a randomly selected hydraulic conductivity field from the Monte Carlo ensemble (realization number 47). The model output for this specific hydraulic conductivity realization is given by "result-47.csv". This file contains the data related to the contaminant's cumulative breakthrough curve at the control plane (represented by the vertical red line). In addition, snapshot files are generated for 400 simulation time steps. For example, "snap-0-47.csv", "snap-400-47.csv", "snap-800-47.csv" and "snap-1200-47.csv" contain the position of each particle at time steps 0, 400, 800 and 1200. For visualization purposes, the position ($x = [x, y, z]$) of the 1^{st}, 2^{nd} and 6478^{th} particles are highlighted by red circles.

have to be specified in this cell's code. Documentation on how to install the FloPy package is available on https://github.com/modflowpy/flopy, in addition to the full description of all the functions available in the just mentioned package, needed to run different MODFLOW's features. The cell gives as output, in the

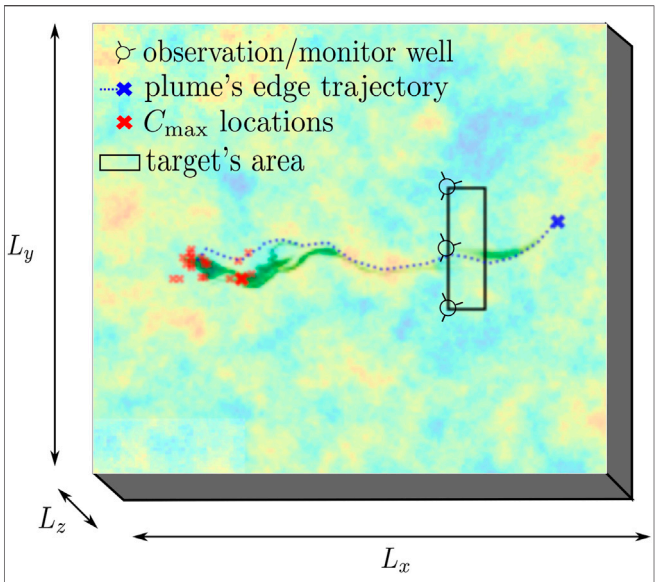

FIGURE 4 | Schematic representation of the contaminant plume, in green, for a given random realization of the hydraulic conductivity field. The figure shows where observations wells (circles) are positioned over the target area (rectangle) and how the locations of the leading edge of the plume (blue cross) and of the maximum contaminant concentration (red cross), experienced at each time steps of the simulation, evolve in time. The trajectory of the leading edge of the plume follows the trajectory given by the dotted blue line. The location of the maximum concentration of the plume bounces from one location to another (see red crosses).

folder named "tmp", N_{MC} ".ftl" files ("model-*.ftl"), each of them containing respectively the flow simulation output related to the "*" MC iteration.

5. *Contaminant Transport Simulations*: PAR^2 (Rizzo et al., 2019) is used to run the contaminant transport simulations. It requires, as input, the groundwater flow velocities originating from the previous code cell output. PAR^2 is a GPU-accelerated solute transport simulator, as explained in **Section 3**, and consequently needs to be run on a platform equipped with an NVIDIA GPU. Simulation parameters can be easily defined through a YAML configuration file. Thus, to run this code cell, the user needs the PAR^2 executable, "par2.exe"(in the GitHub repository, the one for Windows platform) and two YAML files. Note that "config-tmp.yaml" is a temporary file that is modified at each MC iteration, following the structure of "config.yaml," by substituting the "*" symbol with the number of the current MC iteration and the "{}" symbol with the evaluated simulation time step. The user needs to provide the input simulation parameters in both the YAML files. Further information on how to compile these files can be found at "PAR2Info." The output of this code cell, in the folder "output", are N_{MC}.csv files ("result-*.csv"), each of them containing respectively the data related to the contaminant's cumulative breakthrough curves at the selected control planes for the "*" MC iteration. Snap files (i.e., "snap-{}-*.csv"), for each MC iteration, can be an output of the simulation as well. Those files contain respectively the positions of the particles in which the contaminant has been discretized into, at the "{}" time step selected by the user through the YAML file. A schematic representation of the potential output of a contaminant transport simulation with PAR^2 is shown in **Figure 3**. As mentioned above, snapshot files are created for each simulation of the MC ensemble. The time required for a simulation is, among other variables, proportional to the number of generated snapshot files. As a consequence, the user should choose this variable wisely, to avoid too prolonged simulations.
6. *Uncertainties Quantification & Risk Analysis*: This component of the tutorial consists of post-processing the previous data and generating the graphical output that can support the decision-making process. In order to run this last part of the Jupyter Notebook, the ".py" file containing the scripts of the implemented functions is necessary to elaborate the data coming from the previous sections. The file is called "RAUQ_function.py", written following a basic and intuitive structure to allow the user to access and easily modify it accordingly. This section of the package allows the user to visualize the geometry of the model, the location of the source of contaminant and target zone ($\mathcal{A}_T = \Delta\nu_1 \times \Delta\nu_2$) and the positions of observation wells. The user can also visualize the spatiotemporal evolution of the solute plume (including the positions of the leading edge of the plume and the maximum concentration) and the hydraulic conductivity field (see **Figure 4**). For this plume visualization step, the user can choose a specific MC realization and the specific snapshot in time ([days]). The user can also visualize the ensemble statistics of the concentration field in both space and time as well as other risk-related metrics (e.g., probability of contaminant concentration exceedance and the statistics of the maximum concentration, resilience loss, reliability, etc.). Further detail on the generated files and other information can be found in the text included in the Jupyter Notebook.

5 APPLICATION TO RISK AND RESILIENCE

5.1 Probability of Concentration Exceedance and Resilience Loss Maps

We will now use `VisU-HydRA` to investigate the risks associated with an accidental benzene spill. For this hypothetical case study, we will consider a 2D simulation as previously described (therefore, $x_1 = x$ and $x_2 = y$). All parameter values employed in the upcoming results are reported in **Table 1**. Benzene is a wildly used chemical for industrial solvents and for constituents of fossil fuels and is considered to be a major threat to groundwater resources and human health (Sivasankar et al., 2017). Benzene spills are typically associated with transportation and storage tank leakages. Due to its potential health risk (e.g., Logue and Fox, 1986), the State of California (United States) set the Maximum Contaminant Level, i.e. the highest level of a contaminant that is allowed in drinking water, to $C^* = 10^{-3}$ [mg/L] (Proctor et al., 2020). Note that benzene's

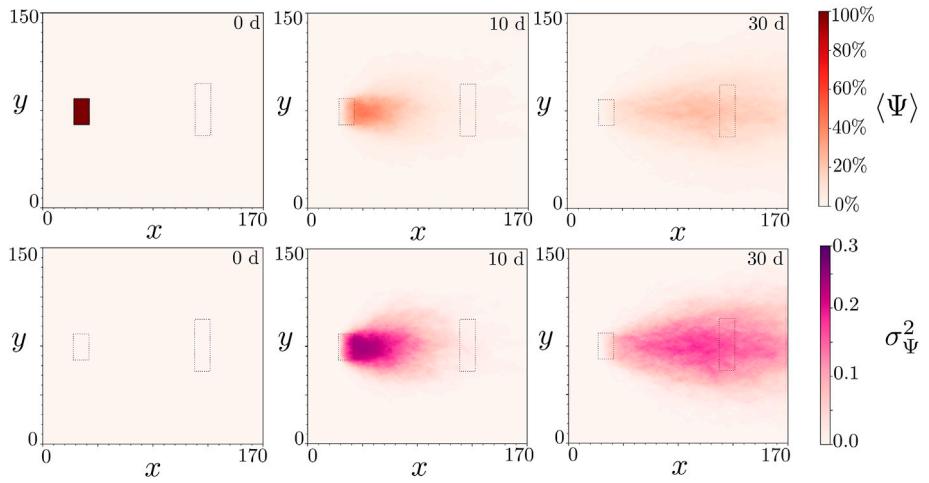

FIGURE 5 | Maps of the probability of concentration exceedance (top), and its uncertainty (bottom) at different instant of time in the simulation (columns). Plots of the top row shows how ⟨Ψ⟩ [-] evolves in space respectively after 0, 10 and 30 days from the beginning of the contamination process. ⟨Ψ⟩ is expressed as a probability, as indicated by its values on the color bar on the right. The bottom row shows σ^2_Ψ, as a measure of the uncertainty related to the information given by the plots in the row above.

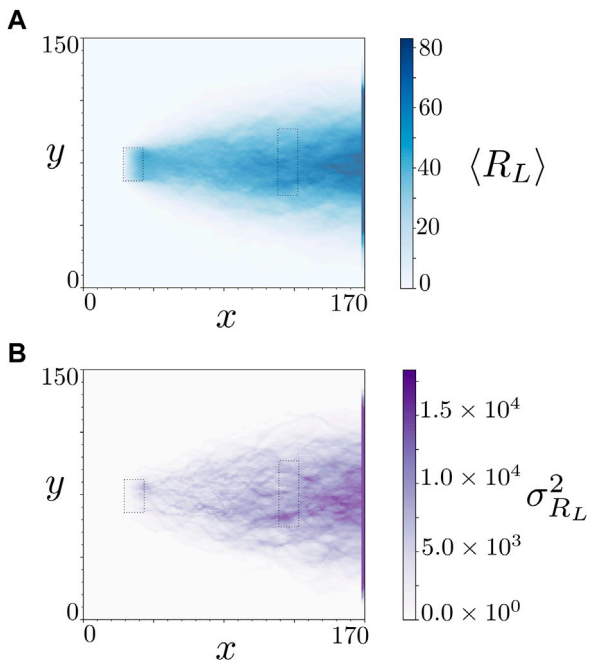

FIGURE 6 | Maps of the average loss of resilience **(A)** in the model domain, and its variance values **(B)** throughout the whole simulation time. The plot on the top shows how ⟨R_L⟩ [d] evolves in space, expressing the amount of days necessary for each point in the domain to recover up to a reliable status. The amount of days associated with the different shades of blue are indicated by the values on the color bar on the right of the top plot. The plot at the bottom shows the values of the variance of R_L, $\sigma^2_{R_L}$, as a measure of the uncertainty related to the information given by the plot above.

degradation in water is extremely slow (i.e., Sivasankar et al., 2017), and consequently, for the purpose of our illustration, we will assume transport to be non-reactive (see **Eq. A2**).

Figure 5 depicts the spatial map of the expected value of Ψ (**Figure 5**, top row), namely ⟨Ψ⟩, as well as its variance σ^2_Ψ (**Figure 5**, bottom row) at different time snapshots. The statistics of Ψ are computed over all N_{MC} realizations of the hydraulic conductivity field. Given that ξ (**Eq. 1**) is a Bernoulli distribution, Ψ (**Eq. 3**) also follows a Bernoulli distribution (see Chapter 2.2. of Mood et al., 1974). Therefore, the expected value of Ψ is equal to the probability of concentration exceedance at the position **x** and instant of time *t*. In other words, $\langle \Psi(\mathbf{x}, t) \rangle \equiv \Pr[C(\mathbf{x}, t) \leq C^*]$. The results shown in **Figure 5** (top row) show that the probability of being at risk decreases with time as an outcome of the enhanced plume dilution due to macroscale spreading (e.g., Dentz et al., 2011; de Barros et al., 2015; Ye et al., 2015; Henri et al., 2016). The maps depicted in **Figure 5** can be used by decision makers for evaluating the risks associated with contamination, identifying sampling locations and allocating resources towards uncertainty reduction.

Next, we analyze the loss of resilience of the aquifer system. **Figure 6** shows the spatial map of first two statistical moments of R_L, see **Eq. 2**. The results illustrated provide information regarding the locations where the expected resilience loss will be the largest (**Figure 6A**) and its corresponding uncertainty (**Figure 6B**). R_L represents a measure of the amount of days necessary for the aquifer to recover up to a state where the risks associated with the contamination can be considered negligible.

As expected, the values of $\langle R_L \rangle$ increase with travel distance from the source location. This is explained by the increase of the macroscale dispersion as the plume moves downstream from the source. An increase in plume dispersion leads to the presence of long tails in the solute breakthrough curve. Therefore, the residence time for the plume while crossing a given location increases thus leading to an increase in the averaged resilience loss. As observed in **Figure 6A**, the maximum number of days necessary for the right boundary of the flow domain to recover is

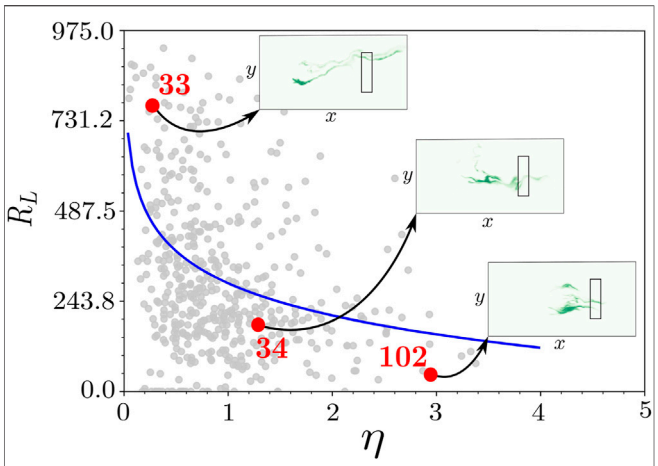

FIGURE 7 | Maximum value of resilience loss (R_L [d]) experienced within the target area versus contaminant source efficiency (η [-]) for all the Monte Carlo realizations. Data (gray dots) are fitted by a logarithmic curve, the blue line in the plot. The red dots highlight three Monte Carlo realizations, the 33^{rd}, 34^{th} and 102^{nd}, in which the spatial configuration of the contaminant plume is investigated at a given time snapshot. The insets are connected, by a black arrow, to the simulated scenario. The black rectangles in the insets represent the environmentally sensitive target location.

approximately 80. The results at the bottom of **Figure 6** quantify the uncertainties related to the estimated resilience loss values. As shown in **Figure 6B**, larger uncertainty in R_L is observed at large distances from the source. One possible explanation for this phenomenon may be due to the fact the contaminant plume has sampled more fluctuations of the velocity field (i.e. the plume travels through many correlation scales) thus leading to increased uncertainty in the solute breakthrough curves at those locations.

5.2 Impact of Contaminant Source Efficiency on Aquifer Reliability

The computational package `VisU-HydRA` can also be employed to improve our understanding on the impact of groundwater flow physics and decision making metrics such as R_L. Here we illustrate how the hydraulics conditions within the contaminant source zone could be used as an indicator of R_L. de Barros and Nowak (2010) introduced the source zone efficiency η

$$\eta = \frac{Q_{sz}}{Q_b}, \tag{4}$$

where Q_{sz} is the volumetric flow rate [m^3/days] crossing the contaminant source zone while Q_b indicates the total background volumetric flow rate passing through the entire aquifer's cross section. de Barros and Nowak (2010) showed that **Eq. 4** controls the overall plume dispersion downstream from the source. Similar results are reported in Henri et al. (2016) in the context of risk analysis, where the authors showed that high water flux crossing the source zone leads to decrease in the magnitude of human health risk due to the presence of chlorinated solvents. Gueting and Englert (2013) also report experimental evidence regarding the importance of source zone hydraulics on transport behavior. Through the use of a Bayesian framework, Nowak et al. (2010) showed that characterizing the flow field surrounding the source zone could significantly reduce the uncertainty of transport observables.

Figure 7 shows the scatter plot of the maximum value of resilience loss R_L experienced within the target area ($\mathcal{A}_T = \Delta v_1 \times \Delta v_2$, see **Table 1**) and η. Each point in **Figure 7** represents the R_L obtained for each realization of hydraulic conductivity ensemble. The data are fitted by a logarithmic curve, the blue line, that has been found through genetic programming (see Im et al., 2021). The data suggest that scenarios characterized by high η values correspond to lower estimates of R_L. For completeness, we include the three plot insets of the plume. These insets belong to the red dots in **Figure 7** which correspond to 33^{rd}, 34^{th} and 102^{nd} Monte Carlo realizations. The results depicted in the plot insets suggest that η has a clear impact in controlling the overall longitudinal macrodispersion of the plume, which in turn will impact the value of R_L. Close inspection of **Figure 7** reveals that a realization characterized by a high value of η, such as realization number 102, is characterized by a compact (along the longitudinal direction) plume and therefore a lower R_L value when compared to realization number 33. When the strength of the contaminant source area decreases ($\eta < 1$), the plume is more dispersed. Increased plume spreading leads to larger plume residence times at an environmentally sensitive target and therefore higher R_L values. Note that the aforementioned conclusions are limited to the groundwater flow and transport scenario adopted, the initial concentration of the contaminant and the threshold concentration C^*. Nevertheless, the analysis carried out in this section re-emphasize the importance of 1) η in decision making and 2) the characterization of the source zone in risk analysis.

6 SUMMARY

In this work we provide `VisU-HydRA`, an open source, documented, computationally efficient toolbox to characterize specific features of the contaminant plume transport. The proposed package serves as a user-ready toolbox and allows to compute the uncertainty associated with metrics typically used in risk analysis. `VisU-HydRA` consists of a collection of rapid and open source software which have been assembled to deliver a rapid, reproducible and transparent modeling framework. Computational efficiency and rapidity are ensured by the usage of a GPU-accelerated solute transport simulator (Rizzo et al., 2019) and the automatized and solid structure of the iterative processes. Reproducibility and transparency are guaranteed by the open source coding language, the user-friendly interface of the Python code and the availability of a well documented and easy to follow tutorial made available on a web-based application.

In order to support the decision making process, the computational toolbox includes a visualization component that allows users to generate probability maps of aquifer resilience loss and risk hot spots. Furthermore, a GitHub repository was created and contains all the material reported to this work. The results reported are limited to a two dimensional application and the hydraulic conductivity field is the only source of uncertainty. As a consequence, the computational toolbox can be expanded to account for other sources of uncertainty as well as three dimensional models. In our work we opted to employ a two-dimensional "reference model" typically encountered in the stochastic hydrogeological community to study dispersion and mixing of solutes in heterogeneous aquifers. However, more realistic scenarios can be incorporated, such as the one reported in Fiori et al. (2019). We point out that the computational efficiency of the proposed toolbox could be improved by making use of other uncertainty quantification methodologies (as opposed to the classic Monte Carlo framework). For example, Olivier et al. (2020) presents an open-source, user-ready, Python package that includes several of the latest approaches for uncertainty estimation that are computationally efficient. The current contribution represents one step towards an integrated framework for analyzing groundwater contamination in risk assessment under uncertainty.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/mariamorvillo/VisU-HydRA.

AUTHOR CONTRIBUTIONS

MM, JI and FdeB were responsible for the conceptualization of the work. MM and FdeB wrote, reviewed and edited the paper. MM and JI performed all the simulations, implemented the computer code and performed the computational analysis. FdeB supervised the work, provided assistance and constructive critical support. All authors contributed to the interpretation of the results, provided critical feedback and helped shape the research, analysis and manuscript.

ACKNOWLEDGMENTS

The authors acknowledge the constructive comments made by the reviewers and Dr. Anneli Guthke. The authors also acknowledge the support provided by NSF Grant Number 1654009.

REFERENCES

Andričević, R., and Cvetković, V. (1996). Evaluation of Risk from Contaminants Migrating by Groundwater. *Water Resour. Res.* 32, 611–621.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting MODFLOW Model Development Using Python and Flopy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Ballio, F., and Guadagnini, A. (2004). Convergence Assessment of Numerical Monte Carlo Simulations in Groundwater Hydrology. *Water Resour. Res.* 40. doi:10.1029/2003wr002876

Bellin, A., and Rubin, Y. (1996). HYDRO_GEN: A Spatially Distributed Random Field Generator for Correlated Properties. *Stoch. Hydrol. Hydraul.* 10, 253–278. doi:10.1007/bf01581869

Boso, F., and Tartakovsky, D. M. (2016). The Method of Distributions for Dispersive Transport in Porous Media with Uncertain Hydraulic Properties. *Water Resour. Res.* 52, 4700–4712. doi:10.1002/2016wr018745

Cardiff, M., Liu, X., Kitanidis, P. K., Parker, J., and Kim, U. (2010). Cost Optimization of DNAPL Source and Plume Remediation under Uncertainty Using a Semi-analytic Model. *J. Contam. Hydrology* 113, 25–43. doi:10.1016/j.jconhyd.2009.11.004

Ciriello, V., and de Barros, F. P. J. (2020). Characterizing the Influence of Multiple Uncertainties on Predictions of Contaminant Discharge in Groundwater within a Lagrangian Stochastic Formulation. *Water Resour. Res.* 56, e2020WR027867. doi:10.1029/2020wr027867

Ciriello, V., Lauriola, I., Bonvicini, S., Cozzani, V., Di Federico, V., and Tartakovsky, D. M. (2017). Impact of Hydrogeological Uncertainty on Estimation of Environmental Risks Posed by Hydrocarbon Transportation Networks. *Water Resour. Res.* 53, 8686–8697. doi:10.1002/2017wr021368

Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. (2012). "Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository," in Proceedings of the ACM 2012 conference on computer supported cooperative work, 1277–1286.

de Barros, F. P. J., Ezzedine, S., and Rubin, Y. (2012). Impact of Hydrogeological Data on Measures of Uncertainty, Site Characterization and Environmental Performance Metrics. *Adv. Water Resour.* 36, 51–63. doi:10.1016/j.advwatres.2011.05.004

de Barros, F. P. J., Fiori, A., Boso, F., and Bellin, A. (2015). A Theoretical Framework for Modeling Dilution Enhancement of Non-reactive Solutes in Heterogeneous Porous Media. *J. Contam. Hydrology* 175-176, 72–83. doi:10.1016/j.jconhyd.2015.01.004

de Barros, F. P. J., and Fiori, A. (2014). First-order Based Cumulative Distribution Function for Solute Concentration in Heterogeneous Aquifers: Theoretical Analysis and Implications for Human Health Risk Assessment. *Water Resour. Res.* 50, 4018–4037. doi:10.1002/2013wr015024

de Barros, F. P. J., and Fiori, A. (2021). On the Maximum Concentration of Contaminants in Natural Aquifers. *Transp. Porous Med.* 140, 273–290. doi:10.1007/s11242-021-01620-3

de Barros, F. P. J., and Nowak, W. (2010). On the Link between Contaminant Source Release Conditions and Plume Prediction Uncertainty. *J. Contam. hydrology* 116, 24–34. doi:10.1016/j.jconhyd.2010.05.004

de Barros, F. P. J., and Rubin, Y. (2008). A Risk-Driven Approach for Subsurface Site Characterization. *Water Resour. Res.* 44. doi:10.1029/2007wr006081

Dentz, M., Le Borgne, T., Englert, A., and Bijeljic, B. (2011). Mixing, Spreading and Reaction in Heterogeneous Media: A Brief Review. *J. Contam. Hydrology* 120-121, 1–17. doi:10.1016/j.jconhyd.2010.05.002

Dentz, M., and Tartakovsky, D. M. (2010). Probability Density Functions for Passive Scalars Dispersed in Random Velocity Fields. *Geophys. Res. Lett.* 37. doi:10.1029/2010gl045748

Fienen, M. N., and Bakker, M. (2016). HESS Opinions: Repeatable Research: what Hydrologists Can Learn from the Duke Cancer Research Scandal. *Hydrol. Earth Syst. Sci.* 20, 3739–3743. doi:10.5194/hess-20-3739-2016

Fienen, M. N., Corson-Dosch, N. T., White, J. T., Leaf, A. T., and Hunt, R. J. (2022). Risk-Based Wellhead Protection Decision Support: A Repeatable Workflow Approach. *Groundwater* 60, 71–86. doi:10.1111/gwat.13129

Fiori, A., Bellin, A., Cvetkovic, V., de Barros, F. P. J., and Dagan, G. (2015). Stochastic Modeling of Solute Transport in Aquifers: From Heterogeneity Characterization to Risk Analysis. *Water Resour. Res.* 51, 6622–6648. doi:10.1002/2015wr017388

Fiori, A., and Dagan, G. (2000). Concentration Fluctuations in Aquifer Transport: A Rigorous First-Order Solution and Applications. *J. Contam. Hydrology* 45, 139–163. doi:10.1016/s0169-7722(00)00123-6

Fiori, A., Zarlenga, A., Bellin, A., Cvetkovic, V., and Dagan, G. (2019). Groundwater Contaminant Transport: Prediction under Uncertainty, with Application to the MADE Transport Experiment. *Front. Environ. Sci.* 7, 79. doi:10.3389/fenvs.2019.00079

Gueting, N., and Englert, A. (2013). Hydraulic Conditions at the Source Zone and Their Impact on Plume Behavior. *Hydrogeol. J.* 21, 829–844. doi:10.1007/s10040-013-0962-7

Hammond, G. E., Lichtner, P. C., and Mills, R. T. (2014). Evaluating the Performance of Parallel Subsurface Simulators: An Illustrative Example with PFLOTRAN. *Water Resour. Res.* 50, 208–228. doi:10.1002/2012wr013483

Harbaugh, A. W. (2005). *MODFLOW-2005, the US Geological Survey Modular Ground-Water Model: The Ground-Water Flow Process*. Reston, VA, USA: US Department of the Interior, US Geological Survey Reston.

Henri, C. V., Fernàndez-Garcia, D., and de Barros, F. P. J. (2016). Assessing the Joint Impact of DNAPL Source-Zone Behavior and Degradation Products on the Probabilistic Characterization of Human Health Risk. *Adv. Water Resour.* 88, 124–138. doi:10.1016/j.advwatres.2015.12.012

Im, J., Rizzo, C. B., de Barros, F. P. J., and Masri, S. F. (2021). Application of Genetic Programming for Model-free Identification of Nonlinear Multi-Physics Systems. *Nonlinear Dyn.* 104, 1781–1800. doi:10.1007/s11071-021-06335-0

Im, J., Rizzo, C. B., and de Barros, F. P. J. (2020). Resilience of Groundwater Systems in the Presence of Bisphenol A under Uncertainty. *Sci. Total Environ.* 727, 138363. doi:10.1016/j.scitotenv.2020.138363

Kapoor, V., and Kitanidis, P. K. (1998). Concentration Fluctuations and Dilution in Aquifers. *Water Resour. Res.* 34, 1181–1193. doi:10.1029/97wr03608

Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge: Cambridge University Press.

Li, S.-G., and Liu, Q. (2006). Interactive Ground Water (IGW). *Environ. Model. Softw.* 21, 417–418. doi:10.1016/j.envsoft.2005.05.010

Logue, J. N., and Fox, J. M. (1986). Residential Health Study of Families Living Near the Drake Chemical Superfund Site in Lock Haven, Pennsylvania. *Archives Environ. Health Int. J.* 41, 222–228. doi:10.1080/00039896.1986.9938337

Maxwell, R. M., Carle, S. F., and Tompson, A. F. B. (2008). Contamination, Risk, and Heterogeneity: on the Effectiveness of Aquifer Remediation. *Environ. Geol.* 54, 1771–1786. doi:10.1007/s00254-007-0955-8

Maxwell, R. M., Condon, L. E., and Kollet, S. J. (2015). A High-Resolution Simulation of Groundwater and Surface Water over Most of the Continental US with the Integrated Hydrologic Model ParFlow V3. *Geosci. Model. Dev.* 8, 923–937. doi:10.5194/gmd-8-923-2015

Meyer, D. W., Tchelepi, H. A., and Jenny, P. (2013). A Fast Simulation Method for Uncertainty Quantification of Subsurface Flow and Transport. *Water Resour. Res.* 49, 2359–2379. doi:10.1002/wrcr.20240

Mood, A. M., Graybill, F. A., and Boes, C. D. (1974). *Introduction to the Theory of Statistics*. International Student Edition. New York, NY, USA: McGraw-Hill.

Neuman, S. P., and Tartakovsky, D. M. (2009). Perspective on Theories of Non-fickian Transport in Heterogeneous Media. *Adv. Water Resour.* 32, 670–680. doi:10.1016/j.advwatres.2008.08.005

Nowak, W., De Barros, F. P. J., and Rubin, Y. (2010). Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. , 46(3). *Water Resour. Res.* 46(3).

Oladyshkin, S., and Nowak, W. (2012). Data-driven Uncertainty Quantification Using the Arbitrary Polynomial Chaos Expansion. *Reliab. Eng. Syst. Saf.* 106, 179–190. doi:10.1016/j.ress.2012.05.002

Oliphant, T. E. (2006). *A Guide to NumPy, Vol. 1*. USA: Trelgol Publishing.

Olivier, A., Giovanis, D. G., Aakash, B. S., Chauhan, M., Vandanapu, L., and Shields, M. D. (2020). Uqpy: A General Purpose python Package and Development Environment for Uncertainty Quantification. *J. Comput. Sci.* 47, 101204. doi:10.1016/j.jocs.2020.101204

Peñuela, A., Hutton, C., and Pianosi, F. (2021). An Open-Source Package with Interactive Jupyter Notebooks to Enhance the Accessibility of Reservoir Operations Simulation and Optimisation. *Environ. Model. Softw.* 145, 105188. doi:10.1016/j.envsoft.2021.105188

Proctor, C. R., Lee, J., Yu, D., Shah, A. D., and Whelton, A. J. (2020). Wildfire Caused Widespread Drinking Water Distribution Network Contamination. *AWWA Water Sci.* 2, e1183. doi:10.1002/aws2.1183

Rizzo, C. B., Nakano, A., and de Barros, F. P. J. (2019). PAR2: Parallel Random Walk Particle Tracking Method for Solute Transport in Porous Media. *Comput. Phys. Commun.* 239, 265–271. doi:10.1016/j.cpc.2019.01.013

Rubin, Y. (2003). *Applied Stochastic Hydrogeology*. Oxford, UK: Oxford University Press.

Siirila, E. R., Navarre-Sitchler, A. K., Maxwell, R. M., and McCray, J. E. (2012). A Quantitative Methodology to Assess the Risks to Human Health from CO2 Leakage into Groundwater. *Adv. Water Resour.* 36, 146–164. doi:10.1016/j.advwatres.2010.11.005

Sivasankar, V., Senthil kumar, M., and Gopalakrishna, G. (2017). Quantification of Benzene in Groundwater Sources and Risk Analysis in a Popular South Indian Pilgrimage City–A GIS Based Approach. *Arabian J. Chem.* 10, S2523–S2533. doi:10.1016/j.arabjc.2013.09.022

vanRossum, G. (1995). *Python Reference Manual*. Blacksburg, VA, USA: Department of Computer Science [CS].

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020). Toward Reproducible Environmental Modeling for Decision Support: a Worked Example. *Front. Earth Sci.* 8, 50. doi:10.3389/feart.2020.00050

Woodruff, M. J., Reed, P. M., and Simpson, T. W. (2013). Many Objective Visual Analytics: Rethinking the Design of Complex Engineered Systems. *Struct. Multidisc Optim.* 48, 201–219. doi:10.1007/s00158-013-0891-z

Ye, Y., Chiogna, G., Cirpka, O. A., Grathwohl, P., and Rolle, M. (2015). Enhancement of Plume Dilution in Two-dimensional and Three-dimensional Porous Media by Flow Focusing in High-permeability Inclusions. *Water Resour. Res.* 51, 5582–5602. doi:10.1002/2015wr016962

Zhang, D., Shi, L., Chang, H., and Yang, J. (2010). A Comparative Study of Numerical Approaches to Risk Assessment of Contaminant Transport. *Stoch. Environ. Res. Risk Assess.* 24, 971–984. doi:10.1007/s00477-010-0400-5

Zheng, C., and Bennett, G. D. (2002). *Applied Contaminant Transport Modeling, Vol. 2*. New York: Wiley-Interscience.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A FLOW AND TRANSPORT EQUATIONS

For computational illustrations, we simulated a steady-state fully saturated incompressible flow in a spatially heterogeneous aquifer in absence of sink or sources. Flow is 2D, and in our computational 2D domain, x_1 denotes the longitudinal dimension and x_2 the transverse one. The steady flow field is governed by:

$$\nabla \cdot [K(\mathbf{x})\nabla h(\mathbf{x})] = 0, \tag{A1}$$

with h denoting the hydraulic head and K the hydraulic conductivity.

For all our simulations, we considered permeameter-like boundary conditions for the flow field. That is, no-flux boundary conditions in the transverse boundaries and constant heads, respectively h_{in} and h_{out}, are adopted in the inflow and outflow boundaries of the computational domain.

For the transport simulation we consider that instantaneous release of Benzene within a rectangular source zone of area $\mathcal{A}_o = \Delta s_1 \times \Delta s_2$. The spatiotemporal evolution of the concentration field is assumed to be governed by the advection-dispersion equation:

$$\frac{\partial C(\mathbf{x},t)}{\partial t} + \mathbf{u}(\mathbf{x}) \cdot \nabla C(\mathbf{x},t) = \nabla \cdot [\mathbf{D}(\mathbf{x})\nabla C(\mathbf{x},t)]. \tag{A2}$$

where C is the resident concentration, $\mathbf{u}$ is the velocity field, $\mathbf{D}$ is the local-scale dispersion tensor assumed to be anisotropic and defined as:

$$\mathbf{D}(\mathbf{x}) = (\alpha_T|\mathbf{u}(\mathbf{x})| + D_m)\mathbf{I} + \frac{\alpha_L - \alpha_T}{|\mathbf{u}(\mathbf{x})|}\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^T \tag{A3}$$

where D_m is the molecular diffusion, α_L is the longitudinal (along x_1) dispersivity and α_T is the transverse (along x_2) dispersivity.

frontiers | Frontiers in Earth Science

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/feart.2022.907609

Application of Time Series Analysis to Estimate Drawdown From Multiple Well Fields

*Davíd A. Brakenhoff[1], Martin A. Vonk[1,2], Raoul A. Collenteur[3], Marco Van Baar[1] and Mark Bakker[2]**

[1]Artesia B.V, Schoonhoven, Netherlands, [2]Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands, [3]NAWI Graz Geocenter, Institute of Earth Sciences, University of Graz, Graz, Austria

In 2018–2020, meteorological droughts over Northwestern Europe caused severe declines in groundwater heads with significant damage to groundwater-dependent ecosystems and agriculture. The response of the groundwater system to different hydrological stresses is valuable information for decision-makers. In this paper, a reproducible, data-driven approach using open-source software is proposed to quantify the effects of different hydrological stresses on heads. A scripted workflow was developed using the open-source Pastas software for time series modeling of heads. For each head time series, the best model structure and relevant hydrological stresses (rainfall, evaporation, river stages, and pumping at one or more well fields) were selected iteratively. A new method was applied to model multiple well fields with a single response function, where the response was scaled by the distances between the pumping and observation wells. Selection of the best model structure was performed through reliability checking based on four criteria. The time series model of each observation well represents an independent estimate of the contribution of different hydrological stresses to the head and is based exclusively on observed data. The approach was applied to estimate the drawdown caused by nearby well fields to 250 observed head time series measured at 122 locations in the eastern part of the Netherlands, a country where summer droughts can cause problems, even though the country is better known for problems with too much water. Reliable models were obtained for 126 head time series of which 78 contain one or more well fields as a contributing stress. The spatial variation of the modeled responses to pumping at the well fields show the expected decline with distance from the well field, even though all responses were modeled independently. An example application at one well field showed how the head response to pumping varies per aquifer. Time series analysis was used to determine the feasibility of reducing pumping rates to mitigate large drawdowns during droughts, which depends on the magnitude and response time of the groundwater system to changes in pumping. This is salient information for decision-makers. This article is part of the special issue “Rapid, Reproducible, and Robust Environmental Modeling for Decision Support: Worked Examples and Open-Source Software Tools”.

Keywords: time series analysis, groundwater, decision support, reproducible, model selection, Hantush response function, well drawdown

OPEN ACCESS

Edited by:
Jeremy White,
Intera, Inc., United States

Reviewed by:
Antonia Longobardi,
University of Salerno, Italy
Andres Gonzalez Quiros,
British Geological Survey, The Lyell Centre, United Kingdom

****Correspondence:***
Mark Bakker
mark.bakker@tudelft.nl

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *29 March 2022*
Accepted: *23 May 2022*
Published: *14 June 2022*

Citation:
Brakenhoff DA, Vonk MA, Collenteur RA, Van Baar M and Bakker M (2022) Application of Time Series Analysis to Estimate Drawdown From Multiple Well Fields. Front. Earth Sci. 10:907609. doi: 10.3389/feart.2022.907609

1 INTRODUCTION

The competition for groundwater resources is fierce, including demands for agricultural production, drinking water supply, groundwater-dependent ecosystems, and for mitigation of land subsidence to maintain the stability of buildings. Dry summers and the growing demand for freshwater increases the pressure on limited groundwater resources. The groundwater table may drop significantly during and after dry summers (e.g., Brakkee et al., 2022) due to a set of stresses on the system including a decrease in precipitation, an increase in evaporation and transpiration, lower surface water levels, and higher groundwater use for both drinking water and irrigation (e.g., Van Loon et al., 2016). The effect of pumping wells on the head is one of the few stresses on the system that can be controlled. Estimates of the head response to pumping wells are therefore salient information for decision makers to manage groundwater resources and possibly mitigate low groundwater tables.

The effect of pumping on the heads in a multi-aquifer system can be estimated with a numerical groundwater model (e.g., Anderson et al., 2015). Such process-based models typically require a large amount of input data to incorporate system and process details (e.g., Hugman and Doherty, 2022). Significant time investment is required to build and calibrate these models, and, even after considerable effort, they are rarely able to simulate the transient head variation with reasonable accuracy. Alternatively, models based on time series analysis are generally much better at simulating the heads measured in an observation well (e.g., Bakker and Schaars, 2019). Additionally, time series models have the advantage of low data requirements and can be developed in a short amount of time.

Many time series analysis approaches are black-box models, for example ARIMA models (e.g., Patle et al., 2015) or deep learning methods (e.g., Wunsch et al., 2018). A disadvantage of such black-box models is that it can be difficult to physically interpret the resulting models. More transparent, gray-box approaches include lumped conceptual models (Mackay et al., 2014) or time series modeling using physically-based response functions (e.g., Von Asmuth et al., 2012; Collenteur et al., 2019). The latter also allows for the differentiation between the stresses causing the head variation (e.g., Von Asmuth et al., 2008).

An important application of time series analysis is the estimation of the drawdown caused by pumping. For example, Von Asmuth et al. (2008), Obergfell et al. (2013), and Shapoori et al. (2015) applied time series analysis to determine the drawdown due to pumping from a single well field. Many observation wells worldwide are impacted by multiple well fields. The pumping rates at these well fields are commonly correlated, which complicates the estimation of the contributions of different well fields and may lead to increased uncertainty in the model outcomes and less robust models.

The objective of this paper is to present a data-driven, reproducible, and robust approach to estimate the head response at observation wells that are potentially affected by variations in rainfall, evaporation, rivers stages, and pumping from multiple well fields. The main objective is to quantify both the magnitude and timing of the head response to the surrounding well fields at each observation well using a new parsimonious approach to incorporate multiple well fields in a time series model. The approach is tested in an area of the Netherlands where the heads are measured at 213 observation wells at multiple depths, resulting in 395 head time series. The heads are potentially affected by four different well fields. A detailed decision tree is developed to determine which stresses have a significant effect on the head variation. Time series analysis is conducted with the open-source Pastas software (version 0.20.0 Collenteur et al., 2019) to determine the response of each well field. The analysis is entirely implemented in Python scripts and is fully reproducible as advocated by Fienen and Bakker (2016) & White et al. (2020).

In the following, the approach to quantify the effects of groundwater pumping using time series analysis is presented. Next, the study area and all available data are described and the results of the analysis are presented including an estimate of the uncertainty. A possible application of the results is presented for the mitigation of low heads in dry summers. The applicability and limitations of the method are discussed, including some challenges faced while performing the study. Concluding remarks are presented at the end of this paper.

2 METHODOLOGY

A time series model represents an independent estimate of the contribution of different stresses on the heads in an observation well that is derived exclusively from observed data. A multi-model approach is applied to determine which hydrological stresses are relevant in describing the head dynamics in an observation well.

Precipitation-excess, river stage, and groundwater pumping are included as potentially relevant hydrological stresses. Eight different model structures are tested for each head time series. The simplest model considers only precipitation-excess, computed from precipitation and potential evaporation. The next model adds the river stage as a stress. In the next three models, up to three well fields are added as potential stresses, starting from the closest well field and moving towards the farthest one. The final three models repeat this last step but leave out the river as a stress.

A set of criteria is used to determine which model structures are deemed reliable. The best model structure is selected from the set of reliable models for each observation well. Split-sample testing, in which a portion of the time series is kept separate, is applied to test the calibrated model.

2.1 Time Series Modeling

The time series modeling approach, also referred to as transfer function noise modeling, uses physically-based impulse response functions that describe the head response to different stresses (Von Asmuth et al., 2002). Simulation of the effect of precipitation-excess, river stage variations, and noise modeling

is based on the standard approach of Von Asmuth et al. (2008). The time series model is written as

$$h(t) = \sum_{m=1}^{M} h_m(t) + d + r(t) \tag{1}$$

where $h(t)$ are the observed heads, $h_m(t)$ is the contribution of stress m to the head, d is the base elevation of the model, and $r(t)$ are the residuals. Each model has an arbitrary number of stresses M, depending on the chosen model structure. The contribution of each stress is computed through convolution as

$$h_m(t) = \int_{-\infty}^{t} S_m(\tau)\theta_m(t-\tau)d\tau \tag{2}$$

where $S_m(t)$ is the time series of a stress m and θ_m is the associated impulse response function. An auto-regressive noise model of order 1 (AR1) is used in an attempt to transform the residuals into a noise time series $n(t)$ that is approximately white noise (Von Asmuth and Bierkens, 2005)

$$n(t_i) = r(t_i) - \mathrm{e}^{(t_i - t_{i-1})/\alpha} \tag{3}$$

where $n(t_i)$ is the remaining noise at time t_i and α is the auto-regressive parameter.

The precipitation-excess, $N(t)$, is modeled as

$$N(t) = P(t) - fE_r(t) \tag{4}$$

where $P(t)$ is the precipitation, $E_r(t)$ is the Makkink reference evaporation (de Bruin and Lablans, 1998), and parameter f is used to scale the reference evaporation to local hydrological conditions. The impulse response of groundwater to precipitation-excess is described using the scaled Gamma distribution (Collenteur et al., 2019)

$$\theta(t) = \frac{A}{a^n \Gamma(n) t^{1-n}} \mathrm{e}^{-t/a} \tag{5}$$

where A, a, and n are fitting parameters. In this formulation of the response function, parameter A is the gain of the response function, i.e., the rise in the head due to a constant unit precipitation-excess. The groundwater response to river stage fluctuations is described with an exponential response function, which is a special case of the scaled Gamma response function with $n = 1$.

The head response to groundwater pumping may be simulated with a response function that has the same mathematical form as the Hantush well function (Hantush and Jacob, 1955). There is a risk of over-parameterization of a time series model when multiple pumping wells are added to the model that each have their own response function and corresponding parameters. For example, adding three pumping wells with a Hantush well function would already add 9 parameters to the model. The use of a single response function is proposed, scaled with the distance to the well field, to quantify the effect of all groundwater pumping wells. The response function is based on the Hantush response function used by Von Asmuth et al. (2008) and is modified to include the distance of the well field to the observation well r explicitly, so that the impulse response function is

$$\theta(r,t) = \frac{A}{2t} \mathrm{e}^{-t/a - abr^2/t} \tag{6}$$

where A, a, and b are fitting parameters. The gain of the response function is $AK_0(2r\sqrt{b})$, where K_0 is the modified Bessel function of the second kind and order zero. It is noted here that the parameter A for this response function does not equal the gain.

The step response $\Theta(t)$, the response to a constant unit stress, is obtained from the impulse response function through integration

$$\Theta(t) = \int_0^t \theta(t-\tau)\mathrm{d}\tau \tag{7}$$

An example of the Hantush step response is shown in **Figure 1**. The t_{50} and t_{95} represent the time when 50 and 95% of the total response has occurred, respectively. For this modified Hantush response function, the t_{50} can be conveniently computed following Veling and Maas (2010).

$$t_{50} = ar\sqrt{b} \tag{8}$$

The calculation of the variances of the gain and the t_{50} is provided in the **Supplementary Material**.

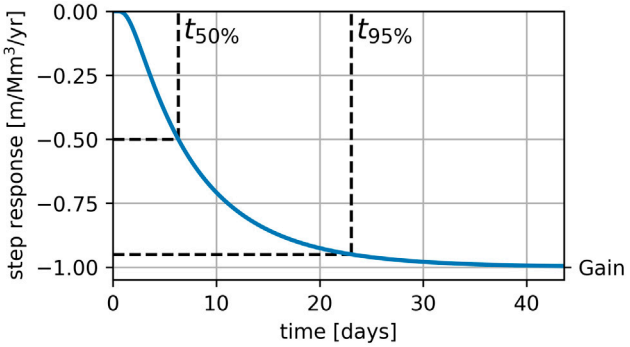

FIGURE 1 | An example of the Hantush step response. The t_{50} and t_{95} represent the time when 50 and 95% of the total response has occurred, respectively.

2.2 Model Calibration, Reliability Criteria, and Selection

The most complex model considered in this paper has a total of eleven parameters: four parameters for the response to precipitation-excess, two parameters for the response to river stages, three parameters for the response to pumping wells, one parameter for the noise model, and one parameter for the base elevation of the model.

The head time series of each observation well is divided into a calibration period and a validation period. The calibration data is used to calibrate each time series model with a two-step optimization approach following Collenteur

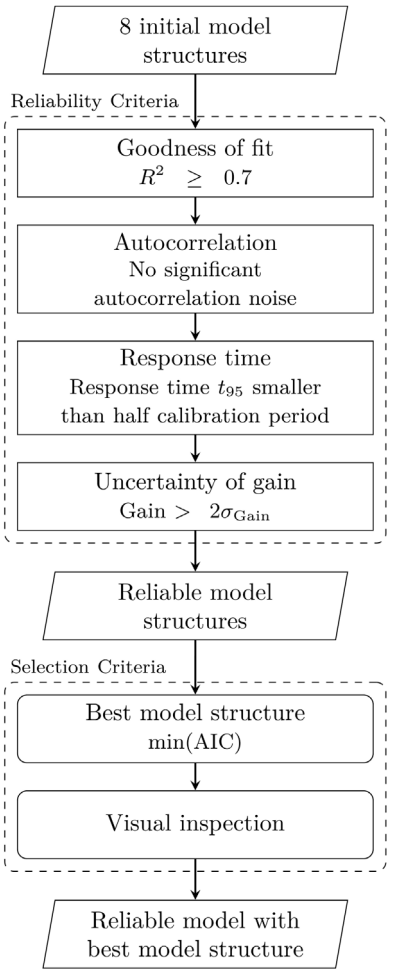

FIGURE 2 | Criteria for reliable models and selection of best model structure.

et al. (2021). In the first step, the parameters are optimized without the use of a noise model by minimizing the sum of squared residuals. In the second step, the noise model is added and the sum of squared noise is minimized using the optimized model parameters from the first step as initial parameter values.

A set of criteria is applied to all eight model structures to determine which model structures are considered reliable for further analysis. A reliable model is defined here as a model meeting four acceptance criteria. From the model structures passing these criteria, a single model structure is selected for each observation well based on two selection criteria. The selection scheme including all reliability criteria is presented in **Figure 2**. The following four acceptance criteria are used:

1) Goodness of fit. The model goodness of fit in the calibration period, measured as the coefficient of determination (R^2), must be equal or larger than 0.7, which means that the model has at least a basic fit.
2) Autocorrelation. There must be no significant autocorrelation in the noise. This is determined with the Runs-test for autocorrelation (Wald and Wolfowitz, 1940) using a significance level of $\alpha = 0.05$. This requirement is important to obtain reliable estimates of the parameter uncertainties (Hipel and McLeod, 1994).
3) Response time. The response time, expressed as the t_{95} (see **Figure 1**), must not exceed half the length calibration period. The calibration time series is potentially too short to accurately estimate the parameters of the response function when the t_{95} of the response is longer than half the length of the calibration period.
4) Uncertainty of gain. The estimated gain of each response function must be significantly different from zero. This is checked by requiring that the estimated gain is larger than twice the estimated standard deviation of the gain (e.g., Collenteur et al., 2019).

When multiple model structures are reliable, the Akaike Information Criterion (AIC; Akaike, 1974) is used to select the best model structure, by selecting the model with the lowest AIC (Burnham et al., 2011). After the AIC selection, the selected model structure is visually inspected for both the calibration and validation periods. Model structure must perform well in both the calibration and the validation period.

The described approach to determine the best model structure for each observation well is applied to all observation wells in a study area. The entire analysis is implemented in Python scripts to ensure reproducibility of the results. All data, scripts, and environment settings required to reproduce the results from this study are available from Zenodo (Brakenhoff et al., 2022).

3 STUDY SITE AND DATA

The study area is the Overbetuwe area in the Netherlands, a polder region of approximately 30 km by 10 km, flanked by two branches of the Rhine river (see **Figure 3**). The land surface elevation varies from around +10 m in the east to around +7 m in the west (all elevations are given relative to the Dutch reference level called NAP, which is approximately equal to mean sea-level). The region is divided into several polders that each strive to keep water levels at a fixed level with a complex system of ditches, canals, weirs, and pumping stations. The land use is a mix of agriculture, nature, and urban environments.

The shallow subsurface is characterized by a low-permeable phreatic layer consisting mostly of clay and clayey sand, underlain by two aquifers, separated by an aquitard (see **Figure 3**). The aquitard consists of clay with a thickness varying from 0 to 15 m. The groundwater is relatively shallow with the depth to water table varying between 0.8 and 4.2 m.

Heads are actively monitored at 213 observation wells in the study area, some measuring heads in multiple filters at different depths, resulting in 395 head time series. Heads

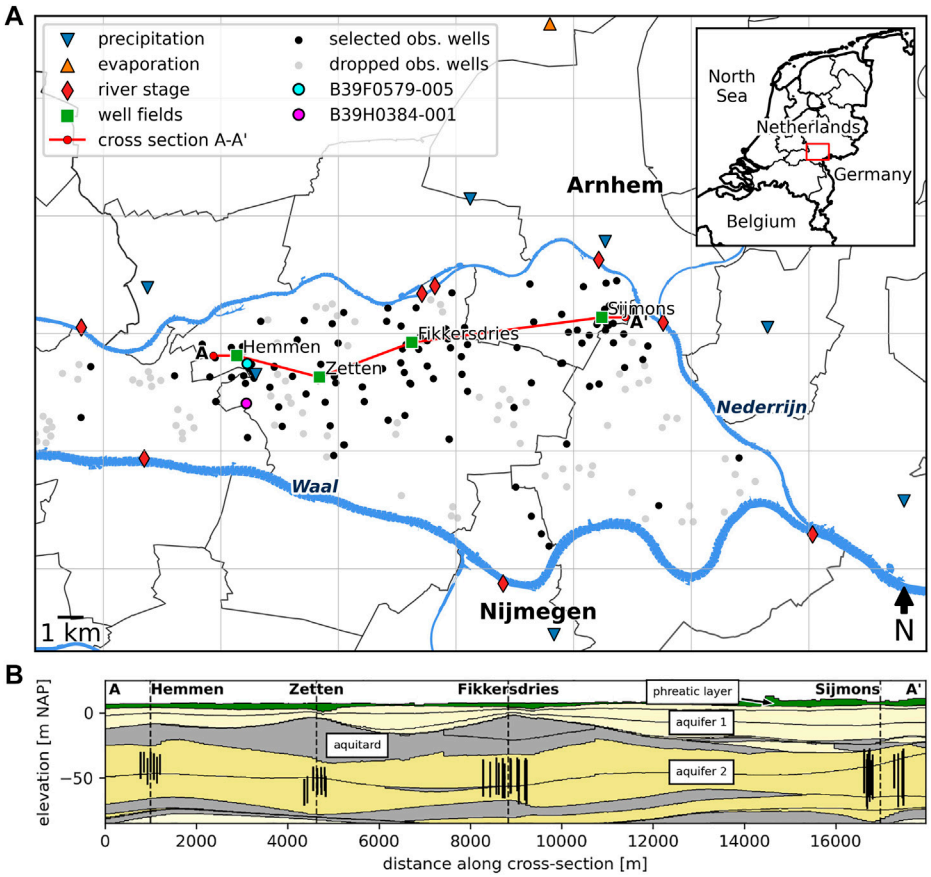

FIGURE 3 | Overview of study area with locations of observation wells, and locations at which stresses are measured **(A)**. Cross-section of the subsurface along well fields showing well screens, aquifers and aquitards **(B)**.

are measured with automatic pressure loggers, with daily, or shorter, measurement intervals for the period 2004–2020. Head measurements prior to 2004 are manual measurements, which are available at lower frequencies for some of the wells.

Daily precipitation data is available at seven measurement stations in the region (KNMI, 2022). Daily Makkink reference evaporation (de Bruin and Lablans, 1998) is available at two automatic weather stations. Mean yearly precipitation for 1990–2021 is 850 mm/year while mean yearly reference evaporation is 584 mm/year.

The river stage is measured at 10-min intervals at eight observation stations along both rivers (**Figure 3**) (Rijkswaterstaat, 2022). The time series are resampled to daily mean values. The maximum recorded daily mean river stage in the period 1990–2021 is 15.8 m, while the minimum recorded stage is 2.9 m.

Drinking water is extracted from the second aquifer at four well fields, at depths between −28 m and −74 m. From west to east, the well fields are Hemmen, Zetten, Fikkersdries, and Sijmons (see **Figure 3**). The date when pumping started, the average well screen depth, and mean pumping discharge are summarized in **Table 1**.

3.1 Data Preparation

The calibration period was selected as 1990–2014 and the validation period as 2015–2021. All head data was pre-screened. Outliers were removed and head data was corrected for sudden unexpected jumps in the time series. Time series were discarded if they had fewer than 6 years with at least 180 measurements per year in the calibration period and/or fewer than one year of at least 180 measurements in the validation period. In addition, time series were discarded that visually showed a strong effect of the on- and off-switching of individual pumping wells in a well field. The resulting dataset consists of 250 head time series at 122 observation wells. Each head time series is assigned to one of the aquifers based on the observation depth.

The river stage is spatially interpolated at the point nearest to the observation well along the center line of the nearest river. Time series are calculated using a distance-weighted average between two observation stations. If the nearest point does not lie between two observation stations, the time series of the nearest observation station is used. The time series of the river stage is normalized by subtracting the mean.

The pumping data was resampled to obtain a time series of daily discharge for each well field. The available data was a mix of

TABLE 1 | Average pumping depth, pumping start date, mean discharge in the period 1990–2021, and the coefficient of variation (CV) for the four well fields. The coefficient of variation is calculated by dividing the standard deviation of the discharge by the mean discharge in the period 1990–2014.

	Start Date	Screen Top	Screen Bottom	Mean Discharge	CV
Well field		**[m]**	**[m]**	**[Mm^3/yr]**	**[-]**
Hemmen	2006–10-01	−32	−49	1.97	0.61
Zetten	2006–10-01	−46	−63	3.62	0.53
Fikkersdries	1961–06-01	−37	−63	12.12	0.11
Sijmons	1980–01-01	−32	−66	4.00	0.28

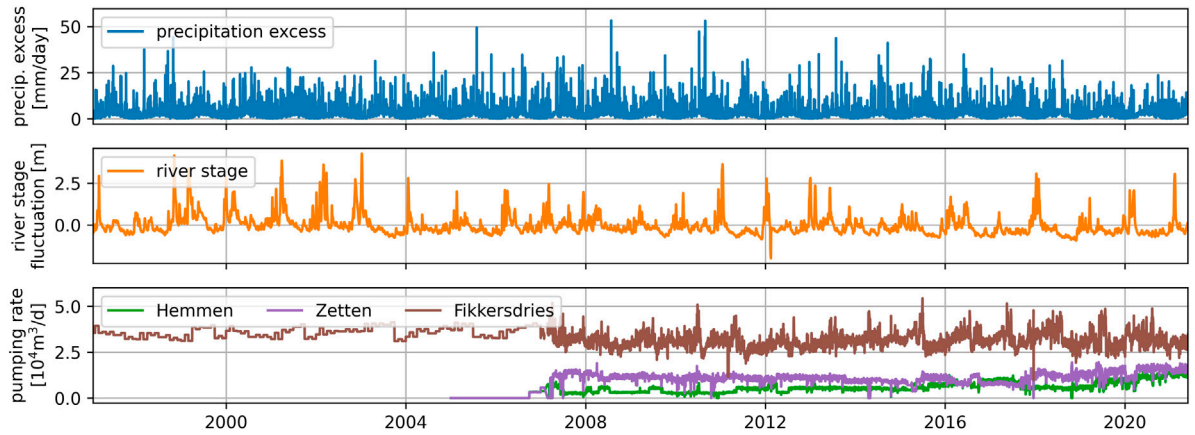

FIGURE 4 | Stresses included in the time series model for observation well B39F0579 (Filter 5).

monthly (before 2007) and daily (after 2007) volumes. The time series prior to 2007 were converted to daily volumes by equally dividing the monthly volumes over each day in the month. The time series of daily pumping discharge for each well field are provided in the **Supplementary Material** (**Supplementary Figure S1**). The average location of all wells in a well field was used to measure the distance between a well field and an observation well (r in **Eq. 6**).

Heads are computed in the calibration period using daily data for all stresses. The noise model (**Eq. 3**) is rarely adequate for obtaining uncorrelated noise when using daily head observations, but works reasonably well for head data at 14-days intervals (e.g., Von Asmuth and Bierkens, 2005; Collenteur et al., 2021). The calibration data is obtained by taking a sample from each head time series on a 14-days interval within the calibration period.

4 RESULTS

4.1 Example Results at One Observation Well

The results obtained for one observation well are discussed here in detail to illustrate the output of the time series model. Consider observation well B39F0579 (highlighted point in **Figure 3**), situated close to pumping station Hemmen (0.6 km) and at larger distances from stations Zetten (3.1 km) and Fikkersdries (7.1 km). Precipitation-excess, reference evaporation, river stage, and pumping rates from all three pumping wells are shown in **Figure 4**. Pumping at the well field in Fikkersdries started in the 1960s, before the start of the head observations. The pumping stations Hemmen and Zetten started operation in late 2006 with a relatively constant pumping rate until 2015, after which the pumping rate varied somewhat, with a significant increase in pumping in Hemmen in the last year of data. Observed heads in screen 5 of well B39F0579 (located in aquifer 2) are shown in the top graph of **Figure 5**. A clear decrease in head is visible from 2007 onwards, which coincides with the start of pumping. A further decline in heads is measured after 2015.

Time series models are developed with eight different model structures, as described in the previous section. Out of the 8 model structures, 6 model structures passed all four reliability checks. The selected best model structure includes precipitation-excess, the variation of the river stage, and three pumping wells. The results are shown in **Figure 5**. The simulated heads show a good fit with the data, as shown by a $R^2 = 0.90$ and $R^2 = 0.79$ in the calibration and the validation period respectively. During the validation period, the model overestimates the head in the summer months. These summers were particularly dry (e.g., Brakkee et al., 2022), and possibly stresses not included in the model (e.g., pumping for irrigation) could explain these deviations, but this has not been investigated here.

The contribution of each stress (precipitation excess, river stage, and pumping at the well fields) to the changes in head and

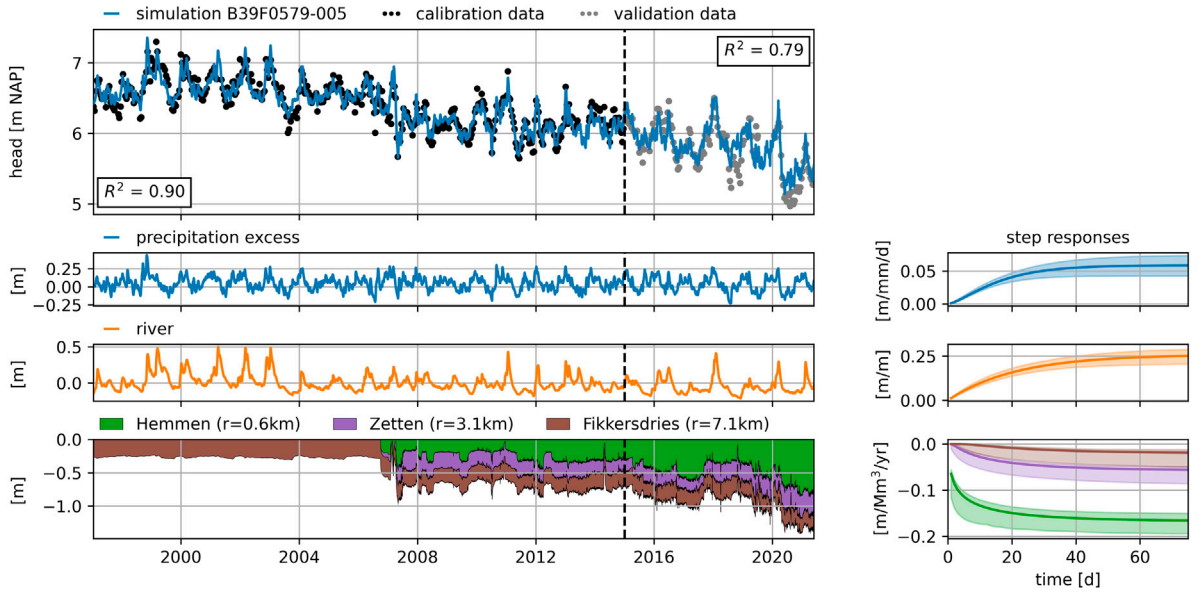

FIGURE 5 | Contribution of the different stresses and the estimated step responses for the example model for observation well B39F0579 (Filter 5) in aquifer 2. The shaded areas around the step responses represent the 95% confidence intervals.

TABLE 2 | Results for the reliability and selection criteria for all 2000 time series models for 250 locations.

	No. of Models
1. Goodness of fit	920
2. Autocorrelation	1558
3. Response time	1269
4. Uncertainty gain	743
Reliable models	247
Best models based on AIC	129
Passed visual inspection	126
Selected Models	126

their associated step response functions, as determined by the time series model, are shown in separate graphs in **Figure 5**. Up to 2006, the well field Fikkersdries caused a small drawdown that was stable over time. The drawdown caused by the other well fields started in late 2006 and stayed relatively constant until 2019, after which the drawdown increased as a result of increased pumping rates at Hemmen and to some extent Zetten. The total drawdown caused by all well fields exceeded 1 m after 2020, according to the model. The calibrated response functions (see plots on the right in **Figure 5**) are used to quantify the magnitude and timing of the drawdown caused by pumping from the three well fields.

4.2 Results for all Observation Wells

For each of the 250 head time series in the data set, 8 models with different structures were created and calibrated. The resulting 2000 time series models were evaluated and a best model structure was selected following the approach outlined in **Figure 2**. The total number of models that meet all four reliability criteria are presented in **Table 2**. 247 models for 129 unique head time series meet all four reliability criteria and are considered reliable. Some time series have multiple reliable models. For 121 (48%) time series no reliable model was present in the set of 8 model structures, and these time series are not considered further. Selection of the best model structure (according to the AIC) at each location, followed by a visual inspection, yields 126 (50%) reliable models that are used for further analysis. Out of these 126 models, 75 models include pumping at one or more well fields as a stress. **Table 3** summarizes the model structures of the selected models, categorized per aquifer.

The steady-state drawdown caused by a well field is computed for all 75 observation wells where at least one well field has a significant effect on the head. The steady-state drawdown is computed using the average discharge of each well field for the period 2015–2021 and is plotted versus the distance between the observation well and the well field in **Figure 6**. The estimated steady-state drawdown in aquifer 2 shows a clear relationship with distance at well fields Hemmen, Zetten, and Sijmons. The estimated drawdown in aquifer 2 decreases with distance from those well fields (green symbols). There are insufficient models that meet all reliability criteria in aquifer 2 near well field Fikkersdries to discern any pattern.

In the phreatic layer and aquifer 1, no spatial pattern in drawdown is discernible. At Hemmen, the phreatic drawdown is in the same order of magnitude as the drawdown in aquifer 1. At Zetten, there are no models for observation well screens located in the top two layers within the first 2 km that passed

TABLE 3 | Summary of model structures for selected time series models, counted per aquifer.

Model Structure	Phreatic	Aquifer 1	Aquifer 2	Total
Precipitation excess	5	6	1	12
Precipitation excess + river	3	23	13	39
Precipitation excess + river + 1 well	6	20	27	53
Precipitation excess + river + 2 wells	0	2	4	6
Precipitation excess + river + 3 wells	0	0	4	4
Precipitation excess + 1 well	4	5	3	12
Precipitation excess + 2 wells	0	0	0	0
Precipitation excess + 3 wells	0	0	0	0
Total selected (percentage)	18 (38%)	56 (45%)	52 (66%)	126 (50%)
Total no. of time series	47	124	79	250

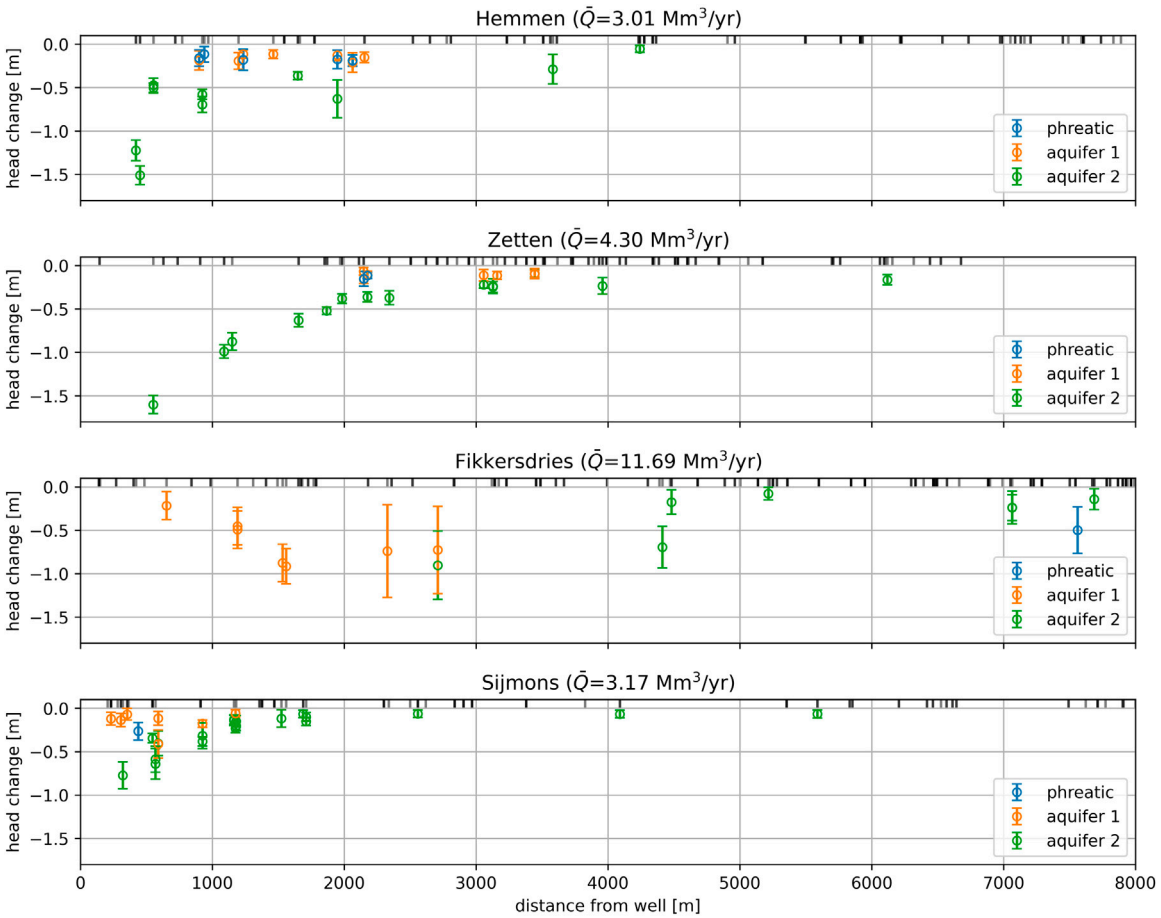

FIGURE 6 | The estimated steady state head change versus the distance between the well field and the observation well using the average pumping rate ($\bar{Q}$) between 2015 and 2020 for each well field. Uncertainty bars are 2 times the estimated standard deviation. The black ticks at the top of each graph indicate the distances at which an observation well is present but either none of the model structures was reliable or the well field was not included in the selected model structure.

all reliability criteria. At Sijmons, there are almost no models for phreatic observation wells and there is no discernible pattern in aquifer 1. The estimated steady state drawdown is much smaller in the top two layers, which are separated by an aquitard from the pumped aquifer. The vertical resistance of the aquitard is lowest around Hemmen and highest around Zetten (see **Supplementary Figure S2** showing aquitard resistance in the **Supplementary Material**). This fits well with the results from time series analysis, where a significant effect of pumping was estimated in shallow observation wells near Hemmen, whereas this is not the case near Zetten.

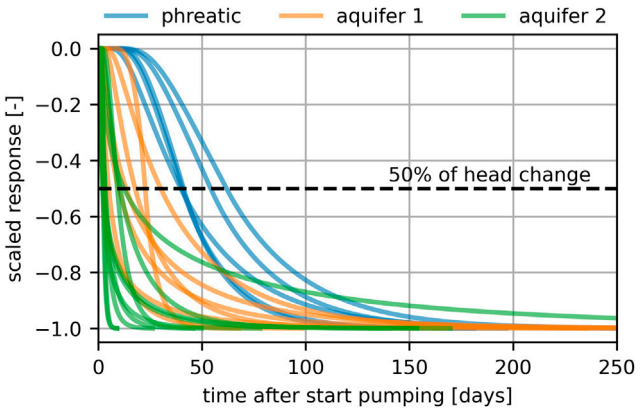

FIGURE 7 | Scaled step responses for well field Hemmen, categorized per aquifer.

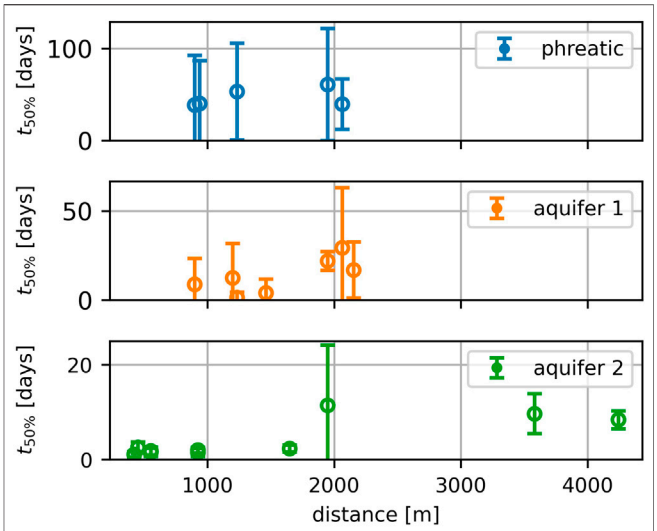

FIGURE 8 | Estimated 50% response time for observations wells near well field Hemmen including estimated uncertainty (2σ).

The drawdown estimates near Fikkersdries show no consistent and plausible pattern. This may be explained through the historic development of pumping at this well field, which has been active since the early 1960s. In the period 2000–2021, the yearly pumping volumes have been relatively constant, varying between 10.5 and 13.6 Mm^3/year. As such, the time series has a low coefficient of variation (CV, see **Table 1**) over the calibration period (CV = 0.11). In contrast, Hemmen (CV = 0.61) and Zetten (CV = 0.53) started pumping in 2006, and Sijmons (CV = 0.28) has seen a reduction in yearly pumping rate from around 5.5 Mm^3/year to about 3 Mm^3/year towards the end of the calibration period. The small variation in the pumping discharge at Fikkersdries makes it difficult to estimate the effect of pumping in observation wells.

5 EXAMPLE APPLICATION: TIMING AND MAGNITUDE OF DRAWDOWN AT HEMMEN

The time series analysis revealed a significant effect of pumping on the heads in 22 observation wells near the well field of Hemmen. Here, the effect of pumping on heads is compared per aquifer. Additionally, it is investigated whether low summer groundwater tables may be mitigated by reducing the pumping rate. Before implementing a potential mitigation measure, a decision-maker would need to determine the effectiveness of such a measure. Information on the magnitude and timing of the head response to reduced pumping is required for this purpose. This information is contained in the response functions associated with the pumping stresses.

The estimated step responses are plotted for all models that include pumping at Hemmen as a stress in **Figure 7**. The step responses are scaled with the gain for comparison purposes. The responses in aquifer 2 are the fastest because the well fields pump from the second aquifer. The responses in the phreatic observation wells are the slowest and show a 15–30-days lag after the start of pumping. Using these results, specifically the magnitude and the timing of the estimated head response to pumping, a decision-maker can determine the effectiveness of reduced pumping as a mitigation measure and what type of pumping strategy is potentially feasible.

The response time, represented by the t_{50}, is plotted versus the distance from well field Hemmen in **Figure 8** for each aquifer. The t_{50} is a linear function of the distance from the well field (**Eq. 8**) for constant values of the fitting parameters a and b in the response function (**Figure 1**). The response times in the second aquifer are the shortest, ranging from about 1 to 11 days. The model at about 2000 m from Hemmen shows a larger uncertainty than other points, casting some doubt on the estimation of the response for this model. In the first aquifer the t_{50} varies between 1 and 30 days. For the five observation wells in the phreatic layer, the t_{50} is between 30 and 60 days. This means that 50% of the maximum reduction in drawdown as a result of a change in pumping rate takes 30–60 days to manifest itself in the phreatic layer. The estimated uncertainties are larger in more shallow aquifers. The general trend is that the independently estimated t_{50} response times increase with the distance, as would be expected, though individual models do show significant variation.

The timing of the response of the phreatic layer to pumping at Hemmen means that reduction in pumping informed by weather forecasts (typically available for a 14-day period) to mitigate low groundwater tables in the summer would not be an effective mitigation measure at Hemmen. As an alternative, a systematic reduction in summer pumping is considered.

Two hypothetical pumping regimes are compared to investigate the effect of reduced pumping during the summer months on the heads. The first regime is a constant pumping rate of 6 Mm^3/year. The second regime pumps 7 Mm^3/year for 9 months per year and a reduced rate of 3 Mm^3/year for three months. The total yearly production volume is equal in both

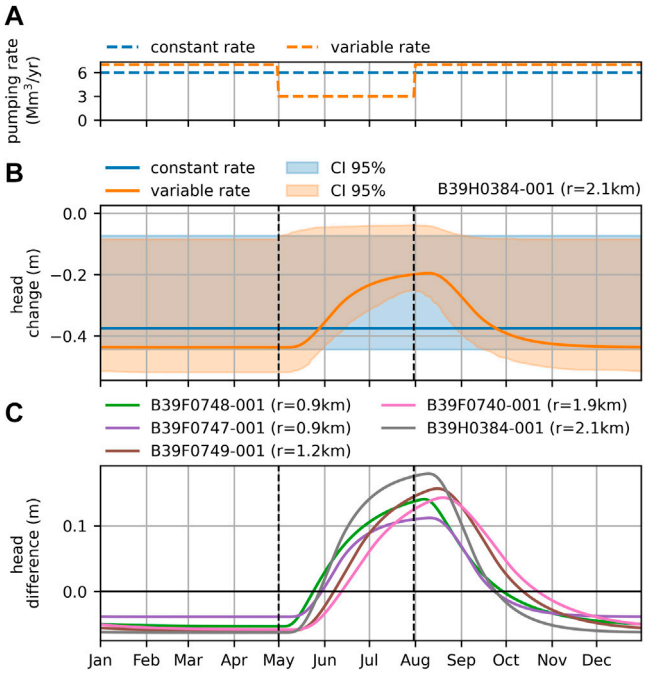

FIGURE 9 | Comparison of the effect of two pumping regimes **(A)** on the calculated drawdowns for well B39H0384 **(B)** and the differences between calculated drawdowns corresponding to different pumping regimes for all 5 phreatic models **(C)**. The drawdown is calculated using the derived response to pumping at well field Hemmen. The shaded areas represent the 95% confidence intervals.

scenarios, and equal to the current permit (6 Mm^3/year). The pumping is reduced in the months May, June, and July such that the drawdown is minimized in June, July, and August, traditionally the driest months in the Netherlands. This variable pumping regime requires additional pumping at, e.g., well field Zetten to compensate for the reduction in drinking water production at Hemmen during the dry summer months. Whether this compensation can be realized in practice is outside the scope of this research.

Figure 9 shows the drawdown calculated with the response function for well field Hemmen for observation well B39H0384-001 (see highlighted point in **Figure 3**) for both pumping regimes, including a 95% confidence interval based on estimated parameter uncertainties. The drawdown at a constant pumping rate of 6 Mm^3/year is 37 cm (with a relatively large 95% confidence interval of 8–45 cm). By reducing the pumping rate from May to July, the drawdown can be reduced in the summer. The maximum effect occurs in August, when the drawdown is reduced to 20 cm (with a 95% confidence interval of 5–27 cm), a reduction of 17 cm as compared to the constant pumping scenario. There is a 5 cm larger drawdown outside the summer months as a result of the increased pumping rate in those periods. The reduction in drawdown caused by the variable pumping regime for all five observation wells is shown in the bottom graph of **Figure 9**. The effect of the variable pumping regime is similar in all observation wells. The largest effect occurs in August and the maximum drawdown reduction varies between 10 and 17 cm.

6 DISCUSSION

The objective of this paper is to develop and apply a data-driven, reproducible, and robust approach to estimate drawdowns as a result of pumping at multiple well fields. The proposed method is based on time series models that are derived exclusively from commonly observed data. The models can be constructed in a limited amount of time with low input data requirements. The described approach is implemented in Python scripts using the open-source software Pastas (Collenteur et al., 2019) to ensure full transparency and reproducibility.

The challenge of any modeling study is that a number of more-or-less subjective modeling decisions must be made. In the following, five major challenges are discussed:

1) Selection of the time periods used for model calibration and validation. A validation period is potentially valuable to test model performance, but the downside is that there is less data available for calibration. Shen et al. (2022) even propose skipping model validation entirely, based on a study of river discharge data in the United States. In the current study, the validation period (2015–2021) contains the driest years on record while at the same time the pumping stations of Hemmen and Zetten show a distinct increase in pumping rates (see **Figure 4**). This dry period may contain information of the head response that is not present in the calibration data. Exclusion of this period from the calibration period means the models might not be able to simulate these periods accurately.

On the other hand, if models perform well despite this choice, this is a strong indication that the models are performing well for the right reasons. The mean R^2 for the selected models in the calibration period is 0.82 and 0.66 in the validation period. Good performance in the validation periods suggests the method and the models are robust. A robust method should yield the same model structure while a robust model should not produce significantly different estimates for the drawdown for an extension of the calibration period.

2) Time interval between head observations. A related challenge was the selection of the time interval between head observations used for model calibration. Higher frequency observations (i.e., daily) mean faster processes are captured by the data, potentially providing additional information to quantify the effects of the different stresses (Kavetski et al., 2011). However, the use of high frequency data increases the autocorrelation in the model residuals, troubling the estimation of parameter uncertainties. Reliable estimates of the parameter standard errors are important in this study, because they are used in the reliability criteria. Theoretically, autocorrelation can be reduced by improving the input data, improving the deterministic model, and/or improving the noise model. There is probably a point, however, where more head observations introduce more problems (e.g., autocorrelation) than they solve (e.g., better models). Different frequencies (daily, weekly, bi-weekly) of head observations were evaluated to calibrate the models (following Collenteur et al., 2021). A time interval of 14 days yielded good models while greatly reducing the number of models with significant autocorrelation in the noise. As a result, 1558 out of 2000 models (78%) passed the autocorrelation check, showcasing the effectiveness of the AR(1) noise model for data at a 14-days interval. The robustness of the method and the models to a different sample of head observations was tested by shifting the sample by 7 days and comparing the results visually. This led to somewhat different drawdown estimates for some models, but most models showed no significant changes in the estimated drawdowns. This analysis can be repeated 14 times to give additional insight into the robustness of the method to the selected sample of head observations (as was done by Collenteur et al., 2021).

3) Goodness-of-fit criterion. One of the four reliability criteria is the goodness-of-fit criterion that $R^2 \geq 0.7$ (**Figure 2**). This is obviously a subjective criterion. It basically means that the model must fit the data reasonably well. It is unclear whether this requirement is really necessary. Other studies, for example Zaadnoordijk et al. (2019), opted for a much lower R^2 cutoffs of 0.1–0.3. When the R^2 criterion is dropped entirely, the number of observations wells where at least one model structure passes the three remaining reliability criteria, increases from 50 to 85%. The underlying question is whether the drawdown of the well fields can be estimated with sufficient accuracy even though the model fits the data poorly. Further research is needed to determine whether, and under what conditions, a goodness-of-fit criterion is needed to estimate the drawdown accurately.

4) Selection of the best model structure. The minimum AIC was used to select the best model structure from all reliable model structures for an observed head time series. In some cases, the differences in AIC values for different model structures were smaller than 2, meaning that multiple model structures are potentially supported by the data (Burnham et al., 2011). This introduces an uncertainty to the model selection step that was not taken into account in this study. Other methods of selecting the best model structure were considered, such as the model goodness-of-fit in the validation period, but occasional dubious observation data in the validation period, or minute differences in model performance, sometimes resulted in the selection of the model with the most parameters, while this did not seem to be warranted.

5) Visual inspection. Visual inspection of model performance in the validation period was deemed necessary as a last step in the selection process. A visual inspection remains valuable to identify models that show odd results, even though they pass all criteria, but a visual inspection is subjective as it is based on the expertise of the modeler. It is desirable to eliminate this subjective step, but this requires additional reliability criteria or fine-tuning of current criteria to local conditions. In each of the three cases where a model was rejected based on visual inspection, the t_{95} of the response to river stage exceeded several years, but was not long enough to be rejected by the model reliability criteria. The model used the long response time to simulate a long-term trend in the data. For this specific study site, an additional reliability criterion can be added to limit the response time of changes in river stage, as such eliminating the necessity for visual inspection. This was not done, however, as such a criterion is applicable only for this local situation, while the four reliability criteria presented in **Figure 2** are broadly applicable.

In groundwater hydrology, drawdowns of well fields are commonly estimated with physically-based groundwater models that are calibrated against head observations by adjusting the (spatial distribution) of the aquifer parameters. As a result, the estimated drawdowns make hydrological sense, as they are derived from basic principles such as continuity of flow and Darcy's law. The approach presented in this paper is fully data-driven. The drawdown is estimated independently at each observation well using physically-based response functions. Each model had to pass four reliability checks and the uncertainty of the modeled drawdowns was estimated and plotted (e.g., **Figure 6**). The resulting spatial pattern of drawdowns makes hydrological sense, with drawdowns decreasing within a limited distance from a pumping station. Collenteur et al. (2019) applied a similar approach for single well fields and showed for an example application in the Netherlands that the drawdowns estimated with time series analysis compared well with the results of an analytic element model.

Application of time series analysis has been shown to be a viable method to estimate drawdowns, for example, through the use of synthetic data generated with a MODFLOW model (Shapoori et al., 2015). The same study also showed, however, that drawdown estimates can be biased if important processes (e.g., groundwater evaporation) are not taken into account in the

model. In the current paper, eight different model structures were tested for each observation well. Different models to compute groundwater recharge (e.g., nonlinear approaches, Peterson and Western, 2014) were not considered, because a linear precipitation-excess model is commonly adequate to simulate head time series with 14-days time intervals in the Netherlands. A final uncertainty in the estimated drawdowns is the possible impact of unknown stresses. For example, pumping for irrigation probably occurs in the area during dry summers. Irrigation wells are largely unmetered, however, so that they can not be included in the model. This may affect the estimated drawdowns, but this is not any different from the application of a physically-based groundwater model.

A reproducible and transparent workflow was presented to develop reliable time series models. Application of this workflow at the study site resulted in reliable models for approximately 50% of the observed head time series. As mentioned above, the goodness-of-fit criterion was responsible for most of the rejections. It is the experience of the authors that a 50% success rate is pretty common in the modeling of transient flow. For example, Zaadnoordijk et al. (2019) obtained 47% decent or good time series models, according to their criteria. There are numerous reasons that can lead to a poor fit varying from data errors and missing stresses to inadequate model structures or local phenomena that affect the head variations. Additional research is needed to increase the percentage of successful models.

7 CONCLUSION

A reproducible, data-driven approach using open-source software is proposed to quantify the effects of different hydrological stresses on heads. A new method was developed to estimate the drawdown caused by pumping at multiple well fields. The data and the code for (re)producing the results presented in this study are available from a dedicated Zenodo repository (Brakenhoff et al., 2022). The method is able to derive reliable models for 50% (126) of the 250 considered head time series and quantify the effect of one or multiple well fields for 78 head time series.

The relative simplicity of the time series models allows the modeler to test multiple model structures (such as stresses and response functions) and model settings (such as the time interval between observations, and the calibration and validation periods) in a short amount of time. This quickly yields valuable insights into the driving hydrological processes affecting the head variations. The data-driven nature of the approach avoids the many approximations that have to be made when analyzing a similar problem using more traditional modeling techniques (e.g., numerical groundwater models). Each time series model is valid only at the specific location of the observation well where the heads are measured. Results at multiple observation wells show clear spatial patterns: drawdowns are larger in the pumped aquifers and decrease with distance from the well field while response times increase with distance from the well field, even though the data at each observation well is analyzed independent from the other observation wells.

The example application at well field Hemmen shows how time series models can be used to estimate the effects of the well field, both spatially and over time. Reduced pumping in May-July can reduce drawdown by about 10–20 cm in the summer months June-August. This is valuable information for decision-makers weighing potential strategies for mitigating low groundwater tables in dry periods.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in an online repository. The name of the repository and accession number can be found at: https://zenodo.org/record/6372578#.YkLTyy8Rr0o.

AUTHOR CONTRIBUTIONS

DB is the lead author, and performed the analysis presented in this paper together with MV. RC and MB contributed significantly with helpful advice throughout, and helped in writing the paper. MvB was involved in reviewing the application of simulating multiple wells with a single response function during the development stage.

FUNDING

The work from Raoul Collenteur was funded by the Austrian Science Fund (FWF) under Research Grant W1256 (Doctoral Program Climate Change: Uncertainties, Thresholds and Coping Strategies).

ACKNOWLEDGMENTS

The authors would like to acknowledge the drinking water supply company Vitens, who allowed us to publish this work using their data. Specifically we want to acknowledge Jelle van Sijl who initiated the project that eventually led to this paper, and was helpful throughout, by providing data, advice, and valuable insights into the complexities of drinking water production. The authors would also like to acknowledge Jeroen Castelijns from Brabant Water, for initiating the project in which the method for simulating multiple well fields with a single response function was developed.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.907609/full#supplementary-material

REFERENCES

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Contr.* 19, 716–723. doi:10.1109/TAC.1974.1100705

Anderson, M. P., Woessner, W. W., and Hunt, R. J. (2015). *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*. San Diego: Academic Press.

Bakker, M., and Schaars, F. (2019). Solving Groundwater Flow Problems with Time Series Analysis: You May Not Even Need Another Model. *Groundwater* 57, 826–833. doi:10.1111/gwat.12927

Brakenhoff, D., Vonk, M., van Baar, M., Collenteur, R., and Bakker, M. (2022). Supplementary materials to "Brakenhoff et al., Application of time series analysis to estimate drawdowns from multiple well fields. [Dataset]. doi:10.5281/zenodo.6372578

Brakkee, E., van Huijgevoort, M. H. J., and Bartholomeus, R. P. (2022). Improved Understanding of Regional Groundwater Drought Development through Time Series Modelling: The 2018-2019 Drought in the Netherlands. *Hydrol. Earth Syst. Sci.* 26, 551–569. doi:10.5194/hess-26-551-2022

Burnham, K. P., Anderson, D. R., and Huyvaert, K. P. (2011). AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons. *Behav. Ecol. Sociobiol.* 65, 23–35. doi:10.1007/s00265-010-1029-6

Collenteur, R. A., Bakker, M., Caljé, R., Klop, S. A., and Schaars, F. (2019). Pastas: Open Source Software for the Analysis of Groundwater Time Series. *Groundwater* 57, 877–885. doi:10.1111/gwat.12925

Collenteur, R. A., Bakker, M., Klammler, G., and Birk, S. (2021). Estimation of Groundwater Recharge from Groundwater Levels Using Nonlinear Transfer Function Noise Models and Comparison to Lysimeter Data. *Hydrol. Earth Syst. Sci.* 25, 2931–2949. doi:10.5194/hess-25-2931-2021

de Bruin, H. a. R., and Lablans, W. N. (1998). Reference Crop Evapotranspiration Determined with a Modified Makkink Equation. *Hydrol. Process.* 12, 1053–1062. doi:10.1002/(SICI)1099-1085(19980615)12:7<1053::AID-HYP639>3.0.CO;2-E

Fienen, M. N., and Bakker, M. (2016). HESS Opinions: Repeatable Research: What Hydrologists Can Learn from the Duke Cancer Research Scandal. *Hydrol. Earth Syst. Sci.* 20, 3739–3743. doi:10.5194/hess-20-3739-2016

Hantush, M. S., and Jacob, C. E. (1955). Non-steady Radial Flow in an Infinite Leaky Aquifer. *Trans. AGU* 36, 95–100. doi:10.1029/TR036i001p00095

Hipel, K. W., and McLeod, A. I. (1994). "Chapter 7 Diagnostic Checking," in *Time Series Modelling of Water Resources and Environmental Systems* (Amsterdam: Elsevier), 45, 235–253. Developments in Water Science. doi:10.1016/S0167-5648(08)70665-8

Hugman, R., and Doherty, J. (2022). Complex or Simple-Does a Model Have to Be One or the Other? *Front. Earth Sci.* 10, 867379. doi:10.3389/feart.2022.867379

Kavetski, D., Fenicia, F., and Clark, M. P. (2011). Impact of Temporal Data Resolution on Parameter Inference and Model Identification in Conceptual Hydrological Modeling: Insights from an Experimental Catchment. *Water Resour. Res.* 47, W05501. doi:10.1029/2010WR009525

KNMI (2022). Dagwaarden Van Weerstations. [Dataset]. Available at: https://www.daggegevens.knmi.nl/klimatologie/daggegevens. (Febuary 22, 2022).

Mackay, J. D., Jackson, C. R., and Wang, L. (2014). A Lumped Conceptual Model to Simulate Groundwater Level Time-Series. *Environ. Model. Softw.* 61, 229–245. doi:10.1016/j.envsoft.2014.06.003

Obergfell, C., Bakker, M., Zaadnoordijk, W. J., and Maas, K. (2013). Deriving Hydrogeological Parameters through Time Series Analysis of Groundwater Head Fluctuations Around Well Fields. *Hydrogeol. J.* 21, 987–999. doi:10.1007/s10040-013-0973-4

Patle, G. T., Singh, D. K., Sarangi, A., Rai, A., Khanna, M., and Sahoo, R. N. (2015). Time Series Analysis of Groundwater Levels and Projection of Future Trend. *J. Geol. Soc. India* 85, 232–242. doi:10.1007/s12594-015-0209-4

Peterson, T. J., and Western, A. W. (2014). Nonlinear Time-Series Modeling of Unconfined Groundwater Head. *Water Resour. Res.* 50, 8330–8355. doi:10.1002/2013WR014800

Rijkswaterstaat (2022). Waterhoogten (Expert). [Dataset]. Available at: https://waterinfo.rws.nl/#!/kaart/Waterhoogten/. (Febuary 22, 2022).

Shapoori, V., Peterson, T. J., Western, A. W., and Costelloe, J. F. (2015). Decomposing Groundwater Head Variations into Meteorological and Pumping Components: A Synthetic Study. *Hydrogeol. J.* 23, 1431–1448. doi:10.1007/s10040-015-1269-7

Shen, H., Tolson, B. A., and Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resour. Res.* 58, e2021WR031523. doi:10.1029/2021WR031523

Van Loon, A. F., Stahl, K., Di Baldassarre, G., Clark, J., Rangecroft, S., Wanders, N., et al. (2016). Drought in a Human-Modified World: Reframing Drought Definitions, Understanding, and Analysis Approaches. *Hydrol. Earth Syst. Sci.* 20, 3631–3650. doi:10.5194/hess-20-3631-2016

Veling, E. J. M., and Maas, C. (2010). Hantush Well Function Revisited. *J. Hydrol.* 393, 381–388. doi:10.1016/j.jhydrol.2010.08.033

Von Asmuth, J. R., and Bierkens, M. F. P. (2005). Modeling Irregularly Spaced Residual Series as a Continuous Stochastic Process. *Water Resour. Res.* 41, W12404. doi:10.1029/2004WR003726

Von Asmuth, J. R., Bierkens, M. F. P., and Maas, K. (2002). Transfer Function-Noise Modeling in Continuous Time Using Predefined Impulse Response Functions. *Water Resour. Res.* 38, 23–31. doi:10.1029/2001WR001136

Von Asmuth, J. R., Maas, K., Bakker, M., and Petersen, J. (2008). Modeling Time Series of Ground Water Head Fluctuations Subjected to Multiple Stresses. *Groundwater* 46, 30–40. doi:10.1111/j.1745-6584.2007.00382.x

Von Asmuth, J. R., Maas, K., Knotters, M., Bierkens, M. F. P., Bakker, M., Olsthoorn, T. N., et al. (2012). Software for Hydrogeologic Time Series Analysis, Interfacing Data with Physical Insight. *Environ. Model. Softw.* 38, 178–190. doi:10.1016/j.envsoft.2012.06.003

Wald, A., and Wolfowitz, J. (1940). On a Test whether Two Samples Are from the Same Population. *Ann. Math. Stat.* 11, 147–162. doi:10.1214/aoms/1177731909

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020). Toward Reproducible Environmental Modeling for Decision Support: A Worked Example. *Front. Earth Sci.* 8, 50. doi:10.3389/feart.2020.00050

Wunsch, A., Liesch, T., and Broda, S. (2018). Forecasting Groundwater Levels Using Nonlinear Autoregressive Networks with Exogenous Input (NARX). *J. Hydrol.* 567, 743–758. doi:10.1016/j.jhydrol.2018.01.045

Zaadnoordijk, W. J., Bus, S. A. R., Lourens, A., and Berendrecht, W. L. (2019). Automated Time Series Modeling for Piezometers in the National Database of the Netherlands. *Groundwater* 57, 834–843. doi:10.1111/gwat.12819

Conflict of Interest: DB, MV, and MvB were employed by the company Artesia B.V.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Earth Science

TECHNOLOGY AND CODE
published: 05 July 2022
doi: 10.3389/feart.2022.907533

Rapid Model Development for GSFLOW With *Python* and pyGSFLOW

Joshua D. Larsen[1]*, *Ayman H. Alzraiee*[1], *Donald Martin*[2] and *Richard G. Niswonger*[3]

[1]U.S. Geological Survey, California Water Science Center, Sacramento, CA, United States, [2]U.S. Geological Survey, California Water Science Center, San Diego, CA, United States, [3]U.S. Geological Survey, Integrated Modeling and Prediction Division, Water Mission Area, Menlo Park, CA, United States

Following the advancement of high-performance computing and sensor technology and the increased availability of larger climate and land-use data sets, hydrologic models have become more sophisticated. Instead of simple boundary conditions, these data sets are incorporated with the aim of providing more accurate insights into hydrologic processes. Integrated surface-water and groundwater models are developed to represent the most important processes that affect the distribution of water in hydrologic systems. GSFLOW is an integrated hydrologic modeling software that couples surface-water processes from PRMS and groundwater processes from MODFLOW and simulates feedbacks between both components of the hydrologic system. Development of GSFLOW models has previously required multiple tools to separately create surface-water and groundwater input files. The use of these multiple tools, custom workflows, and manual processing complicates reproducibility and confidence in model results. Based on a need for rapid, reproduceable, and robust methods, we present two example problems that showcase the latest updates to pyGSFLOW. The software package, pyGSFLOW, is an end-to-end data processing tool made from open-source *Python* libraries that enables the user to edit, write input files, run models, and postprocess model output. The first example showcases pyGSFLOW's capabilities by developing a streamflow network in the Russian River watershed with an area of 3,850 km^2 located on the coast of northern California. A second example examines the effects of model discretization on hydrologic prediction for the Sagehen Creek watershed with an area of 28 km^2, near Lake Tahoe, California, in the northern Sierra Nevada.

Keywords: groundwater, surface-water, integrated hydrologic modeling, GSFLOW, PRMS, MODFLOW, python

OPEN ACCESS

Edited by:
Jeremy White,
Intera, Inc., United States

Reviewed by:
Matteo Camporese,
University of Padua, Italy
Daniel Partington,
Flinders University, Australia
Jason Bellino,
United States Geological Survey,
United States

***Correspondence:**
Joshua D. Larsen
jlarsen@usgs.gov

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *29 March 2022*
Accepted: *23 May 2022*
Published: *05 July 2022*

Citation:
Larsen JD, Alzraiee AH, Martin D and Niswonger RG (2022) Rapid Model Development for GSFLOW With Python and pyGSFLOW. Front. Earth Sci. 10:907533. doi: 10.3389/feart.2022.907533

INTRODUCTION

Water resources are dynamic and managing them, given competing demands as supplies of both surface water and groundwater change rapidly, is difficult. The challenges are compounded as human reliance on groundwater grows faster than the ability to monitor groundwater supplies (Konikow and Kendy, 2005; Wada et al., 2010). With this background as context, integrated hydrologic models such as GSFLOW (Markstrom et al., 2008) can be used to evaluate management strategies. Integrated models that can simulate surface water and groundwater have potential to improve decision making. However, their benefits are often not fully realized, due in part to errors in model predictions caused by data limitations and incomplete process understanding (Beven, 2019; Blöschl et al., 2019).

Model calibration and application requires hypothesis testing to better represent important processes impacting water storage and flow in hydrologic systems (Clark et al., 2011). Hypothesis

testing necessitates rapid processing and construction of input data for multiple models to integrate soft knowledge, test multiple parameter sets, and different conceptualizations; without automation, developing such hydrologic models for river basins is onerous. Automated data processing tools can improve the value of hydrologic models because they reduce the occurrence of data input errors, improve reproducibility, and reduce model construction time and effort (Gardner et al., 2018; Ng et al., 2018).

Techniques for developing and applying coupled surface-water and groundwater models to represent conditions within hydrologic systems have developed along with the design and complexity of hydrologic simulators. In recent decades, hydrologic models have been developed over much larger regions, including river basins (Schoups et al., 2005; Werner et al., 2006; Huntington and Niswonger, 2012; Kitlasten et al., 2021) and continental scales (Wood et al., 1997; Condon and Maxwell, 2015; Regan et al., 2019; Shin et al., 2019). Regional to continental scale models require processing of massive geographic data sets to provide realistic representation of distributed drainage networks. Furthermore, simulating surface-water and groundwater interactions require hydraulic gradients which are sensitive to the relative positioning of streams and topographic features like river canyons, flood plains, and riparian forests (Gardner et al., 2018; Leaf et al., 2021).

Models constructed at river basin and larger scales require sampling digital elevation models (DEMs) at the model grid scale to represent topography. Surface-water networks should be consistent with the both the DEM used to represent the model surface boundary and the model grid scale. Stream networks generated from fine-scale DEMs and overlayed onto coarse DEMs can cause streams to become incongruent, for example streams that are offset from river canyons or unaligned with watershed boundaries. These types of scale mismatches can lead to erroneous surface-water and groundwater exchanges (e.g., reversed hydrologic gradients between surface-water and groundwater systems) and numerical problems (Kampf and Burges, 2007; Schoups et al., 2010; Gardner et al., 2018). Consequently, stream networks often need to be developed consistent with the model surface discretization. Existing software used to build input data sets for large scale hydrologic models have most often relied on stream hydrography data sets built on DEMs at their native scale (e.g., 30 m or finer; Ng et al., 2018; Leaf et al., 2021).

In contrast to using established national stream networks like NHDPlus (Buto and Anderson, 2020), topographic analysis can be used to develop stream networks using DEMs at any resolution. These analyses rely on two geographic data sets used exclusively to develop drainage networks: flow direction and contributing area, also called flow accumulation (O'Callaghan and Mark, 1984; Jenson and Domingue, 1988; Mark, 1988; Tarboton, 1997). Flow direction methods typically choose 1 of 8 possible outflow directions for each grid cell, including directions perpendicular and diagonal to cell faces (D8; O'Callaghan and Mark, 1984) or by using the direction of greatest slope from triangular facets at the center of each grid cell (D-infinity; Tarboton, 1997). Methods that calculate multiple outflow directions for each grid cell and variable flow partitioning also have been applied to address dispersion (Qin et al., 2007). Flow direction methods are then combined with an algorithm for calculating upslope contributing area (Mark, 1988). These approaches form the basis for the popular Arc-Hydro toolset (Maidment and Morehouse, 2002).

Topographic analysis methods used to develop stream networks from DEMs have their own limitations. Digital artifacts in DEMs related to spatial averaging can misassign a flow network and associated model grid cell altitudes. Increasing DEM resolution can reduce these artifacts (Goodchild, 2011); however, finer-scale data are not always available for a geographic area, and computational costs can require coarse spatial discretization for hydrologic models applied to large regions. In low relief basins, spatial averaging can lead to digitally closed sinks, create uncertainty in flow direction assignment, or even lead to digital flow directions that contradict the natural system. Hydrologic conditioning methods such as digitally filling (Jenson and Domingue, 1988; Garbrecht and Martz, 1997; Metz et al., 2011) and outlet breaching (Martz and Garbrecht, 1999) can be applied to the DEM or flow-direction system to remove sink artifacts from flow accumulation. Uncertainty in flow direction assignment generally occurs in flat areas of DEMs associated with low relief watersheds and create maze-like conditions where flow directions cannot be determined without additional information. Previous approaches to solving digitally flat areas have included iterative methods to connect uncertain flow directions (Jenson and Domingue, 1988), incrementally raising flat portions of a DEM to create a gradient from higher terrain to lower terrain (Garbrecht and Martz, 1997), cost function solutions (Metz et al., 2011), and weighted topological methods (Zhang et al., 2017). Given these issues and solutions, stream networks produced for hydrologic modeling are a simplification of the actual network limited primarily by DEM scale and model scale resolution.

Previous approaches for developing coupled surface-water/groundwater systems that can be simulated with GSFLOW (Markstrom et al., 2008) have relied on multiple software tools. GSFLOW is an integrated hydrologic modeling software that couples surface-water processes from the Precipitation Runoff Modeling System (PRMS; Markstrom et al., 2015) and groundwater processes from MODFLOW (Harbaugh, 2005; Niswonger et al., 2011) and is often applied to build models at scales of 10s to 1,000s of km^2. GSFlow-ArcPy provides functionality to transform raster data sets, such as a DEM or land use/land cover data sets, into PRMS input files (Gardner et al., 2018). GSFLOW-GRASS allows users to build GSFLOW input files from raster and vector data through a command line scripting process within a GRASS GIS environment (Ng et al., 2018). SFRBuilder (Leaf et al., 2021) overlays vector data of streamlines provided by NHDPlus or another custom hydrography dataset onto a DEM to develop the Streamflow Routing Package (SFR) in MODFLOW. PRMS-Python allows the user to load, modify, and run simulation scenarios with existing PRMS input data sets (Volk and Turner, 2019). Finally, the FloPy *Python* package allows users to build and modify most MODFLOW package files that represent the groundwater system in GSFLOW (Bakker et al., 2016; Bakker et al., 2022).

TABLE 1 | Overview of the pyGSFLOW process types and the required input data to create a simple GSFLOW model.

Process	Required input(s)
model grid creation	Raster, shapefile, or extent of model grid
	Cell dimensions in *x* and *y* direction
raster resampling	model grid and input raster dataset
flow directions	model grid
	resampled Digital Elevation Model
flow accumulation	flow direction array
watershed delineation	flow direction array
	watershed pour point location (xy coordinate or row column location)
	model grid
subbasin delineation	flow direction array
	watershed boundary
	subbasin pour point locations (xy coordinates or row column location)
	model grid
stream network generation	watershed boundary
	flow direction array
	flow accumulation array
	number of contributing grid cells for determining streamflow
Cascade routing	flow direction array
	stream network information
Model inputs (MODFLOW)	model grid. resampled Digital Elevation Model. user supplied model name. stream network information
	optional inputs. model bottom elevations. UZF infiltration array watershed boundary
Model inputs (PRMS)	stream network information
	cascade routing information
	model grid
	Digital Elevation Model
	watershed boundary
	Climate information
Model inputs (GSFLOW)	model start time
	model end time
	MODFLOW time zero
	Climate module and data information
Editing GSFLOW parameters	GSFLOW input files
	Ancillary data

Although these tools can be used to complement each other and create most of the GSFLOW input, proprietary software is sometimes required and manual edits to input files and additional scripts are needed to build, edit, run models, and process output data. These requirements create a disconnected process which requires multiple tools and/or scripts to develop input files and estimate parameters for the surface-water and groundwater systems and hinders reproducibility.

Based on a need for a rapid, reproducible, and robust GSFLOW model building pipeline, we seek to address the needs of constructing complex stream networks from large geographic data sets by presenting the latest version of pyGSFLOW (*pronounced "pie-g-s-flow"*). The pyGSFLOW *Python* scripting library allows users to develop input files and set boundary conditions, climate forcing, and hydrologic parameters, edit existing GSFLOW input files, and postprocess model results (Larsen et al., 2021; 2022). Instead of creating an entirely new approach, pyGSFLOW improves upon previous conceptual frameworks and leverages existing tools to create a framework for constructing GSFLOW models (Henson et al., 2013; Gardner et al., 2018; Larsen et al., 2021; Bakker et al., 2022; Larsen et al., 2022). The pyGSFLOW package leverages FloPy (Bakker et al., 2022) for interfacing with MODFLOW model files and integrates this functionality with custom features for interfacing with PRMS and GSFLOW model files. This work extends the existing pyGSFLOW software (Larsen et al., 2021; 2022) with new open-source methods to construct flow direction, flow accumulation, watershed and subbasin delineation, and stream network data sets. These advancements add robustness to the model development process.

METHODS

The approach presented here improves upon the existing pyGSFLOW package (Larsen et al., 2021; 2022) to include model-building tools for rapidly generating GSFLOW (Markstrom et al., 2008) models from external raster data. The pyGSFLOW model building tools are part of an open-source python toolkit and can ingest and process ancillary data sets to produce intermediate and primary data required by GSFLOW. **Table 1** provides a list of data sets that are required to build a simple GSFLOW model; additional data not included in this list can be used to modify the model. This process can be started or stopped at any point to allow modifications to pre-existing models or to update a single part of the workflow. The steps are presented in the following sections and an overview describing the required data for each step is outlined in **Table 1**.

Model Grid Creation and Raster Resampling

Model spatial discretization, referred to herein as a model grid, serves as the foundation for building hydrologic models in GSFLOW. GSFLOW and MODFLOW-NWT model grids are structured cellular grids, composed of rectangular cells that are discretized on the basis of row and column spacing. Model grid discretization choices affect not only the location of boundary conditions but also the numerical solution of a model (Markstrom et al., 2008). The pyGSFLOW package provides support for quickly creating rectilinear model grids from geographic information. A raster, shapefile, or bounding box can be supplied with grid spacing information to produce a model grid. Upon creation of the model grid, raster resampling to the model grid scale is accomplished with a simple Python function from FloPy's "Raster" class (Bakker et al., 2016; Bakker et al., 2022).

The model grid produced by pyGSFLOW is a FloPy "StructuredGrid" (Bakker et al., 2022) object that contains polyline, polygon, and geographic feature information. Resampling the native DEM to the model grid produces an array of raster values consistent in shape and size to the model grid. Model grid land-surface elevations can be calculated for each model grid cell using mean, median, minimum, maximum, or interpolated elevations resampled from the native DEM. The model grid and the resampled DEM are used to develop the flow direction, flow accumulation, and stream network, as described in the next sections.

Flow Direction and Flow Accumulation

Gridded flow-direction calculations determine the direction of flow based on the slopes between a cell and each of its neighboring cells. For 8-direction (D8) flow direction calculations, information for a cell and its 8 adjacent cells are compared (Jenson and Domingue, 1988). Calculations are performed on elevation data resampled from a DEM, and a one-to-one outflow connection is assumed (Jenson and Domingue, 1988). The flow direction is set by calculating the cell index (i_{cell}) using **Eq. 1**.

$$i_{cell} = argmax\left(\frac{\Delta\vec{e}}{\sqrt{\Delta\vec{x}^2 + \Delta\vec{y}^2}}\right) \quad (1)$$

where $\Delta\vec{x}$ and $\Delta\vec{y}$ are the difference between a cell's center coordinates and a vector of neighboring cell's center coordinates, and $\Delta\vec{e}$ is the difference between a cell's elevation and a vector of its neighbors' elevations. Once i_{cell} is known, the flow direction is encoded with a digital number that describes the outflow direction using the convention:

$$\begin{bmatrix} 64 & 128 & 1 \\ 32 & -1 & 2 \\ 16 & 8 & 4 \end{bmatrix}$$

where −1 represents the model cell for which calculations are being applied and the digital numbers 1 through 128 represent the specific direction of neighboring cells relative to the model cell. When i_{cell} is not a unique value, flow direction is undefined based on the slope between cells. The pyGSFLOW package's default method of solving flow direction for undefined cells is a topological method that maps each undefined cell to the nearest outlet and attempts to minimize the absolute distance of the flow direction to the outlet. In cases where the default method does not perform well (e.g., in large, complex, digitally flat areas), a modified version of Dijkstra's (1959) algorithm can be used to solve the flow direction problem. The pyGSFLOW implementation of Dijkstra's (1959) algorithm first creates a connectivity graph of all cells and their potential flow paths within a digitally flat area. The algorithm then solves the digitally flat area from the outlet location and minimizes the routing distance for each cell within the graph by weighting each potential flow direction by the routing distance to the outlet. Although hydrologic conditioning is recommended to fill sinks prior to calculating flow directions, a breaching stage threshold can be applied for cases where small digital artifacts in the DEM data create sinks or produce flow directions that conflict with the hydrologic flow system. The breaching stage threshold is a small user defined value that can be used to smooth out differences in resampled DEM elevation values caused by artifacts. Elevation differences between neighboring cells smaller than the breaching stage threshold are considered as equal in elevation which allows the flow direction to pass over a slightly higher cell. The "FlowAccumulation" object in pyGSFLOW performs the flow direction, as well as flow accumulation, calculations from a model grid object and a DEM (shown in "Sagehen Creek Watershed Example" section).

Flow accumulation calculates the number of upslope cells that drain to each cell within the watershed. The flow direction array defines the connectivity of cells and drainage pattern for the flow accumulation calculation. For the D8 flow direction model, each cell can have flow drain into it from multiple neighbors; however, the flow from each cell only can drain to a single neighbor. Flow accumulation numbers are calculated using a queue where accumulation numbers of downstream cells are increased as each cell is taken off the queue, and the number of input drainage paths for the given cell is decreased by one (Wang et al., 2011). If the number of input drainage paths for a cell equals zero, it is added back to the queue. The algorithm completes when the queue is empty.

Watershed and Subbasin Delineation

Watershed boundary delineation is calculated from flow direction arrays following Jenson and Domingue (1988). In-lieu of automated subbasin delineation, user supplied watershed outlets, called "pour points," are used to define the outlet locations for both the watershed and subbasin delineation calculations. From a single pour point, a topological diagram that includes connection information from the flow direction array is produced, and a watershed is classified from topographic divide information inherent in connection data from the flow direction array. Subbasin delineation is performed in a similar manner as watershed delineation. Multiple pour points are supplied by the user and subbasins, within a watershed, are classified with a unique value. Upstream subbasin boundaries are respected while delineating downstream subbasins. This

TABLE 2 | GSFLOW model input files that are produced with pyGSFLOW's automated model building methods. Note that climate input files are not automatically populated with default values and instead the user must specify their climate representation.

GSFLOW Input Files
Control file (Markstrom et al., 2008)
PRMS input files
Parameter file (Markstrom et al., 2015)
MODFLOW package input files
Discretization package (DIS; Harbaugh, 2005)
Basic package (BAS6; Harbaugh, 2005)
Upstream weighting package (UPW; Niswonger et al., 2011)
Streamflow Routing package (SFR2; Niswonger and Prudic, 2005)
Unsaturated Zone Flow package (UZF; Niswonger et al., 2006)
Output control package (OC; Harbaugh et al., 2000)
Newton solver package (NWT; Niswonger et al., 2011)

approach is preferable to fully automated subbasin delineation because hydrologic areas of interest and contributing areas for gaged flows can be isolated by the user for subsequent parameterization, calibration, and output analysis.

Stream Network Generation

Stream network generation is a critical step in defining model boundary conditions for both PRMS and the Streamflow Routing Package (SFR; Niswonger and Prudic, 2005) components of GSFLOW. Streams in GSFLOW are discretized into reaches and segments, where a reach is a part of stream spanning a grid cell, and a segment is generally defined as a stream spanning two confluences or the start of a stream to a confluence. Stream segments can also be further divided at the user discretion to represent other surface-water features such as diversions. Network generation has three distinct parts: classifying grid cells that contain streams, defining the connectivity between stream cells, and defining the number of cells that drain to each stream cell. Stream cell classification is performed by comparing the contributing number of cells in the flow accumulation array data to a user-specified threshold. If the number of accumulated cells is greater than the user-specified threshold value, the cell will be classified as a stream cell. Stream connectivity and flow direction are determined using information from the flow direction array to ensure that streams are continuously connected and that flow occurs in the downslope direction. Clusters of adjacent stream cells are grouped into segments based on the connectivity of the flow direction array. Stream segments either begin at the most upstream cells based on the flow directions or at locations where more than one stream cell drains into a cell. Stream segments end either at a confluence of stream cells, the watershed outlet, or at the watershed boundary. After grouping, topological sorting (Kahn, 1962) and renumbering is performed on the stream network to ensure upstream flows are calculated before downstream flows and to provide optimal calculation order for GSFLOW. Finally, a graph of landscape flow connectivity is created from the flow direction array and stream cells. Each cell is then assigned to a specific stream segment to which it drains.

Once the stream network has been produced, cell to cell routing information for PRMS, referred to as cascade routing, can be extracted from the existing flow direction arrays and the stream network. Because flow direction calculations rely on a D8 method, the cascade routing calculation allows many cells to contribute flow to a single cell, but an individual cell can only drain to 1 cell.

Model Input Generation

PRMS and MODFLOW packages reliant on the stream network can be generated after the stream connectivity has been determined. A set of default model input parameters are stored in a JavaScript Object Notation (JSON) file that can either be loaded, edited by the user, and supplied to the package generation classes or be automatically applied by the package generation classes. These parameters are then passed to their respective PRMS and MODFLOW packages and by default, a single-layer model is created with a GSFLOW control file, PRMS parameter file, and MODFLOW packages (**Table 2**). After *Python* input objects are generated, the user can write these inputs to file, edit existing model parameters, and/or add additional information and packages to the model. Because integrated hydrologic modeling includes many more processes than DEM information can provide alone, this approach allows the user to define many additional processes—e.g., vegetative cover, soil zone, climate, pumping, general head boundaries, etc.—outside of the automated model builder methods. Groundwater flow processes can be added or edited with FloPy (Bakker et al., 2016; Bakker et al., 2022), and surface-water water processes including simulation modules can be adjusted with built in functionality from pyGSFLOW (Larsen et al., 2021; 2022). Climate information can be applied as daily time-series information from one or multiple climate stations to the *data file* or as arrays in climate by hydrologic response unit (grid cell; HRU) files and adjustment factors can be specified in the parameter file. The pyGSFLOW approach gives the user flexibility on how best to represent the climate of the simulated watershed and provides python tools to aid in the processing and writing of these input files. After input objects are generated, users can write the packages to GSFLOW compatible input files and then run the model.

Editing GSFLOW Model Parameters

After model creation, surface-water and groundwater parameters are commonly added or adjusted in the model calibration process. The pyGSFLOW package allows the user to easily add new parameters, remove unused parameters, and edit existing ones within a python environment. Surface-water parameters, such as land cover, impervious surfaces, and soil physical properties, can be sampled from existing raster data using the raster resampling methods described earlier. Once a raster has been sampled into an array, it can be set directly as a parameter, be scaled or masked, or be used in a mathematical relationship to derive one or multiple parameters. Groundwater parameters can be added and edited using FloPy's built in features (Bakker et al., 2016; Bakker et al., 2022). After model parameters have been adjusted, new input files can be written for subsequent model runs and analysis. More

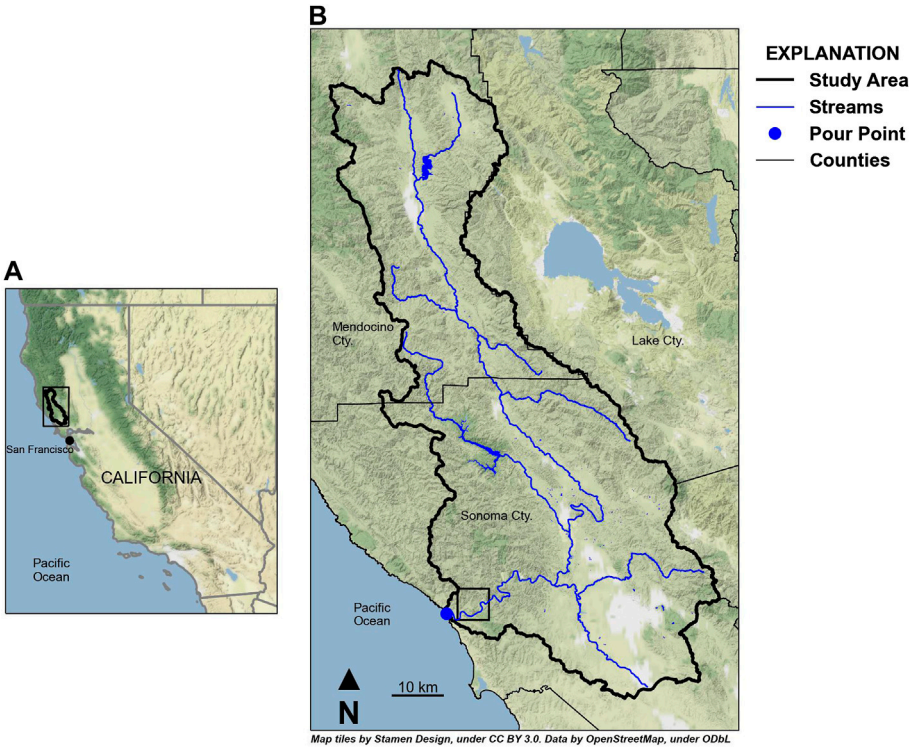

FIGURE 1 | Location of the Russian River watershed study area in northwestern California **(A)**. Main drainages of the Russian River watershed provided by NHDPlus dataset (Buto and Anderson, 2020) within the study area boundary **(B)**. Black box inset in **(B)** defines the boundary for flow direction vectors shown in **Figure 3**.

information and detailed instructions on pyGSFLOW's usage can be found in Larsen et al. (2021).

Example Problems

Two example problems are presented in this section where the first example illustrates the robustness of pyGSFLOW to develop model input for a complex watershed and the second example demonstrates the utility of pyGSFLOW to rapidly produce multiple conceptualizations of a hydrologic model. The first example is of the Russian River watershed. Only a cursory description is provided for the "Russian River Watershed Example"; instead, the example focuses on demonstrating the robustness of pyGSFLOW for automatically developing a stream network and watershed boundary in a complex watershed. The second example is of the undeveloped Sagehen Creek watershed. This example demonstrates the complete process of building and running the model, including descriptions of all the required ancillary data. The "Sagehen Creek Watershed Example" provides detail to convey each step in the GSFLOW model construction process for a simpler watershed that is computationally inexpensive to build for testing and evaluating results across different users and computing environments.

Russian River Watershed Example

The Russian River watershed, in northwestern California, was chosen as an example to illustrate the stream network generation methods (**Figure 1**). Construction of the stream network for the Russian River provides an opportunity to showcase pyGSFLOW's capabilities in a complex system; however, because this system has extensive anthropogenic modifications, it is beyond the scope of this work to present a fully functional model of the Russian River watershed. For this example, only details regarding the grid and stream network generation are provided.

The Russian River's main channel flows from north to south for about 180 km and drains about 3,850 sq. km. of Mendocino and Sonoma Counties to the Pacific Ocean (**Figure 1**). The watershed contains both steep terrain from the northern coastal range and gentle to flat terrain within the valley regions. Large sections of the watershed are digitally flat at 30 m DEM resolution, parts of the watershed near the Russian River have steep canyon walls with low relief river drainage that creates digital artifacts, and large sections of the watershed have significant topographic relief. The process for generating a flow network for the Russian River watershed begins with DEM selection, spatial discretization, and resampling the DEM to the spatial discretization of the model. A 1 arc-second DEM product was selected for the Russian River watershed (U.S. Geological Survey, 2020). Grid cell size for the Russian River watershed was set to 300 m × 300 m (410 rows, 252 columns; 103,320 grid cells) to adequately represent the topography of the watershed without creating a model that is too computationally demanding. GSFLOW is most often applied to regional-scale systems ranging in size from 10 s to 1,000 s of km^2 to answer questions about water resources, and the total number of model

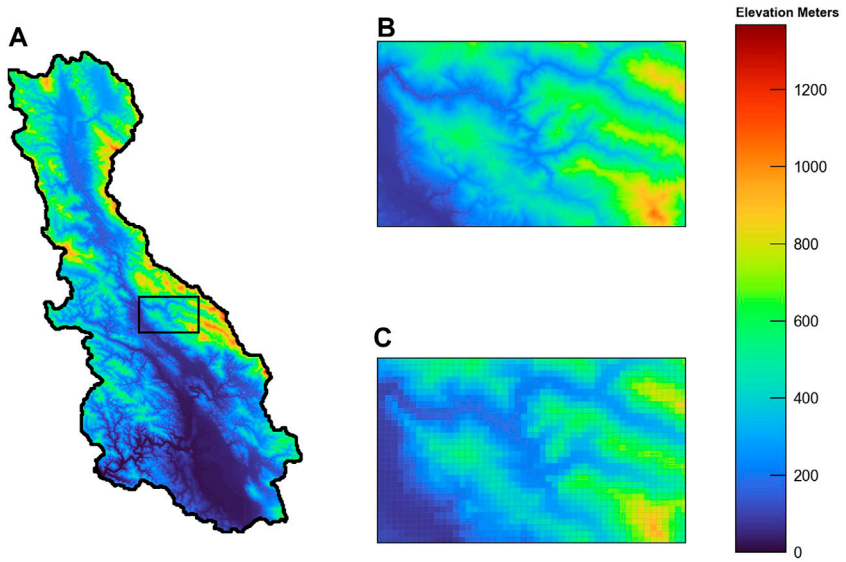

FIGURE 2 | Digital elevation model (27.28 m x 27.28 m; DEM; U.S. Geological Survey, 2020) of the Russian River study area **(A)**. DEM values were resampled by minimum elevation to model grid size (300 m × 300 m) and are shown for a subset of the watershed: the raw DEM data (27.8 m resolution) **(B)** and the resampled DEM values (300 m resolution) **(C)** correspond to the inset box in **(A)**.

cells were determined to meet the computational constraints for the project. The DEM was resampled using the minimum elevation in each model grid cell to produce a model-elevation profile that can be used to construct the GSFLOW model (**Figure 2**).

Preparation of the Russian River DEM for flow direction and flow accumulation processes involved filling large sinks and pits prior to resampling to the grid scale (Wang and Liu, 2006). Groups of cells in the model grid that have a lower elevation than their surrounding cells are referred to as sinks that must be filled to provide continuous pathways across all cells from all ridges in the watershed to the watershed outlet at the Pacific Ocean. However, this sink-filling process did not remove all digital artifacts within the watershed. Near the watershed outlet, artifacts from high-relief canyon walls that border low-relief valley floors persisted. In parts of the DEM, large digitally flat areas were also present. Model grid cell elevations adjacent to the Russian River's outlet, in the Pacific Ocean, were slightly lowered to create a unique condition that Dijkstra's (1959) algorithm could solve. In the case of a non-unique watershed outlet (all cells are the same elevation), Dijkstra's algorithm is unable to automatically identify an outlet cell and will try to choose an outlet cell that minimizes the number of uncertain connections, which can yield unexpected results.

Flow direction and flow accumulation processes were applied on the resampled DEM. Flow direction vectors, created from the flow direction array, show that the solution generally follows existing NHDPlus flowlines (**Figure 3A**). A comparison of the flow direction arrays produced by using pyGSFLOW's Dijkstra algorithm to solve digitally flat areas (**Figure 3A**) and pyGSFLOW's default topographic method shows that 1) the Dijkstra algorithm is well suited to solve this problem and 2) the simple topological method is unable to solve this complex scenario (**Figure 3B**). Furthermore, a small breaching threshold (1.52e-3 m) was applied to resolve the diverging flow directions caused by digital artifacts. Flow accumulation processes were run to calculate the contributing area to each cell. For the Russian River watershed this process creates an array of values that represent the drainage watershed area for each cell. A threshold of 30 grid cells or about 2.7 km^2 of contributing area was applied to define the Russian River watershed stream network (**Figure 4**).

The final steps to prepare the Russian River surface-water network were to define the watershed and subbasin boundaries within the watershed. A single pour point was selected at the outflow location of the Russian River to the Pacific Ocean to define the watershed. The "define_watershed" method in pyGSFLOW (method shown in the "Sagehen Creek Watershed Example") was applied to the flow direction array to binarize the grid into active and inactive model cells (watershed outline shown in **Figure 4**). Subbasin delineation was then run. Pour points were selected based on the locations of streamgages throughout the watershed to isolate contributing areas to the streamgages. These contributing areas were grouped using a unique numerical identifier and are presented in **Figure 4**.

Results from stream generation were compared to the flow accumulation methods available in ArcGIS Pro. Both methods were able to produce representations of the Russian River stream network from the same pour point (**Figure 5**), and in much of the watershed, both methods follow the same path. Some differences are present in low relief and digitally flat parts the watershed. For example, in the southern part of the Russian River watershed, the ArcGIS Pro stream representation has sections follow a completely straight path, whereas our method produces a

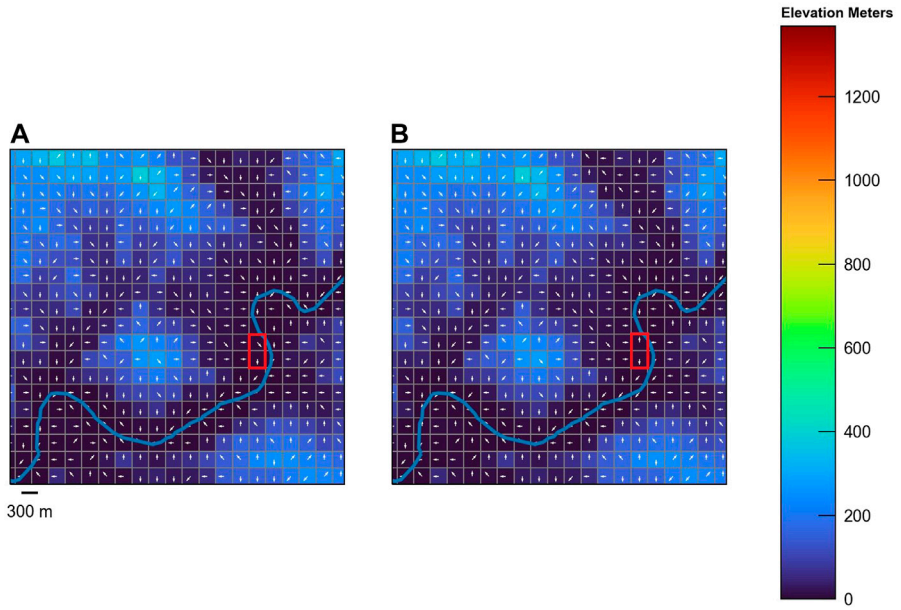

FIGURE 3 | Maps showing calculated flow direction vectors for a part of the Russian River watershed (inset box in **Figure 1B**) with a low relief valley bounded by steep topography. The modified Dijkstra Algorithm (Dijkstra, 1959) produces flow direction vectors that generally follow the NHDPlus streamline (blue; Buto and Anderson, 2020) **(A)**. The red box shows a problem section of the watershed that contains digital artifacts. Stream direction vectors within the red box drain downstream in **(A)**. The red box in **(B)** shows stream direction vectors that diverge from the watershed drainage pattern and create a condition where the flow direction array does provide a continuous drainage path.

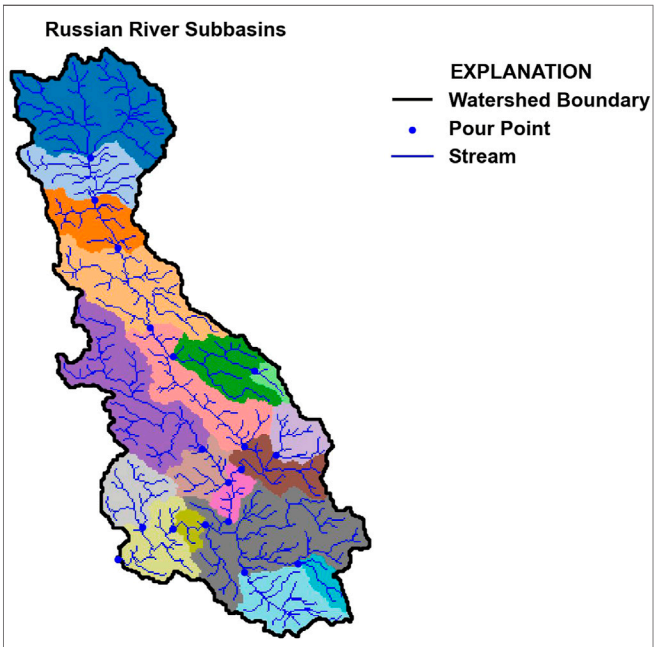

FIGURE 4 | Watershed and subbasin delineation results for the Russian River watershed. U.S. Geological Survey streamgages (U.S. Geological Survey, 2021) were supplied to the script as pour points for watershed and subbasin delineation. The black outline represents the Russian River watershed, and unique colors represent individual subbasins calculated by pyGSFLOW. Blue lines show the stream network generated from 1 Arc-second (27.28 m) digital elevation data (U.S. Geological Survey, 2020) by pyGSFLOW, Russian River watershed, northwestern California.

sinuous section of stream cells (**Figure 5**). Both representations of the stream network would likely be suited for hydrologic modeling; however, pyGSFLOW is open-source, and additions and inputs can be suggested by the larger modeling community.

Sagehen Creek Watershed Example

The Sagehen Creek watershed is located near Lake Tahoe, California, in the northern Sierra Nevada (**Figure 6**). The watershed drains an area of about 27 km^2 and has an east facing aspect with about 720 m of relief. The Sagehen Creek watershed has been described in detail and documented as a GSFLOW example problem by Markstrom et al. (2008). In this example, pyGSFLOW model building tools are applied to the Sagehen Creek watershed to create two separate GSFLOW models from raster data with model grid discretization of 50 m × 50 m and 90 m × 90 m to illustrate the utility of pyGSFLOW for creating multiple model frameworks or conceptualizations of the same hydrologic system; specifically, the grid-cell size is evaluated with respect to simulated streamflow and surface-water/groundwater exchanges. This version of the Sagehen Creek watershed model has a different spatial discretization compared to that presented by Markstrom et al., 2008 and consequently has a different set of parameter values and solution.

Ancillary data sets used to develop the Sagehen Creek GSFLOW model with pyGSFLOW include a 1 arc-second (30 m) resolution DEM for the Sagehen Creek watershed area (U.S. Geological Survey, 2022), a pour point located at the USGS streamgage near the outlet of the watershed (10,343,500

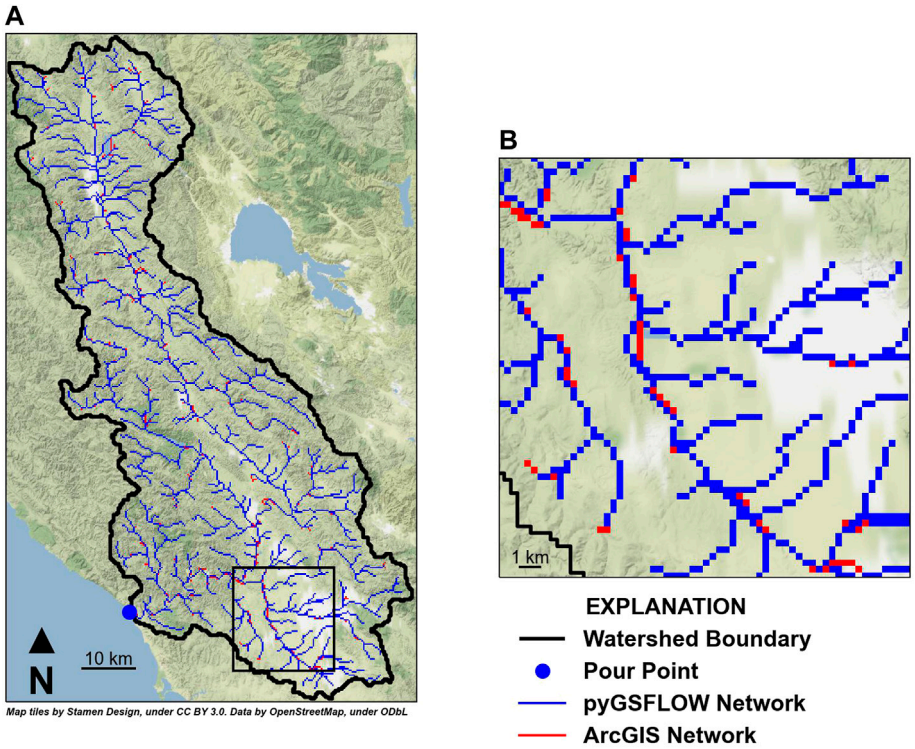

FIGURE 5 | Stream network generation from pyGSFLOW (blue lines) and ArcGIS Pro's Flow Direction and Flow Accumulation tools (red lines) show that both methods are able to create representations of the Russian River watershed's drainage pattern **(A)**. Some differences in the two stream network representations are observed in low-relief areas throughout the watershed **(B)**.

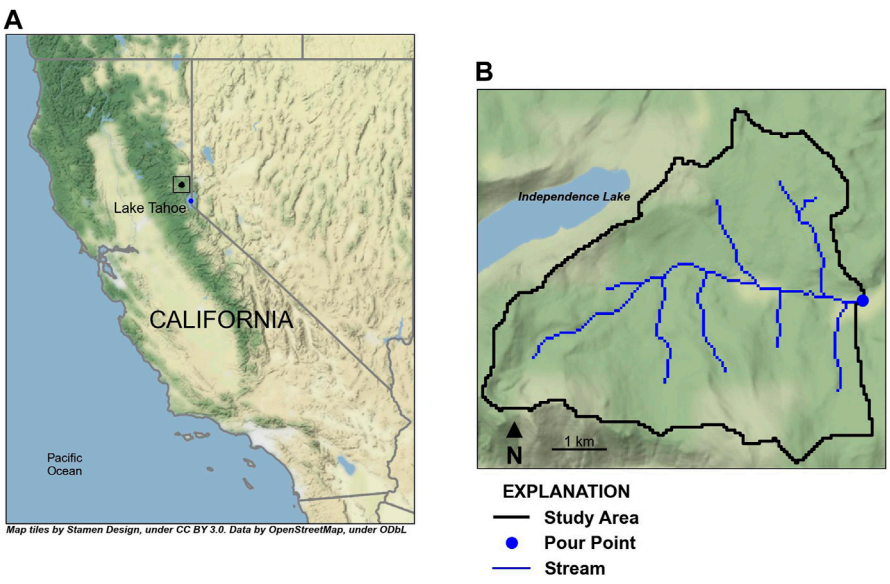

FIGURE 6 | Location of the Sagehen Creek watershed study area near Lake Tahoe, California, in the northern Sierra Nevada **(A)**; the study area extent (black outline), Sagehen Creek watershed streamlines (blue lines), and USGS streamgage (10,343,500 SAGEHEN C NR TRUCKEE CA; U.S. Geological Survey, 2021; blue marker) is shown in **(B)**.

SAGEHEN C NR TRUCKEE CA), LANDFIRE existing vegetation layers (LANDFIRE, 2016), SSURGO 1:24,000 inventory of soil and non-soil layers (USDA, 2021), National Land Cover Database (NLCD) impervious cover data layer (National Land Cover Database, 2020), and PRISM 30-years normals (PRISM Climate Group, 2014). Data from Sagehen Creek co-operative station were provided by University of California, Berkeley (2008), and these daily values of minimum and maximum air temperature and precipitation were distributed to all model grid cells using the PRMS modules temp_1sta and precip_1sta and adjustment factors calculated from the PRISM 30-years normals (800 m resolution), due to spatial resolution constraints from PRISM daily data (4 km resolution).

Two scripts were written specifically for the Sagehen Creek watershed example to construct models with a constant 50 m × 50 m grid cell size (50 m model) and a constant 90 m × 90 m grid cell size (90 m model). These scripts processed the ancillary data sets to construct the GSFLOW models with pyGSFLOW. The outline of the Sagehen Creek watershed script is provided below to illustrate the process. However, some of the coding is not included here for brevity, and users can refer to the two python scripts, Sagehen_50 m py and Sagehen_90 m py or example notebooks, included in the pyGSFLOW repository for details:

1) Generate a structured model grid and resample the native DEM to the model grid:

```
dem_file = os.path.join(ws,'dem.img')
cellsize = 50
modelgrid = GenerateFishnet(dem_file, xcellsize=cellsize, ycellsize=cellsize)
raster = Raster.load(dem_file)
dem = raster.resample_to_grid(modelgrid, band=raster.bands[0], method="median",
multithread=True, thread_pool=12)
```

2) Generate the flow direction and flow accumulation data sets for the model grid:

```
hru_type = np.ones((modelgrid.nrow, modelgrid.ncol), dtype=int)
fa = FlowAccumulation(dem, modelgrid.xcellcenters, modelgrid.ycellcenters,
hru_type=hru_type, verbose=True)
flow_dir = fa.flow_directions(dijkstra=True, breach=0.001)
flow_acc = fa.flow_accumulation()
```

3) Define the watershed boundary using the pour point located at the USGS streamgage:

```
watershed = fa.define_watershed(pour_point, modelgrid, fmt='xy')
```

4) Generate the stream network and cascade directions used to route flow from overland runoff and interflow to streams:

```
strm_obj = fa.make_streams(flow_dir, flow_acc, threshold)
cascades = fa.get_cascades(strm_obj)
```

5) Generate the MODFLOW component of the GSFLOW input files using the Fishnet and stream network:

```
mfbuild = ModflowBuilder(modelgrid, dem, "sagehen_50m")
ml = mfbuild.build_all(strm_obj.reach_data, strm_obj.segment_data,
strm_obj.irunbnd, finf=np.ones(dem.shape), botm=botm,
ibound=watershed.astype(int), iuzfbnd=watershed.astype(int))
```

6) Generate the PRMS component of the GSFLOW input files:

```
prmsbuild = PrmsBuilder(strm_obj, cascades, modelgrid, fa.get_dem_data().ravel(),
hru_type=watershed, hru_subbasin=watershed)
param_obj = prmsbuild.build()
```

In addition to building the MODFLOW and PRMS components, the GSFLOW control and climate data file also must be built using pyGSFLOW, as shown in the Saghen_50 m py script. Because Sagehen Creek has a relatively small watershed, manual calibration was used here by adjusting MODFLOW input

TABLE 3 | Parameter values for models using two different spatial resolutions. Parameter values were modified from their default values to calibrate the model with 50 m by 50 m horizontal discretization, and calibrated parameters for the 90 m by 90 m model. Simulated snowpack, temperature, and horizontal hydraulic conductivity values were sensitive to changes in model discretization.

Input Data/Parameter	Sagehen_50 m model	Sagehen_90 m model
Grid cell dimension (in meters)	50	90
Number of layers, rows, and columns	1,149,138	1,77,83
Horizontal hydraulic conductivity of aquifer (in meters per day)	0.018	0.022
Aquifer specific storage (in per meter)	1 × 10-7	1 × 10-7
Aquifer specific yield	0.2	0.2
Model layer thickness (in meters)	100	100
Saturated water content of unsaturated zone	0.25	0.25
Brooks-Corey exponent	3.5	3.5
Vertical hydraulic conductivity of the unsaturated zone (in meters per day)	1	1
Streambed hydraulic conductivity (in meters per day)	1	1
Average stream cross-sectional width (in meters)	10	10
Mannings roughness coefficient	0.04	0.04
Depth water holding capacity of the soil zone held in tension (soil_moist_max, in centimeters)	9–15.12	9–15.12
Depth of water holding capacity of the soil zone drained by gravity (sat_threshold, in centimeters)	333	333
Jensen-Haise potential evapotranspiration coefficient (in per degrees Fahrenheit)	0.03	0.03
Lapse rates for minimum and maximum air temperatures (in degrees Celsius per 1,000 m)	1.2	1
Maximum air temperature when precipitation is assumed to be all snow (in degrees Celsius)	0.7	0.3
Maximum air temperature when precipitation is assumed to be rain (in degrees Celsius)	2.1	3.1
Maximum snowmelt infiltration rate (in inches per day)	10	4

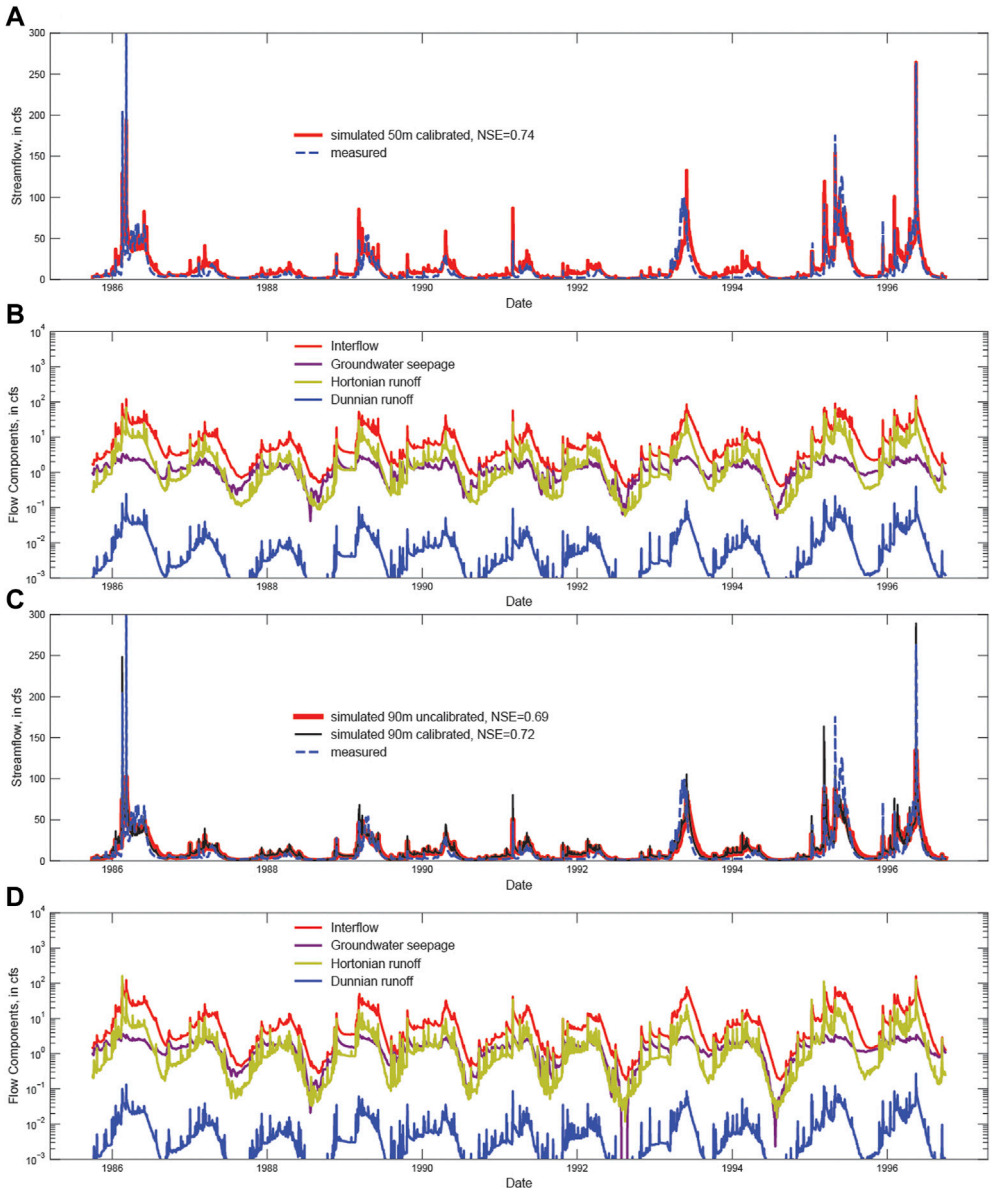

FIGURE 7 | Sagehen Creek test model (Larsen et al., 2021) results showing comparisons between simulated and measured streamflow (10,343,500 SAGEHEN C NR TRUCKEE CA; U.S. Geological Survey, 2021) and contributions to streamflow for **(A,B)** 50 m × 50 m, and **(C,D)** 90 m × 90 m model cell sizes, respectively. All flows correspond to the location of the streamgage at the outlet of the watershed shown in **Figure 6**. Part **(C)** shows simulated versus measured flows for parameters unchanged from the 50 m calibration used in the 90 m calibration and the calibrated 90 m simulation.

data and PRMS parameters using pyGSFLOW. Manual calibration of the Sagehen Creek watershed for the 50 and 90 m models followed a stepwise procedure outlined by Hay et al. (2006). The MODFLOW input data sets and PRMS parameters determined from the calibration of the 50 m model were first directly applied to the 90 m Sagehen Creek watershed model and then calibrated to evaluate the effects of grid cell size on simulated streamflow (**Table 3**).

Both the 50 and 90 m models were run for the 14-year period 1 October 1982, to 31 September 1996, and the first 3 years of the simulation were used to develop equilibrium storage conditions, often referred to as the "spin-up" period for which results are not included in the calibration or in the results shown in **Figure 7**. Model results were analyzed by comparing the streamflow between the measured and simulated values at the outlet of the watershed (**Figure 7**) and by comparing the components of streamflow for the 50 and 90 m models. GSFLOW simulates several components that contribute to the total streamflow, including interflow that flows laterally to streams through soils, groundwater seepage from aquifers to the stream, Dunnian overland flow generated by saturation excess, and Hortonian overland flow generated by snowmelt and rainfall in excess of the soil infiltration capacity.

Because the 50 m model was calibrated to the measured streamflow, and these parameter values were transferred directly to the 90 m model, the effect of grid size on simulated streamflow can be isolated and evaluated. The results confirm previous research that indicates larger grid cells tend to slightly increase dispersion and attenuate peak flow events (Sulis et al., 2010), and because the model was calibrated using the 50 m model, this attenuation results in a slight underprediction of peak flow in the 90 m model. The daily Nash-Sutcliffe Efficiency (NSE) values are equal to 0.74 and 0.69 for the 50 and 90 m models, respectively, when the 50 m calibrated parameters are transferred unchanged to the 90 m model. Calibrated results for the 90 m model show that a slightly better NSE is obtained with the finer discretization, where NSE is 0.72 and 0.74 for the calibrated 90 m and calibrated 50 m models (comparison between subparts 7a and 7c). Note that the slight reduction in the NSE value for the 90 m model is a result of changes in the magnitude of Hortonian runoff and groundwater discharge to the stream network shown in **Figures 7B,D**. This example illustrates the value of using automated model construction approaches like that provided by pyGSFLOW to quickly evaluate how different model resolutions impact hydrologic prediction, a process that is time consuming and prohibitive using conventional model construction approaches. The pyGSFLOW package provides an automated and efficient approach for evaluating the impacts of model conceptualization on predictions that allow the user to optimize model construction and quickly balance different factors, such as the tradeoff between fine spatial discretization and accuracy versus model computational costs. Increases in cell size from 50 to 90 m resulted in slightly lower Nash-Sutcliffe Efficiency, 0.74 to 0.72; however, this small sacrifice in accuracy is balanced by the significant reduction in computation time from 603 to 294 s.

SUMMARY AND CONCLUSION

GSFLOW models simulate complex interactions between surface-water and groundwater flow systems and require large data sets from many sources to fully parameterize. This paper presents methods that outline GSFLOW integrated hydrologic model development from raster digital elevation data to running model with pyGSFLOW. This approach builds on previous works (Bakker et al., 2016; Gardner et al., 2018; Larsen et al., 2021; Bakker et al., 2022) to create an open-source method for building PRMS, MODFLOW, and ultimately GSFLOW models. Flow direction, flow accumulation, watershed and subbasin delineation, and model building methods were developed specifically for use with tightly coupled GSFLOW models. Two example problems are presented to illustrate the robustness of the approach and illustrate the model construction process using pyGSFLOW.

The "Russian River Watershed Example" presents a regional system that is characterized by large areas of digitally flat digital elevation model (DEM), low-relief terrain with steep canyon walls that creates digital artifacts in the DEM data, and areas with high-relief topography. The D8 flow direction algorithm implemented in pyGSFLOW was used to define flow vectors within the watershed and ultimately be used to define subbasin boundaries and a model stream network. Results from this study showed that 1) pyGSFLOW's modified Dijkstra algorithm is well suited for solving systems with large digitally flat expanses, like the Russian River; 2) the standard topological D8 flow direction method is ill suited for performing this task; and 3) the results are comparable but slightly different than both NHDPlus streamlines and ArcGIS's flow accumulation methods. Differences between NHDPlus and pyGSFLOW's results are explained by DEM scale mismatches between the model spatial discretization and NHDPlus streamlines.

The "Sagehen Creek Watershed Example" illustrates the step-by-step approach to developing input data required for a GSFLOW application. The python scripts are summarized here and provided in the pyGSFLOW repository to walk the user through the model development process using pyGSFLOW, including the processing of raster data to provide model parameters for both the PRMS and MODFLOW components of GSFLOW. This example compares two models created by varying spatial discretization with pyGSFLOW's model building tools (**Table 3**). The first model has 50 m × 50 m grid cells, and the second model has 90 m × 90 m grid cells. The 90 m × 90 m model was quickly produced from a copy of the 50 m × 50 m model by changing only the spatial discretization of the model. Comparison of the two models shows that larger grid cells impact both the surface-water and groundwater components of simulated streamflow. Additional calibration beyond the parameterization of the 50 m model could be applied to the 90 m model to compensate for the deterioration in model fit when compared to the finer discretization model.

The pyGSFLOW package is currently being used to develop GSFLOW models of hydrologic watersheds in California to support groundwater sustainability and for tools that water managers can use to better manage surface water and groundwater as a single resource. Some example applications include, but are not limited to, evaluating different land-management or land-use scenarios, evaluating climate scenarios under historical or future conditions, parameter estimation, and sensitivity analysis. Because pyGSFLOW is a programmatic method for model creation, editing, and postprocessing, these applications can be accomplished by either creating a comprehensive script or with a series of scripts. Both options can be used for repeatable, transferable, and transparent model development.

The pyGSFLOW package is an open-source project that welcomes community input and involvement. The Russian River watershed and Sagehen Creek watershed example problems discussed in this paper can be found as python scripts in the pyGSFLOW repository (Larsen et al., 2021). Installation instructions, example problems, and links to documentation that demonstrate model building, editing existing models, and output data visualization can be accessed from the pyGSFLOW repository.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://code.usgs.gov/water/pyGSFLOW/-/tree/master/examples/frontiers and https://code.usgs.gov/water/pyGSFLOW/-/tree/master/examples.

AUTHOR CONTRIBUTIONS

JL worked as lead software developer and lead author. AA has been a lead developer on the pyGSFLOW methods and authored sections of this article. DM developed approximately 50% of the new pyGSFLOW model building methods described and created the figures for this work. RN provided guidence on this work, calibrated the example problems, and authored parts of the article.

FUNDING

This work was funded by the USGS's Water Availability and Use Program, the California State Water Resources Control Board, and Sonoma County Water Agency.

ACKNOWLEDGMENTS

The authors welcome additions, suggestions, and assistance from the scientific community, and thank all past contributors for their work.

REFERENCES

Bakker, M., Post, V., Hughes, J. D., Langevin, C. D., White, J. T., Leaf, A. T., et al. (2022). *FloPy v3.3.6 — Release Candidate*. U.S. Geological Survey Software Release. doi:10.5066/F7BK19FH

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting MODFLOW Model Development Using Python and FloPy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Beven, K. (2019). How to Make Advances in Hydrological Modelling. *Hydrology Res.* 50, 1481–1494. doi:10.2166/nh.2019.134

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., and Renner, M. (2019). Twenty-three Unsolved Problems in Hydrology (UPH)–a Community Perspective. *Hydrological Sci. J.* 64, 10. doi:10.1080/02626667.2019.1620507

Buto, S. G., and Anderson, R. D. (2020). *NHDPlus High Resolution (NHDPlus HR)—A Hydrography Framework for the Nation*. Reston, VA: U.S. Geological Survey. Fact Sheet 2020-3033. doi:10.3133/fs20203033

Clark, M. P., Kavetski, D., and Fenicia, F. (2011). Pursuing the Method of Multiple Working Hypotheses for Hydrological Modeling. *Water Resour. Res.* 47, 9. doi:10.1029/2010WR009827

Condon, L. E., and Maxwell, R. M. (2015). Evaluating the Relationship between Topography and Groundwater Using Outputs from a Continental-Scale Integrated Hydrology Model. *Water Resour. Res.* 51, 6602–6621. doi:10.1002/2014WR016774

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. *Numer. Math.* 1, 269–271. doi:10.1007/BF01386390

Garbrecht, J., and Martz, L. W. (1997). The Assignment of Drainage Direction over Flat Surfaces in Raster Digital Elevation Models. *J. Hydrology* 193, 204–213. doi:10.1016/S0022-1694(96)03138-1

Gardner, M. A., Morton, C. G., Huntington, J. L., Niswonger, R. G., and Henson, W. R. (2018). Input Data Processing Tools for the Integrated Hydrologic Model GSFLOW. *Environ. Model. Softw.* 109, 41–53. doi:10.1016/j.envsoft.2018.07.020

Goodchild, M. F. (2011). Scale in GIS: An Overview. *Geomorphology* 130, 5–9. doi:10.1016/j.geomorph.2010.10.004

Harbaugh, A. W., Banta, E. R., Hill, M. C., and McDonald, M. G. (2000). *MODFLOW-2000, the U.S. Geological Survey Modular Ground-Water Model – User Guide to Modularization Concepts and the Ground-Water Flow Process*. Reston, VA: U.S. Geological Survey. Open-File Report 00-92. doi:10.3133/ofr200092

Harbaugh, A. W. (2005). "MODFLOW-2005 : the U.S. Geological Survey Modular Ground-Water Model-Tthe Ground-Water Flow Process," in *Techniques and Methods 6-A16* (Reston, VA: U.S. Geological Survey). doi:10.3133/tm6A16

Hay, L. E., Leavesley, G. H., Clark, M. P., Markstrom, S. L., Viger, R. J., and Umemoto, M. (2006). Step Wise, Multiple Objective Calibration of a Hydrologic Model for a Snowmelt Dominated Basin. *J. Am. Water Resour. Assoc.* 42 (4), 877–890. doi:10.1111/j.1752-1688.2006.tb04501.x

Henson, W. R., Medina, R. L., Mayers, C. J., Niswonger, R. G., and Regan, R. S. (2013). "CRT—Cascade Routing Tool to Define and Visualize Flow Paths for Grid-Based Watershed Models," in *Techniques and Methods 6-D2* (Reston, VA: U.S. Geological Survey), 28.

Huntington, J. L., and Niswonger, R. G. (2012). Role of Surface-Water and Groundwater Interactions on Projected Summertime Streamflow in Snow Dominated Regions: An Integrated Modeling Approach. *Water Resour. Res.* 48, 11. doi:10.1029/2012WR012319

Jenson, S. K., and Domingue, J. O. (1988). Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis. *Photogrammetric Eng. remote Sens.* 54 (11), 1593–1600.

Kahn, A. B. (1962). Topological Sorting of Large Networks. *Commun. ACM* 5, 558–562. doi:10.1145/368996.369025

Kampf, S. K., and Burges, S. J. (2007). A Framework for Classifying and Comparing Distributed Hillslope and Catchment Hydrologic Models. *Water Resour. Res.* 43, 5. doi:10.1029/2006WR005370

Kitlasten, W., Morway, E. D., Niswonger, R. G., Gardner, M., White, J. T., Triana, E., et al. (2021). Integrated Hydrology and Operations Modeling to Evaluate Climate Change Impacts in an Agricultural Valley Irrigated with Snowmelt Runoff. *Water Resour. Res.* 57, 6. doi:10.1029/2020WR027924

Konikow, L. F., and Kendy, E. (2005). Groundwater Depletion: A Global Problem. *Hydrogeol. J.* 13 (1), 317–320. doi:10.1007/s10040-004-0411-8

LANDFIRE (2016). *Data from: LANDFIRE Existing Vegetation Type Layer*. Reston, VA: U.S.Department of Interior, Geological Survey, and U.S. Department of Agriculture. Available at: http://landfire.cr.usgs.gov/viewer/.

Larsen, J. D., Alzraiee, A., and Niswonger, R. G. (2022). Integrated Hydrologic Model Development and Postprocessing for GSFLOW Using pyGSFLOW. *J. Open Source Softw.* 7, 3852. 72. doi:10.21105/joss.03852

Larsen, J. D., Alzraiee, A., and Niswonger, R. (2021). *pyGSFLOW v1.0.0*. U.S. Geological Survey Software Release. doi:10.5066/P9NPZ5AD

Leaf, A. T., Fienen, M. N., and Reeves, H. W. (2021). SFRmaker and Linesink-Maker: Rapid Construction of Streamflow Routing Networks from Hydrography Data. *Groundwater* 59, 761–771. doi:10.1111/gwat.13095

Maidment, D. R., and Morehouse, S. (2002). *Arc Hydro: GIS for Water Resources*. Redlands, CA: ESRI, Inc.

Mark, D. M. (1988). "Network Models in Geomorphology," in *Modelling in Geomorphological Systems*. Editor M. G. Anderson (Chichester [West Sussex], NY: John Wiley), 73–97.

Markstrom, S. L., Niswonger, R. G., Regan, R. S., Prudic, D. E., and Barlow, P. M. (2008). *GSFLOW-coupled Ground-Water and Surface-Water FLOW Model Based on the Integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005)*. Reston, VA: U.S. Geological Survey, 240. Techniques and Methods 6-D1.

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M. T., Payn, R. A., et al. (2015). *PRMS-IV, the Precipitation-Runoff Modeling System, Version 4.*

Reston, VA: U.S. Geological Survey, 158. Techniques and Methods, book 6, chap. B7.

Martz, L. W., and Garbrecht, J. (1999). An Outlet Breaching Algorithm for the Treatment of Closed Depressions in a Raster DEM. *Comput. Geosciences* 25, 835–844. doi:10.1016/s0098-3004(99)00018-7

Metz, M., Mitasova, H., and Harmon, R. S. (2011). Efficient Extraction of Drainage Networks from Massive, Radar-Based Elevation Models with Least Cost Path Search. *Hydrol. Earth Syst. Sci.* 15, 667–678. doi:10.5194/hess-15-667-2011

National Land Cover Database (NLCD). (2020). Data from: 2016 Shrubland Fractional Components for the Western U.S. doi:10.5066/P9MJVQSQ

Ng, G.-H. C., Wickert, A. D., Somers, L. D., Saberi, L., Cronkite-Ratcliff, C., Niswonger, R. G., et al. (2018). GSFLOW-GRASS v1.0.0: GIS-Enabled Hydrologic Modeling of Coupled Groundwater-Surface-Water Systems. *Geosci. Model Dev.* 11, 4755–4777. doi:10.5194/gmd-11-4755-2018

Niswonger, R. G., Panday, S., and Ibaraki, M. (2011). *MODFLOW-NWT, A Newton Formulation for MODFLOW-2005*. Reston, VA: Geological Survey, 44. Techniques and Methods 6-A37. doi:10.3133/tm6A37

Niswonger, R. G., and Prudic, D. E. (2005). *Documentation of the Streamflow-Routing (SFR2) Package to Include Unsaturated Flow beneath Streams - A Modification to SFR1*. Reston, VA: Geological Survey, 47. Techniques and Methods 6-A13. doi:10.3133/tm6A13

Niswonger, R. G., Prudic, D. E., and Regan, R. S. (2006). *Documentation of the Unsaturated-Zone Flow (UZF1) Package for Modeling Unsaturated Flow between the Land Surface and the Water Table with MODFLOW-2005*. Reston, VA: U.S. Geological Survey, 62. Techniques and Methods 6-A19. doi:10.3133/tm6A19

O'Callaghan, J. F., and Mark, D. M. (1984). The Extraction of Drainage Networks from Digital Elevation Data. *Comput. Vis. Graph. Image Process.* 28, 323–344. doi:10.1016/S0734-189X(84)80011-0

PRISM Climate Group (2014). *Data from: Prism Climate Group*. Corvallis, OR: Oregon State University. Available at: https://prism.oregonstate.edu.

Qin, C., Zhu, A. X., Pei, T., Li, B., Zhou, C., and Yang, L. (2007). An Adaptive Approach to Selecting a Flow-Partition Exponent for a Multiple-Flow-Direction Algorithm. *Int. J. Geogr. Inf. Sci.* 21, 443–458. doi:10.1080/13658810601073240

Regan, R. S., Juracek, K. E., Hay, L. E., Markstrom, S. L., Viger, R. J., Driscoll, J. M., LaFontaine, J. H., and Norton, P. A. (2019). The U. S. Geological Survey National Hydrologic Model Infrastructure: Rationale, Description, and Application of a Watershed-Scale Model for the Conterminous United States. *Environ. Model. Softw.* 111, 192–203. doi:10.1016/j.envsoft.2018.09.023

Schoups, G., Hopmans, J. W., Young, C. A., Vrugt, J. A., Wallender, W. W., Tanji, K. K., et al. (2005). Sustainability of Irrigated Agriculture in the San Joaquin Valley, California. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15352–15356. doi:10.1073/pnas.0507723102

Schoups, G., Vrugt, J. A., Fenicia, F., and Van De Giesen, N. C. (2010). Corruption of Accuracy and Efficiency of Markov Chain Monte Carlo Simulation by Inaccurate Numerical Implementation of Conceptual Hydrologic Models. *Water Resour. Res.* 46, 10. doi:10.1029/2009WR008648

Shin, S., Pokhrel, Y., and Miguez-Macho, G. (2019). High-Resolution Modeling of Reservoir Release and Storage Dynamics at the Continental Scale. *Water Resour. Res.* 55, 787–810. doi:10.1029/2018WR023025

Sulis, M., Paniconi, C., and Camporese, M. (2010). Impact of grid resolution on the integrated and distributed response of a coupled surface-subsurface hydrological model for the des Anglais catchment, Quebec. *Hydrol. Process.* 25, 1853–1865. doi:10.1002/hyp.7941

Tarboton, D. G. (1997). A New Method for the Determination of Flow Directions and Upslope Areas in Grid Digital Elevation Models. *Water Resour. Res.* 33, 309–319. doi:10.1029/96WR03137

University of California Berkeley (2008). *Data from: Sagehen Creek Field Station, UC Berkeley*. Berkeley, CA: Historical Weather Data. Available at: https://wrcc.dri.edu/cgi-bin/rawMAIN.pl?nvsagh.

U.S. Geological Survey (2020). Data from: 3D Elevation Program 1-Meter Resolution Digital Elevation Model. Available at: https://elevation.nationalmap.gov/arcgis/rest/services/3DEPElevation/ImageServer.

U.S. Geological Survey (2021). Data from: National Water Information System Data Available on the World Wide Web (Water Data for the Nation). doi:10.5066/F7P55KJN

U.S. Geological Survey (2022). *Data from: USGS 3D Elevation Program Digital Elevation Model*. Available at: https://elevation.nationalmap.gov/arcgis/rest/services/3DEPElevation/ImageServer.

Volk, J. M., and Turner, M. A. (2019). PRMS-Python: A Python Framework for Programmatic PRMS Modeling and Access to Its Data Structures. *Environ. Model. Softw.* 114, 152–165. doi:10.1016/j.envsoft.2019.01.006

Wada, Y., Van Beek, L. P. H., Van Kempen, C. M., Reckman, J. W. T. M., Vasak, S., and Bierkens, M. F. P. (2010). Global Depletion of Groundwater Resources. *Geophys. Res. Lett.* 37, L20402. doi:10.1029/2010GL044571

Wang, L., and Liu, H. (2006). An Efficient Method for Identifying and Filling Surface Depressions in Digital Elevation Models for Hydrologic Analysis and Modelling. *Int. J. Geogr. Inf. Sci.* 20, 193–213. doi:10.1080/13658810500433453

Wang, Y., Liu, Y., Xie, H., and Xiang, Z. (2011). "A Quick Algorithm of Counting Flow Accumulation Matrix for Deriving Drainage Networks from a DEM," in Proceedings on the Third International Conference on Digital Image Processing. doi:10.1117/12.896274

Werner, A. D., Gallagher, M. R., and Weeks, S. W. (2006). Regional-scale, Fully Coupled Modelling of Stream-Aquifer Interaction in a Tropical Catchment. *J. Hydrology* 328, 497–510. doi:10.1016/j.jhydrol.2005.12.034

Wood, E. F., Lettenmaier, D., Liang, X., Nijssen, B., and Wetzel, S. W. (1997). Hydrological Modeling of Continental-Scale Basins. *Annu. Rev. Earth Planet. Sci.* 25, 279–300. doi:10.1146/annurev.earth.25.1.279

Zhang, H., Yao, Z., Yang, Q., Li, S., Baartman, J. E. M., Gai, L., et al. (2017). An Integrated Algorithm to Evaluate Flow Direction and Flow Accumulation in Flat Regions of Hydrologically Corrected DEMs. *Catena* 151. doi:10.1016/j.catena.2016.12.009

Author Disclaimer: Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors AA.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Earth Science

TYPE Technology and Code
PUBLISHED 23 August 2022
DOI 10.3389/feart.2022.907614

OPEN ACCESS

EDITED BY
Michael Fienen,
United States Geological Survey,
United States

REVIEWED BY
Andrew Leaf,
United States Geological Survey,
United States
Urminder Singh,
Iowa State University, United States

*CORRESPONDENCE
Stephanie R. James,
sjames@usgs.gov

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Earth Science

RECEIVED 30 March 2022
ACCEPTED 11 July 2022
PUBLISHED 23 August 2022

CITATION
James SR, Foks NL and Minsley BJ (2022), GSPy: A new toolbox and data standard for Geophysical Datasets.
Front. Earth Sci. 10:907614.
doi: 10.3389/feart.2022.907614

GSPy: A new toolbox and data standard for Geophysical Datasets

Stephanie R. James[1]*, Nathan Leon Foks[2] and Burke J. Minsley[1]

[1]U.S. Geological Survey, Geology, Geophysics, and Geochemistry Science Center, Denver, CO, United States, [2]Apogee Engineering LLC. Contracted to U.S. Geological Survey, Science Analytics and Synthesis, Advanced Research Computing, Denver, CO, United States

The diversity of geophysical methods and datatypes, as well as the isolated nature of various specialties (e.g., electromagnetic, seismic, potential fields) leads to a profusion of separate data file formats and documentation conventions. This can hinder cooperation and reduce the impact of datasets researchers have invested in heavily to collect and prepare. An open, portable, and well-supported community data standard could greatly improve the interoperability, transferability, and long-term archival of geophysical data. Airborne geophysical methods particularly need an open and accessible data standard, and they exemplify the complexity that is common in geophysical datasets where critical auxiliary information on the survey and system parameters are required to fully utilize and understand the data. Here, we propose a new Geophysical Standard, termed the GS convention, that leverages the well-established and widely used NetCDF file format and builds on the Climate and Forecasts (CF) metadata convention. We also present an accompanying open-source Python package, GSPy, to provide methods and workflows for building the GS-standardized NetCDF files, importing and exporting between common data formats, preparing input files for geophysical inversion software, and visualizing data and inverted models. By using the NetCDF format, handled through the Xarray Python package, and following the CF conventions, we standardize how metadata is recorded and directly stored with the data, from general survey and system information down to specific variable attributes. Utilizing the hierarchical nature of NetCDF, GS-formatted files are organized with a root *Survey* group that contains global metadata about the geophysical survey. Data are then organized into subgroups beneath *Survey* and are categorized as *Tabular* or *Raster* depending on the geometry and point of origin for the data. Lastly, the standard ensures consistency in constructing and tracking coordinate reference systems, which is vital for accurate portability and analysis. Development and adoption of a NetCDF-based data standard for geophysical surveys can greatly improve how these complex datasets are shared and utilized, making the data more accessible to a broader science community. The architecture of GSPy can be easily transferred to additional geophysical datatypes and methods in future releases.

KEYWORDS

data standards, NetCDF, open-source software, geophysics, airborne geophysics

1 Introduction

Accurate management and usage of scientific data is fundamentally dependent on how the data are stored and documented. Community-agreed-upon standards in data formatting and organization are a natural and necessary step in simplifying the transfer and analysis of complex datasets, both within and across disciplines. In the Earth Sciences, many communities of practice have evolved, such as Cooperative Ocean/Atmosphere Research Data Service (COARDS) from the National Oceanic and Atmospheric Administration (NOAA), Common Data Form (CDF) from the National Aeronautics and Space Administration (NASA), or Hierarchical Data Format (HDF) originally developed by the National Center for Supercomputing Applications (NCSA) but currently maintained by The HDF Group (NOAA, 1995; Folk et al., 1999; Yang et al., 2005; NASA, 2019). Notably, the Network Common Data Form (NetCDF) architecture has become the basis for many modern data standards (Rew et al., 2006; Hankin et al., 2010; Unidata, 2021b). Overall, the purpose of a data standard is to control how data and metadata are documented, formatted, and stored such that datasets can be shared, displayed, and operated on with minimal user intervention across platforms and software (Eaton et al., 2020).

Geophysical datasets are widely used in Earth system studies to interrogate subsurface properties and processes. Methods vary considerably, each relying on different physics and are sensitive to different physical properties of Earth materials (e.g., rocks, sediments, and fluids). Geophysical data are commonly acquired using instruments on land, on or beneath water, from airborne platforms, or in boreholes. Broad categories of geophysical methods (e.g., electrical, magnetic, seismic, electromagnetic, radiometric, gravity) have specific measurement modalities (e.g., frequency-domain or time-domain electromagnetics), each of which can have many unique instruments with differing designs and configurations. In addition to the values measured by an instrument's sensors, a host of other auxiliary information is often needed but contained in separate supplementary files, field notes, or contractor's reports and not directly attached to the data. The supplementary information includes fundamental positioning information, general survey metadata, as well as details about acquisition parameters or instrument characteristics needed to interpret the measured data. Without this accompanying supplementary information, acquiring meaningful results and interpretations would be a challenge.

Although geophysical datasets have much in common at a basic level—recorded data values, system information, coordinate information, and auxiliary metadata—data formats vary widely by method and by instrument. Probably the most established geophysical formats relate to the Society of Exploration Geophysicists (SEG) digital tape standards used for seismic data, owing to the vast amount of industrial seismic data collection (Northwood et al., 1967; Hagelund and Levin, 2017). Yet, even within data formats that are more widely used in the geophysical community, none meet the criteria of 1) being an open format that allow for publication according to Findability, Accessibility, Interoperability, and Reuse (FAIR) principles in public repositories, 2) attaching important system information and metadata to the data in a single file, and 3) incorporate a file structure that facilitates transferability between open-source computational software, web services, and geospatial systems. The lack of a common open data standard leads to inefficiencies where processing or interpretation software must be customized to read specific formats from different instruments, and data need to be re-formatted before they can be used by software and/or published according to FAIR standards (Wilkinson et al., 2016; Salman et al., 2022).

Similar to seismic acquisitions, airborne geophysical surveys are often acquired by industry for a wide range of government, academic, and private clients. Airborne geophysical surveys are becoming more commonplace, providing cost-effective, high resolution, and multi-scale subsurface imaging not easily obtained with ground-based observations over large areas. As with the field of geophysics overall, there is currently no open community standard that is widely used for sharing and releasing airborne geophysical datasets. Furthermore, airborne datasets entail significant supplementary information on survey design, system and acquisition parameters, and post-processing details that are often included in PDFs or other report documents separate from the digital data, posing a risk to the long-term integrity of the data. The large size and complexity of airborne geophysical data, as well as their broad community value, necessitates accessible tools and standards be developed to keep pace with rising demands and usage.

Efforts have been made in the past to standardize airborne data formats, along with interoperable inversion software for working with airborne electromagnetic (AEM) datasets (Møller et al., 2009; Brodie, 2017). The Australian Society of Exploration Geophysicists (ASEG) established the ASEG-GDF2 (General Data Format Revision 2) data standard (Dampney et al., 1985; Pratt, 2003), an ASCII-based data structure for general point and line data, with particular focus on large airborne geophysical datasets such as magnetic, radiometric, electromagnetic, and gravity. Tabular ASCII data, such as ASEG-GDF2 or CSV, have the advantage of being both human and machine readable for easy usage, but these formats result in larger file sizes compared with binary formats. ASCII formats are also limited in how datasets can be structured, grouped, and documented. For example, the ASEG-GDF2 structure includes general and variable-specific metadata information in separate definition files that accompany the data, but this design requires users to always maintain multiple files. In Denmark, a national, publicly accessible geophysical database (GERDA) hosts numerous types of airborne and ground-based geophysical datasets in a structured relational database (Møller et al.,

2009); however, GERDA databases are not easily used or accessed outside of proprietary software. Geosoft databases are also an industry standard for delivery and storage of airborne geophysical tabular datasets. Their binary format has advantages in data compression and file size, and Geosoft databases are supported by sophisticated software such as Oasis Montaj (Seequent Ltd. https://www.seequent.com/products-solutions/) for processing, analysis, and visualization. However, use of this software requires a commercial subscription, and the binary Geosoft databases do not meet open standards for publication. Lastly, gridded data and products often accompany airborne datasets and can be provided in many binary and ASCII raster formats (e.g., TIF/GeoTIFF, ARC/INFO, GXF, Geosoft GRD, Surfer GRD, etc.), each compatible with one or more of the commonly used software tools. However, some tools are open while others are proprietary and require paid subscription.

Here, we present a data standard using the NetCDF file format that provides a structure for storing geophysical data, metadata, and survey information in a single file. The proposed geophysical standard (GS) balances the need to require information for certain datatypes be stored in a well-defined structure, while also allowing for flexibility with optional information. In addition to recorded data, we use the hierarchical group structure within the NetCDF file to store multiple related datasets or products together. For example, separate groups might contain raw data, processed data, and physical property models determined through inversion or other analyses. Storing digital data along with associated coordinate and system information in a single self-describing open file structure with well-established standards can greatly improve the interoperability, transferability, and impact of geophysical datasets. The underlying HDF data structure is computationally advantageous when compared to human-readable ASCII files (Yang et al., 2005; Rew et al., 2006).

Along with the new GS data convention, we developed a Python package (GSPy) as a community tool which facilitates use of the NetCDF file structure. A basic function of GSPy is conversion, either reading original input files into our proposed data structure and creating the standardized NetCDF file or converting content from the standard structure into a different format needed to work with specific software or for cooperator and end-user needs. Beyond this basic input-output functionality, GSPy can also be incorporated into processing and visualization workflows utilizing the GS structure. Though GSPy is not required to work with the GS data model—any tools capable of interacting with a NetCDF file can be used—we developed GSPy as a building block to make the process of transforming datasets into the GS structure easy and straightforward to maximize their usability.

In this paper, we define the proposed data standard and provide an overview of the GSPy software structure and functionality. Our focus in the initial stage of development of the GS model and associated GSPy tools has been on airborne geophysical data due to their immediate need for an open-source community standard, while also keeping in mind flexibility in design to allow future accommodation of other types of geophysical data in the same model. We use an existing airborne geophysical dataset from Wisconsin as a case study to exemplify the GS convention and demonstrate usage of the GSPy package (Minsley et al., 2022). Finally, we discuss the scalability, limitations, and opportunities provided by a NetCDF-based community geophysics data standard.

2 Methods

Our goal with the GS data model and GSPy software tool is to assimilate data from a variety of file formats, geometries, and geophysical methods into a common and open data structure that can be broadly shared and utilized (Figure 1). The GS data model provides a common, open, and standardized framework for geophysical datasets, which is disconnected and independent from the original source formatting.

In airborne surveys, data from one or more geophysical sensors (e.g., electromagnetic, magnetic, radiometric, gravity) are acquired along relatively linear flight lines covering large areas. Data are stored at a regular sampling interval typically in tabular format, and published in ASCII files such as CSV or ASEG-GDF2 (e.g., Ley-Cooper et al., 2019; Drenth and Brown, 2020; Shah, 2020; Minsley et al., 2021). Data from multiple sensors acquired at the same time (e.g., electromagnetic and magnetic) are often combined in a singular tabular dataset at the same sample interval. Two-dimensional rasterized data, typically gridded maps of measured values (e.g., flight altitude or powerline monitor) and/or multi-dimensional interpreted products (e.g., resistivity depth slices or residual magnetic intensity), are often included with contractor-delivered datasets or as publicly archived products. In addition to geophysical sensor data, each measurement also includes important auxiliary information needed for quality control, processing, interpretation, and visualization. Auxiliary metadata includes information such as the position and attitude of the aircraft and geophysical sensors during acquisition, flight line numbers and fiducials, timestamps, noise channels (e.g., powerline monitoring channel for AEM data), and processed or corrected data channels. The GS convention, through GSPy, integrates airborne geophysical data and auxiliary metadata from these various input formats and geometries into a standardized NetCDF file that can be publicly released and shared through data repositories like ScienceBase (https://www.sciencebase.gov), and is portable to common geospatial and visualization software (Figure 1).

2.1 Geophysical data standard

To support efficient metadata documentation, combined storage of related datasets, and transferability to multiple software tools and web services, the GS data model is founded

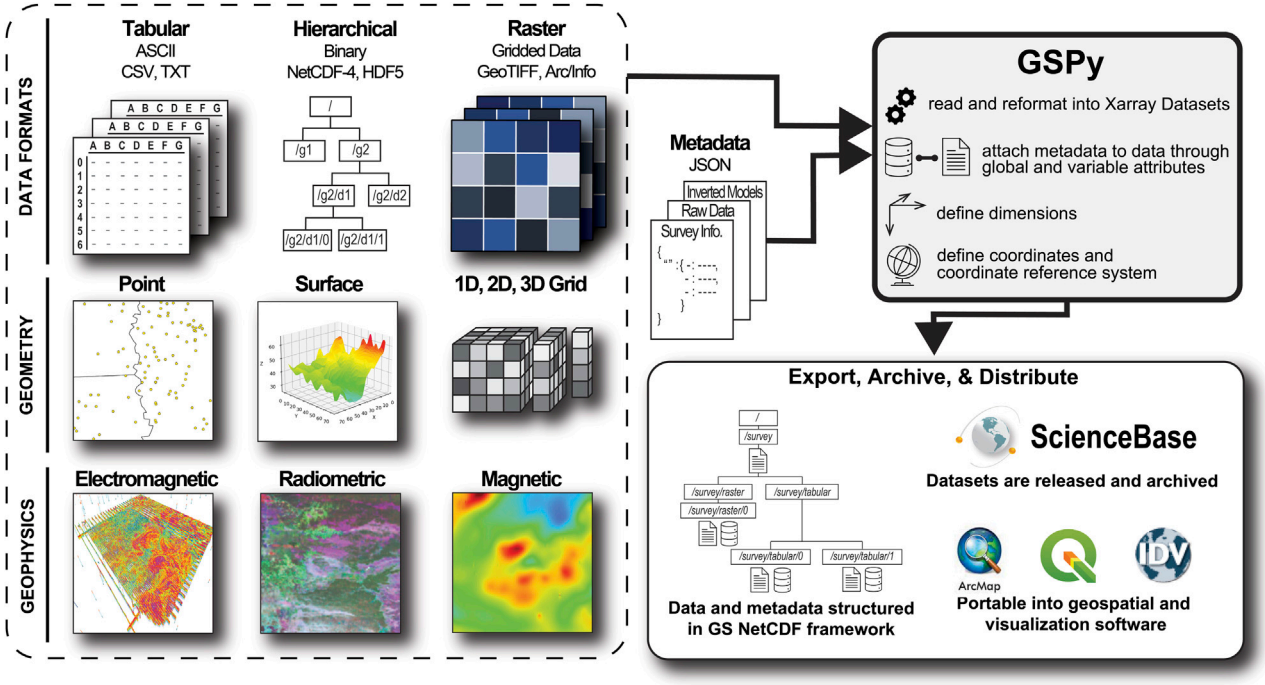

FIGURE 1
Conceptual diagram of GSPy workflow. Data from a variety of formats and types are read into GSPy, along with required metadata files. Through the GSPy software, data are converted into a standardized NetCDF file containing the dataset and metadata appropriate for archiving and sharing.

within the NetCDF file format. NetCDF was established in 1989 by the University Corporation for Atmospheric Research (UCAR)'s Unidata program (Rew and Davis, 1990), who continue to provide development and support for newer NetCDF versions and related software (Rew et al., 2006; Unidata, 2021b). The latest version, NetCDF-4 is built on the HDF5 storage layer and format (Rew et al., 2006). As modern datasets are becoming larger and more complex, e.g., studies are more often data-rich and/or employ "big data" approaches (Vermeesch and Garzanti, 2015; Shelestov et al., 2017; Reichstein et al., 2019; Li and Choi, 2021) the appeal of NetCDF is growing. Organizations such as NASA, NOAA, National Snow and Ice Data Center (NSIDC), and National Center for Atmospheric Research (NCAR) have adopted NetCDF as one of their preferred formats (Ramapriyan and Leonard, 2021, see complete list of users at https://www.unidata.ucar.edu/software/netcdf/usage.html). We have chosen to follow the same path, recognizing the many advantages provided by the NetCDF format:

- **Self-describing**: Metadata are directly attached to datasets. This architecture eliminates any risk of critical metadata becoming separated from the data, which can severely reduce dataset usability. This structure is especially important in geophysical datasets, where auxiliary system information such as transmitter waveforms, time gates, or transmitter-receiver coil orientations are essential for accurate analysis and interpretation of the data.
- **Space-saving**: The binary format has a smaller file size compared to ASCII files. Extra packing and compression options can further reduce file sizes.
- **Accessible**: Subsets of large datasets can be accessed directly without needing to read in the full dataset, thereby minimizing memory requirements.
- **Portable**: Files are platform-independent, meaning datasets are represented uniformly across different computer operating systems.
- **Hierarchical**: Multiple datasets can be stored in a single file following a tiered group organization. This structure provides a clean and efficient mechanism for archiving and sharing related datasets in a single file, such as raw versus processed data, in addition to inverted models and any products derived from those models.
- **Scalable**: Files can be read from and written to using large-scale distributed memory machines, allowing fast access at massive computational scales.

The GS design builds on existing conventions in other Earth science disciplines. Specifically, we adapt and extend the Climate and Forecast (CF) Metadata Conventions (hereafter, the CF conventions;

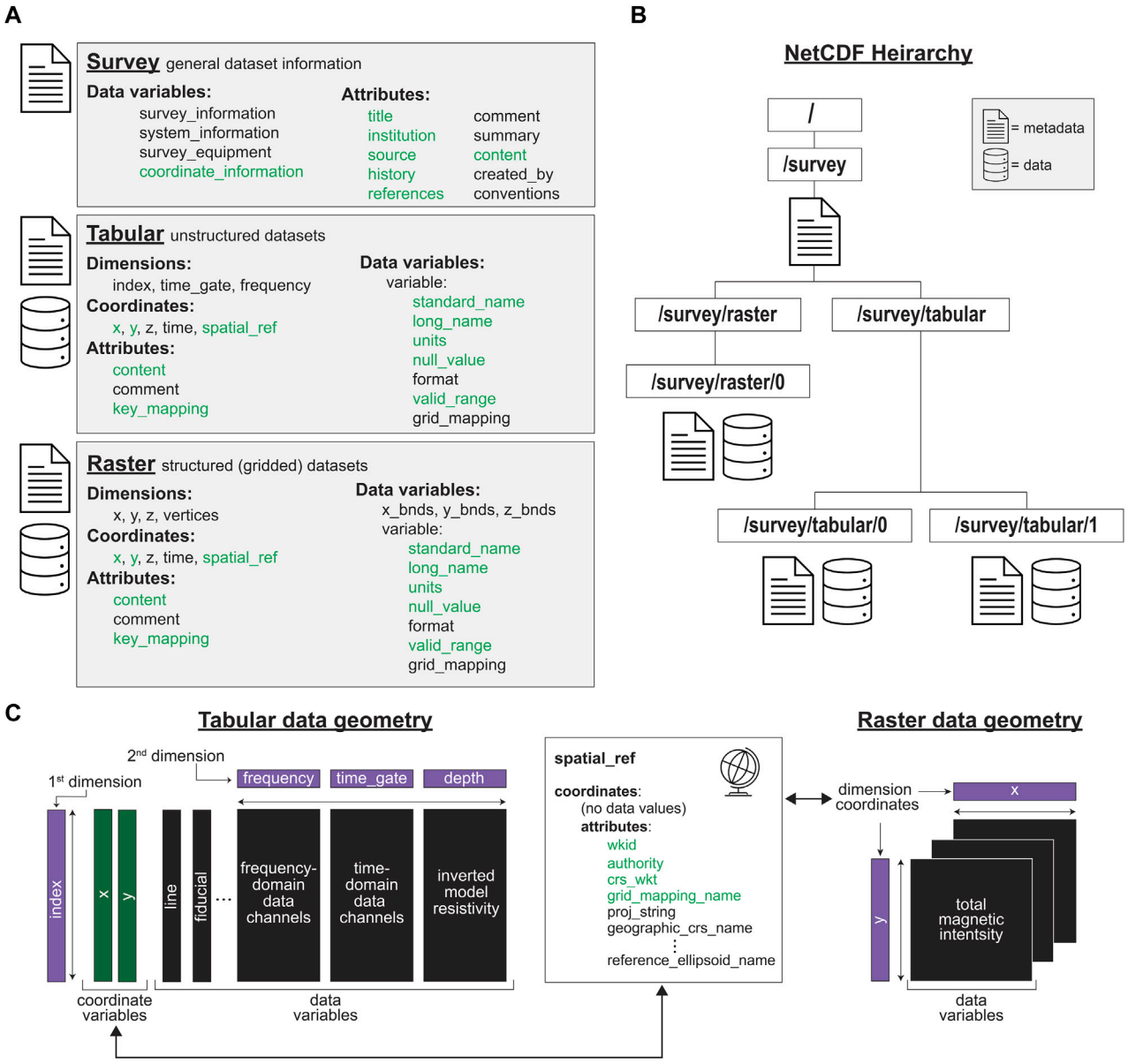

FIGURE 2
GS data convention. **(A)** Datasets are structured into three fundamental group types based on content and data geometry. The *Survey* group contains general metadata about the dataset. Unstructured datasets, such as from CSV or TXT files, form *Tabular* groups, whereas structured (gridded) datasets are categorized under the *Raster* group. Metadata is attached to all groups, with various required attributes (green text) that expands on the CF-1.8 convention. **(B)** Groups follow a strict hierarchy in the NetCDF file, with a single *Survey* group at the top to which all data groups are attached. Datasets are indexed within their respective group type. **(C)** *Tabular* and *Raster* data groups must contain clearly defined dimensions, such as index or x, y, z, as well as coordinate variables. *Raster* groups are distinct in that dimensions are also coordinates, whereas *Tabular* datasets are assigned spatial coordinates that align with the index dimension. Lastly, the coordinate variable "spatial_ref" is required for all data groups, which expands on the "coordinate_information" variable required in the *Survey* metadata.

Eaton et al., 2020) to satisfy the needs of geophysical datasets. The CF conventions ensure datasets conform to a minimum standard of description, with common fields and elements, and guarantee data values can be accurately located in time and space (Eaton et al., 2020). The CF conventions originated as an extension of the COARDS NetCDF convention from NOAA (NOAA, 1995), and are Unidata's recommended standard of choice. The GS convention follows the rules and guidelines of the CF standard, with additional constraints in grouping datasets and metadata while also allowing for nuances inherent to geophysical datasets. Specific details on GS-

specific metadata requirements are outlined in the GSPy documentation pages (Foks et al., 2022).

2.1.1 GS group structure

The hierarchical nature of NetCDF allows for groups of multiple self-described datasets within a single file, where each dataset can have differing structure or dimensions and can be accessed separately using a defined path, similar to a file system directory path (e.g., /group1/group2). The GS data model contains three fundamental categories for grouping data and metadata (Figure 2). General metadata for the dataset(s) as a whole are contained within the *Survey* group. Every file is required to have a *Survey* group which sits at the root of the hierarchical group structure and contains all data subgroups (Figure 2B). Data are then categorized by the nature and geometry of the values. Unstructured data, such as scattered points or lists of values, are contained within the *Tabular* group. Structured data, i.e., gridded data, are contained within the *Raster* group. Having two separate data groups is meant to ease import/export of datasets with minimal manipulation or alteration, thereby ensuring transparency and accountability in cataloging processing steps as well as improving accessibility in data handling. For example, a *Raster* dataset can immediately be exported to a GeoTIFF file, whereas a *Tabular* dataset would require modification such as interpolation onto a regular grid. When there are multiple datasets attached to a single group, we separate them with a simple integer index (e.g., /survey/tabular/0 and /survey/tabular/1, in the case where two tabular entries are attached).

1) *Survey*: This group contains general metadata about the dataset, or collection of related datasets, within the NetCDF file. General information about where the data was collected, acquisition start and end dates, who collected the data, any clients or contractors involved, system specifications, equipment details, and so on are contained within data variables of the *Survey* group. Information included in the *Survey* group is often provided or recorded separately from the data, such as in contractor PDF reports or field notes. Attaching this digital metadata preserves important survey details and facilitates processing and analysis, for example, by including instrument parameters needed for visualization or geophysical inversion. Users are allowed to add as much or little information to the *Survey* data variables as they choose. However, following the CF convention, we require a set of global attributes [e.g., title, institution, source, history, references, see section 2.6.2. of Eaton et al. (2020)]. In the GS standard, we add an additional "content" key that provides a brief summary of what datasets are included in the file and their locations, e.g., "raw data at /survey/tabular/0". Secondly, a "coordinate_information" variable is required within *Survey* and should contain all relevant information about the coordinate reference system. More details on handling coordinate reference systems are described in section 2.1.2.
2) *Tabular*: Data that is organized in a tabular format, such as a CSV file with discrete locations along rows and measurement values along columns, are read and categorized into a *Tabular* group. In the case of airborne geophysics this would include data collected at discrete points along flight lines, inverted physical property models determined from measured data, or any other type of scattered point data.
3) *Raster*: Data that is structured into predefined grids are categorized into the *Raster* group. Generally, this includes two-dimensional (2D) and three-dimensional (3D) gridded data, such as interpolated geophysical models or surfaces.

Data groups are located a level below the *Survey* group in the NetCDF file and have access to the same global metadata (Figure 2B). The hierarchical group structure allows for multiple related datasets to be stored and shared together, such as raw data, processed data, inverted models, and any products derived from those models. This structure also inherently provides an audit trail for users, thereby encouraging transparency and dataset integrity. It is best practice to provide meaningful variable and dimension names and follow established conventions (e.g., CF) or community norms whenever possible. A small set of global attributes are required for all data groups, as well as required variable attributes, and a defined "spatial_ref" variable containing the coordinate system information (Figure 2A).

The relationship between dimensions, coordinates, and data values differs between *Tabular* and *Raster* groups (Figure 2C). For *Tabular* datasets, data variables are more often one-dimensional (1D), such as columns in a CSV, which are by default given an "index" dimension. For 2D or 3D variables, the second or third dimensions are defined and attached to the dataset, such as measurement time gates for time-domain AEM data channels, or frequencies for frequency-domain data channels. All data groups require spatial coordinate variables, standardized as "x" and "y". In the case of *Tabular* data, the coordinate variables match the size of the 1D index dimension and are sourced from corresponding input data variables, e.g., the longitude and latitude of data points, through the "key_mapping" attributes. In contrast, *Raster* datasets are gridded such that the dimensions of the data are also the coordinates (Figure 2C). A *Raster* group may contain multiple variables (e.g., total magnetic intensity and residual magnetic field) if all variables within the dataset share the same dimensions, otherwise separate *Raster* groups are encouraged (e.g., /survey/raster/0 and /survey/raster/1).

2.1.2 Coordinate reference systems

All datasets are required to have a defined coordinate reference system to maintain accurate representation of data values for both visualization and analysis purposes. Information about the coordinate system, such as a Well-known ID (WKID; Esri, 2016) and corresponding authority (e.g., EPSG), if it is geographic or projected, horizontal and vertical datums, and so on are stored within the *Survey* group's required "coordinate_information" variable. Any *Tabular* or *Raster* datasets attached to the *Survey* must have a matching variable "spatial_ref" and adhere to the same coordinate reference system. Following CF conventions (see section 5.6 of Eaton et al. (2020)), the "spatial_ref" coordinate variable must have the attribute "grid_mapping_name" which ties to a corresponding "grid_mapping" attribute within the data variables. Additionally, the "x" and "y" coordinate variables require certain attributes, such as "GeoX" and "GeoY" for "_CoordinateAxisType" which connects to a related key in the "spatial_ref" variable. If the coordinate system is a projection, then the "standard_name" keys for "x" and "y" should be "projection_x_coordinate" and "projection_y_coordinate". These details ensure that datasets are portable and accurately represented within geospatial systems (Eaton et al., 2020; Esri, 2022).

2.1.3 NcML

The last piece of the GS convention is the NetCDF eXtensible Markup Language (XML), NcML, metadata file, which is an XML representation of the metadata and group structure within the NetCDF file. NcML files are commonly used to allow simple updates or corrections to the metadata contained within NetCDF files (Nativi et al., 2005). For example, the Thematic Real-time Environmental Distributed Data Services (THREDDS) data server (TDS) employs NcML to define new NetCDF files, or augment and correct existing files hosted on their web service (Caron et al., 2006; Unidata, 2021c). NcML files also serve as a quick means for users to gain an overview of NetCDF file contents without needing to access the binary files. The NcML is not required to understand the data or metadata, but are an optional component that we recommend including when sharing or archiving GS NetCDF files.

2.2 GSPy v0.1.0

To implement this new GS data convention, we developed an open-source Python package, GSPy, which provides a basic toolkit to build, interface with, and export standardized geophysical datasets. GSPy utilizes the extensive Xarray Python package to assemble the GS groups and read/write the NetCDF files (Hoyer and Hamman, 2017). Xarray's architecture consists of DataArrays and Datasets. An Xarray DataArray is a labeled, multi-dimensional array containing 1) "data": an N-dimensional array of data values, 2) "coords": a dictionary container of the data coordinates, 3) "dims": the dimensions for each axis of the data array, and 4) "attrs": an attribute dictionary of key metadata (e.g., units, null values, descriptions) (Hoyer and Hamman, 2017). An Xarray Dataset is a collection of DataArrays, and similarly has the components of "dims" and "coords" which reflect those of the DataArrays (categorized as "data_vars" in the Dataset) and "attrs" for global metadata attributes that describe the collection. In the GS structure, each *Tabular* and *Raster* data group, as well as the *Survey* group, are individual Xarray Datasets. The data variables (DataArrays) within the *Survey* group's Dataset are unique in that they contain no data values, only variable attributes of *Survey* metadata information.

The GSPy package can be found at https://doi.org/10.5066/P9XNQVGQ, and requires Python version 3.5 or later (Foks et al., 2022). The software is platform independent (operates on both Windows and Unix operating systems) and has been released under the CC-0 license as per U.S. Geological Survey (USGS) software release policy. In this initial version, GSPy primarily serves as a data conversion tool, with functionality to interface with multiple input data formats and output to a GS-structured NetCDF file. Metadata is currently documented and input to GSPy through user-prepared JSON files.

2.2.1 Classes

GSPy contains *Survey* and *Data* classes, and the *Data* class is extended to the *Tabular* and *Raster* classes allowing for specific handling of those data types. The code requires a *Survey* object be instantiated as the first step to building a GS dataset (Figure 3A). A JSON metadata file (Figure 3B) is required to initialize the *Survey* object, where dictionaries such as "system_information," "survey_equipment," and "coordinate_information," for example, become data-less DataArray variables within the *Survey*'s Dataset, consisting primarily of metadata within the variable attributes. The required dictionary "dataset_attrs" populates the Dataset attributes, most of which follow the CF convention required inputs.

Each data assemblage, typically contained within a single tabular text file or a collection of related raster files, are attached to the established *Survey* object as the appropriate *Data* class using the "add_tabular" or "add_raster" methods of the *Survey* (Figure 4A, Figure 5A). Each instance of "add_tabular" and "add_raster" appends a new class object, *Tabular* or *Raster*, respectively, to the *Survey* with an incremented index for each location once written to disk, e.g., /survey/tabular/0, /survey/tabular/1, and /survey/tabular/2. The code is ignorant of any meaningful descriptions of data type, e.g., raw data vs. inverted models, and instead handles data purely based on the input format type and geometry. Therefore, it is up to the user to ensure the metadata—we recommend the "content" attribute field—provide sufficient description of what each Dataset within a *Survey* contains.

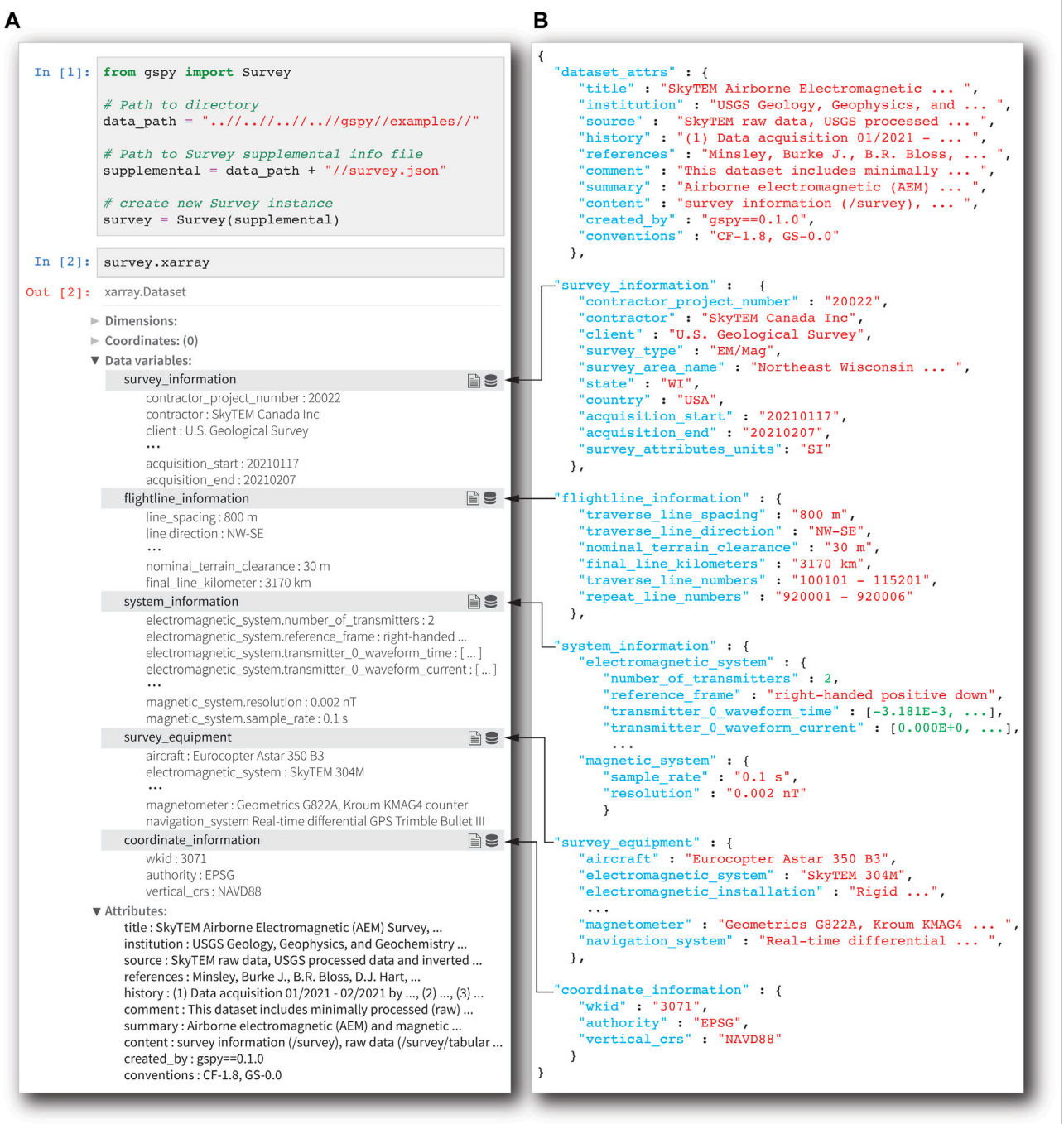

FIGURE 3
Starting a GS dataset with the GSPy package. **(A)** A GS dataset is always initialized with an instance of the *Survey* class. General metadata about the project and survey are read from a JSON file **(B)** and formatted into an Xarray Dataset attached to the *Survey* class object. The "dataset_attrs" dictionary contains required fields that populate the general attributes of the Xarray Dataset, while all other dictionaries become DataArray variables (black arrows).

GSPy version 0.1.0 supports CSV and ASEG-GDF2 text file formats for *Tabular* groups. For CSV files, a "variable_metadata" dictionary needs to be passed through the JSON file (Figure 4B). In contrast, ASEG-GDF2 files allow for variable attributes to be populated from the structured ASEG definition (.dfn) metadata file (Pratt, 2003). The "variable_metadata" dictionary can be optionally included for ASEG files to add or overwrite metadata values. For both input file types, GSPy executes one-to-one

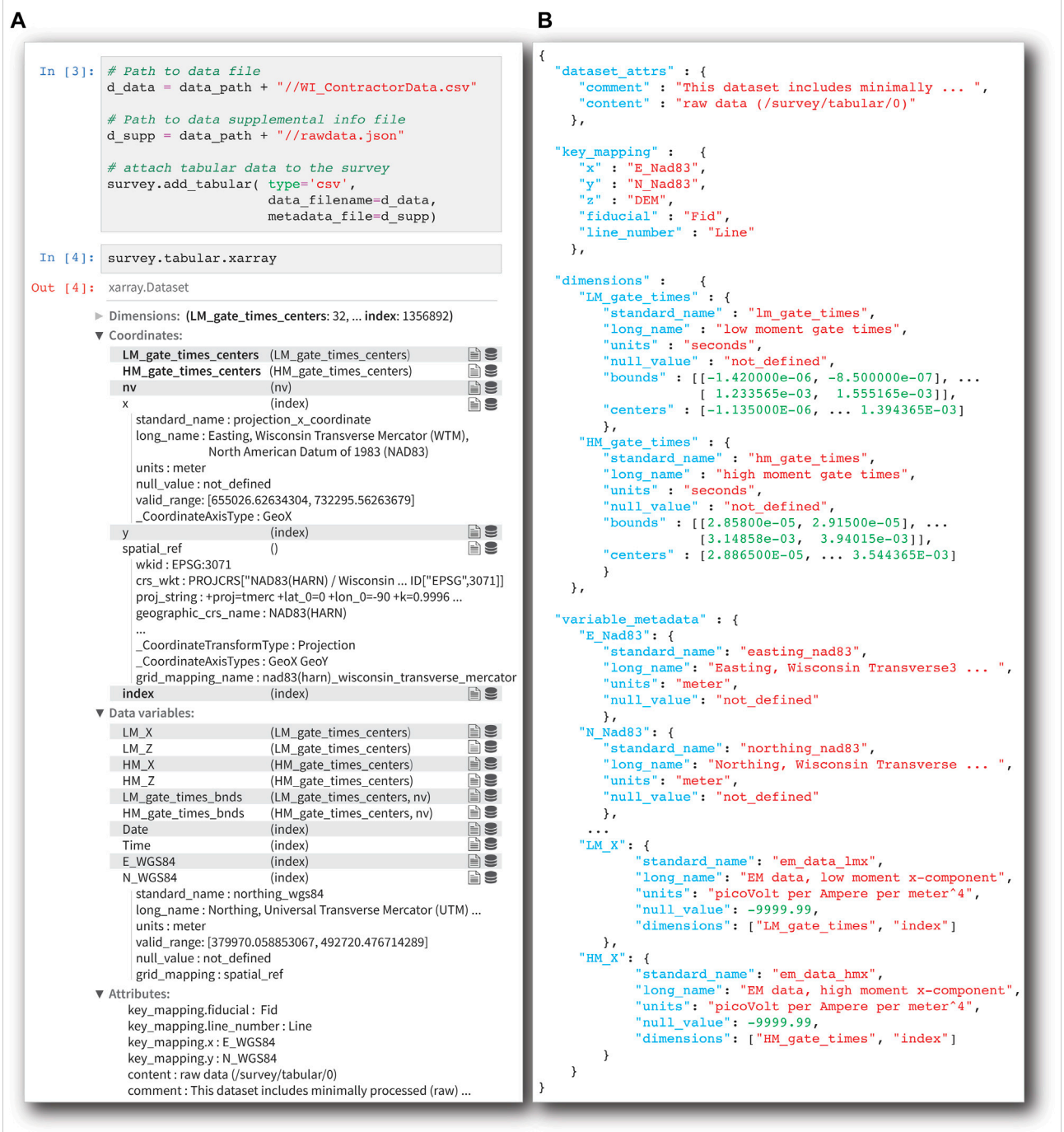

FIGURE 4
Attaching a *Tabular* group to a *Survey*. **(A)** The "add_tabular" method is used to add a *Tabular* Dataset to *Survey*. **(B)** The Dataset attributes, coordinate "key_mapping," variable-specific metadata, and any second-dimension variables are passed through a required JSON file when attaching the *Tabular* group.

mapping of columns into DataArrays by default. Variables that comprise multiple columns are handled in one of two ways. First, if the columns contain an incrementor following the variable name formatted as [0], [1], . . . [N] for N number of columns—a common format for geophysical datasets—then the columns are concatenated in order and labeled by the root column name. For example, a time-domain AEM variable that appears as "EMX_HPRG [0]", "EMX_HPRG [1]", "EMX_HPRG [2]" etc.

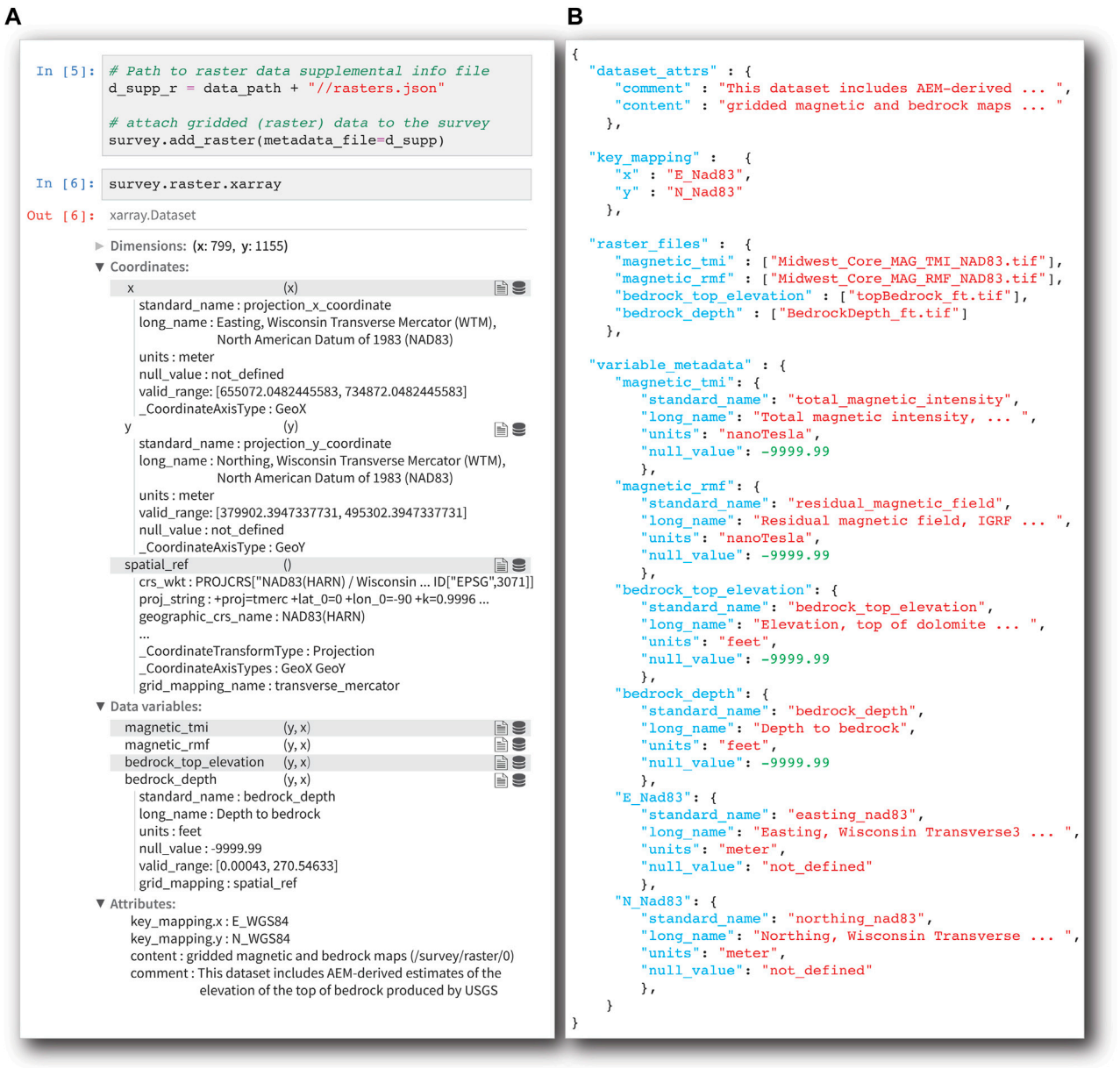

FIGURE 5
Attaching a *Raster* group to a *Survey*. **(A)** The "add_raster" method is used to add a *Raster* Dataset to *Survey*. **(B)** As with *Tabular* groups, the Dataset attributes, coordinate "key_mapping", and variable-specific metadata are passed through a required JSON file. The *Raster* class allows for a one-to-one mapping of GeoTIFF files to DataArray variables within the Dataset, or multiple files can be stacked into a single variable. The "raster_files" dictionary maps the desired variables to its input file(s).

within the contractor-provided data file would be combined into a 2D DataArray variable, "EMX_HPRG", within the GSPy Dataset. In this example, the dimensions of the "EMX_HPRG" variable would be "index" and "gate_times." The "gate_times" dimension values and metadata are also defined through the JSON file in the "dimensions" dictionary (Figure 4B). For *Tabular* groups, users also have the option to provide bounds on dimensions, when appropriate, such as the start and end times for each time gate. We follow the CF conventions' approach to bounding variables, such that a rank 1 dimension of length N will have bounds of shape (N, 2), where each value along the first axis has 2 vertices corresponding to its bounds (Figure 4).

The second approach to multi-dimensional column variables is to pass a "raw_data_columns" key within the "variable_metadata" dictionary of the JSON for the desired

output variable name, where the values of "raw_data_columns" points to the original column names in the data file in the order they should be concatenated. For example, a frequency-domain AEM variable for in-phase filtered data can often appear in the raw data file with unique columns named by frequency, such as "cpi400_filt", "cpi1800_filt", etc. A sorted list of these data columns should be passed through the metadata of a new variable, such as "ip_filtered," which would have the dimensions "index" and "frequency." As before, the "frequency" dimension would be defined and described through a "dimensions" dictionary. Coordinates for *Tabular* data are defined through the "key_mapping" dictionary of the JSON. As stated previously, *Tabular* variables have coordinates of dimension "index" and the "key_mapping" allows GSPy to create the coordinate variables based on named input variables, e.g., {"x": "Longitude", "y": "Latitude"}.

GSPy v0.1.0 supports GeoTIFF files as the primary input/output format for *Raster* groups. In contrast to *Tabular* groups, variables are added either as 2D variables from single GeoTIFF files (1 file = 1 DataArray) or 3D variables by stacking multiple files along a named dimension (e.g., individual depth slices). In the JSON metadata file, the "raster_files" dictionary maps each DataArray variable to a file or list of files. As before, a "variable_metadata" dictionary is needed to complete the attributes of each variable. The dimensions of the data are by default the coordinates defined by the input file, thus no "dimensions" dictionary is needed. The "key_mapping" dictionary is still needed for *Raster* datasets to update the metadata of the dimension coordinates ("x" and "y"). We use the Rioxarray module (http://github.com/corteva/rioxarray) to go between GeoTIFF files and Xarray DataArrays. Upon reading in a GeoTIFF file, GSPy compares the input coordinate reference system with that of the *Survey*. If the input reference system does not match, the DataArray is reprojected using Rioxarray. Future versions can follow the same procedures for other standard raster data file formats.

For all data types regardless of geometry (*Tabular* and *Raster*), the JSON metadata file is required to contain a "dataset_attrs" dictionary, which populates the attributes of the Dataset. Since data groups are contained within the *Survey* group of the NetCDF file, the globally required attributes of the *Survey* apply to all data groups, per CF conventions (Eaton et al., 2020). Therefore, the attributes of data groups only require the "content" key and any "key_mapping", with additional keys such as "comment" optionally included at user-discretion. Lastly, the coordinate reference system of the *Survey* is used to create the "spatial_ref" coordinate variable to accompany each Dataset, thereby requiring all groups under a *Survey* to have matching coordinate systems. Either a well-known identification (WKID) number and associated authority, e.g., EPSG:4326, or a coordinate reference system well-known text (CRS_WKT) string are needed to then generate the complete "spatial_ref" variable using the GDAL and Pyproj packages (GDAL/OGR contributors, 2022; https://github.com/pyproj4/pyproj). We follow CF conventions and ArcGIS guidelines (e.g., Esri, 2022) to ensure proper transferability of datasets into common geospatial and NetCDF-supported software.

2.2.2 Class properties and methods

GSPy provides many helpful properties and methods for working with datasets. Here we highlight some essential functions, and refer readers to the GSPy documentation pages for a complete description of all classes, methods, and functionality, along with code examples (Foks et al., 2022).

First, all classes share the property of "xarray" to return the GS-formatted Xarray Dataset (Figures 3, 4, 5A). The "read_metadatafile" method is common to each class and attaches the full dictionary read from the provided JSON file to the property "json_metadata". If a metadata file does not get passed or is missing required dictionaries, the "write_metadata_template" method is called to generate a template file that users can then edit. This function is useful for large CSV datasets with many variables, as it will generate a "variable_metadata" dictionary based on the column names. All attributes are given "not_defined" values that users can then update.

Once all groups have been attached to a *Survey*, the "write_netcdf" and "write_ncml" methods will write the GS-structured NetCDF file and accompanying NcML file, respectively (Figure 6). The data classes, *Tabular* and *Raster*, also contain "write_netcdf" functions to export groups in separate files; however, we recommend always using the *Survey* class "write_netcdf" function to adhere to the standard with all groups written to a single file. The *Tabular* and *Raster* classes also contain export methods such as "to_csv" and "to_tif", respectively. Lastly, some simple plotting methods are provided for both *Raster* and *Tabular* classes using Xarray's scatter and pcolor functions (Figure 6).

3 Results

To demonstrate the proposed GS convention and the functionality provided by GSPy, we converted a recently acquired airborne geophysical dataset into the new standard through GSPy workflows. This dataset provides the opportunity to showcase examples of diverse input data formats (CSV and GeoTIFF) and geometries (*Tabular* and *Raster*) within the proposed GS architecture. In January and February 2021, the U.S. Geological Survey oversaw collection of 3,170 line kilometers of AEM and magnetic data over northeast Wisconsin through collaboration with the Wisconsin Department of Agriculture, Trade, and Consumer Protection (DATCP) and Wisconsin Geological and Natural History Survey (WGNHS) (Minsley et al., 2022). The primary purpose of this effort was to improve understanding of the depth to bedrock

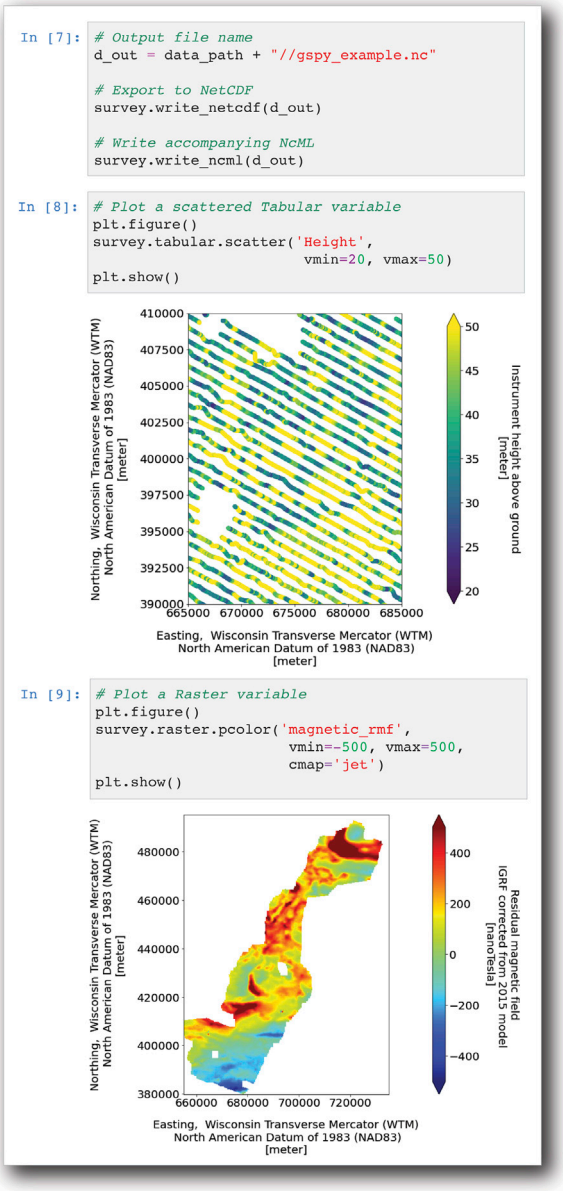

FIGURE 6
Writing and plotting examples. Once all groups have been attached to a *Survey*, the "write_netcdf" and "write_ncml" methods will write the GS NetCDF and NcML files, respectively. GSPy also provides methods to generate scatter and pcolor plots for variables.

across the study area. The airborne data were acquired by SkyTEM Canada Inc. with the SkyTEM 304M time-domain helicopter-borne electromagnetic system together with a Geometrics G822A cesium vapor magnetometer.

Input data consisted of 1) a CSV file (3.17 GB) of contractor-provided raw AEM and magnetic data along with auxiliary flight data; 2) a CSV file (123.9 MB) of processed AEM data; 3) a CSV file (145.3 MB) of inverted resistivity models; 4) a CSV file (4.4 MB) of AEM-derived point estimates of the elevation of the top of bedrock; and 5) four GeoTIFF files containing gridded magnetic data (total magnetic intensity: 7.4 MB, residual magnetic field: 7.4 MB), AEM-derived gridded depth to the top of bedrock (3.7 MB) and top of bedrock elevation (3.7 MB). We created JSON files for the *Survey* group, pulling critical information on the flightlines, system parameters, and equipment from the contractor-provided report (Figure 3B). Each CSV file was added as a separate *Tabular* group with an individual JSON metadata file (e.g., Figure 4B). The four GeoTIFF files were added as variables within a single *Raster* group with an accompanying JSON file (Figure 5B).

With the GSPy workflow, and proper documentation in the JSON files, all datasets were assembled under the *Survey* group with complete dataset- and variable-specific metadata, mapping of dimensions and coordinates, and standardized coordinate reference systems variables. We then used GSPy methods to export the combined datasets into a single NetCDF file and generate the NcML metadata file. Figure 7 shows a simplified version of the NcML file, with essential elements represented. Both NetCDF and NcML files were publicly released in ScienceBase (Minsley et al., 2022). The size of the final GS NetCDF file was 1.93 GB, corresponding to a file size reduction of 44% relative to the original input files without utilizing further compression. The complete file and its contents were accurately imported into common NetCDF software such as Unidata's Integrated Data Viewer (IDV) (Unidata, 2021a). *Raster* variables from the full GS NetCDF file were accurately imported into Quantum Geographic Information System (QGIS), with correct placement, coordinate reference system, and null value representation. ArcMap was unable to import the full NetCDF file comprising multiple groups, but datasets exported to individual files, at the root group position, were accurately imported. Notably, both scattered *Tabular* data and gridded *Raster* data were successfully viewed in ArcMap, but we were unable to view scattered data in QGIS.

4 Discussion

The GS data convention improves the accessibility and functionality of geophysical datasets by providing much-needed standards for the storage of both data and metadata built on the established NetCDF open data structure and existing CF conventions. By building on the NetCDF CF conventions, the GS model has several advantageous characteristics summarized earlier: it is self-describing, space-saving, accessible, portable, scalable, and hierarchical. Most importantly, the GS model allows multiple types of geophysical data and incremental data processing steps to be stored together in a single self-

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2" location="gspy_example.nc">
<group name="/survey">
  <attribute name="title" value="SkyTEM Airborne Electromagnetic (AEM) Survey, Northeast ... "/>
  ...
  <attribute name="conventions" value="CF-1.8, GS-0.0"/>
  <variable name="coordinate_information" shape="" type="Float">
    <attribute name="wkid" type="String" value="3071"/>
    <attribute name="vertical_crs" type="String" value="NAVD88"/>
  </variable>
  ...
  <group name="/tabular">
    <group name="/0">
    <dimension name="HM_gate_times_centers" length="32"/>
    <dimension name="LM_gate_times_centers" length="28"/>
    <dimension name="nv" length="2"/>
    <dimension name="index" length="1356892"/>
    <attribute name="content" value="raw data"/>
    ...
    <attribute name="comment" value="This dataset includes minimally processed (raw) AEM ... "/>
    <variable name="HM_gate_times_centers" shape="HM_gate_times_centers" type="Float">
      <attribute name="standard_name" type="String" value="hm_gate_times_centers"/>
      <attribute name="long_name" type="String" value="high moment gate times centers"/>
      <attribute name="units" type="String" value="seconds"/>
      <attribute name="null_value" type="String" value="not_defined"/>
      <attribute name="bounds" type="String" value="HM_gate_times_bnds"/>
    </variable>
    ...
    </group>
  </group>
  <group name="/raster">
    <group name="/0">
    <dimension name="x" length="799"/>
    <dimension name="y" length="1155"/>
    <attribute name="content" value="gridded magnetic and bedrock maps"/>
    ...
    <attribute name="key_mapping.y" value="N_Nad83"/>
    <variable name="x" shape="x" type="Float">
      <attribute name="standard_name" type="String" value="easting_nad83"/>
      <attribute name="long_name" type="String" value="Easting, Wisconsin Transverse Mercator ... "/>
      <attribute name="units" type="String" value="meter"/>
      <attribute name="null_value" type="String" value="not_defined"/>
    </variable>
    ...
    <variable name="bedrock_depth" shape="y x" type="Float">
      <attribute name="standard_name" type="String" value="bedrock_depth"/>
      <attribute name="long_name" type="String" value="Depth to bedrock"/>
      <attribute name="units" type="String" value="foot"/>
      <attribute name="null_value" type="String" value="-9999.99"/>
    </variable>
    </group>
  </group>
</group>
</netcdf>
```

FIGURE 7
Example NcML file. Due to space constraints, only essential elements are shown here for example representations. Gaps in variable and attribute lists are noted by ellipses.

described file with all variable-specific and general survey metadata attached. Our support for unstructured point data (tabular) within the NetCDF is particularly novel, as both historical and modern implementations of the NetCDF format have dominantly been for gridded (raster) datasets (e.g., Hankin et al., 2010; Eaton et al., 2020; Morim et al., 2020). These characteristics are important for both the long-term accessibility and interoperability of geophysical datasets. While

our focus here is on airborne geophysical surveys, this data model can be readily extended to other survey data that can be described in tabular or raster formats.

Application of the GSPy workflow and GS data model to a real airborne geophysical dataset resulted in several successful outcomes and insights. First, what began as several disconnected, undocumented, and uniquely formatted data files became a single NetCDF file with related, self-described datasets clearly categorized and standardized. This improved the shareability and usability of the data, as every variable and dataset group were fully documented and easily accessed within the single file. Second, the NetCDF file, and its accompanying NcML file, was all that was needed to be archived for public release. This resulted in a significantly simplified data release process, i.e., file preparation, metadata documentation, and the review process were all streamlined compared to a traditional release of the original data files and incomplete metadata documentation. Lastly, the standardized datasets within the NetCDF file were accurately viewed and represented within common NetCDF and GIS software, signifying the broad transferability and interoperability of the GS format.

We recognize the aforementioned advantages of using the NetCDF file structure also comes with some challenges. Accessing information in binary NetCDF files may be a barrier for users not familiar with this format, especially compared with ASCII-based file formats. The accompanying GSPy software tools include methods for exporting to common tabular or raster formats if those are needed for specific end-users. Additionally, raising awareness about common GIS or other software tools that can read NetCDF files, along with their current limitations, will be important. Preparing the JSON metadata files can be time consuming, but once prepared executing the GSPy workflow is straightforward and efficient. Furthermore, datasets being published in an open repository would need much of the same metadata information, prepared here in JSON input files, to instead be produced in XML or other online metadata records. Thus, we recognize that documentation of metadata can be a tedious endeavour but a necessary one nevertheless. While accessibility and ease-of-use need to be continually improved upon, such as changing to a slightly more user-friendly metadata input format like Yet Another Markup Language (YAML), for example, the additional complexity of the GS convention is outweighed by its broader advantages discussed above. Upfront time costs with the GSPy workflow will likely balance out with time savings during archival, as well as improve overall dataset usability and impact.

The first version of GSPy has focused on an implementation of the GS data model for airborne geophysical data; however, we have developed the software, data classes, and functions with the intention of being generalized and adaptable to all types of geophysical methods. We plan to layer new functionality for ground-based and airborne geophysical data alike in future versions, such as method-specific converters for ground resistivity data and models or seismic timeseries. A guiding principle is to build a strong foundation for the data standard and software tools that can be readily extended to other datatypes without changing the basic structure. Any number or type of classes can be attached as groups within the hierarchical NetCDF file structure, always falling under a general metadata *Survey* group. Most geophysical datasets and related products can be described by the generic *Tabular* or *Raster* classes, and additional classes can be developed as needs are identified. By developing GSPy as an open-source package, our goal is to enable a broad community of users to improve its functionality and capabilities.

New GSPy functionality is planned for future versions to simplify import and export workflows, such as automatically recognizing different datatypes and routing to customized methods that handle different datatype requirements. Support for other data formats and software interfaces is also planned, for example leveraging existing packages such as gxpy (https://github.com/GeosoftInc/gxpy) to directly import data from commonly used binary Geosoft databases and sciencebasepy (https://github.com/usgs/sciencebasepy) to automate the publication process to the USGS ScienceBase repository. Accessibility can also be broadened by including documentation and links in future versions for common software programs that can read GS-structured NetCDF files.

Additional worked examples of other airborne geophysical datasets and data types are needed to continue refining the structural details of how data and metadata are imported to and stored in the GS data model. For example, identifying and revising required versus recommended versus optional attributes and variables, defining generic and adaptable structures for storing *Survey* metadata information, and standardizing JSON templates for various data types will improve the overall usability of the data standard. Future GSPy functionality can also be added to aid in data processing and visualization—eventually with GSPy serving as a central platform for importing datasets, processing, exploring, reformatting, interfacing with various inversion software, and exporting in a standardized format for public release. Additionally, we plan to explore the use of web-based tools such as the THREDDS Data Server (Caron et al., 2006; Unidata, 2021c) for accessing and subsetting content from GS-structured files stored in online repositories, without needing to download entire datasets.

If adopted as a common standard for geophysical datasets, further efficiency could be realized by having instruments or contractor-delivered datasets directly create GS-structured files, or at least the information needed to readily create them. Likewise, processing, visualization, and inversion software tools could directly read files in the GS convention without having to export other specialized input formats. For example, the study presented in this paper required multiple file format conversion steps throughout the workflow: contractor-provided databases and PDF reports, processed data, inverted geophysical models, and

bedrock elevation picks were all exported from proprietary software tools into CSV and JSON formats to prepare them for publication in open formats. Significant improvements in workflow efficiency and interoperability can be achieved by using the GS convention as a link that connects instrument-recorded data and metadata to processing, visualization, and interpretation tools as well as archival-ready data structures.

5 Conclusion

The field of geophysics encompasses diverse and complex data formats that can vary between methods, techniques, and from one collection to another. Inconsistencies in data and metadata documentation reduce the longevity and impact of geophysical datasets. To address the pressing need for a community-supported geophysical data standard, we have developed the GS convention, based on the NetCDF file format and CF metadata conventions. The GS convention meets the goals we set out to achieve in a geophysical data standard:

- The format is open source meeting the requirements of FAIR data publication standards.
- The file format allows for multiple related and self-described datasets to be grouped together under a clear and standardized hierarchical structure.
- Dataset- and variable-specific attributes join important auxiliary information and metadata directly to the digital data, ensuring dataset integrity, longevity, and interoperability.
- Data dimensions and coordinates are clearly defined, along with a well-defined coordinate reference system for accurate visualization and representation.
- The format is transferable between open-source computational software, web services, and geospatial systems.

The accompanying open-source Python package, GSPy, facilitates efficient data conversion between common data formats (e.g., CSV, ASEG-GDF2, GeoTIFF), proper metadata documentation through JSON supporting files, and export of GS NetCDF files. We demonstrated the GS structure and GSPy workflow using an example airborne geophysical dataset from Wisconsin. The single resulting GS NetCDF file was significantly reduced in size compared to the multiple ASCII-text and GeoTIFF input files. Furthermore, metadata that was previously distributed throughout a contractor-provided PDF report was cleanly incorporated and appropriately attached to specific dataset groups and variables. Aside from a few limitations identified, such as the group structure in ArcMap or scattered data in QGIS, the GS-formatted file and/or individual data groups were successfully loaded and accurately represented in geospatial software.

Adoption of the GS standard for airborne geophysical data fills a particular need for an open-source, community-wide standard that ensures accurate archival of critical metadata jointly with digital datasets. Moreover, establishment of a NetCDF-based open data standard for a broad range of geophysical survey types can help to greatly improve how these complex datasets are shared and utilized, making the data more accessible to a broader science community and the public. File formats and functionality supported by GSPy v0.1.0 is limited; however, by developing the standard and package as open source, we aim to leverage the broad geophysical community to contribute to the continued development of robust data standard requirements and tools to facilitate their use.

Data availability statement

The dataset used in this work to demonstrate the GSPy code implementation and GS convention can be found on the U.S. Geological Survey's data repository, ScienceBase, located at https://doi.org/10.5066/P93SY9LI. The GSPy code repository is located at https://doi.org/10.5066/P9XNQVGQ.

Author contributions

BM led the conceptualization and overall design of the geophysical data standard. NF and SJ developed the GSPy software and worked with BM on designing the GS structure and its practical implementation. SJ led the writing effort for this article, which was contributed to by all authors.

Funding

This work was jointly supported by the USGS Water Availability and Use Science Program and the USGS Mineral Resources Program.

Acknowledgments

The authors thank JR Rigby (USGS) for supporting this effort. We also thank Bennett Hoogenboom (USGS) for helpful comments and review of the software, as well as Jade Crosbie (USGS), Andrew Leaf (USGS), and Urminder Singh (Iowa State University) for their helpful reviews of the manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of interest

Author NF is employed by Apogee Engineering LLC as a contractor to the U.S. Geological Survey.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Brodie, R. (2017). ga-aem: Modelling and inversion of airborne electromagnetic (AEM) data in 1D. *Geosci. Aust.* Available at: https://github.com/GeoscienceAustralia/ga-aem.

Caron, J., Davis, E., Ho, Y., and Kambic, R. (2006). "UNIDATA's THREDDS data server," in 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology.

Dampney, C. N. G., Pilkington, G., and Pratt, D. A. (1985). ASEG-GDF: The ASEG standard for digital transfer of geophysical data. *Explor. Geophys.* 16, 123–138. doi:10.1071/EG985123

Drenth, B. J., and Brown, P. J. (2020). *Airborne magnetic survey, iron mountain-chatham region, central upper peninsula, Michigan, 2018*. Reston, VA: U.S. Geol. Surv. data release. doi:10.5066/P91EF3CI

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2020). *NetCDF Climate and Forecast (CF) metadata conventions version 1.8*. Available at: https://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.html.

Esri (2016). *Faq: What does "authority: EPSG" mean in an ArcGIS desktop .prj file?* Esri Tech. Support. Available at: https://support.esri.com/en/technical-article/000011199 (Accessed May 23, 2022).

Esri (2022). Spatial reference for netCDF data. ArcGIS Deskt. - ArcMap 10.8. Available at: https://desktop.arcgis.com/en/arcmap/latest/manage-data/netcdf/spatial-reference-for-netcdf-data.htm (Accessed May 19, 2022).

Foks, N. L., James, S. R., and Minsley, B. J. (2022). *GSPy: Geophysical data standard in Python*. U.S. Geol. Surv. softw. release. doi:10.5066/P9XNQVGQ

Folk, M., McGrath, R. E., and Yeager, N. (1999). Hdf: An update and future directions. in IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No. 99CH36293), 273–275.

GDAL/OGR contributors (2022). *{GDAL/OGR} geospatial data abstraction software library*. doi:10.5281/zenodo.5884351

Hagelund, R., and Levin, S. A. (2017). SEG-Y revision 2.0 data exchange format. *Soc. Explor. Geophys. Houst.*

Hankin, S. C., Blower, J. D., Carval, T., Casey, K. S., Donlon, C., Lauret, O., et al. (2010). "NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes," in Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Venice, Italy, 21-25 September 2009. Editors HallJ, D. E. Harrison, and D. Stammer (Venice, Italy: ESA Publication WPP-306). doi:10.5270/OceanObs09.cwp.41Vol. 2

Hoyer, S., and Hamman, J. (2017). xarray: ND labeled arrays and datasets in Python. *J. Open Res. Softw.* 5, 10. doi:10.5334/jors.148

Ley-Cooper, Y., Roach, I., and Brodie, R. C. (2019). Geological insights of Northern Australia's AusAEM airborne EM survey. *ASEG Ext. Abstr.*, 1–4. doi:10.1080/22020586.2019.12073170

Li, G., and Choi, Y. (2021). HPC cluster-based user-defined data integration platform for deep learning in geoscience applications. *Comput. Geosci.* 155, 104868. doi:10.1016/j.cageo.2021.104868

Minsley, B. J., Bloss, B. R., Hart, D. J., Fitzpatrick, W., Muldoon, M. A., Stewart, E. K., et al. (2022). *Airborne electromagnetic and magnetic survey data, northeast Wisconsin (ver. 1.1, June 2022)*. Reston, VA: U.S. Geol. Surv. data release. doi:10.5066/P93SY9LI

Minsley, B. J., James, S. R., Bedrosian, P. A., Pace, M. D., Hoogenboom, B. E., and Burton, B. L. (2021). *Airborne electromagnetic, magnetic, and radiometric survey of the Mississippi Alluvial Plain*. Reston, VA: U.S. Geol. Surv. data release. doi:10.5066/P9E44CTQ

Møller, I., Søndergaard, V. H., Jørgensen, F., Auken, E., and Christiansen, A. V. (2009). Integrated management and utilization of hydrogeophysical data on a national scale. *Near Surf. Geophys.* 7, 647–659. doi:10.3997/1873-0604.2009031

Morim, J., Trenham, C., Hemer, M., Wang, X. L., Mori, N., Casas-Prat, M., et al. (2020). A global ensemble of ocean wave climate projections from CMIP5-driven models. *Sci. Data* 7, 105. doi:10.1038/s41597-020-0446-2

NASA (2019). *CDF user's guide, version 3.8.0. Sp. Phys. Data facil. NASA/goddard sp. Flight cent., 1–164.* Available at: https://spdf.gsfc.nasa.gov/pub/software/cdf/doc/cdf380/cdf380ug.pdf.

Nativi, S., Caron, J., Davis, E., and Domenico, B. (2005). Design and implementation of netCDF markup language (NcML) and its GML-based extension (NcML-GML). *Comput. Geosci.* 31, 1104–1118. doi:10.1016/j.cageo.2004.12.006

NOAA (1995). Cooperative Ocean/Atmosphere research data service. *Natl. Ocean. Atmos. Adm.* Available at: https://ferret.pmel.noaa.gov/Ferret/documentation/coards-netcdf-conventions.

Northwood, E. J., Weisinger, R. C., and Bradlei, J. J. (1967). Recommended standards for digital tape formats. *Geophysics* 32, 1073–1084. doi:10.1190/1.32060004.1

Pratt, D. A. (2003). ASEG-GDF2 A standard for point located data exchange. *Aust. Soc. Explor. Geophys.* 4, 1–34. Available at: https://www.aseg.org.au/sites/default/files/pdf/ASEG-GDF2-REV4.pdf.

Ramapriyan, H. K., and Leonard, P. J. T. (2021). Data product development guide (DPDG) for data producers version 1.1. *NASA Earth Sci. Data Inf. Syst. Stand. Off.* doi:10.5067/DOC/ESO/RFC-041VERSION1

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 195–204. doi:10.1038/s41586-019-0912-1

Rew, R., and Davis, G. (1990). NetCDF: An interface for scientific data access. *IEEE Comput. Graph. Appl.* 10, 76–82. doi:10.1109/38.56302

Rew, R., Hartnett, E., and Caron, J. (2006). "NetCDF-4: Software implementing an enhanced data model for the geosciences," in 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanograph, and Hydrology.

Salman, M., Slater, L., Briggs, M., and Li, L. (2022). Near-surface geophysics perspectives on integrated, coordinated, open, networked (ICON) science. *Earth Space Sci.* 9, e2021EA002140. doi:10.1029/2021EA002140

Shah, A. K. (2020). *Airborne magnetic and radiometric survey, Charleston, South Carolina and surrounds, 2019*. Reston, VA: U.S. Geol. Surv. data release. doi:10.5066/P9EWQ08L

Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., and Skakun, S. (2017). Exploring google earth engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping. *Front. Earth Sci.* 5. doi:10.3389/feart.2017.00017

Unidata (2021a). *Integrated data viewer (IDV) version 6.0*. BoulderCent: CO UCAR/Unidata Progr. doi:10.5065/D6RN35XM

Unidata (2021b). *Network common data form (netCDF)*. BoulderCent: CO UCAR/Unidata Progr. version 4.8.1. doi:10.5065/D6H70CW6

Unidata (2021c). *THREDDS data server (TDS) version 5.3*. BoulderCent: CO UCAR/Unidata Progr. doi:10.5065/D6N014KG

Vermeesch, P., and Garzanti, E. (2015). Making geological sense of 'Big Data' in sedimentary provenance analysis. *Chem. Geol.* 409, 20–27. doi:10.1016/j.chemgeo.2015.05.004

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. data* 3, 160018. doi:10.1038/sdata.2016.18

Yang, M., McGrath, R. E., and Folk, M. (2005). "HDF5-a high performance data format for Earth science," in Proceedings of the International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology.

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/feart.2022.975156

OPEN ACCESS

EDITED BY
Michael Fienen,
United States Geological Survey,
United States

REVIEWED BY
Matteo Camporese,
University of Padua, Italy
Howard W. Reeves,
United States Geological Survey,
United States

*CORRESPONDENCE
Alexandre Pryet,
alexandre.pryet@bordeaux-inp.fr

SPECIALTY SECTION
This article was submitted to Hydrosphere,
a section of the journal
Frontiers in Earth Science

RECEIVED 21 June 2022
ACCEPTED 29 July 2022
PUBLISHED 25 August 2022

CITATION
Pryet A, Matran P, Cousquer Y and Roubinet D (2022), Particle tracking as a vulnerability assessment tool for drinking water production.
Front. Earth Sci. 10:975156.
doi: 10.3389/feart.2022.975156

Particle tracking as a vulnerability assessment tool for drinking water production

Alexandre Pryet[1]*, Pierre Matran[1], Yohann Cousquer[2] and Delphine Roubinet[3]

[1]EPOC (UMR 5805), CNRS, Univ. Bordeaux and Bordeaux INP, Pessac, France, [2]HSM, Univ. Montpellier, CNRS, IMT, IRD, Montpellier, France, [3]Geosciences Montpellier (UMR 5243), CNRS, Univ. Montpellier, Montpellier, France

The simulation of concentration values and use of such data for history-matching is often impeded by the computation time of groundwater transport models based on the resolution of the advection-dispersion equation. This is unfortunate because such data are often rich in information and the prediction of concentration values is of great interest for decision making. Particle tracking can be used as an efficient alternative under a series of simplifying assumptions, which are often reasonable at groundwater sinks (wells and drains). Our approach consists of seeding particles around a sink and tracking particles backward, up to the source boundary condition, such as a contaminated stream. This particle tracking approach allows the use of parameter estimation and optimization methods requiring numerous model calls. We present a Python module facilitating the pre- and post-processing operations of a modeling workflow based on the widely used USGS MODFLOW6 and MODPATH7 programs. The module handles particle seeding around the sink and estimation of the mixing ratio of water withdrawn from the sink. This ratio is computed with a mixing law from the particle endpoints, accounting for particle velocities and mixing in the source model cells. We investigate the best practice to obtain robust derivatives with this approach, which is a benefit for the screening methods based on linear analysis. We illustrate the interest of the approach with a real world case study, considering a drinking water well field vulnerable to a contaminated stream. The configuration is typical of many other drinking water production sites. The modeling workflow is fully script-based to make the approach easily reproducible in similar cases.

KEYWORDS

particle-tracking, advective transport, steady state, surrogate model, groundwater contamination, stream-aquifer flow, well vulnerability

1 Introduction

Drinking water contamination is a major matter of concern for worldwide public health, with worldwide economic and social effects (Daud et al., 2017; Sharma and Bhattacharya, 2017; Gwimbi et al., 2019; Turner et al., 2021). This emphasizes the need for decision-support modeling tools that could provide 1) reliable predictions of the risk associated with water contamination in prospective scenarios and 2) optimized production settings to mitigate the impact of reported contamination. To this end, models may be of great interest provided a series of conditions are satisfied (Doherty, 2015; Hermans, 2017; Doherty and Moore, 2020). First, that models should properly account for the processes governing the outputs of interest. Second, that an appropriate and reliable observational dataset constrains the parameters of importance. Third, that the model is "practical", which implies reasonable computation time and effective tools to automatize time consuming operations (Bakker et al., 2016; White et al., 2020a). Such conditions may not be easy to fulfill, in particular when dealing with the complex process of contaminant transport in heterogeneous aquifers.

In such a context, a thorough analysis of the simplicity-complexity trade-off becomes a critical step for which guidelines are now available (Hrachowitz et al., 2014; Guthke, 2017; Schwartz et al., 2017; Hugman and Doherty, 2022). In practice, appropriate choices have to be made on 1) the parameterization of hydraulic and transport properties and 2) the simulated physio-chemical processes driving the propagation of contaminant. For the parameterization of hydraulic properties, the options range from homogeneous equivalent hydraulic conductivity zones to heterogeneous fields with discrete features (de Marsily et al., 2005; Carniato et al., 2015; Pool et al., 2015). For the selection of processes governing transport, the options range from simplified models that rely on assumptions such as steady-state and dominant processes (e.g., advection), to complex, advection-dispersion reactive transport models to provide a detailed description of the physics of the problem (Anderson et al., 2015). Several studies highlighted that the best balance in terms of robustness, efficiency and reliability may be achieved with relatively simple "surrogate" models based on simplified representations of the physics (Razavi et al., 2012; Asher et al., 2015; Burrows and Doherty, 2015). Surrogate transport models present fast run times allowing for thousands of model executions of transport calculations, which are necessary for history matching of highly parameterized models. The interest of including a diverse observational data types has been highlighted by Hunt et al. (2006) and concentration in particular has been recently discussed by Schilling et al. (2019) and Knowling et al. (2020). Furthermore, fast run time allows the use of more robust but demanding algorithms for uncertainty quantification (Rajabi et al., 2018).

During the last decades, parameter estimation and uncertainty quantification algorithms have been made available with an ever growing variety of approaches (Doherty, 2016; White et al., 2020b). Their use is now facilitated by Python interfaces (White et al., 2016). More recently, studies provided repeatable script-based workflows which greatly facilitate the replication of the presented approaches (White et al., 2020a; Fienen et al., 2022). Though facilitated, the interfacing of parameter estimation and uncertainty quantification algorithms with complex models remains difficult for transport models, which are characterized by long computation times and numerical instabilities.

Focusing on the widely reported risk of contaminant transfer from contaminated rivers to groundwater production units, we present a framework based on particle tracking as a fast and effective surrogate model for contaminant transport. The approach, initially presented by (Cousquer et al., 2018), is made available in a newly developed Python module, TrackTools which will facilitate its replication. The module provides particle seeding capabilities and post-processing options that are valuable for analyzing drinking water vulnerability of production wells or drains to pre-defined sources of contamination. The script facilitates exploratory parametric analysis, which can be useful to investigate the system response to different configurations. The interfacing with the PEST + + suite (White et al., 2020b) is detailed on a real-world case study which paves the way for history matching, hypothesis testing, and optimization of decision variables, which are essential to support decision related to the definition of wellhead protection area and the optimization of production settings.

The theoretical background and numerical tools related to the developed Python module are described in Section 2. A simple synthetic model is presented in Section 3 and a parametric study is conducted to investigate the driving factors of mixing ratios computed with this approach. In Section 4, the interest of the method is illustrated on a drinking water production site with a fully script-based approach from model setup to pumping optimization through parameter estimation.

2 Methodology

A common practice in vulnerability analysis is to investigate the origin of water withdrawn at a groundwater sink (well or drain) originating from one, or a series of potential or effective contaminant sources. This can be conducted from the analysis of the flow contributions of each source to the discharge rate of the sink. The methodology described hereafter is an extension of the approach described by Cousquer et al. (2018) initially designed for a single river reach. The method may now be applicable to

multiple weak or strong source boundary conditions (e.g., fixed head, river, general head boundary condition).

Consider a groundwater flow model with a groundwater sink (well or drain) subject to a potential are effective contamination originating from one of the model boundary conditions. A set of N starting points is seeded around the model cells where the sink condition is applied. Given the flow field, the origin of flow to the sink can be described by backward particle tracking.

The sink discharge rate Q can then be decomposed as follows: $Q = \sum_{i=1}^{N} q_i$, where q_i is the contribution of the i-th particle to the discharge rate at the sink. This can be rewritten considering particle velocities, v_i: $Q = \sum_{i=1}^{N} v_i \cdot S_i$ where S_i is the area of the surface crossed by q_i. When particles are evenly distributed around the model cell, it can be assumed that $S_i = \frac{S}{N}$ for all i.

Assume that from the N particles seeded around the sink, n_j originate from the $j - th$ boundary condition, subject to contamination. The contribution of this boundary condition to the sink contamination can be written as follows:

$$\alpha_j = \frac{1}{Q} \sum_{k=1}^{n_j} (v_k \cdot \beta_k S_k) \quad (1)$$

where β_k is the mixing ratio in the source model cell where the k-th particle ends. $\beta_k = 1$ for "strong" sources and $0 \leq \beta_k \leq 1$ for "weak sources" where a mixing occurs (Pollock, 2016). In the latter case, β_k can be quantified from the cell water budget at the endpoint cell (Cousquer et al., 2018). This allows the method to account for mixing in the source aquifer cell, where the boundary condition (typically a river) is applied.

The approach is now integrated in the TrackTools Python module. Our approach relies on the USGS MODFLOW6 (Langevin et al., 2017) and MODPATH7 (Pollock, 2016) programs to simulate fluid flow and transport processes, respectively. We rely on the FloPy module (Bakker et al., 2016) for the processing of model input and output files.

Pre-processing capabilities of the TrackTools module are provided in the ParticleGenerator class, which can handle:

- Particle seeding around groundwater sinks, which can be described by a Python geometry or ESRI™ shapefile,
- Adding, removing and merging particles into groups for FloPy and MODPATH7.

After the simulation of flow with MODFLOW and particle tracking with MODPATH, the TrackingAnalyzer class of TrackTools can be used to process particle pathline data and derive the values of mixing ratios at each groundwater sink. The MODFLOW cell-by-cell water budget file is also required to derive the source cell mixing ratio β_k. Results can be provided in the form of a data frame and plotting options are proposed for the description of the origin of groundwater withdrawn in the series of sinks where particles were seeded.

3 Synthetic case

3.1 Model description

In order to illustrate the presented Python module, a synthetic 2D case is considered with a production well in an unconfined aquifer in interaction with a stream. The domain (3,750 × 5,000 m) is discretized with a 2D regular mesh (25 × 25 m cells) with a local refinement around the stream and the well (6.5 × 6.5 m cells). A Cauchy-type boundary condition is applied to the upper limit of the model domain with a stage of 40 m and conductance of 10^{-3} m^2 s^{-1}, while a Dirichlet-type condition is prescribed to the lower boundary with a stage of 20 m. They are simulated with the *General Head Boundary* (GHB) and *Fixed Head* (FH) MODFLOW packages, respectively. The left and right domain boundaries are considered as impermeable. The stream is simulated with a head-dependent flux (Cauchy-type) boundary condition with the dedicated MODFLOW river package (RIV), with a head ranging from 35 m to 25 m from the upper to the lower boundary condition. The stream conductance is set to 10^{-3} m^2 s^{-1}. Steady-state flow conditions are considered and the transmissivity is assumed to be independent of water level fluctuations, which corresponds to the Boussinesq assumption. The well is considered as fully penetrating and is located at position (x, y) = (1212, 1363).

We consider both homogeneous and heterogeneous hydraulic conductivity fields, the heterogeneous cases being generated with an isotropic exponential variogram for log 10(k) defined by a sill of one and a nugget of 0.1 for two different ranges of 200 and 500 m as presented in Figure 1. These fields are generated using the geostatistical python package GSTools (Müller et al., 2021). During the following simulations, the mean hydraulic conductivity value will evolve in heterogeneity patterns (Figures 1B,C).

3.2 Parametric study

The synthetic case previously described is used to illustrate the behavior of the mixing ratio α depending on key parameters such as the hydraulic conductivity K and pumping discharge Q. β from Eq. (1) is quantified from the river cell water budget and can vary from 0 to 1. Figure 2 shows the distribution of α for parameters K and Q ranging from 10^{-5} to 10^{-3} m s^{-1} and from 0 to 200 m^3 h^{-1}, respectively, considering a homogeneous distribution of K. Boundary conditions remain the same throughout this parametric study. We also present in Figure 3 three examples of the spatial distribution of the hydraulic head H with K set to (A) 10^{-5}, (B) 10^{-4} and (C) 10^{-3} m s^{-1} and the same configuration but with a pumping discharge of 90 m^3 h^{-1} for D, E and F. These values have been chosen as representative of the three main tendencies that are observed for α: 1) the yellow

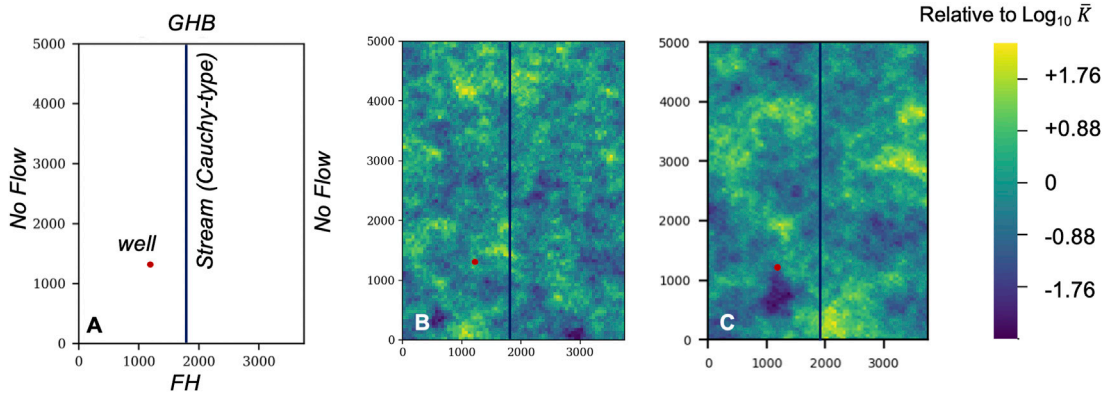

FIGURE 1
(A) Synthetic case model structure and boundary conditions with a General Head Boundary Condition (GHB) of 40 m to the upper boundary and a Fixed Head (FH) of 20 m to the lower boundary. No flow is imposed at the left and right domain boundaries. The stream is simulated with a head-dependent flux (Cauchy-type) boundary condition and the well position is represented with a red dot. The hydraulic conductivity field patterns correspond to an exponential variogram with a sill of 1, a nugget of 0.1 and a range of **(B)** 200 m and **(C)** 500 m.

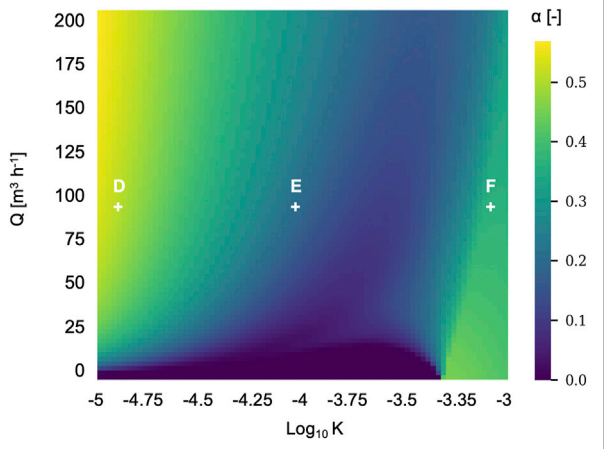

FIGURE 2
Distribution of the mixing ratio α depending on the hydraulic conductivity K and pumping discharge Q.

region in which example D is located corresponding to the highest values of α, 2) the dark blue region in which example E is located corresponding to the lowest values of α and, 3) the green region in which example F is located corresponding to intermediate values of α. This shows that increasing K for a given value of Q results in decreasing (from D to E) and then increasing (from E to F) α, except for very small values of Q for which α directly increases (from E to F). These different behaviors are explained by the different sources of water that contribute to the pumped water, which are characterized by the particle paths represented on the hydraulic head distributions in blue. For small values of K (example D), most of the particles comes from the portion of the river that is located right next to the well, implying that the river provides a substantial proportion of water to the pumping well. Increasing K (example E) results in reducing the impact of the pumping on the natural hydraulic head distribution, implying that the top boundary condition contributes to the pumping. When we keep increasing K (example F), we reach configurations that do not depend on the pumping rate (green triangular region on the right side of Figure 2). For these cases, the pumping does not modify the natural hydraulic head distribution and the particle paths are fully driven by the boundary conditions, resulting in particles coming from the top of the river and thus high values of α. For small values of Q and K (black region of the bottom of the figure), the impact of the pumping on the natural hydraulic head distribution is still negligible but the values of K imply that there is no contribution of the river to the pumping ($\alpha = 0$).

Figure 4 shows the distribution of α when considering the heterogeneous hydraulic conductivity fields provided in Figure 1. The three main tendencies described above for the homogeneous case are also observed for the heterogeneous cases, showing that these behaviors are mostly driven by the boundary, river and pumping conditions. However, small differences are noticeable, in particular for small values of Q and large values of $\bar{K}$, which correspond to configurations with a small impact of the pumping on the natural hydraulic head distribution. For example, the extent of the dark blue region observed at the bottom of Figure 3 is reduced and increased in Figures 4A,B, respectively, showing that the river does not contribute to the pumping for different values of Q and $\bar{K}$. We also observe the presence of yellow areas for large values of $\bar{K}$ in the heterogeneous cases, corresponding to high values of α and thus a strong contribution of the river to the pumping. These different

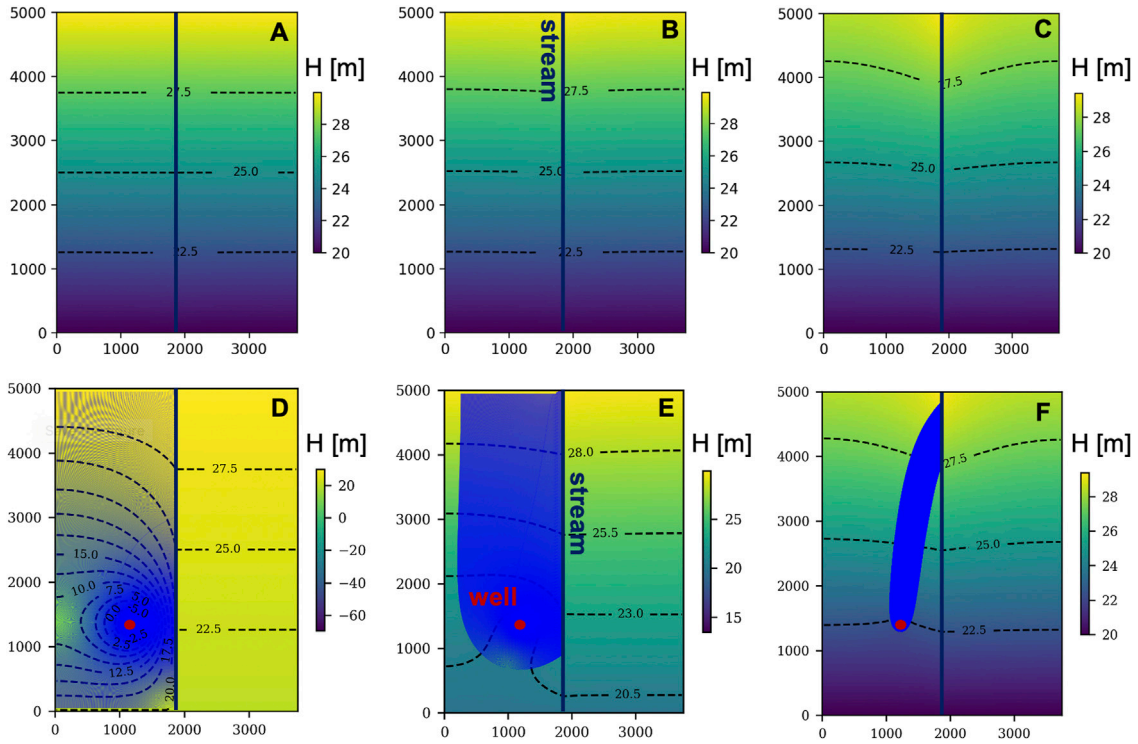

FIGURE 3
The spatial distribution of the hydraulic head $H(x, y)$ for hydraulic conductivity K set to 10^{-5} m s^{-1} **(A,D)**, 10^{-4} m s^{-1} **(B,E)** and 10^{-3} m s^{-1} **(C,F)**. Cases **(A,B,C)** are without pumping, while a pumping rate of 90 m^3 h^{-1} is set for cases **(D,E,F)**. The red dot represent the pumping well locations, the vertical bold black line represents the stream and the blue lines in cases **(D,E,F)** are the path of the particles from the backward particle tracking.

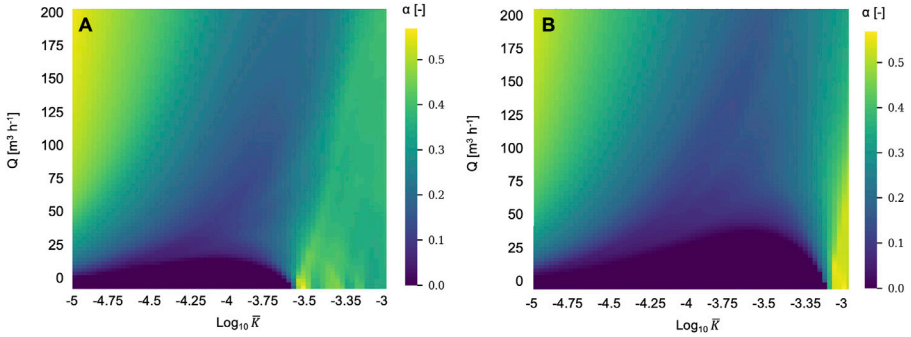

FIGURE 4
Distribution of the mixing ratio α depending on the average hydraulic conductivity $\bar{K}$ and pumping discharge Q for the hydraulic conductivity patterns presented in **(A)** Figures 1B and **(B)** Figure 1C.

behaviors are due to different flow paths induced by the heterogeneities of the hydraulic conductivity fields. Note that, as mentioned before, the impact of these heterogeneities is reduced when increasing the pumping rate.

3.3 Derivative analysis

This synthetic case is also used to set some rules in order to obtain robust derivatives which is essential for parameter estimation and model analysis based on linearization methods.

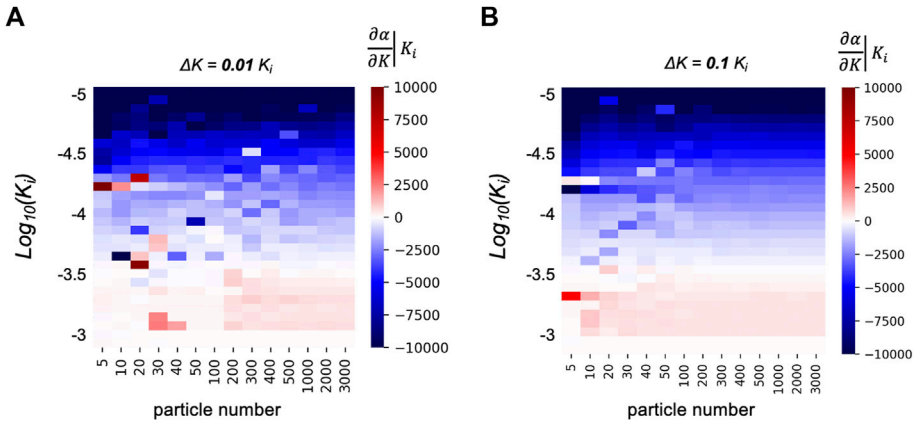

FIGURE 5
Exploration of derivatives quality behaviour regarding the number of particles with the hydraulic conductivity K_i ranging from 10^{-5} to 10^{-3} m s^{-1} and the increment ΔK set to **(A)** 0.01 × K_i and **(B)** 0.1 × K_i.

The sufficient number of particles corresponding to this synthetic case is evaluated with the derivative calculation of $\frac{\partial \alpha}{\partial K}$ which constitutes the Jacobian matrix used for model linearization. The use of gradient based methods for parameter estimation (e.g., the Gauss-Marquardt-Levenberg method on which PEST and PEST++ are based (Doherty, 2015; White et al., 2018)) requires that model outputs are differentiable or at least continuous with regard to model parameters (Doherty, 2010). Doherty and Hunt (2010) mentioned that this assumption often does not hold because of poor model performance and the lack of differentiability can be due to, among others reasons, a low number of particles in the MODPATH particle tracking model. In order to avoid this issue, we explore the differentiability of model output regarding particles number. Let the mixing ratio α be the output of interest, the hydraulic conductivity K is the parameter and ΔK is the parameter increment. The sensitivity of α to K is approximated as follows:

$$\frac{\partial \alpha}{\partial K}|_{K_i} \sim \frac{\alpha(K_i + \Delta K) - \alpha(K_i)}{\Delta K}. \quad (2)$$

The estimate of $\frac{\partial \alpha}{\partial K}$ may be biased when 1) ΔK is too small, due to numerical noise, 2) when ΔK is too large due to model non-linearity, or 3) when the number of particles is not high enough to correctly simulate the mixing ratio. These three points are evaluated by calculating the value of the derivative for several numbers of particles and different values of K_i considering two values of ΔK set to 0.01 × K_i and 0.1 × K_i (Figure 5).

For both ΔK = 0.01 × K_i and ΔK = 0.1 × K_i, Figure 5 shows changes in the derivative sign that are visible when the color representing the derivative goes from blue (negative values) to red (positive values) and vice versa. For a given line, these changes depend on the number of particles N_p used in the model, showing wrong estimations of the derivative when N_p is too small (due to wrong estimations of the mixing ratio). This results in instabilities of the derivative along K for small number of particles (N_p < 200–500, columns that are located on the left side of Figure 5). For parameter estimation and inversion purposes, these instabilities could be interpreted as local minima and lead to wrong estimates and optimized values. When N_p is high enough (columns on the right side), the derivative values are stable in the sense that they do not change with the value of N_p. In this case, changes in $\frac{\partial \alpha}{\partial K}$ along K are due to the non-linearity of the model, which needs to be taken into account for parameter estimation and optimization.

Furthermore, comparing Figures 5A,B shows that the choice of the increment of K influences the stability of the derivative. The number of particles that is required to observe stable values of $\frac{\partial \alpha}{\partial K}$ is lower when increasing ΔK.

Although they are expected, these results clearly show that the derivative calculation requires to carefully choose the parameters N_p and ΔK with 1) a sufficient number of particles to correctly simulate the mixing ratios and its derivative, and 2) an increment that is neither too small to avoid that the derivative calculation is tainted by numerical problems, nor too large to take into account the non-linearity of the model. In all cases, increasing the number of particles is a systematic method to improve the definition of the derivative whatever the value of the discretization step of the derivative.

4 Case study

4.1 Model description

The method described above is embedded in a workflow designed for model-based decision making and applied to a

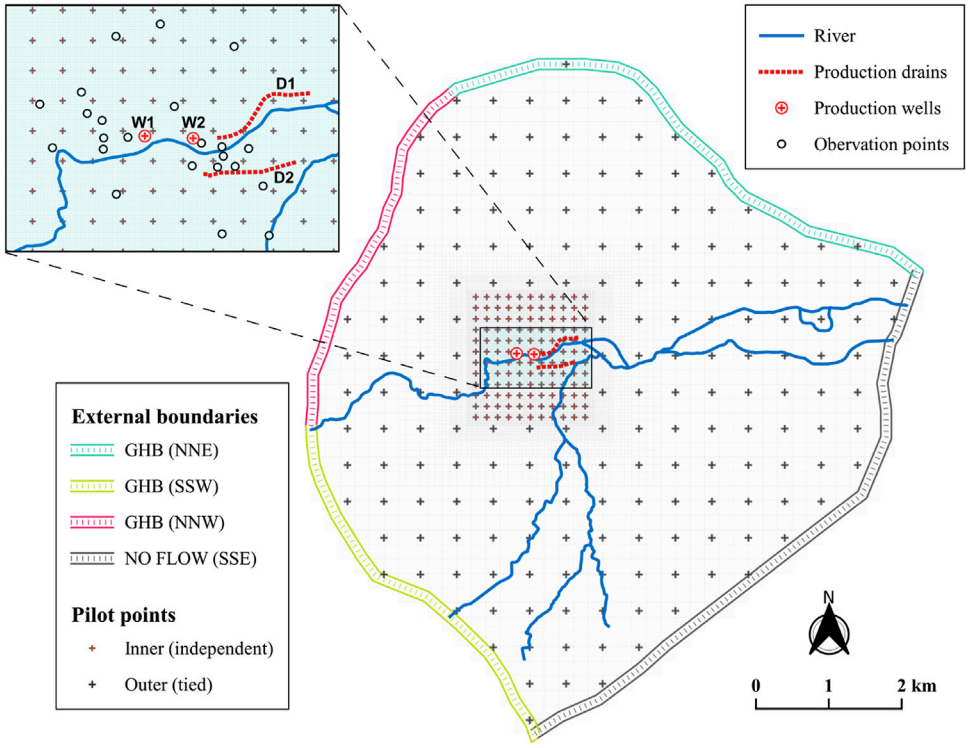

FIGURE 6
Model domain with boundary conditions inferred from the regional groundwater flows. Mesh refinement is conducted in the area of interest around the production and observation wells and drains (see inset). The parameterization of hydraulic conductivity has been conducted with pilot points.

complex real world case study. We describe the script-based modeling workflow and interfacing with the PEST + +, which provides an illustration of the interest of the method and facilitates the replication of the approach to other cases. All scripts are available on a GitHub® repository.

The well field is located in South-West France and is critical to the water supply for the city of Bordeaux (Cousquer, 2017; Valois et al., 2018; Delbart et al., 2021) as it accounts for about 20% of the needs. The aquifer lies in Oligocene limestone overlain by sandy and alluvial deposits. Limestones are locally subject to karstification, leading to strongly heterogeneous hydraulic properties. Two wells (W1, W2) and two horizontal drains (D1, D2) are used for groundwater extraction. Due to industrial activities, the main stream crossing the well field from west to east is prone to contamination. The vicinity of pumping wells and drains to the stream is favorable to stream-aquifer exchanges. The end-purpose of this modeling exercise is to maximize groundwater production while meeting drinking water quality standards.

For this model purpose, the geographical data describing the domain and geometry of boundary conditions was first processed with QGIS (QGIS Development Team, 2022). The model setup was fully script-based with the FloPy Python library (Bakker et al., 2016) for the pre- and post-processing of MODFLOW6 and MODPATH7 input and output files. A 2D single layer was considered under the Dupuit-Forchheimer approximation to represent the aquifer of interest. The model domain was discretized by Gridgen (Lien et al., 2015) with a quadtree grid with four horizontal refinement levels so that square cell dimensions range from 200 m to 12.5 m (Figure 6). External boundaries have been defined from the regional groundwater levels as no-flow or 3rd type head-dependent flow boundary condition. The streams and drains are simulated with head-dependent (Cauchy-type) flux boundary conditions and the wells are represented by sink terms in the aquifer cells corresponding to their location.

In order to consider contrasting settings while avoiding the burden of transient simulations, pseudo steady-state flow conditions are considered (Haitjema, 2006; Moore and Doherty, 2021). This hypothesis is supported in the present case since the permeable aquifer responds quickly to changes Cousquer (2017). The TrackTools module is used to seed particles around the two wells and the two drains and to derive the mixing ratio of the water withdrawn. It was found that

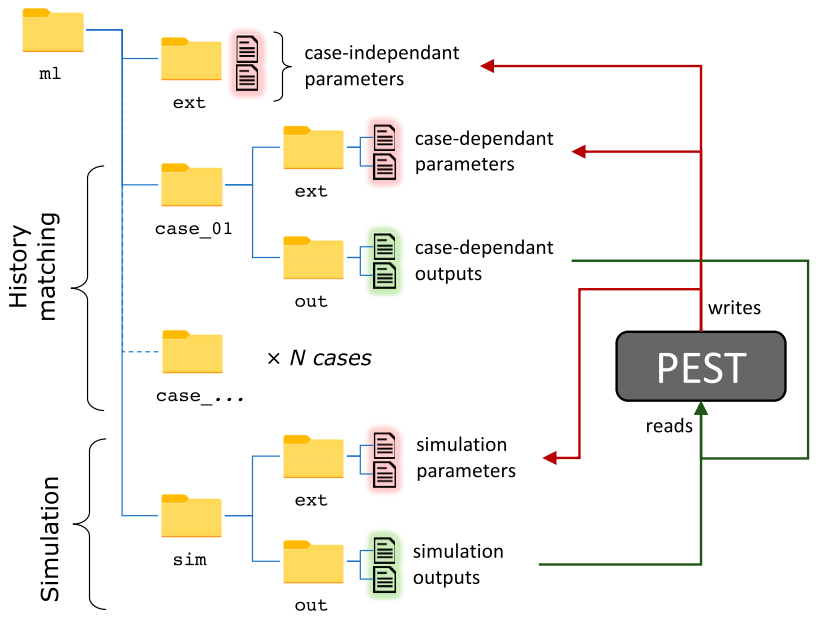

FIGURE 7
Directory tree structure used for parameter estimation and optimization workflow.

simulated mixing ratios were relatively stable with 500 particles seeded around each production unit (wells and drains). As detailed in Section 2 and references herein, the mixing ratios at source cells where particles stop (β) is inferred from the cell water budget. This allows to account for mixing of aquifer and river water in model cells where a river boundary condition is applied. Note that only the main river is prone to contamination, water originating from the tributaries is not considered as contaminated.

4.2 Parameter estimation and optimization

The model interfacing with the PEST + + suite has been performed by means of PyEMU (White et al., 2016). The hydraulic conductivity field was parameterized with a dense set of pilot points. Hydraulic conductivity values at pilot points lying at the center of the study area were all considered as independent, but they were tied in the outer portion of model domain (Figure 6). This is motivated by the lack of observations in the outer zone and it is unlikely that model predictions are sensitive to details in this remote area.

A series of 10 field surveys, spread over 4 years (2014–2018) have been considered for history matching, considering pseudo steady-state conditions for contrasting stream levels and operating conditions. Though it was challenging for a well field in activity, we have made our best for operating conditions (well discharge rates and drain levels) to remain relatively stable on the period preceding each of the surveys. The observation data set is composed of hydraulic heads, drain discharge rates, and mixing ratios. The latter were derived by end-member analysis from HCO^{3-} and Ca^{2+} concentrations (Delbart et al., 2021).

In order to reduce run times, all model files for each of these surveys were setup in separate folders (Figure 7). Doing so, file operations during parameter estimation were limited to writing parameter files and reading output files. These operations were all conducted with the dedicated methods of the PyEMU Python package.

The parameter estimation was conducted with the widely used Gauss-Levenberg-Marquardt Algorithm (GLMA) as implemented in PEST + + suite (White et al., 2020b). Parameter increments (DERINC) values were adjusted by trial and error. Best results were obtained with relative parameter increments of 15% for hydraulic conductivities and 10% for all the other parameters. Both zero order (preferred value) and first order Tikhonov regularizations were employed (Doherty, 2015).

After parameter estimation, the resulting model was used to optimize the total abstraction rate. The optimization problem consists in maximizing the total abstraction rate Q while verifying the drinking water quality standards after mixing. It can be expressed as follows:

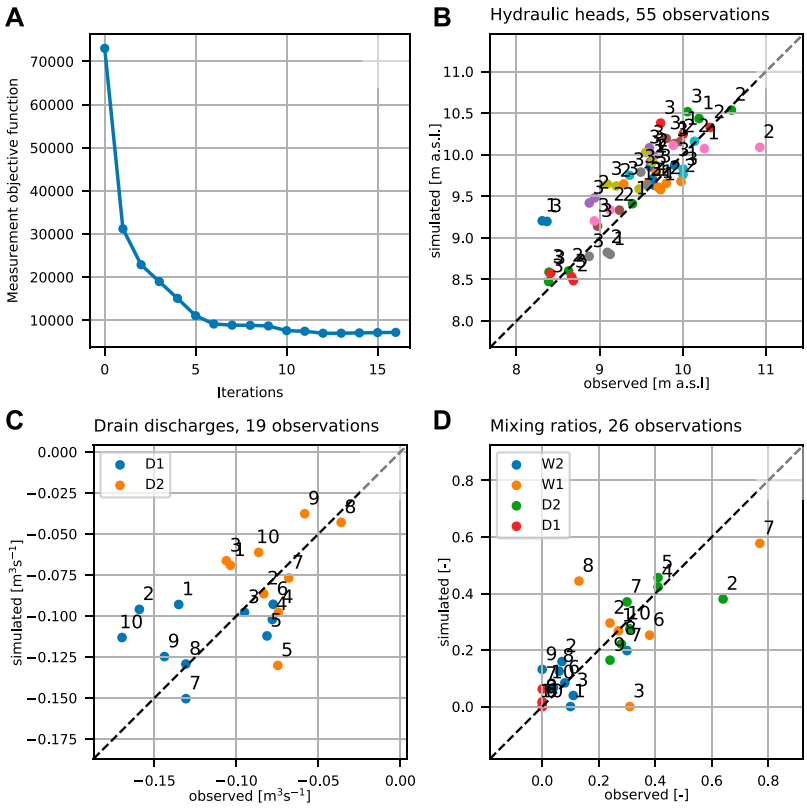

FIGURE 8
Results of the parameter estimation with the GLMA. **(A)** Evolution of the measurement objective function throughout GLMA iterations. Simulated versus measured hydraulic heads **(B)**, drain discharges **(C)**, and mixing ratios **(D)**. Numbers in **(B,C,D)** refer to the surveys; colors in **(B)** refer to the observation wells.

$$\begin{aligned} \max_{Q_i,h_i} \quad & Q = \sum_{i=1}^{N} Q_i \\ \text{s.t.} \quad & \alpha = \frac{1}{\sum_i Q_i} \sum_{i=1}^{N} Q_i \alpha_i \le \alpha_{\text{crit}} \\ & Q_i \le Q_{\max,i} \quad i \in \{1;2\} \\ & h_i \ge h_{\min,i} \quad j \in \{1;2\}. \end{aligned} \tag{3}$$

where the decision variables Q_i and h_i correspond to well discharge rate and drain level, respectively, and the constraint on water quality is expressed as a critical stream-aquifer mixing ratio α_{crit}.

For this illustrative exercise, the optimization is solved with a sequential version of linear programming as implemented in PESTPP-OPT (White et al., 2018) considering a 50% risk (maximum likelihood) configuration. This algorithm is fast but sensitive to model non-linearity.

4.3 Results

The history matching conducted with the GLMA converged in approximately 10 iterations (Figure 8A). The fit between simulated values with their observed counterparts can be considered as satisfying for heads (Figure 8B) and reasonable for discharge rates and mixing ratios (Figures 8C,D). Measurements of drain discharge rates can be considered as reliable and the misfit may rather be explained by the static, linear relation that is considered with the MODFLOW drain package. In contrast, measurement of mixing ratios derived from concentrations by end-member analysis are prone to uncertainties (Delbart et al., 2021). The structural error of this simplified model is therefore strong and surely contributes to the misfit, but uncertainties in the observation data and historic operating variables are also important. In this context, it is unclear whether more detailed process-modeling could improve the predictive capacity of the model.

The estimated hydraulic conductivity field is highly heterogeneous (Figure 9A), as expected for partly karstified limestones. Attempts to smooth these contrasts with stronger regularization constraints lead to an important increase of the misfit with historical data. Karst conduits are expected to be narrow and of limited extension but can be at the origin of the high values of estimated hydraulic

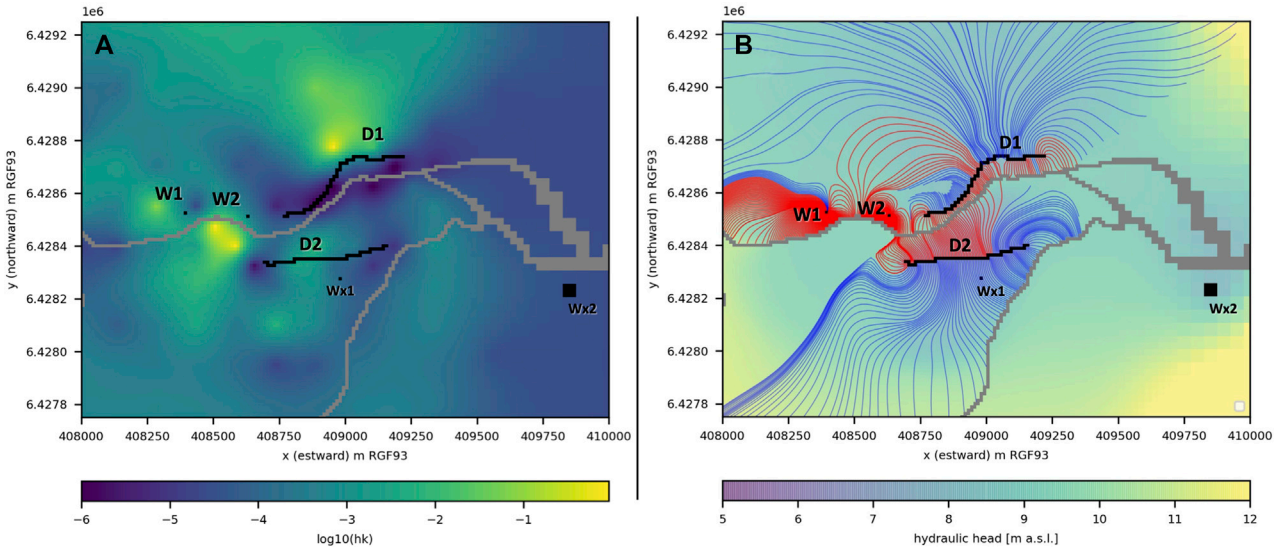

FIGURE 9
(A) Hydraulic conductivity field after parameter estimation. **(B)** Simulated groundwater table level and particle tracking trajectory for the initial values of parameters Q and α. Particle trajectory from the contaminated stream are colored in red and those from uncontaminated boundaries in blue. Particle tracking was conducted from production wells W1, W2 and drains D1, D2 shown in black. Pumping from wells Wx1 and Wx2 was considered in the flow model but they do not pertain to the studied drinking water facility.

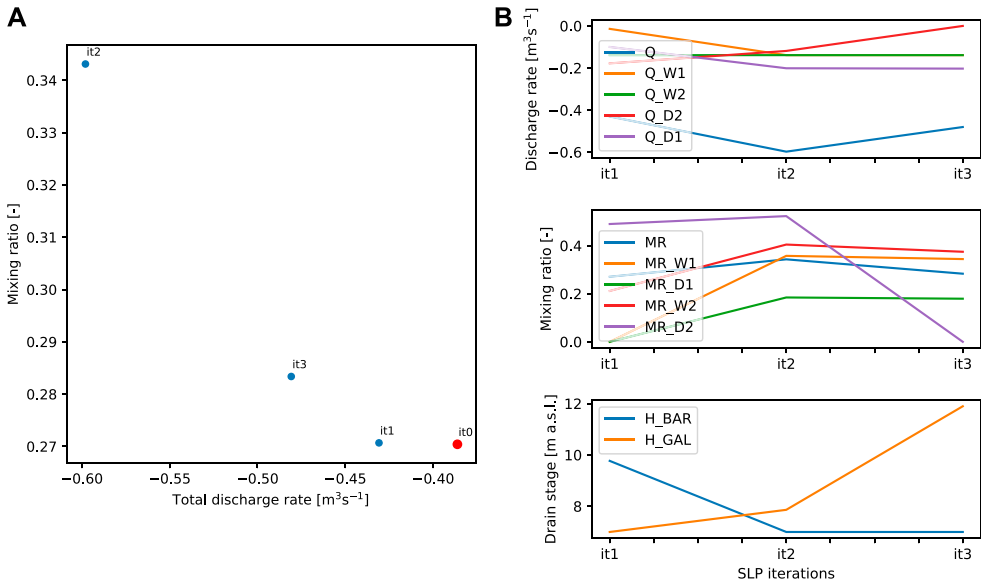

FIGURE 10
Optimization results with **(A)** total discharge rate and mixing ratio for each optimization iteration of PESTPP-OPT and **(B)** discharge rate (Q), mixing ratio (MR), and drain stage (H) for each optimization iteration. it0 represents the initial parameter values, it1, it2 and it3 the values for the first, second and third iterations, respectively.

conductivity (Cousquer, 2017). On the opposite, low values can be explained by fine deposits in the riverbed (Delbart et al., 2021).

We followed a deterministic approach, which can be considered as a first exploratory step for more robust methods. The "best" calibrated parameter set was therefore

used to conduct the optimization of well and drain discharge given a quality constraint expressed as a global mixing ratio α_{crit}=25%. This corresponds to a 4-fold reduction of contaminant concentration with respect to river contamination levels. The optimization algorithm was run from a situation close to the current operating configuration over three iterations of the sequential linear programming algorithm (Figure 10). Compared to the initial settings (iteration 0), the third and last iteration lead to an increase of production rate by + 25% for a similar contamination level. The particle pathlines for this configuration are presented in Figure 9B.

5 Discussion and conclusion

In the context of drinking water contamination issues, we provided a decision-making tool to investigate the vulnerability of groundwater production units. The TrackTools Python module offers pre- and post-processing functions of particle tracking data simulated with MODFLOW and MODPATH. Mixing ratios are inferred from flow contributions of each contaminant source to the discharge rate of a groundwater sink considering particle velocities and mixing in source cells. This can be appropriate where a production well or drain is vulnerable to a boundary condition such as a river. However, the method is not relevant when the source of the contamination is diffuse, such as agricultural contaminants driven by groundwater recharge. In the latter case, forward particle tracking is more appropriate [see e.g., Fienen et al. (2022)].

The synthetic case presented in this work illustrates that contrasting behaviors can be observed even for simple configurations. It highlights the interest of the method to explore the sensitivity of mixing ratios to operating variables (pumping rates, drain levels) and model parameters. The impact of the number of particles seeded around each sink, and the size of the increment used for model linearization was also evaluated through the analysis of the derivatives of α. Derivative instabilities are mitigated when increasing the number of particles or increasing the parameter increment for estimating model derivatives (DERINC). However, adding particles tends to increase the computation time, so that a reasonable particle number should be considered.

Relatively fast and easy to implement, this approach will facilitate the use of contaminant transport models for history matching, uncertainty quantification or optimization. It also presents the advantage to be didactic and can be used to represent contaminant transfer in a visual and potentially interactive manner, which is important for end-users.

The method was implemented on a real world case study and is provided with a series a script for the interfacing with the PEST + + suite. The parameter estimation with the GLMA and optimization by linear programming illustrated the interest of the approach. Difficulties in reproducing observations can be explained by errors in the historic dataset (observations and operating variables) and model structural error. In spite of our efforts to obtain a robust linear version of the model, the methods based on the Jacobian matrix have shown their limitations. This advocates for the use of more robust ensemble methods such as the Iterative Ensemble Smoother (IES (White et al., 2018)) for parameter estimation and uncertainty quantification, and evolutionary algorithms for optimization [PESTPP-MOU (White et al., 2022)]. The presented workflow may easily be extended to these methods recently available in the PEST + + Suite.

The method was illustrated on 2D horizontal models assuming steady state flow conditions. This simplifies the implementation and leads to particularly fast run times, but 3D models and transient conditions may have to be considered in other contexts. The method may be extended to these configurations with some additional processing. The implementation of the method on 3D models will be straightforward so long the wells and drains penetrate a single model layer. Otherwise, particles should be seeded in each of the intersected layers and mixing in the multi-layer well should be accounted considering the respective flow contributions provided by the dedicated Multi-Aquifer Well package (MAW) from MODFLOW6 (Langevin et al., 2017). The impact of transient flow dynamics on the mixing ratio at a specific instant will be easy to consider with backward particle tracking over historic conditions. However, it would be more challenging to quantify mixing ratios averaged over a period as it would require multiple runs of the particle tracking algorithm.

Among the simplifying assumptions of this approach, we assumed the respective flow contributions of particles evenly placed around the sink to be proportional to their velocities (Eq. 1). The validity of this approach has been investigated by Cousquer et al. (2018) with an advective-dispersive model on a synthetic, homogeneous model. Results were encouraging but the performance of the simplified model based on particle tracking may be put into question in other contexts. The use of a paired simple-complex model approach (Doherty and Christensen, 2011) can be suggested, where the complex model would be based on the resolution of the advection-dispersion equation.

Future work will also focus on improving the reliability of the simplified transport models by considering additional processes, such as dispersion. For this matter, various recent particle-based methods (Noetinger et al., 2016; Gouze et al., 2020; Roubinet et al., 2022) and machine learning techniques (Kang et al., 2021; Zhou et al., 2021) could be considered in order to consider more realistic configurations while keeping the low computational cost of the forward models.

As mentioned in the introduction, the optimal complexity level is to great extent, case- and purpose-specific, as it will depend on the purpose of the modeling exercise and the quality of the available observation dataset. In this perspective, it is likely that simple and effective methods such as the one presented in this study remain of interest in numerous situations.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/tracktools/.

Author contributions

AP: methodology, software development, case study, writing and editing. PM: software development, data processing and visualization. YC: methodology, synthetic case, writing DR: synthetic case, writing, reviewing, editing.

Funding

This work was initially conducted in the framework of the Mhyqadeau project supported by Suez Environnement (LyRE) and the French Aquitaine regional council.

Acknowledgments

The authors are thankful to the editors for handling this special issue and to the two anonymous reviewers for their corrections and valuable suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

M. P. Anderson, W. W. Woessner, and R. J. Hunt (Editors) (2015). *Applied groundwater modeling*. Second Edition (San Diego: Academic Press). doi:10.1016/B978-0-08-091638-5.00018-3

Asher, M. J., Croke, B. F., Jakeman, A. J., and Peeters, L. J. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resour. Res.* 51, 5957–5973.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting modflow model development using python and flopy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Burrows, W., and Doherty, J. (2015). Efficient calibration/uncertainty analysis using paired complex/surrogate models. *Groundwater* 53, 531–541.

Carniato, L., Schoups, G., van de Giesen, N., Seuntjens, P., Bastiaens, L., and Sapion, H. (2015). Highly parameterized inversion of groundwater reactive transport for a complex field site. *J. Contam. hydrology* 173, 38–58.

Cousquer, Y. (2017). *Modélisation des échanges nappe-rivière à l'échelle intermédiaire : Conceptualisation, calibration, simulation*. Ph.D. thesis (Pessac, France: Géoressources et Environnement, Bordeaux INP et Univ. Bordeaux Montaigne).

Cousquer, Y., Pryet, A., Atteia, O., Ferré, T. P., Delbart, C., Valois, R., et al. (2018). Developing a particle tracking surrogate model to improve inversion of ground water – surface water models. *J. Hydrology* 558, 356–365. doi:10.1016/j.jhydrol.2018.01.043

Daud, M. K., Nafees, M., Ali, S., Rizwan, M., Bajwa, R. A., Shakoor, M. B., et al. (2017). Drinking water quality status and contamination in Pakistan. *BioMed Res. Int.* 2017, 7908183. doi:10.1155/2017/7908183

de Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005). Dealing with spatial heterogeneity. *Hydrogeology J.* 13, 161–183. doi:10.1007/s10040-004-0432-3

Delbart, C., Pryet, A., Atteia, O., Cousquer, Y., Valois, R., Franceschi, M., et al. (2021). When perchlorate degradation in the riverbank cannot impede the contamination of drinking water wells. *Hydrogeology J.* 29, 1925–1938.

Doherty, J. (2010). *Addendum to the pest manual*. Brisbane, Australia: Watermark Numerical Computing, 131.

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental modelsWatermark Numerical Computing*. Brisbane, Australia.

Doherty, J., and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resour. Res.* 47. doi:10.1029/2011WR010763

Doherty, J. E., and Hunt, R. J. (2010). *Approaches to highly parameterized inversion: A guide to using PEST for groundwater-model calibration, vol. 2010*. Reston, VA, USA: US Department of the Interior, US Geological Survey.

Doherty, J. (2016). *Model-independent parameter estimation user manual part i: Pest, sensan and global optimisers*. Brisbane, Australia: Watermark Numerical Computing, 390.

Doherty, J., and Moore, C. (2020). Decision support modeling: Data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater* 58, 327–337. doi:10.1111/gwat.12969

Fienen, M. N., Corson-Dosch, N. T., White, J. T., Leaf, A. T., and Hunt, R. J. (2022). Risk-based wellhead protection decision support: A repeatable workflow approach. *Groundwater* 60, 71–86. doi:10.1111/gwat.13129

Gouze, P., Puyguiraud, A., Roubinet, D., and Dentz, M. (2020). Characterization and upscaling of hydrodynamic transport in heterogeneous dual porosity media. *Adv. Water Resour.* 146. doi:10.1016/j.advwatres.2020.103781

Guthke, A. (2017). Defensible model complexity: a call for data-based and goal-oriented model choice. *Groundwater* 55, 646–650. doi:10.1111/gwat.12554

Gwimbi, P., George, M., and Ramphalile, M. (2019). Bacterial contamination of drinking water sources in rural villages of mohale basin, Lesotho: Exposures through neighbourhood sanitation and hygiene practices. *Environ. Health Prev. Med.* 24, 33. doi:10.1186/s12199-019-0790-z

Haitjema, H. (2006). The role of hand calculations in ground water flow modeling. *Groundwater* 44, 786–791. doi:10.1111/j.1745-6584.2006.00189.x

Hermans, T. (2017). Prediction-focused approaches: An opportunity for hydrology. *Groundwater* 55, 683–687. doi:10.1111/gwat.12548

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). Process consistency in models: the importance of system signatures, expert knowledge, and process complexity. *Water Resour. Res.* 50, 7445–7469. doi:10.1002/2014wr015484

Hugman, R., and Doherty, J. (2022). Complex or simple—Does a model have to be one or the other? *Front. Earth Sci.* 10, 867379. doi:10.3389/feart.2022.867379

Hunt, R. J., Feinstein, D. T., Pint, C. D., and Anderson, M. P. (2006). The importance of diverse data types to calibrate a watershed model of the trout lake basin, northern Wisconsin, USA. *J. Hydrology* 321, 286–296. doi:10.1016/j.jhydrol.2005.08.005

Kang, X., Kokkinaki, A., Kitanidis, P. K., Shi, X., Lee, J., Mo, S., et al. (2021). Hydrogeophysical characterization of nonstationary dnapl source zones by

integrating a convolutional variational autoencoder and ensemble smoother. *Water Resour. Res.* 57, e2020WR028538. doi:10.1029/2020WR028538

Knowling, M. J., White, J. T., Moore, C. R., Rakowski, P., and Hayley, K. (2020). On the assimilation of environmental tracer observations for model-based decision support. *Hydrology Earth Syst. Sci.* 24, 1677–1689. doi:10.5194/hess-24-1677-2020

Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., and Provost, A. M. (2017). "Documentation for the MODFLOW 6 groundwater flow model," in *Tech. rep.* (Reston, VA, USA: US Geological Survey).

Lien, J.-M., Liu, G., and Langevin, C. D. (2015). *GRIDGEN version 1.0: A computer program for generating unstructured finite-volume grids*. Reston, VA, USA: US Department of the Interior, US Geological Survey. doi:10.3133/ofr20141109

Moore, C. R., and Doherty, J. (2021). Exploring the adequacy of steady-state-only calibration. *Front. Earth Sci.* 9, 692671. doi:10.3389/feart.2021.692671

Müller, S., Schüler, L., Zech, A., and Heße, F. (2021). Gstools v1. 3: A toolbox for geostatistical modelling in python. *Geosci. Model. Dev. Discuss.* 2021, 1–33.

Noetinger, B., Roubinet, D., Russian, A., Le Borgne, T., Delay, F., Dentz, M., et al. (2016). RussianRandom walk methods for modeling hydrodynamic transport in porous and fractured media from pore to reservoir scale. *Transp. Porous Media* 2016, 1–41. doi:10.1007/s11242-016-0693-z

Pollock, D. W. (2016). *User guide for MODPATH version 7—a particle- tracking model for MODFLOW*. Reston, VA, USA: U.S. Geological Survey Open-File Report 2016–1086. doi:10.3133/ofr20161086

Pool, M., Carrera, J., Alcolea, A., and Bocanegra, E. (2015). A comparison of deterministic and stochastic approaches for regional scale inverse modeling on the mar del plata aquifer. *J. Hydrology* 531, 214–229.

QGIS Development Team (2022). *Geographic Information System Developers Manual. QGIS Association*. Electronic document: https://docs.qgis.org/3.22/en/docs/developers_guide/index.html.

Rajabi, M. M., Ataie-Ashtiani, B., and Simmons, C. T. (2018). Model-data interaction in groundwater studies: Review of methods, applications and future directions. *J. Hydrology* 567, 457–477.

Razavi, S., Tolson, B. A., and Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resour. Res.* 48. doi:10.1029/2011WR011527

Roubinet, D., Gouze, P., Puyguiraud, A., and Dentz, M. (2022). Multi-scale random walk models for reactive transport processes in fracture-matrix systems. *Adv. Water Resour.* 164, 104183. doi:10.1016/j.advwatres.2022.104183

Schilling, O. S., Cook, P. G., and Brunner, P. (2019). Beyond classical observations in hydrogeology: The advantages of including exchange flux, temperature, tracer concentration, residence time, and soil moisture observations in groundwater model calibration. *Rev. Geophys.* 57, 146–182. doi:10.1029/2018RG000619

Schwartz, F. W., Liu, G., Aggarwal, P., and Schwartz, C. M. (2017). Naïve simplicity: The overlooked piece of the complexity-simplicity paradigm. *Groundwater* 55, 703–711. doi:10.1111/gwat.12570

Sharma, S., and Bhattacharya, A. (2017). Drinking water contamination and treatment techniques. *Appl. Water Sci.* 7, 1043–1067. doi:10.1007/s13201-016-0455-7

Turner, S. W. D., Rice, J. S., Nelson, K. D., Vernon, C. R., McManamay, R., Dickson, K., et al. (2021). Comparison of potential drinking water source contamination across one hundred U.S. cities. *Nat. Commun.* 12, 7254. doi:10.1038/s41467-021-27509-9

Valois, R., Cousquer, Y., Schmutz, M., Pryet, A., Delbart, C., and Dupuy, A. (2018). Characterizing stream-aquifer exchanges with self-potential measurements. *Groundwater* 56, 437–450. doi:10.1111/gwat.12594

White, J. T., Fienen, M. N., Barlow, P. M., and Welter, D. E. (2018). A tool for efficient, model-independent management optimization under uncertainty. *Environ. Model. Softw.* 100, 213–221. doi:10.1016/j.envsoft.2017.11.019

White, J. T., Fienen, M. N., and Doherty, J. E. (2016). A python framework for environmental model uncertainty analysis. *Environ. Model. Softw.* 85, 217–228. doi:10.1016/j.envsoft.2016.08.017

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020a). Toward reproducible environmental modeling for decision support: A worked example. *Front. Earth Sci.* 8. doi:10.3389/feart.2020.00050

White, J. T., Hunt, R. J., Fienen, M. N., and Doherty, J. E. (2020b). "Approaches to highly parameterized inversion: PEST++ version 5, a software suite for parameter estimation, uncertainty analysis, management optimization and sensitivity analysis," in *Tech. rep.* (Reston, VA, USA: US Geological Survey). doi:10.3133/tm7C26

White, J. T., Knowling, M. J., Fienen, M. N., Siade, A., Rea, O., and Martinez, G. (2022). A model-independent tool for evolutionary constrained multi-objective optimization under uncertainty. *Environ. Model. Softw.* 149, 105316. doi:10.1016/j.envsoft.2022.105316

Zhou, Z., Roubinet, D., and Tartakovsky, D. M. (2021). Thermal experiments for fractured rock characterization: Theoretical analysis and inverse modeling. *Water Resour. Res.* 57, e2021WR030608. doi:10.1029/2021WR030608

frontiers | Frontiers in Earth Science

TYPE Technology and Code
PUBLISHED 30 September 2022
DOI 10.3389/feart.2022.903965

OPEN ACCESS

EDITED BY
Francesca Pianosi,
University of Bristol, United Kingdom

REVIEWED BY
Richard Niswonger,
United States Geological Survey (USGS), United States
Norman Jones,
Brigham Young University, United States

*CORRESPONDENCE
Andrew T. Leaf,
aleaf@usgs.gov

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 24 March 2022
ACCEPTED 29 July 2022
PUBLISHED 30 September 2022

CITATION
Leaf AT and Fienen MN (2022), Modflow-setup: Robust automation of groundwater model construction. *Front. Earth Sci.* 10:903965. doi: 10.3389/feart.2022.903965

Modflow-setup: Robust automation of groundwater model construction

Andrew T. Leaf* and Michael N. Fienen

U.S. Geological Survey Upper Midwest Water Science Center, Middleton, WI, United States

In an age of both big data and increasing strain on water resources, sound management decisions often rely on numerical models. Numerical models provide a physics-based framework for assimilating and making sense of information that by itself only provides a limited description of the hydrologic system. Often, numerical models are the best option for quantifying even intuitively obvious connections between human activities and water resource impacts. However, despite many recent advances in model data assimilation and uncertainty quantification, the process of constructing numerical models remains laborious, expensive, and opaque, often precluding their use in decision making. Modflow-setup aims to provide rapid and consistent construction of MODFLOW groundwater models through robust and repeatable automation. Common model construction tasks are distilled in an open-source, online code base that is tested and extensible through collaborative version control. Input to Modflow-setup consists of a single configuration file that summarizes the workflow for building a model, including source data, construction options, and output packages. Source data providing model structure and parameter information including shapefiles, rasters, NetCDF files, tables, and other (geolocated) sources to MODFLOW models are read in and mapped to the model discretization, using Flopy and other general open-source scientific Python libraries. In a few minutes, an external array-based MODFLOW model amenable to parameter estimation and uncertainty quantification is produced. This paper describes the core functionality of Modflow-setup, including a worked example of a MODFLOW 6 model for evaluating pumping impacts to a lake in central Wisconsin, United States.

KEYWORDS

groundwater modeling, Python, MODFLOW, Flopy, numerical modeling, software, automation

Introduction

Numerical groundwater models can provide water managers and other stakeholders with a powerful physics-based framework for evaluating complex hydrologic systems, which may be difficult or impossible to represent analytically (e.g., Anderson et al., 2015). In comparison to analytical methods, numerical models provide flexibility in their ability

to finely discretize natural heterogeneity and complex boundaries or structures, and in their ability to represent transient effects. In many real-world systems, these capabilities may be essential to representing and understanding questions of interest (e.g., Leaf et al., 2015; Fienen et al., 2022,2021a). Just as importantly, numerical models allow for higher dimensional parametrization that can be critical for effective data assimilation, associated model error reduction, and meaningful consideration of model uncertainty (e.g., Moore and Doherty 2005; Hunt et al., 2007; White et al., 2021, 2014). Because of these advantages, numerical groundwater models are used widely. MODFLOW (e.g., Niswonger et al., 2011; Langevin et al., 2017) and related codes are the most popular framework for numerical modeling of groundwater flow and transport worldwide.

The flexibility of numerical models, however, comes with steep costs. Disparate input data must be mapped to thousands or millions of computational cells, a process that can be cumbersome, labor-intensive, and error-prone. The number and complexity of operations presents a fundamental challenge to scientific reproducibility (e.g., Peng 2011; Fienen and Bakker, 2016), step-wise modeling (Haitjema, 1995), and the modeler's own cognitive load (e.g., Sweller 1988). The inherent difficulty of, for example, changing discretization or model structure makes it difficult to revisit these choices later in a project in response to what is learned, and carrying alternative conceptual models through a project is seldom feasible. As noted by White et al. (2021), realistic representation of model uncertainty presents an additional set of challenges that may be out of reach if the basic model inputs cannot be efficiently built. As a result of these costs, numerical groundwater models are not only expensive but can often fall short of expectations (e.g., Donoho et al., 2008; Moran 2016; Doherty and Moore, 2020). A key goal, then, is to automate repetitive, but crucial, model construction tasks such that a modeler can focus their efforts on the underlying problem and conceptualization rather than model mechanics.

Numerical groundwater models are often constructed with the help of a graphical user interface (GUI). GUIs provide an interactive environment for building and post-processing models that is especially helpful for visualization and handling of model input and output formats. Some GUIs even support grid-independent input. The reader is referred to Anderson et al. (2015) for a more thorough discussion of GUI options. Although many consider the "point and click" approach afforded by GUIs to be more intuitive than direct manipulation of model input files, most GUI workflows are not readily automatable, and therefore prone to the issues mentioned earlier. Without automation, meaningful documentation of the workflow requires additional effort on the modeler's part and may not be feasible under typical project constraints.

In recent years, open-source software tools to automate the mapping of disparate data to computational grids have become readily available and easy to use. These include Python packages for working with MODFLOW files (Bakker et al., 2016), GIS file formats and geoprocessing (Gillies 2022a,b,c,d), NetCDF data (Hoyer and Hamman, 2017), coordinate transformations (Snow et al., 2022), and general scientific algorithms (Virtanen et al., 2020); as well as software development tools that facilitate collaborative version control (e.g., Git; https://git-scm.com/and GitHub; https://github.com/), automated testing (e.g., Pytest; https://pytest.org), continuous integration, and online documentation (e.g., Sphinx; https://www.sphinx-doc.org/); and accessible tutorials that show domain scientists how to use them (e.g., https://nsls-ii.github.io/scientific-python-cookiecutter).

Script-based development of model input with a high-level language such as Python has therefore been proposed as a solution to overcome model construction challenges (e.g., Bakker et al., 2016), but in practice this is easier said than done. Ad hoc scripts must be assembled into a carefully documented workflow that can have many steps and interdependencies and is itself subject to the "ubiquity of error" (Donoho et al., 2008). Even the most well documented workflows depend on the quality of the underlying code and therefore, the fastidiousness and programming abilities of the modeler. In the end, a fully scripted workflow may be no easier to understand, repeat, or reproduce than a sequence of manual operations in a GUI or spreadsheet environment.

Fisher et al. (2016) presented what may be the best published example of a fastidious model construction workflow. In development of a groundwater model for a project in Idaho, United States, they developed code functions and assembled them into a formal R package complete with code documentation and *vignettes* (R Core Team, 2014) walking users through the workflow. Although this approach almost certainly improved reproducibility and likely carried other advantages, it was focused on a single project, and likely required considerable overhead effort that may not be readily transferable to other work.

Quality code development for robust and reproducible workflows takes time (e.g., Donoho et al., 2008; Wilson et al., 2014) that is most efficiently spent developing general code that can be reused in many different contexts. Functions and other objects that are dedicated to specific tasks and can be readily imported into a script or called repeatedly in a loop provide a local means for reusability. Functions carry the added benefit of breaking complex workflows into easily understandable pieces that can also be readily tested, thereby reducing error. At a higher level, software packages provide a well-understood framework for developing, testing, and sharing collections of functions and other objects. The Flopy project (Bakker et al., 2016) provides such a package, but at a low level that typically requires extensive ad hoc scripting and geoprocessing outside of Flopy to develop a MODFLOW model in a real-world context. MODFLOW models are not internally geolocated, so Flopy is always referenced to

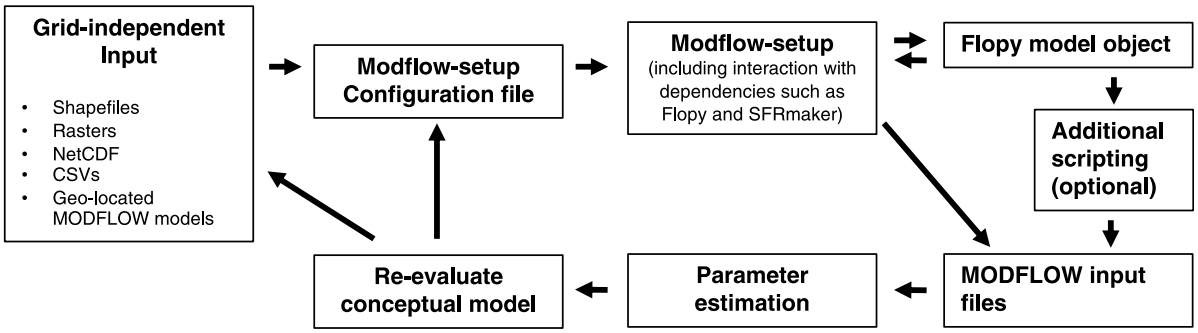

FIGURE 1
Modflow-setup in the groundwater modeling workflow. Modflow input can be written directly, or additional scripting can be performed with Flopy to customize the model. By automating the discretization of input data, Modflow-setup allows the conceptual model and model structure to be more readily revised in response to parameter estimation or new information.

MODFLOW model grids rather than geospatial coordinates. As a result, all source data must be mapped to the MODFLOW grid.

Modflow-setup provides a formal, tested, and documented Python package that builds on Flopy and the other packages referenced earlier to provide a robust, fully automated workflow for constructing MODFLOW models in a wide variety of contexts. Source data can include shapefiles, rasters, NetCDF files, and other MODFLOW models that are geolocated. We chose MODFLOW as the endpoint because it is free, open-source, easy to use (and widely used), well documented and tested, and well supported by Flopy. Modflow-setup extends the datatypes of Flopy to facilitate reading and writing MODFLOW package input and handle inter-package dependencies in memory. A key advance of Modflow-setup is the configuration file, which succinctly summarizes the data sources to a groundwater model and the methods used to process the data into model input. The information in the configuration file can be used to drive a fully automated model construction workflow, reducing the scripting needed to build a MODFLOW model to as little as a few lines of Python. This paper gives an overview of Modflow-setup, including a working example based on a published study in Wisconsin, United States (Fienen et al., 2022; 2021b).

Methods

Overview of the Modflow-setup workflow

Figure 1 illustrates the use of Modflow-setup in a groundwater modeling workflow. Grid-independent source data are preprocessed as needed and specified in a configuration file for input to Modflow-setup, along with other settings such as space and time discretization. Modflow-setup reads the configuration file, maps the input data to the model grid, and produces a modified Flopy model object. Some model inputs, such as external array text files, are written directly by Modflow-setup; other input, such as MODFLOW package input files, are written by Flopy. Prior to writing any files, additional scripting can be performed on the Flopy model object as needed, to prepare any input not supported by Modflow-setup. Parameter estimation can then be performed on the working model, which may lead to re-evaluation of the conceptual model and changes to the model structure or discretization. Modflow-setup can rapidly regenerate a new model incorporating the changes.

General paradigms

Modflow-setup supports the construction of MODFLOW 6 (Langevin et al., 2017) or MODFLOW-NWT (Niswonger et al., 2011) models from scratch (i.e., from grid-independent source data) or as an "inset" model that is coupled in one direction to a "parent" model *via* specified head or flux perimeter boundaries from the parent model solution. Parent and inset models can be mixed between MODFLOW 6 and MODFLOW-NWT. An additional "local grid refinement" (LGR) option for MODFLOW 6 models allows for specification of an inset model that is dynamically linked (in both directions) to the parent model at a finer grid resolution. Unlike previous versions of LGR (e.g., Mehl et al., 2006; Vilhelmsen et al., 2012), this inset model formulation is coupled to the parent model at the matrix level (Langevin et al., 2017), making this an efficient option for simulating both regional flow and a detailed area of interest. To facilitate array resampling and dereferencing, currently; only uniform structured grids (that may be rotated) are supported. Temporal discretization is specified in blocks that are piecewise-constant, allowing, for example, for longer spin-up periods early in a simulation, followed by a finer temporal discretization in a

period of interest. The model grid is referenced internally to a specified projected coordinate reference system (CRS; with units of feet or meters), but source data can be in any CRS; reprojection is handled automatically as needed *via* the Pyproj package (Snow et al., 2022). Similarly, length and time units can be specified for the inset and parent models, and any source data and unit conversions are handled automatically.

Time-varying specified head or specified flux boundaries can be applied to the perimeter of an inset model from a parent model solution, *via* the Constant Head and Well Packages, respectively. The parent and inset model grids need not align, but spatial alignment of the two grids can be beneficial for preserving mass when resampling recharge from one grid to another. Parent and inset model grids are located relative to one another using their respective CRS. Currently, transient inset models must have either a steady-state parent, or align temporally with a subset of the parent model stress periods.

Currently, the Streamflow Routing (SFR), Lake, and basic stress packages (Constant Head, Drain, General Head, River, and Well Packages) are fully supported for internal boundaries, with some limited support for the Multi-node Well 2 (MNW2) Package in MODFLOW-NWT. Unlike previous inset model translators such as MODTMR (Leake and Claar, 1999), internal boundary conditions are always re-discretized from their grid-independent source data (typically shapefiles), as inset models will usually carry a finer discretization than the parent. An exception is an option to translate the Well Package based on the nearest neighbor location of the model cell centers. Preparation of SFR input is handled by SFRmaker (Leaf et al., 2021). In general, the geographic extents of surface water features are specified *via* shapefiles, and any transient data such as stream inflows or pumping rates are specified *via* comma separated variable (CSV) files. Well locations can be specified with CSV or shapefiles. Transient input data are mapped to the model stress periods by computing a specified statistic (usually the mean) for values falling within each model stress period, or within a user-specified timeframe (for example, a long-term average period representing steady-state conditions).

Array-based input can be specified from rasters, shapefiles, NetCDF files, or the parent MODFLOW model. Rasters can be used to assign values to specific layers or stress periods; input is resampled to the model grid at the cell center locations using either a nearest neighbor or linear interpolation approach. Shapefiles are generally only used for delineating discrete features such as the active model area and are mapped using the `rasterize` method in the Rasterio package (Gillies, 2022b). NetCDF files provide a convenient mechanism for array-based input with many two-dimensional time slices, for example, daily estimates of net infiltration from Soil-Water-Balance code (SWB; Westenbroek et al., 2018). Similar to other transient inputs, NetCDF time slices are mapped to the model time discretization by computing period statistics. As with other data, unit conversions are performed automatically if the units are specified. Finally, arrays from a parent MODFLOW model can be resampled to the inset grid in time or space. Inset-parent layer mapping is typically specified for static inputs such as aquifer properties or cell top and bottom elevations. The coarser parent model values are then upsampled by layer to the inset model resolution, using a barycentric scheme similar to the `griddata` method in Scipy (Virtanen et al., 2020). In the case of perimeter boundary conditions, fields of head or flux components from the parent model are upsampled by stress period to the inset model grid, using the same barycentric interpolation scheme but in three dimensions, which simplifies specification of the system state along the inset model perimeter when the inset and parent grids do not align exactly (e.g., Leake and Claar, 1999).

Another key feature of Modflow-setup is the creation of MODFLOW observation input. Head observation locations can be supplied *via* a CSV file and are then mapped to the closest model cell center. Observations are set up in each model layer at the mapped locations, to allow for subsequent post-processing of model output to derive simulated head equivalents for the well open intervals (for example, using the transmissivity-based weighting functionality in Modflow-obs; https://github.com/aleaf/modflow-obs). Head observation input is created for the observation utility in MODFLOW 6 (Langevin et al., 2017) or the HYDMOD Package in MODFLOW-NWT (Hanson et al., 1999). Streamflow observations can also be supplied, with either coordinate locations or unique identifiers referencing them to a specific flowline within the input hydrography. SFRmaker will then locate the observations within the SFR package and create the relevant SFR package observation input (Leaf et al., 2021). Finally, other types of observations can be set up automatically based on lake numbers (for the Lake Package) or boundnames (for the basic stress packages in MODFLOW 6; Langevin et al., 2017).

Version control presents a fundamental challenge to reproducibility and robustness in numerical models. Even if a version control system such as Git is used to track model construction, the model files may inevitably get copied or modified outside of the Git framework, leading to confusion about their provenance. Modflow-setup records up to three levels of version information in the comment headers of the produced MODFLOW input files: 1) the Flopy version, 2) the Modflow-setup version (including the commit hash), and 3) the model version, if the model is being tracked by Git, or if a version is specified in the configuration file. In the former case, Git versioning information is read by Modflow-setup using an approach similar to the Versioneer package (https://github.com/python-versioneer/python-versioneer). This way, the methods used to generate a particular model input file can be understood and reproduced, even if the code base and model have changed.

Software implementation

Modflow-setup is implemented as a Python package that works on Linux, OSX, or Windows. The version of the code documented in

this study is available as a USGS software release (Leaf et al., 2022); the current development version that incorporates bug fixes and other improvements is available through GitHub (https://github.com/doi-usgs/modflow-setup) or the Python Package Index (PyPI). It should be noted that Modflow-setup has software dependencies that must be installed prior to its use. Detailed instructions on how to install the dependencies and Modflow-setup are available in the online documentation (https://doi-usgs.github.io/modflow-setup). Similar to any Python package, Modflow-setup consists of objects that can be imported into a Python session and therefore used within scripts or other Python code. The use of Modflow-setup does not require extensive knowledge of Python, however. In the simplest use case, input can be specified in a configuration file, and a MODFLOW model can be built from the configuration file using only a few lines of Python, as illustrated in the example.

Code organization

The core function of Modflow-setup is to automate mapping of disparate, grid-independent data to a finite difference grid. At a basic level, the Modflow-setup package houses general objects (functions, classes, and methods) to do this. The objects are organized into modules that loosely correspond to the components of a MODFLOW model (e.g., "oc.py" for output control) or specific functionality (e.g., "interpolate.py" for interpolation). Ideally, each module has a corresponding test module in the "tests" folder, and each object a corresponding test within that test module. In practice, much of the testing follows an integration approach where entire packages or models are built within a single test, effectively testing the interactions of multiple objects at once.

While Modflow-setup may evolve to include more functionality as a library of stand-alone components, the current development focus is on an integrated workflow that builds Flopy model objects from information provided in a configuration file. Three model classes, each contained in their own module, are central to this focus. The `MF6model` and `MFnwtModel` classes subclass the Flopy `ModflowGwf` and `Modflow` classes, respectively, to add additional model construction functionality for MODFLOW 6 and MODFLOW-NWT models. Both `MF6model` and `MFnwtModel` also subclass a shared `MFsetupMixin` class that contains core functionality common to any MODFLOW version. The model classes themselves contain a number of methods centered around various arrays and packages, which, in turn, interact with functions and other objects in the remaining Modflow-setup modules.

The configuration file

Most user interaction with Modflow-setup is through the configuration file, which is specified in the YAML format (yaml.org). YAML maps key:value pairs similar to a Python dictionary (https://docs.python.org/3/tutorial/datastructures.html), except that whitespace and newlines can often be used in the place of commas and brackets to delimit structures. YAML input in the configuration file is organized into blocks that generally follow the MODFLOW input structure, with primary blocks representing specific MODFLOW packages or model components, and sub-blocks representing MODFLOW 6 input blocks or features in Modflow-setup. The naming of blocks and variables is intended to follow MODFLOW and Flopy conventions as closely as possible, with MODFLOW given preference where these conflict. For example, this block (from the example problem discussed below) describes the MODFLOW 6 simulation:

```yaml
simulation:
  sim_name: 'pleasant_lgr'
  version: 'mf6'
  sim_ws: 'pleasant_lgr/'
```

In the model setup workflow, input from the configuration file is loaded by Modflow-setup into a configuration dictionary attached to the model object. For example, the `simulation:` block shown above would be loaded as:

```
cfg['simulation'] = {'sim_name: 'mfsim',
                     'version': 'mf6',
                     'sim_ws': 'pleasant_lgr/'}
```

The above dictionary would then be fed to the Flopy `MFSimulation` class constructor to create a simulation instance. Within package blocks, input to MODFLOW can be specified directly using the appropriate variables and structures described in the MODFLOW input instructions (Niswonger et al., 2011; Langevin et al., 2017). For example, in the block below, the dimensions: and griddata: sub-blocks would be fed directly to the MODFLOW 6 Discretization Package constructor in Flopy:

```yaml
dis:
  options:
    length_units: 'meters'
  dimensions:
    nlay: 2
    nrow: 30
    ncol: 35
  griddata:
    delr: 1000.
    delc: 1000.
    top: 2.
    botm: [1, 0]
```

Such direct input might also contain paths to external text file arrays that are consistent with the model grid. Alternatively, *source_data*: sub-blocks can be used to reference grid-independent data (shapefiles, rasters, or comma separated variable files, etc.) that need to be mapped to the model grid. The Pleasant Lake example described below includes a DIS package block that references GeoTIFF rasters as input for layer tops and bottoms. More details on configuration file input options are available in the online documentation (https://doi-usgs.

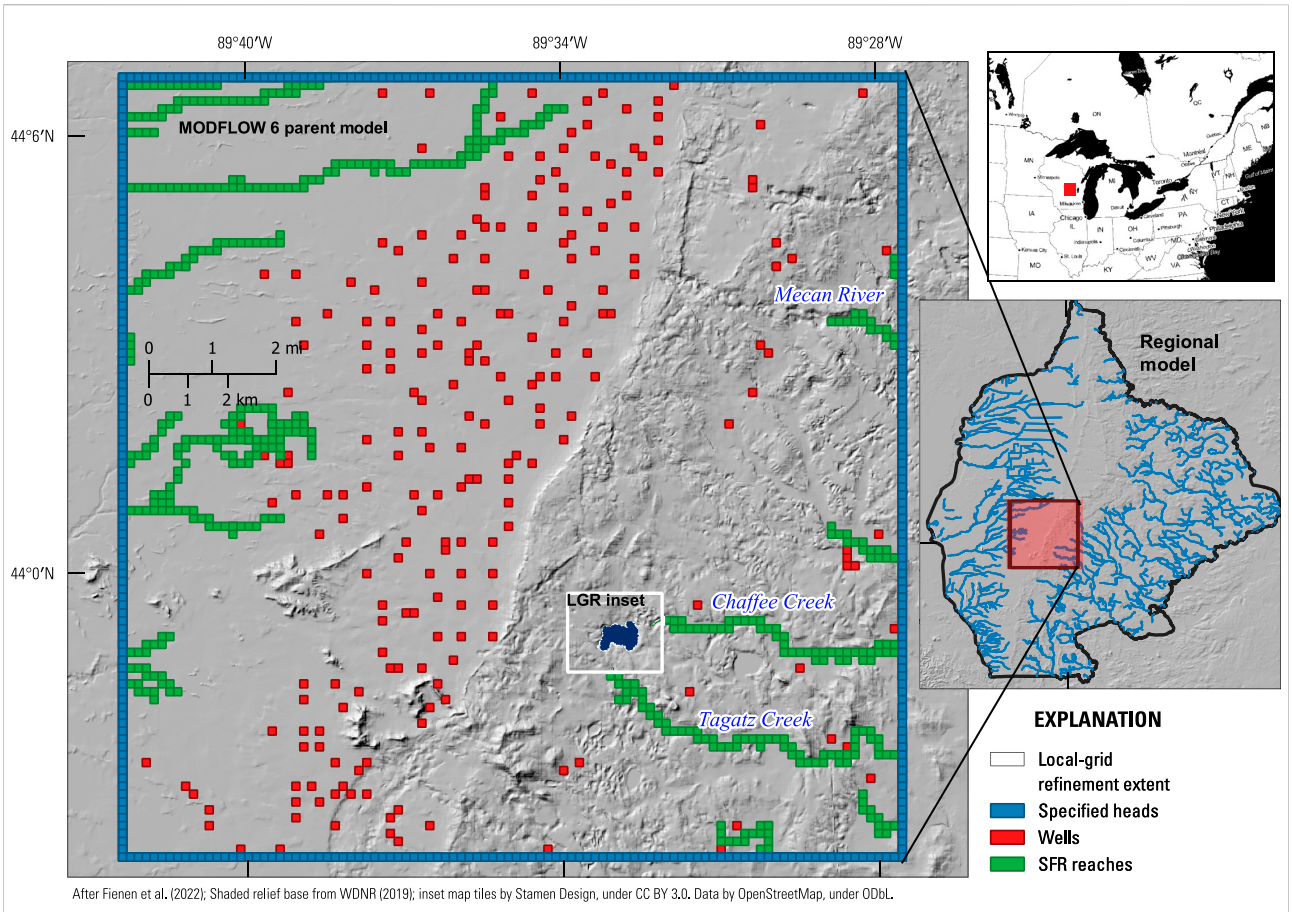

FIGURE 2
The full Pleasant Lake model domain with location map, showing the relationship between the regional, intermediate and LGR inset models, as well as the irrigation wells considered.

github.io/modflow-setup), which includes a gallery of working configuration files for various models in the Modflow-setup test suite.

An example model build script

With an appropriate configuration file, a Python script to build a MODFLOW model can be as simple as the following three lines of Python:

```
from mfsetup import MF6model
model = MF6model.setup_from_yaml('config_file.yml')
model.write_input()
```

In this example, the model object class is imported, similar to Flopy, and the `setup_from_yaml` constructor method is called with the configuration file. An `MF6model` instance, which is essentially a Flopy model object with additional functionality, is returned. The `MF6model` instance can be used to write the model input files or as the basis for additional custom scripting.

Example: Setup of the Pleasant Lake model

The Pleasant Lake model (Fienen et al., 2022) is a MODFLOW 6 simulation that was constructed using Modflow-setup. We show this example both because this project motivated the development of the Modflow-setup code and because it highlights a complex workflow that benefits greatly from the scripting approach. A simplified version of this workflow with a smaller model domain is available on the Modflow-setup GitHub site (https://github.com/doi-usgs/modflow-setup); the published, fully detailed models for Pleasant Lake are available from Fienen et al. (2021b). Another worked example including uncertainty analysis and a decision support outcome is available from Fienen and Corson-Dosch (2021; https://github.com/usgs/neversink_workflow).

The goal of the Pleasant Lake model, part of the Central Sands Lake Study (Fienen et al., 2022), was to address connections between groundwater abstraction and the ecological function of a lake in central Wisconsin, United States (WDNR 2021; Figure 2). This required modeling at multiple scales. Fine discretization was needed near the

lake for accurate simulation of water levels and groundwater–lake flux. A large model domain was also needed to simulate farfield water-use activity (chiefly irrigated agriculture), in order to delineate a limit of connection, as well as to incorporate distant hydrologic boundaries. Adopting a fine enough discretization for the lake detail throughout the farfield would have resulted in a model with more cells than could be practically managed. To mitigate this, three models were combined: a large regional model built with MODFLOW-NWT, an intermediate MODFLOW 6 model inset within the regional model to simulate the irrigated agriculture area, and a refined MODFLOW 6 inset model (nested within the intermediate model) to simulate the lake (Figure 2; Fienen et al., 2022). Regional groundwater flow and the effects of distant boundaries were simulated with the MODFLOW-NWT model, which was coupled sequentially (one-way) to the MODFLOW 6 models through time-varying specified head boundaries along the intermediate MODFLOW 6 model perimeter. The two MODFLOW 6 models were coupled dynamically (both ways) within the groundwater flow solution, allowing for feedback between the models. Estimates of groundwater recharge for the MODFLOW models were provided by a SWB simulation that could represent alternative assumptions of climate and land use. Net infiltration estimates from the SWB model in the NetCDF format were read directly by Modflow-setup to produce Recharge Packages for the MODFLOW models. Climate-based estimates of irrigation demand from SWB were also passed to the Well Package for simulations that considered future scenarios.

The MODFLOW 6 models were set up using the Modflow-setup LGR feature, which uses the LGR utility in Flopy (Bakker et al., 2016) to create input for the Groundwater Flow Exchange Package, which dynamically links MODFLOW 6 models within the same matrix solution *via* fluxes across their shared boundaries (Langevin et al., 2017). The Modflow-setup LGR feature also sets up the Water Mover Package to maintain continuity in the SFR Package streamflow across the linked model boundaries. Construction of the LGR inset model is activated by an `lgr:` sublock within the parent model configuration file, which points to a second configuration file for the LGR inset. Full versions of the parent and inset model configuration files for the example are available in the online documentation (https://doi-usgs.github.io/modflow-setup). An abbreviated version of the example LGR inset model configuration file is reproduced in snippets here for illustration.

As noted earlier, the `simulation:` block provides input to the Flopy MFSimulation constructor, and, critically, the `version:` argument that also tells Modflow-setup which version of MODFLOW to use. Similarly, the `model:` block contains input to the Flopy `ModflowGwf` constructor and ultimately, the MODFLOW 6 Name file (Langevin et al., 2017). The `packages:` argument tells Modflow-setup which packages to build. Since this model is an LGR inset, the parent model is already known to Modflow-setup and does not need to be re-specified here. Similarly, any packages included in the package list, but not specified in the inset model configuration file, are simply built (on the inset model grid) from the input in the parent model configuration file.

```yaml
simulation:
  sim_name: 'pleasant_lgr'
  version: 'mf6'
  sim_ws: 'pleasant_lgr/'

model:
  simulation: 'pleasant_lgr'
  modelname: 'plsnt_lgr_inset'
  options:
    print_input: True
    save_flows: True
    newton: True
    newton_under_relaxation: True
  packages: ['dis', 'ic', 'npf', 'oc', 'sto', 'rch', 'sfr',
             'lak', 'obs', 'wel', 'ims']
```

The `setup_grid:` block specifies the orientation and discretization of the LGR inset grid. Model grids in Modflow-setup can be defined explicitly or using a buffer around a feature of interest. If the model is associated with a parent model, the model discretization is aligned with the parent model grid by default (this is required for LGR models). A `snap_to_parent:` option allows for unaligned grids. Unrotated models with a grid spacing that is a factor of 1,000 m can also be aligned with the National Hydrogeologic Grid, a framework intended to facilitate the development and use of national-scale hydrogeologic datasets in the United States (Clark et al., 2018).

In this case, a polygon for Pleasant Lake is provided *via* a shapefile, and Modflow-setup is instructed to create a regular 40-m mesh within a 1000-m buffer of the lake. The projected CRS for the model grid is Wisconsin Transverse Mercator (indicated by EPSG code 3070). The vertical discretization is specified in a `dis:` (Discretization Package) block. A digital elevation model (DEM) in units of meters is specified for the model top. As in Python, numbering for layers or stress periods is zero-based. Since no bottom elevation grid is supplied for layer 0, the bottom of that layer will be set halfway between the model top and the specified bottom of layer 1. Additional layers could be similarly subdivided by specifying the desired layer number for the next bottom surface elevation.

```yaml
setup_grid:
  source_data:
    features_shapefile:
      filename: 'data/pleasant/source_data/shps/all_lakes.shp'
      id_column: 'HYDROID'
      include_ids: [600059060]
  dxy: 40
  buffer: 1000
  epsg: 3070

dis:
  options:
    length_units: 'meters'
  dimensions:
    nlay: 5
  source_data:

top:
  filename: 'data/pleasant/source_data/rasters/dem40m.tif'
  elevation_units: 'meters'
botm:
  filenames:
    1: 'data/pleasant/source_data/rasters/botm0.tif'
    2: 'data/pleasant/source_data/rasters/botm1.tif'
    3: 'data/pleasant/source_data/rasters/botm2.tif'
    4: 'data/pleasant/source_data/rasters/botm3.tif'
```

The Lake Package (`lak:`) block includes shapefile input to delineate the horizontal extent of the lake, and optionally, a

bathymetry_raster: input to delineate bottom depths that are subtracted off the initial model top (which is assumed to represent the water surface, typically the case for DEMs). Alternatively, a *stage_area_volume_file:* can be specified to allow for more accurate representation of lake volume and surface area as lake levels change. Initial values for lakebed leakance can be input for both a littoral zone extending a specified distance from shore around the perimeter of the lake, and a lower permeability profundal zone in the lake interior (e.g., after Hunt et al., 2013; Leaf and Haserodt, 2020). Finally, climate input, including daily precipitation and mean air temperatures, are supplied in a text file downloaded from the PRISM Climate Group (2019); PRISM provides modeled climate time series at any point location within the United States. Precipitation is used directly by the Lake Package to compute the lake water balance; Modflow-setup uses the Hamon (1961) method to convert daily mean air temperatures to estimates of lake surface evaporation (Harwell, 2012). A *period_stats:* sub-block specifies how the climate input should be aggregated to the model stress periods. For the initial steady-state period, 2012–2018 averages of the daily precipitation and lake surface evaporation are used; subsequently, the average values within each monthly stress period are used. Alternatively, lake climate information can be input directly or supplied in a general CSV format.

```
lak:
  options:
    boundnames: True
    save_flows: True
    surfdep: 0.1
  source_data:
    littoral_leakance: 0.045 # 1/d
    profundal_leakance: 0.025 # 1/d
    littoral_buffer_zone_width: 40
    lakes_shapefile:
      filename: 'data/pleasant/source_data/shps/all_lakes.shp'
      id_column: 'HYDROID'
      include_ids: [600059060] # pleasant lake
    climate:
      filenames:
        600059060: 'data/pleasant/source_data/PRISM_ppt_tmean_stable_4km.csv'
      format: 'prism'
      period_stats:
        0: ['mean', '2012-01-01', '2018-12-31']
        1: 'mean'
    bathymetry_raster:
      filename: 'data/pleasant/source_data/rasters/pleasant_bathymetry.tif'
      length_units: 'meters'
    stage_area_volume_file:
      filename: 'data/pleasant/source_data/tables/area_stage_vol_Pleasant.csv'
      length_units: 'meters'
        id_column: 'hydroid'
        column_mappings:
          volume_m3: 'volume'
    external_files: False  # option to write connectiondata table to external file
```

The *sfr:* block instructs Modflow-setup to generate an SFR Package for the LGR inset model area, using SFRmaker (Leaf et al., 2021). Since this is an LGR inset model, Modflow-setup will automatically set up the Water Mover Package as needed to connect the SFR network across the boundary with the enclosing parent model.

```
sfr:
  options:
    save_flows: True
  source_data:
    flowlines:
      nhdplus_paths: ['data/pleasant/source_data/shps']
    dem:
      filename: 'data/pleasant/source_data/rasters/dem40m.tif'
      elevation_units: 'meters'
  sfrmaker_options:
    set_streambed_top_elevations_from_dem: True
```

To simplify input as much as possible, Modflow-setup includes configuration files of default settings for MODFLOW 6 and MODFLOW-NWT models. In constructing a model, the default configuration files are read first, and the settings within them are recursively updated with user-specified input. Therefore, many settings are optional. For example, `save_flows: True` in the *sfr:* block earlier is also specified in the default configuration, making it technically redundant, although perhaps useful as a placeholder to turn the setting on or off. Other examples of default configurations include the Output Control Package, which is generated by default to save output on the last timestep of each stress period, and initial heads, which are set to the model top by default if no configuration is specified. The default configuration files can be viewed in the online documentation.

The *obs:* block here illustrates how head observation locations can be supplied from multiple CSV files. In this case, no x and y column arguments are needed, because both files have the default column names of "x" and "y." Non-default column names can be specified with the *column_mappings:* argument. In this example, the column names "obsprefix" and "common_name" are mapped to the default "obsname," column for observation names.

```
obs:
  source_data:
    filenames: ['data/pleasant/source_data/tables/lake_sites.csv'
                'data/pleasant/source_data/tables/wdnr_gw_sites.csv']
    column_mappings:
      obsname: ['obsprefix', 'common_name']
  drop_observations: ['10019209_lk']
```

Since this is an LGR inset model that shares a MODFLOW 6 simulation with the enclosing parent model, the simulation-level Temporal Discretization and Iterative Model Solution Packages are specified in the parent model configuration file. The remaining unspecified packages (Initial Conditions, Output Control, Node Property Flow, Storage, Recharge, and Well packages) are generated for the LGR inset model using the input blocks specified for the parent (MODFLOW 6) model, as described previously. The simplified example version of the Pleasant Lake model from the online documentation is shown in Figure 3.

Discussion

After setting up this framework for model construction and linkage, it is straightforward to evaluate some of the many decisions that are often made once in a modeling workflow and not revisited again, such as spatial discretization, time discretization, changing data sources, or hypothesis testing. In the Pleasant Lake example, the key goal of establishing a causal connection between human water use (chiefly irrigation abstraction) and lake levels required evaluation of multiple conditions. Under a unified representative climate, we evaluated recharge and irrigation-required water abstraction for three land use scenarios: 1) no irrigated agriculture, 2) irrigated agriculture in the footprint of current conditions, and 3) potential maximum irrigated agriculture.

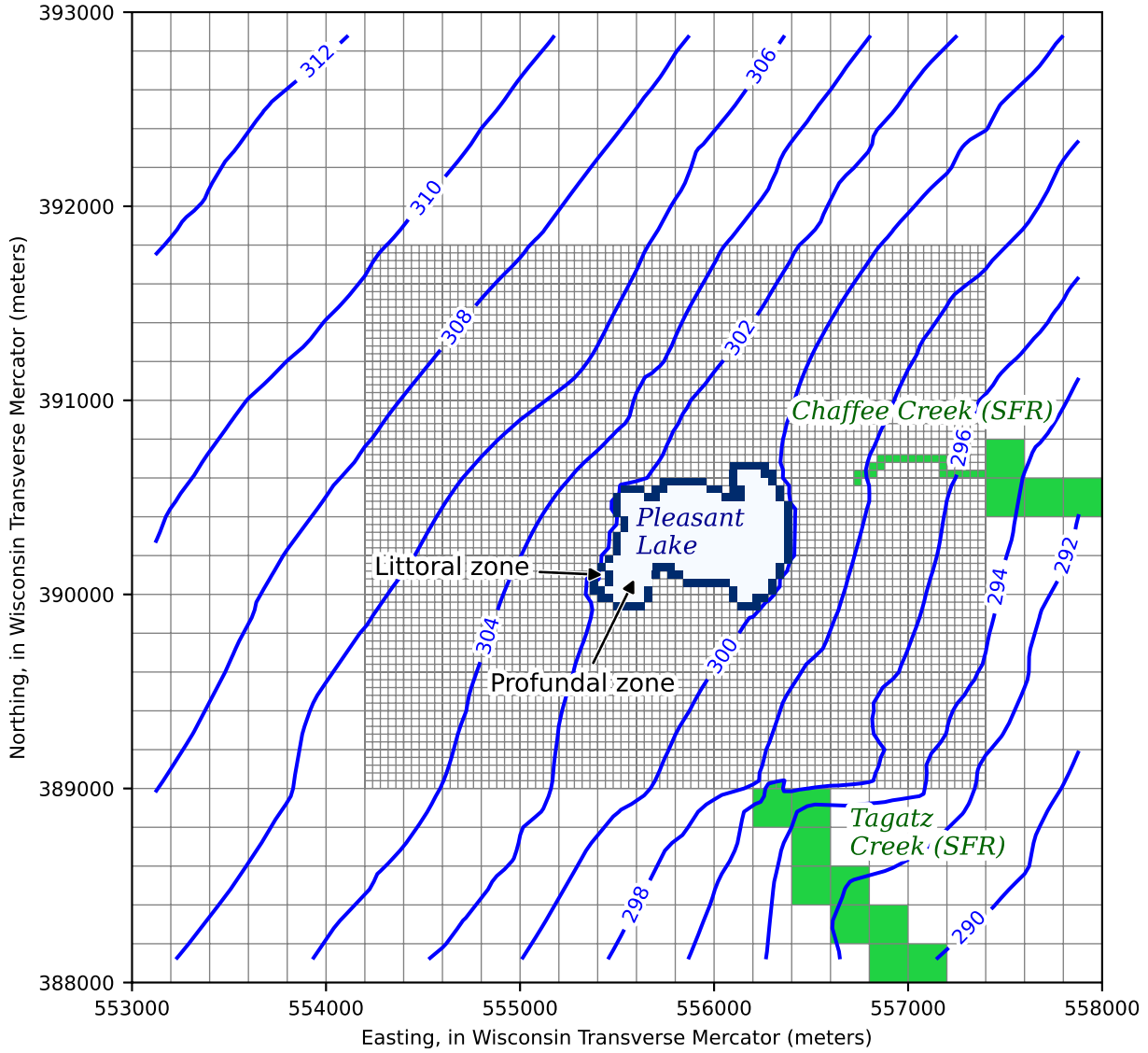

FIGURE 3
Close-up of the example local grid refinement inset model configuration, showing the discretization and combined water table solution, as well as Streamflow Routing (SFR) Package cells and the littoral and profundal lakebed leakance zones.

Consideration of these multiple hypotheses required multiple instances of the SWB model, with the outputs from each instance ingested as recharge and water-use inputs to multiple MODFLOW 6 models. The robust and repeatable nature of the Modflow-setup framework enabled efficient evaluation of the scenarios and has yielded similar benefits to other projects involving multiple numerical models or advanced analyses (e.g., Fienen et al., 2022; 2021a).

As is typical in modeling projects, a single iteration of this workflow was insufficient, as all modeling requires refinement of datasets, testing of hypotheses, and incorporation of lessons learned (e.g., Anderson et al., 2015). For example, examination of model history matching results pointed to the need to better represent headwater springs near the lake. This required rebuilding the SFR package, a task that would be prohibitively time-consuming in a traditional modeling workflow but that was easily done with Modflow-setup. In addition, Modflow-setup allowed for multiple updates to the layering and geological structures represented in the model as new data became available during the course of the project. By opening the numerical model structure to testing and improvement, the

automated workflow enabled by the Modflow-setup can maximize the assimilation of data and ultimately provide models that are better suited for decision support.

It is important to note that while Modflow-setup aims to be general, development is ongoing on the project GitHub site, and to date has focused primarily on meeting project needs through iterative improvement, instead of building a comprehensive tool from the ground up. Some features are incomplete, and others haven't been developed yet. While the configuration file interface is mostly established, it may change somewhat going forward to accommodate new features or improve the user experience. The internal code structure is almost certain to change. While the online documentation is also a work in progress, it aims to accurately describe the current state of the project and how to use it. Contributions and ideas at all levels are encouraged and can be submitted through issues and pull requests on the project GitHub page, or *via* email. In any case, the integration of Modflow-setup with the general Python interface provided by Flopy allows for custom code to be added to a model construction workflow as needed.

Finally, like many open-source software projects, Modflow-setup depends on a large "stack" of other software that is constantly changing. Regular continuous integration testing helps ensure functionality by executing the test suite in freshly built Python environments encompassing the last two minor versions of Python (e.g., 3.10 and 3.9), across the supported platforms. For reproducibility, a project-specific Python environment built from a configuration file works well (for example, a Conda environment file; https://docs.conda.io/). Long-term archives that are meant to persist over years may consider packaging this environment into a stand-alone Python distribution, for example using Conda-pack (https://github.com/conda/conda-pack).

Conclusions

Modflow-setup provides a rapid, reproducible, and robust framework for building MODFLOW models from grid-independent source data. Common model construction tasks are distilled in an open-source, online code base that is tested and extensible through collaborative version control. The workflow for building the model—including input data, construction options, and output packages—is summarized in a single configuration file in the human-readable YAML format. Integration with Flopy allows for additional customization of the model construction workflow as needed. The benefits of Modflow-setup include reduced time and labor required to build a groundwater model, reduced potential for error, improved reproducibility, expanded ability to explore alternative conceptual models or hypotheses, and a reduction in cognitive load that allows the modeler to focus on the most important aspects of the analysis. In the case of the Pleasant Lake model, the robust automation enabled by Modflow-setup allowed for efficient exploration of cumulative pumping impacts to lake levels from hundreds of wells, across multiple scenarios.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

AL has led the code development of Modflow-setup, and contributed approximately 2/3 of the writing to the manuscript. MF has contributed code and ideas to the Modflow-setup project and has tested/applied it extensively in the project work that is cited in the manuscript. MF contributed approximately 1/3 of the writing in the manuscript.

Funding

This project was funded by the U.S. Geological Survey Mississippi Alluvial Plain Project, the Wisconsin Department of Natural Resources, and the Aarhus University Department of Geoscience.

Acknowledgments

The authors would like to thank Aaron Pruitt, Jonathan Traylor, Leslie Duncan, Meg Haserodt, Moussa Guira, Nick Corson-Dosch, Rasmus Frederiksen, Troels Vilhelmsen, and J.R. Rigby for their enthusiastic support and beta testing of Modflow-setup. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anderson, M. P., Woessner, W. W., and Hunt, R. J. (2015). *Applied groundwater modeling*. Second Edition. San Diego: Academic Press.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting modflow model development using python and flopy. *Groundwater* 54 (5), 733–739. doi:10.1111/gwat.12413

Clark, B. R., Barlow, P. M., Peterson, S. M., Hughes, J. D., Reeves, H. W., and Viger, R. J. (2018). *National-scale grid to support regional groundwater availability studies and a national hydrogeologic database*. New York: U.S. Geological Survey data release. doi:10.5066/F7P84B24

Doherty, J., and Moore, C. (2020). Decision support modeling: Data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater* 58, 327–337. doi:10.1111/gwat.12969

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., and Stodden, V. (2008). Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* 11 (1), 8–18. doi:10.1109/MCSE.2009.15

Fienen, M., and Corson-Dosch, N. (2021). *Groundwater model archive and workflow for neversink/rondout basin*. New York: Source Water Delineation: U.S. Geological Survey Data Release. doi:10.5066/P9HWSOHP

Fienen, M. N., and Bakker, M. (2016). Hess opinions: Repeatable research: What hydrologists can learn from the duke cancer research scandal. *Hydrol. Earth Syst. Sci.* 20 (9), 3739–3743. doi:10.5194/hess-20-3739-2016

Fienen, M. N., Corson-Dosch, N. T., White, J. T., Leaf, A. T., and Hunt, R. J. (2021a). Risk-based wellhead protection decision support: A repeatable workflow approach. *Groundwater* 60, 71–86. doi:10.1111/gwat.13129

Fienen, M. N., Haserodt, M. J., and Leaf, A. T. (2021b). *MODFLOW models used to simulate groundwater flow in the Wisconsin Central Sands Study Area, 2012-2018*. New York: U.S. Geological Survey Data Release. doi:10.5066/P9BVFSGJ

Fienen, M. N., Haserodt, M. J., Leaf, A. T., and Westenbroek, S. M. (2022). *Simulation of regional groundwater flow and groundwater/lake interactions in the central Sands, Wisconsin*. U.S. Geological Survey Scientific Investigations Report 2022-5046. doi:10.3133/sir20225046

Fisher, J. C., Bartolino, J. R., Wylie, A. H., Sukow, J., and McVay, M. (2016). *Groundwater-flow model of the Wood River Valley aquifer system, south-central Idaho*. U.S. Geological Survey Scientific Investigations Report 2016– 5080, 71. doi:10.3133/sir20165080

Gillies, S. (2022b). Rasterio: Access to geospatial raster data. Available at: https://rasterio.readthedocs.io/en/latest/(Accessed January 28, 2022).

Gillies, S. (2022c). Rtree: Spatial indexing for python. Available at: https://toblerity.org/rtree/(Accessed January 28, 2022).

Gillies, S. (2022a). The fiona user manual. Available at: https://fiona.readthedocs.io/en/latest/manual.html (Accessed January 28, 2022).

Gillies, S. (2022d). The shapley user manual. Available at: https://shapely.readthedocs.io/en/latest/manual.html (Accessed January 28, 2022).

Haitjema, H. M. (1995). *Analytic element modeling of groundwater flow*. San Diego, California: Academic Press.

Hamon, W. R. (1961). Estimating potential evapotranspiration: Journal of hydraulics division. *J. Hydr. Div.* 87, 107–120. doi:10.1061/JYCEAJ.0000599

Hanson, R. T., and Leake, S. A. (1999). *Documentation for HYDMOD, a program for extracting and processing time-series data from the U.S. Geological Survey's modular three-dimensional finite-difference ground-water flow model*. U.S. Geological Survey Open-File Report 98-564, 57. doi:10.3133/ofr98564

Harwell, G. R. (2012). *Estimation of evaporation from open water—a review of selected studies, summary of U.S. Army Corps of Engineers data collection and methods, and evaluation of two methods for estimation of evaporation from five reservoirs in Texas*. U.S. Geological Survey Scientific Investigations Report 2012– 5202, 96. doi:10.3133/sir20125202

Hoyer, S., and Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *J. Open Res. Softw.* 5 (1), 10. doi:10.5334/jors.148

Hunt, R. J., Doherty, J., and Tonkin, M. J. (2007). Are models too simple? *Ground Water* 45 (3), 254–262. doi:10.1111/j.1745-6584.2007.00316.x

Hunt, R. J., Walker, J. F., Selbig, W. R., Westenbroek, S. M., and Regan, R. S. (2013). *Simulation of climate-change effects on streamflow, lake water budgets, and stream temperature using GSFLOW and SNTEMP, Trout Lake Watershed, Wisconsin*. U.S. Geological Survey Scientific Investigations Report 2013– 5159, 118. doi:10.3133/sir20135159

Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., and Provost, A. M. (2017). *Documentation for the MODFLOW 6 groundwater flow model*. U.S. Geological Survey Techniques and Methods, book 6, 197. chap. A55. doi:10.3133/tm6A55

Leaf, A. T., Fienen, M. N., Hunt, R. J., and Buchwald, C. A. (2015). *Groundwater/surface-water interactions in the bad river watershed, Wisconsin*. U.S. Geological Survey Scientific Investigations Report 2015– 5162, 110. doi:10.3133/sir20155162

Leaf, A. T., and Fienen, M. N. (2022). *Modflow-setup version 0.1*. U.S. Geological Survey Software Release, 1 Aug. 2022. doi:10.5066/P9O3QWQ1

Leaf, A. T., Fienen, M. N., and Reeves, H. W. (2021). SFRmaker and linesink-maker: Rapid construction of streamflow routing networks from hydrography data. *Groundwater* 59, 761–771. doi:10.1111/gwat.13095

Leaf, A. T., and Haserodt, M. J., 2020, Hydrology of haskell lake and investigation of a groundwater contamination plume, lac du Flambeau reservation, Wisconsin: U.S. Geological Survey Scientific Investigations Report 2020– 5024, 79. doi:10.3133/sir20205024

Leake, S. A., and Claar, D. V. (1999). *Procedures and computer programs for telescopic mesh refinement using MODFLOW*. U.S. Geological Survey Open-File Report 99-238, 53. doi:10.3133/ofr99238

Mehl, S., Hill, M. C., and Leake, S. A. (2006). Comparison of local grid refinement methods for MODFLOW. *Ground Water* 44, 792–796. doi:10.1111/j.1745-6584.2006.00192.x

Moore, C., and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resour. Res.* 41, W05020. doi:10.1029/2004WR003501

Moran, T. (2016). *Projecting forward: A framework for groundwater model development under the sustainable groundwater management act*. Stanford, CA, USA: Stanford Woods Institute for the Environment, 56. Available at: https://waterinthewest.stanford.edu/sites/default/files/Groundwater-Model-Report.pdf (Accessed July 25, 2022).

Niswonger, R. G., Panday, S., and Ibaraki, M., 2011, *MODFLOW–NWT—a Newton formulation for MODFLOW–2005*. U.S. Geological Survey Techniques and Methods, book 6, 44. chap. A37. doi:10.3133/tm6A45

Peng, R. D. (2011). Reproducible research in computational science. *Science* 334 (6060), 1226–1227. doi:10.1126/science.1213847

PRISM Climate GroupOregon State University (2019). Time series values for individual locations. Available at: https://prism.oregonstate.edu/explorer/(Accessed December 10, 2019).

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at:http://www.R-project.org/(Accessed March 9, 2016).

Snow, A. D., Whitaker, J., et al. (2020). *Pyproj documentation*. Available at:https://pyproj4.github.io/pyproj/stable/ (Accessed August 5, 2020).

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* 12 (2), 257–285. doi:10.1207/s15516709cog1202_4

Vilhelmsen, T. N., Christensen, S., and Mehl, S. W. (2012). Evaluation of MODFLOW-LGR in connection with a synthetic regional-scale model. *Ground Water* 50, 118–132. doi:10.1111/j.1745-6584.2011.00826.x

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

Westenbroek, S. M., Engott, J. A., Kelson, V. A., and Hunt, R. J. (2018). *SWB Version 2.0—a soil-water-balance code for estimating net infiltration and other water-budget components*. U.S. Geological Survey Techniques and Methods, book 6, 118. chap. A59. doi:10.3133/tm6A59

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50, 1152–1173. doi:10.1002/2013WR014767

White, J. T., Hemmings, B., Fienen, M. N., and Knowling, M. J. (2021). Towards improved environmental modeling outcomes: Enabling low-cost access to high-dimensional, geostatistical-based decision-support analyses. *Environ. Model. Softw.* 139, 105022. doi:10.1016/j.envsoft.2021.105022

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., et al. (2014). Best practices for scientific computing. *PLoS Biol.* 12 (1), e1001745. doi:10.1371/journal.pbio.1001745

Wisconsin Department of Natural Resources (WDNR) (2021). Central Sands Lake study report: Findings and recommendations. *Rep. Wis. State Legislature*. doi:10.5281/zenodo.5708791

Wisconsin Department of Natural Resources [WDNR] (2019). Digital elevation model (DEM) – 10 meter. Available at: https://data-wi-dnr.opendata.arcgis.com/search?q=DEM (Accessed July 25, 2022).

frontiers | Frontiers in Earth Science

TYPE Technology and Code
PUBLISHED 30 September 2022
DOI 10.3389/feart.2022.884766

OPEN ACCESS

EDITED BY
Anneli Guthke,
University of Stuttgart, Germany

REVIEWED BY
Laurene Bouaziz,
Deltares, Netherlands
Julianne Quinn,
University of Virginia, United States

*CORRESPONDENCE
Edom Moges,
edom.moges@berkeley.edu

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 27 February 2022
ACCEPTED 26 July 2022
PUBLISHED 30 September 2022

CITATION
Moges E, Ruddell BL, Zhang L, Driscoll JM, Norton P, Perez F and Larsen LG (2022), HydroBench: Jupyter supported reproducible hydrological model benchmarking and diagnostic tool.
Front. Earth Sci. 10:884766.
doi: 10.3389/feart.2022.884766

HydroBench: Jupyter supported reproducible hydrological model benchmarking and diagnostic tool

Edom Moges[1]*, Benjamin L. Ruddell[2], Liang Zhang[1], Jessica M. Driscoll[3], Parker Norton[3], Fernando Perez[1] and Laurel G. Larsen[1]

[1]University of California, Berkeley, Berkeley, CA, United States, [2]Northern Arizona University, Flagstaff, AZ, United States, [3]U.S. Geological Survey, Denver, CO, United States

Evaluating whether hydrological models are right for the right reasons demands reproducible model benchmarking and diagnostics that evaluate not just statistical predictive model performance but also internal processes. Such model benchmarking and diagnostic efforts will benefit from standardized methods and ready-to-use toolkits. Using the Jupyter platform, this work presents HydroBench, a model-agnostic benchmarking tool consisting of three sets of metrics: 1) common statistical predictive measures, 2) hydrological signature-based process metrics, including a new time-linked flow duration curve and 3) information-theoretic diagnostics that measure the flow of information among model variables. As a test case, HydroBench was applied to compare two model products (calibrated and uncalibrated) of the National Hydrologic Model - Precipitation Runoff Modeling System (NHM-PRMS) at the Cedar River watershed, WA, United States. Although the uncalibrated model has the highest predictive performance, particularly for high flows, the signature-based diagnostics showed that the model overestimates low flows and poorly represents the recession processes. Elucidating why low flows may have been overestimated, the information-theoretic diagnostics indicated a higher flow of information from precipitation to snowmelt to streamflow in the uncalibrated model compared to the calibrated model, where information flowed more directly from precipitation to streamflow. This test case demonstrated the capability of HydroBench in process diagnostics and model predictive and functional performance evaluations, along with their tradeoffs. Having such a model benchmarking tool not only provides modelers with a comprehensive model evaluation system but also provides an open-source tool that can further be developed by the hydrological community.

KEYWORDS

Hydrological Modeling, Model Evaluation, Model Benchmarking, Model Diagnostics, Uncertainty Analysis, Nash Sutcliffe, Kling-Gupta, Reproducibility

Introduction

Supported by advances in computational capacity, there is a proliferation of hydrological models ranging from simple black box data-driven models to complex integrated models. Similarly, the application of these models ranges from local to regional and continental-domain hydrological decision support tools. In this regard, the U.S. Geological Survey's National Hydrologic Model-Precipitation Runoff Modeling System (NHM-PRMS) (Regan et al., 2018, 2019) and National Oceanic and Atmospheric Administration's National Water Model (Cohen et al., 2018) are examples of continental-domain models that strive to address national-scale water balance, water supply, and flood risk analyses. Although model adoption can be more of a function of legacy than adequacy, models' reliability rests on performance evaluation (Adorr and Melsen, 2019). Performance evaluation, which includes model benchmarking and diagnostic efforts, benefits from standardized methods and ready-to-use toolkits that implement those methods (Kollet et al., 2017; Nearing et al., 2018; Lane et al., 2019; Saxe et al., 2021; Tijerina et al., 2021). Standardized methods and toolkits also help modeling communities and model users build trust in a model's operational reliability. As such, having a ready-to-use, organized, and comprehensive model-agnostic (i.e., model-independent) benchmarking tool is critical for advancing modeling communities and modeling practice.

Hydrologic model performance evaluations often rely on statistical metrics such as Nash-Sutcliffe efficiency and correlation coefficient. However, as these metrics are indicative of focused aspects of model performance, there is a call of comprehensive model evaluation that includes process-based model diagnostics (Gupta et al., 2008; McMillan, 2020, 2021) and functional model evaluations (Weijs et al., 2010; Ruddell et al., 2019). Process-based model diagnostics evaluate the hydrological consistency of the model with observations (e.g., through examination of hydrological signatures that capture dominant processes), while the functional model performance evaluation focuses on the interactions or information flows among internal flux and state variables (e.g., uncertainty reduction of streamflow by precipitation data). Thus, a comprehensive model benchmarking tool may need to include at least three types of metrics that 1) quantify model predictive performances by comparing observations and their corresponding model outputs, 2) reveal hydrological process consistency and 3) assess the functional performance of the model. As a whole, such a benchmarking practice helps evaluate not only predictive performance but also reveals whether the models are right for the right reasons (Kirchner, 2006).

Hydrologic model consistency, which refers to the representation of dominant processes by the model, can be evaluated by using hydrological process signatures. This benchmarking strategy reveals a model's ability to reproduce observed process-informative signatures such as flow duration curve, runoff coefficient, and recession curves. For instance, Yilmaz et al. (2008) used flow duration curves to diagnose model performance in capturing the different segments of a hydrograph, while De Boer-Euser et al. (2017) showed the use of flow duration curves in diagnosing model inadequacy. Similarly, recession curves are employed to evaluate and derive models that characterize subsurface processes (Clark et al., 2009; Kirchner, 2009). Meanwhile, numerous studies used a mixture of different signature measures (e.g., McMillan et al., 2011; Tian et al., 2012; Moges et al., 2016). These studies have shown that hydrological signatures can highlight how well the model is capturing the causal processes rather than being a mere predictive tool that may suffer in out-of-sample tests.

Model functional performances can be evaluated using information-theoretic metrics that quantify information flows between flux and state variables. These metrics are used as 1) a better measure of dependence between simulations and observations than linear metrics such as the Pearson correlation coefficient and similar L-norm based metrics (Pechlivanidis et al., 2010, 2014; Weijs et al., 2010), 2) tools that reveal model internal interactions among all variables (termed "process networks") (Ruddell and Kumar, 2009; Bennett et al., 2019; Moges et al., 2022), and 3) quantitative measures of the synergies or tradeoffs between predictive and functional performance in a model. L-norm based metrics quantify the actual differences between observed and simulated values as opposed to information flow metrics that quantify differences in probabilistic distributions. Here, synergies refer to simultaneous improvements in both predictive and functional performance, while tradeoffs refer to gains in either functional or predictive performance leading to a loss in the other (i.e, between "right answers" versus "right reasons") (Kirchner, 2006; Ruddell et al., 2019). The use of functional model performance metrics, particularly a model's process network, helps to evaluate the validity of the model's constitutive functional hypotheses in light of both expert judgment and model intercomparisons. However, as some of these tools were developed only recently, there is a lack of widespread application and ready-to-use interfaces accessible to the wider community.

Reproducibility is central to science and one of the key features of the geosciences paper of the future (Gil et al., 2016). It involves the full documentation, description, and sharing of research data, software, and workflows that underpin published results. However, multiple disciplines including hydrology have indicated that there is a reproducibility crisis (Stagge et al., 2019). Thus, similar to the call for model diagnostics and benchmarking, there is a drive towards hydrological research reproducibility. Hutton et al. (2016) indicated that the lack of common standards that facilitate code readability and reuse, well-documenting workflows, open availability of codes with metadata, and

citation of codes are key challenges in hydrological computational reproducibility. As a potential solution, recent tools in computer science are enabling ease of documenting, collaborating, self-descriptiveness, and sharing of codes and workflows. These tools can likewise be used to support reproducibility in computational hydrology. Furthermore, as these tools are user-friendly and interactive, they can be used to support not only modelers but also decision-makers who are not as equally code-adept and trained as modelers.

One way to meet the call for an organized (less fragmented) system of comprehensive model evaluation and reproducibility is to have a readily available tool. For instance, the Toolbox for Streamflow Signatures in Hydrology (TOSSH) was recently developed as a Matlab ® toolbox that provides a variety of hydrological process signatures (Gnann et al., 2021). Similarly, Hydroeval focuses on statistical predictor metrics (Hallouin, 2021). Although these tools are available, they are limited in their focus to one set of diagnostics and lack interactivity. For instance, Hydroeval is focused on multiple predictive performance measures such as the Nash-Sutcliffe coefficient while TOSSH provides an extended list of hydrological signature measures to evaluate process consistency. Furthermore, they do not incorporate the recent information-theoretic toolsets that quantify model functional performances. On the other hand, although various Jupyter based tools that support reproducibility are being developed in hydrology (for example, Peñuela et al. (2021) on reservoir management), they cannot typically produce benchmarking and diagnostic metrics.

Building on the existing model benchmarking and diagnostic tools, HydroBench (https://emscience.github.io/HydroBenchJBook/HydroBenchIntroduction.html) serves as an open-source, model agnostic hydrological diagnostics platform that emphasizes reproducibility. As a comprehensive model performance evaluation tool, HydroBench consists of three sets of metrics that include 1) predictive performance metrics, 2) hydrological signatures, and 3) functional performance metrics that use information-theoretic concepts. The tool can be used to help modelers diagnose potential issues with their models, users to reproduce model performance evaluations, decision-makers to quickly evaluate and understand model performances interactively, and educators to teach hydrological science students about both model diagnostics and reproducibility. In order to demonstrate its usefulness and application, HydroBench is applied to the NHM-PRMS product at the watershed scale near Cedar River, WA.

Methods

HydroBench helps answer the following model performance evaluation questions in a reproducible manner:

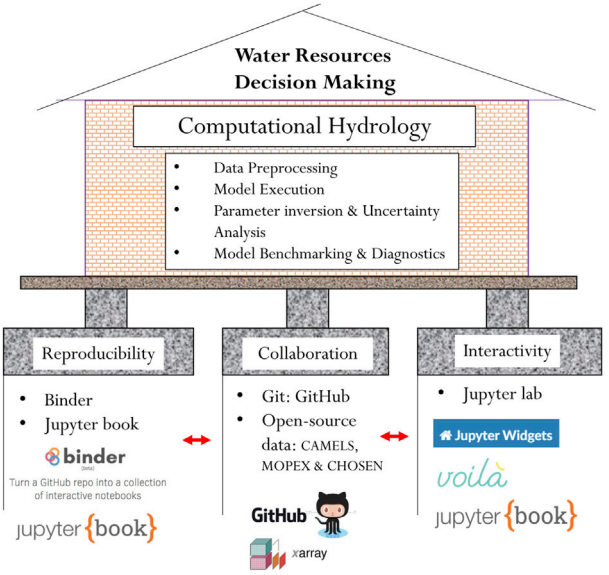

FIGURE 1
HydroBench supports computational hydrology and water resources decision making by facilitating reproducibility, collaboration, and computational interactivity. Voila is a widget rendering tool that enables interactive computation (https://github.com/voila-dashboards/voila), while Xarray is a standardized data format that eases working on multidimensional datasets (https://docs.xarray.dev/en/stable/).

1) How good a predictor is the model with respect to statistical predictive performance measures?
2) How consistent is the model with a suite of observed hydrological behaviors (i.e., signatures)?
3) How well do the model's internal dynamics replicate interactions among observed system variables?

These three questions are addressed within HydroBench through three types of hydrological benchmarking metrics that aid in model performance diagnostics. In this section, we first highlight the software ecosystem that underlies HydroBench and supports reproducible research and then discuss the three sets of benchmarking metrics.

Reproducibility and the jupyter ecosystem

Model diagnosis and benchmarking require evaluation strategies that are applicable to any watershed or model (i.e., "model-agnostic"). Standardizing model benchmarking and diagnostics in a reproducible and collaborative manner will allow modelers to better focus their time on research development, rather than on reinventing the model evaluation wheel. In this regard, the Jupyter ecosystem (https://jupyter.org/)

provides foundational tools that are intended to facilitate reproducibility and collaboration.

In HydroBench, we followed a three-pillar scheme to support hydrological model benchmarking and diagnostics: 1) reproducibility, 2) collaboration and 3) interactive computation (Figure 1). To support reproducibility, we used Jupyter Notebook, Binder and JupyterBook (Project Jupyter 2022 | https://jupyter.org/). Jupyter Notebooks are open-source documents that merge code, results, texts and interactive widgets to narrate a computational story (Pérez and Granger, 2007). By narrating a computational story rather than presenting mere codes or results, notebooks make computational workflows self-descriptive. Furthermore, as notebooks can be viewed and shared easily, they also facilitate collaboration and reproducibility. For a detailed description of Jupyter Notebooks, the ten best practices of using Jupyter Notebooks are outlined in Rule et al. (2019) while ten best practices of reproducible research are outlined by Sandve et al. (2013).

Hydrological computations may require the use of more than one Jupyter Notebook or a very long single notebook. Having long or multiple notebooks leads to story fragmentation. To avoid this fragmentation, a Jupyter Book can be used to bind together multiple notebooks (Community, 2020 - https://jupyterbook.org/intro.html). A Jupyter Book is a compilation of notebooks and markdown (text) readme-files. This compilation can then be published as a traditional book narrating the computational story from its multiple components.

One way to facilitate scientific reproducibility is by openly sharing a complete, re-runnable workflow over the cloud. Binder is a web-based cloud platform that enables sharing and executing codes by recreating the computational environment without installing packages locally (Jupyter et al., 2018). Since the computational environment is recreated on the cloud, Binder makes reproducing codes and their results a single-click task. Thus, Binder not only provides a reproducible environment but also simplifies the user experience.

Collaboration is key in both model development and diagnostics. Git is a version control state-of-the art tool for code development and collaboration, while GitHub and other similar platforms are online repositories that enable sharing and collaboration on codes. Through its version-control features, Git enables a reproducible workflow among groups of collaborators on a project. In addition to collaboration on code developments, open-source hydrological data are also critical for community-wide model benchmarking, as they enable modelers to test their hypotheses beyond local watersheds and over a broad range of time against consistent information. Examples of large-sample open-source data in hydrology include the MOPEX, CAMELS, EMDNA, and CHOSEN datasets (Duan et al., 2006; Addor et al., 2017; Tang et al., 2021; Zhang et al., 2021).

The third pillar of HydroBench is interactive computation. Although sharing codes, executables and data is critical in reproducibility, codes are not always user-friendly, as their use is impossible without baseline expertise. In contrast, widgets are user-friendly tools that can be intuitively executed with clicks and slider bars. As a result, they can support most users and stakeholders across the spectrum of computing skills. In addition, widgets clear up code blocks and can facilitate interpretation through informative visualizations.

Model benchmarking and diagnostics

Statistical predictive metrics

Numerous model predictive performance metrics are used in hydrological model evaluation to compare hydrological responses such as observed and modeled streamflow (and/or water table, or evapotranspiration) data. Each metric has a different skill in its evaluation. For instance, the Pearson correlation coefficient is effective in revealing the linear relationship between observed and modeled output, while the log-transformed Nash-Sutcliffe coefficient is more sensitive to low flow regimes than high flows. A detailed skills description of these metrics can be found in Krause et al. (2005), Gupta et al. (2009), and Moriasi et al. (2015). Due to their variation in skill, it is recommended to evaluate models using multiple metrics (Bennett et al., 2013). As a result, HydroBench includes multiple statistical metrics as indicators of models' predictive performances. Table 1 provides the list of HydroBench's model predictive performance metrics and their corresponding skills. These metrics are selected according to their skill, widespread use in hydrology, complementarity, and avoidance of redundancy. In terms of skill, they cover high and low flows, volume, and overall hydrograph characteristics (Table 1 and Figure 2).

Process-based hydrological signature metrics

Statistical predictive performance metrics lack hydrological rigor and are not sufficient in diagnosing model performances (Gupta et al., 2008; McMillan, 2021). In contrast, the use of hydrological signature metrics can help diagnose model performances by indicating the model's ability to reproduce specific hydrological processes such as high/low flows or subsurface flows. Multiple process-based signature metrics are implemented in HydroBench (Table 2). Table 2 provides a description and relative skills of the signature metrics, which are complementary to each other in characterizing subsurface flow, different segments of a hydrograph and water balance. In addition, we have also created an interface between TOSSH and HydroBench to support the full access of the TOSSH hydrological signature metrics to HydroBench users. A detailed guide of the interface is provided in the example notebook included in HydroBench. For an extended list, skill, and computation of hydrological signatures, we refer users to the TOSSH toolbox and the references therein (Gnann et al., 2021).

TABLE 1 List and description of predictive performance evaluation metrics in HydroBench. Here, Q represents streamflow, an example of the dependent variable, P represents precipitation, as an example of an input flux variable, mod = model, and obs = observed.

Name	Equation	Description and skill
Nash-Sutcliffe efficiency (NSE)	$NSE = 1 - \sum_{i=1}^{n}(Q_{obs,i} - Q_{mod,i})^2 / \sum_{i=1}^{n}(Q_{obs,i} - \underline{Q}_{obs})^2$	NSE is relatively skilled in revealing model performance in capturing high flows, while it has limited skill in capturing low flows, as it is an L^2 norm-derived metric
Log transformed (logNSE)	Similar to NSE but with Q_{obs} and Q_{mod} in the logarithm space	logNSE is similar to the Nash Sutcliffe efficiency but with the inputs being transformed to the logarithm space. As it is computed based on log-transformed inputs, it is skilled in capturing model predictive performances of low flows
Percent Bias (PBIAS)	$PBIAS = \sum_{i=}^{n}(Q_{obs,i} - Q_{mod,i}) / \sum_{i=1}^{n} Q_{obs,i}$	Compared to the L^2 norm-derived NSE, PBIAS is an L^1-derived metric that is less sensitive to peaks and suitable to reveal predictive performances of total streamflow volume Moriasi et al. (2015)
Pearson correlation coefficient (r)	$r = \sum_{i=1}^{n}(Q_{obs,i} - \underline{Q}_{obs})(Q_{mod,i} - \underline{Q}_{mod}) / \sqrt{\sum_{i=1}^{n}(Q_{obs,i} - \underline{Q}_{obs})^2}\sqrt{\sum_{i=1}^{n}(Q_{mod,i} - \underline{Q}_{mod})^2}$	r is a linear measure of model performance. It quantifies the linear relationship between observed and model prediction
Kling-Gupta efficiency (KGE)	$\alpha = stdev(Q_{mod})/stdev(Q_{obs})$ $\beta = mean(Q_{mod})/mean(Q_{obs})$ $KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$	KGE addresses NSE's biases and better evaluates model performance in capturing both high and low flows (Gupta et al., 2009)

A

Streamflow Observed	Streamflow Model	Precipitation	Air Temperature	Soil Moisture	Snowmelt	Actual ET	Potential ET

B

Statistical performance metrics	Hydrological Signature metrics	Information-theoretic Functional performance metrics
Nash Sutcliffe coefficeint	Runoff coefficient	Tradeoffs between functional and predictive performance
Kling-Gupta coefficeint	Flow duration curve (FDC)	Mutual information
Log-transformed Nash Sutcliffe coefficeint	Recession curve	Entropy and conditional entropies
Percent Bias	Time linked FDC	Informaton flow Process Networks
Correlation coefficient		

Key: Sensitivity of metrics	
High flow	Water balance
Volume	Mixture
Low flow	Linearity of relationships
Functional performance	

FIGURE 2
(A) Example of a standard input table to HydroBench. The empty cells refer to user provided input data, and **(B)** Summary of the output metrics of HydroBench and their sensitivities (color-coded). Color codes, described in the lower table ("Key: sensitivity of metrics") indicate the hydrological feature to which the metric is most sensitive.

HydroBench includes Hydrograph and Flow Duration Curve (FDC) as part of the signature metrics. However, a hydrograph becomes cumbersome and difficult to interpret when the time-series being evaluated is long (i.e., multiple years of fine resolution data). Similarly, as FDC is purely probabilistic, it delinks the temporal dimension of the streamflow magnitude. That is, as long as the model preserves the exceedance probability of the observed data, FDC suggests high model performance,

TABLE 2 List and description of hydrological signature-based model diagnostic metrics in HydroBench. Here, Q denotes streamflow, an example of the dependent variable, P denotes precipitation, an example of an input flux variable, and r denotes rank based on a decreasing sorting of a time series.

Name	Equation/ Function	Description and skill
Runoff coefficient (RC)	$RC = \Sigma Q/\Sigma P$	RC deals with the flow of mass from precipitation to streamflow and helps in diagnosing water balance discrepancies between the observed and model time series at the annual scale. Namely, it measures to what extent the model captures the observed annual water balance
Flow duration curve (FDC)	$Q_r = f(Q_{rank})$ $Q_{rank} = r/n + 1$	FDC provides visual diagnostics of model performance in capturing both high- and low-flow segments of a hydrograph in a temporally delinked manner
Recession curve	$dQ/dt = f(Q)$	Recession curves help evaluate model performance in the absence of precipitation. Their shape is most sensitive to the rate at which water is released from catchment storage. Consequently, recession curves can indicate a model's performance in characterizing subsurface processes
Time Linked Flow Duration curve (T-FDC)	$f(Q, bin\ size)$	Because FDC does not have a time component in revealing under- and overestimation of flows, we developed T-FDC, which complements FDC by incorporating a time component. For a given day observed streamflow, T-FDC tracks whether a model estimate results in the same, higher or lower bin. This is analogous to the confusion matrix and requires binning of the data according to the observed minimum and maximum values. T-FDC is a (visual) metric between FDC and hydrograph. Thus T-FDC eases the interpretation of a hydrograph by simplifying it to be within a specific bin count

regardless of the time coincidence of the model simulation. Complementing the hydrograph and FDC, we developed a signature metric that is probabilistic like FDC but also preserves the time correspondence of the simulation like a hydrograph. The metric is called Time linked Flow Duration Curve (T-FDC), and it inherits the characteristics of both FDC and a hydrograph.

T-FDC is a heatmap-based model performance evaluation hydrological signature metric. In constructing the heatmap, T-FDC first lets users define a bin size for segmenting streamflow. Second, it bins the observed streamflow to the predefined bin size and sets it as the *y*-axis. Then, for its *x*-axis, T-FDC tracks whether the time corresponding model-simulated streamflow is binned in the same bin class as the observed streamflow or other bin classes. Finally, it generates a heatmap based on the time-tracked counts of the simulated streamflow in each bin class. A perfect model with a high number of data counts in the same bin as the observed values will only populate the main diagonal of the heatmap. In contrast, a high number of data counts below the diagonal indicate an underestimating model, while an overestimating model will have a high number of counts above the diagonal. This makes T-FDC's visual interpretation intuitive. In addition to the visual interpretation, we have included a numerical quantification of model performance based on T-FDC using the percentage of data counts in the diagonal. Higher percentages indicate higher performance and vice versa.

Information-theoretic metrics

Beyond the predictive metrics and signature measures, recent developments in hydrological model diagnostics involve the use of information-theoretic metrics (Nearing et al., 2018, 2020). Compared to the predictive and hydrological signature metrics, the information-theoretic metrics require longer hydrological records. However, the diagnostic information they provide about why a model may be exhibiting poor performance, or whether it exhibits good performance for the right reasons, can be more powerful. Specifically, HydroBench provides a suite of information theoretic-based metrics (Table 3) that reveal 1) functional model performance, 2) predictive model performance and 3) the tradeoff between functional and predictive performances. Functional performance can be quantified by comparing observed transfer entropy (TE) with modeled TE and visualized using information flow process network (PN) illustrating functional relationships within the model (Ruddell et al., 2019). TE is a measure of time-lagged information flow from a "source" to a "sink" variable that accounts for autocorrelation in the "sink" time series. Unlike the runoff coefficient, which quantifies the flow of mass from precipitation (P) to streamflow (Q), PNs quantify information flow (i.e., uncertainty reduction of Q by P) between these and other variables. On the other hand, the predictive performance of a model can be quantified as the mutual information (MI) between the observed and modeled time series, which functions similarly to a correlation coefficient but is robust to nonlinearity (Ruddell et al., 2019). By providing visualizations of these metrics and how they vary across alternative models, HydroBench helps reveal the tradeoffs between predictive and functional performances.

In HydroBench, predictive performance is quantified based on the similarity between the observed and predicted streamflow time series, computed through their mutual (i.e., shared)

information (1-MI). Functional performance is evaluated as a comparison of information flows from a forcing variable (e.g., temperature, precipitation) to a sink variable (e.g., streamflow) in the model versus observations ($TE_{source \rightarrow sink:\ model} - TE_{source \rightarrow sink:\ observed}$). Ideally, information will flow similarly among modeled variables as in observations leading to a zero score in functional performance. Negative values of functional performance indicate that the model does not extract enough information from the forcing variable, with extreme negative values being indicative of an overly-random fit. Positive values of functional performance indicate that the model extracts too much information from the forcing variable of interest, resulting in an overly-deterministic fit. Further details of this interpretation can be found in Ruddell et al. (2019).

HydroBench additionally provides one, two and three-dimensional entropy measures for the given random variables X, Y, and Z as H(X), H(X, Y) and H(X, Y, Z), that quantify the information content of a single variable or its simultaneous interactions with multiple other variables. However, higher-dimensional quantities require longer data record lengths than the metrics discussed above. Along with their data length requirements, information theoretic metrics have a few shortcomings or caveats in comparison to the other metrics. As information theoretic metrics are dependent on probability distributions rather than on actual variable values, it is important to use them along with hydrological signatures and statistical performance measures that are a function of the actual values of the variables. Moreover, the computation of these information-theoretic metrics involves subjective parameters such as the number of bins and the statistical significance threshold. The Jupyter notebook accompanying HydroBench describes these parameters and their computation, including the number of bins and statistical significance.

HydroBench interface—Input and output data structure

HydroBench is a model-agnostic platform that requires basic Python programming skills. It can be downloaded/cloned from the following GitHub link https://github.com/EMscience/HydroBench with multiple application test cases and a particular focus on the Cedar River, WA. HydroBench accepts model and observed data in a predefined structure. The input structure is a table of data that consists of at least two data columns (e.g., observed streamflow, and model streamflow), along with their start and end dates (Figure 2). The model that generated the data can be lumped or distributed, as HydroBench requires inputs of time series variables. With these inputs, basic benchmarking results can be obtained. The basic results are the predictive performance metrics, plus FDC and T-FDC diagnostics. With an extended input table that contains one or more additional columns of independent variables (e.g., precipitation), HydroBench can provide all three types of metrics - predictive, hydrological signature, and functional (Figure 2). Since HydroBench has a modular design, it can easily be called into any notebooks that host model results and generate a table of inputs (e.g., Figure 2A). Additionally, any single metric can be employed depending on users' preferences.

Case study description

HydroBench was applied to a 103.5-km^2, relatively low-gradient watershed near Cedar River, WA (Figure 3), which was extracted from the NHM infrastructure (Regan et al., 2018) for this case study. The Cedar River watershed was selected for the case study because it is considered undisturbed according to the GAGES II classification (Falcone et al., 2010) and because NHM-PRMS predictions of its streamflow strongly contrast between the calibrated and uncalibrated version of the model (Section 3). The catchment's land cover is dominated by a coniferous forest (Falcone et al., 2010). Comparing the long-term (1980–2016) average monthly precipitation and catchment area-normalized streamflow volume, streamflow is higher than precipitation from April to July, indicating that most of the streamflow is a function of storage during these months, while the remaining months are dominated by precipitation, meaning that water enters storage. The catchment resides in a humid climate, where 53% of precipitation falls as snow (Figure 3 and Falcone et al., 2010).

The model under consideration is NHM-PRMS. NHM-PRMS provides two hydrological model products based on two model parameter sets: a nationally calibrated set and the uncalibrated set (Driscoll et al., 2018; Hay, 2019). In the NHM-PRMS uncalibrated model (Driscoll et al., 2018), parameters are estimated from both catchment and climatic characteristics (Markstrom et al., 2015; Regan et al., 2018; Regan et al., 2018). In cases where estimation is impossible, the uncalibrated product is based on model default parameter values from Markstrom et al. (2015). This approach has its advantages and limitations. Primarily, it is fast compared to automatic calibration schemes and can be used to initialize the PRMS model for a further automatic calibration. Additionally, the approach might also be beneficial for parameter estimation in ungauged watersheds and nonstationary systems, as it does not rely on historical climatic/meteorological data. However, the approach becomes poor in cases where local data is sparse and in regions where the model is not tested before, as the default values may not be relevant. An extended description of the uncalibrated NHM-PRMS model parameter estimation and its product can be found at Regan et al. (2018) and Driscoll et al. (2018).

The calibrated version of NHM-PRMS employed a multivariable stepwise parameter estimation using the Shuffle Complex Evolution algorithm (Hay and Umemoto, 2007; Hay

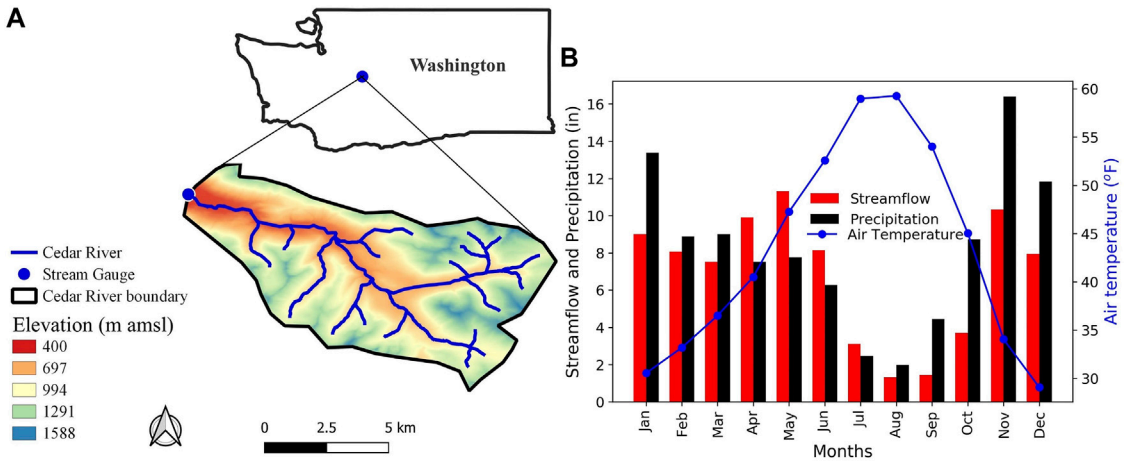

FIGURE 3
Case study watershed near Cedar River, WA: **(A)** location and elevation, and **(B)** annual hydrometeorological characteristics. The WA basemap source data is from https://geo.wa.gov/.

et al., 2006 & 2019). In starting the calibration, the parameters were initialized at their uncalibrated NHM-PRMS value. The calibration uses multiple variables, including daily streamflow from 1980 to 2010 and for the same period, monthly snow cover area (SCA, from SNODAS; National Operational Hydrologic Remote Sensing Center, 2004), potential evapotranspiration and solar radiation (PET and SR, from Farnsworth and Thompson, 1982, the DAYMET climate data and Regan et al., 2018), actual evapotranspiration (AET, from Cao et al., 2006; Rietjes et al., 2013) and soil moisture estimates (SM, from Campo et al., 2006; Thorstensen et al., 2016). These data are derived from national scale remotely sensed datasets and other model products. The sensitivity of the different model parameters to these variables is assessed, and parameters are then sequentially calibrated with an objective function defined as the normalized root mean square error between the observed and simulated values of the output variables in decreasing order of sensitivity (Markstrom et al., 2016). That is, in calibrating PRMS to these variables, sensitivity analysis guides the identification of which parameters are calibrated by which variable in a stepwise manner. Stepwise calibration starts with 1) PET and SR, followed by 2) SM and AET, and finally, 3) streamflow. For a detailed description of the model calibration and the optimization employed, please refer to Hay et al. (2006), Hay and Umemoto (2007) and LaFontaine et al. (2019).

In demonstrating the application of HydroBench at the Cedar River, we evaluated model performance with respect to the input, state, and output variables of the calibrated and uncalibrated NHM-PRMS model. Namely, as NHM-PRMS computes hydrologic fluxes using inputs of daily precipitation and maximum and minimum air temperature, these variables were included in our analysis. Similarly, we extracted the predicted variables of streamflow, snowmelt, basin soil moisture, and actual evapotranspiration from 1980 to 2016 at a daily time step for our model benchmarking and diagnostics at the Cedar River, WA.

Results

Facilitating reproducibility, all inputs and the results presented in this section are available on GitHub (https://github.com/EMscience/HydroBench). As a Binder link is also included, the analysis can be fully reproduced, and the different widgets can also be used for further interactive computation on the cloud. Thus, users of HydroBench can emulate and adapt the workflow easily.

Statistical predictive performance metrics

At the Cedar River watershed, the uncalibrated model shows better statistical predictive performance than the calibrated model, according to the HydroBench-provided statistics, except for the KGE metric under the log-transformed flow condition (Table 4 and Figure 4). Regardless of the skills of the metrics in representing the different hydrograph segments (low or high flows), most of the predictive performance metrics suggest that the uncalibrated model is a preferred choice (Tables 1–3). However, the predictive performance metrics do not explain why and how the uncalibrated model exhibits better predictive performance than the calibrated model. In addition, it is important to note that the calibration of NHM-PRMS does not only focus on the prediction of streamflow but also on capturing remotely sensed ET and other variables with a stepwise calibration method.

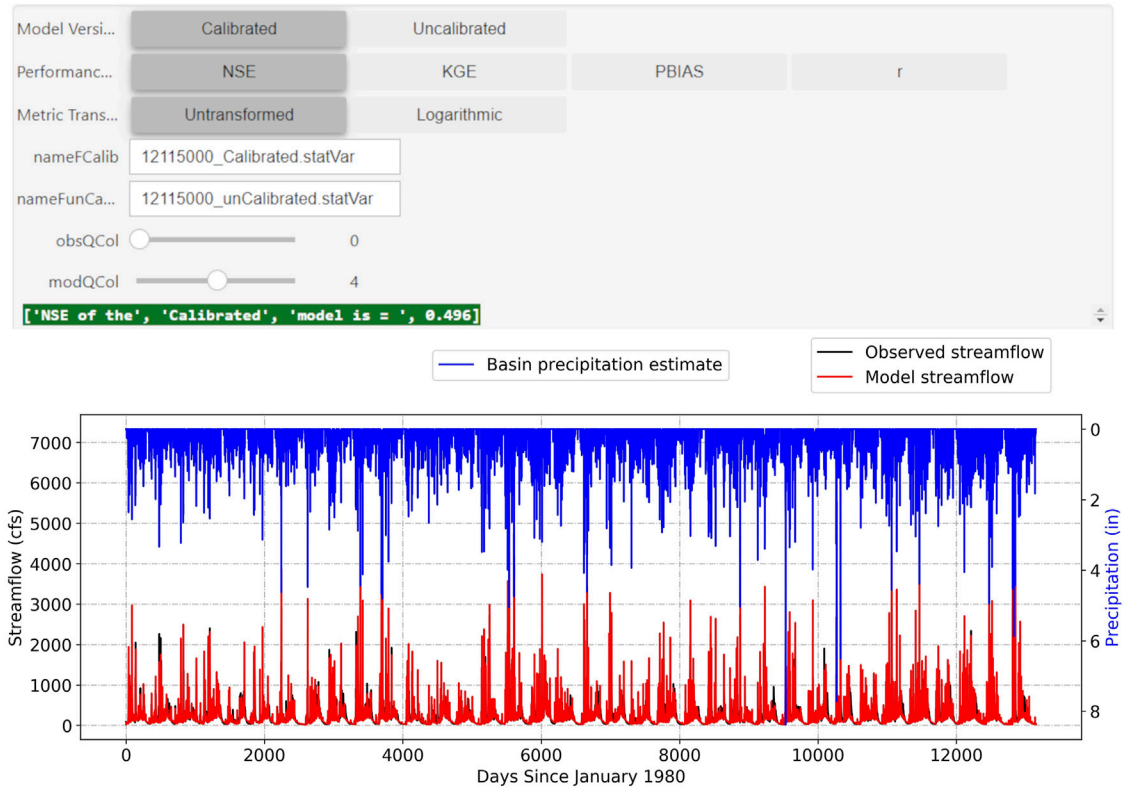

FIGURE 4
Screenshot of HydroBench interactive window showing predictive performance metrics and the corresponding hydrograph for NHM-PRMS simulations of the Cedar River watershed daily streamflow, 1980–2016, compared to observations (USGS stream gage, 12115000).

TABLE 3 List and description of information-theoretic model diagnostic metrics. Here, Q denotes streamflow, an example of the dependent variable of interest, and P denotes precipitation, an example of an input flux variable.

Name	Equation	Description and skill
Entropy (H(x))	$H(Q) = -1 * \sum_{i=1}^{i=1} p(Q_i) * log(p(Q_i))$	Provides a measure of the uncertainty of the indicated flux or store variable(s) Shannon (1948)
Mutual Information (MI)	$MI(P,Q) = \sum_{P,Q}^{n} p(P,Q) log(\frac{p(P,Q)}{p(P)p(Q)})$	MI quantifies the predictive performance of a model. It measures the shared information content of the observed and modeled dependent variable
Transfer Entropy (TE)	$TE(P \rightarrow Q) = MI(Q_t, P_t \mid Q_{t-1})$	TE quantifies the shared information between two variables (typically thought of as an independent and dependent variable) conditioned on the history of the dependent variable Schreiber (2000). In HydroBench, the variables can be any flux or store variables as chosen by expert's (user's) choice
The trade-off between functional and predictive performances	$f(MI, TE)$	The tradeoffs between functional and predictive performance metrics across models are visualized through a bivariate plot showing MI and TE Ruddell et al. (2019); see also Figure 7C here for an example)
Process networks (PN)	$PN = f(TE)$	PNs provide a visual web of the model internal information flow between different flux and store variables as computed by TE

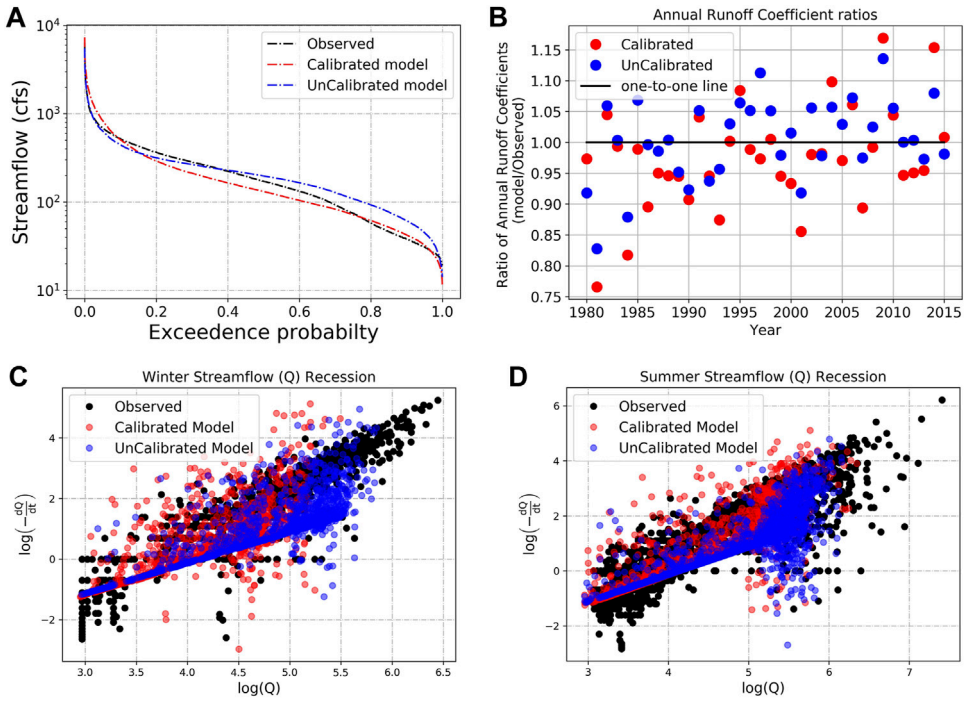

FIGURE 5
Hydrological signature-based evaluation of NHM-PRMS predictions of daily streamflow at Cedar River, WA over 1980–2016: **(A)** flow duration curve, **(B)** annual (i.e., October to September water year) runoff coefficient, **(C)** winter/cold season (months October to March) recession curves and, **(D)** summer/warm season (months April to September) recession curves. The seasons and the corresponding months can be adaptively defined in HydroBench.

Hydrological process consistency using hydrological signature metrics

Both the FDC and T-FDC indicate that high flows are better represented by the uncalibrated model (Figures 5A, 6; Table 5). In contrast, the recession curves indicate that the subsurface release of water from storage over extended periods is better represented by the calibrated model (Figures 5C,D; Table 5), as it has a scatter (slope and intercept) more similar to the observations than does the partly near-linear (in a semi-log space) uncalibrated model. Figure 6 along with Table 5 shows that the calibrated model is closely related to the observed data (35% than 33%). On the other hand, the runoff coefficient (RC) comparison between both model versions indicates strong similarity between the models, with an RC model to RC observed ratio of 0.969 for the calibrated and 1.003 for the uncalibrated model (Figure 5B). The similarity in RC may suggest that the annual mass flow (precipitation to streamflow) of the two models is similar, with slightly more precipitation converted into streamflow in the uncalibrated model.

Despite the high statistical predictive performance reports of the uncalibrated model (Table 4), the hydrological signature metrics revealed that the calibrated model better represents the low-flow segments of the Cedar River hydrograph. This comparison of predictive and hydrological signature metrics underscores the need for both types of performance evaluations. Although hydrological process signature metrics illuminate the failure or success of each model in representing different processes, neither they nor the statistical predictive metrics can reveal what type of model input and output interactions lead to the model results, underscoring the need for functional performance evaluations.

Model functional performances using information-theoretic metrics

The calibrated and uncalibrated models have a similar pattern of information flows, depicted in their process networks (PN), with a few exceptions (Figures 7A,B; Table 6). For example, the PNs depict high transfer entropy (TE) from precipitation to snowmelt in the uncalibrated model. In contrast, the calibrated model has high TE from precipitation directly to streamflow. Although observations of daily snowmelt are not available for this watershed for

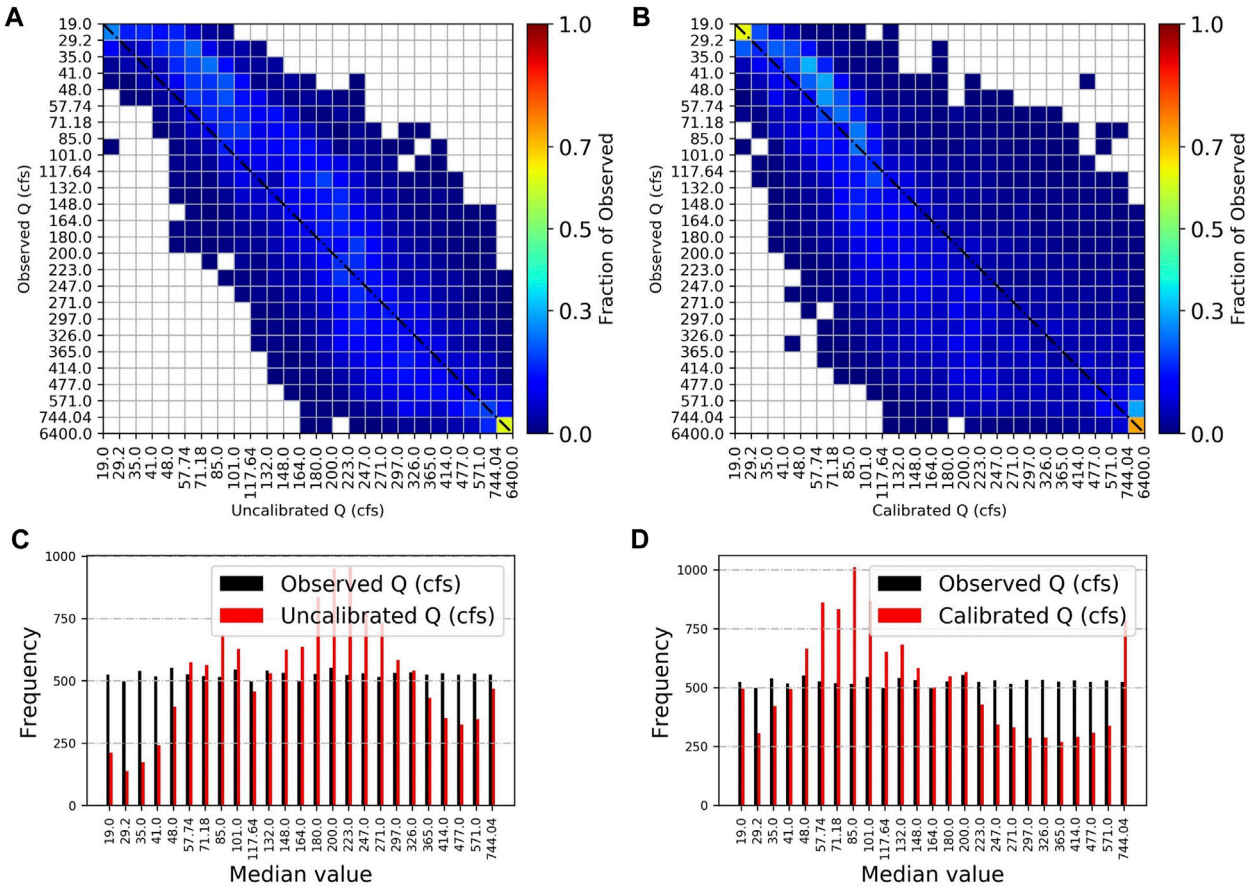

FIGURE 6
Time-linked flow duration curve for **(A)** the uncalibrated model and **(B)** the calibrated model **(C)** the sum of the number of simulated flows in the same flow range bin as the observed for the uncalibrated model and **(D)** the same as C but for the calibrated model. Figures **(A,B)** show how the observed flows in each bin are distributed across the bins of the model estimated flows. The number of bins, a user-defined value, is 25 here. Ideally, hot colors would populate the diagonal, implying minimum over/underestimations.

TABLE 4 Summary of statistical predictive performance metrics for the uncalibrated and calibrated NHM-PRMS model of a watershed near Cedar River, WA, based on daily streamflow, 1980–2016.

	NSE		KGE		PBIAS		r	
Model Versions	**Calibrated**	**Uncalibrated**	**Calibrated**	**Uncalibrated**	**Calibrated**	**Uncalibrated**	**Calibrated**	**Uncalibrated**
Untransformed flow	0.50	0.76	0.66	0.85	2.6%	−0.35%	0.84	0.88
Log transformed flow	0.69	0.78	0.85	0.79	N/A	N/A	0.85	0.90

comparison to an observed PN, the PN difference noted by the models suggests that snowmelt contributions in the uncalibrated model could be the cause of low flow overestimation in the FDC. Following these insights from the PN plots, we explored the day of the year (DoY) averages, minimums and maximums of snowmelt, actual evapotranspiration and soil moisture of the two models (Figure 8). The figure showed that the uncalibrated model leads to snowmelt processes even in the late summer months, which is not likely.

The visualization of tradeoffs between predictive and functional performance metrics (Figure 7C) shows that

TABLE 5 Numerical scores of hydrological signature metrics. For this test case, we chose the mid slope of the FDC (25–45% exceedance probability). Similarly, we chose the main diagonal in T-FDC as a strict measure and 'Dry' months (April—September) for recession score as a representative of subsurface flow dominant season. HydroBench allows users to choose the exceedance probabilities, the number of diagonals in T-FDC and seasons for recession curve scores.

	FDC slope at exceedance probability of 0.25–0.45	T-FDC main diagonal	Recession coefficients		Annual runoff coefficient ratio (model/observed)
			Slope	Intercept	
Observed	14.27	N/A	1.384	−5.087	N/A
Calibrated	14.86	35%	1.396	−5.561	0.969
Uncalibrated	7.63	33%	1.179	−4.816	1.003

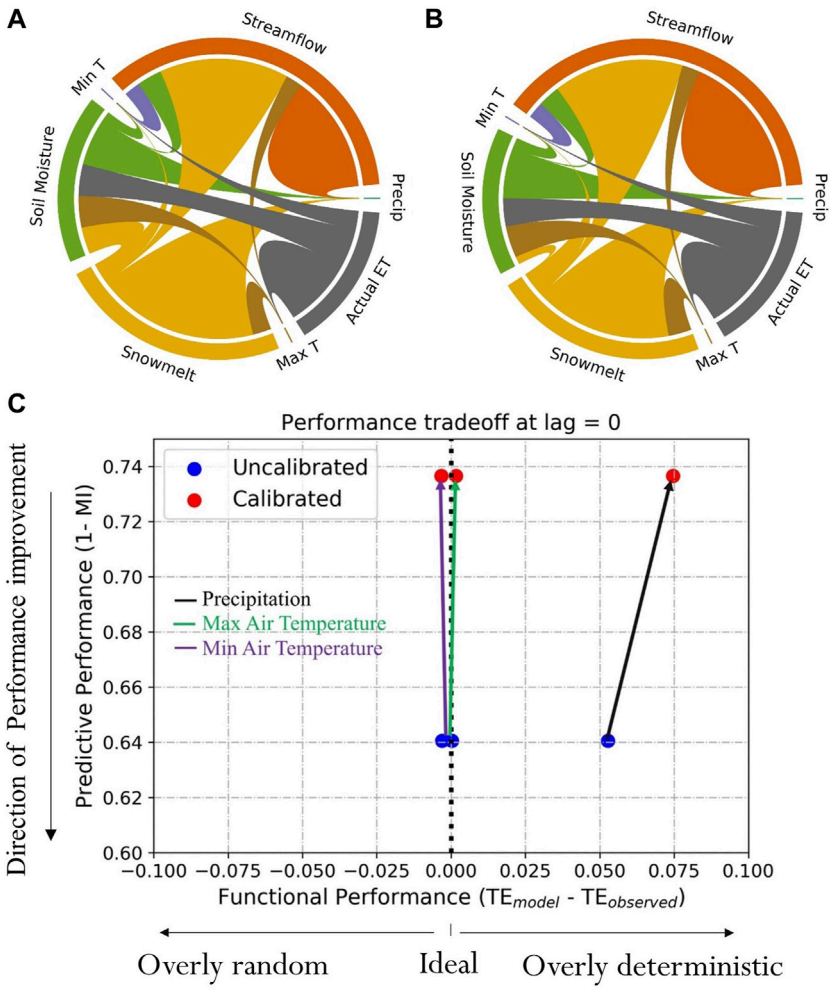

FIGURE 7
Functional performance metrics based on evaluation of NHM-PRMS at the Cedar River watershed. **(A)** uncalibrated model and **(B)** calibrated model, and **(C)** tradeoff between functional and predictive performance metrics. In interpreting PN plots, the outer colored circle indicates the interacting variables. The width of the chords linking the interacting variables corresponds to the TE magnitudes. In **(C)**, the change in predictive performance and functional performance from the uncalibrated model (origin of the arrow, blue) to the calibrated model (point of the arrow, red) is plotted. Thus, the arrows show the effect of calibration. The difference between the two figures is presented in Table 6.

TABLE 6 TE difference between Calibrated and Uncalibrated model (%) ($(TE_{cal} - TE_{uncal})^{*}100$

	Sink				
Source	**Streamflow**	**Soil moisture**	**Snowmelt**	**Actual ET**	**Potential ET**
Precipitation	3.216	0.306	−3.310	—	—
Min Air Temperature	−0.011	0.121	0.518	0.112	−0.053
Max Air Temperature	0.132	−0.046	0.537	−0.139	0.155
Soil Moisture	−0.352	—	—	0.185	−0.316
Snow melt	−0.061	−2.400	—	—	—
Actual ET	—	−0.482	—	—	—
Potential ET	—	−0.213	—	−0.124	—

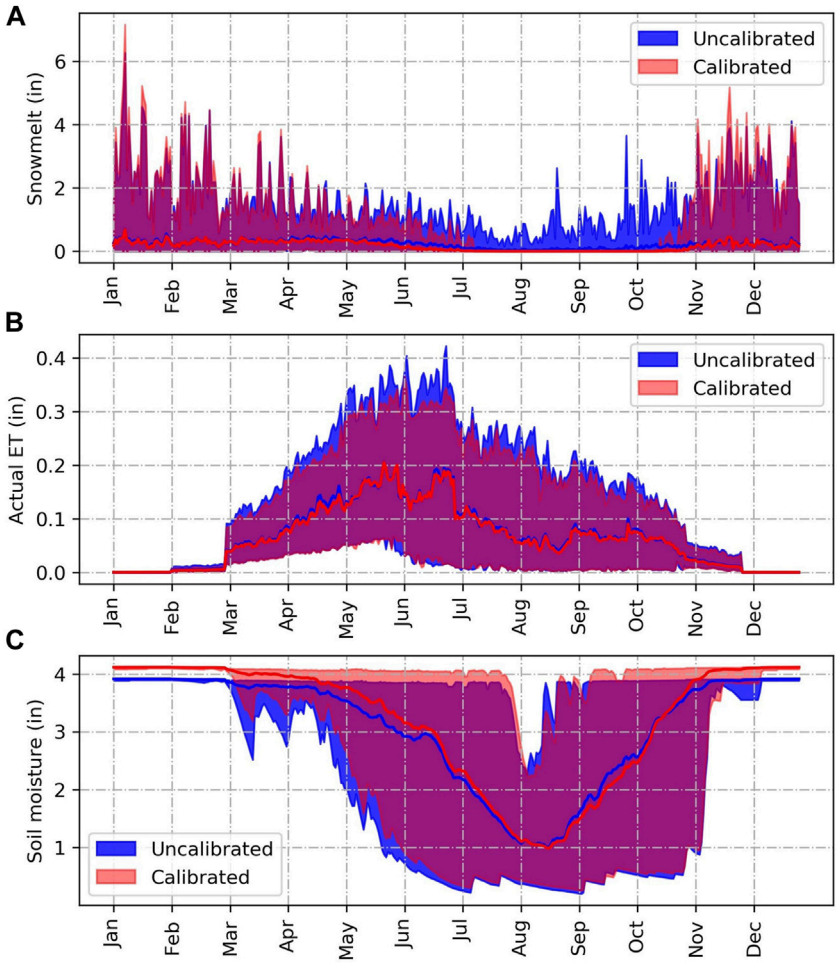

FIGURE 8
Day of the year averages of **(A)** snowmelt, **(B)** actual evapotranspiration and **(C)** soil moisture for both the calibrated and uncalibrated model.

calibration decreased the predictive performance of the model, primarily by over-extracting information from precipitation to inform streamflow (as seen in the higher transfer entropy from precipitation to model streamflow, i.e., $TE_{P \to Q\,model}$ compared to observed streamflow $TE_{P \to Q\,observed}$). However, both the uncalibrated and calibrated

models have $TE_{P\rightarrow Q\,model}$ greater than $TE_{P\rightarrow Q\,observed}$ for precipitation (i.e., overly-deterministic fitting), suggesting that other processes involved in water balance partitioning (e.g., evapotranspiration) or through which precipitation is routed to streams (*via* subsurface or snow storage) may be imperfectly represented in the model structure and/or parameter values. In contrast to the information flows originating from precipitation, information flows from both maximum and minimum air temperatures to streamflow are close to the observed information flows and near the 'ideal fit' point. Given the dominant role of temperature as a driver of evapotranspiration, this similarity of temperature-to-streamflow information flows between models may suggest, by elimination, that the overly-deterministic information flow from precipitation to streamflow observed in the calibrated model is likely attributable to its representation (or lack thereof) of storage processes. Namely, a more direct translation of precipitation to streamflow in the calibrated model may neglect some of the contributions of snow storage to peak flow that are better reflected in the uncalibrated model. However, larger flows of information from snowmelt and soil moisture to streamflow in the uncalibrated model may underlie its poorer performance (relative to the calibrated model) during periods of baseflow and suggest that too much water is extracted from storage over longer time periods.

Discussion

Case-study reflections: Example of how a hydrologist may use HydroBench results

Overall, HydroBench showed that the calibrated and uncalibrated NHM-PRMS model products at the Cedar River watershed have different skills. Although long-term snow and moisture observational data were not available to support the diagnosis of performance discrepancies, HydroBench produced a set of insights into the mechanisms underlying performance differences. In summary, the uncalibrated model exhibited better statistical predictive performance than the calibrated model, particularly during high flows. However, the uncalibrated model was less skilled at capturing low flows and streamflow recession processes, based on the hydrological signature metrics. Functional metrics suggested that routing of precipitation through snow storage and melt differs between the two models, with the calibrated model abstracting too much information directly from precipitation. Thus, it is likely that the uncalibrated model does a better job of capturing peak flows than the calibrated model because it better represents the initial release of water from the snowpack. However, the tradeoff is that the release of water from storage from the uncalibrated simulation is too high during baseflow-dominated periods, in comparison to the calibrated model.

In general, information about whether the relationship between variables is overly random or overly deterministic, as in the Cedar River, can provide useful insight into the next steps. In an overly-random system, although the process information is contained in the observations, it is under-utilized, meaning the model might not have extracted it effectively. Structural changes to the model to represent hydrologic processes more realistically, a better calibration strategy, and/or better objective function may help extract the process information contained in the observations. In contrast, in an overly-deterministic system where there is 'over extraction', it might be better to reduce the dependency of the model on the observed input data. The reduction in dependency might be achieved through diversifying the input data by, for example, incorporating new data (e.g., adding snow and soil moisture data into a model that was forced by precipitation and temperature inputs). Additionally, the user may consider changing the model optimization strategy. Alternative strategies may include calibrating and validating the model in contrasting seasons and hydrograph regimes, using transformed data, and/or changing calibration and validation objective functions in a way that penalizes models in which training data have substantially higher performance than test data. These approaches may lead to less reliance of the model on specific variables or aspects of a variable that have resulted in the overly-deterministic fit.

For the Cedar River case study, the insight provided by HydroBench suggests that further calibration would be a logical next step. Though the calibrated model exhibited poorer predictive performance, its improved ability to capture low flow dynamics may indicate that performance gains can be obtained without changing the model structure. The parameters of focus may be those relevant to snow and soil storage, and the objective function of the calibration may need to be adjusted further to upweight peak flows. Alternatively, the tradeoff in better low-flow performance at the expense of high-flow performance seen in the calibrated model may suggest that rather than an 'absolute best model' parameter set, there exists a Pareto front (i.e., an unavoidable tradeoff). However, this possibility would need to be tested using a multi-objective optimization scheme for calibration that provides the Pareto front. Finally, if further parameter calibration attempts failed to improve the predictive performance of the model while maintaining acceptable functional performance, the modeler may wish to revisit the fundamental structure (i.e., equations) of the model. In this case, the representation of snow storage and melt processes in PRMS might need to be revised to better reflect the Cedar River catchment response.

Alternatively, given the two tested models, a user may decide to opt for the uncalibrated model if most interested in outcomes related to high flows, or the calibrated model if most interested in low flows. Additionally, users or developers may decide to adopt model averaging techniques such as Bayesian Model Averaging -or Hierarchical Mixture of Experts to derive a consensus

prediction (Marshall et al., 2006; Duan et al., 2007; Moges et al., 2016). Importantly, the application of HydroBench to the test case proves that relying only on high performance statistical predictive measures can be misleading as shown by the high predictive performance but the poor functional performance of the uncalibrated model. Thus, a holistic performance evaluation is critical.

The value of a systematic framework for model benchmarking

Model benchmarking and diagnostics are not only at the core of model trust and reliability but also serve as guides for future model development and improvements. HydroBench was designed as a model diagnostic and benchmarking tool in a practice of open, reproducible science. The tool relies on the Jupyter ecosystem for reproducible, collaborative, and interactive computation. HydroBench enables model performance evaluation and diagnosis of performance discrepancies by providing three sets of complementary metrics, including statistical performance metrics, process-based hydrological signatures, and information theoretic-based tools. As demonstrated in the test case, this tool produces insight into many different aspects of a model's performance and helps diagnose performance shortfalls.

The metrics in HydroBench support the different aspects of model evaluation outlined in Gleeson et al. (2021), including a comparison of model results against 1) observations, 2) other models, and/or 3) expert-based expectations. All of the metric categories in HydroBench (predictive, process diagnostics, and functional performances) facilitate comparison against observations in watersheds that have observed data. The information theoretic-based model functional performance metric using PN supports model comparisons even in the absence of observed data, though availability of observed data strengthens such comparisons (e.g., Figures 7A,B). Similarly, PNs and the hydrological signatures can facilitate expert-based model evaluation as they highlight the key hydrological processes and model hypotheses. The graphical representation of a PN can be interpreted as an imprint of the models' process conceptualization. HydroBench can be used to formalize and standardize the ad-hoc expert-based model evaluation approaches commonly applied by the hydrologic science community.

Although all the three categories of metrics in HydroBench are designed to be used in concert, HydroBench is modular and supports the use of any of the metrics individually. For instance, in watersheds with abundant data, all capabilities of HydroBench can be utilized. However, in cases of limited record length or data diversity, a user may decline to use information-theoretic metrics because they are not reliable in limited record lengths.

Choice of calibration objective functions dictates model performance and sensitivity analysis results (Diskin and Simon, 1977; Jie et al., 2016; Markstrom et al., 2016; Garcia et al., 2017). For instance, a model calibrated using root mean square error may not result in better performance in logNSE. Thus, in using HydroBench, we suggest a careful choice of performance metrics that reflect the modeling objective. For instance, for pure predictive purposes, such as short term flow forecasts, relying on predictive performance metrics is beneficial. On the other hand, water balance projections and quantifications can better be served by signature based diagnostics and functional performance evaluation metrics as they seek to get the right answer for the right reasons. Furthermore, in modeling works that start with a sensitivity analysis, the sensitivity analysis result can also be used to align sensitive parameters, modeling objectives and evaluation metrics. That is, evaluating models based on a metric that reflects the objective function set for the sensitivity analysis. Although this approach is consistent with the user's modeling objective, the approach is susceptible to getting the right answer for the wrong reasons. For instance, in a non-stationary system, an insensitive parameter or process can be activated and the prediction and evaluations can be misplaced. In this regard, multi-objective calibration and comprehensive model evaluation across the three categories of HydroBench can be beneficial in diagnosing whether the model is right for the right reasons.

In addition to its utility in hydrologic research and applications, HydroBench can be used to support hydrological teaching that focuses on modeling and model evaluations (Wagener and Mcintyre, 2007; Wagener et al., 2012). Last, HydroBench is an open source project and can be extended by the community and also integrated with other benchmarking tools, as TOSSH is interfaced with HydroBench.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/EMscience/HydroBench, and https://zenodo.org/badge/latestdoi/375593287.

Author contributions

EM, BR, FP, JD, and LL: motivation and framing of the work. EM: first draft manuscript development and writing. BR: MatLab code for the Information theory metrics which was translated to Python by EM. PN and JD: dataset for NHM-PRMS. All authors: discussions, manuscript editing and improvements.

Funding

This work is partially supported by the U.S. Geological Survey Powell Center for Analysis and Synthesis, a Gordon and Betty Moore Foundation Data-Driven Discovery Investigator grant to LL, and the Jupyter Meets the Earth project, funded by NSF grant numbers 1928406 and 1928374 to LL and FP.

Acknowledgments

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Addor, N., and Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resour. Res.* 55, 378–390. doi:10.1029/2018WR022958

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.* 21, 5293–5313. doi:10.5194/hess-21-5293-2017

Bennett, A., Nijssen, B., Ou, G., Clark, M., and Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resour. Res.* 55, 4613–4629. doi:10.1029/2018WR024555

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environ. Model. Softw.* 40, 1–20. doi:10.1016/J.ENVSOFT.2012.09.011

Campo, L., Caparrini, F., and Castelli, F. (2006). Use of multi-platform, multi-temporal remote-sensing data for calibration of a distributed hydrological model: An application in the arno basin, Italy. *Hydrol. Process.* 20, 2693–2712. doi:10.1002/HYP.6061

Caol, W., Sun, G., Chen, J., Noormets, A., and Skaggs, R. W. (2006). Evapotranspiration of a Mid-Rotation Loblolly Pine Plantation and a Recently Harvested Stands on the Coastal Plain of North Carolina, U.S.A. Williams, Thomas, eds. Hydrol. Manag. For. Wetl. Proc. Int. Conf. St. Joseph, MI Am. Soc. Agric. Biol. Eng. 27-33.

Clark, M. P., Rupp, D. E., Woods, R. A., Tromp-van Meerveld, H. J., Peters, N. E., and Freer, J. E. (2009). Consistency between hydrological models and field observations: Linking processes at the hillslope scale to hydrological responses at the watershed scale. *Hydrol. Process.* 23, 311–319. doi:10.1002/hyp.7154

Cohen, S., Praskievicz, S., and Maidment, D. R. (2018). Featured collection introduction: National water model. *J. Am. Water Resour. Assoc.* 54, 767–769. doi:10.1111/1752-1688.12664

Community, E. B. (2020). *Jupyter Book*. Zenodo. doi:10.5281/ZENODO.4539666

De Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., et al. (2017). Looking beyond general metrics for model comparison – lessons from an international model intercomparison study. *Hydrol. Earth Syst. Sci.* 21, 423–440. doi:10.5194/hess-21-423-2017

Diskin, M. H., and Simon, E. (1977). A procedure for the selection of objective functions for hydrologic simulation models. *J. Hydrol. X.* J34, 129–149. doi:10.1016/0022-1694(77)90066-X

Driscoll, J. M., Regan, R. S., Markstrom, S. L., and Hay, L. E. (2018). Application of the national hydrologic model infrastructure with the precipitation-runoff modeling system (NHM-PRMS), uncalibrated version. *U.S. Geol. Surv.* doi:10.5066/P9USHPMJ

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* 30, 1371–1386. doi:10.1016/j.advwatres.2006.11.014

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol. X.* 320, 3–17. doi:10.1016/j.jhydrol.2005.07.031

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R. (2010). Gages: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. *Ecology* 91, 621. doi:10.1890/09-0889.1

Farnsworth, R. K., and Thompson, E. S. (1982). *Mean monthly, seasonal, and annual pan evaporation for the United States*: Washington, D.C., National Oceanic and Atmospheric Administration Technical Report NWS 34, 82 p

Garcia, F., Folton, N., and Oudin, L. (2017). Which objective function to calibrate rainfall–runoff models for low-flow index simulations? *Hydrological Sci. J.* 62 (7), 1149–1166. doi:10.1080/02626667.2017.1308511

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., et al. (2016). Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth Space Sci.* 3, 388–415. doi:10.1002/2015EA000136

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., et al. (2021). GMD perspective: The quest to improve the evaluation of groundwater representation in continental-to global-scale models. *Geosci. Model Dev.* 14, 7545–7571. doi:10.5194/GMD-14-7545-2021

Gnann, S. J., Coxon, G., Woods, R. A., Howden, N. J. K., and McMillan, H. K. (2021). Tossh: A toolbox for streamflow signatures in hydrology. *Environ. Model. Softw.* 138, 104983. doi:10.1016/J.ENVSOFT.2021.104983

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol. X.* 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Process.* 22, 3802–3813. doi:10.1002/hyp.6989

Hallouin, T. (2021). *hydroeval: an evaluator for streamflow time series in Python*. Zenodo. doi:10.5281/ZENODO.4709652

Hay, L. (2019). Application of the national hydrologic model infrastructure with the precipitation-runoff modeling system (NHM-PRMS), by HRU calibrated version - ScienceBase-catalog. *U.S. Geol. Surv.* Available at: doi:10.5066/P9NM8K8WAccessed February 23, 2022)

Hay, L. E., Leavesley, G. H., Clark, M. P., Markstrom, S. L., Viger, R. J., and Umemoto, M. (2006). Step wise, multiple objective calibration of a hydrologic model for a snowmelt dominated basin. *J. Am. Water Resour. Assoc.* 42, 877–890. doi:10.1111/J.1752-1688.2006.TB04501.X

Hay, L. E., and Umemoto, M. (2007). Multiple-objective stepwise calibration using luca. Available at: http://www.usgs.gov/pubprod [Accessed June 7, 2022].

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., and Arheimer, B. (2016). Most computational hydrology is not reproducible, so is it really science? *Water Resour. Res.* 52, 7548–7555. doi:10.1002/2016WR019285

Jie, M. X., Chen, H., Xu, C. Y., Zeng, Q., and Tao, X. E. (2016). A comparative study of different objective functions to improve the flood forecasting accuracy. *Hydrology Res.* 47, 718–735. doi:10.2166/NH.2015.078

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., et al. (2018). Binder 2.0 - reproducible, interactive, sharable environments for science at scale. Proceedings of the 17th Python in Science Conference, July 9-15, 2018 Austin, Texas, 113–120. doi:10.25080/MAJORA-4AF1F417-011

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resour. Res.* 45, n/a. doi:10.1029/2008WR006912

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* 42, n/a. doi:10.1029/2005WR004362

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resour. Res.* 53, 867–890. doi:10.1002/2016WR019191

Krause, P., Boyle, D. P., and Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* 5, 89–97. Available at: doi:10.5194/adgeo-5-89-2005https://hal.archives-ouvertes.fr/hal-00296842/ (Accessed February 27, 2015)

LaFontaine, J. H., Hart, R. M., Hay, L. E., Farmer, W. H., Bock, A. R., Viger, R. J., et al. (2019). Simulation of water availability in the Southeastern United States for historical and potential future climate and land-cover conditions. *Sci. Investig. Rep.* 2019–5039, 83 doi:10.3133/SIR20195039

Lane, R., Coxon, G., E Freer, J., Wagener, T., J Johnes, P., P Bloomfield, J., et al. (2019). Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. *Hydrol. Earth Syst. Sci.* 23, 4011–4032. doi:10.5194/HESS-23-4011-2019

Markstrom, S. L., Hay, L. E., and Clark, M. P. (2016). Towards simplification of hydrologic modeling: Identification of dominant processes. *Hydrol. Earth Syst. Sci. Discuss.*, 20–4655-4671. doi:10.5194/hess-2015-508

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M. T., Payn, R. A., et al. (2015). PRMS-IV, the precipitation-runoff modeling system, version 4. Available at: https://pubs.usgs.gov/tm/6b7/[Accessed September 5, 2017].

Marshall, L., Sharma, A., and Nott, D. (2006). Modeling the catchment *via* mixtures: Issues of model specification and validation. *Water Resour. Res.* 42, 11409. doi:10.1029/2005WR004613

McMillan, H. K. (2021). A review of hydrologic signatures and their applications. *WIREs Water* 8, e1499. doi:10.1002/WAT2.1499

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., and Woods, R. A. (2011). Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure. *Hydrol. Process.* 25, 511–522. doi:10.1002/hyp.7841

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. *Hydrol. Process.* 34, 1393–1409. doi:10.1002/HYP.13632

Moges, E., Demissie, Y., and Li, H.-Y. (2016). Hierarchical mixture of experts and diagnostic modeling approach to reduce hydrologic model structural uncertainty. *Water Resour. Res.* 52, 2551–2570. doi:10.1002/2015WR018266

Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., and Larsen, L. G. (2022). Strength and memory of precipitation's control over streamflow across the conterminous United States. *Water Resour. Res.* 58, e2021WR030186. doi:10.1029/2021WR030186

Moriasi, D. N., Gitau, M. W., Pai, N., Daggupati, P., Gitau, M. W., Member, A., et al. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Trans. ASABE* 58, 1763–1785. doi:10.13031/TRANS.58.10715

National Operational Hydrologic Remote Sensing Center (2004). *Snow Data Assimilation System (SNODAS) Data Products at NSIDC, Version 1*. Boulder, Colorado USA: NSIDC: National Snow and Ice Data Center. doi:10.7265/N5TB14TC

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., and Gupta, H. V. (2020). Does information theory provide a new paradigm for Earth science? Hypothesis testing. *Water Resour. Res.* 56, e2019WR024918. doi:10.1029/2019WR024918

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., and Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *J. Hydrometeorol.* 19, 1835–1852. doi:10.1175/JHM-D-17-0209.1

Pechlivanidis, I. G., Jackson, B., McMillan, H., and Gupta, H. (2014). Use of an entropy-based metric in multiobjective calibration to improve model performance. *Water Resour. Res.* 50, 8066–8083. doi:10.1002/2013WR014537

Pechlivanidis, I. G., Jackson, B., and Mcmillan, H. (2010). "The use of entropy as a model diagnostic in rainfall-runoff modelling," in International Congress on Environmental Modelling and Software, July 5-8, 2010, Ottawa, Canada. Available at: http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings (Accessed February 23, 2022).

Peñuela, A., Hutton, C., and Pianosi, F. (2021). An open-source package with interactive Jupyter Notebooks to enhance the accessibility of reservoir operations simulation and optimisation. *Environ. Model. Softw.* 145, 105188. doi:10.1016/J.ENVSOFT.2021.105188

Pérez, F., and Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Comput. Sci. Eng.* 9, 21–29. doi:10.1109/MCSE.2007.53

Project Jupyter 2022 | Home Available at: https://jupyter.org/[Accessed February 24, 2022].

Regan, R. S., Juracek, K. E., Hay, L. E., Markstrom, S. L., Viger, R. J., Driscoll, J. M., et al. (2019). The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. *Environ. Model. Softw.* 111, 192–203. doi:10.1016/j.envsoft.2018.09.023

Regan, R. S., Markstrom, S. L., Hay, L. E., Viger, R. J., Norton, P. A., Driscoll, J. M., et al. (2018). Description of the national hydrologic model for use with the precipitation-runoff modeling system (PRMS). *Tech. Methods* 6, 38. doi:10.3133/tm6B9

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., and Bhatti, H. A. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *J. Hydrol. X.* 505, 276–290. doi:10.1016/J.JHYDROL.2013.10.006

Ruddell, B. L., Drewry, D. T., and Nearing, G. S. (2019). Information theory for model diagnostics: Structural error is indicated by trade-off between functional and predictive performance. *Water Resour. Res.* 55, 6534–6554. doi:10.1029/2018WR023692

Ruddell, B. L., and Kumar, P. (2009). Ecohydrologic process networks: 1. Identification. *Water Resour. Res.* 45. doi:10.1029/2008WR007279

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S. C., Knight, R., et al. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLOS Comput. Biol.* 15, e1007007. doi:10.1371/JOURNAL.PCBI.1007007

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLOS Comput. Biol.* 9, e1003285. doi:10.1371/JOURNAL.PCBI.1003285

Saxe, S., Farmer, W., Driscoll, J., and Hogue, T. S. (2021). Implications of model selection: A comparison of publicly available, conterminous US-extent hydrologic component estimates. *Hydrol. Earth Syst. Sci.* 25, 1529–1568. doi:10.5194/HESS-25-1529-2021

Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. doi:10.1103/PhysRevLett.85.461

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., and James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Sci. Data* 61 (6), 190030–190112. doi:10.1038/sdata.2019.30

Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., et al. (2021). Emdna: An ensemble meteorological dataset for North America. *Earth Syst. Sci. Data* 13, 3337–3362. doi:10.5194/ESSD-13-3337-2021

Thorstensen, A., Nguyen, P., Hsu, K., and Sorooshian, S. (2016). Using densely distributed soil moisture observations for calibration of a hydrologic model. *J. Hydrometeorol.* 17, 571–590. doi:10.1175/JHM-D-15-0071.1

Tian, F., Li, H., and Sivapalan, M. (2012). Model diagnostic analysis of seasonal switching of runoff generation mechanisms in the Blue River basin, Oklahoma. *J. Hydrol. X.* 418–419, 136–149. doi:10.1016/j.jhydrol.2010.03.011

Tijerina, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M. M., Sampson, K., et al. (2021). Continental hydrologic intercomparison project, phase 1: A large-scale hydrologic model comparison over the continental United States. *Water Resour. Res.* 57, e2020WR028931. doi:10.1029/2020WR028931

Wagener, T., Kelleher, C., Weiler, M., McGlynn, B., Gooseff, M., Marshall, L., et al. (2012). It takes a community to raise a hydrologist: The Modular Curriculum for Hydrologic Advancement (MOCHA). *Hydrol. Earth Syst. Sci.* 16, 3405–3418. doi:10.5194/HESS-16-3405-2012

Wagener, T., and Mcintyre, N. (2007). Tools for teaching hydrological and environmental modeling. *Comput. Educ. J.* 17 (3).

Weijs, S. V., Schoups, G., and van de Giesen, N. (2010). Why hydrological predictions should be evaluated using information theory. *Hydrol. Earth Syst. Sci.* 14, 2545–2558. doi:10.5194/hess-14-2545-2010

Yilmaz, K. K., Gupta, H. V., and Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resour. Res.* 44, n/a. doi:10.1029/2007WR006716

Zhang, L., Moges, E., Kirchner, J., Coda, E., Liu, T., Wymore, A. S., et al. (2021). Chosen: A synthesis of hydrometeorological data from intensively monitored catchments and comparative analysis of hydrologic extremes. *Hydrol. Process.* 35, e14429. doi:10.1002/HYP.14429

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 12 October 2022
DOI 10.3389/feart.2022.979823

OPEN ACCESS

EDITED BY
Daniele Ganora,
Politecnico di Torino, Italy

REVIEWED BY
Qiang Guo,
China Jiliang University, China
Cong Luo,
Hohai University, China
Ilaria Butera,
Politecnico di Torino, Italy

*CORRESPONDENCE
Catherine Moore,
c.moore@gns.cri.nz

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 28 June 2022
ACCEPTED 29 September 2022
PUBLISHED 12 October 2022

CITATION
Moore C, Scott D, Burbery L and Close M (2022), Using sequential conditioning to explore uncertainties in geostatistical characterization and in groundwater transport predictions. *Front. Earth Sci.* 10:979823. doi: 10.3389/feart.2022.979823

Using sequential conditioning to explore uncertainties in geostatistical characterization and in groundwater transport predictions

Catherine Moore[1]*, David Scott[2], Lee Burbery[2] and Murray Close[2]

[1]GNS Science, Wellington, New Zealand, [2]ESR, Christchurch, New Zealand

Rapid transmission of contaminants in groundwater can occur in alluvial gravel aquifers that are permeated by highly conductive small-scale open framework gravels (OFGs). This open framework gravel structure and the associated distribution of hydraulic properties is complex, and so assessments of contamination risks in these aquifers are highly uncertain. Geostatistical models, based on lithological data, can be used to quantitatively characterize this structure. These models can then be used to support analyses of the risks of contamination in groundwater systems. However, these geostatistical models are themselves accompanied by significant uncertainty. This is seldom considered when assessing risks to groundwater systems. Geostatistical model uncertainty can be reduced by assimilating information from hydraulic system response data, but this process can be computationally challenging. We developed a sequential conditioning method designed to address these challenges. This method is demonstrated on a transition probability based geostatistical simulation model (TP), which has been shown to be superior for representing the connectivity of high permeability pathways, such as OFGs. The results demonstrate that the common modelling practice of adopting a single geostatistical model may result in realistic predictions being overlooked, and significantly underestimate the uncertainties of groundwater transport predictions. This has important repercussions for uncertainty quantification in general. It also has repercussions if using ensemble-based methods for history matching, since it also relies on geostatistical models to generate prior parameter distributions. This work highlights the need to explore the uncertainty of geostatistical models in the context of the predictions being made.

KEYWORDS

uncertainty, geostatistical characterization, decision-support, sequential conditioning, groundwater, pathogens, contaminants

1 Introduction

Alluvial gravel aquifers are a valuable source of freshwater globally (Alsharhan and Rizk, 2020; De Luca et al., 2020). Due to their alluvial nature, such aquifers are inherently heterogeneous, being composed of various sedimentary textural classes that include open framework gravels (OFGs) (Cary, 1951; Lunt et al., 2004; Bridge & Lunt, 2006; Lunt & Bridge, 2007). A characteristic trait of OFGs is their macroporosity, which makes them highly permeable (Klingbeil et al., 1999; Ferreira et al., 2010). Dann et al. (2008), and Jussel et al. (1994) have shown through field tests and numerical modelling that the connectedness of OFGs has a profound effect on the hydraulic function of alluvial gravel aquifers, by facilitating preferential flow and rapid solute transport. OFGs also have a low capacity for microbial removal (Rossi et al., 1994; Pang, 2009; Flynn et al., 2015) and therefore pose a significant risk in regard to human exposure to pathogenic disease in situations where untreated drinking water is sourced from alluvial gravel aquifers. The 2016 [ground]waterborne Campylobacteriosis outbreak (around 7,000 cases) that occurred at Havelock North, New Zealand is a case example (Government Inquiry into Havelock North Drinking Water, 2017; Gilpin et al., 2020).

The work presented in this paper was motivated by the need for robust methods to evaluate groundwater contamination risks associated with alluvial gravel aquifer settings that incorporate OFG. Robust model-based assessments of contaminant risk in these groundwater systems are based on geostatistical models that characterize the structure of these rapid transport pathways. In this paper we focus on the geostatistical characterization of these most permeable pathways, and the implications of uncertainty in this characterization.

1.1 Geostatistical methods in hydrogeological modelling

There are many practical limitations to mapping the structure of alluvial gravels at a resolution that can detect OFGs. Heterogeneous aquifer datasets are almost always sparse and incomplete, particularly in lateral dimensions (Sanchez-Vila & Fernandez-Garcia, 2016). Consequently, stochastic inversion methods, even when coupled with distributed parameterizations, are unlikely to identify small-scale highly heterogeneous pathways (Doherty & Moore, 2021). Because of this, we use stochastic frameworks to characterize heterogeneity structure, based on geostatistical or physical process-based modelling methods (e.g. Riva et al., 2006; Riva et al., 2008; Ritzi & Soltanian, 2015; Scheibe et al., 2015; Siena & Riva, 2020).

Physical process-based modelling methods simulate structural aspects of alluvial deposits based on probabilistic representations of lithological categories within a meandering river geometry, and are informed by the sedimentary disposition of the system. Examples include BCS-3D (Webb & Anderson, 1996), FLUVSIM (Deutsch & Tran, 2002) and ALLUVSIM (Pyrcz et al., 2009). Geostatistical models can be based on covariance or variogram structures for continuously variable hydraulic properties (Deutsch & Journel, 1998). Other options, such as training image methods, including Multiple-Point Statistics, can be used to represent more complex geological environments that cannot be fully represented by two-point covariance relationships (Strebelle, 2002; Huysmans & Dassargues, 2009).

Where sharp interfaces occur between high and low conductivity media, such as in aquifers with OFGs, geostatistical models based on categorical variables can be used to generate realizations of aquifer media. Categorical methods include Sequential Indicator Simulation (SIS) which relies on indicator variograms based on borehole lithological data (Goovaerts, 1997; Deutsch & Journel, 1998), and Transition Probability (TP) Simulation (Carle, 1999). TP simulation has the advantage that it honors volumetric proportions, mean dimensions and the connectivity patterns of the categorical variables.

The choice of the most appropriate structural model of heterogeneity largely depends on the features that control the predictive response of concern (e.g., Jafarpour & Tarrahi, 2011; Ciriello et al., 2013; Riva et al., 2015). We opted to use TP simulation which has been shown to be superior for representing the connectivity of high permeability pathways (Siena & Riva, 2020). TP simulation is also well-established and used in numerous modelling studies, including groundwater modelling studies, that rely on a geostatistical description of the spatial dependencies of selected categories (e.g. Park et al., 2004; Engdahl & Weissmann, 2010; Hansen et al., 2014; He et al., 2014).

1.2 Direct and indirect data and sequential conditioning

The sparsity of information with which to develop geostatistical models of heterogeneity structure has motivated efforts to combine ‘direct’ observations made of mapped lithological properties with ‘indirect’ data (Refsgaard et al., 2012; Carle & Fogg, 2020) that require another level of interpretation that carries with it uncertainty. Pumping test data (Harp et al., 2008; Harp and Vessilinov, 2010; Harp & Vessilinov, 2012) and geophysical data (Engdahl & Weissmann, 2010; Koch, 2013; He et al., 2014; Zhu et al., 2016) are examples of ‘indirect’ observational data often used to condition and reduce the uncertainty of geostatistical models. Indirect data relating to dynamical physical processes such as flow and transport can be extremely informative, since they provide a measure of connectivity within the hydrogeological model (Renard and Allard, 2013).

Using information from indirect hydrogeologic data can nonetheless impose a significant computational burden, as this involves the comparison of field observations with groundwater model simulation outputs within a stochastic or Bayesian workflow (Jafarpour & Tarrahi, 2011; Ciriello et al., 2013; Linde et al., 2015; Riva et al., 2015). Selection of the conditioning method can also be problematic, potentially degrading the geological realism of the conditioned realizations when spatially distributed parameters are adjusted in order to provide a match to field observations (Oliveira et al., 2017; Chan & Elsheikh, 2020).

Sequential conditioning can be used to address the computational burden described above, where some data requires more computational processing effort than others (Feyen et al., 2003; Hassan et al., 2009; Dorn et al., 2012). Sequential conditioning commences by history matching to datasets that require the least computational effort. Each subsequent conditioning step is focused on a selection of observations requiring increasingly greater computational processing. Using this process allows the prior distribution to gradually morph into a posterior distribution. We develop an approach that harnesses the strengths, and mitigates the weaknesses, of two distinct conditioning methods: stochastic inversion and rejection sampling.

The stochastic inversion method uses a history matching approach to condition the TP model parameters through minimizing the residuals between the transition probabilities derived from the TP geostatistical model and those derived from the direct lithological data. The TP model parameters are further conditioned using "greater than" and "less than" constraints, *via* a rejection sampling methodology applied to the indirect observations.

Conditioning of geostatistical models is easier to discuss when adopting Bayesian nomenclature (Kennedy and O'Hagan, 2002). Therefore, in the sections that follow the geostatistical model parameters (facies lengths, and volumetric proportions) are referred to as 'hyperparameters', to distinguish them from 'model parameters' i.e., hydraulic properties of the underlying aquifer system being analyzed. Probability density functions of hyperparameters are thus used to describe the uncertainty of the geostatistical model. Note that if alternative geostatistical models were adopted the hyperparameters would differ: e.g. for a variogram based geostatistical model, the hyperparameters would comprise the sill, range and nugget parameters.

1.3 Research objectives

This paper explores the implications of geostatistical model uncertainty for a particle transport modelling problem in an alluvial gravel aquifer, where transport function is determined by the connected, small-scale OFG textural class (Dann et al., 2008; Burbery et al., 2017; Theel et al., 2020). It also demonstrates the potential of a Sequential Conditioning Approach, using a case study which contains a uniquely detailed field dataset (consisting of direct and indirect observations) that was initially described by Burbery et al. (2017). This case study includes data from novel smoke tracing experiments designed to characterize the connectivity of OFG pathways. These data are particularly valuable, given the advantages of conditioning to prediction-salient information (White et al., 2014; Doherty, 2015). Using the geostatistical model hyperparameter distribution derived from the analysis, we explored the predictive implications of the common practice of adopting a single most likely geostatistical model to underpin a groundwater contamination risk assessment.

The structure of the paper is as follows: in Section 2 we provide some background to the case study field site and describe the direct and indirect field observational datasets that were compiled for the alluvial gravel aquifer model used in the study. The mathematical methodologies and framework developed and tested in this study are described in Section 3. In Section 4 we present the results from our modelling analyses, whilst also exploring and discussing some implications that the equifinality of structural heterogeneity models have for predictions of travel time in alluvial gravel aquifers. Section 5 presents our conclusions and discusses implications of these for stochastic decision support modelling in practice.

2 Case study

The Canterbury Plains aquifer on the South Island, New Zealand (NZ) covers an area of approximately 8,000 km^2 and consists of sets of coalesced alluvial fans that were active during the Quaternary period (Leckie, 1994; Bal, 1996; Ashworth et al., 1999; Browne & Naish, 2003; Leckie, 2003). This extensive aquifer, used for irrigation, industrial and potable water supply (Bal, 1996; Brown, 2001) is ranked as the most valuable groundwater resource in NZ (White, 2001). The general sedimentary structure of the aquifer is characteristic of gravel outwash deposits formed from large braided-river systems. The Kyle field site (43.94338 S, 172.06788 E), from which the observational datasets used in this study were obtained, is located on the Rakaia River fan, on the coastal boundary of the Canterbury Plains. At the last glacial maxima, the site would have been positioned approximately mid-point on the Rakaia fan (Browne & Naish, 2003).

2.1 Lithological mapping (direct observational data)

A 3D portion of the alluvial deposits at Kyle has been mapped, covering an area measuring 28 m x 20 m and to a

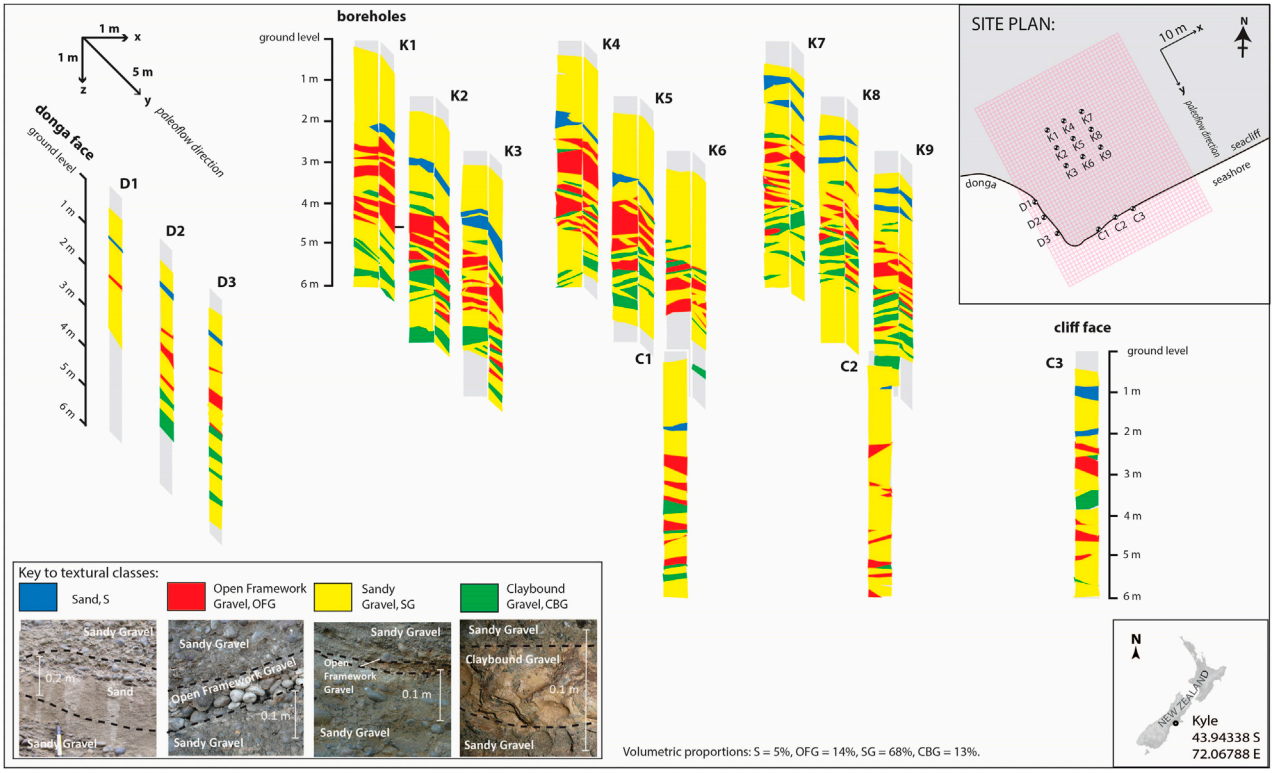

FIGURE 1
Box plot map of the Kyle site showing the spatial distribution of the four main lithological categories (textural classes), as identified on cliff exposures and borehole walls. Site location, and the model domain and grid that is used in the geostatistical and groundwater modelling analyses, are shown in inset of site plan. Representative photographic images of textural classes are included in the legend.

depth of 6 m below the Rakaia fan surface. Mapping was conducted from two cliff exposures (a sea cliff oriented perpendicular to the presumed paleoflow direction and a 'donga', i.e. a steep-sided gully created by soil erosion, with a face aligned 90° to the sea cliff), and nine large (1.2 m) diameter boreholes (coded K1 to K9) that were drilled 5 m apart on a 3x3 uniform grid, 18 m inland from the cliffs (see Figure 1 inset). The lithological examination is limited to the unsaturated zone. However, these vertically-stacked gravel packages mapped near the surface at Kyle represent a sample of the alluvial sequence that forms the Canterbury Plains aquifer system. Therefore we are able to make the assumption that this sample provides a useful analogue model of a saturated gravel aquifer system. Further details of the Kyle field site and investigative methods are provided in Burbery et al. (2017).

Employing descriptive methods such as those described by Koltermann & Gorelick (1996), Burbery et al. (2017) compiled a map of the Kyle site adopting four lithological categories, being: sand (S); sandy gravel (SG); open framework gravel (OFG) and clay-bound gravel (CBG). Photographic examples of the four lithological categories, as imaged at Kyle, are presented in Figure 1. Particle size distribution data, and a description of the geological depositional history for this alluvial system can be found in Burbery et al. (2017). The relative compositions of the lithological categories at the Kyle site are: S 5.1%, SG 67.8%, OFG 14.2% and CBG 12.9%.

OFG at the Kyle site predominantly occur as cross-strata comprising packaged sets of alternating OFG and SG. The thickness of the OFG lithological category is dependent on the angle of the foresets and was observed to vary between 0.2 m and 1 m. The lateral extent of OFG seen in cliff exposures was more than 25 m in some cases. Although less common, OFGs at Kyle also feature as planar beds up to 0.3 m thick and 4 m–5 m wide. From exposures observed on the donga face that is orientated along the paleoflow direction, it is apparent that the planar beds can extend for at least 15 m in length (Burbery et al., 2017).

On the basis of pumping and tracer test data, Dann et al. (2008) established that typical hydraulic conductivities of OFG in the Canterbury Plains alluvial aquifer system are two to three orders of magnitude greater (i.e. more permeable) than those of the other three categories (i.e. OFG have a saturated conductivity of 1,498–10,646 m/day compared with a range for the other three categories from 5.5 to 117.28 m/day). This is consistent with Klingbeil et al. (1999) who described the average hydraulic

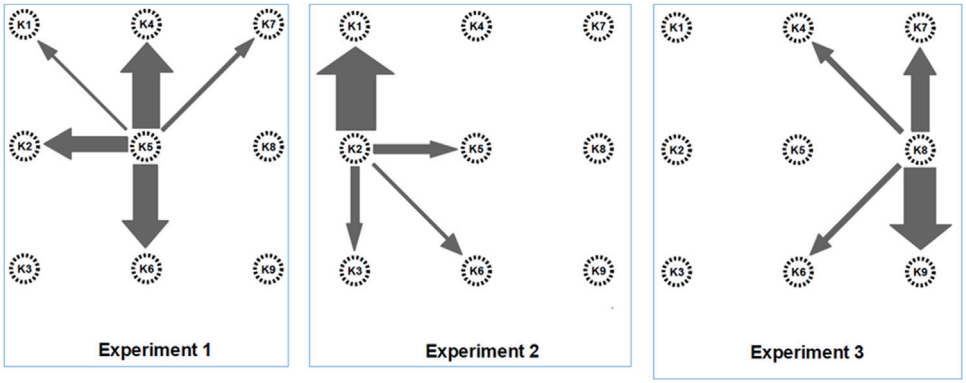

FIGURE 2
Schematic of smoke tracer test results. Arrows indicate connections between injection boreholes and observation points. The width of the arrows indicates the relative strength of those connections of high conductivity pathways as inferred from the observed arrival times.

conductivity of OFG being around 100 times greater than the other alluvial categories studied. Dann et al. (2008) estimated that this hydraulic conductivity contrast between lithological categories results in approximately 98% of aquifer flow occurring through these permeable connected OFGs. Therefore, it is the connectivity of these rapid transport OFG pathways that is relevant to the representation of pathogen transport in aquifers (Fiori et al., 2013; Hunt & Johnson 2017).

We adopt a 'forecast first' approach to the development of the geostatistical model (Doherty 2015; White 2017) and focus our analysis on the connectivity of the OFG category identified from observations to generate random realizations of the OFG and other lithological categories at the scale of the numerical grid. Assignment of hydraulic conductivity values reflected the order of magnitude differences in conductivity between the most permeable (OFG) and the next most permeable facies (S, SG, CBG), resulting in a binary characterization of the hydraulic properties within the aquifer system. While at first glance this grouping may seem to overly simplify the geological characterization described in Burbery et al. (2017), this simplification has no impact on the predictions we are making, given the contrast in permeability between the OFG and the other lithological categories, which are considered to be analogous to hydrofacies (Soltanian & Ritzi, 2014; Theel et al., 2020).

2.2 Smoke tracer tests to determine connectivity of OFG (indirect observational data)

Three smoke tracer tests were conducted using the array of open boreholes described to examine the interconnectedness of OFGs at Kyle. The tests are documented in detail in Burbery et al. (2017). In brief, they involved injecting smoke under a low positive pressure into each of the centrally-located boreholes K2, K5 and K8 for an extended period. The arrival time and position of smoke emerging from OFG in neighboring boreholes was then recorded. The results of the smoke experiments confirmed that OFGs are truly 'open' since smoke was able to travel rapidly between boreholes through OFGs. Connectivity was found to be non-uniform in direction, reflecting both the heterogeneity and anisotropy of the alluvial sediments. Figure 2 illustrates the connectivity between OFGs, as inferred from the three smoke tracer tests at boreholes K2, K5 and K8.

The earliest arrival time for smoke transmitted between two boreholes located 5 m apart was 48 s, along a co-set of planar OFG strata running between K2 and K1 that were aligned with the paleoflow direction. For all tests, fastest velocities corresponded to the mean paleoflow direction, i.e. NNW-SSE or *y*-direction (Figure 1). In some cases, no smoke was detected to have travelled between adjacent boreholes, suggestive of no apparent connectivity between the observed OFG strata. When transmission between boreholes occurred, the latest observed smoke arrival time was 30 min, between boreholes K5 and K1. A specific result of the smoke tracer tests was the lack of any observable hydraulic connection between OFG in test borehole K5 and OFG mapped in both K8 and K9, to the north (Figure 2).

The lithological data from borehole-logs, outcrops and geological characterization (Figure 1), and the observed cross-borehole connectivity from these smoke tracer experiment results (Figure 2), provided the respective direct, and indirect, observational data that were used to derive the geostatistical model of the most permeable and least permeable categories in this aquifer structure. The following sections describe the processes and mathematical methods that utilized these different types of information.

3 Methodology

3.1 Sequential conditioning of heterogeneity structure models

The sequential conditioning approach provides a Monte Carlo implementation of Bayes theorem and involves a sequence of history matching steps. The initial step focusses on the data which is the most rapid to process. Subsequent conditioning steps are only applied to those parameter ranges identified as plausible from the preceding step, and targets data that is increasingly slower to process. The approach has computational advantages over joint inversion if assimilating information from multiple datasets with one dataset involving a simulation model with long run times (Feyen et al., 2003; Hassan et al., 2009).

Previous sequential conditioning studies have adopted a single conditioning approach (Feyen et al., 2003; Hassan et al., 2009; Dorn et al., 2012). We combine two stochastic inversion approaches to further reduce the computation burden of history matching to disparate datasets. Computationally efficient stochastic inversion methods, such as randomized maximum likelihood approaches, are used to condition observation groups where possible. However, rejection sampling is used if stochastic inversion risks degrading the representation of important spatially defined geological features, such as a connected flow pathway, as demonstrated by Dorn et al. (2012) with observations of cross-borehole connectivity in a fractured rock aquifer. While rejection sampling is too computationally inefficient for most groundwater modelling contexts, this burden is alleviated when using it only in final conditioning steps (Dorn et al., 2012; Linde et al., 2015; Cirpka & Valocchi, 2016; Carle & Fogg, 2020). In this way the different strengths of conditioning methods can be employed where appropriate, while the respective weaknesses of each method are mitigated.

We applied this sequential conditioning approach using direct and indirect geological observations. Direct observations were comprised of lithological log data and were processed using a stochastic inversion approach. Indirect observations of cross-borehole connectivity, derived from the case study tracer test, were processed using rejection sampling.

3.2 Geostatistical model

Geostatistical models based on transition probability (TP) simulation are used in a number of fields (e.g. Huang et al., 2017; Li & Zhang, 2019) to characterize the distribution and juxtapositional characteristics of heterogeneity, such as the connected high permeability features of interest to this case study. We adopted the T-PROGS software for our TP model implementation (Carle (1999), which has been used widely in the hydrogeological field. Carle (1999) defines a transition probability, t_{ij}, as:

'Given that a facies j is present at location x, what is the probability that another (or the same) facies i occurs at location $x+h$', or:

$$t_{ij}(h) = P\{j \text{ occurs at } x + h \mid i \text{ occurs at } x\} \tag{1}$$

Borehole-log and outcrop data is catalogued into categories, at regular depth intervals, allowing the juxtapositional probabilities of lithological categories to be calculated. These are summarized in matrices of transition probabilities at specific lags (h), in the vertical (z) and horizontal directions, i.e. along the mean paleoflow direction (y or dip direction) and transverse to this direction (x or strike direction). The collation of these transition probabilities at specified separation distances (lags) can also be depicted as a transiogram (Figure 5). These transition probability matrices form the lithological constraints in the sequential conditioning approach.

From Carle & Fogg (1997), the mean length $\bar{L}$ of the ith category unit in a particular direction can be calculated as:

$$\bar{L}_i = \frac{1}{r_{ii}} = \left[\frac{\partial t_{ii}(0)}{\partial h}\right]^{-1} \tag{2}$$

where r_{ii} is the auto-transition rate, and t_{ii} is the auto-transition probability.

A range of assumptions can be adopted to simplify the calculation of transition probability matrices. Given the emphasis of this study is on the impact of the geostatistical representation of the connectivity of the high permeability OFG category on the uncertainty, we adopt the simplifying constraint of assuming that the probability of transition from the ith to the jth category is solely dependent on the volumetric probability of the jth facies as discussed in Harp and Vesselinov (2010).

This allows the probability of a transition from one category to another to be expressed as a function of the volumetric proportions, p, and the mean lengths of the categories:

$$t_{ij}(h) = (1 - t_{ii}(h))\frac{p_j}{1 - p_i} \; for \; j \neq i \tag{3a}$$

Which can be expressed for transition rates as:

$$r_{ij} = -r_{ii}\frac{p_j}{1 - p_i} \; for \; j \neq i \tag{3b}$$

3.3 Conditioning using direct lithological constraints with stochastic inversion

We adopted the Null Space Monte Carlo (NSMC) procedure (Tonkin & Doherty, 2009), as implemented in the PEST software suite (Doherty, 2016), which provides a close approximation to randomized maximum likelihood methods. This automated the task of exploring the range of TP model hyperparameters that provides a fit to the empirical transition probabilities. The objective function, $J_{lithology}$, used in the

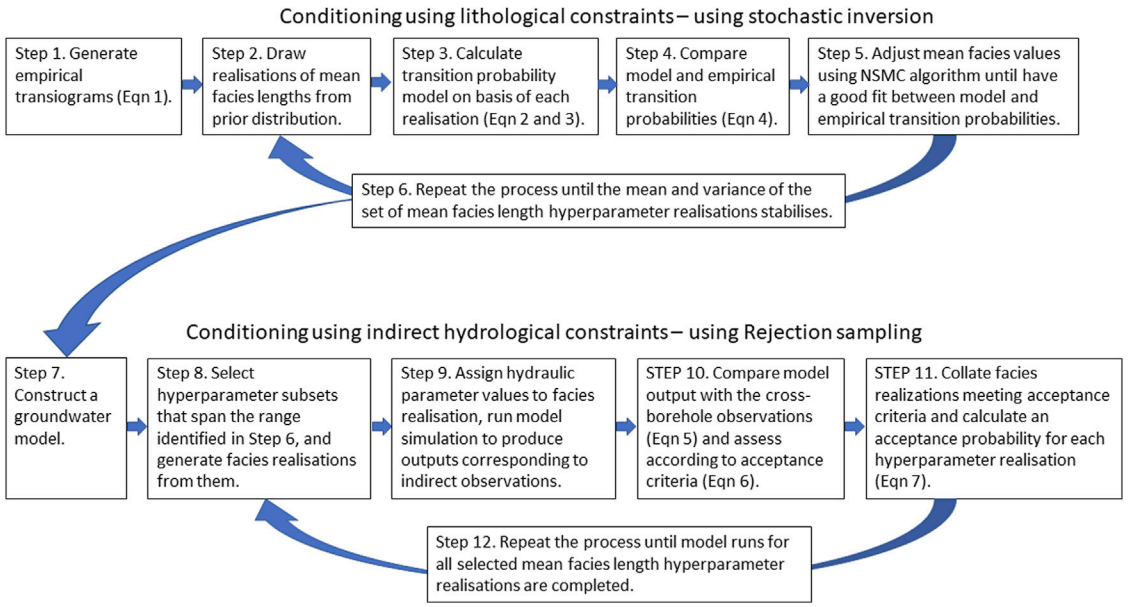

FIGURE 3
Flow chart showing the steps involved in the sequential Monte Carlo modelling procedure.

inversion procedure is the L2 norm, or sum of squared residuals between the transition probabilities derived from the geostatistical model and those derived from the lithological data:

$$J_{lithology} = min_{\beta \epsilon B} \sum_{i=1}^{M} \left(t_{ki}\widehat{(h_{\varnothing})}(\boldsymbol{B}) - t_{ki}(h_{\varnothing}) \right)^2 \quad (4)$$

where empirical transition probabilities $t_{ki}(h_{\varnothing})$ are calculated directly from the lithology. The modelled transition probabilities $t_{ki}\widehat{(h_{\varnothing})}(\boldsymbol{B})$, are derived from the TP model with hyperparameters in the vector β which has upper and lower limits as defined by B. M refers to the number of specified lags $h_{\varnothing}$ at which transition probabilities are calculated. The parameter bounds B are informed by the estimates of mean category lengths from the borehole and cliff observations and analyses in Burbery et al. (2017). These mean lengths in the paleoflow, transverse and vertical directions are the TP model hyperparameters being estimated by the sequential conditioning approach (refer to Table 1).

The following steps summarize this process of conditioning the TP models, using stochastic inversion (also depicted in Figure 3).

1. Generate empirical transiograms from the lithological logs at specified lags (Eq. 1).
2. Draw realizations from a prior hyperparameter distribution of the mean lengths in the paleoflow, transverse and vertical directions of the OFG lithological category.
3. Derive TP model based on the mean lengths, and the calculated textural class proportions from lithological logs for the same specified lags as in Eq. 1) above (Eqs 3a and 3b).
4. Compare modelled and empirical transition probabilities at specified lag distances (Eq. 4).
5. Adjust hyperparameter values using stochastic inversion algorithm until a good fit between modelled and measured transition probabilities is obtained.
6. Repeat the process until first and second moments of the distribution of mean OFG lengths have stabilized.

3.4 Conditioning using indirect hydrological constraints with rejection sampling

We adopted a straightforward implementation of rejection sampling, consistent with the Generalized likelihood uncertainty estimation (GLUE) method introduced by Beven & Binley (1992). It relies on generating multiple realizations from a prior probability distribution, running the model with each realization, and comparing model outputs with the observations. Each realization that does not provide a good fit to measured observations is rejected.

However, the 'indirect' nature of the observations of cross-borehole connectivity requires many additional steps before the information in these observations can be used to condition TP model hyperparameters. For every hyperparameter set we generate multiple aquifer heterogeneity realizations. Each realization is then

TABLE 1 Rate of producing plausible fields for 21 alternative geostatistical model hyperparameters, and the OFG mean lengths in the paleoflow (L_y) and perpendicular to paleoflow (L_x) directions. Representative samples from the range of plausible geostatistical model hyperparameters identified in the initial conditioning step were selected for prediction uncertainty analysis are identified by a suffix (a to i).

OFG mean length		L_y^2/L_x, (m)[a]	Acceptance probability $\widehat{P_A}$
Paleoflow direction, L_y (m)	Orthogonal to paleoflow direction, L_x (m)		
1.99	4.86	0.81[a]	0
0.06	0.01	0.3	0
1.26	10.2	0.16	0.005
1.39	3.09	0.6	0.03
1.26	10.41	0.15	0.04
1.25	3.93	0.4	0.08
1.47	2.19	0.98	0.12
1.27	2.05	0.8	0.2
1.26	2.21	0.71	0.22
1.26	2.06	0.77	0.28
1.26	1.95	0.81	0.44
2.72	1.67	4.43	1.06
0.99	0.02	56.1[b]	2.6
2.49	1.34	4.6[c]	8.0
1.08	0.02	73.2[d]	11.7
1.12	0.02	56.9[e]	17.4
0.93	0.01	87[f]	23.0
0.96	0.03	31[g]	26.7
2.84	0.23	35[h]	38.5
6.47	0.26	158.6	80.0
2.37	0.03	178.1[i]	80.1

*The suffix identifies the TP, models evaluated in particle track modelling.

converted to a hydraulic parameter field and used in a groundwater model simulation. Based on work in Dann et al. (2008 and 2009), this assignment of hydraulic conductivity values for OFG is 100 times that of the other three categories. The groundwater model simulates a flow process that provides outputs that correspond to the observations of cross-borehole connectivity. The details of this groundwater model are discussed in the following section.

Cross-borehole connectivity observations can be denoted as $flux_i$. When a model-to-measurement comparison is made, it is assessed according to an acceptance metric $\varnothing_{hydrology(\theta)}$which defines the sum of differences between the observed and modelled cross-borehole fluxes across all n_{obs} observed fluxes, i.e. the sum of squared residuals or L2 norm:

$$\varnothing_{\boldsymbol{hydrology}(\boldsymbol{\theta})} = \sum_{i=1}^{n_{obs}} \left(\boldsymbol{flux}_i(\boldsymbol{\theta}) - \boldsymbol{flux}_i\right)^2 \quad (5)$$

where θ defines the vector of hydraulic parameters defining the realization, which are generated from a specific geostatistical hyperparameter set and random seed. An acceptable fit threshold is applied to the fits between modelled and observed cross-borehole fluxes, defined as $\varnothing_{hydrology(max)}$. This is used to define a set of acceptable realizations, $\Omega(\varnothing_{hydrology(\theta)})$ as:

$$\Omega\left(\varnothing_{hydrology(\theta)}\right) = \begin{cases} 0, \varnothing_{hydrology(\theta)} > \varnothing_{hydrology(max)} \\ 1, \varnothing_{hydrology(\theta)} \le \varnothing_{hydrology(max)} \end{cases} \quad (6)$$

Note that this set will vary if the acceptance threshold is varied.

Collating the heterogeneity realizations that meet the acceptance threshold in Eq. 6 provides an acceptance probability $\widehat{P_A}$ for each mean hydrofacies length hyperparameter set Eq. 7:

$$\widehat{P_A} = \frac{1}{N_r} = \sum_{i=1}^{N_r} \Omega\left(\varnothing_{hydrology(i)}\right) \quad (7)$$

where N_r is the total number of realizations.

In summary, this rejection sampling process requires the generation of both hyperparameter realizations (also described as hyperparameter sets in the discussion that follows to avoid confusion with the aquifer heterogeneity realizations) as well as heterogeneity realizations and the hydraulic conductivity realizations based on them. The following steps summarize this process (as depicted in Figure 3):

7. Construct a groundwater flow model that simulates a flow field to represent the cross-borehole connectivity observations revealed by the smoke tracer test.
8. Select a subset of hyperparameter realizations that span the range of mean OFG length values defined in step 6 and generate an ensemble of heterogeneity realizations on the basis of each selected hyperparameter realization.
9. Assign hydraulic parameter values to the high and low conductivity categories of each realization, import these into the flow model, and run the flow model simulation to provide model outputs that correspond to cross-borehole observations.
10. Compare model outputs with these indirect observations (Eq. 5) and retain or reject the heterogeneity realization depending on whether the model-to-observation fit is sufficient to meet the selected acceptance criteria (Eq. 6).
11. Collate those heterogeneity realizations that meet the acceptance criteria and calculate an acceptance probability for each geostatistical model hyperparameter set.
12. For each of the selected subset of geostatistical hyperparameter realizations, return to Step 9 and continue until all realizations have been completed.

3.5 The flow model

Heterogeneity realizations were generated at a regular fine-scale grid discretization which covered the case study site (Step

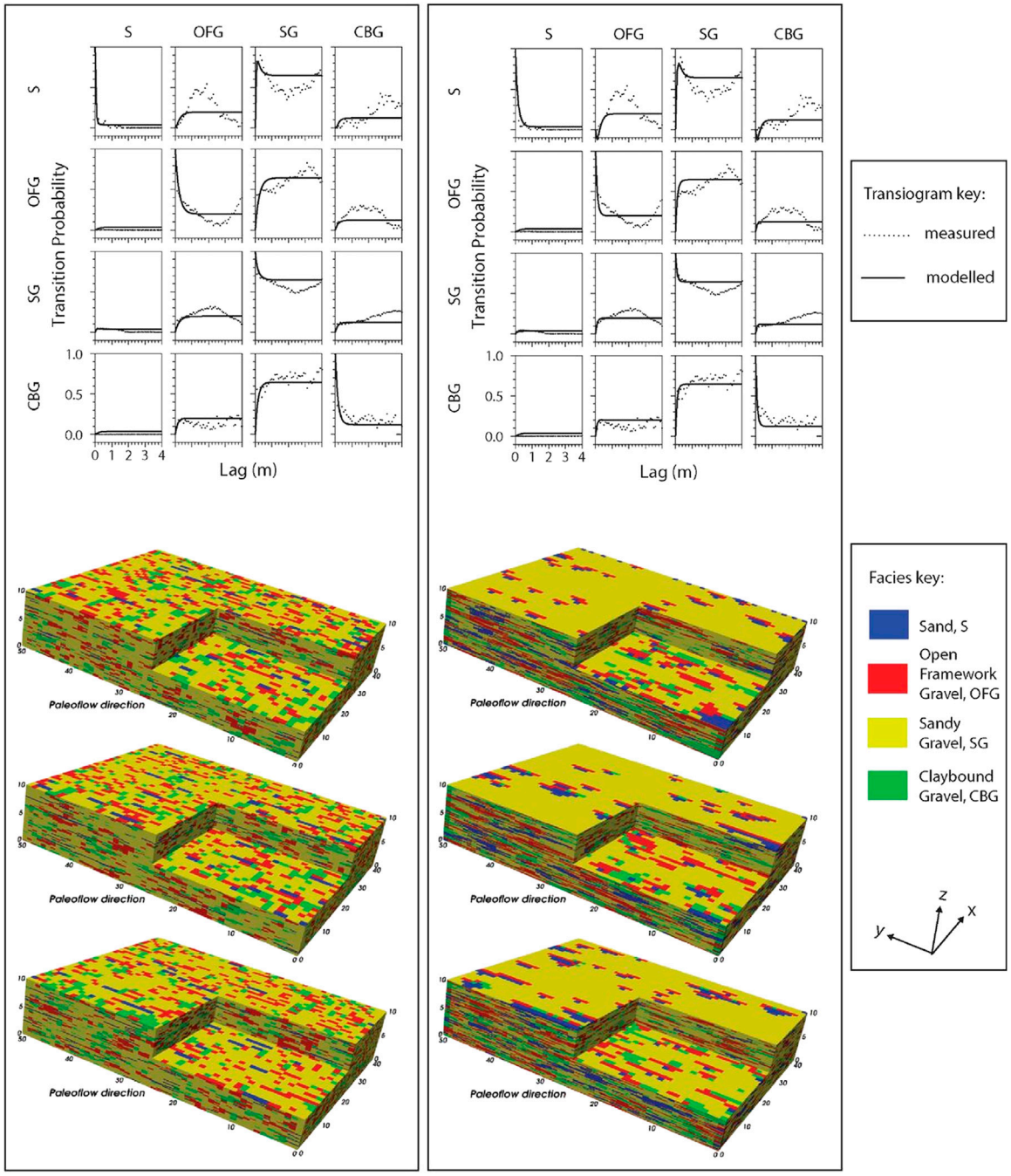

FIGURE 4
Modelled vertical transition probabilities (black line), and calculated transition probabilities from the lithological data (black dots) at the Kyle site for two statistically valid TP models are shown in the left and right columns. The left and right-hand columns correspond to suffix 'b', and 'i' in Table 1. Below the transiograms are three heterogeneity realizations relating to the TP models shown in the transiograms.

8 above): an area of 40 m by 50 m by 10 m, with a grid cell dimension of 1 m by 1 m in the horizontal direction, and 0.1 m in the vertical direction. As noted above, each category (i.e. OFG and the combined SG, S and CBG category) was assigned a hydraulic conductivity value that represented the mean value for that category, transforming the heterogeneity realizations to hydraulic conductivity field realizations. Note that the cross-borehole flow observations are dependent on the relative hydraulic parameter values, rather than absolute hydraulic parameter values, and in this case the OFGs have a

conductivity value that is three orders of magnitude higher than the other three textural classes.

Each hydraulic conductivity realization was used in a series of flow model simulations, implemented using the USGS software MODFLOW (Harbaugh et al., 2000), where cross-borehole connectivity observations derived from the smoke tracer test were modelled as being analogous to groundwater flow. The relative connectivity derived from the smoke tracer tests, was simulated using a steady-state flow field. In this flow field the boreholes were represented as constant head cells. For each of the smoke tracer experiments, the smoke injection borehole was assigned a positive head and remaining boreholes were assigned a zero head. No-flow boundaries were assigned around the edge of the model domain. The resulting pattern of cross-borehole flows was evaluated. This approach was possible because it was not necessary to simulate the smoke particle movement, as only the cross-borehole connectivity observations are used to screen out improbable heterogeneity structures in this study.

Using MODFLOW to simulate the flow of fluids other than groundwater, such as vapor transport in the unsaturated zone, relies on the analogies between these two flow problems (USEPA, 1995), and is appropriate where differential pressures are low, as demonstrated in Massmann (1989). This approach has also been used when simulating oil flows (Hsieh, 2011), when simulating multiphase flow in coal seam gas problems (Herckenrath et al., 2015), and can be used in advection-dispersion contexts (Rubbab et al., 2016).

Using this approach, three flow simulations were undertaken, representing each of the smoke injection bore tests depicted in Figure 2. This approach was both sufficient for simulation of the cross-borehole flow connectivity observations and provided a convenient and rapid-to-deploy surrogate for smoke transport simulation. Individual realizations were considered 'plausible' if the sum of the simulated head dependent discharges corresponding to the three most dominant connections, depicted in Figure 2, comprised 50% or more of the total discharge. The approximate and categorical nature of this acceptance criteria implicitly accounted for measurement and conceptual uncertainty that would impact on the precision of model to measurement fits and resembles Dorn et al. (2012) who used observations of the degree of fracture connectivity with similarly approximate acceptance criteria.

3.6 Assessment of prediction uncertainty using single and multiple plausible TP models

The implications of adopting a single geostatistical model are explored for simulations of groundwater transit times. The transit time prediction was simulated with a particle moving through a saturated steady-state flow field, with fixed head boundaries at opposite west and east sides of the model domain, and no flow boundaries at the north and south sides. The flow field was derived from a hydraulic property field relating to the aquifer heterogeneity realizations generated in the same manner as Step 9 above. These predictions were generated for 10 hyperparameter realizations which spanned the range of mean length hyperparameters identified with low to high acceptance criteria in the rejection sampling step. A total of 1,000 heterogeneity realizations were generated for each of the ten selected hyperparameter sets.

The groundwater model domain and hydraulic property values were the same as those used to simulate the cross-borehole connectivity observations described above. Constant head cells were placed at the upstream and downstream extent of the model domain. MODFLOW (Harbaugh et al., 2000) was used to simulate this groundwater flow field, and the transit time of the particle was simulated using the MODFLOW ADV package (Anderman & Hill, 2001). Transit time probability distributions for selected geostatistical model hyperparameter realizations were then explored.

4 Results and discussion

We examine and discuss the results of the numerical experiments from this study within a contaminant transport predictive context. Contaminant transport is very sensitive to the disposition of highly permeable pathways in aquifer systems (Lee et al., 2007; Fiori, & Jankovic, 2012; Soltanian & Ritzi, 2014; De Barros et al., 2016; Sanchez-Vila & Fernàndez-Garcia, 2016; Theel et al., 2020). This is particularly so for pathogen transport where risks are largely related to the fastest transit times through an aquifer, as pathogen numbers reduce over time at a rate governed by the half-life of the pathogen of concern (Hunt & Johnson, 2017).

Specifically, this section discusses the performance of the methodology used to condition the geostatistical model of OFG rapid transport pathways. The stochastic exploration of the ill-posed geostatistical model discussed in this section has similarities to the approaches described in Zhu et al. (2016), and Harp & Vesselinov (2012). This section also discusses the implications of the remaining geostatistical model uncertainty for pathogen transport predictions. The OFG lithological category controls how much and how quickly groundwater flows in our case study (Dann et al., 2009), and the geostatistical model hyperparameters correspond to the mean lengths of this category in the paleoflow direction (y or dip direction), transverse to this direction (x or strike direction), and the z-direction.

4.1 Sequential conditioning: Performance of initial stochastic inversion step

A total of 130 TP model hyperparameter realizations were generated from the stochastic inversion process, with the mean

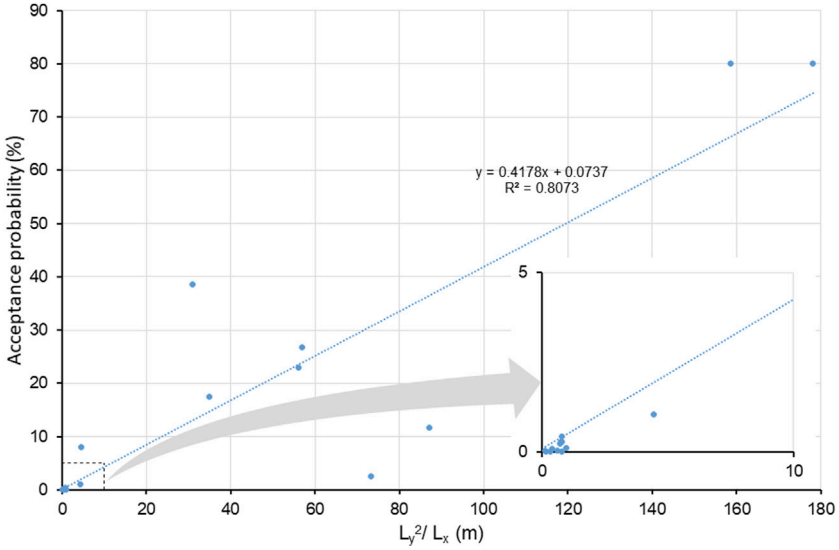

FIGURE 5
Relationship between the proportion of realizations meeting the smoke tracer plausibility test (Acceptance probability) and a ratio L_y^2/L_x, which is the squared OFG category mean length in the paleoflow direction (L_y) and perpendicular to the paleoflow direction (L_x). This relationship is shown for 21 TP models examined for the Kyle case study as summarized in Table 1.

and standard deviations of the initial conditioning of the hyperparameter distribution having stabilized with this number of conditioned realizations. Each of the conditioned hyperparameter realizations resulted in modelled-to-measured transition probability fits with correlation coefficients of approximately 0.8. The standard deviations of the TP model hyperparameters (mean lengths of OFG in the paleoflow, transverse and vertical directions) were significantly reduced in this initial conditioning step from the reasonably unconstrained prior in the log domain of three log(m) in the x and y directions and one log(m) in the z-direction to 0.14 log(m) in the x-direction, 0.2 log(m) in the y-direction and 0.22 log(m) in the z-direction.

Despite the substantial reduction in the uncertainty achieved through the initial conditioning step, the remaining hyperparameter uncertainty can result in very different heterogeneity structures (Figure 4). For example, Figure 4 (upper part of Figure 4), shows two modelled-to-measured transiogram fits, where OFG mean lengths in the paleoflow y-direction (L_y) were 0.99 m for the left-hand side of the figure and 2.37 m on the right-hand side. A comparison of the heterogeneity realizations corresponding to these two hyperparameter sets (lower part of Figure 4) shows more elongated connected OFGs in the right-hand side of the plot, than those on the left.

The non-uniqueness of the TP model hyperparameters, as depicted in Figure 4, reflects the lack of information regarding the OFG category mean lengths in the lateral direction (L_x). Non-uniqueness of geostatistical models is acknowledged and discussed in several studies (Harp & Vessilinov, 2012; Koch, 2013; He et al., 2014; Siena & Riva, 2020). Despite this, the impact of geostatistical model equifinality on the quantification of the uncertainty of groundwater model predictions of interest is seldom considered by practitioners in decision support models (Sanchez-Villa & Fernandez-Garcia, 2016). A component of the barrier to the uptake of stochastic methods relates to the computational burden of their implementation (Linde et al., 2015).

4.2 Sequential conditioning: Performance of rejection sampling step

A selection of 21 hyperparameter sets were selected so that they spanned the OFG hyperparameter range identified in the first conditioning step. Details of these 21 hyperparameter sets are listed in Table 1. Heterogeneity realizations were then generated, with more than 17,000 heterogeneity realizations being generated for each of the selected hyperparameter sets. Each realization was then used in a flow simulation of cross-borehole connectivity. The proportion of flow simulations, associated with each heterogeneity realization, that met the cross-borehole connectivity acceptance criteria ranged from 0 to 81%. This approach of ranking the plausibility of TP model hyperparameter sets using a system response-based acceptance criteria was also used by Harp &

Vesselinov (2010) and Dorn et al. (2012), when conditioning geostatistical facies and fracture network models respectively.

Examination of the relationship between acceptance probabilities and the disposition of the OFG provided important information. The range of the mean vertical thickness (L_z) of the OFG category was reasonably well constrained on the basis of the lithological data, and varied from 0.13 m to 0.43 m, with an average of 0.23 m. These small variations in these OFG thicknesses did not impact on the plausibility of hyperparameter sets, whereas the lateral (L_y) and transverse (L_x) OFG dimensions did.

Figure 5 shows the relationship between OFG category mean lengths and acceptance probabilities, expressed using a ratio of the squared mean length in the paleoflow y-direction (L_y) and the x-direction (L_x). The ratio between L_y and L_x can convey the relationship between the relative mean lengths in the x and y-directions, whereas the absolute magnitude of the mean length is incorporated into the ratio by squaring the L_y term. By ranking the mean OFG dimensions by their propensity to produce realizations which generate model outputs that are consistent with the indirect observations, additional information is provided about the spatial disposition of the OFGs.

Various relationships between acceptance probability and the mean length hyperparameters were analyzed to explore this information. The absolute magnitude of these ratios was found to be important when simulating the cross-borehole flows, correlating strongly with acceptance probabilities. Long and narrow OFGs, elongated in the paleoflow direction, were found to be positively correlated with higher acceptance probabilities (Figure 5; Table 1). In contrast, small OFG lengths in the paleoflow direction, or large OFG widths (direction orthogonal to paleoflow), had very low acceptance probabilities.

Table 1 lists the OFG mean lengths in the paleoflow direction, (L_y), and orthogonal to this direction, (L_x), for the 21 selected hyperparameter sets. It also lists the L_y^2/L_x ratio and acceptance probabilities, ($\widehat{P_A}$) for the 21 selected hyperparameter sets, which are also depicted in Figure 5. Of the realizations summarized in Table 1, two different geostatistical models achieved the highest acceptance probability of 80%; these occurred for the models which had a L_y^2/L_x ratio of 178 and 158.6. These ratios corresponded to mean lengths between 2.37 m and 6.47 m in the paleoflow y-direction (L_y) and mean widths of 0.03 m–0.26 m in the x-direction (L_x). These acceptance probability figures indicate that the lack of OFG connectivity in the direction orthogonal to paleoflow (x-direction) was as important as the connectivity in the paleoflow y-direction when reproducing the cross-borehole connectivity observations. This strongly directional dependent nature of the OFG connectivity was not indicated by the lithological logs alone. Some of the other TP models listed in Table 1, while not producing such high acceptance probabilities, still provided some realizations which meet the acceptance criteria. Therefore, those geostatistical model hyperparameter sets cannot be discounted as valid.

The overall greater connectivity of the OFG in the paleoflow direction across the range of hyperparameters explored is depicted in probability plots as shown in Figure 6. Figure 6A shows the probability of OFGs occurring for all selected TP models, denoted 'a' to 'i' in Table 1, spanning the range of acceptance probabilities from 0 to 80%. A probability of one occurs where OFGs were observed directly in the lithological logs, as depicted in Figures 6A and 6B, with probabilities approaching one clustered around borehole locations. With greater distance from the boreholes, these probabilities gradually reduce to background levels of 0.14, representing the bulk proportion of OFGs at this site.

Figure 6B depicts the probability of OFGs occurring in any model cell, for a single TP model (denoted as 'a' in Table 1) which corresponds to an acceptance probability of zero. Figure 6B depicts wider OFGs in the direction orthogonal to the paleoflow direction. These wider OFG's would tend to allow greater cross-bore connectivity orthogonal to the paleoflow direction, which is inconsistent with the smoke tracer test observations.

Harter (2005), Fogg et al. (2000), and Fogg & Zhang (2016) discuss how the upper 12–28% of a hydraulic conductivity distribution will tend to be fully connected in 3-D created random fields. This proportional range of high conductivity facies encompasses the bulk proportion of OFGs in this study. Fogg & Zhang (2016) assert that this connectivity will lead to laterally and vertically extensive rapid transport pathways. The OFG probabilities shown in Figure 6A provide an indication of the disposition of such connected pathways.

4.3 Sequential conditioning summary

The sequential conditioning approach adopted, combining stochastic inversion and rejection sampling was able to support the conditioning to direct and indirect geological observations. Rejection sampling revealed the anisotropy of the OFGs more fully and was required only when conditioning to the cross-bore flow observations, to adhere to the conceptual model for geologic heterogeneity. Many more model runs would have been required if rejection sampling had been applied as part of a joint inversion approach. This study provides a demonstration of using a sequential conditioning approach comprised of two history matching methods to successfully negotiate the pitfalls of numerical inefficiency on one hand and degradation of the geological model on the other.

4.4 Implications of geostatistical model equifinality for contaminant risk assessments

Particle tracking simulations are often used to assess the risks associated with rapid transport rates of pathogens in groundwater (Hunt & Johnson, 2017). We adopted this approach when assessing the impact on transit time predictions incurred by adopting a single geostatistical model. In total, 1,000 realizations were used from selected geostatistical

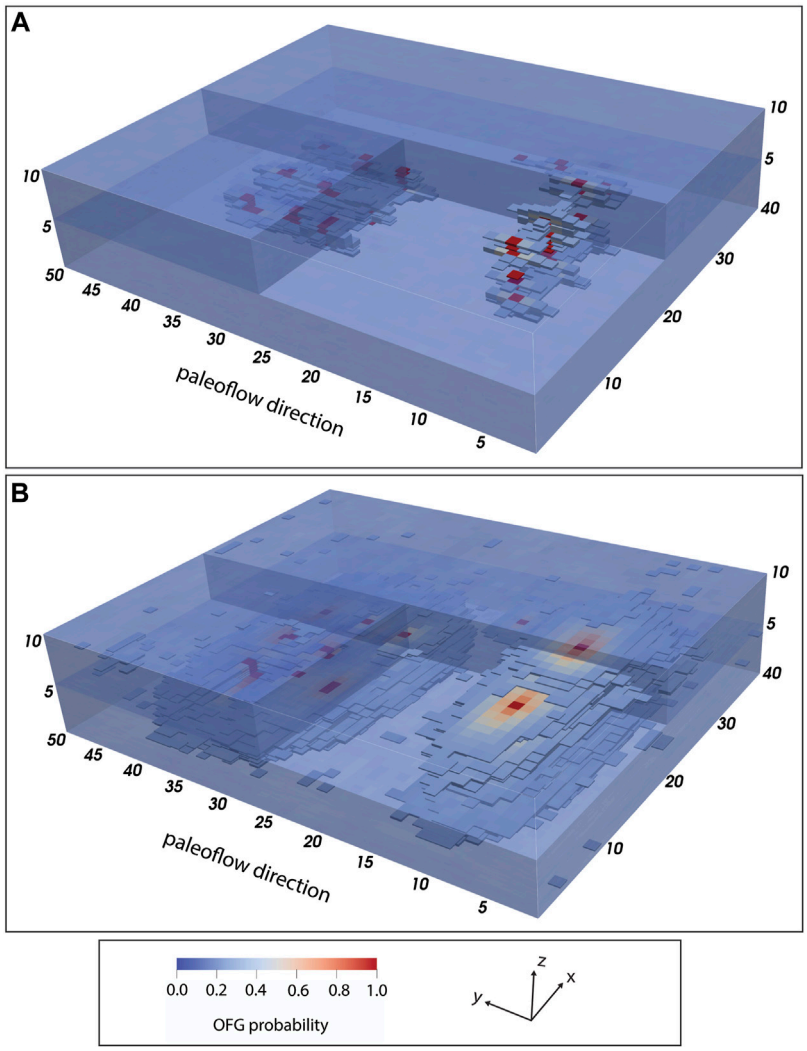

FIGURE 6
OFG probability models showing the spatial probability of OFGs occurring at a probability of greater than 0.25: **(A)** for combined realizations from TP models labelled 'a' to 'l' in Table 1; and **(B)** for realizations from a single TP model labelled a0 in Table 1.

models (identified as (a) to (i) in Table 1). This subset of nine out of 21 hyperparameters sets explored were selected to span the range of acceptance probabilities identified.

The distribution of particle paths simulated on the basis of each realization, that were generated from the nine selected TP model hyperparameter sets, are mapped in plan-view in Figure 7. The corresponding acceptance probability listed in Table 1 is also noted in Figure 7 for each model, with model (i) having the highest acceptance probability of 80%. Blue particle paths correspond to all realizations generated for each hyperparameter set, while red particle paths correspond to only those realizations which met the cross-borehole acceptance criteria. Figure 7 shows particle tracks have less lateral spread and are more closely aligned with the groundwater flow direction for those models associated with higher acceptance probabilities. This is consistent with the narrow and longer disposition of OFG pathways identified through conditioning to the cross-borehole connectivity observations.

The transit times corresponding to the particle tracks shown in Figure 7 were collated and are summarized in the box and whisker plot of Figure 8. Note that these times are normalized by using a porosity value that scales the maximum travel times for hyperparameter set (i) to be approximately 1 day, allowing the relativity of these travel times to be depicted rather than their absolute magnitude. For example, the hyperparameter set (c) results in a maximum travel time that is 500% greater than that of the hyperparameter set (i). The hyperparameter set (i) in the

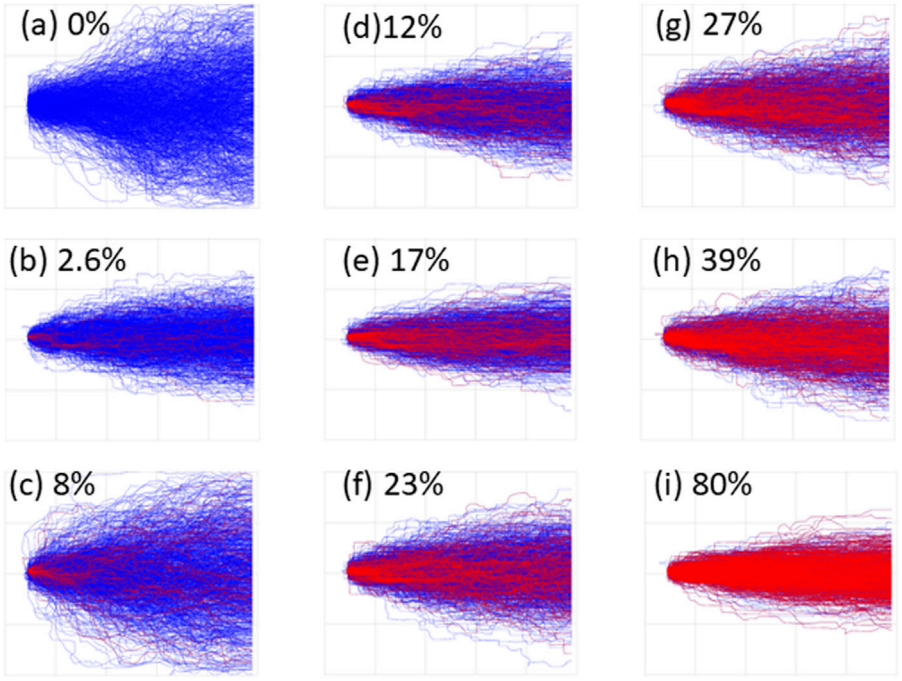

FIGURE 7
Particle tracks for selected TP models **(a–i)** are shown in blue. Particle tracks for plausible realizations for each model are shown in red. The acceptance probability for each of the figures are shown alongside the relevant model label.

bottom right of Figure 8 provides the fastest particle travel times, and also provides the highest acceptance probability.

A comparison of the results from the other eight hyperparameter sets (a to h), is illuminating. It shows that particle transit times which met the cross-borehole acceptance criteria can fall well outside the range defined by the hyperparameter set (i) which had the highest acceptance probability. This has implications for analysis of predictive uncertainty. An analysis of transit time based on a single geostatistical model hyperparameter set with the highest acceptance probability, could significantly under-estimate the transit time uncertainty. Instead, a robust uncertainty analysis may need to consider predictions simulated from realizations generated from TP model hyperparameter sets with lower and higher acceptance probabilities. Therefore, while conditioning can reduce geostatistical model uncertainty, the results indicate that multiple plausible geostatistical models need to be considered for a prediction of concern in decision support modelling. This issue was also raised in Harp & Vessilinov (2012) when examining the quantification of the uncertainty of aquifer drawdown predictions.

This has practical implications for model-based risk assessments, and highlights the importance of considering the specific prediction being made when selecting geostatistical models. For this case study, when simulating pathogen contamination risks, the greatest risks would be exposed using model hyperparameter set (i), and would not have been exposed using model (b). If instead the concern had been related to inefficient use of land, due to over estimation of source protection zone areas, the greatest risk would be exposed by adopting model hyperparameter sets (b) or (f).

4.5 Additional considerations

In other prediction uncertainty studies based on TP geostatistical models, it may be important to also represent the variability of hydraulic properties within a defined category. Riva et al. (2006) found that representing hydraulic conductivity variability within a category could led to elongated capture zones, where this variability was being used to represent small-scale preferred flow paths. In contrast, Copty & Findikakis (2002) found that internal variability of hydraulic conductivity within defined categories had no significant impact on the predictions of a solute particle transport.

Those two studies exemplify the fact that any requirements for representing inter-category variability are context specific. Where realizations have been generated at significantly larger scales than the hydraulic property variability occurs, the

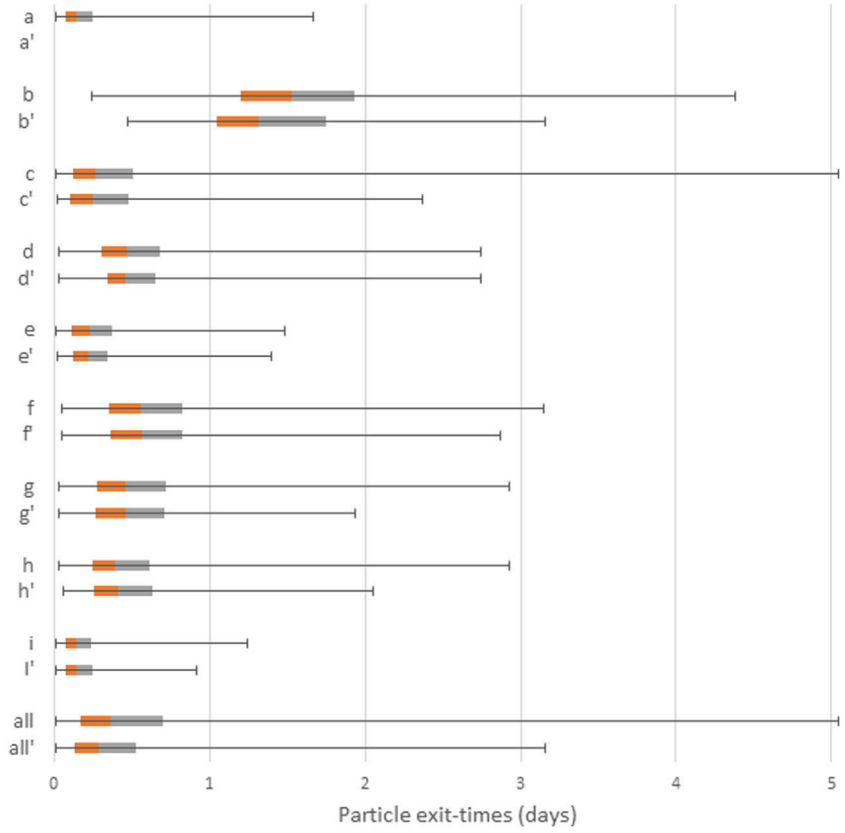

FIGURE 8
Box and whisker plots showing relative particle exit-times for TP models **(a–i)** together with their combined times (all). The 'suffix represents the exit-times derived from the plausible realizations for each TP model. Box and whiskers represent the minimum, maximum and lower and upper quartile of the particle exit-times.

geostatistical model is upscaled, and hence inter-category variability may be required if the prediction is sensitive to this variability. The study reported here sought to represent the fine scale detail of the permeable connected OFG category, populating a very fine model grid to explicitly represent preferred pathways. Therefore, the need to represent inter category variability was avoided in this study.

The 1 m by 1 m grid discretization used in this study required only slight upscaling of the transition probabilities in the *x*-direction, perpendicular to the paleoflow direction. The stochastic nature of prediction specific heterogeneity at field scales was not addressed in this study and remains a research challenge (Fogg et al., 2000; Fogg & Zhang, 2016; Doherty & Moore, 2019). Initial explorations into this field of research include hierarchical nested model frameworks (Li et al., 2006; Sreekanth & Moore, 2018), or the hybrid multiscale methods outlined in Scheibe et al. (2015). Local and global upscaling approaches can be used to derive the field scale versions of these stochastic models of aquifer heterogeneity if a robust fine-scale characterization of the heterogeneity exists (Fengjun et al., 2003; Chen, 2009; Zhou et al., 2010; Li et al., 2015; Soltanian et al., 2015; Li & Durlofsky, 2016).

5 Conclusion

Groundwater model predictions of contaminant transport, particularly pathogen contaminants, are sensitive to small-scale high permeability pathways. The ability to improve the characterization of geostatistical models that are used to describe these rapid transport pathway distributions is central to improving models used for water management decision support. However, the extent to which uncertainties in geostatistical models can influence the outcomes of a contaminant transport risk assessment is not well understood and is typically neglected in modelling practice.

In this study we have put considerable effort into conditioning geostatistical model hyperparameters and exploring their uncertainty, but it is common practice to adopt a single geostatistical model parameterization. We demonstrate that the selection of a single

geostatistical model, without consideration of its relevance to the predictions that matter, may prevent the simulation of real predictive possibilities, undermining the quantification of risks. Risk based assessments may need to consider alternative geostatistical models in the context of particular types of predictions (e.g. predictions dependent on time of travel) to guarantee robust decision support.

Conditioning with both direct and indirect aquifer information can mitigate the impact of sparse geological data, allowing the uncertainty of geostatistical models to be significantly reduced. In this study the detailed lithological study documented in Burbery et al. (2017) supplemented by the observations of cross-borehole connectivity from smoke tracer tests, enabled geostatistical models of the risk salient aspects of OFG pathways in alluvial gravels to be defined i.e. the connectivity of the OFG rapid transport pathways. To the best of our knowledge, no studies combining fine detailed lithological logs from alluvial deposits with *in-situ* measurements of the connectivity of rapid transport pathways, have previously been used to derive a geostatistical model of small-scale high permeability groundwater pathways. The fine-scale characterization of highly permeable OFG pathway structure, made accessible by this study, provides a much-needed basis for the assessments of risk in these contexts.

The significant computational burden involved when conditioning geostatistical models with direct and indirect data is made more challenging if efficient conditioning methods risk degrading the geological realism of the geostatistical model. To address this, we developed a sequential conditioning approach that combines alternative history matching methods. This approach enables each dataset to be conditioned with the history matching method best suited to that data. Datasets that involve more processing are scheduled for processing in later steps, thereby supporting better management of the computational load.

This work also provides a basis for future research. The fine-scale characterization made accessible by this study provides a much-needed basis for the analysis of upscaled hydraulic parameters to field scales that account for the presence of OFG pathways when assessing the risk of early arrival times of pathogen contaminants. With better characterization of geostatistical models of rapid transport pathways, we can more reliably model groundwater flow and contaminant transport and provide improved environmental decision support at a range of scales.

The particle tracking predictive scenario discussed in this paper demonstrates the implications of the geostatistical model uncertainty in one specific predictive context. Future research may also explore an advection-dispersion transport predictive simulation to assess the implications of geostatistical model uncertainty for transport predictions that are more affected by factors such as dispersion or chemical reactions.

In summary, the results of this study illustrate the importance of considering geostatistical uncertainty, in the context of specific predictions. Adoption of a single geostatistical model can result in realistic predictions being overlooked. Prediction specific geostatistical models need to be selected, such as those of rapid transport pathways explored in this study, to ensure robust assessments of risk. Combining acceptable realizations from multiple credible geostatistical models, to ensure that the true predictive uncertainty range is conveyed, may be required. Alternatively, selection of a worse-case geostatistical model for a particular prediction could be adopted. This has important practical implications for uncertainty quantification and history matching when using ensemble-based methods, which are based on geostatistical models to generate prior parameter distributions.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

CM was the primary author of this article, with all co-authors also contributing to the manuscript. CM and DS undertook the numerical experiments documented in the paper. The ideas and concepts in the manuscript were co-developed principally by CM and DS with contributions from LB and MC. The case study details were all supplied by LB. The original concept and funding for this work was secured by MC. All authors have seen and approved the final version of the manuscript being submitted. They warrant that the article is the authors' original work, has not received prior publication and is not under consideration for publication elsewhere.

Acknowledgments

This paper represents a collaboration between ESR and GNS Science, both in New Zealand, and was funded by the New Zealand Ministry of Business Innovation and Employment (grant nos. C03X1001 and C05X1803, and was also supported by GNS Science Groundwater Strategic Science Investment Fund (SSIF)).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alsharhan, A. S., and Rizk, Z. E. (2020). "Gravel aquifers," in *Water resources and integrated management of the United Arab Emirates* (Cham: Springer), Vol. 3. doi:10.1007/978-3-030-31684-6_11

Anderman, E. R., and Hill, M. C. (2001). *MODFLOW-2000: The U.S. Geological survey modular ground-water model – documentation of the advective-transport observation (ADV2) package, version 2: U.S*, 54. Geological Survey Open-File Report, 69. doi:10.3133/ofr0154

Ashworth, P., Best, J. L., Peakall, J., and Lorsong, J. A. (1999). "The influence of aggradation on braided alluvial architecture: Field study and physical scale modelling of the ashburton gravels, Canterbury plains, New Zealand," in *Fluvial sedimentology VI*. Editors N. D. Smith and J. Rogers (Special Publication of International Association of Sedimentologists), 28, 333–346.

Bal, A. A. (1996). Valley fills and coastal cliffs buried beneath an alluvial plain: Evidence from variation of permeabilities in gravel aquifers, Canterbury plains, New Zealand. *J. Hydrology (New Zealand)* 35 (1), 1–27.

Beven, K. J., and Binley, A. M. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* 6, 279–298. doi:10.1002/hyp.3360060305

Bridge, J. S., and Lunt, I. A. (2006). "Depositional models of braided rivers," in *Braided rivers: Process, deposits, ecology and management*. Editors G. S. Sambrook-Smith, J. L. Best, C. S. Bristow, and G. E. Petts (Oxford: Blackwell), 11–49. doi:10.1002/9781444304374.ch2

Brown, L. J. (2001). "Canterbury,". *Groundwaters of New Zealand*. Editors M. R. Rosen and P. A. White (Wellington: The New Zealand Hydrological Society), 441–459.

Browne, G. H., and Naish, T. R. (2003). Facies development and sequence architecture of a late quaternary fluvial-marine transition, Canterbury plains and shelf, New Zealand: Implications for forced regressive deposits. *Sediment. Geol.* 158 (1-2), 57–86. doi:10.1016/S0037-0738(02)00258-0

Burbery, L. F., Jones, M. A., Moore, C. R., Abraham, P., Humphries, B., and Close, M. E. (2017). "Study of connectivity of open framework gravel facies in the Canterbury Plains aquifer using smoke as a tracer," in *Geology and geomorphology of alluvial and fluvial fans: Terrestrial and planetary perspectives* (London: Geological Society). doi:10.1144/SP440.10

Carle, S. F., and Fogg, G. E. (2020). Integration of soft data into geostatistical simulation of categorical variables. *Front. Earth Sci.* 8. doi:10.3389/feart.2020.565707

Carle, S. F., and Fogg, G. E. (1997). Modeling spatial variability with one and multidimensional continuous-lag Markov chains. *Math. Geol.* 29 (7), 891–918. doi:10.1023/a:1022303706942

Carle, S. (1999). *T-PROGS: Transition probability geostatistical software*. Davis: University of California. http://gmsdocs.aquaveo.com/t-progs.pdf.

Cary, A. S. (1951). Origin and significance of openwork gravel. *Trans. Am. Soc. Civ. Eng.* 116 (1), 1296–1308. doi:10.1061/TACEAT.0006486

Chan, S., and Elsheikh, A. H. (2020). Parametrization of stochastic inputs using generative adversarial networks with application in geology. *Front. Water* 2, 5. doi:10.3389/frwa.2020.00005

Chen, T. (2009). New methods for accurate upscaling with full-tensor effects. PhD thesis, Energy and Resources Engineering, Stanford University. Available at https://pangea.stanford.edu/ERE/pdf/pereports/PhD/Chen09.pdf.

Ciriello, V., Di Federico, V., Riva, M., Cadini, F., de Sanctis, J., Zio, E., et al. (2013). Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. *Stoch. Environ. Res. Risk Assess.* 27, 945–954. doi:10.1007/s00477-012-0616-7

Cirpka, O. A., and Valocchi, A. J. (2016). Debates—stochastic subsurface hydrology from theory to practice: Does stochastic subsurface hydrology help solving practical problems of contaminant hydrogeology? *Water Resour. Res.* 52 (12), 9218–9227. doi:10.1002/2016WR019087

Copty, N. K., and Findikakis, A. N. (2002). Uncertainty analysis of a well capture zone under multiple scales of heterogeneity. in "*Calibration and reliability in groundwater modelling: A few steps closer to reality: Proceedings of ModelCARE'2002*". Prague, Czech Republic: IAHS Publ.

Dann, R., Close, M. E., Flintoft, M. J., Hector, R., Barlow, H., Thomas, S., et al. (2009). Characterization and estimation of hydraulic properties in an alluvial gravel vadose zone. *Vadose Zone J.* 8 (3), 651–663. doi:10.2136/vzj2008.0174

Dann, R., Close, M. E., Pang, L., Flintoft, M. J., and Hector, R. (2008). Complementary use of tracer and pumping tests to characterize a heterogeneous channelized aquifer system in New Zealand. *Hydrogeol. J.* 16, 1177–1191. doi:10.1007/s10040-008-0291-4

De Barros, F. P. J., Bellin, A., Cvetkovic, V., Dagan, G., and Fiori, A. (2016). Aquifer heterogeneity controls on adverse human health effects and the concept of the hazard attenuation factor. *Water Resour. Res.* 52 (8), 5911–5922. doi:10.1002/2016wr018933

De Luca, D. A., Lasagna, M., and Debernardi, L. (2020). Hydrogeology of the Western Po plain (piedmont, NW Italy). *J. Maps* 16 (2), 265–273. doi:10.1080/17445647.2020.1738280

Deutsch, C. V., and Journel, A. G. (1998). *Geostatistical software library and user's guide*. 2nd Edn. New York, NY: Oxford University Press.

Deutsch, C. V., and Tran, T. T. (2002). Fluvsim: A program for object-based stochastic modeling of fluvial depositional systems. *Comput. Geosciences* 28 (4), 525–535. doi:10.1016/S0098-3004(01)00075-9

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models*. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., and Moore, C. (2021). "Decision support modelling viewed through the lens of model complexity," in *A GMDSI monograph* (South Australia: National Centre for Groundwater Research and Training, Flinders University). doi:10.25957/p25g-0f58

Doherty, J., and Moore, C. R. (2019). Decision support modeling: Data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater* 58 (3), 327–337. doi:10.1111/gwat.12969

Doherty, J. (2016). *User's manual for PEST version 14.2*. Brisbane, Queensland, Australia: Watermark Numerical Computing, 339.

Dorn, C., Linde, N., Le Borgne, T., Bour, O., and Klepikova, M. (2012). Inferring transport characteristics in a fractured rock aquifer by combining single-hole ground-penetrating radar reflection monitoring and tracer test data. *Water Resour. Res.* 48, W11521. doi:10.1029/2011WR011739

Engdahl, N. B., and Weissmann, G. S. (2010). Anisotropic transport rates in heterogeneous porous media. *Water Resour. Res.* 46, W02507. doi:10.1029/2009WR007910

Fengjun, Z., Reynolds, A. C., and Oliver, D. S. (2003). The impact of upscaling errors on conditioning a stochastic channel to pressure data. *SPE J.* 8, 13–21. doi:10.2118/83679-PA

Ferreira, J. T., Ritzi, R. W., and Dominic, D. F. (2010). Measuring the permeability of open-framework gravel. *Ground Water* 48 (4), 593–597. doi:10.1111/j.1745-6584.2010.00675.x

Feyen, L., Gómez-Hernández, J. J., Ribeiro, P. J., Jr., Beven, K. J., and De Smedt, F. (2003). A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations. *Water Resour. Res.* 39 (5). doi:10.1029/2002WR001544

Fiori, A., Dagan, G., Jankovic, I., and Zarlenga, A. (2013). The plume spreading in the MADE transport experiment: Could it be predicted by stochastic models? *Water Resour. Res.* 49, 2497–2507. doi:10.1002/wrcr.20128

Fiori, A., and Jankovic, I. (2012). On preferential flow, channeling and connectivity in heterogeneous porous formations. *Math. Geosci.* 44 (2), 133–145. doi:10.1007/s11004-011-9365-2

Flynn, R. M., Mallen, G., Engel, M., Ahmed, A., and Rossi, P. (2015). Characterizing aquifer heterogeneity using bacterial and bacteriophage tracers. *J. Environ. Qual.* 44 (5), 1448–1458. doi:10.2134/jeq2015.02.0117

Fogg, G. E., Carle, S. F., and Green, C. (2000). "A connected-network paradigm for the alluvial aquifer system," in *Theory, modelling and field investigation in hydrogeology: A special volume in honor of shlomo P. Neuman's 60th birthday, GSA special paper 348*. Editors D. Zhang and C. L. Winter (Boulder, CO: Geological Society of America), 348, 25–42.

Fogg, G. E., and Zhang, Y. (2016). Debates—stochastic subsurface hydrology from theory to practice: A geologic perspective. *Water Resour. Res.* 52, 9235–9245. doi:10.1002/2016WR019699

Gilpin, B. J., Walker, T., Paine, S., Sherwood, J., Mackereth, G., Wood, T., et al. (2020). A large scale waterborne Campylobacteriosis outbreak, Havelock North, New Zealand. *J. Infect.* 81 (3), 390–395. doi:10.1016/j.jinf.2020.06.065

Goovaerts, P. (1997). "Geostatistics for natural resources evaluation," in *Applied geostatistics series* (New York, NY: Oxford University Press).

Government Inquiry into Havelock North Drinking Water (2017). Report of the Havelock North drinking water Inquiry: Stage 1. Available at https://www.dia.govt.nz/Stage-1-of-the-Water-Inquiry.

Hansen, A. L., Gunderman, D., He, X., and Refsgaard, J. (2014). Uncertainty assessment of spatially distributed nitrate reduction potential in groundwater using multiple geological realizations. *J. Hydrology* 519, 225–237. doi:10.1016/j.jhydrol.2014.07.013

Harbaugh, A. W., Banta, E. R., Hill, M. C., and McDonald, M. G., (2000). *MODFLOW-2000: The U.S. Geological survey modular ground-water model – user guide to modularization concepts and the ground-water flow process*. U.S Geological Survey Open-File, 121. doi:10.3133/ofr200092

Harp, D., Dai, Z., Wolfsberg, A., Vrugt, J., Robinson, B., and Vesselinov, V. (2008). Aquifer structure identification using stochastic inversion. *Geophys. Res. Lett.* 35, L08404. doi:10.1029/2008GL033585

Harp, D. H., and Vesselinov, V. V. (2012). Analysis of hydrogeological structure uncertainty by estimation of hydrogeological acceptance probability of geostatistical models. *Adv. Water Resour.* 36, 64–74. doi:10.1016/j.advwatres.2011.06.007

Harp, D. H., and Vesselinov, V. V. (2010). Stochastic inverse method for estimation of geostatistical representation of hydrogeologic stratigraphy using borehole logs and pressure observations. *Stoch. Environ. Res. Risk Assess.* 24 (7), 1023–1042. doi:10.1007/S00477-010-0403-2

Harter, T. (2005). Finite-size scaling analysis of percolation in three-dimensional correlated binary Markov chain random fields. *Phys. Rev. E* 72 (2), 026120. doi:10.1103/PhysRevE.72.026120

Hassan, A., Bekhit, H., and Chapman, J. (2009). Using Markov chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model. *Environ. Model. Softw.* (24), 749–763. doi:10.1016/j.envsoft.2008.11.002

He, X., Koch, J., Sonnenborg, T. O., Jorgensen, F., Schamper, C., and Refsgaard, J. C. (2014). Transition probability-based stochastic geological modeling using airborne geophysical data and borehole data. *Water Resour. Res.* 50, 3147–3169. doi:10.1002/2013WR014593

Herckenrath, D., Doherty, J., and Panday, S. (2015). Incorporating the effect of gas in modelling the impact of CBM extraction on regional groundwater systems. *J. Hydrology* 523 (3–4), 587–601. doi:10.1016/j.jhydrol.2015.02.012

Hsieh, P. A. (2011). Application of MODFLOW for oil reservoir simulation during the deepwater horizon crisis. *Ground Water* 49 (3), 319–323. doi:10.1111/j.1745-6584.2011.00813.x

Huang, X., Li, J., Liang, Y., Wang, Z., Guo, J., and Jiao, P. (2017). Spatial hidden Markov chain models for estimation of petroleum reservoir categorical variables. *J. Pet. Explor. Prod. Technol.* 7, 11–22. doi:10.1007/s13202-016-0251-9

Hunt, R. J., and Johnson, W. P. (2017). Pathogen transport in groundwater systems: Contrasts with traditional solute transport. *Hydrogeol. J.* 25, 921–930. doi:10.1007/s10040-016-1502-z

Huysmans, M., and Dassargues, A. (2009). Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer (Belgium). *Hydrogeol. J.* 17, 1901–1911. doi:10.1007/s10040-009-0495-2

Jafarpour, B., and Tarrahi, M. (2011). Assehe performance of the ensemble Kalman filter for subsurface flow data integration under variogram uncertainty. *Water Resour. Res.* 47 (5), W05537. doi:10.1029/2010WR009090

Jussel, P., Stauffer, F., and Dracos, T. (1994). Transport modeling in heterogeneous aquifers: 1. Statistical description and numerical generation of gravel deposits. *Water Resour. Res.* 30, 1803–1817. doi:10.1029/94WR00162

Kennedy, M. C., and O'Hagan, A. (2002). Bayesian calibration of computer models. *J. R. Stat. Soc. B* 63, 425–464. doi:10.1111/1467-9868.00294

Klingbeil, R., Kleineidam, S., Asprion, U., Aigner, T., and Teutsch, G. (1999). Relating lithofacies to hydrofacies: Outcrop-based hydrogeological characterisation of quaternary gravel deposits. *Sediment. Geol.* 129 (3), 299–310. doi:10.1016/S0037-0738(99)00067-6

Koch, J. (2013). "Geological heterogeneity in the Norsminde catchment – a hydrogeological perspective," Master's thesis: University of Copenhagen.

Koltermann, C. E., and Gorelick, S. M. (1996). Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches. *Water Resour. Res.* 32 (9), 2617–2658. doi:10.1029/96WR00025

Leckie, D. A. (1994). Canterbury Plains, New Zealand: Implications for sequence stratigraphic models. *Am. Assoc. Pet. Geol. Bull.* 78 (8), 1240–1256. doi:10.1306/A25FEABD-171B-11D7-8645000102C1865D

Leckie, D. A. (2003). Modern environments of the Canterbury plains and adjacent offshore areas, New Zealand — An analog for ancient conglomeratic depositional systems in nonmarine and coastal zone settings. *Bull. Can. Petroleum Geol.* 51 (4), 389–425. doi:10.2113/51.4.389

Lee, S-Y., Carle, S. F., and Fogg, G. E. (2007). Geologic heterogeneity and a comparison of two geostatistical models: Sequential Gaussian and transition probability-based geostatistical simulation. *Adv. Water Resour.* 30, 1914–1932. doi:10.1016/j.advwatres.2007.03.005

Li, H., and Durlofsky, L. J. (2016). Local–global upscaling for compositional subsurface flow simulation. *Transp. Porous Med.* 111, 701–730. doi:10.1007/s11242-015-0621-7

Li, J., Lei, Z., Qin, G., and Gong, B. (2015). Effective local-global upscaling of fractured reservoirs under discrete fractured discretization. *Energies* 8, 10178–10197. doi:10.3390/en80910178

Li, S.-G., Liu, Q., and Afshari, S. (2006). An object-oriented hierarchical patch dynamics paradigm (HPDP) for modeling complex groundwater systems across multiple-scales. *Environ. Model. Softw.* 21 (5), 744–749. doi:10.1016/j.envsoft.2005.11.001

Li, W., and Zhang, C. (2019). Markov chain random fields in the perspective of spatial Bayesian networks and optimal neighborhoods for simulation of categorical fields. *Comput. Geosci.* 23, 1087–1106. doi:10.1007/s10596-019-09874-z

Linde, N., Renard, P., Mukerji, T., and Caers, J. (2015). Geological realism in hydrogeological and geophysical inverse modeling: A review. *Adv. Water Resour.* 86, 86–101. doi:10.1016/j.advwatres.2015.09.019

Lunt, I. A., and Bridge, J. S. (2007). Formation and preservation of open-framework gravel strata in unidirectional flows. *Sedimentology* 54 (1), 71–87. doi:10.1111/j.1365-3091.2006.00829.x

Lunt, I. A., Bridge, J. S., and Tye, R. S. (2004). "Development of a 3D depositional model of braided river deposits,". in "*Aquifer characterization*" Editors J. S. Bridge and D. W. Hyndman (Society for Sedimentary Geology, Special Publications), 80, 39–169. doi:10.2110/pec.04.80.0139

Massmann, J. W. (1989). Applying groundwater flow models in vapor extraction system design. *J. Environ. Eng. New. York.* 115, 129–149. doi:10.1061/(asce)0733-9372(1989)115:1(129)

Oliveira, G. S., Soares, A. O., Schiozer, D. J., and Maschio, C. (2017). Reducing uncertainty in reservoir parameters combining history matching and conditioned geostatistical realizations. *J. Petroleum Sci. Eng.* 156, 75–90. doi:10.1016/j.petrol.2017.05.003

Pang, L. (2009). Microbial removal rates in subsurface media estimated from published studies of field experiments and large intact soil cores. *J. Environ. Qual.* 38 (4), 1531–1559. doi:10.2134/jeq2008.0379

Park, Y. J., Sudicky, E. A., McLaren, R. G., and Sykes, J. F. (2004). Analysis of hydraulic and tracer response tests within moderately fractured rock based on a transition probability geostatistical approach. *Water Resour. Res.* 40 (12). doi:10.1029/2004WR003188

Pyrcz, M., Boisvert, J., and Deutsch, C. (2009). Alluvsim: A program for event-based stochastic modeling of fluvial depositional systems. *Comput. Geosciences* 35 (8), 1671–1685. doi:10.1016/j.cageo.2008.09.012

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A., and Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Adv. Water Resour.* 36, 36–50. doi:10.1016/j.advwatres.2011.04.006

Renard, P., and Allard, D. (2013). Connectivity metrics for subsurface flow and transport. *Adv. Water Resour.* 51, 168–196. doi:10.1016/j.advwatres.2011.12.001

Ritzi, R. W., and Soltanian, M. R. (2015). What have we learned from deterministic geostatistics at highly resolved field sites, as relevant to mass transport processes in sedimentary aquifers? *J. Hydrology* 531 (1), 31–39. doi:10.1016/j.jhydrol.2015.07.049

Riva, M., Gaudagnini, L., Gaudagnini, A., Ptak, T., Martac, E., and GuAdAgnini, A. (2006). Probabilistic study of well capture zones distribution at the Lauswiesen field site. *J. Contam. Hydrol.* 88, 92–118. doi:10.1016/j.jconhyd.2006.06.005

Riva, M., Guadagnini, A., Fernandez-Garcia, D., Sanchez-Vila, X., and Ptak, T. (2008). Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the Lauswiesen site. *J. Contam. Hydrol.* 101, 1–13. doi:10.1016/j.jconhyd.2008.07.004

Riva, M., Neuman, S. P., and Gaudagnini, A. (2015). New scaling model for variables and increments with heavy-tailed distributions. *Water Resour. Res.* 51 (6), 4623–4634. doi:10.1002/2015WR016998

Rossi, P., De Carvaldho-Dill, A., Muller, I., and Aragno, M. (1994). Comparative tracing experiments in a porous aquifer using bacteriophages and fluorescent dye on a test field located at Wilerwald (Switzerland) and simultaneously surveyed in detail on a local scale by radio-magneto-tellury (12–240 kHz). *Environ. Geol.* 23, 192–200. doi:10.1007/BF00771788

Rubbab, Q., Mirza, I. A., and Qureshi, M. Z. A. (2016). Analytical solutions to the fractional advection-diffusion equation with time-dependent pulses on the boundary. *AIP Adv.* 6, 075318. doi:10.1063/1.4960108

Sanchez-Vila, X., and Fernandez-Garcia, D. (2016). Debates—stochastic subsurface hydrology from theory to practice: Why stochastic modeling has not yet permeated into practitioners? *Water Resour. Res.* 52 (12), 9246–9258. doi:10.1002/2016WR019302

Scheibe, T. D., Yang, X., Chen, X., and Hammond, G. (2015). A hybrid multiscale framework for subsurface flow and transport simulations. *Procedia Comput. Sci.* 51, 1098–1107. doi:10.1016/j.procs.2015.05.276

Siena, M., and Riva, M. (2020). Impact of geostatistical reconstruction approaches on model calibration for flow in highly heterogeneous aquifers. *Stoch. Environ. Res. Risk Assess.* 34, 1591–1606. doi:10.1007/s00477-020-01865-2

Soltanian, M. R., and Ritzi, R. W. (2014). A new method for analysis of variance of the hydraulic and reactive attributes of aquifers as linked to hierarchical and multiscaled sedimentary architecture. *Water Resour. Res.* 50, 9766–9776. doi:10.1002/2014WR015468

Soltanian, M. R., Ritzi, R. W., Dai, Z., and Huang, C. C. (2015). Reactive solute transport in physically and chemically heterogeneous porous media with multimodal reactive mineral facies: The Lagrangian approach. *Chemosphere* 122, 235–244. doi:10.1016/j.chemosphere.2014.11.064

Sreekanth, J., and Moore, C. (2018). Novel Patch Modelling method for efficient simulation and prediction uncertainty analysis of multi-scale groundwater flow and transport processes. *J. Hydrology* 559, 122–135. doi:10.1016/j.jhydrol.2018.02.028

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.* 34 (1), 21. doi:10.1023/A:1014009426274

Theel, M., Huggenberger, P., and Zosseder, K. (2020). Assessment of the heterogeneity of hydraulic properties in gravelly outwash plains: A regionally scaled sedimentological analysis in the munich gravel plain, Germany. *Hydrogeol. J.* 28, 2657–2674. doi:10.1007/s10040-020-02205-y

Tonkin, M., and Doherty, J. (2009). Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques. *Water Resour. Res.* 45, W00B10. doi:10.1029/2007WR006678

USEPA (1995). Innovative site remediation technology – vapour extraction. https://nepis.epa.gov/Exe/ZyPDF.cgi/2000ITX8.PDF?Dockey=2000ITX8.PDF downloaded.

Webb, E. K., and Anderson, M. P. (1996). Simulation of preferential flow in three-dimensional, heterogeneous conductivity fields with realistic internal architecture. *Water Resour. Res.* 32 (3), 533–545. doi:10.1029/95WR03399

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50 (2), 1152–1173. doi:10.1002/2013WR014767

White, J. T. (2017). Forecast first: An argument for groundwater modeling in reverse. *Groundwater* 55, 660–664. doi:10.1111/gwat.12558

White, P. A. (2001). "Groundwater resources in New Zealand," in *Groundwaters of New Zealand*. Editors M. R. Rosen and P. A. White (New ZealandWellington: Hydrological Society Inc.), 45–75.

Zhou, H., Li, L., and Gómez-Hernández, J. (2010). Three-dimensional hydraulic conductivity upscaling in groundwater modeling. *Comput. Geosciences* 36 (10), 1224–1235. doi:10.1016/j.cageo.2010.03.008

Zhu, L., Dai, Z., Gong, H., Gable, C., and Teatini, P. (2016). Statistic inversion of multi-zone transition probability models for aquifer characterization in alluvial fans. *Stoch. Environ. Res. Risk Assess.* 30, 1005–1016. doi:10.1007/s00477-015-1089-2

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/feart.2022.1026479

OPEN ACCESS

EDITED BY
Jeremy White,
Intera, Inc., United States

REVIEWED BY
Guoqiang Tang,
University of Saskatchewan, Canada
Zhongfan Zhu,
Beijing Normal University, China
Ayman Alzraiee,
California Water Science Center (USGS), United States

*CORRESPONDENCE
Maruti K. Mudunuru,
maruti@pnnl.gov

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 23 August 2022
ACCEPTED 03 November 2022
PUBLISHED 24 November 2022

CITATION
Mudunuru MK, Son K, Jiang P, Hammond G and Chen X (2022), Scalable deep learning for watershed model calibration.
Front. Earth Sci. 10:1026479.
doi: 10.3389/feart.2022.1026479

Scalable deep learning for watershed model calibration

Maruti K. Mudunuru*, Kyongho Son, Peishi Jiang, Glenn Hammond and Xingyuan Chen

Pacific Northwest National Laboratory, Richland, WA, United States

Watershed models such as the Soil and Water Assessment Tool (SWAT) consist of high-dimensional physical and empirical parameters. These parameters often need to be estimated/calibrated through inverse modeling to produce reliable predictions on hydrological fluxes and states. Existing parameter estimation methods can be time consuming, inefficient, and computationally expensive for high-dimensional problems. In this paper, we present an accurate and robust method to calibrate the SWAT model (i.e., 20 parameters) using scalable deep learning (DL). We developed inverse models based on convolutional neural networks (CNN) to assimilate observed streamflow data and estimate the SWAT model parameters. Scalable hyperparameter tuning is performed using high-performance computing resources to identify the top 50 optimal neural network architectures. We used ensemble SWAT simulations to train, validate, and test the CNN models. We estimated the parameters of the SWAT model using observed streamflow data and assessed the impact of measurement errors on SWAT model calibration. We tested and validated the proposed scalable DL methodology on the American River Watershed, located in the Pacific Northwest-based Yakima River basin. Our results show that the CNN-based calibration is better than two popular parameter estimation methods (i.e., the generalized likelihood uncertainty estimation [GLUE] and the dynamically dimensioned search [DDS], which is a global optimization algorithm). For the set of parameters that are sensitive to the observations, our proposed method yields narrower ranges than the GLUE method but broader ranges than values produced using the DDS method within the sampling range even under high relative observational errors. The SWAT model calibration performance using the CNNs, GLUE, and DDS methods are compared using R^2 and a set of efficiency metrics, including Nash-Sutcliffe, logarithmic Nash-Sutcliffe, Kling-Gupta, modified Kling-Gupta, and non-parametric Kling-Gupta scores, computed on the observed and simulated watershed responses. The best CNN-based calibrated set has scores of 0.71, 0.75, 0.85, 0.85, 0.86, and 0.91. The best DDS-based calibrated set has scores of 0.62, 0.69, 0.8, 0.77, 0.79, and 0.82. The best GLUE-based calibrated set has scores of 0.56, 0.58, 0.71, 0.7, 0.71, and 0.8. The scores above show that the CNN-based calibration leads to more accurate low and high streamflow predictions than the GLUE and DDS sets. Our research demonstrates that the proposed method has high potential to improve our current practice in calibrating large-scale integrated hydrologic models.

KEYWORDS

SWAT calibration, watershed modeling, parameter estimation, inverse problems, convolutional neural networks, scalable deep learning

1 Highlights

- We developed a scalable deep learning (DL) methodology to estimate SWAT model parameters.
- Our DL methodology is based on convolutional neural networks (CNN).
- Our CNN-enabled SWAT model calibration shows higher streamflow prediction accuracy than traditional parameter estimation methods such as the Generalized Likelihood Uncertainty Estimation (GLUE) and the Dynamically Dimensioned Search (DDS) algorithms.
- Estimated SWAT model parameters from observed discharges are within the sampling range of ensemble simulations even when high-observational errors exist.
- An added benefit is that CNN-enabled parameter estimation after training is at least $\mathcal{O}(10^3)$ times faster than GLUE- and DDS-based methods.
- However, the hyperparameter tuning to discover reasonably accurate CNN models is computationally expensive, which is in $\mathcal{O}(10^5)$ processor hours.

2 Introduction

Watershed models frequently are used to predict streamflow and other components in the terrestrial water cycle. These components are affected by a wide range of anthropogenic activities (e.g., agricultural intensification), climate perturbations (e.g., rain-on-snow, rising temperatures and increasing precipitation, earlier occurrence of snow melt in mountainous regions), and disturbances (e.g., wildfire) (Singh and Frevert, 2003; Singh and Frevert, 2010; Daniel et al., 2011). Watershed models also have been used to assess the sustainability of the water supply for effective water resource management. Some popular and open-source watershed modeling software that can accurately simulate various components of water cycling in intensively managed watersheds include the Soil and Water Assessment Tool (SWAT) and its variants (e.g., SWAT-MRMT-R) (Mankin et al., 2010; Neitsch et al., 2011; Fang et al., 2020), the Advanced Terrestrial Simulator (ATS) (Coon et al., 2020), the Precipitation Runoff Modeling System (PRMS) (Leavesley et al., 1983; Markstrom et al., 2015), the Weather Research and Forecasting Model Hydrological modeling system (WRF-Hydro) (Sampson and Gochis, 2018; Wu et al., 2021), etc (Donigian et al., 1995; Tague and Band, 2004; Graham and Butts, 2005; Cuo et al., 2008; Hamman et al., 2018).

Watershed models adopt physical laws (e.g., mass and energy balance) or known empirical relationships to simulate the watersheds' different hydrological components (e.g., infiltration, evapotranspiration, groundwater flow, streamflow). These models feature two types of parameters (Johnston and Pilgrim, 1976; Mein and Brown, 1978; Nakshatrala and Joshaghani, 2019). The first type includes parameters with physical characteristics (e.g., permeability, porosity). The second type includes conceptual or empirical parameters, which are currently impossible or difficult to measure directly. Most watershed simulators (e.g., SWAT, PRMS) consist of parameters that fall in the second category (Singh and Frevert, 2010). As a result, observed data, such as streamflow collected at the watershed outlet, are used to estimate the conceptual parameters through model calibration. Many semi-distributed or bucket models can only achieve adequately accurate predictions after calibrating their parameters with available observations, making them less ideal for ungauged watersheds. On the other hand, advanced fully integrated watershed models (e.g., ATS) can predict watershed responses with reasonable accuracy without undergoing intensive model calibration; however, running those models is computationally expensive (Chen et al., 2021; Cromwell et al., 2021). Certain parameters in these mechanistic models (e.g., ATS) are measurable and physically significant while others are empirically similar to the SWAT.

Various techniques and software tools for calibrating watershed models have been reported in the literature (Duan et al., 2004). Popular methods include generalized likelihood uncertainty estimation (GLUE) (Blasone et al., 2008; Nott et al., 2012), the dynamically dimensioned search (DDS), maximum likelihood estimation (Myung, 2003), the shuffled complex evolution method developed at the University of Arizona (SCE-UA) (Duan et al., 1994), Bayesian parameter estimation methods (Thiemann et al., 2001; Gupta et al., 2003; Misirli et al., 2003), ensemble-based data assimilation methods (e.g., ensemble Kalman filter, ensemble smoother) (Evensen, 1994; Van Leeuwen and Evensen, 1996; Evensen, 2003; Chen et al., 2013; Evensen, 2018; Jiang et al., 2021), and adjoint-based methods (Tarantola, 2005; Aster et al., 2018). These techniques underpin popular software packages such as PEST (Doherty and Hunt, 2010), DAKOTA (Adams et al., 2009), SWAT-CUP (Abbaspour, 2013), MATK (Model Analysis ToolKit, 2021), MADS (MADS, 2021), and DART (Anderson et al., 2009), which are developed to facilitate model calibration. Using these existing calibration methods and tools can be time consuming (e.g., slow convergence), require good initial guesses, and can be computationally intensive (e.g., may require many forward model runs or runs using high-performance computing clusters) (Rouholahnejad et al., 2012; Zhang et al., 2016; Bacu et al., 2017). Moreover, calibration using such tools can potentially result in reduced accuracy when estimating high-

dimensional parameters (> 10) (Duan et al., 2004; Eckhardt et al., 2005). New PEST tools have been developed to handle high dimensional inverse modeling like PESTPP-ies and PESTPP-DA. However, many of the methods mentioned above have challenges (see Supplementary Text S1) in properly capturing the strong nonlinear relationships between parameters and observed responses (Franco and Bonumá, 2017). Recent advances in deep learning (DL) (e.g., deep neural networks [DNNs], convolutional neural networks [CNNs]) show promise for developing reliable model calibration methods that overcome the challenges described above (Gabrielli et al., 2017; Cromwell et al., 2021).

Deep learning shows promise in aiding inverse modeling associated with highly nonlinear relationships (Zhang et al., 2009; Gabrielli et al., 2017; Marçais and de Dreuzy, 2017; Afzaal et al., 2020; Sit et al., 2020; Nearing et al., 2021). It uses multiple neural layers to extract features that are representative of inputs, and DL-enabled inverse models for parameter estimation are known to be robust even when observed errors or noise exist (Rolnick et al., 2017; Edwards, 2018; Gupta and Gupta, 2019; Rudi et al., 2020). In hydrology, neural networks (e.g., deep, convolutional, recurrent) have been used to model and predict streamflow, water quality, and precipitation (Shen, 2018; Khandelwal et al., 2020; Bhasme et al., 2021). Recently Tsai and co-workers (Tsai et al., 2021) developed a novel differentiable parameter learning framework that efficiently learns a global mapping between inputs and process model parameters. They applied this framework to estimate Variable Infiltration Capacity (VIC) land surface hydrologic model. The trained DL models produced parameters which allow VIC to best match surface soil moisture observations from NASA's Soil Moisture Active Passive satellite mission. In this paper, we present a scalable, DL methodology that uses observed streamflow data to estimate high-dimensional SWAT model parameters efficiently and reliably with reasonable accuracy. By scalable, we mean that the CNNs can be trained and tuned at any scale (e.g., from laptop computers to high-performance computers at leadership-class computing facilities) without any changes in the proposed method or developed code. This study uses CNNs, which are frequently used in hydrological applications (Sadeghi et al., 2019; Van et al., 2020; Jagtap et al., 2021).

CNNs offer many advantages over DNNs (Read et al., 2019; Dagon et al., 2020; Jia et al., 2021; Rahmani et al., 2021; Willard et al., 2022). A significant advantage of CNNs is that they explicitly learn local representations (or patterns). As a result, CNNs are best suited to produce image or time series data where the neighboring dependencies are important. This superior performance of CNNs can be attributed to the multiple convolutional layers that learn hierarchical patterns from the inputs. The resulting broader set of abstract patterns are used to develop nonlinear mappings between streamflow and the SWAT model parameters. Another benefit of CNN-enabled inverse models is their low inference time for parameter estimation compared to traditional methods; however, data requirements and associated training time (e.g., hyperparameter tuning) needed to develop such inverse models can be substantial. Once the CNN-enabled inverse model is trained, it can allow assimilation of observed data, thereby significantly reducing the time required to estimate parameters in high-dimensional space (Cromwell et al., 2021).

2.1 Main contributions

The main contribution of this study is development of an accurate parameter estimation methodology using CNNs that calibrates watershed models better than traditional methods (e.g., GLUE, DDS). The CNN-enabled inverse mappings are built on ensemble simulations generated by the SWAT model. Scalable hyperparameter tuning is performed to identify the top 50 architectures based on mean squared error and other performance metrics[1]. Further, we test the influence of errors in observed streamflow on parameter estimation and streamflow prediction accuracy. A significant advantage of the proposed DL method is that it estimates sensitive parameters with reasonably good accuracy even at high observation error levels (e.g., 100% relative observational errors). Moreover, these estimated parameters are within the prior sampling range, showing the proposed methodology's robustness to observational errors. Compared to the GLUE and DDS optimization methods, parameters estimated by the CNN-enabled inverse model provide more accurate streamflow predictions within and beyond the calibration period. The GLUE method identified a set of behavioral parameters within the ensemble parameter combinations. By "behavioral" parameters, we mean to signify parameter sets for which SWAT model simulations are deemed to be "acceptable" upon satisfying certain user-defined performance metrics (e.g., KGE greater than 0.5) on observed data (Blasone et al., 2008). Based on a cutoff threshold that uses metrics such as the KGE, the entire set of simulations then is split into behavioral and non-behavioral parameter combinations. The behavioral parameter set provides better accurate predictions than the non-behavioral set. Our analysis also showed that the CNN estimated parameter sets are narrower than the GLUE-based behavioral sets but wider than estimations obtained using the DDS method. As the DDS method is a global optimization, it searches for a best parameter value based on a performance metric (e.g., KGE). Hence, the obtained parameter ranges from the DDS method can be narrower than those

1 Popular objective functions such as R^2-score, Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE), and their modifications (e.g., logNSE, mKGE, and npKGE) are used to evaluate the fit between observed and simulated streamflow time series.

obtained from the CNN and GLUE methods. Another advantage of the proposed CNN-based inverse models is that it is at least $\mathcal{O}(10^3)$ times faster than the GLUE and DDS methods. From a computational cost perspective, traditional parameter estimation using local and global optimization algorithms (e.g., using PEST, DAKOTA) requires multiple forward model runs. As a result, inverse modeling may require source code modifications and also high-performance computational resources, which can be prohibitively expensive. We acknowledge that hyperparameter tuning can be expensive. However, such tuning is needed for finding optimal CNN architectures. Once CNNs are trained, the savings in computational cost enable our DL-enabled parameter estimation to be inclusive [i.e., easy to adapt using transfer learning (Zhuang et al., 2020; Song and Tartakovsky, 2021)] and ideal for calibrating multi-fidelity models (e.g., ATS, PFLOTRAN, WRF-Hydro, PRMS) at spatial scales of watersheds and basins.

2.2 Outline of the paper

The paper is organized as follows: Section 2 discusses state-of-the-art methods for parameter estimation and their limitations. We also demonstrate the need for developing DL method to better calibrate hydrological models, such as SWAT. Section 3 describes the study site and SWAT model developed using a National Hydrography Dataset PLUS (NHDPLUS v2)-based watershed delineation (Moore and Dewald, 2016). We discuss data generation to develop CNN-enabled inverse models. We also compared observed data with the SWAT model ensemble simulations. Section 4 introduces the proposed scalable DL methodology for estimating SWAT parameters. We performed sensitivity analysis to rank the sensitivities of SWAT model parameters. We performed scalable hyperparameter tuning to identify the optimal CNN architectures and described the associated computational costs for training the DL models and generating inferences (e.g., on test and observational data). Section 5 presents the training, validation, and testing results of the CNNs. We compared the performance of CNN-estimated parameters with those of the GLUE and DDS methods. Performance of the calibration model within and beyond calibration period is provided. Sections 6 and 7 present our future work and conclusion.

3 Study site and data generation

This section first describes the study site, the American River Watershed (ARW) in the Yakima River Basin (YRB), before discussing the SWAT model, its parameters, and specifics on ensemble runs needed to develop CNN-enabled inverse models. We also compare the observed streamflow data used to calibrate the SWAT model with the ensemble runs within the calibration period (i.e., from water years [WYs] 2014 to 2016 [1 October 2013, through 30 September 2016]). Each SWAT model run produces daily simulated streamflow values.

3.1 Study site

The YRB (see Figure 1), situated in Eastern Washington State, has a drainage area of about 16,057 km^2 (Mastin and Vaccaro, 2002; Qiu et al., 2019). The daily averaged flow for the YRB is about 95 m^3s^{-1} over a period of 40 years. This averaged flow is computed using the data collected from 1/1/1980 to 12/31/2021 at the Kiona gauge station, which is the closest to the outlet of the YRB. A major tributary of the Yakima River is the American River, which is a third-order stream, with a watershed of about 205 km^2. According to 30-year normalized PRISM data, the mean annual precipitation and temperature within the ARW range from 978 to 2,164 mm and 2.8–4.9°C, respectively (Daly et al., 2000; Daly and Bryant, 2013; PRISM, 2021). The climate within the ARW exhibits strong seasonal patterns, including cold, wet winters and hot, dry summers. About 60% of precipitation occurs in the winter as snow, with snowmelt occurring from April to June the following year. Peak snow accumulation and flow occur in April and May, respectively. This prior site-specific knowledge shows that the snow process parameters in the SWAT model are essential. Guided by the information mentioned above and sensitivity analysis, our results demonstrate that we can better estimate such important process model parameters using our DL method rather than DDS and GLUE methods.

The slope of the ARW varies from 0° to 83°, with a mean slope of 23°. The major surface geology types are andesitev (72%), granodiorite (20%), and alluvium (8%). The primary soil texture is gravelly loamy sand with a maximum soil depth of 1,524 mm based on U.S. Department of Agriculture State Soil Geographic Data (STATSGO) (Schwarz and Alexander, 1995). This soil is classified as hydrologic group B with moderate runoff potential and infiltration rates. Evergreen trees (83%) and shrub (11%) dominate the land cover. Other types of land cover include urban, grass, and wetlands. The ARW has a U.S. Geological Survey (USGS) gauging station (USGS 12488500) located in the watershed outlet. This station has been recording the daily observed streamflow from 16 July 1988, to the present. A snow telemetry (SNOTEL) station (site name: Morse lake) is located northwest of the watershed. This SNOTEL station has measured the snow water, daily precipitation, and air maximum/mean/minimum temperatures from 1 October1979, to the present.

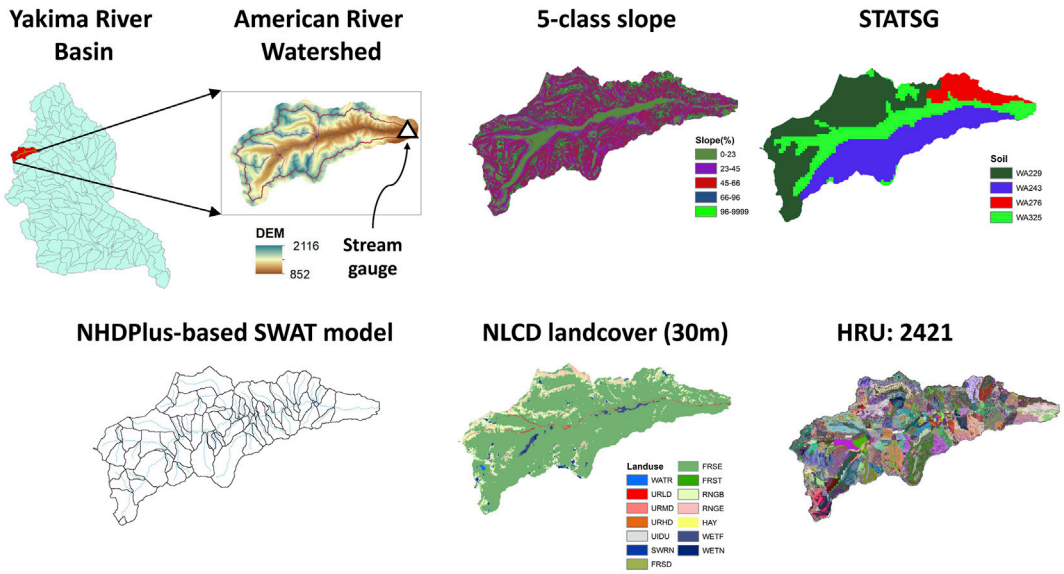

FIGURE 1
NHDPlus-based SWAT model for ARW: This figure shows the watershed delineation and data products used to develop the SWAT model for our study site. The top left figures show the delineation of the YRB into different watersheds, including our ARW study site. It also shows the DEM used for modeling the ARW and the associated stream gauge at the watershed outlet. The top right figures show the spatially varying slope and soil data within the study site. The bottom left figure shows the delineation of third-order streams in the ARW. The bottom right figures show the spatially varying land cover (primarily evergreen) and the number of HRU employed in the SWAT model.

3.2 Brief description of the SWAT model

SWAT is a semi-distributed eco-hydrological model. It can simulate both subsurface and surface hydrological processes, soil or plant bio-geochemistry, and in-stream processes (Arnold et al., 2012). The SWAT model requires various spatial Geographic Information System data to represent the different watershed characteristics (e.g., topography, land cover, and soil). The USGS 10-m digital elevation model (DEM) is used to compute the topographic parameters (e.g., drainage area, slope, slope length) with the ARW basin and sub-basin boundaries and stream networks defined by the National Hydrography Dataset Plus (NHDPlus) catchment/streams. Previous studies (Chiang and Yuan, 2015; Moore and Dewald, 2016) have demonstrated that NHDPlus-based catchment/streams outperformed modeled streamflows that did not account for such delineation.

Figure 1 shows the location of the ARW, and key inputs of NHDPlus-based SWAT model used to simulate streamflow at the ARW study site. This model is composed of 87 sub-basins with five slope classes (percent rise of the slope): 1) 0–26, 2) 26–51, 3) 51–80, 4) 80–129, and 5) 129–999. Data from the USGS National Landcover Database 2016 (30-m resolution) and the Department of Agriculture STATSGO database are used to estimate the land cover/use and soil parameters, respectively. Hydrologic response unit (HRU) maps are developed by fully combining unique slope class, land cover/use, and soil type, resulting in a total of 2,421 HRUs for the ARW (see Figure 1). Supplementary Text S1 provides additional details on SWAT model development for our study site. Daily precipitation, maximum and minimum air temperatures, radiation, and relative humidity from a daily Daymet (Daymet, 2021) with 1-km spatial resolution are used to prepare the climate input data for the SWAT model simulations. Wind speed data are generated using weather generators in the SWAT.

3.3 Data for the SWAT model calibration

Table 1 summarizes the 20 parameters and their associated sample ranges (e.g., minimum and maximum values) we calibrated in the SWAT model to generate simulation data. The table clearly shows seven groups/types of SWAT model parameters: 1) landscape, 2) soil, 3) groundwater, 4) channel, 5) snow, 6) plant, and 7) climate. Each parameter in a specified group is calibrated at different spatial scales. For example, snow group parameters such as SFTMP and SMTMP represent basin-scale snow processes. Channel group parameters are at the sub-basin level, and soil/groundwater/plant/climate group parameters represent HRU level spatial variation. Even though some parameters differ at the HRU level, we calibrate basin-scale scaling coefficients that vary within [−0.3, 0.3] and are the same for all HRUs.

TABLE 1 This table provides a list of 20 different SWAT model parameters that are calibrated using the proposed scalable DL methodology. The associated lower and upper limits of parameter values also are specified. Boldfaced descriptors are mutual information (MI)-identified sensitive parameters (Jiang et al., 2022b).

Parameter group/type	Parameter[2]	Lower limit	Upper limit	Brief description (units)	Parameter modification[3]	Spatial variability
Landscape	**CN2**	−0.3	0.3	% change in SCS runoff curve number	R	Varying across HRUs
Groundwater	**RCHRG_DP**	0	1	Deep aquifer percolation fraction	V	Constant across HRUs
Groundwater	GWQMN	0	5,000	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	V	Constant across HRUs
Groundwater	GW_REVAP	0	0.2	Groundwater "revap" coefficient	V	Constant across HRUs
Groundwater	REVAPMN	1	500	Threshold depth of water in the shallow aquifer for "revap" to occur (mm)	V	Constant across HRUs
Groundwater	**GW_DELAY**	1	100	Groundwater delay (days)	V	Constant across HRUs
Groundwater	**ALPHA_BF**	0.01	0.99	Baseflow alpha factor	V	Constant across HRUs
Soil	SOL_K	−0.3	0.3	% change in saturated hydraulic conductivity (mm h^{-1})	R	Varying across HRUs
Soil	SOL_AWC	−0.3	0.3	% change in available water change in capacity of the soil layer (mm H_2O mm $soil^{-1}$)	R	Varying across HRUs
Soil	ESCO	0.01	1	Soil evaporation compensation factor	V	Constant across HRUs
Soil	OV_N	−0.3	0.3	% change in Manning's "n" value for overland flow	R	Varying across HRUs
Channel	**CH_K2**	0	200	Effective hydraulic conductivity in main channel alluvium (mm h^{-1})	V	Constant across sub-basins
Channel	**CH_N2**	0.02	0.15	Manning's "n" value for the main channel	V	Constant across sub-basins
Snow	**SFTMP**	−5	5	Snowfall temperature (°C)	V	Constant in the basin
Snow	**SMTMP**	−5	5	Snow melt base temperature (°C)	V	Constant in the basin
Snow	**SMFMX**	1.4	6.9	Maximum melt rate for snow during the year (mm H_2O°C day^{-1})	V	Constant in the basin
Snow	**TIMP**	0.01	1	Snowpack temperature lag factor	V	Constant in the basin
Plant	**EPCO**	0.01	1	Plant uptake compensation factor	V	Constant in the basin
Climate	PLAPS	343.3	964	Precipitation lapse rate (mm km^{-1})	V	Constant in the basin
Climate	TLAPS	-4.86	3.353	Temperature lapse rate (°C km^{-1})	V	Constant in the basin

Using a Sobol quasi-random[2] sequence sampling method (Herman and Usher, 2017), we generated 1,000 sets of these 20 parameters to develop CNN-enabled inverse models. Sobol sequence is quasi-random low-discrepancy sequences (Sobol', 1967; Herman and Usher, 2017). Compared with random sampling from a uniform distribution, Sobol sequence guarantee better uniform coverage of the samples. We adopted Sobol sequences to generate the ensemble realizations of standardized parameters within [0,1], which were then scaled back to the parameter ranges shown in Table 1 The daily streamflow data and flow duration curves simulated using the SWAT model for these 1,000 realizations are shown in Figure 2. The simulation time for the SWAT model calibration is between the beginning of WY 2014 to the end of WY 2016 (i.e., 1 October 2013—30 September 2016), which is referred to as the calibration

2 Table 1: Note that the sensitive parameters are identified using the MI method. These sensitive parameters are presented in boldface in this parameter column.

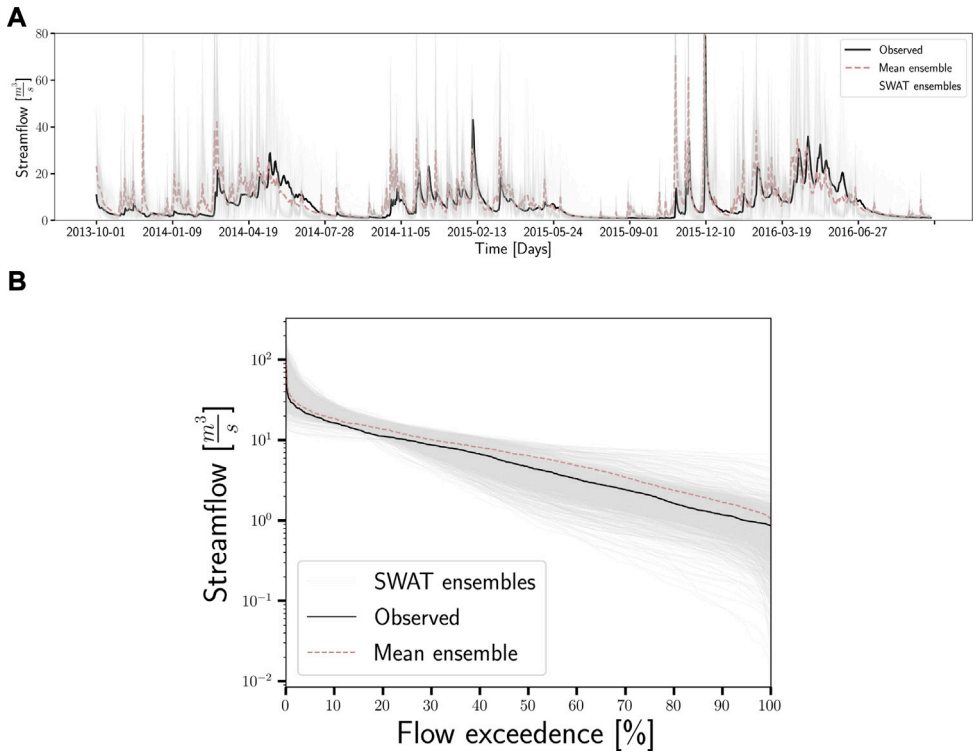

FIGURE 2
SWAT model simulations vs. observational data (within calibration period): This figure compares the modeled streamflow data generated based on NHDPlus-based SWAT model with observed flow for the ARW study site. The top figure **(A)** shows the streamflow time-series and the bottom figure **(B)** shows the flow duration curves. The dark brown color dashed line represents the ensemble mean of 1000 SWAT model simulations. The grey color lines represents the modeled streamflow ensembles. The black-colored line corresponds to the observed streamflow data.

period. The validation period is from[3] WY 2000 through WY 2013 (i.e., 1 October 1999—30 September 2013). The calibrated SWAT model is run during the validation period, and its performance is then compared with the observed data. Figure 2 compares the ensemble mean of simulated discharge (i.e., 1,000 realizations) with the observational data. The grey color represents each of the 1,000 simulated discharge realizations. This streamflow time-series and flow duration curve qualitatively shows the similarities of the trends in the simulated discharge and observed data. However, the comparison against observations also show the over/under predictions of peak/low flows that can be due to structural deficiencies of the model. The SWAT model fidelity may need to be enhanced to overcome these structural deficiencies. The generated data are used to estimate SWAT parameters by the CNN-based calibration, GLUE, and DDS methods. The GLUE- and DDS-based SWAT model calibrations also are compared with the observed data for both periods. The behavioral model parameter sets (i.e., from the GLUE method) are selected based on KGE metrics. We also use the other accuracy measures (e.g., NSE, logNSE, R^2-score) to evaluate the calibrated SWAT model performance, which are described in Section 4.4.

4 Proposed methodology

This section presents the overall methodology consisting of data pre-processing, scalable hyperparameter tuning (Mudunuru et al., 2022), and computational cost of constructing the CNN-enabled inverse models. We also briefly describe the GLUE and DDS optimization methods that are used to compare the performance of CNNs. The comparison of the DL method performance against the most commonly applied algorithms for calibration of

3 In Table 1, the parameter modification column indicates how SWAT model parameters are modified during calibration and the training data generation for CNN-enabled inverse modeling. The term "V" indicates that existing SWAT model parameter values are replaced with values in the provided range. The term "R" indicates relative changes in parameters by multiplying existing values with 1+ calibrated parameter values in the range (Qiu et al., 2019). The CN2, SOL_K, and SOL_AWC parameter modifications are "R," whose absolute values as (Eckhardt et al., 2005; Rouholahnejad et al., 2012), [0.001, 1,000], and [0.01, 0.35], respectively.

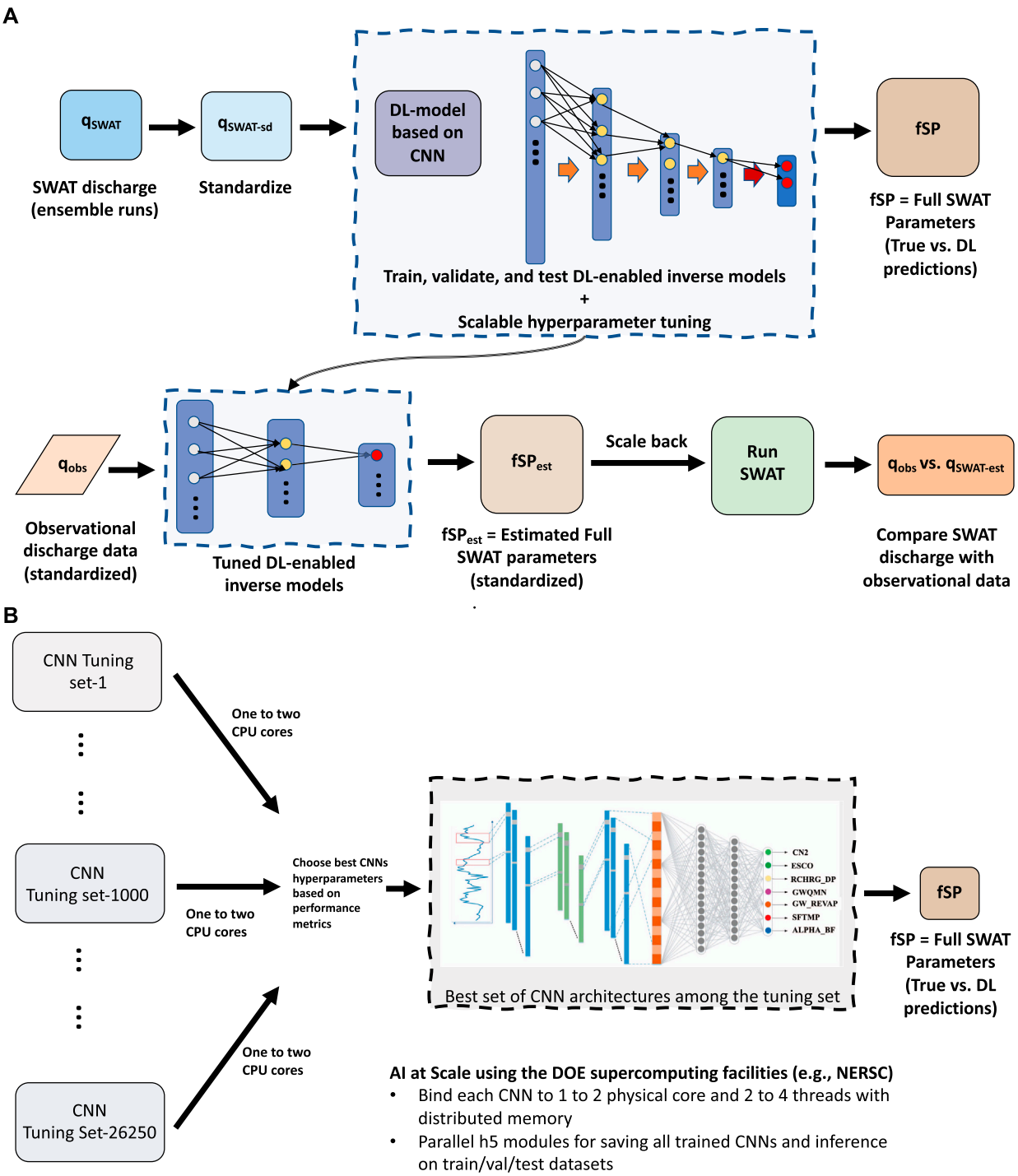

FIGURE 3
Proposed scalable deep learning workflow for the SWAT model calibration: A pictorial description of the proposed DL methodology to estimate parameters and calibrate the SWAT model using observational discharge. Ensemble simulations generated by the SWAT model are used to train, validate, and test the CNN-enabled inverse models, as shown in the top figure **(A)**. The observed streamflow is then provided as an input to the developed DL models to estimate site-specific parameters. These parameters are then used by the SWAT model to simulate discharge for comparison with observational data. The bottom figure **(B)** shows a scalable hyperparameter tuning approach to identify optimal CNN architectures using high-performance computing resources at NERSC. Each explored CNN architecture is trained on one/two CPU physical cores and its performance is estimated using validation loss and streamflow prediction metrics (e.g., R^2-score, NSE, logNSE, KGE and its variants). From the explored space, top 50 CNNs are chosen for inverse modeling and analysis.

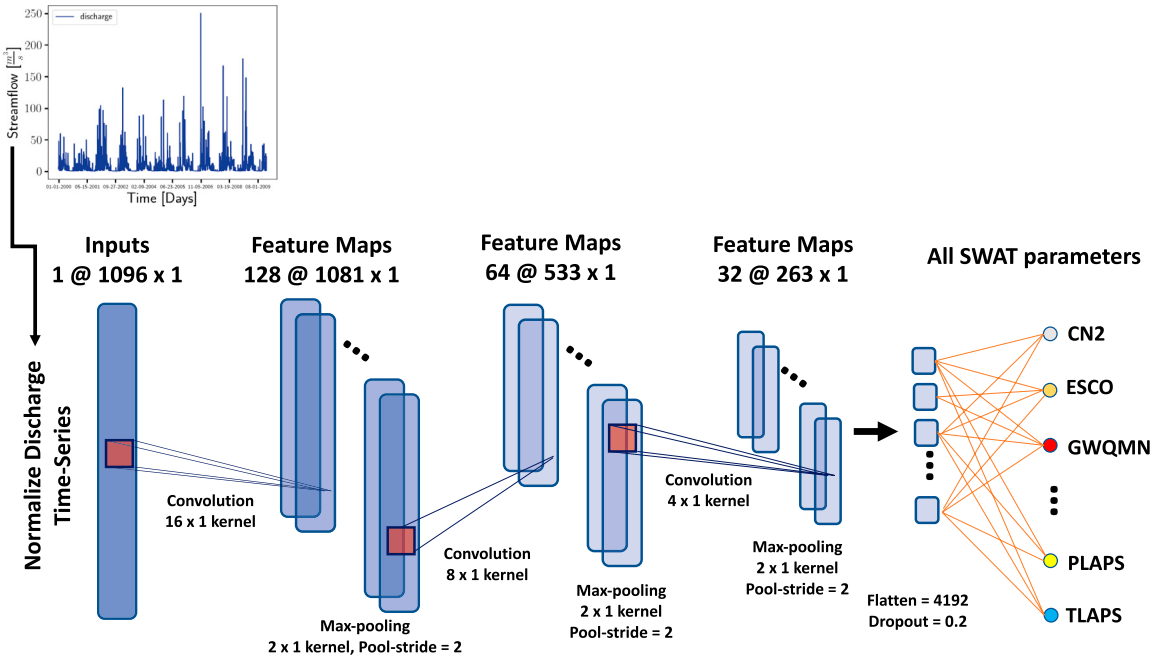

FIGURE 4
A top-50 CNN-enabled inverse model architecture: This figure shows a pictorial description of a deep CNN for estimating SWAT model parameters. This CNN model inversely maps simulated discharge to 20 different conceptual parameters. Hyperparameter tuning is then performed to arrive at this CNN architecture. This CNN-enabled inverse model is a multi-task learning model that accounts for the correlation between different SWAT parameters during training.

watershed simulation models (i.e., DDS and GLUE) gives better insight into CNN's capability in providing accurate parameter estimations and uncertainty on streamflow predictions.

4.1 Proposed scalable deep learning methodology

Figure 3 summarizes our proposed DL method for training the inverse models and then inferring the SWAT parameters. We train, validate, and test CNN-enabled inverse models (Schmidhuber, 2015; Goodfellow et al., 2016; Chollet, 2017) using SWAT model ensemble runs. The proposed DL methodology can be divided into multiple steps, which is described below in a step-by-step approach.

1) The inputs to the CNNs are the modeled daily streamflow time-series data and outputs are the SWAT parameters. Both the inputs and outputs are normalized for training CNNs.
2) The CNN-enabled inverse models are developed to estimate all 20 of the SWAT process model parameters. We used the Keras API in Tensorflow package (Keras API, 2021) to build our CNN-enabled inverse models.
3) The simulated streamflow and parameter sets are assembled into a data matrix and then partitioned into training (80%), validation (10%), and testing (10%) sets, of which each SWAT run contains 1,096 daily data points. The training and validation sets are jointly used in hyperparameter tuning to find the optimal CNN architecture.
4) The dataset is normalized, which is necessary for CNN model development as CNNs are filter/kernel-based methods that benefit from normalization of their inputs to make accurate predictions (Anysz et al., 2016; Gu et al., 2018). The normalization is done by first removing the mean and scaling the training dataset to unit variance and then applying the same pre-processing normalizer to transform the validation and testing sets.
5) Hyperparameter tuning is performed to identify the optimal CNN architectures whose performances were evaluated against validation dataset during model training. By CNN architecture, we mean convolutional and pooling layers that needs to be tuned for optimal performance.
6) The testing step includes performance evaluation (e.g., mean-squared error) of the tuned CNNs on test data.
7) The observed data are standardized using the pre-processing normalizer (Pedregosa et al., 2011) that we trained on simulation data. This normalized data is input to the tuned

TABLE 2 This table provides the hyperparameter space used to explore CNN architectures for developing reliable DL-enabled inverse mappings for the SWAT model calibration.

Hyperparameter type	Description	Explored options
Layers	Number of 1D convolutional layers	[1, 2, 3, 4, 5]
Filters	The number of output filters in the 1D convolution	[16, 32, 64, 128, 256]
Kernel size	An integer to specify the length of the 1D convolution window	[2, 4, 8, 16, 32]
Dropout rate	Applies dropout to the input[5]	[0.0, 0.1, 0.2, 0.3, 0.4]
Learning rate	The value of the optimizer in the Adam algorithm	[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]
Batch size	The number of training samples seen by CNN per gradient update	[4, 8, 16, 32, 64]
Epochs	The number of times the algorithm sees the training data	[50, 100, 200, 300, 400, 500]

CNN-enabled inverse model to estimate the study site SWAT model parameters. We also add errors to observed streamflow data and assess the performance of CNNs for SWAT model calibration.

8) Finally, these calibrated parameter sets are given to the SWAT model to obtain daily streamflow values in the calibration and validation periods. The predicted discharge is then compared with the observed data to evaluate the performance of the CNN-calibrated SWAT model in both calibration (WY 2014–2016) and validation time periods (WY 2000–2013).

Hyperparameter tuning is a crucial step in obtaining reliable and accurate CNN-enabled inverse models. The search for hyperparameters is performed in parallel at the National Energy Research Scientific Center (NERSC) (NERSC, 2021), a high performance computing user facility operated by Lawrence Berkeley National Laboratory for the U.S. Department of Energy Office of Science. Scalable hyperparameter tuning is achieved by combining mpi4py (MPI for Python) package with Tensorflow package and parallel HDF5 modules to train each CNN architecture on at least one physical central processing unit (CPU) (see Figure 3B). As tuning is embarrassingly parallel, the CNN architectural search space is distributed across the processes employed and run simultaneously on one to two cores each. All trained CNN models and their inferences are written to their individual HDF5 files. This tuning is necessary as the training process and predictions of the CNN-enabled inverse model are controlled by the parameters and topology of the CNN architecture. We tested two types of hyperparameters: 1) model hyperparameters and 2) algorithm hyperparameters. Model hyperparameters define the neural network architecture. For instance, the selection of CNN topology is influenced by model hyperparameters such as the number and width of hidden layers. Algorithm hyperparameters influence the training process after the architecture is established. The values of the trainable weights of a CNN architecture are controlled by algorithm hyperparameters such as learning rate and the number of epochs. Table 2 shows the search space that we explored. Supplementary Table S1 shows the model and algorithm hyperparameters for the top 50 CNN architectures identified through this scalable approach. During the tuning process, we used ReLU as the activation function, and max pooling is taken to be equal to 2. The optimal hyperparameter set is chosen based on the validation mean squared error along with streamflow prediction metrics using the grid search tuning method[4]. In addition to identifying the optimal hyperparameter set, we also identified the next 50 best candidates. In Section 5, we show the predictions of these 50 best models and the associated uncertainty in their streamflow predictions[5].

Figure 4 shows a pictorial description of a tuned CNN architecture from the grid search. The CNN filters are initialized with the Glorot uniform initializer (Gu et al., 2018; Keras API, 2021). This Glorot uniform allows us to initialize the weights so the variance of the activations are the same across every neural layer. Moreover, this constant variance initialization helps prevent the gradient from exploding or vanishing. After each convolution, a max-pooling operation is applied, and the final convolutional layer is flattened. After the dropout layer, the remaining features are mapped to the SWAT parameters. The entire CNN is compiled using an Adam optimizer, with the loss being the mean squared error. The resulting tuned CNN architectures (each of the top 50 models) have approximately 1 M trainable weights.

4.2 Dynamically dimensioned search method

The DDS method is a global optimization algorithm developed to automatically calibrate highly parameterized

4 Grid search is an exhaustive search technique performed on a specific hyperparameter values of the CNN architecture.

5 Table 2: To reduce model over-fitting, we randomly set the last convolutional layer units that connect to the output to 0 at each step during training time. The rate value controls the frequency of dropping the units.

hydrologic models. Typically, the total number of evaluations available for SWAT model calibration is always limited and is also case-study dependent because of the curse of dimensionality. The DDS method is designed from this calibration perspective to find practical or high-quality parameter sets. It is well known that the DDS method outperforms methods such as SCE-UA (available in PEST package) when the number of calibrated parameters is high (i.e., 10 or more) (Tolson and Shoemaker, 2007). Below, we summarize the steps involved in executing the DDS method to calibrate the SWAT model.

First, we define DDS algorithm inputs such as neighborhood perturbation size parameter (0.2 as the default value), maximum number of function evaluations (a total of 500 for each random seed), number of random seeds (a total of 10), bounds on all the SWAT model parameters (as mentioned in Table 1), and initial guesses/solutions for these parameters. Second, for the initial guess, we construct and evaluate an objective function (e.g., KGE) that minimizes differences between the simulated and observed data. Third, we perturb the initial guess by using a vector sampled from a standard normal random distribution with zero mean and unit standard deviation. We ensure that the perturbed values are within the physical bounds, which is the SWAT parameter range. Fourth, we evaluate the objective function and update the best solution until all user-defined evaluations are exhausted or a stopping criterion is met. We executed these steps for 10 different random seeds, which resulted in a total of 5,000 DDS calibration sets (i.e., 10×500). Then, we selected the top-50 from this total of 5,000 DDS calibration sets.

4.3 Generalized likelihood uncertainty estimation method

The GLUE method (Beven and Binley, 2014) used in hydrology provides a framework for evaluating model performance and quantifying the impact of various uncertainty sources on predictive uncertainty. For its simplicity and flexibility, the GLUE method (Beven and Binley, 2014) has been applied to various watershed models. The method uses a Monte Carlo approach to evaluate different model structure/parameter sets by comparing observed data with modeled values. In many cases, the different models or parameter sets show similar model performance (e.g., NSE), which is called as an equifinality. Thus, instead of searching for an optimum model, searching for a behavioral parameter and model structure is a general practice. In this study, we use the GLUE method to select the behavioral parameter sets for the SWAT model by comparing the observed streamflow and modeled value. Because of the lack of prior knowledge of the distribution of each parameter, the 20 parameters used in the SWAT model are assumed to follow uniform distributions, and we use a Sobol sequence method to efficiently sample the parameters values. The behavioral parameter sets are the top-50 sets selected from a total of 1,000 simulations based on the accuracy of the KGE metric. The selected KGE values of the behavioral parameter sets range from 0.5 to 0.7. They are shown in Section 5 and also in Supplementary Table S2. Also, to evaluate the impact of total number of model simulations on model performance, we also increased the number of model simulations from 1,000 to 5,000, and the results obtained from 5,000 simulations remain very similar to the results from 1,000 simulations.

4.4 Performance metrics

The evaluation criteria for SWAT model calibration using the CNN, DDS, and GLUE estimated sets include R^2-score, NSE, logNSE, KGE, and its variants (i.e., mKGE and npKGE) (Hydroeval, 2021). For instance, NSE, logNSE, and KGE are evaluated as follows:

$$\mathrm{NSE}(\mathbf{q},\hat{\mathbf{q}}) = 1 - \frac{\sum_{i=1}^{n}(q_i - \hat{q}_i)^2}{\sum_{i=1}^{n}(q_i - \mu_{\mathbf{q}})^2} \quad \text{where} \quad \mu_{\mathbf{q}} = \frac{1}{n}\sum_{i=1}^{n} q_i \tag{1}$$

$$\mathrm{logNSE}(\mathbf{q},\hat{\mathbf{q}}) = 1 - \frac{\sum_{i=1}^{n}(\log[q_i] - \log[\hat{q}_i])^2}{\sum_{i=1}^{n}(\log[q_i] - \log[\bar{q}])^2} \tag{2}$$

$$\mathrm{KGE}(\mathbf{q},\hat{\mathbf{q}}) = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\hat{\mathbf{q}}}}{\sigma_{\mathbf{q}}} - 1\right)^2 + \left(\frac{\mu_{\hat{\mathbf{q}}}}{\mu_{\mathbf{q}}} - 1\right)^2} \tag{3}$$

Where $\hat{q}_i \in \hat{\mathbf{q}}$ is the SWAT model prediction and $q_i \in \mathbf{q}$ is the observational streamflow. n is the dimension of $\hat{\mathbf{q}}$ and $\mathbf{q}$, which is the total number of time-steps. r is the Pearson product-moment correlation coefficient. $\sigma_{\hat{\mathbf{q}}}$ and $\sigma_{\mathbf{q}}$ are the standard deviations in the SWAT model predictions and observations, respectively. $\mu_{\hat{\mathbf{q}}}$ and $\mu_{\mathbf{q}}$ are the mean values in the SWAT model predictions and observations, respectively. The objective functions for computing mKGE and npKGE metrics are described in References (Kling et al., 2012; Pool et al., 2018).

Each metric takes into account different aspects of calibration performance (Liu, 2020). The R^2-score indicates the goodness of fit, which measures how close the streamflow predictions from the CNN-enabled calibration are to observed data. NSE evaluates how well the calibrated SWAT model predictions capture high flows. Complementary to NSE, logNSE determines the accuracy of model predictions for low flows. KGE combines these three different components of NSE (i.e., 1) correlation, 2) bias, and 3) a ratio of variances or coefficients of variation) in a more balanced way (e.g., more weight on low flows and less weight on extreme flows) to assess the SWAT model calibration. mKGE makes sure the bias and variability ratios are not cross-correlated, which otherwise may occur when (for instance the precipitation) inputs are biased. npKGE provides the variability and the correlation term in KGE in a non-parametric form. This reformulation of

KGE as npKGE allows us to estimate non-parametric components (i.e., the Spearman rank correlation and the normalized flow–duration curve), which are necessary for watershed model calibrations aiming at multiple hydrograph aspects. Hence, including multiple accuracy metrics when evaluating a calibrated model has obvious advantages. In addition to the above metrics, we quantify the uncertainty of the modeled streamflow for each method. Uncertainty is measured by the averaged width of maximum and minimum modeled streamflow results over the simulation periods and how well the modeled uncertainty boundary contains the observed streamflow. We evaluate this predictive uncertainty and associated probability that streamflow is contained within this boundary for the top-50 sets estimated by the CNN, DDS, and GLUE methods.

4.5 Computational cost

The wall clock time to run the WY 2014 to 2016 SWAT model simulation (each realization) is approximately an hour on a four-core processor (Intel(R) i7-8650U CPU at 1.90 GHz), which is a standard desktop computer. The ensemble run simulations for training the CNN-enabled inverse models were developed using a cluster of 56 cores (Intel(R) Xeon(R) Gold 5120 CPU at 2.20 GHz) and 256 GB DDR4 RAM. We trained a total of 26,250 CNN architectures by using 400 Cori's KNL CPU nodes at NERSC. Each KNL node comprises of 68 physical 1.40 GHz Intel Xeon Phi Processor 7,250 (Knights Landing) with four threads per core, 96 GB DDR4, and 16 GB MCDRAM memory. The scalable hyperparameter tuning to identify top-50 CNN architectures led to the use of approximately 520,000 processor hours. This is the total computational cost to calibrate the SWAT process model using CNNs. The time to calibrate SWAT process model using DDS and GLUE is equal to 20,000 processor hours (5,000 realizations × 4 cores). Even though model calibration using CNN is expensive (≈ 18 hours/architecture), it is embarrassingly parallel, allowing us to efficiently use supercomputing resources. The DDS method is generally sequential, as the parameter update depends on the previous estimation. Also, the DDS-based estimations depend on initial random guesses similar to CNN training. Hence, the algorithm needs to be run multiple times to remove the effect of randomness. Like the GLUE method, DDS allowed us to calibrate parameters in approximately 10 h.

From this training time, it is evident that a thorough hyperparameter tuning can be computationally expensive and requires high-performance computing resources. This high training time is mainly due to the slow training of CNN models on CPUs, which can be accelerated by using graphic processing units (GPUs). Despite the expensive computational cost to develop the proposed CNN-enabled inverse models, the inference cost to estimate the SWAT model parameters takes only 0.16 s. Moreover, hyperparameter tuning allows us to find CNNs that are highly accurate. The tuned CNNs allow us to make ensemble estimations quickly without the need to retrain the model. The GLUE and DDS algorithms need to be re-run on each discharge input to estimate SWAT parameters, which makes the trained CNNs attractive for inference. This low inference time is attractive for estimating the SWAT model parameters using streamflow data (with and without observational errors/noise). The wall clock time for making a prediction/inference shows that our CNN-enabled parameter estimation is at least $\mathcal{O}(10^3)$ times faster than the GLUE and DDS-based methods (e.g., may require thousands of forward model runs for each observational time series), in addition to its predictive capability.

5 Results

This section presents results on the overall accuracy and efficiency of the proposed DL methodology. First, we provide results from sensitivity analysis performed using a mutual information theory on the SWAT ensembles (Jiang et al., 2022b). Second, we describe the CNN-enabled inverse modeling results from the ensemble runs. Third, we show the SWAT model parameters estimated from observed discharges and compare the performance of CNN-enabled parameter estimation with the results from application of the GLUE and DDS methods. We also compare the streamflow predictions from the calibrated SWAT model with observed discharges for both the calibration and validation periods for all three methods. Finally, we give the performance metrics and calibration uncertainties for CNN-enabled, DDS, and GLUE estimated parameters.

5.1 Sensitivity analysis results

Table 1 identifies sensitive parameters that influence simulated discharge at the ARW study site. Figure 5 shows that the simulated discharge is sensitive to 11 out of the 20 parameters: 1) SFTMP, 2) CH_K2, 3) ALPHA_BF, 4) RCHRG_DP, 5) CH_N2, 6) SMTMP, 7) TIMP, 8) CN, 9) SMFMX, 10) GW_DELAY, and 11) EPCO. These 11 parameters correspond to landscape, groundwater, channel, plant, and snow groups. The important parameters mentioned above are identified using the MI methodology as described in (Cover and Thomas, 2006; Jiang et al., 2022b). Mutual information is a non-negative value that measures the dependency between the SWAT model parameters and its outputs. Zero MI means that streamflow is not affected by that parameter, and higher values of MI mean higher dependency. We note that discharge is primarily influenced

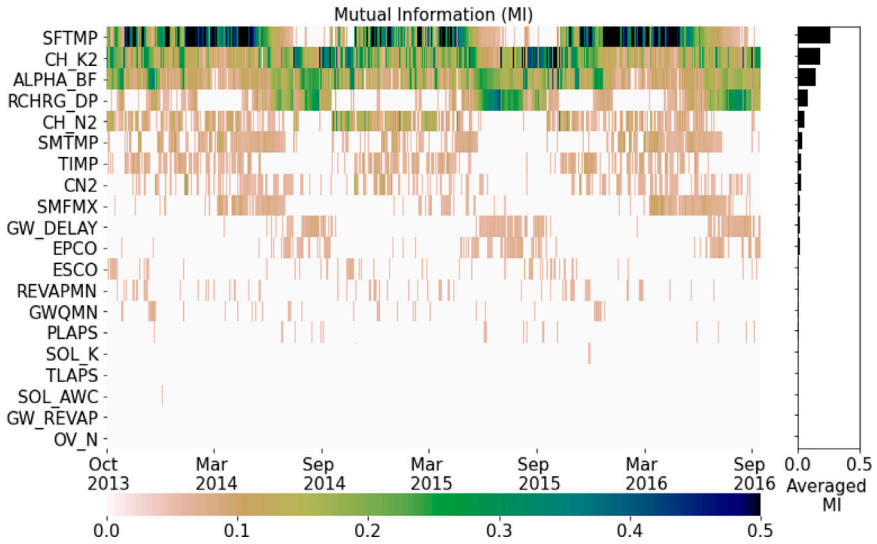

FIGURE 5
MI analysis on the SWAT model simulations: This figure shows the ranking of the SWAT model parameters based on MI. The analysis is performed for 1,000 realizations generated using a Sobol sequence. Among the sensitive parameters, it is evident that SFTMP is the dominant parameter and OV_N is the least important parameter for this ARW study site. In addition to snow parameters, we also see channel and groundwater parameter types are sensitive to streamflow.

by the snowfall temperature (SFTMP; the most sensitive), whose sensitivity shows the seasonality pattern consistent with the site description in Section 3.1. The importance of SFTMP in determining streamflow verifies the critical role of the snow process in this watershed.

5.2 Training, validation, and testing results

Figure 6A shows the training and validation loss of the best CNN-enabled inverse model in estimating the SWAT parameters. Validation loss plateaus even as the training loss decreases due to the lack of valuable information in the streamflow data to constrain the lesser and insensitive parameters (e.g., soil, climate, and other groundwater variables such as GW_REVAP). Figures 6B–D shows the prediction of the tuned CNN-enabled inverse model for estimating SFTMP. Supplementary Figures S1–S3 provide one-to-one plots for the remaining 10 sensitive parameters. Some one-to-one plots between the estimated and true parameters are closely distributed along the one-to-one line, which shows that the most sensitive parameters (e.g., SFTMP, CH_K2, ALPHA_BF) are predicted with reasonably good accuracy. The accuracy of the training predictions is lower for other less sensitive parameters (e.g., RCHRG_DP, EPCO). This reduced accuracy is evident from the more scattered drift away from the one-to-one straight line, seen in Supplementary Figure S3. The reduced accuracy is comparable to the training results where we see an increased deviation of data scatters from the one-to-one straight line. Similar results are obtained for other tuned CNN architectures. This scattered deviation indicates that these less-sensitive parameters are hard to predict using the discharge time series.

5.3 Sensitivity of estimated SWAT parameters to observation noise

We selected all test realizations to evaluate the parameter estimation sensitivity of the CNN-enabled inverse models to observed errors. We added random observation errors to the synthetic observed discharge time series for each test realization. We then generated 100 different observation realizations for parameter estimation, $\mathbf{q}_n$, which is given by (Cromwell et al., 2021)

$$\mathbf{q}_n = \mathbf{q} + \epsilon \times \mathbf{q} \times \mathbf{r} \tag{4}$$

where ϵ is the standard deviation of the noise, usually taken as $\frac{1}{3}$ of the observation error, and $\mathbf{r}$ is a random vector of the same size as $\mathbf{q}$. The elements of the random vector contain samples drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1. We tested different levels of observation errors (i.e., 5%, 10%, 25%, 50%, and 100%) relative to the observed values. These noisy discharge data (both synthetic and observations) are provided as input to the best CNN-enabled inverse models to estimate the SWAT model parameters.

Figure 7 shows the variability in estimated SFTMP results from the CNNs results as box plots. It also shows CNN model predictions for all noisy test realizations (Figure 7A) as well as

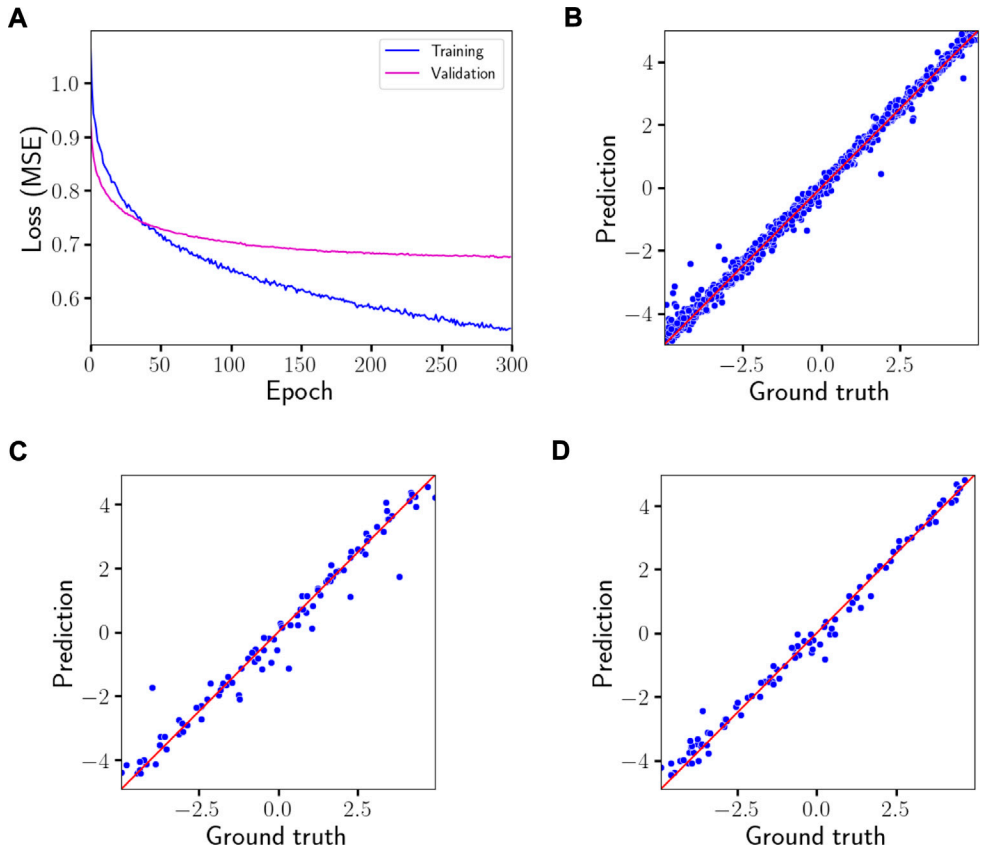

FIGURE 6
Loss metrics of the best CNN-enabled inverse model and its predictions for the SFTMP: The top left figure **(A)** figure shows the overall training and validation loss of the best CNN architecture. The top right **(B)**, lower left **(C)**, and lower right **(D)** figures show one-to-one plots for the most sensitive parameter, SFTMP (units in °C). It compares the CNN estimation with the ground truth for the training, validation, and test datasets. We did not use test data for finding the tuned CNN architectures. Only the validation set is used for hyperparameter tuning. Each blue dot represents a realization from the corresponding train/validation/test set of ensembles. The red line is the one-to-one line.

observed data (Figure 7B). Supplementary Figures S4–S8 provide estimates for the other sensitive parameters. We note that the parameter estimates are within the prior sampling range even after adding high relative noise levels (i.e., 100%), which instills confidence in the predictive capabilities of CNN models. From Figure 7A, it is evident that CNN-enabled inverse models are reasonably robust to noise in estimating the most sensitive parameter. This shows that the SFTMP predictions are not that sensitive to noise, as the performance of CNNs is stable even after adding high errors to data. This predictive capability when noise is applied to discharge time series also provides the insight that CNNs can effectively learn underlying representations in the streamflow data rather than noise in the observed data. Similar assessments can be made for the other sensitive parameters (e.g., CH_K2, ALPHA_BF). However, as model parameter sensitivity decreases, the CNN predictions are more prone to be influenced by noise. The performance of CNN estimation for EPCO, which is the least sensitive parameter among the top 11 parameters, is lower than that of sensitive parameters such as SFTMP. As discussed in Section 5.2 and from MI analysis, it is evident that streamflow provides little information to estimate this parameter. This reduced performance is the result of less valuable information being available in the streamflow data to accurately estimate less sensitive parameters, such as EPCO.

5.4 Calibrated SWAT model based on observed discharge

Trained CNN-enabled inverse models are used to estimate the SWAT parameters at the ARW study site based on observed discharge data. We provide streamflow predictions of the calibrated SWAT model based on the best CNN architecture and the following 49 best candidates. We also compare the

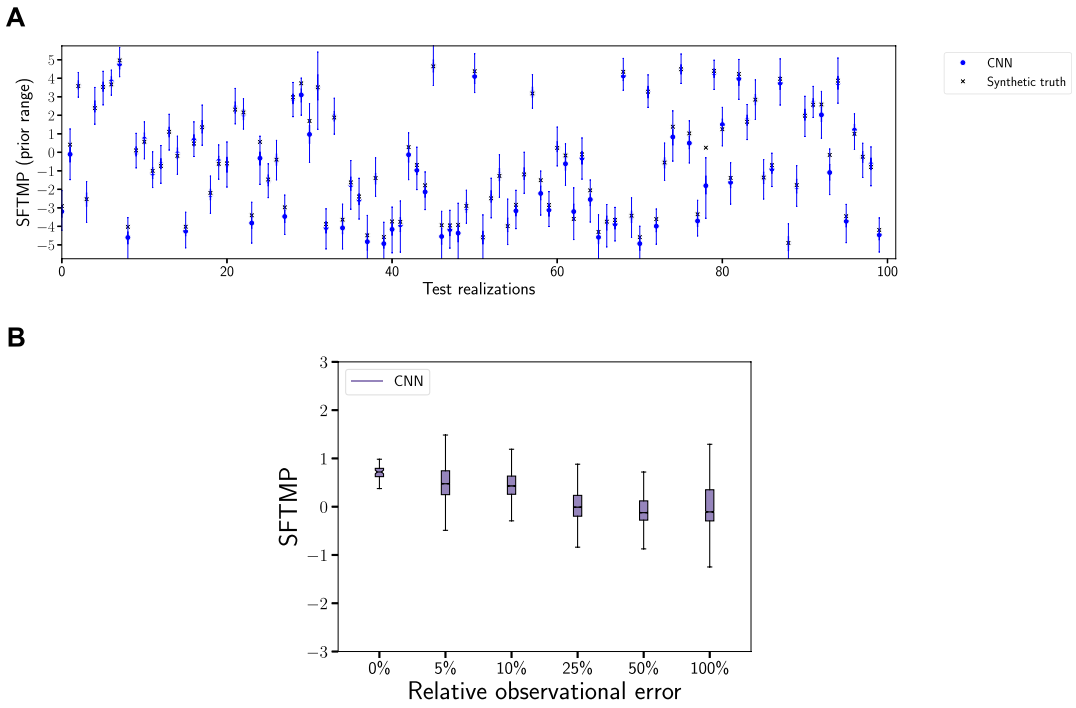

FIGURE 7
Top-50 CNN model estimation under influence of noise (SFTMP): The top figure **(A)** shows the sensitivities of the CNN-enabled inverse models to 100% noise added to the test realizations. The bottom figure **(B)** shows the CNN estimations on the observational discharge represented by the colors filling the box plots. Both of these figures show the variation in SFTMP (units in °C) with different noise levels. Moreover, the CNN estimations are closer to ground truth for all synthetic predictions as shown in the top figure. From the bottom figure, we observe that even under high relative observational errors, the estimations of the most sensitive parameter SFTMP are narrower compared to GLUE as seen in Figure 8.

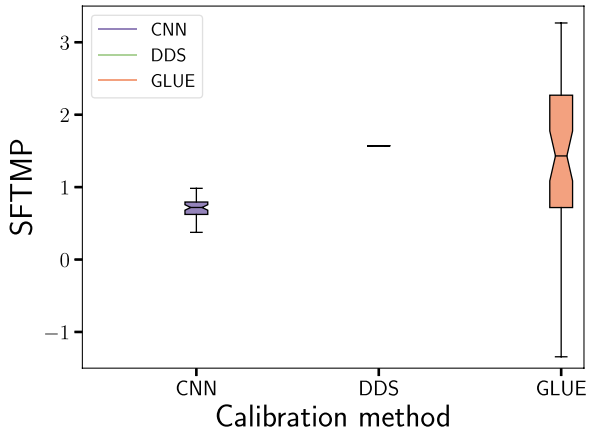

FIGURE 8
Comparison of top-50 CNN, DDS, and GLUE estimated sets (SFTMP): This figure compares the estimations of CNN, DDS, and GLUE on observational discharge for SFTMP (units in °C). From this figure, it is evident that DDS has the narrower range than CNN and GLUE.

performance of CNN predictions against predictions provided by the DDS and GLUE methods. Figure 8 shows the estimated SFTMP parameter range for CNN, DDS, and GLUE for the observed data. We see that the DDS method has a narrower range than the CNN and GLUE methods. The reason for this narrower range is DDS uses a global optimization algorithm that iteratively searches for a parameter set that produces a unique value. On the other hand, the loss function in the CNN method is non-convex, meaning that in all likelihood, gradient descent converges to sub-optimal valleys or local minima. Hence, the CNN method has a slightly broader range compared to the DDS method but a narrower range than the GLUE method. Similar inferences can be made on other parameters as shown in Supplementary Figure S9 provided in the supplementary information (e.g., ALPHA_BF, CH_K2, RCHRG_DP).

Figure 9 shows the calibration performance of top-50 CNN, DDS, and GLUE calibration set predictions using six different metrics. It is clear that CNN-enabled parameter estimation is better than behavioral parameter sets estimated by the GLUE and DDS methods for all six studied metrics. Additionally, in

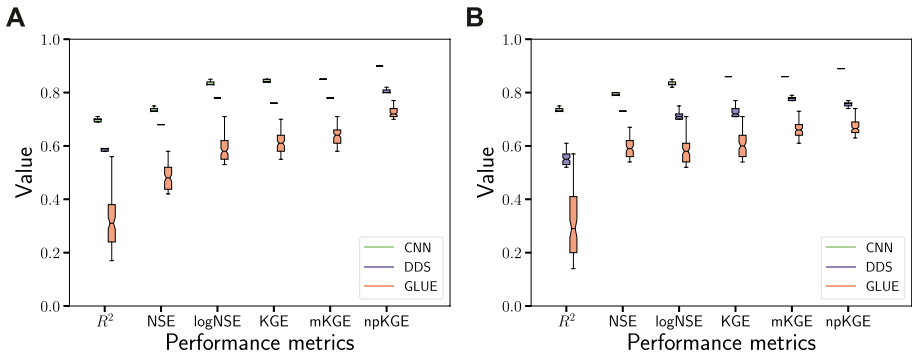

FIGURE 9
Performance metrics of top-50 estimated sets using CNN, DDS, and GLUE methods: This figure compares different performance metrics of the CNN, DDS, and GLUE calibrated sets. The left **(A)** and right **(B)** figures show the performances in calibration and validation periods, respectively. The green, blue, and red whiskers represent the CNN estimation, DDS, and GLUE. Top-50 best performance sets are identified and evaluated for each method within and beyond calibration period. The performance metrics (e.g., NSE, logNSE, npKGE) focus on the predictive capability of CNN-, DDS-, and GLUE-based calibrated SWAT models in both low and high flow scenarios. Across all performance metrics, it is evident that estimation using the CNN-enabled inverse models outperforms DDS and GLUE.

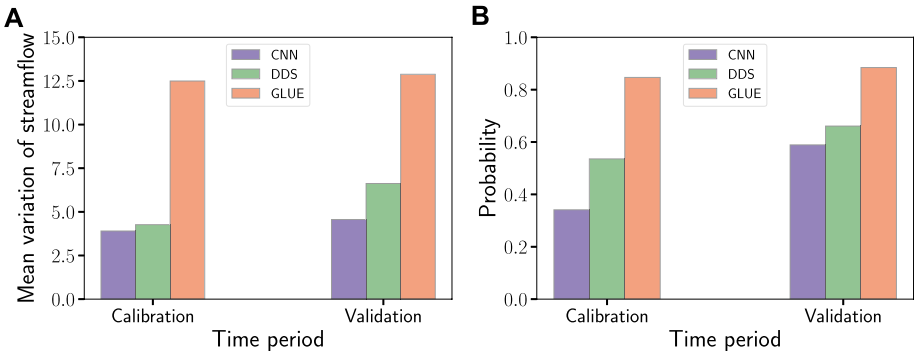

FIGURE 10
Comparison of top-50 CNN, DDS, and GLUE's streamflow variations: The left figure **(A)** shows the size of the mean modeled streamflow variation (i.e., a representation of predictive uncertainty). The right figure **(B)** provides the probability that the observed flow is contained within the predicted bounds of the streamflow (e.g., the light blue colored region in Figure 11) estimated by the calibrated SWAT model. The uncertainty in the GLUE-based calibration sets prediction, and associated probability is higher than DDS and CNN.

Supplementary Figure S15, we show one-to-one scatter plots for the best CNN, DDS, and GLUE streamflow predictions with observed data both in calibration and validation periods. In Supplementary Figure S15, each dot corresponds to daily streamflow. The predictions are based on the best sets calibrated by the CNN, GLUE, or DDS methods. The best CNN-based calibrated set has R^2, NSE, logNSE, KGE, mKGE, and npKGE scores of 0.71, 0.75, 0.85, 0.85, 0.86, and 0.91, respectively. The best DDS-based calibrated set has scores of 0.62, 0.69, 0.8, 0.77, 0.79, and 0.82. The best GLUE-based calibrated set has scores of 0.56, 0.58, 0.71, 0.7, 0.71, and 0.8. Supplementary Table S2 also provides the metric values for the CNN, GLUE, and DDS sets. From these values, it is clear that the CNN-enabled inverse model estimations are more accurate for SWAT model calibration than the GLUE and DDS estimations. Therefore, CNNs show promise for parameter estimation, especially in nonlinearly relating streamflow data to conceptual parameters.

Figure 10 shows smaller uncertainty ranges for CNN sets in both calibration and validation periods than the GLUE and DDS estimations. The probability that the prediction intervals estimated by the CNN sets contain the observed streamflow also is lower than the GLUE and DDS sets. This shows that top-50 CNN sets are not sufficient to capture the predictive boundary of streamflow variations. If we include all the CNN sets (as shown in Supplementary Figures S12–S15G,H), the probability that

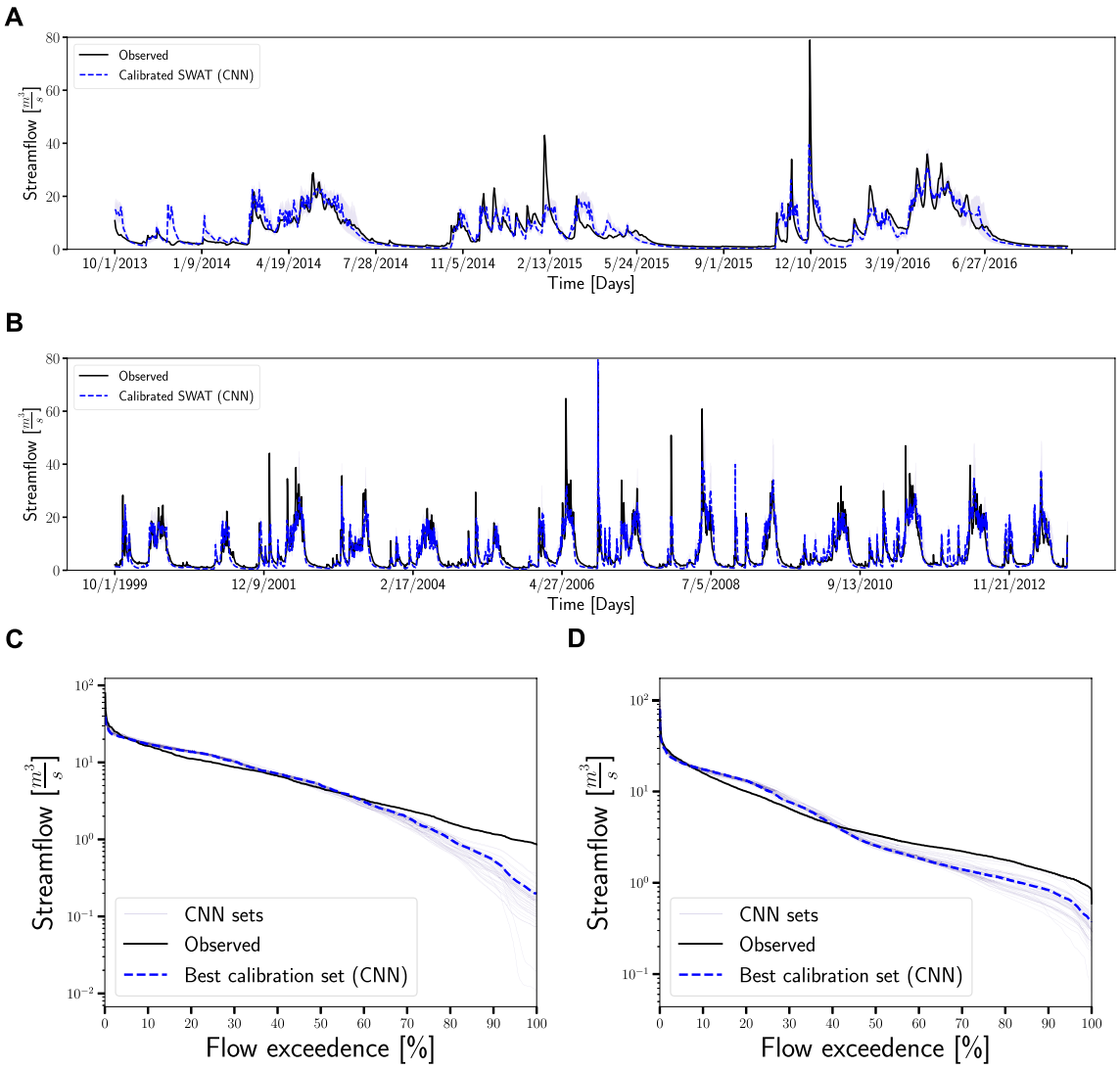

FIGURE 11
Comparison of the calibrated SWAT model (top-50 CNN) with observation data: This figure **(A)** and **(B)** compares the predictions of the calibrated SWAT model with observational data within and beyond the calibration period. The solid black line represents the observational data. The dashed colored (blue) line represents the predictions based on the best calibrated set using CNN. The light colored region in the streamflow plots represents the prediction uncertainty. This region is calculated by running the SWAT model using the calibration sets obtained using CNN. The bottom figures **(C)** and **(D)** show the flow duration curves in both calibration and validation period. It is clear that CNN estimation sets produce curves that are reasonably closer to observational data.

observational data is contained within the prediction bounds is greater than 0.95. However, the mean variation size of the streamflow increases fourfold to accommodate this increase in probability. One of our next steps is to improve this top-50 CNN estimation probability while keeping predictive uncertainty low. This can be achieved through ensemble DL, knowledge-guided DL, and probabilistic BNNs (Lu et al., 2021; Jiang et al., 2022a). These types of networks can account for uncertainty so that CNN-enabled inverse models can assign lower confidence levels to incorrect predictions. Figure 11 compares the streamflow predictions from the calibrated SWAT with the observed data using the top-50 CNN model sets. Supplementary Figures S10, S11 shows the predictions from both the top-50 GLUE and DDS methods. The CNN estimations capture the various high and low streamflows better than the GLUE and DDS methods during both the calibration and validation periods. However, the calibrated SWAT model over predicts in certain parts of WY 2014 (e.g., 9 January 2014) and WY 2015. This lower predictive performance may imply potential deficiencies (i.e., structural errors) in the underlying SWAT model representation of

watershed processes. Additional investigations are necessary to identify other processes and parameters that reduce structural errors and discrepancies in streamflow predictions.

6 Possible extensions of current work

Our results demonstrate the applicability of using scalable deep learning to calibrate the SWAT model. We note that the proposed methodology is general and can be used to calibrate other watershed models such as ATS and PRMS. This extensibility for calibrating other models and study sites can be achieved using transfer learning methods (Zhuang et al., 2020), which will allow us to reuse the CNNs developed in this study and leverage them for new, similar problems. Minimal re-training is necessary to fine tune the trained CNNs (Song and Tartakovsky, 2021) and apply them to calibrate watershed models for other study sites. Such a transfer of knowledge across study sites usually is performed when generating the large amount of training data needed to develop a full-scale CNN and tuning its trainable weights from the start is too computationally expensive (e.g., when using ATS). Additionally, we can improve our DL methodology to calibrate the SWAT model by incorporating other multi-source data streams (e.g., evapotranspiration (ET), and snow water equivalent (SWE)) along with streamflow. Our next step is to use such data streams to further investigate the deficiency of the model structure or processes in the SWAT model by ingesting streamflow, ET, and SWE into CNNs.

Figure 10B shows the DL method's probability that observational data contained within the prediction bounds are lower than probabilities provided by the DDS and GLUE methods. There are multiple ways in which we can improve our CNN-based parameter estimation and predictive uncertainty. A possible approach involves accelerating the training process using GPUs available at leadership class supercomputing resources (e.g., NERSC, Oak Ridge Leadership Computing Facility, and Argonne Leadership Computing Facility user facilities) (ALCF, 2021; NERSC, 2021; OLCF, 2021). This accelerated CNN training allows us to developed ensemble learning models through bootstrapping, which are known to provide better generalization performance than a final CNN.

As discussed in Section 5.4, improved uncertainty intervals can be achieved through ensemble learning (e.g., combining predictions of different types of neural networks such as DNNs and BNNs). Additionally, developing CNNs tailored to estimate the SWAT model parameters under different hydrological seasons (McMillan, 2020) (e.g., winter vs. summer) may enhance the calibration process. For example, comparing CNN-estimated sets from wet and dry periods of the year can provide better insights into the SWAT model parameters that control streamflow predictions across different seasons. When making such comparisons between real data and model predictions, hydrological signatures and their associated metrics (Westerberg and McMillan, 2015; McMillan et al., 2017; Fatehifar et al., 2021; Gnann et al., 2021; McMillan, 2021) can be used to elucidate the structural deficiencies of the SWAT model. Hydrological signatures on which we can evaluate performance metrics include the slope of the flow duration curve, rising limb density, recession shape, and baseflow index of streamflow time-series data (McMillan, 2021).

In addition to the data-driven methodology presented in this paper[6], the efficacy of the proposed DL methodology also can be improved by embedding domain knowledge into DNNs (Read et al., 2019; Khandelwal et al., 2020; Bhasme et al., 2021; Jia et al., 2021). Recent advances in knowledge-guided machine learning (Jiang et al., 2022a) provide a way to incorporate model states/fluxes and water balances as part of recurrent neural network architectures (Khandelwal et al., 2020). The papers mentioned above used such neural architectures to develop forward emulators for watershed models. One can extend the methods presented in those works to incorporate process model knowledge into our proposed CNNs to improve SWAT model calibration. Also, Explainable AI (XAI) methods such as deep Taylor decomposition (Kindermans et al., 2016), SHAPley values (Messalas et al., 2019), and integrated gradients (Sundararajan et al., 2017) can be used to explain the CNN predictions. These XAI methods not only allow us to explain why CNNs provide results that are understandable for the domain experts (Leduc et al., 2020) but also extract informative signals (e.g., precursors) from the streamflow time-series data (McMillan et al., 2017; McMillan, 2020; McMillan, 2021).

7 Conclusion

In this paper, we describe an accurate and reliable DL methodology that we developed to calibrate the SWAT model. We used CNN-enabled inverse models to estimate the SWAT parameters for the ARW study site in the YRB. Our approach leverages recent advances in CNNs to extract representations from streamflow data and then map them to the SWAT model parameters. Scalable hyperparameter tuning was performed to identify optimal CNN architectures. Ensemble runs from the SWAT model were used to train, validate, and test the CNN-enabled inverse models. We performed sensitivity analyses to identify the dominant parameters that influence streamflow. Our results show that CNN models are able to estimate the sensitive

6 Or by combining Markov chain Monte Carlo methods with forward emulators (Dagon et al., 2020) for model calibration.

parameters reasonably well. The parameters estimated from the trained CNNs were robust to high observed errors. We then compared the SWAT parameters estimated by our DL method with parameters generated by the GLUE and DDS optimization algorithms. We found that all the methods estimated SWAT parameters within the sampling range of the ensemble runs. As DDS is a global optimization method, its estimated range of parameters are narrower compared to parameters estimated by the GLUE and DL methods. Furthermore, this comparison also showed that predictions of the calibrated SWAT model based on CNNs performs better than the GLUE and DDS methods. Key performance metrics (e.g., R^2-score, NSE, logNSE, KGE, and its variants) showed that the best CNN-based calibration sets capture low and high flows better than the GLUE and DDS methods. This improvement in predictive performance is probably because CNNs can more effectively use the information (e.g., learning representative features from streamflow) provided in ensemble runs than the GLUE and DDS methods. By capturing the nonlinear relationships between SWAT model inputs and outputs through multiple convolutional neural layers, CNNs yielded more realistic predictions for the ARW and a better calibrated SWAT model. This improvement resulted in a closer match between model-predicted and observed stream discharges. Our results showed that the probability that the observed data are contained within the prediction bounds estimated by top-50 CNN sets is lower than that of DDS and GLUE sets. This lower probability shows that the top-50 CNN sets alone are insufficient to capture the variations in streamflow. If all the CNN estimations are included, we are able to capture the observed data within the prediction bounds. However, including all the CNN estimations resulted in higher mean variation of streamflow (i.e., fourfold increase when compared to the top-50 CNN sets). Our future work involves further improving the accuracy the CNN method while keeping the predictive uncertainty (i.e., size of streamflow variation) lower.

From a computational cost perspective, the time needed to infer parameters based on the DL method is at least $\mathcal{O}(10^3)$ faster than that of the GLUE and DDS methods, which makes extending this method to complex watershed models (e.g., ATS) attractive. However, the computational cost of identifying optimal CNN architectures is high compared to the GLUE and DDS methods. The training time needed to develop CNN models can be improved further by using GPUs and TPUs (Bisong, 2019). Reducing the computational cost of developing CNN-enabled inverse models is one of our next steps, with a focus on using the distributed deep learning training framework (e.g., using Horovod (Sergeev and Del Balso, 2018) or DeepHyper (Balaprakash et al., 2018)) that already shows promise in the training speedup. This improves the efficiency during training process by using asynchronous distributed Bayesian optimization algorithms, which are known to be much more efficient than the grid search that has to exhaust all the hyperparameter space.

Our methodology is general and can be used to calibrate complex watershed models (i.e., through transfer learning methods (Zhuang et al., 2020)) with minimal re-training. For example, using transfer learning. Transfer learning consists of using pre-trained deep learning models such as CNNs on one watershed and leveraging them on a new and similar watershed. Specifically, transfer learning (Oruche et al., 2021) allows us to transfer knowledge from gauged (e.g., ARW) to ungauged basins (e.g., YRB) or watersheds (Westerberg et al., 2016; Guo et al., 2021). This knowledge transfer is usually done when training a full-scale CNN from scratch is challenging due to the availability of limited simulation data or when regions are data sparse, observationally. In such scenarios, a watershed classification scheme is first used to identify a new watershed with characteristics similar to ARW. Then, the neural features from the pre-trained CNN that has learned to extract patterns from ARW's streamflow data can be adapted to that new paired watershed. Finally, fine-tuning is performed to achieve meaningful improvements by incrementally adapting the pre-trained CNN's features to the new simulation data. For fine-tuning to be successful, minimal simulation data on the newly selected watershed is needed. Additional future work involves modifying the proposed method to incorporate multi-source datasets (e.g., by combining streamflow, ET, and SWE) to further enhance SWAT model calibration (Moriasi et al., 2007; Samimi et al., 2020), and transfer the knowledge gained on ARW to the entire Yakima river basin (i.e., by transfer learning).

Data availability statement

The data generated for the proposed DL model development uses open-science principles. Specifically, we make the data Findable Accessible Interoperable and Reusable (FAIR). FAIR principles expedite community-based data generation, modeling, and interdisciplinary collaboration and provides a means to test new hypotheses. The datasets generated and analyzed as well as scripts for this study, will be made available on this GitHub repository: ˜\url {https://github.com/maruti-iitm/DL4SWAT.git} upon publication. \texttt{SWAT} open-source code can be downloaded at \url {https://swat.tamu.edu/}.

Author contributions

MM: Conceptualization, methodology, software, data curation, visualization, investigation, writing—original draft, writing—review and editing. KS: Methodology, data generation, writing—review and editing. PJ: Sensitivity analysis, writing—review and editing. GH: Methodology,

writing—review and editing. XC: writing—review and editing and funding.

Funding

This research was supported by the U.S. Department of Energy (DOE), Office of Science (SC) Biological and Environmental Research (BER) program, as part of BER's Environmental System Science program.

Acknowledgments

This contribution originates from the River Corridor Scientific Focus Area at Pacific Northwest National Laboratory (PNNL). This research used resources from the National Energy Research Scientific Computing Center, a DOE-SC User Facility. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. The authors thank the reviewers whose feedback helped in substantially improving the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.1026479/full#supplementary-material

References

Abbaspour, K. C. (2013). *Swat-cup 2012. SWAT calibration and uncertainty program–A user manual.*

Adams, B. M., Bohnhoff, W. J., Dalbey, K. R., Eddy, J. P., Eldred, M. S., Gay, D. M., et al. (2009). *Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 5.0 user's manual*. Tech. Rep. SAND2010-2183.*Sandia Natl. Lab.*

Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., and Esau, T. (2020). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. *Water* 12, 5. doi:10.3390/w12010005

ALCF (2021). *Argonne leadership computing facility*. Available at: https://www.alcf.anl.gov/ (Accessed on 07, 202121).

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., et al. (2009). The data assimilation research testbed: A community facility. *Bull. Am. Meteorol. Soc.* 90, 1283–1296. doi:10.1175/2009bams2618.1

Anysz, H., Zbiciak, A., and Ibadov, N. (2016). The influence of input data standardization method on prediction accuracy of artificial neural networks. *Procedia Eng.* 153, 66–70. doi:10.1016/j.proeng.2016.08.081

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., et al. (2012). Swat: Model use, calibration, and validation. *Trans. ASABE* 55, 1491–1508. doi:10.13031/2013.42256

Aster, R. C., Borchers, B., and Thurber, C. H. (2018). *Parameter estimation and inverse problems*. Elsevier.

Bacu, V., Nandra, C., Stefanut, T., and Gorgan, D. (2017). SWAT model calibration over Cloud infrastructures using the BigEarth platform. *13th IEEE Int. Conf. Intelligent Comput. Commun. Process. (ICCP)*, 453–460.

Balaprakash, P., Salim, M., Uram, T., Vishwanath, V., and Wild, S. (2018). Deephyper: Asynchronous hyperparameter search for deep neural networks. *IEEE 25th Int. Conf. high Perform. Comput. (HiPC)*, 42–51.

Beven, K., and Binley, A. (2014). Glue: 20 years on. *Hydrol. Process.* 28, 5897–5918. doi:10.1002/hyp.10082

Bhasme, P., Vagadiya, J., and Bhatia, U. (2021). *Enhancing predictive skills in physically-consistent way: Physics informed machine learning for hydrological processes*. arXiv preprint arXiv:2104.11009.

Bisong, E. (2019). "Google colaboratory," in *Building machine learning and deep learning models on google cloud platform* (Berlin: Springer), 59–64.

Blasone, R.-S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., and Zyvoloski, G. A. (2008). Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling. *Adv. Water Resour.* 31, 630–648. doi:10.1016/j.advwatres.2007.12.003

Chen, X., Hammond, G. E., Murray, C. J., Rockhold, M. L., Vermeul, V. R., and Zachara, J. M. (2013). Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at Hanford 300 area. *Water Resour. Res.* 49, 7064–7076. doi:10.1002/2012wr013285

Chen, X., Shuai, P., Son, K., Jiang, P., Mudunuru, M., Coon, E., et al. (2021). AGU fall meeting abstracts. In *What can we learn from multiple watershed models and observations?* 2021. H23H–01.

Chiang, L.-C., and Yuan, Y. (2015). The NHDPlus dataset, watershed subdivision and SWAT model performance. *Hydrological Sci. J.* 60, 1690–1708. doi:10.1080/02626667.2014.916408

Chollet, F. (2017). *Deep learning with Python*. Shelter Island, NY: Manning Publications Company.

Coon, E. T., Berndt, M., Jan, A., Svyatsky, D., Atchley, A. L., Kikinzon, E., et al. (2020). *Advanced terrestrial simulator*. USA: U.S. Department of Energy. Version 1.0. doi:10.11578/dc.20190911.1

Cover, T. M., and Thomas, J. A. (2006). *Wiley Series in Telecommunications and Signal Processing*. Elements of information theory.

Cromwell, E. L. D., Shuai, P., Jiang, P., Coon, E., Painter, S. L., Moulton, D., et al. (2021). Estimating watershed subsurface permeability from stream discharge data using deep neural networks. *Front. Earth Sci. (Lausanne).* 9. doi:10.3389/feart.2021.613011

Cuo, L., Lettenmaier, D. P., Mattheussen, B. V., Storck, P., and Wiley, M. (2008). Hydrologic prediction for urban watersheds with the distributed hydrology–soil–vegetation model. *Hydrol. Process.* 22, 4205–4213. doi:10.1002/hyp.7023

Dagon, K., Sanderson, B. M., Fisher, R. A., and Lawrence, D. M. (2020). A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. *Adv. Stat. Climatol. Meteorol. Oceanogr.* 6, 223–244. doi:10.5194/ascmo-6-223-2020

Daly, C., and Bryant, K. (2013). *The PRISM climate and weather system–An introduction*. Corvallis, OR: PRISM climate group.

Daly, C., Taylor, G. H., Gibson, W. P., Parzybok, T. W., Johnson, G. L., and Pasteris, P. A. (2000). High-quality spatial climate data sets for the United States and beyond. *Trans. ASAE* 43, 1957–1962. doi:10.13031/2013.3101

Daniel, E. B., Camp, J. V., LeBoeuf, E. J., Penrod, J. R., Dobbins, J. P., and Abkowitz, M. D. (2011). Watershed modeling and its applications: A state-of-the-art review. *Open Hydrology J.* 5, 26–50. doi:10.2174/1874378101105010026

Daymet (2021). *Daily surface weather and climatological summaries*. Available at: https://daymet.ornl.gov/ (Accessed on 07, 202121).

Doherty, J. E., and Hunt, R. J. (2010). *Approaches to highly parameterized inversion: A guide to using PEST for groundwater-model calibration*, 2010. Middleton, WI: U.S. Geological Survey.

Donigian, A. S., Jr, Bicknell, B. R., and Imhoff, J. C. (1995). Hydrological simulation program-fortran (HSPF). *Comput. models watershed hydrology*, 395–442.

Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., and Turcotte, R. (2004). *Calibration of watershed models*. American Geophysical Union.

Duan, Q., Sorooshian, S., and Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrology* 158, 265–284. doi:10.1016/0022-1694(94)90057-4

Eckhardt, K., Fohrer, N., and Frede, H.-G. (2005). Automatic model calibration. *Hydrol. Process.* 19, 651–658. doi:10.1002/hyp.5613

Edwards, C. (2018). Deep learning hunts for signals among the noise. *Commun. ACM* 61, 13–14. doi:10.1145/3204445

Evensen, G. (2018). Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.* 22, 885–908. doi:10.1007/s10596-018-9731-y

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* 99, 10143–10162. doi:10.1029/94jc00572

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean. Dyn.* 53, 343–367. doi:10.1007/s10236-003-0036-9

Fang, Y., Chen, X., Velez, J. G., Zhang, X., Duan, Z., Hammond, G. E., et al. (2020). A multirate mass transfer model to represent the interaction of multicomponent biogeochemical processes between surface water and hyporheic zones (SWAT-MRMT-R 1.0). *Geosci. Model. Dev.* 13, 3553–3569. doi:10.5194/gmd-13-3553-2020

Fatehifar, A., Goodarzi, M. R., Montazeri, H., S. S., and Dastjerdi, S. (2021). Assessing watershed hydrological response to climate change based on signature indices. *J. Water Clim. Change* 12, 2579–2593. doi:10.2166/wcc.2021.293

Franco, A. C. L., and Bonumá, N. B. (2017). Multi-variable SWAT model calibration with remotely sensed evapotranspiration and observed flow. *RBRH* 22. doi:10.1590/2318-0331.011716090

Gabrielli, L., Tomassetti, S., Squartini, S., and Zinato, C. (2017). "Introducing deep machine learning for parameter estimation in physical modelling," in *Proceedings of the 20th international conference on digital audio effects*.

Gnann, S. J., Coxon, G., Woods, R. A., Howden, N. J. K., and McMillan, H. K. (2021). Tossh: A toolbox for streamflow signatures in hydrology. *Environ. Model. Softw.* 138, 104983. doi:10.1016/j.envsoft.2021.104983

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.

Graham, D. N., and Butts, M. B. (2005). Flexible, integrated watershed modelling with MIKE SHE. *Watershed models*. 849336090, 245–272.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognit.* 77, 354–377. doi:10.1016/j.patcog.2017.10.013

Guo, Y., Zhang, Y., Zhang, L., and Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *WIREs Water* 8, e1487. doi:10.1002/wat2.1487

Gupta, H. V., Sorooshian, S., Hogue, T. S., and Boyle, D. P. (2003). Advances in automatic calibration of watershed models. *Calibration Watershed Models* 6, 9–28.

Gupta, S., and Gupta, A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Comput. Sci.* 161, 466–474. doi:10.1016/j.procs.2019.11.146

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., and Mao, Y. (2018). The Variable Infiltration Capacity model version 5 (VIC-5): Infrastructure improvements for new applications and reproducibility. *Geosci. Model. Dev.* 11, 3481–3496. doi:10.5194/gmd-11-3481-2018

Herman, J., and Usher, W. (2017). SALib: An open-source Python library for sensitivity analysis. *J. Open Source Softw.* 2, 97. doi:10.21105/joss.00097

Hydroeval (2021). *An evaluator for streamflow time series in Python*. Available at: https://github.com/ThibHlln/hydroeval.git (Accessed on 0803, 2022).

Jagtap, N. V., Mudunuru, M. K., and Nakshatrala, K. B. (2021). A deep learning modeling framework to capture mixing patterns in reactive-transport systems. *Commun. Comput. Phys.* 0.4208/cicp.OA-2021-0088.

Jia, X., Willard, J. D., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., et al. (2021). Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM. IMS. Trans. Data Sci.* 2, 1–26. doi:10.1145/3447814

Jiang, P., Chen, X., Chen, K., Anderson, J., Collins, N., and Gharamti, M. E. (2021). DART-PFLOTRAN: An ensemble-based data assimilation system for estimating subsurface flow and transport model parameters. *Environ. Model. Softw.* 142, 105074. doi:10.1016/j.envsoft.2021.105074

Jiang, P., Shuai, P., Sun, A., Mudunuru, M. K., and Chen, X. (2022a). Knowledge-informed deep learning for hydrological model calibration: An application to coal creek watershed in Colorado. *Hydrology Earth Syst. Sci. Discuss.*, 1–31. doi:10.5194/hess-2022-282

Jiang, P., Son, K., Mudunuru, M. K., and Chen, X. (2022b). *Using mutual information for global sensitivity analysis on watershed modeling*. Malden, MA: John Wiley & Sons Inc.

Johnston, P. R., and Pilgrim, D. H. (1976). Parameter optimization for watershed models. *Water Resour. Res.* 12, 477–486. doi:10.1029/wr012i003p00477

Keras API (2021). *The high-level API of Tensorflow*. Available at: https://www.tensorflow.org/api_docs/python/tf/keras (Accessed on 07, 202121).

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., et al. (2020). *Physics guided machine learning methods for hydrology*. arXiv preprint arXiv: 2012.02854.

Kindermans, P.-J., Schütt, K., Müller, K.-R., and Dähne, S. (2016). *Investigating the influence of noise and distractors on the interpretation of neural networks*. arXiv preprint arXiv:1611.07270.

Kling, H., Fuchs, M., and Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. hydrology* 424, 264–277. doi:10.1016/j.jhydrol.2012.01.011

Leavesley, G. H., Lichty, R. W., Troutman, B. M., and Saindon, L. G. (1983). Precipitation-runoff modeling system: User's manual. *Water-resources Investig. Rep.* 83, 207.

Leduc, R., Hulbert, C., McBrearty, I. W., and Johnson, P. A. (2020). Probing slow earthquakes with deep learning. *Geophys. Res. Lett.* 47, e2019GL085870. doi:10.1029/2019gl085870

Liu, D. (2020). A rational performance criterion for hydrological model. *J. Hydrology* 590, 125488. doi:10.1016/j.jhydrol.2020.125488

Lu, D., Konapala, G., Painter, S. L., Kao, S.-C., and Gangrade, S. (2021). Streamflow simulation in data-Scarce basins sing Bayesian and physics-informed machine learning models. *J. Hydrometeorol.* 22, 1421–1438.

MADS (2021). *Model analysis & decision Support*. Available at: https://mads.lanl.gov/ (Accessed on 07, 202121).

Mankin, D., Srinivasan, R., and Arnold, J. G. (2010). Soil and water assessment tool (SWAT) model: Current developments and applications. *Trans. ASABE* 53, 1423–1431. doi:10.13031/2013.34915

Marçais, J., and de Dreuzy, J.-R. (2017). Prospective interest of deep learning for hydrological inference. *Groundwater* 55, 688–692. doi:10.1111/gwat.12557

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M. T., Payn, R. A., et al. (2015). PRMS-IV, the precipitation-runoff modeling system, version 4. *U. S. Geol. Surv. Tech. Methods* 6, B7.

Mastin, M. C., and Vaccaro, J. J. (2002). *Tech. Rep., open-file report 02-404*. Washington DC, USA: U.S. Department of Interior.Watershed models for decision support in the Yakima river basin, Washington

McMillan, H. K. (2021). A review of hydrologic signatures and their applications. *WIREs Water* 8, e1499. doi:10.1002/wat2.1499

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. *Hydrol. Process.* 34, 1393–1409. doi:10.1002/hyp.13632

McMillan, H., Westerberg, I., and Branger, F. (2017). Five guidelines for selecting hydrological signatures. *Hydrol. Process.* 31, 4757–4761. doi:10.1002/hyp.11300

Mein, R. G., and Brown, B. M. (1978). Sensitivity of optimized parameters in watershed models. *Water Resour. Res.* 14, 299–303. doi:10.1029/wr014i002p00299

Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). "Model-agnostic interpretability with SHAPley values," in *2019 10th international conference on information, intelligence, systems and applications (IISA)*, 1–7.

Misirli, F., Gupta, H. V., Sorooshian, S., and Thiemann, M. (2003). Bayesian recursive estimation of parameter and output uncertainty for watershed models. *Calibration Watershed Models, Water Sci. Appl. Ser* 6, 113–124.

Model Analysis ToolKit (2021). *Python toolkit for model analysis*. Available at: http://dharp.github.io/matk/ (Accessed on 07, 202121).

Moore, R. B., and Dewald, T. G. (2016). The road to NHDPlus-advancements in digital stream networks and associated catchments. *J. Am. Water Resour. Assoc.* 52, 890–900. doi:10.1111/1752-1688.12389

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50, 885–900. doi:10.13031/2013.23153

Mudunuru, M. K., Cromwell, E. L. D., Wang, H., and Chen, X. (2022). Deep learning to estimate permeability using geophysical data. *Adv. Water Resour.* 167, 104272. doi:10.1016/j.advwatres.2022.104272

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *J. Math. Psychol.* 47, 90–100. doi:10.1016/s0022-2496(02)00028-7

Nakshatrala, K. B., and Joshaghani, M. S. (2019). On interface conditions for flows in coupled free-porous media. *Transp. Porous Media* 130, 577–609. doi:10.1007/s11242-019-01326-7

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resour. Res.* 57, e2020WR028091. doi:10.1029/2020wr028091

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R. (2011). *Soil & water assessment tool theoretical documentation, version 2009, Grassland, soil and water research laboratory-agricultural research service*. Temple, TX: Blackland Research Center-Texas AgriLife Research.

NERSC (2021). *National energy research scientific computing center*. Available at: https://www.nersc.gov/ (Accessed on 07, 202121).

Nott, D. J., Marshall, L., and Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resour. Res.* 48. doi:10.1029/2011wr011128

OLCF (2021). *Oak Ridge leadership computing facility*. Available at: https://www.olcf.ornl.gov/ (Accessed on 07, 202121).

Oruche, R., Egede, L., Baker, T., and O'Donncha, F. (2021). *Transfer learning to improve streamflow forecasts in data sparse regions. arXiv preprint arXiv:2112.03088.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pool, S., Vis, M., and Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sci. J.* 63, 1941–1953. doi:10.1080/02626667.2018.1552002

PRISM (2021). *A high-resolution spatial climate data for the United States*. Available at: https://prism.oregonstate.edu/ (Accessed on 07, 202121).

Qiu, J., Yang, Q., Zhang, X., Huang, M., Adam, J. C., and Malek, K. (2019). Implications of water management representations for watershed hydrologic modeling in the Yakima River basin. *Hydrol. Earth Syst. Sci.* 23, 35–49. doi:10.5194/hess-23-35-2019

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.* 16, 024025. doi:10.1088/1748-9326/abd501

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resour. Res.* 55, 9173–9190. doi:10.1029/2019wr024922

Rolnick, D., Veit, A., Belongie, S., and Shavit, N. (2017). *Deep learning is robust to massive label noise*. arXiv preprint arXiv:1705.10694.

Rouholahnejad, E., Abbaspour, K. C., Vejdani, M., Srinivasan, R., Schulin, R., and Lehmann, A. (2012). A parallelization framework for calibration of hydrological models. *Environ. Model. Softw.* 31, 28–36. doi:10.1016/j.envsoft.2011.12.001

Rudi, J., Bessac, J., and Lenzi, A. (2020). *Parameter estimation with dense and convolutional neural networks applied to the FitzHugh-Nagumo ODE*. arXiv preprint arXiv:2012.06691.

Sadeghi, M., Asanjan, A. A., Faridzad, M., Nguyen, P., Hsu, K., Sorooshian, S., et al. (2019). PERSIANN-CNN: Precipitation estimation from remotely sensed information using artificial neural networks–convolutional neural networks. *J. Hydrometeorol.* 20, 2273–2289. doi:10.1175/jhm-d-19-0110.1

Samimi, M., Mirchi, A., Moriasi, D., Ahn, S., Alian, S., Taghvaeian, S., et al. (2020). Modeling arid/semi-arid irrigated agricultural watersheds with SWAT: Applications, challenges, and solution strategies. *J. Hydrology* 590, 125418. doi:10.1016/j.jhydrol.2020.125418

Sampson, K., and Gochis, D. (2018). *RF Hydro GIS pre-processing tools, version 5.0, documentation*. Boulder, CO: National Center for Atmospheric Research, Research Applications Laboratory.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Schwarz, G. E., and Alexander, R. B. (1995). State soil geographic (STATSGO) data base for the conterminous United States. *Tech. Rep.*

Sergeev, A., and Del Balso, M. (2018). *Horovod: Fast and easy distributed deep learning in Tensorflow*. arXiv preprint: 1802.05799.

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* 54, 8558–8593. doi:10.1029/2018wr022643

Singh, V. P., and Frevert, D. K. (2003). "Watershed modeling," in *World water & environmental resources congress 2003*, 1–37.

Singh, V. P., and Frevert, D. K. (2010). *Watershed models*. Boca Raton, FL: CRC Press.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* 82, 2635–2670. doi:10.2166/wst.2020.369

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* 7, 86–112. doi:10.1016/0041-5553(67)90144-9

Song, D. H., and Tartakovsky, D. M. (2021). *Transfer learning on multi-fidelity data*. arXiv preprint arXiv:2105.00856.

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks," in *International conference on machine learning*, 3319–3328.

Tague, C. L., and Band, L. E. (2004). RHESSys: Regional Hydro-Ecologic Simulation System-An object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling. *Earth Interact.* 8, 1–42. doi:10.1175/1087-3562(2004)8<1:rrhsso>2.0.co;2

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Philadelphia, PA: SIAM.

Thiemann, M., Trosset, M., Gupta, H. V., and Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resour. Res.* 37, 2521–2535. doi:10.1029/2000wr900405

Tolson, B. A., and Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.* 43. doi:10.1029/2005wr004723

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.* 12, 5988–6013. doi:10.1038/s41467-021-26107-z

Van Leeuwen, P. J., and Evensen, G. (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Weather Rev.* 124, 2898–2913. doi:10.1175/1520-0493(1996)124<2898:daaimi>2.0.co;2

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H., and Anh, D. T. (2020). Deep learning convolutional neural network in rainfall–runoff modelling. *J. Hydroinformatics* 22, 541–561. doi:10.2166/hydro.2020.095

Westerberg, I. K., and McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrol. Earth Syst. Sci.* 19, 3951–3968. doi:10.5194/hess-19-3951-2015

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., et al. (2016). Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resour. Res.* 52, 1847–1865. doi:10.1002/2015wr017635

Willard, J. D., Read, J. S., Appling, A. P., Oliver, S. K., Jia, X., and Kumar, V. (2022). *Predicting water temperature dynamics of unmonitored lakes with meta transfer learning*. Malden, MA: John Wiley & Sons Inc. e2021WR029579.

Wu, R., Chen, X., Hammond, G. E., Bisht, G., Song, X., Huang, M., et al. (2021). *Coupling surface flow with high-performance subsurface reactive flow and transport code PFLOTRAN*, 137. Environmental Modelling & Software.

Zhang, D., Chen, X., Yao, H., and James, A. (2016). Moving SWAT model calibration and uncertainty analysis to an enterprise Hadoop-based cloud. *Environ. Model. Softw.* 84, 140–148. doi:10.1016/j.envsoft.2016.06.024

Zhang, X., Srinivasan, R., and Van Liew, M. (2009). Approximating SWAT model using artificial neural network and support vector machine. *JAWRA J. Am. Water Resour. Assoc.* 45, 460–474. doi:10.1111/j.1752-1688.2009.00302.x

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76. doi:10.1109/jproc.2020.3004555

Nomenclature

ARW American River Watershed

ATS Advanced Terrestrial Simulator

BNN Bayesian Neural Networks

CN Curve Number

CNN Convolutional Neural Network

DART The Data Assimilation Research Testbed

DHSVM The Distributed Hydrology Soil Vegetation Model

DDS Dynamically Dimensioned Search

DEM Digital Elevation Model

DNN Deep Neural Network

DL Deep Learning

ET Evapotranspiration

FAIR Findable Accessible Interoperable and Reusable

GIS Geographic Information System

GLUE Generalized Likelihood Uncertainty Estimation

GSA Global Sensitivity Analysis

GPU Graphical Processing Unit

HRU Hydrologic Response Unit

HSPF Hydrological Simulation Program-Fortran

MADS Model Analysis & Decision Support

KGE Kling-Gupta Efficiency

MATK Model Analysis ToolKit

mKGE Modified Kling-Gupta Efficiency

MI Mutual Information

ModEx Model-Experimentation

MTL Multi-Task Learning

NLCD National Land cover Database

NHDPlus National Hydrography Dataset Plus

npKGE Non-Parametric Kling-Gupta Efficiency

NSE Nash-Sutcliffe efficiency

logNSE Logarithmic Nash-Sutcliffe Efficiency

NWM The National Water Model

PET Potential Evapotranspiration

PEST Parameter Estimation Software

PRISM Parameter Elevation Regression on Independent Slopes Model

PRMS Precipitation Runoff Modeling System

RHESSys Regional Hydro-Ecologic Simulation System

SCE-UA Shuffled Complex Evolution Method developed at The University of Arizona

SWAT Soil and Water Assessment Tool

SWAT-CUP SWAT Calibration and Uncertainty Programs

SNOTEL Snow Telemetry

STATSGO Soil Maps for the State Soil Geographic

STL Single-Task Learning

TPU Tensor Processing Unit

USGS United States Geological Survey

WRF-Hydro The Weather Research and Forecasting Model Hydrological Modeling System

VIC The Variable Infiltration Capacity model

XAI Explainable AI

YRB Yakima River Basin

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 19 December 2022
DOI 10.3389/feart.2022.972305

OPEN ACCESS

EDITED BY
E. Bruce Pitman,
University at Buffalo, United States

REVIEWED BY
Subodh Chandra Pal,
University of Burdwan, India
Ryan Thomas Bailey,
Colorado State University, United States

*CORRESPONDENCE
Wesley Kitlasten,
w.kitlasten@gns.cri.nz

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 18 June 2022
ACCEPTED 05 December 2022
PUBLISHED 19 December 2022

CITATION
Kitlasten W, Moore CR and Hemmings B (2022), Model structure and ensemble size: Implications for predictions of groundwater age.
Front. Earth Sci. 10:972305.
doi: 10.3389/feart.2022.972305

Model structure and ensemble size: Implications for predictions of groundwater age

Wesley Kitlasten[1]*, Catherine R. Moore[2] and Brioch Hemmings[1]

[1]Wairakei Research Centre, GNS Science, Taupō, New Zealand, [2]GNS Science, Lower Hutt, New Zealand

This paper examines the influence of simplified vertical discretization using 50- to four- layer models and ensemble size on history matching and predictions of groundwater age for a national scale model of New Zealand (approximately 265,000 km^2). A reproducible workflow using a combination of opensource tools and custom python scripts is used to generate three models that use the same model domain and underlying data with only the vertical discretization changing between the models. The iterative ensemble smoother approach is used for history matching each model to the same synthetic dataset. The results show that: 1) the ensemble based mean objective function is not a good indicator of model predictive ability, 2) predictive failure from model structural errors in the simplified models are compounded by history matching, especially when small (<100 member) ensembles are used, 3) predictive failure rates increase with iteration, 4) predictive failure rates for the simplified model reach 30–65% using 50-member ensembles, but stabilize at relatively low values (<10%) using the 300 member ensemble, 5) small (50 member) ensembles contribute to predictive failure of 22–30% after six iterations even in structurally "perfect" models, 6) correlation-based localization methods can help reduce prediction failure associated with small ensembles by up to 45%, 7) the deleterious effects of model simplification and ensemble size are problem specific. Systematic investigation of these issues is an important part of the model design, and this investigation process benefits greatly from a scripted, reproducible workflow using flexible, opensource tools.

KEYWORDS

groundwater age, discretization, predictive uncertainty, model structure, iterative ensemble smoother, particle tracking, MODFLOW, PEST++

1 Introduction

Groundwater accounts for approximately 97% of all accessible fresh water, supplies drinking water for nearly half the world's population, and accounts for 43% of the global water consumption for agriculture (Siebert et al., 2010; Guppy et al., 2018). Physically based numerical models (as opposed to data-driven models such as are used in Ruidas et al., 2021. or Jaydhar et al., 2022), combined with subsurface properties inferred from sparse observations can help extend our understanding of groundwater systems (e.g., Singh, 2014), providing an essential tool to help inform resource management decisions

(Jakeman et al., 2016). However, all models require simplification of real-world properties and processes. Identifying the appropriate level of simplification for modelling groundwater systems remains challenging. Appropriate simplification depends on the intended use of the model (Watson et al., 2013; Guthke, 2017; White, 2017). We explore this important issue in the context of simulating groundwater age at a national scale across Aotearoa/New Zealand, to inform national water management policy. Note, the objective of the study presented here is not to provide definitive maps for groundwater age across Aotearoa/New Zealand, but rather to explore and highlight the implications of model and methodological simplification on groundwater age predictions at large scale.

Groundwater age provides a convenient method for evaluating the potential for groundwater recharge and hence contamination from recent sources (Sanford 2011; Morgenstern et al., 2015). The utility of decision support models based on groundwater age, where "young" groundwater suggests a potential groundwater contamination risk and "old" groundwater suggests a smaller component of modern recharge, would clearly be compromised by the presence of model structural errors that bias simulated groundwater age (e.g., Knowling et al., 2020). This study reveals that predictions of groundwater age can be biased by the inability to represent parameter complexity with simplified (upscaled) layering. Due to the relationship between flow depth and groundwater age, where deeply circulating water is generally older, the range of ages impacted depends on the depth of these structural simplifications.

Increased and wide-spread human impacts on climate and natural resources can warrant national government consideration and oversight of environmental processes and resource management activities over larger spatial extents, often in data-scarce areas (e.g., Regan et al., 2019). Maintaining national oversite of the effectiveness of policy requires an understanding of the broad range of natural processes and resource management activities that affect water resources extending from the mountains to the sea. This understanding also includes consideration of interactions between climate, ecosystems, lakes, rivers, aquifers, land use, land management, and water allocation.

However, the desire for models with continuous coverage over large spatial scales presents several modelling challenges: 1) trade-offs between model resolution and computational burden, 2) upscaling of hydraulic properties to a representative elemental volume (REV; the volume within which properties are assumed to be constant to facilitate numerical modelling), 3) representation of local processes over a large REV (e.g., upscaling stream-aquifer interactions), 4) representation of high variations in permeability (e.g., bedrock–aquifer contacts which typically form model boundaries in "traditional" groundwater models), 5) large changes in topography (e.g., Southern Alps rising 3,700 km from sea level over 30 km and/or deeply incised streams), and 6) limited subsurface data makes characterization of the groundwater system difficult, especially in areas with complex topography and geology like Aotearoa/New Zealand. We explore these modelling challenges within this paper.

2 Background

2.1 Model structure and parameterization challenges

One of the most fundamental techniques for simplifying processes and properties in numerical groundwater models is the subdivision of the model domain into discrete volumes with representative properties (REV). This requires heterogeneous and potentially scale dependent properties (e.g., hydraulic conductivity, porosity) within each REV to be represented by a single value in each cell. Also, complex processes (e.g., stream-aquifer interactions) need to be conceptualized and simplified in a way that allows them to be effectively represented over the entire cell.

The choice of model discretization provides the underlying structure to support the parameter representation (parameterization) of hydraulic properties. It also imposes a limit on the level of parameterization a numerical groundwater model can accommodate for history matching and predictions. Coarse discretization can reduce the computational burden and may ease the parameter estimation and inversion process, but it also increases the potential for structural deficiencies caused by homogenising processes and properties over larger areas which can bias model results (e.g., Wildemeersch et al., 2014; Knowling et al., 2019).

Doherty and Moore (2021) discuss how the model structure, and the accompanying parameterization approach, need not be more detailed than is required to make the prediction of interest, despite resulting in a more abstract (less "realistic") representation of hydraulic properties. On the other hand, parameter compensation resulting from deficiencies in structural and/or parameterization detail may impose bias in predictions, especially if those predictions are significantly different than data used for history matching (Doherty and Welter, 2010; Doherty and Christensen, 2011; White et al., 2014; Doherty, 2015). White et al. (2019a) explored the impact of truncating the vertical representation of a regional groundwater system, by comparing a 7-layer representation of a regional aquifer system, with truncated 2- and 4-layer representations. Knowling et al. (2020), showed that the inappropriate vertical truncation limited the ability of that model to assimilate information in tritium data, imposing a history matching induced parameter and predictive bias.

None of the previous work investigating the impact of model discretization on prediction uncertainty has specifically isolated

the influence of vertical discretization/layering while keeping all other factors the same (e.g., aquifer thickness). This research specifically focusses on issues associated with using simplified vertical discretisation approaches to represent complex parameter fields and its impact on the uncertainty of groundwater age model predictions after history matching. We use a paired complex-simple model methodology to explore the propensity for bias using various vertical discretisation structures (Doherty and Christensen, 2011; White et al., 2019a; Gosses and Wöhling, 2019; Knowling et al., 2019).

In this study groundwater flow is simulated using MODFLOW and advective transport, used as a surrogate for age, is simulated using MODPATH (particle tracking). We show that the inability of the coarse discretization to represent the appropriate level of heterogeneity during the history matching process results in model bias when compared to more refined discretization schemes.

2.2 History matching challenges

Highly parameterized approaches to model inversion can provide the flexibility to match observations but can also incur a large computational cost when using finite difference methods, which require one model run per adjustable parameter to fill a sensitivity matrix (Jacobian). Here instead we use the iterative ensemble smoother (IES; Chen and Oliver, 2013) method as implemented in the PEST++ suite (White, 2018). The IES method calculates an empirical Jacobian based on an ensemble of stochastic realizations. The number of realizations in the ensemble is generally much less than the number of adjustable parameters, resulting in significant gains in computational efficiency (e.g., Hunt et al., 2021).

The size of the IES ensemble should reflect the dimensionality of the solution space (i.e., the extent to which history matching targets inform various parameters), and therefore it is problem dependent. Spurious correlations can compromise parameter upgrade calculations when the ensemble size is small compared to the number of independent observations that span the solution space. Determining the appropriate ensemble size is challenging in that it depends on the relationship between the history matching dataset, and the representation of relevant real-world detail in the model (e.g., discretization or resolution of the computational grid), the predictions of interest, and the scale of the processes being simulated. Systematic explorations of this issue appear to be absent in the literature.

This research explores the size of the stochastic ensembles used for history matching in IES. Smaller ensembles combined with simplified model structures compromise the predictive ability of the calibrated model, despite a simple, synthetic dataset used for history matching. In some cases, this compromise is exacerbated as a better fit to the calibration dataset is sought through more iterations. The automatic adaptive localization (Luo et al., 2018) option implemented in PEST++ is shown to improve history matching and prediction.

2.3 Research objectives

In Aotearoa/New Zealand groundwater accounts for nearly 70% of consented freshwater takes and supplies approximately 30% of the population with drinking water (White, 2001; Rajanayaka et al., 2010). Land use changes over the last 40-year have resulted in increased groundwater contamination (e.g., nitrogen, pathogens, etc) prompting a national scale evaluation of groundwater resources and threats (Ministry for the Environment and Stats, 2021). In responses to these changes the National Policy Statement for Freshwater Management in New Zealand (NPSFWM) calls for the management of freshwater in a way that gives effect to Te Mana o te Wai ("the fundamental importance of water and the recognition that protecting the health of freshwater protects the health and well-being of the wider environment"; Ministry for the Environment, 2020).

We present a series of national scale models (approximately 268,000 km^2) that simulate groundwater flow and groundwater age (derived from particle tracking), embracing the extensive nature of New Zealand's NPSFWM. These models use the best available nationwide data and estimates of uncertainty for groundwater recharge, hydrogeology, and the location of stream networks. This model represents the spatially continuous groundwater system in Aotearoa/New Zealand and a consistent starting point for the development of regional or local scale models that may include more detailed representation of the processes of interest. However, the complexity of the natural world and the spatial extent of this model require significant abstraction and simplification of many processes. This simplification is necessary to ensure numerical stability and reasonable simulation times that enable history matching and inversion. This study specifically investigates the uncertainty and bias imposed by simplification of model layering on predictions of groundwater age.

3 Methods

The models presented herein are designed to evaluate: 1) the effects of different vertical discretization approaches on simulations of particle travel times in large scale groundwater models and 2) the effects of ensemble size on the ability of the model to match predictions. The model calculates particle travel times from a surface water source (i.e., stream or rainfall recharge) to an observation location *via* backward particle tracking. We use these particle travel times as an estimate of groundwater age, which in turn can be used to infer the potential

TABLE 1 List of characteristics resulting from the layering approach and parameterization for each model, including simulation times and times for parameter upgrades using automatic adaptive localization (AAL).

	Description	Complex	Fine	Even
North Island	Active cells	218,426	111,592	111,068
	Number of Parameters	37,087	16,769	16,753
	Simulation (minutes)	7.2	3.7	3.9
	AAL upgrade (minutes)	27.2	5.4	5
South Island	Active cells	321,271	148,597	148,048
	Number of Parameters	61,442	25,433	25,379
	Simulation (minutes)	14.8	7.0	7.5
	AAL upgrade (minutes)	81.5	17.8	17.1

susceptibility of groundwater to contamination from recent surface sources (Stauffer et al., 2005) and estimate sustainable groundwater recharge rates (McMahon et al., 2011). Vertical discretization is explored using three layering schemes: up to 50 evenly spaced layers ("complex" model), up to four layers with fine discretization in the upper layers and a single layer at depth ("fine" model), and up to four layers with evenly spaced layers at depth ("even" model). See below for detailed descriptions.

The IES method is used to match model outputs to a synthetic dataset (i.e., "truth") using models with alternative vertical discretization. One realization is chosen from a 300-member ensemble of the complex model to serve as the truth, based on the minimum sum of squared differences between the realization age at each location (age_i) and the simulated mean age at each observation location ($\overline{age_i}$):

$$\min \sum_i^n \left(age_i - \overline{age_i}\right)^2 \quad (1)$$

where n is the number of observation locations in gravel and sand. The realization chosen to represent the truth is removed from the parameter ensembles used for history matching.

We consider a failure or conflict to occur if the true value of an observation (plus or minus a representative measurement noise) falls outside the range of the simulated observation ensemble. The percentage of locations for which the model fails to capture the truth (Pf) is the ratio of the number of observation (parameter) values that fail to total number of observations (parameters), times 100. This is the same approach used to identify prior data conflict (PDC) in PEST++ and requires no assumptions about the shape of the posterior probability density function (PDF). More thorough analysis of the PDFs and more precise statistical tests are warranted to determine criteria for model failure in real world applications with specific management objectives.

Simulating groundwater age older than the true age ("overestimation") represents a failure of the model in a management context when groundwater age is used as a proxy for potential contamination from recent sources. Conversely, simulating groundwater age younger than the true age ("underestimation") represents a failure of the model in a management context when groundwater age is used as an indicator for the presence of modern recharge, leading to an overestimate of sustainable aquifer yield and potential for groundwater contamination. Underestimation and overestimation Pf generally follow the same trend (see Supplementary Material; "SM"). We report total Pf for observations used in history matching, predictions, and parameters for each model structure–ensemble size combination. Details for observations, predictions, and parameters are described below.

3.1 Models

Groundwater models often have finer discretization near the surface and coarser discretization at depth, reflecting the availability of data and the desire to represent important surface boundary conditions (e.g., surface water-groundwater interactions, recharge, etc) while still meeting reasonable computation requirements. Coarse discretization reduces the ability of the model to represent heterogeneity and more complex flow paths, potentially affecting simulated groundwater ages. We isolate the influence of vertical discretization on mean age by presenting a series of equivalent models where only the vertical discretization, and the parameterization supported by that discretization, is changed.

As noted above, three versions of a steady-state groundwater flow and particle tracking model of developed using MODFLOW v6.2.2 (Langevin et al., 2021) are presented in this study. MODPATH v7.2.002 (provisional at the time of writing) was used for all particle tracking simulations. Each version of the model is produced with the same scripts and

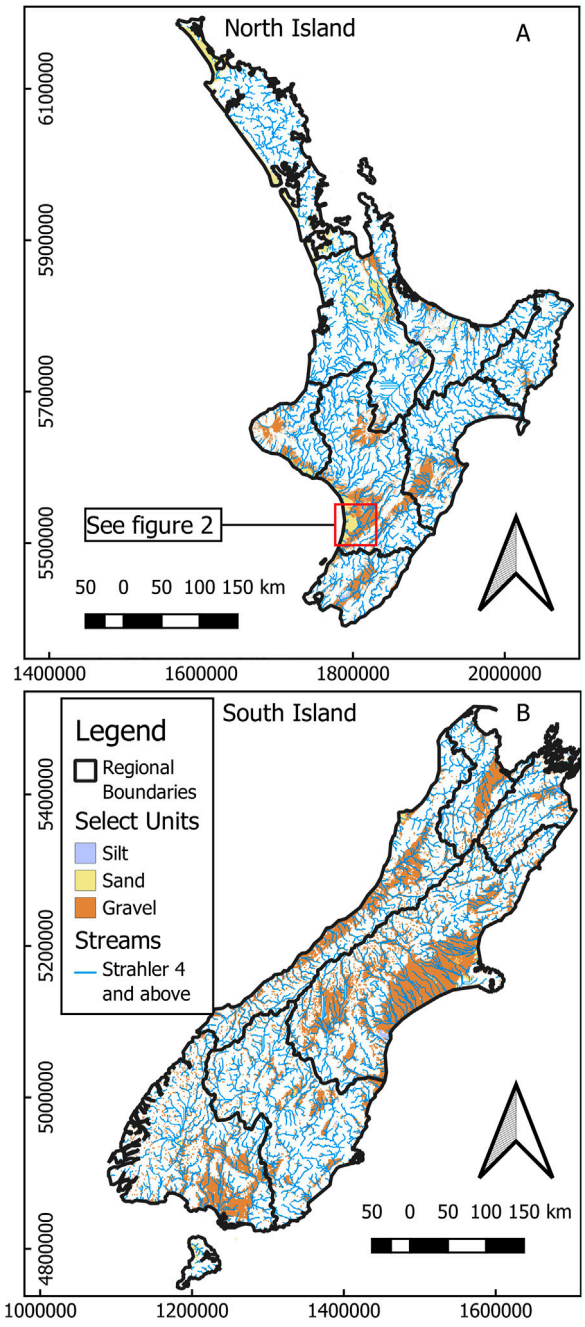

FIGURE 1
Map of the North Island **(A)** and South Island **(B)** of Aotearoa/New Zealand showing Strahler order four streams and above, regional boundaries, and hydrogeologic units of interest in this study (silt, sand, and gravel). Regional boundaries are used for manual localization in the IES method. Hydrogeologic units of interest are used to determine locations for history matching observations and predictions. See Figure 2 for a plan view of model specifics in the area shown by the red box.

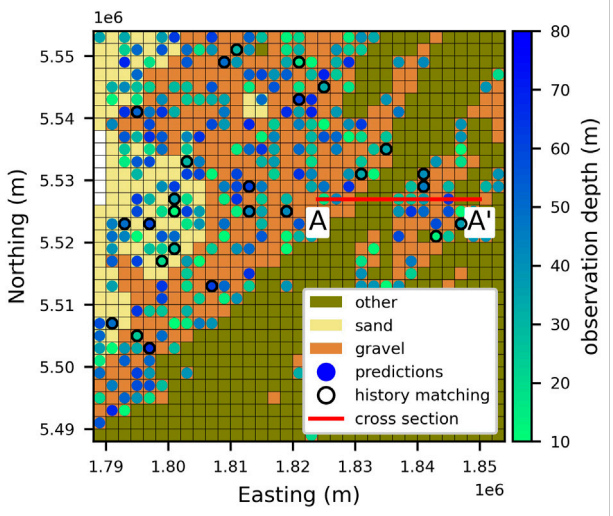

FIGURE 2
Detailed portion of the Aotearoa/New Zealand national groundwater model showing regional boundaries, hydrogeologic units, and observation locations (circles) used for history matching (black boarder) and predictions (no boarder). Observation depths are show by the colour bar. The red line **(A,A')** shows the location of the cross section shown in Figure 3.

underlying data. The number of active cells in the MODFLOW domain and the number of parameters for each model are reported in Table 1. The model domain and underlying data are based on Aotearoa/New Zealand. However this study is designed to explore the trade-offs between model simplification and predictive ability in the context of history matching large scale models to age tracer data, rather than reproduce real-world observations. We use synthetic data generated by the complex model in order to isolate vertical discretisation simplification errors from other sources of error inherent in real-world data (e.g., model conceptualization, measurement). The high number of parameters, wide prior parameter distributions, and flexible boundary conditions ensure a statistically feasible representation of the real-world system. The results presented in this study reveal important considerations for future history matching efforts using real-world data.

The open-source python package FloPy 3.3.5 (Bakker et al., 2021) was used to construct most of the MODFLOW input files. The Surface Water Network tool (SWN; Toews and Hemmings, 2019) was used to generate inputs for the Streamflow Routing Package (SFR2; Niswonger and Prudic, 2005) in MODFLOW. The PstFrom class (White et al., 2021) in the python package pyEMU (White et al., 2016) was used to ensure a consistent approach to representing adjustable parameters, observations, and predictions between the various models (see "Parameterization" section below and Supplementary Material). Additional python package libraries including NumPy, Pandas, and SciPy were used to pre-process data and post-process model results.

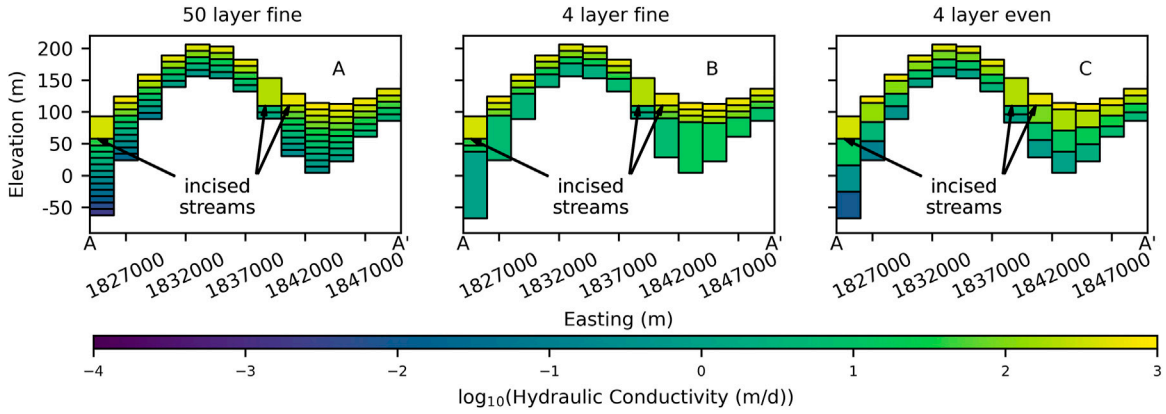

FIGURE 3
Cross-sections of A-A′ shown in Figure 2 illustrating discretization and upscaled hydraulic conductivity values (log(K)) for the **(A)** complex model, **(B)** fine model, **(C)** even model.

3.2 Discretization

Each of the model vertical structures explored in this study represents the same subsurface domain (horizontal and vertical extents) with a horizontal discretization of 2 km (Figures 1, 2). The specific depth and thickness of each layer is dependent on the spatially distributed depth to hydrogeologic basement (DHGB) as described in Westerhoff et al. (2019) and the layering scheme (Figure 3). A minimum layer thickness of 10 m and a minimum model thickness of 50 m is enforced for all models. The bottom of the top layer in all models is nominally 10 m below the surface. Routines in the SWN package that ensure stream reach elevations progress downstream from high elevation to low elevation can result in stream bed elevation being significantly lower than surface elevation, especially in steep terrain. While this is reflective of the often deep incision of streams in many parts of Aotearoa/New Zealand, it may require that the bottom of the surface layer is shifted down to accommodate the stream. The top of the model is unchanged to honour the elevation data, resulting in a thicker upper layer where streams are deeply incised (Figure 3).

The vertical model structure with a constant vertical discretization of 10 m and up to 50 layers is used as the complex version of the system (Figure 3A; "complex" model). The actual number of layers depends on the depth of the model (DHGB) and any adjustments to the top layer needed to accommodate incised channels. Two additional layering approaches are investigated: 1) a four-layer model with three thin (nominally 10 m) upper layers and one deeper layer (Figure 3B; "fine" model), and 2) a four-layer model with three evenly distributed deeper layers (Figure 3C; "even" model).

3.3 Boundary conditions

Surface water sources in our models are either distributed recharge along the top surface of the model representing rainfall recharge or losing streams. In this study we use the Streamflow Routing Package (SFR2) which provides a more realistic and flexible way to simulate streamflow than other packages. For example, in the RIV package cells with a river boundary condition essentially act as a general head boundary when the groundwater head falls below the bottom of the streambed. This can lead to higher groundwater recharge compared to SFR2 (e.g., Foglia et al., 2018), creating higher gradients near streams, and incorrect simulation of streams as sources. The input data for the SFR2 package is generated for Strahler order four and above streams contained in the River Environment Classification database from the National Institute of Water and Atmospheric Research (National Institute of Water and Atmospheric Research, 2019) using the Surface Water Network (SWN) tool developed by the Institute of Geological and Nuclear Sciences (GNS; Toews and Hemmings, 2019).

Spatially distributed recharge from the nationwide model of groundwater recharge for Aotearoa/New Zealand (NGRM; Westerhoff et al., 2018) is added to the model using the RCHA package. The NGRM model considers the effects of precipitation, evapotranspiration, vegetation, topography, soils, and geology on groundwater recharge. However, overland flow due to saturation from below (i.e., Dunnian flow) is not considered in the NGRM model because the groundwater flow system is not well represented. Dunnian flow is simulated in our models by applying a head dependent flux boundary condition to the upper surface of the model using the drain package (DRN) and routing groundwater discharge to the surface or rejected recharge in cells where groundwater reaches the surface to the nearest SFR segment using the mover package

(MVR). This also prevents unrealistically high groundwater heads in areas of high recharge and low conductivity. The general head boundary package (GHB) is used to represent the edge of the active model domain at the coast. See the MODFLOW documentation (Langevin et al., 2021) for a detailed description of these packages.

3.4 Age simulations

Locations of history matching targets and predictions (i.e., weighted and unweighted observations, respectively) in this study are limited to areas mapped as sand or gravel in the model, the materials that make up the most extensive and productive aquifers in Aotearoa/New Zealand (White P. A. et al., 2019). To avoid potential boundary effects, coastal boundary cells were excluded from the observation dataset. A total of 6,970 locations mapped as sand or gravel are randomly selected as observation locations: 2,056 for the North Island and 4,914 for the South Island, reflecting the relative abundance of sand and gravel aquifers on each island (Figure 1). The distribution of observations also reflects the relative abundance of these aquifer materials within each region. The mean depth of each observation is selected from a random distribution between the bottom of the top layer and either 80 m below the surface or 10 m above the bottom of the model, whichever is shallower. Observation locations were limited to 10 m above the bottom of the model to avoid potential stagnant conditions along the bottom of the model. The distribution of observation locations per layer for each model is listed in Supplementary Material. Each observation point is populated with 100 particles evenly distributed along the surface of a cylinder with a radius of 10 m and a height of 2 m. Particles are tracked from the observation location to the source, as determined by the steady-state flow field and the IFACE parameter in MODPATH (Pollock, 2012; see below).

Mean age is calculated from the travel times of particles originating from each location described above. The IFACE parameter in MODPATH specifies which cell face is considered the source for each boundary cell. For upper boundary cells in the RCH package the IFACE parameter is set to six indicating the source (i.e., zero age) is at the top of the cell. Abrams et al. (2013) showed that travel times to weak sink streams in a simple 1-layer model can be accurately simulated if the bottom of the stream channel is aligned with the top of the model and IFACE is set to 6 (i.e., the top face of the cell). However, in the current models where a stream may be incised hundreds of meters below the surface elevation, using the top of an SFR cell as the source can result in ages over 100 years higher than if the bottom of the cell is used (i.e., 0.5 m below the stream bed). Therefore, we set the IFACE parameter to 0 for all cells with SFR segment, indicating the source (i.e., zero age) is along the face of the boundary cell that is first intersected by the particle path during the backward particle tracking simulation.

3.5 Parameterization

The prior values of horizontal hydraulic conductivity (Kh), vertical hydraulic conductivity (K33), streambed hydraulic conductivity (Ksb), drain conductance (Cd), GHB conductance (Cghb), and porosity (ϕ) for all models are assigned consistently based on the main rock type in QMAP (GNS Science, 2012) and representative values found in the literature. The surface geology is assumed to extend to the DHGB (Westerhoff et al., 2019), except for units mapped as silt. Silt deposits are assumed to be 10 m thick and overly gravels with thickness determined by adjacent deposits. The hydraulic conductivity and porosity of the gravels overlain by silt is reduced by 10%. The hydraulic conductivity and porosity of all materials decrease as an exponential function of depth following Westerhoff et al. (2018).

Uncertainty for all parameters is addressed using parameter multipliers over four spatial scales (two scales geostatisical interpolation, zone multipliers, and layer multipliers; see Supplementary Material). Model inputs are the product of the multipliers and the "native" values. The initial value of all multipliers is one. The limits of each multiplier are reported in Table 2. The large range in parameter values accommodates: 1) the potential for inaccuracies in the mapping QMAP hydrofacies to the model grid, 2) the large uncertainty in hydraulic conductivity values for geologic materials (e.g., Domenico and Schwartz, 1998), and 3) the potential for parameters taking on physically unrealistic values to accommodate structural defects in the model, including due to averaging properties to accommodate different discretization approaches. Further details of model parameterization can be found in the Supplementary Material.

The PstFrom utility in the pyEMU package is used to create the interface and input files for the PEST++ suite and generate the initial parameter ensemble. The parameter ranges are used to define a wide prior parameter distribution, representing ±3σ. The PstFrom.draw() method in pyEMU is used to draw an ensemble of stochastic parameter vectors (realizations) assuming multi-variate Gaussian distributions. We limit parameter values to physically realistic and numerically stable values by enforcing an "ultimate" upper and lower bound for "native" parameter values *via* the PstFrom utility (Table 2).

3.6 History matching

Simulated ages from a single stochastic realization of the complex model is used to define a set of observations representing the target values (i.e., "truth"). This dataset is free from real-world complication such as measurement noise and transience. Since the data were generated by the complex model, the complex model is endowed with precisely the appropriate parameter complexity to reproduce the results. This end-member

TABLE 2 Values of multiplier parameters, potential for combined multipliers, and native value bounds enforced for each parameter group.

Name	Parameters	Multipliers			Potential Combined		Native value Bounds		
		Initial	Max	Min	Max	Min	Max	Min	Units
Conductivity	Kh, Ksb, K33	1	10	0.1	10,000	1.0E-04	2000	1.E-10	m d-1
Conductance	Cd, Cghb	1	10	0.1	10,000	1.0E-04	2000	1.E-10	m2 d-1
Porosity	ϕ	1	3	0.3	81	0.012	0.3	1.E-10	-
Recharge	Rp	1	1.5	0.5	2.3	0.25	0.008	1.E-10	m d-1

case is compared to the simplified models to isolate the impacts of coarse vertical discretization. History matching to this data using the complex model shows how small ensembles can bias predictions, despite using a structurally perfect model.

A random sample of approximately 10% of the observations on each island are assigned a weight of one during the history matching process (204 out of 2,056 for the North Island, 504 out of 4,914 for the South Island). The other 90% (1,852 and 4,410, respectively) are retained as predictions with zero weights. This allows us to evaluate the implications of model simplification and the associated parameterization on model predictions following history matching. The number observations used for history matching in each layer of each model is listed in Supplementary Material.

The IES method, as implemented in the PEST++ suite, is used for history matching (White 2018; White et al., 2020; Welter et al., 2015). The IES method uses an empirical Jacobian matrix calculated using cross-covariances between ensembles of stochastic realizations of parameter vectors and simulated equivalents of historical observations constituting the history matching dataset. Too few realizations in the ensemble, compared to the span of the observations which determine the dimensions of the solution space, can cause spurious correlations. These spurious correlations for infeasible or impossible parameter-observation relationships can be "zeroed out" using localization (see below). The history matching process in IES can be further improved by using more realizations than the dimensionality of the calibration solution space to increase the rank of the empirical Jacobian.

Methods exist for estimating the dimensionality of the solution space using a high-fidelity, perturbation-based Jacobian (e.g., Doherty and Hunt, 2009). However, we are not aware of a similar method for estimating the solution space using an empirical Jacobian. Practitioners typically use ensembles of 50–150 realizations for parameter estimation. Hunt et al. (2021) use 300 realizations for a parameter estimation problem with 1,777 adjustable parameters and a diverse set of approximately 30,000 history matching targets to "ensure the solution space was fully represented and results were free from adverse effects of ensemble collapse." Here we test the effects of using ensembles of 50, 100, 150, 200, and 300 realizations for history matching to a relatively simple dataset using each of the three model variations.

By default, PEST++ identifies prior data conflict (PDC) for weighted observation when the ensemble of observation values plus noise does not cover the ensemble of simulated values using the prior parameter ensemble. Observations with PDC are likely to cause bias as the history matching process seeks extreme parameter values to satisfy those observations. In this study, we retain observations with PDC in order to explore the potential impact on model predictions.

3.7 Localization

Localization masks spurious correlations between parameters and observations that can result from the use of a low-order ensemble. In this study, localization is initially based on groups defined by the 16 regions in Aotearoa/New Zealand, the boundaries of which typically follow major watershed boundaries. This groupwise localization scheme breaks correlations established between parameters in one region and observations in another. Zone and layer multiplier parameters are not included in this level of localization, meaning observations on a given island can influence zone and layer multipliers anywhere on that island. This groupwise localisation is very efficiently defined and implemented within PEST++ (White et al., 2021).

Localization also has the effect of increasing the rank of the empirical Jacobian used in the IES scheme, beyond that set by the size of the ensemble. Hence localization can mitigate the effects of truncation of the solution space if the ensemble size is too small. An alternative and automated localisation scheme can also be implemented in PEST++ using "automatic adaptive localization" (AAL; Luo, et al., 2018; White et al., 2021). AAL attempts to identify and mask spurious parameter-observation correlations generated by the stochastic nature of the ensembles for every parameter and observation pair. This process of localization results in a highly disjointed Jacobian matrix requiring numerous "local" parameter upgrade solves, which can become numerically expensive (Table 1). We explore the effectiveness of AAL using the lowest order ensemble (50 realizations).

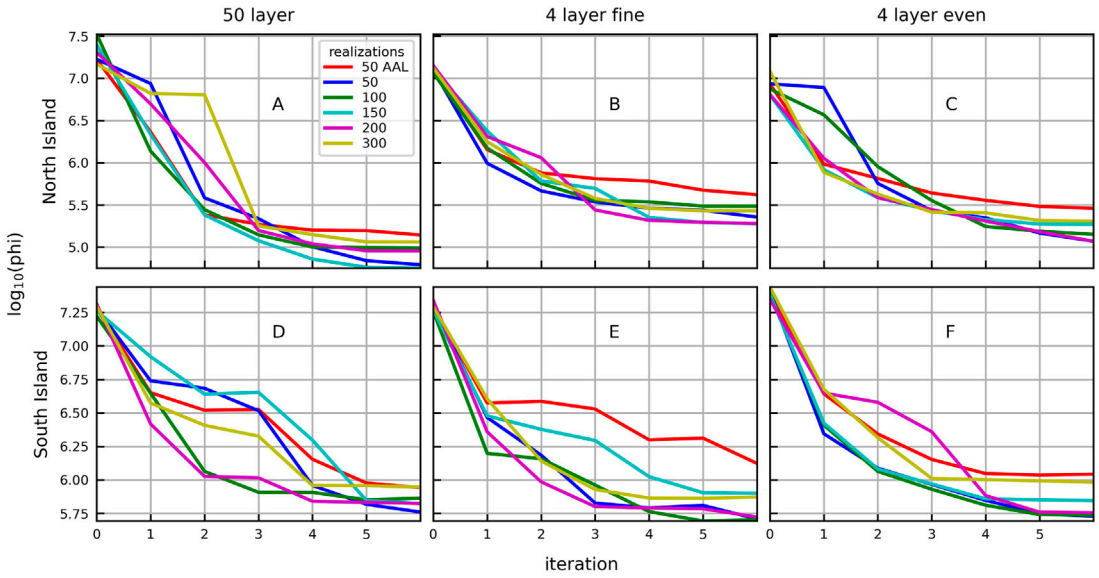

FIGURE 4
Plots of mean phi with iteration for the complex models **(A,D)**, simplified fine models **(B,E)**, and simplified even models **(C,F)** of the North Island **(A–C)** shown in Figure 3 (columns). Different ensemble sizes used in the IES history matching are shown by the colours.

4 Results

4.1 Mean phi

In general, model to measurement fits, as summarised by the mean objective function (phi or the L2-norm), decrease rapidly within the first three iterations (Figure 4). All three model vertical discretisation approaches display significant reductions in mean phi for all ensemble sizes, with all reducing to the same order of magnitude over six IES iterations. Generally, the simplified models of the North Island (Figures 4B, C) do not achieve the same reduction in phi as the complex model (Figure 4A) after six IES iterations. The simplified models of the South Island (Figures 4E, F) achieve similar values of mean phi as the complex model (Figure 4D) after six IES iterations. Interestingly the lower order ensembles often achieve the lowest values of mean phi for all models. Other than this, there is no apparent relationship between the rate of decrease and the level of model simplification or the size of the ensemble. Instead, the ensemble size and the number of iterations needed to attain a particular value of mean phi depends on the system being modelled (e.g., North Island vs. South Island) and the size of the ensemble. This seems particularly true for the structurally 'perfect' complex models (i.e., iteration 2 in Figure 4A and iterations 2 and 3 in 4D).

The complex models contain the appropriate level of structural complexity and parameterization to adequately reproduce the calibration targets and predictions, since this 50-layer model was used to generate the 'truth' target observations using a single parameter vector chosen from the prior probability distribution. History matching using the complex models of the North Island results in a mean phi value that is lower than the simplified models after the third iteration, regardless of ensemble size (Figures 4A–C). However, this is not the case for the South Island. History matching of the South Island model using the 4-layer models produces mean phi values similar to, and occasionally lower than, the complex model after six iterations, depending on ensemble size (Figures 4D–F).

The ensembles with 50 realizations and AAL result in the highest mean phi (worst fit) after six iterations for all models. The ensemble with 300 realizations also results in a relatively high mean phi after six iterations for most of the models. Conversely, the ensemble with 50 realizations results in a relatively low mean phi after six iterations for most of the models.

4.2 History matching observations

The history matching targets are captured by the prior parameter ensemble for more than 92% of the weighted observation locations, for all ensemble sizes and all models (Pf < 8%; Figure 5 iteration 0). The complex model with 300 realizations performed the best in terms of history matching, with less than 0.5% failure for all iterations (Figures 5A, D). The highest prior failure (PDC) occurs with the 50-realization ensembles (2% < Pf < 8%; Figure 5, iteration 0), except for the complex model of the South Island (Pf = 0.2%; Figure 5D). The history matching process with 50 realizations and no AAL significantly increases the percentage of history matching

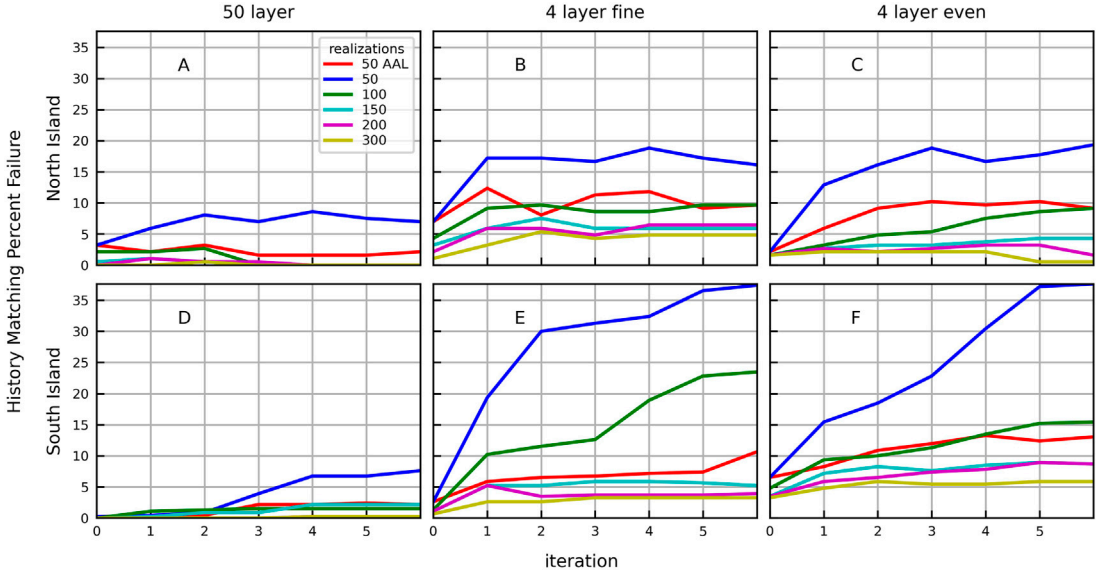

FIGURE 5
Percent failure (Pf) by iteration, ensemble size, and model structure for history matching targets (weighted observations) for the complex models **(A,D)**, simplified fine models **(B,E)**, and simplified even models **(C,F)** of the North Island **(A–C)** and South Island **(D–F)**. Ensemble sizes are indicated by colours.

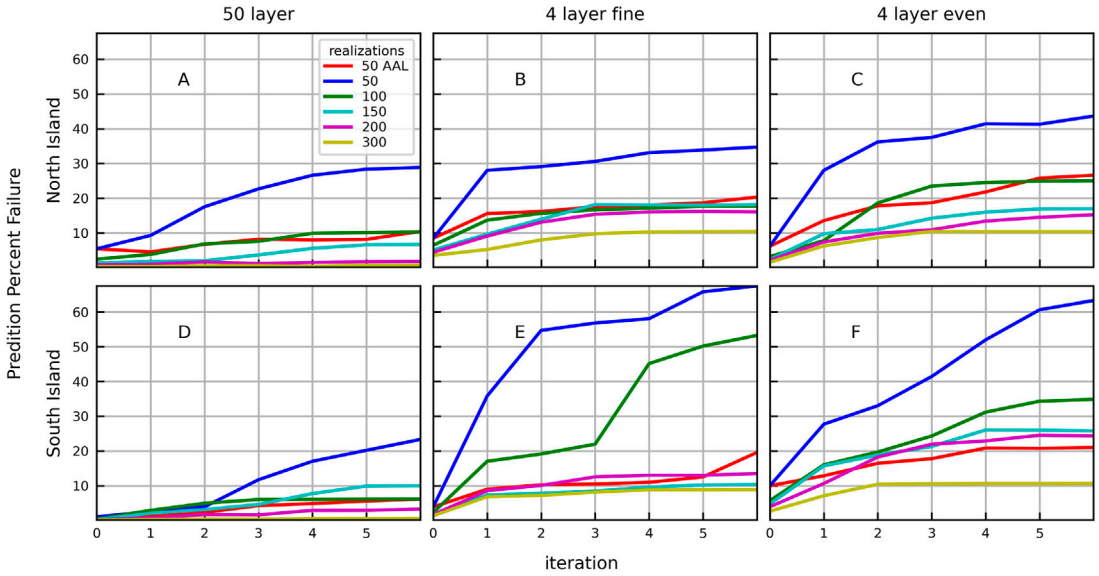

FIGURE 6
Percent failure (Pf) by iteration, ensemble size, and model structure for predictions (unweighted observations) for the complex models **(A,D)**, simplified fine models **(B,E)**, and simplified even models **(C,F)** of the North Island **(A–C)** and South Island **(D–F)**. Ensemble sizes are indicated by colours.

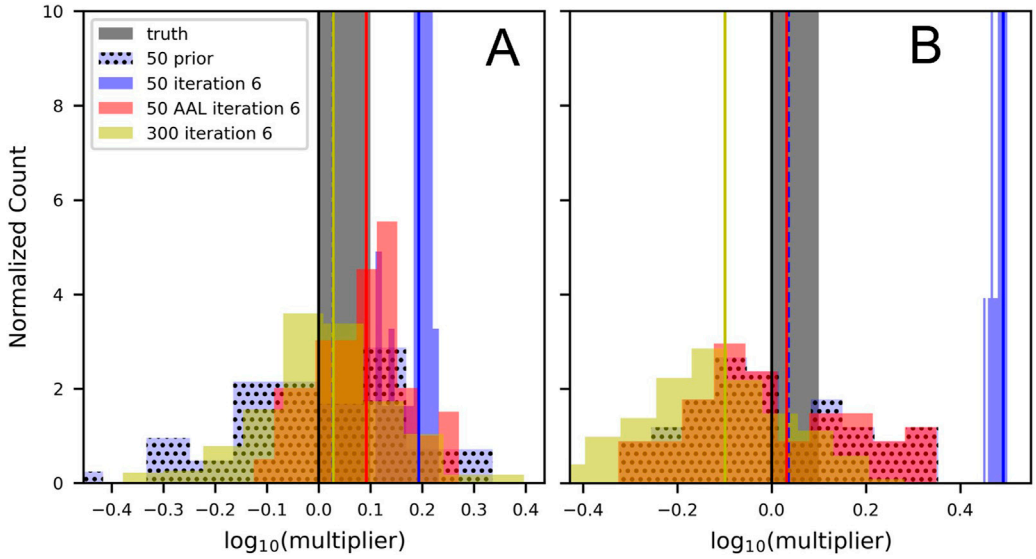

FIGURE 7
Probability density functions for hydraulic conductivity multipliers for the **(A)** complex model and **(B)** fine model of the South Island. The prior PDF for 50 realizations is shown in light blue with stipples. The mean of the posterior PDF for 50 realizations (blue), 50 realizations with AAL (red), and 300 realizations (yellow) ensembles are shown by vertical lines.

observation locations for which the model fails to capture the truth (Pf >15%), except for the structurally perfect complex models.

We can isolate the influence of structural errors from the history matching process by examining the percent failure at iteration 0 in Figure 5 for all model structures. This corresponds to prior data conflict (PDC) reported by PEST++. The PDC suggests structural defects in the fine model affect the North Island (Figure 5B) more than the even model (Figure 5C), while the opposite is true for the South Island (Figures 5E, 6F, respectively). The structural defects in the simplified models implied by the PDC are compounded through the history matching process, resulting in higher model failure rates with more iterations. This is particularly true using 50 realizations without AAL. The simplified models of the South Island with 50 realizations and no AAL have higher percentage of failure than the simplified models of the North Island for any given iteration, reaching 37.6% and 24.7% failure, respectively.

The ability to capture the true value of the history matching dataset is improved with AAL. The failure rate of the predictions after six iterations using the structurally perfect complex models using 50 realizations with no AAL is 7.6% for the North Island and 8.6% for the South Island; this is reduced to 3.2% and 2.4%, respectively, using AAL.

4.3 Predictions

The predictions are captured by the prior parameter ensemble for more than 86% of the locations, for all ensemble sizes and all models (Pf < 14%; Figure 6 iteration 0). The complex model with 300 realizations performed the best in terms of prediction (Figures 6A, D), with less than 0.7% failure for all iterations, with no systematic change in prediction failure rates over iterations. The history matching process significantly increases prediction failure for all other models, particularly when using 50 realizations without AAL. Similar to the history matching targets, the simplified models of the South Island (Figures 6E, F) with 50 realizations and no AAL tend to have a higher percentage of failure for predictions than the North Island (Figures 6B, C), reaching 43.6% and 63.3%, respectively.

The ability to capture the true value of predictions is improved with AAL. The failure rate of the predictions after six iterations using the structurally perfect complex models and 50 realizations with no AAL is 28.8% for the North Island and 23.3% for the South Island; this is reduced to 10.4% and 6.1%, respectively, using AAL.

4.4 Parameter estimation: Prior and posterior distributions

The three model structures presented here support different levels of parameterization at depth, making it difficult to make direct comparisons of individual parameter adjustments for each model during the history matching process. However, for a single model structure we can examine how parameter ensembles of different sizes

morph from prior to posterior through history matching in different ways.

Figure 7 provides an example of this for hydraulic conductivity multipliers from the complex (A) and fine (B) models of the South Island. After six iterations the history matching process with the complex model maintains a broad PDF using 300 realizations (Figure 7A yellow, 10-0.38–100.4) and 50 realizations with AAL (Figure 7A red, 10-0.12–100.28). These PDFs encompass the initial value of unity and hence still represent the initial value of hydraulic conductivity in native parameter space. However, using 50 realizations without AAL (Figure 7A solid blue) shows a narrower PDF that does not encompass the initial value of unity (100.11–100.24) despite achieving a lower value of phi. The posterior PDF does still fall within the prior PDF (Figure 7A, blue with stipples).

After six iterations the history matching process with the fine model shows similar behaviour using 300 realizations and 50 realizations with AAL (Figure 7B). However, the even the structurally complex model with a small ensemble without AAL shows a much narrower PDF that falls outside the prior PDF. This example illustrates how small parameter ensembles can result in collapse of the posterior parameter PDF. It also demonstrates the role of large ensembles and localization to prevent ensemble collapse. Figure 7 also shows how structural defects can corrupt the posterior PDF as parameters take on surrogate roles that accommodate for the missing parameters as model output to measurement matches are sought.

5 Discussion and conclusions

The numerical experiments described in this paper focus on the predictive performance implications of adopting structurally simple models and history matching with reduced ensemble sizes. The implications of the results of these experiments are considered in a decision support modelling context that relies on groundwater age simulations at a national scale. For this specific decision support context, we adopted a similar paired complex and simple model approach as documented in Doherty and Christensen (2011), Knowling et al. (2019), White et al. (2019a) and others. This method assesses the performance of simpler model structures in relation to a complex model structure. In synthetic experiments such as are documented in this paper, this complex model structure can represent the nominal "truth", for the purposes of the study.

5.1 Model to measurement fits and predictive performance

On the basis of model to measurement fits, as summarised by the mean objective function (phi), one might consider in some cases that the simplified models is as effective as the complex model for simulating the system, e.g., the South Island simplified models examples. At first glance it may also appear that low realisation numbers are more than sufficient for conditioning parameters to system observations. However, the higher prediction failure rates (Pf) of the simplified models are not consistent with how well the model was able to fit the data (as reflected by the associated mean phi values). Many of the configurations that produce the best fits, or lowest mean phi, also produce the highest prediction failure rate.

The implications of good fits being a poor indicator of good predictive performance are not well understood in the larger modelling community. The demonstration of this issue in this paper is consistent with the recent discussions in Hunt et al. (2021) and Doherty and Moore (2021) in a numerical physically based modelling context, and Ruidas et al. (2021) in a data-driven modelling context. The interplay of predictive performance with model structural errors and ensemble size is discussed below.

5.2 Model structural errors and predictive performance

The predictive failure results relating to the 4-layer models conflate structural deficiencies with those arising from inadequate ensemble size. However, because the 300-realisation ensemble can be inferred to span the solution space (Hunt et al., 2021; Doherty and Moore 2019), the predictive impact from structural deficiencies can be isolated when exploring the 300-realisation ensemble results. Prediction failure occurs because the simplified (coarse) vertical discretization inhibits the ability of the model to represent the hydraulic property heterogeneity that occurs with depth; this heterogeneity places controls on groundwater flow paths and hence groundwater age. This simpler structure therefore compromises the ability of the model to process information from the history matching observations to the model parameters in a way that adequately informs the predictions, as evidenced by the higher prediction failure rate of the simplified models compared to the complex model.

The predictive performance of simplified models and smaller ensemble sizes is problem specific, as illustrated by differences in the geological contexts of the two models; the North and South Island models. In general, the South Island has more extensive

and deeper gravel aquifers than the North Island. As such, parameters in the simplified models of the South Island represent parameters lumped over a greater depth interval. We observe that the greater the extent of parameter lumping (or upscaling), the greater structural related prediction errors will be, wherever predictions are sensitive to the hydraulic property detail that has been obscured by the lumping process.

5.3 Ensemble size, number of history matching iterations and predictive performance

History matching using smaller ensembles (particularly those without AAL) significantly increases failure for both history matching targets (phi) and predictions (Pf). This is evident for all versions of vertical model structures examined. These results again emphasise that good model to measurement fits are an insufficient criterion for predictive model efficacy, as described above.

The results also show that there is an increase in model predictive failure rate through history matching iterations for all discretisation versions, and across all ensemble sizes, except for the complex models with the 300 realizations ensemble, which remains below 1% over all iterations. This trend is especially evident in the simplified models and where the ensemble size is small. The 300-realization ensembles consistently provide the minimum predictive failure trend in all models, reaching a fairly constant value of 10% in the simplified models after the first few iterations. This indicates that the history matching process involving simpler models and/or inadequately sized ensembles, is forcing parameters to take on surrogate roles that can lead to parameter and predictive bias (Doherty and Moore 2019; Knowling et al., 2019).

This becomes clearer when examining the complex model with 300 realizations, for which we can assume that there is no structural model error, as the structure is the same as the 'truth model'. For this complex model, because history matching does not appear to incur any increase in predictive failure, we can also assume that the 300 realizations sufficiently span the solution space. Therefore, the history matching and predictive ability of the complex model presented is compromised only by rank deficient Jacobian matrices associated with smaller ensembles. This effectively allows us to isolate the impact of deficiencies in model structure from those resulting from the history matching implementation with a rank deficient Jacobian. These rank deficiency related errors result from the smaller ensembles and hence insufficient dimensions in parameter space to realistically convey predictive error, i.e., some parameter combinations that the observations and predictions are sensitive to are not well represented in the smaller ensembles.

Localisation methods can address this to varying extents by increasing the rank of the Jacobian matrices. The automatic adaptive localization (AAL) was demonstrated to reduce model failure, which becomes more extreme with smaller ensembles. This is as it should be as AAL provides a method for mitigating the impacts of adopting small ensembles to some extent, which is a compromise that is often made when model run times are larger as discussed in Chen and Oliver 2017. This mitigation is achieved by removing spurious correlations from the parameter update calculations. It is this process that helps to guard against failure to capture the true values of both history matching targets and predictions. However, it should be noted that using AAL can incur a significant computational cost due to the potentially disjointed Jacobian.

5.4 Implications for design of large-scale groundwater age models

Results in this study suggest simplified layering schemes appropriate for large, national scale models may produce adequate results, provided large enough ensembles are used. However, history matching with simplified models and small ensembles is likely to produce unacceptably high failure rates. Acceptable model simplifications and adequate ensemble size is problem specific, as illustrated by the difference between the North and South Island models. This study suggests a reasonable combination of model simplification and ensemble size may be identified by a stable failure rate of weighted observations with iteration, as seen with the 300-realization ensemble for all models presented. In contrast, increasing failure rate of weighted observations with iteration, as seen in the lower order ensembles and simplified models, suggests a concomitant increase in prediction failure rate.

Finally, we note that while other national groundwater models exist (Döll and Fiedler, 2008; De Lange et al., 2014), the development of a national groundwater age model, which to the authors knowledge is a world first in terms of scale, represents an extensive modelling effort. This type of development includes the running of numerous numerical experiments as part of the model design process, one of which is documented in this paper. The number of moving parts is enormous, and the cognitive load of a modeller is limited, and hence we believe that this effort would likely not be possible without adopting a scripted modelling workflow that spans task ranging from model discretization to highly parameterized inversion (Leaf and Fienen 2022). This workflow benefits enormously from the existence of opensource software packages and the community that contributes to their development and maintenance (Bakker et al., 2021; White et al., 2016; White et al., 2021).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

WK is responsible for the model construction and analysis of results. CM and WK are responsible for the experimental design. CM and BH provided important theoretical and practical input throughout the experiment. All authors contributed to the writing of this manuscript.

Funding

This work was funded by the New Zealand Ministry of Business Innovation and Employment (Grant No. C05X1803) and was also supported by GNS Science Groundwater Strategic Science Investment Fund (SSIF).

Acknowledgments

The authors would like to thank all those who support, develop, and contribute to the ecosystem of opensource software and tools that made this research possible.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.972305/full#supplementary-material

References

Abrams, D., Haitjema, H., and Kauffman, L. (2013). On modeling weak sinks in MODPATH. *Groundwater* 51 (4), 597–602.

Bakker, M., Post, V., Hughes, J. D., Langevin, C. D., White, J. T., Leaf, A. T., et al. (2021). *FloPy v3.3.5 — release candidate*. U.S. Geological Survey Software Release. doi:10.5066/F7BK19FH

Chen, Y., and Oliver, D. S. (2013). Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantificationcient history-matching and uncertainty quantification. *Comput. Geosci.* 17 (4), 689–703. doi:10.1007/s10596-013-9351-5

Chen, Y., and Oliver, D. S. (2017). Localization and regularization for iterative ensemble smoothers. *Comput. Geosci.* 21, 13–30. doi:10.1007/s10596-016-9599-7

De Lange, W. J., Prinsen, G. F., Hoogewoud, J. C., Veldhuizen, A. A., Verkaik, J., Essink, G. H. P. O., et al. (2014). An operational, multi-scale, multi-model system for consensus-based, integrated water management and policy analysis: The Netherlands hydrological instrument. *Environ. Model. Softw.* 59, 98–108. doi:10.1016/j.envsoft.2014.05.009

Doherty, J., and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resour. Res.* 47 (12), 12534. doi:10.1029/2011WR010763

Doherty, J. E. (2015). *PEST and its utility support software, Theory*. Watermark Numerical Publ.

Doherty, J., and Hunt, R. J. (2009). Two statistics for evaluating parameter identifiability and error reduction. *J. Hydrology* 366 (1-4), 119–127. doi:10.1016/j.jhydrol.2008.12.018

Doherty, J., and Moore, C. R. (2019). Decision support modeling: Data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater* 58, 327–337. doi:10.1111/gwat.12969

Doherty, J., and Moore, C. R. (2021). Decision-support modelling viewed through the lens of model complexity. Available at: https://gmdsi.org/blog/monograph-decision-support-modelling-viewed-through-the-lens-of-model-complexity/.

Doherty, J., and Welter, D. (2010). A short exploration of structural noise. *Water Resour. Res.* 46. doi:10.1029/2009WR008377

Döll, P., and Fiedler, K. (2008). Global-scale modeling of groundwater recharge. *Hydrol. Earth Syst. Sci.* 12, 863–885. doi:10.5194/hess-12-863-2008

Domenico, P. A., and Schwartz, F. W. (1998). *Physical and chemical hydrogeology*, 506. New York: Wiley.

Foglia, L., Neumann, J., Tolley, D. G., Orlo_, S. G., Snyder, R. L., and Harter, T. (2018). Modeling guides Groundwater management in a basin with river–aquifer interactions. *Calif. Agric. (Berkeley).* 72, 84–95. doi:10.3733/ca.2018a0011

GNS Science (2012). Qmap. Available at: https://www.gns.cri.nz/Home/Our-Science/Land-and-Marine-Geoscience/Regional-Geology/Geological-Maps/1-250-000-Geological-Map-of-New-Zealand-QMAP (accessed on January 3, 2017).

Gosses, M., and Wöhling, T. (2019). Simplification error analysis for groundwater predictions with reduced order models. *Adv. Water Resour.* 125, 41–56. doi:10.1016/j.advwatres.2019.01.006

Guppy, L., Uyttendaele, P., Villholth, K. G., and Smakhtin, V. (2018). *Groundwater and sustainable development goals: Analysis of interlinkages. UNU-INWEH report series, issue 04*. Hamilton, Canada: United Nations University Institute for Water, Environment and Health.

Guthke, A. (2017). Defensible model complexity: A call for data-based and goal-oriented model choice. *Groundwater* 55 (5), 646–650. doi:10.1111/gwat.12554

Hunt, R. J., Fienen, M. N., and White, J. T. (2020). Revisiting 'an exercise in groundwater model calibration and prediction' after 30 Years: Insights and New directions. *Groundwater* 58, 168–182. doi:10.1111/gwat.12907

Hunt, R. J., White, J. T., Duncan, L. L., Haugh, C. J., and Doherty, J. (2021). Evaluating lower computational burden approaches for calibration of large environmental models. *Groundwater* 59, 788–798. doi:10.1111/gwat.13106

Jakeman, A. J., Barreteau, O., Hunt, R. J., Rinaudo, J., Ross, A., Arshad, M., and Hamilton, S. (2016). "Integrated groundwater management: An overview of concepts and challenges," in *Integrated groundwater management: Concepts, approaches and challenges*. A. J. Jakeman, O. Barreteau, R. J. Hunt, et al. (Cham: Springer International Publishing), 3–20.

Jaydhar, A. K., Pal, S. C., Saha, A., Islam, A. R. M. T., and Ruidas, D. (2022). Hydrogeochemical evaluation and corresponding health risk from elevated arsenic and fluoride contamination in recurrent coastal multi-aquifers of eastern India. *J. Clean. Prod.* 369, 133150. doi:10.1016/j.jclepro.2022.133150

Knowling, M. J., White, J. T., Moore, C. R., Rakowski, P., and Hayley, K. (2020). On the assimilation of environmental tracer observations for model-based decision support. *Hydrol. Earth Syst. Sci.* 24, 1677–1689. doi:10.5194/hess-24-1677-2020

Knowling, M. J., White, J. T., and Moore, C. R. (2019). Role of model parameterization in risk-based decision support: An empirical exploration. *Adv. Water Resour.* 128, 59–73. doi:10.1016/j.advwatres.2019.04.010

Langevin, C. D., Hughes, J. D., Banta, E. R., Provost, A. M., Niswonger, R. G., and Panday, S. (2021). *MODFLOW 6 modular hydrologic model version 6.2.2*. U.S. Geological Survey Software Release. doi:10.5066/F76Q1VQV

Leaf, A. T., and Fienen, M. N. (2022). Modflow-setup: Robust automation of groundwater model construction. *Front. Earth Sci. (Lausanne).* 10. doi:10.3389/feart.2022.903965

Luo, X., Bhakta, T., and Naevdal, G. (2018). Correlation-based adaptive localization with applications to ensemble-based 4d seismic history-matching. *SPE J.* 23, 396–427. doi:10.2118/185936-pa

McMahon, P. B., Plummer, L. N., Bohlke, J. K., Shapiro, S. D., and Hinkle, S. R. (2011). A comparison of recharge rates in aquifers of the United States based on groundwater-age data. *Hydrogeol. J.* 19, 779–800. doi:10.1007/s10040-011-0722-5

Ministry for the Environment (2020). National policy statement for freshwater management. Available from https://environment.govt.nz/publications/national-policy-statement-for-freshwater-management-2020 (accessed Jan 13, 2022).

Ministry for the Environmentand Stats, N. Z. (2021). New Zealand's environmental reporting series: Our land 2021. Available from www.stats.govt.nz (accessed June 13, 2022).

Morgenstern, U., Daughney, C. J., Leonard, G., Gordon, D., Donath, F. M., and Reeves, R. (2015). Using groundwater age and hydrochemistry to understand sources and dynamics of nutrient contamination through the catchment into Lake Rotorua, New Zealand. *Hydrol. Earth Syst. Sci.* 19, 803–822. doi:10.5194/hess-19-803-2015

National Institute of Water and Atmospheric Research (2019). freshwater-and-estuaries/management-tools/river-environment-classification. Available from https://niwa.co.nz/freshwater-and-estuaries/management-tools/river-environment-classification-0 (accessed May 29, 2019).

Niswonger, R. G., and Prudic, D. E. (2005). Documentation of the streamflow-routing (SFR2) package to include unsaturated flow beneath streams—a modification to SFR1. *U.S. Geol. Surv.* 50.

Pollock, D. W. (2012). User guide for MODPATH version 6—a particle-tracking model for MODFLOW: U.S. *Geol. Surv. Tech. Methods* 6–A41, 58.

Rajanayaka, C., Donaggio, J., and McEwan, H. (2010). *Update of water allocation data and estimate of actual water use of consented takes 2009-10*. Ministry for theEnrivornment.

Regan, R. S., Juracek, K. E., Hay, L. E., Markstrom, S. L., Viger, R. J., Driscoll, J. M., et al. (2019). The US Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. *Environ. Model. Softw.* 111, 192–203.

Ruidas, D., Pal, S. C., Islam, A. R. M. T., and Saha, A. (2021). Characterization of groundwater potential zones in water-scarce hardrock regions using data driven model. *Environ. Earth Sci.* 80, 809. doi:10.1007/s12665-021-10116-8

Sanford, W. E. (2011). Calibration of models using groundwater age. *Hydrogeol. J.* 19, 13–16. doi:10.1007/s10040-010-0637-6

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., et al. (2010). Groundwater use for irrigation – A global inventory. *Hydrol. Earth Syst. Sci.* 14, 1863–1880. doi:10.5194/hess-14-1863-2010

Singh, A. (2014). Groundwater resources management through the applications of simulation modeling: A review. *Sci. total Environ.* v499, 414–423. doi:10.1016/j.scitotenv.2014.05.048

Stauffer, F., Guadagnini, A., Butler, A., Franssen, H. H., Wiel, N., Bakr, M., et al. (2005). Delineation of source protection zones using statistical methods. *Water Resour. manage.* 19, 163–185. doi:10.1007/s11269-005-3182-7

Toews, M. W., and Hemmings, B. (2019). "A surface water network method for generalising streams and rapid groundwater model development," in New Zealand Hydrological Society Conference, Rotorua, 3-6 December 2019, 166–169.

Watson, T. A., Doherty, J. E., and Christensen, S. (2013). Parameter and predictive outcomes of model simplification. *Water Resour. Res.* 49 (7), 3952–3977. doi:10.1002/wrcr.20145

Welter, D. E., White, J. T., Doherty, J. E., and Hunt, R. J. (2015). "PEST++ version 3, a parameter estimation and uncertainty analysis software suite optimized for large environmental models," in *U.S. Geological Survey Techniques and Methods Report* 7–C12, 54.

Westerhoff, R., Rawlinson, J., and Tschritter, C. (2019). New Zealand groundwater atlas: Depth to hydrogeological basement. Lower hutt (NZ). *GNS Sci.* 19

Westerhoff, R., White, P., and Rawlinson, Z. (2018). Incorporation of satellite data and uncertainty in a nationwide groundwater recharge model in New Zealand. *Remote Sens.* 10 (1), 58. doi:10.3390/rs10010058

White, J. T., Hemmings, B., Fienen, M. N., and Knowling, M. J. (2021). Towards improved environmental modeling outcomes: Enabling low-cost access to high-dimensional, geostatistical-based decision-support analyses. *Environ. Model. Softw.* 139, 105022. doi:10.1016/j.envsoft.2021.105022

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50 (2), 1152–1173. doi:10.1002/2013wr014767

White, J. T., Fienen, M. N., and Doherty, J. E. (2016). A python framework for environmental model uncertainty analysis. *Environ. Model. Softw.* 85, 217–228. doi:10.1016/j.envsoft.2016.08.017

White, J. T. (2017). Forecast first: An argument for groundwater modeling in reverse. *Groundwater* 55, 660–664. doi:10.1111/gwat.12558

White, J. T., Knowling, M. J., and Moore, C. R. (2019a). Consequences of groundwater-model vertical discretization in risk-based decision-making. *Ground Water* 58, 695–709. doi:10.1111/gwat.12957

White, P. A. (2001). "Groundwater resources in New Zealand," in *Groundwaters of New Zealand*. Editors M. R. Rosen and P. A. White (Wellington: New Zealand Hydrological Society), 45–75.

White, P. A., Moreau, M., Mourot, F., and Rawlinson, Z. J. (2019b). *New Zealand groundwater atlas:hydrogeological-unit map of New Zealand*. Lower Hutt (NZ): GNS Science, 88.

Wildemeersch, S., Goderniaux, P., Orban, Ph., Brouyére, S., and Dassargues, A. (2014). Assessing the effects of spatial discretization on large-scale flow model performance and prediction uncertainty. *J. Hydrology* 510, 10–25. doi:10.1016/j.jhydrol.2013.12.020

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/feart.2022.1079287

OPEN ACCESS

EDITED BY
Catherine Moore,
GNS Science, New Zealand

REVIEWED BY
Xavier Sanchez-Vila,
Universitat Politecnica de Catalunya, Spain
Erwan Gloaguen,
Université du Québec, Canada

*CORRESPONDENCE
Neil Manewell,
n.manewell@uq.net.au

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 25 October 2022
ACCEPTED 09 December 2022
PUBLISHED 04 January 2023

CITATION
Manewell N, Doherty J and Hayes P (2023), Spatial averaging implied in aquifer test interpretation: The meaning of estimated hydraulic properties. *Front. Earth Sci.* 10:1079287. doi: 10.3389/feart.2022.1079287

Spatial averaging implied in aquifer test interpretation: The meaning of estimated hydraulic properties

Neil Manewell[1]*, John Doherty[2] and Phil Hayes[3]

[1]Chemical Engineering, University of Queensland, Brisbane, QLD, Australia, [2]Watermark Numerical Computing, Brisbane, QLD, Australia, [3]Centre for Natural Gas, University of Queensland, Brisbane, QLD, Australia

Processing of aquifer test drawdowns to obtain estimates of transmissivity, and sometimes storativity, is an integral part of hydrogeological site investigations. Analysis of these data often relies on an assumption of hydraulic property uniformity. Aquifer properties are often estimated by fitting a Theis curve to measured drawdowns. Where an aquifer exhibits heterogeneity, quantities that are forthcoming from such analyses are assumed to represent spatially-averaged properties. However the nature of the averaging process, and the area over which averaging takes place are unknown. In this study we derive spatial averaging functions that link inferred hydraulic properties to real-world hydraulic properties. These functions employ Fréchet integrals derived by previous investigators that link observation well drawdowns to aquifer properties under an assumption of mild aquifer heterogeneity. It is shown that these hydraulic property spatial averaging functions are complex, especially at times that immediately follow the commencement of pumping. Furthermore, they cross hydraulic property boundaries, so that estimates of storativity can be contaminated by heterogeneities in real-world transmissivity, and *vice versa*. Because of its greater averaging area at later times, estimates of transmissivity are generally more immune to the effects of local hydraulic property heterogeneity than are those of storativity. They are therefore more reflective of broadscale real-world hydraulic properties, particularly those that prevail in areas that are removed from the immediate vicinity of the pumping and observation wells.

KEYWORDS

pumping tests, resolution matrix, sensitivity coefficient, Fréchet kernel, inversion, Theis equation, spatial averaging function

1 Introduction

Aquifer tests comprise an essential component of site characterisation studies. A well is pumped, often at a constant rate, for a certain amount of time. Drawdowns are measured in the pumped well and possibly in one or a number of observation wells. Local hydraulic properties are inferred from these drawdowns.

Interpretation of aquifer test data is generally based on a number of simplifying assumptions. In the simplest case, the pumping well is assumed to fully penetrate a confined aquifer. The aquifer is imagined to be homogeneous; groundwater flow that is induced by pumping is therefore presumed to be radial.

Under these circumstances, drawdown can be calculated using the Theis equation (Theis 1935). Back-calculation of aquifer transmissivity (T) and storativity (S) can therefore be achieved by finding values of T and S for which the Theis curve provides the best fit with observed drawdowns. This can be done manually, or it can be automated. Where drawdowns are measured in a number of observation wells, it is commonplace to subject each set of well-specific drawdowns to this kind of analysis. While values inferred for T and S may differ between wells, all of them are reported. Differences between them are taken as a measure of local hydraulic property heterogeneity.

The Theis assumption of hydraulic property homogeneity over the entire drawdown-affected area can rarely be justified. It is made in order to attain uniqueness of an inverse problem, and to permit use of a simplified forward model for computation of drawdown. It is presumed that solution of this simplified inverse problem yields values of T and S that are "representative" of the area in which drawdown has been induced.

A number of authors have inquired into the nature of the relationship between real and inferred hydraulic properties. These include Butler (1988), Butler (1990), Oliver (1990), Oliver (1993), Sánchez-Vila, et al. (1999), Leven and Dietrich (2006) and Copty et al. (2011). Most of these studies focussed on the relationship between drawdown in a pumped or observation well and hydraulic properties that characterise pumping-affected areas. Linearization of this relationship enables rapid evaluation of drawdown-to-parameter sensitivities. It is argued that greater sensitivity of drawdown to hydraulic properties that prevail in one area over those that prevail in another area implies that values of T and S that are inferred from these drawdowns are more reflective of properties in the former area than those in the latter area.

In this paper we extend the utility of linear analysis in order to derive equations that directly relate values for T and S that are forthcoming from Theis-based analysis of drawdowns to values of T and S that characterise an aquifer; the former are referred to as "apparent values" by Sanchez-Vila et al. (2006). The methodology that we employ can be readily extended to other aquifer test contexts where forward modelling of pumping-induced drawdown relies on fewer assumptions than those that are required by the Theis equation. However, linear analysis under Theis assumptions is rendered particularly easy by the availability of analytical formulae for calculation of drawdown-to-parameter sensitivities.

2 Theory

2.1 Fréchet kernels

Consider a pumping well situated at $(-a/2,\ 0)$ and an observation well at $(a/2,\ 0)$; they are separated by a distance a. At time zero, extraction of water begins at a rate of q_0. The situation is depicted in Figure 1.

Suppose that the medium which these wells penetrate is homogeneous, with a pervasive transmissivity of T_o and a pervasive storativity of S_o. Under these circumstances, drawdown s at the observation well can be calculated using the Theis equation:

$$s\left(\frac{a}{2},t\right)=\frac{q_0}{4\pi T_0}E_1\left(\frac{S_0a^2}{4T_0t}\right) \tag{1}$$

where E_1 is the exponential integral function.

Now suppose that the aquifer test host medium is not homogeneous, and that transmissivity and storativity are functions of location $\boldsymbol{x}$ i.e. $(x,\ y)$. We further suppose that heterogeneities in transmissivity and storativity can be viewed as perturbations of background T_o and S_o. We denote differences between actual and background transmissivity and storativity by T and S. That is:

$$T(\boldsymbol{x})=T_a(\boldsymbol{x})-T_0 \tag{2a}$$

$$S(\boldsymbol{x})=S_a(\boldsymbol{x})-S_0 \tag{2b}$$

where $T_a(\mathbf{x})$ and $S_a(\mathbf{x})$ are the actual values of transmissivity and storativity at location $\boldsymbol{x}$. If $T(\boldsymbol{x})$ and $S(\boldsymbol{x})$ are small, then the drawdown perturbation $h(t)$ at the pumping well arising from these hydraulic property perturbations can be formulated as a convolution integral as follows:

$$h(t)=\int_A T(\boldsymbol{x})F_T(\boldsymbol{x},\ t)d\boldsymbol{x}+\int_A S(\boldsymbol{x})F_S(\boldsymbol{x},t)d\boldsymbol{x} \tag{3}$$

The functions $F_T(\mathbf{x},t)$ and $F_S(\boldsymbol{x},t)$ comprise so-called Fréchet kernels for transmissivity and storativity respectively. Knight and Kluitenberg (2005) derived the following analytical expressions for them:

$$F_T(\boldsymbol{x},t)=-\frac{q_0\left(r^2-a^2/4\right)}{8\pi^2DT_0^2r_1r_2t}K_1\left(\frac{r_1r_2}{2Dt}\right)\exp\left(-\frac{r^2+a^2/4}{2Dt}\right) \tag{4a}$$

$$F_S(\boldsymbol{x},t)=-\frac{q_0}{8\pi^2T_0^2t}K_0\left(\frac{r_1r_2}{2Dt}\right)\exp\left(-\frac{r^2+a^2/4}{2Dt}\right) \tag{4b}$$

In these equations K_0 and K_1 are modified Bessel functions of order 0 and 1, while:

$$r=\sqrt{\left(x^2+y^2\right)} \tag{5}$$

and

$$D=\frac{T_0}{S_0} \tag{6}$$

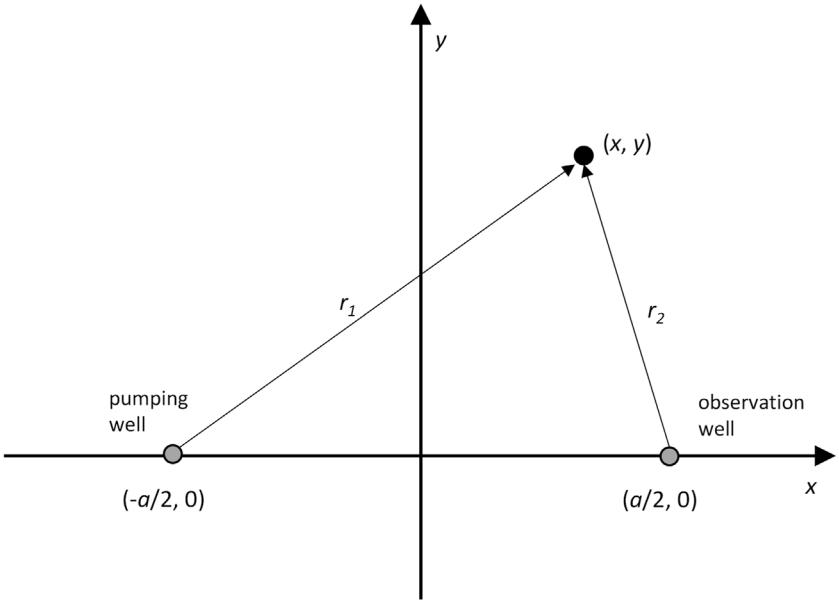

FIGURE 1
A pumping and observation well. Also shown is a general point in two-dimensional space. The hydraulic properties at this point are functions of location (*x*, *y*).

r_1 and r_2 are depicted in Figure 1. Similar equations were derived by Zha et al. (2020).

The sensitivities of observation well drawdown to domain-wide transmissivity and storativity are obtained by areal integration of the respective Fréchet kernels. From Knight and Kluitenberg:

$$M_T(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_T(\boldsymbol{x},t)d\boldsymbol{x} = -\frac{q_0}{4\pi T_0^2}\left[E_1\left(\frac{S_0a^2}{4T_0t}\right) - \exp\left(-\frac{S_0a^2}{4T_0t}\right)\right] \quad (7a)$$

$$M_s(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_S(\boldsymbol{x},t)d\boldsymbol{x} = -\frac{q_0}{4\pi S_0T_0}\exp\left(-\frac{S_0a^2}{4T_0t}\right) \quad (7b)$$

In the Supplementary Material we show how Knight and Kluitenberg's Fréchet kernels can be extended to accommodate the T_x and T_y components of directional transmissivity. The extended kernels are:

$$F_{T_x}(\boldsymbol{x},t) = -\frac{q_0(x+a/2)(x-a/2)}{8\pi^2DT_0^2r_1r_2t}K_1\left(\frac{r_1r_2}{2Dt}\right)\exp\left(-\frac{r^2+a^2/4}{2Dt}\right) \quad (8a)$$

$$F_{T_y}(\boldsymbol{x},t) = -\frac{q_0y^2}{8\pi^2DT_0^2r_1r_2t}K_1\left(\frac{r_1r_2}{2Dt}\right)\exp\left(-\frac{r^2+a^2/4}{2Dt}\right) \quad (8b)$$

Note that these sum to $F_T(\boldsymbol{x},t)$. With T_x and T_y treated separately, Eq. 3 becomes:

$$h(t) = \int_A T_x(\boldsymbol{x})F_{T_x}(\boldsymbol{x},t)d\boldsymbol{x} + \int_A T_y(\boldsymbol{x})F_{T_y}(\boldsymbol{x},t)d\boldsymbol{x} + \int_A S(\boldsymbol{x})F_S(\boldsymbol{x},t)d\boldsymbol{x} \quad (9)$$

2.2 Parameter estimation

Suppose that we wish to back-calculate transmissivity and storativity from drawdowns measured in an observation well. This comprises an ill-posed inverse problem as it is impossible to assign unique values of transmissivity and storativity to all drawdown-affected points within a heterogeneous aquifer. If uniqueness is sought, it must be attained through regularisation. In aquifer test analysis, regularisation is usually achieved by assuming hydraulic property uniformity. In the present case, this reduces inverse problem complexity to that of estimating just two parameters, namely those that represent the transmissivity and storativity of the entire medium. This simplifies the analysis considerably. Meanwhile, it is hoped that the values of domain-wide transmissivity and storativity that emerge from this process are not too different from the "average" transmissivity and storativity of the porous medium which hosts the pumping test. Shortly, we examine whether this hope is well-placed.

We continue to assume a linear relationship between drawdown and aquifer hydraulic properties. This is in accordance with the

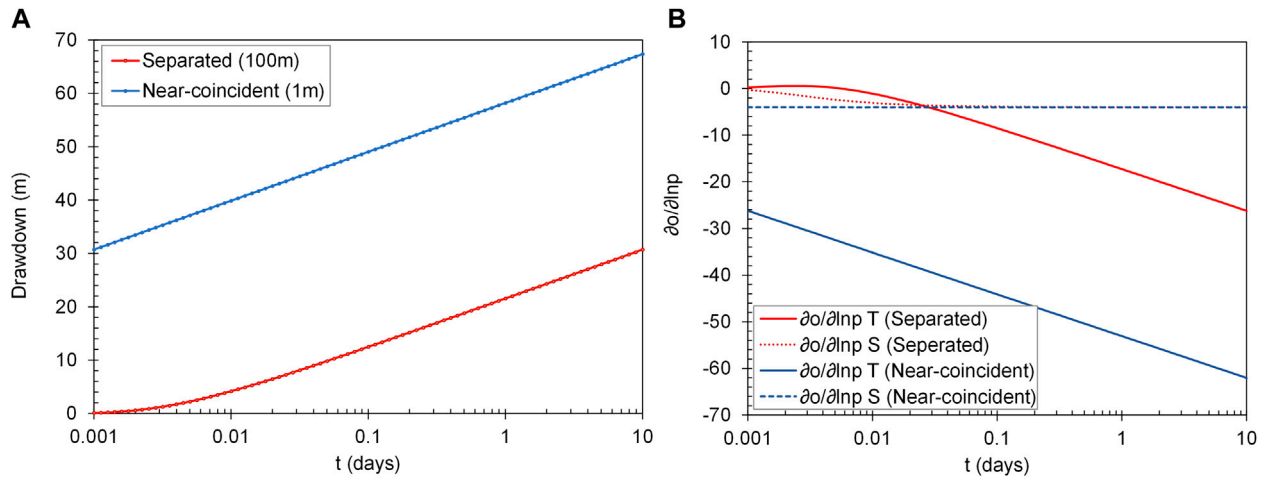

FIGURE 2
(A) Drawdown in the observation well. **(B)** The derivative of drawdown with respect to the natural logs of global T and S.

theory on which most parameter estimation methods are based. In practical parameter estimation, non-linearities in this relationship are accommodated through iterative updating of sensitivities as estimated parameter values change.

The matrix-vector equation on which linearized parameter estimation is based can be written as follows:

$$\mathbf{h} = \mathbf{M}\mathbf{p} + \boldsymbol{\varepsilon} \tag{10}$$

In Eq. 10 the $\mathbf{h}$ vector contains differences between measured drawdowns in the observation well and those that are calculated using the domain-wide background values T_0 and S_0. Let us suppose that there are n such drawdown measurements. The vector $\mathbf{p}$ contains adjustments T and S to T_0 and S_0 (Eqs 2a, 2b). That is:

$$\mathbf{p} = \begin{bmatrix} T \\ S \end{bmatrix} \tag{11}$$

$\boldsymbol{\varepsilon}$ (another n-dimensional vector) encapsulates random noise that is associated with measurements of drawdown. The $n \times 2$ matrix $\mathbf{M}$ can be written as follows:

$$\mathbf{M} = \begin{bmatrix} M_T(t_1) & M_S(t_1) \\ M_T(t_2) & M_S(t_2) \\ . & . \\ M_T(t_n) & M_S(t_n) \end{bmatrix} \tag{12}$$

where $M_T(t_i)$ and $M_S(t_i)$ are Eqs 7a, 7b calculated at time t_i.

The least squares solution to the inverse problem posed by Eq. 10 [see, for example, Draper and Smith (1998)] is:

$$\mathbf{p} = (\mathbf{M}^t\mathbf{Q}\mathbf{M})^{-1}\mathbf{M}^t\mathbf{Q}\mathbf{h} \tag{13}$$

where $\mathbf{Q}$ is an observation weight matrix. Ideally $\mathbf{Q}$ is proportional to the inverse of the covariance matrix of measurement noise C($\boldsymbol{\varepsilon}$). The latter is normally assumed to be diagonal; so too is $\mathbf{Q}$.

The elements of $\mathbf{h}$ can be calculated from distributed transmissivity and storativity using Eq. 3 or Eq. 9. The choice depends on whether or not we wish to characterise transmissivity using T alone, or T_x and T_y separately. To reduce complexity of the following equations, we first consider T on its own.

Equation 3 can be written in vector form as:

$$\mathbf{h} = \mathbf{F}\mathbf{k}. \tag{14}$$

where $\mathbf{k}$ is the vector:

$$\mathbf{k} = \begin{bmatrix} T_1 \\ . \\ T_m \\ S_1 \\ . \\ S_m \end{bmatrix} \tag{15}$$

The subscripts that accompany T and S in the $\mathbf{k}$ vector of Eq. 15 signify discretisation of x-y space into m elements (where m is a large number) for the purpose of numerical integration. In the present study we employ equal-sized, square cells and apply the midpoint rule.

To specify the $\mathbf{F}$ matrix, we write the integrals in Eq. 3 as summations:

$$h_i = \sum_A F^T_{i,j} T_j + \sum_A F^S_{i,j} S_j \tag{16}$$

where i denotes the i'th time at which drawdown measurements were made, and j denotes the j'th cell that is used for spatial integration. With $\mathbf{k}$ defined by Eq. 15, $\mathbf{F}$ becomes:

$$\mathbf{F} = \begin{bmatrix} F^T_{1,1} & F^T_{1,m} & F^S_{1,1} & F^S_{1,m} \\ F^T_{n,1} & F^T_{n,m} & F^S_{n,1} & F^S_{n,m} \end{bmatrix} \tag{17}$$

We now substitute Eq. 14 into Eq. 13 to obtain:

$$\mathbf{p} = (\mathbf{M}^t\mathbf{Q}\mathbf{M})^{-1}\mathbf{M}^t\mathbf{F}\mathbf{k} \tag{18}$$

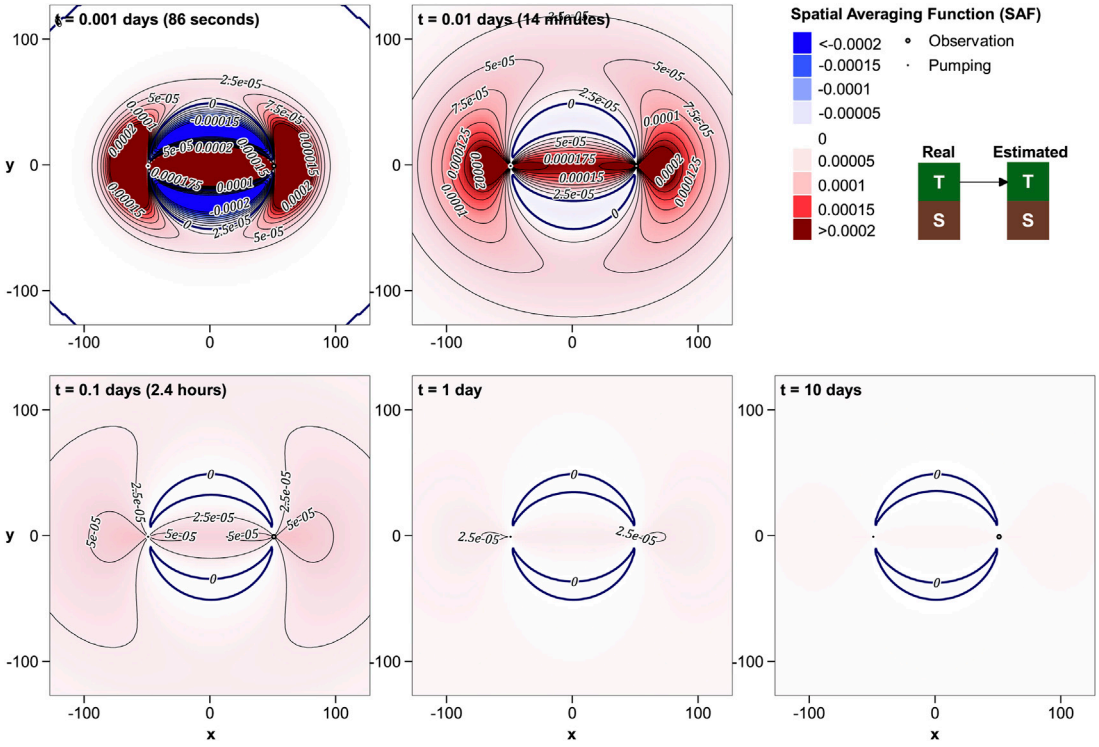

FIGURE 3
$R'_{\underline{T}-T}$ for separated wells at five different times.

This can be re-written as:

$$\mathbf{p} = \mathbf{R}'\mathbf{k} \tag{19}$$

where $\mathbf{R}'$ is defined through this equation.

The matrix $\mathbf{R}'$ has two rows and $2m$ columns. Each row of this matrix depicts the manner in which elements of $\mathbf{k}$ are summed (i.e., spatially integrated) in order to calculate the pertinent element of $\mathbf{p}$. That is, each row of $\mathbf{R}'$ shows how an estimated, global (or apparent) parameter ($\underline{T}$ or $\underline{S}$) is related to spatially-distributed, real-world hydraulic properties T and S. We make this relationship explicit by re-writing $\mathbf{R}'$ as follows:

$$\mathbf{R}' = \begin{bmatrix} R'_{\underline{T}-T_1} & R'_{\underline{T}-T_m} & R'_{\underline{T}-S_1} & R'_{\underline{T}-S_m} \\ R'_{\underline{S}-T_1} & R'_{\underline{S}-T_m} & R'_{\underline{S}-S_1} & R'_{\underline{S}-S_m} \end{bmatrix} \tag{20}$$

It is apparent from Eq. 20 that part of the estimated value of global $\underline{T}$ is inherited from real-world values of S, and *vice versa*. We refer to this phenomenon as "parameter contamination" herein. The mapping of real-world, spatially-distributed T and S to estimated $\underline{T}$ and estimated $\underline{S}$ can be visualized by plotting respective elements of the $\mathbf{R}'$ matrix at the locations in space to which they pertain. Four such maps are implied in the $\mathbf{R}'$ matrix - two for the mapping of real-world T and S to $\underline{T}$, and two for the mapping of real-world T and S to $\underline{S}$. To allow easier identification of maps that are presented in the next section, we re-write $\mathbf{R}'$ as a composite matrix in which each sub-matrix pertains to such a map.

$$\mathbf{R}' = \begin{bmatrix} \mathbf{R}'_{\underline{T}-T} & \mathbf{R}'_{\underline{T}-S} \\ \mathbf{R}'_{\underline{S}-T} & \mathbf{R}'_{\underline{S}-S} \end{bmatrix} \tag{21}$$

Each of the submatrices $\mathbf{R}'_{\underline{X}-Y}$ that appear in Eq. 21 has one row and m columns. In the discussion that follows, we refer to the contents of these columns as a "spatial averaging function."

Where transmissivity is considered to be directional, Eq. 15 becomes:

$$\mathbf{k} = \begin{bmatrix} T_{x1} \\ . \\ T_{xm} \\ T_{y1} \\ . \\ T_{ym} \\ S_1 \\ . \\ S_m \end{bmatrix} \tag{22}$$

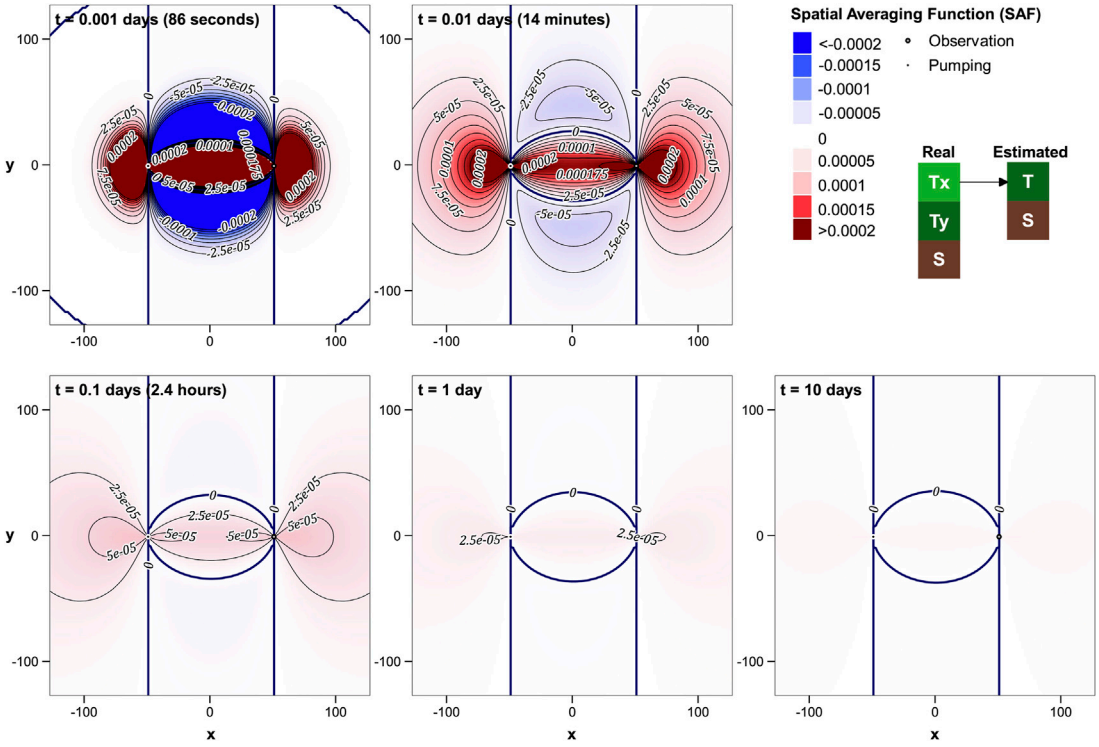

FIGURE 4
$R'_{\underline{T}-Tx}$ for separated wells at five different times.

so that Eq. 21 becomes:

$$\mathbf{R}' = \begin{bmatrix} \mathbf{R}'_{\underline{T}-T_x} & \mathbf{R}'_{\underline{T}-T_y} & \mathbf{R}'_{\underline{T}-S} \\ \mathbf{R}'_{\underline{S}-T_x} & \mathbf{R}'_{\underline{S}-T_y} & \mathbf{R}'_{\underline{S}-S} \end{bmatrix} \tag{23}$$

2.3 A note on regularisation

The matrix $\mathbf{R}'$ that is defined above bears some relationship to the so-called "resolution matrix" that plays a prominent role in the theory of regularised inversion; see, for example, Menke (2018) and Aster, et al. (2019). However, a true resolution matrix is square and generally rank-deficient; it relates fine-scale parameter estimates to fine-scale, real-world hydraulic properties. The name of our $\mathbf{R}'$ matrix includes a prime in order to distinguish it from the conventional resolution matrix.

Inversion theory makes it clear that an inevitable consequence of inverse problem ill-posedness is that the value that is estimated for a parameter at one particular location is a spatial integral of parameter values over many locations, and that this integration process can cross parameter boundaries where parameters of more than one type are simultaneously estimated. This is a "cost of uniqueness" (Moore & Doherty, 2006). It is incurred regardless of the adopted regularisation strategy. Regularisation that is based on an assumption of hydraulic property uniformity cannot evade this cost. Nor is hydraulic property uniformity necessarily the best regularisation strategy to use, if "best" is defined as a proclivity to yield predictions whose error variance is minimized (Doherty, 2015). However, it is generally the most convenient strategy to use for aquifer test analysis.

It is important to understand that the averaging function that relates estimated to real-world parameters is an outcome of the adopted regularisation strategy. It is not a foregone conclusion that this averaging function is either "clean" or desirable, or yields hydraulic property estimates that are immediately useful for other purposes (for example, parameterization of a groundwater model).

Conceptually, it is possible to design an inversion process that specifically seeks estimates of hydraulic properties that are averaged over space in a user-specified manner. However, this process is somewhat cumbersome. It requires pre-inversion construction of a matrix that characterizes "structural noise" incurred by departures of real-world hydraulic properties from uniformity. This, in turn, requires prior statistical characterization of hydraulic property heterogeneity. For details see Cooley (2004) and Cooley and Christensen (2006).

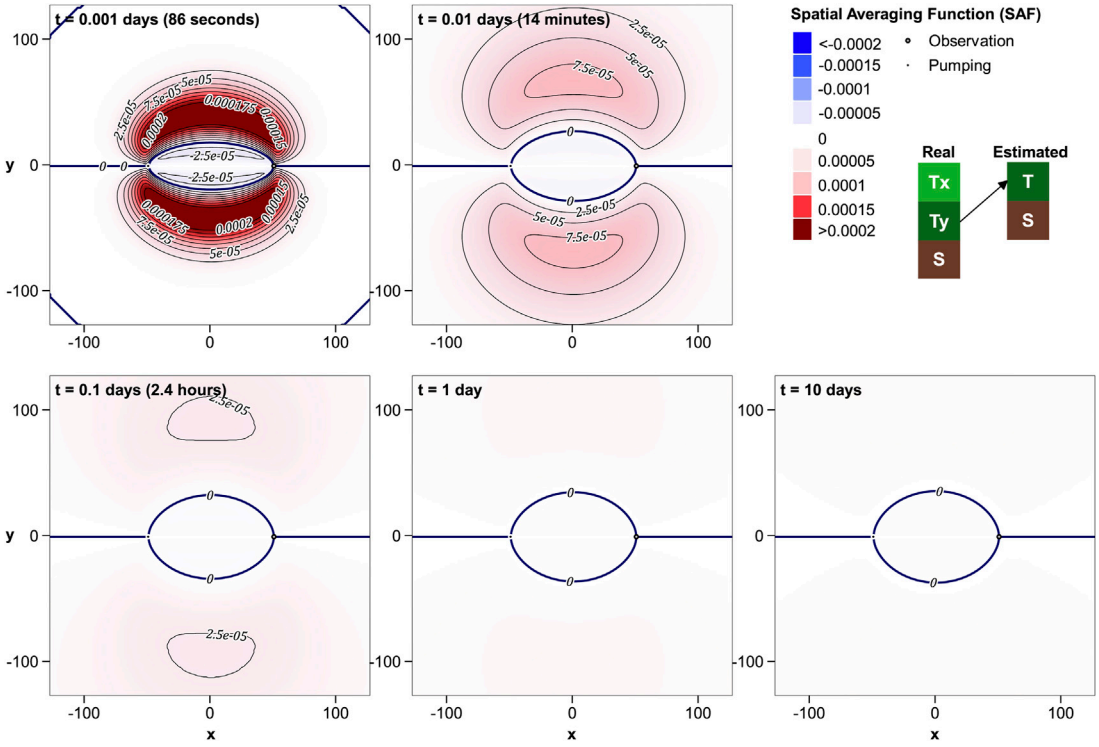

FIGURE 5
$R'_{\underline{T}-Ty}$ for separated wells at five different times.

In most cases we have no option but to pose an inverse problem in a way that is amenable to rapid solution, and then to "take what we can get" as far as hydraulic property averaging is concerned.

The theory that is outlined above is linear. It is presented in terms of departures from uniformity of real-world hydraulic properties. Each element of $\mathbf{R}'$ can therefore be viewed as a derivative; as such, it specifies the change in the estimated value of $\underline{T}$ or $\underline{S}$ incurred by a change in real-world T or S at a specified subsurface location. In actual fact, the value of this derivative is dependent on the spatially variable values of T and S; however in practice, the elements of R' are computed using T_o and S_o in accordance with the above theory. Despite these limitations, maps based on the submatrices that appear in Eqs 21, 23 can be loosely viewed as depicting contributions to estimated $\underline{T}$ and $\underline{S}$ by real-world, spatially-distributed T and S. Thus they address the question of what estimated values of $\underline{T}$ and $\underline{S}$ really mean. The linearity assumption yields an approximate answer to this question that would be difficult to obtain in any other way.

2.4 A note on the methodology

It can be argued that the introduction of heterogeneity to a model domain erodes the applicability of Theis-type aquifer test analysis. Evidence of its invalidity may be visible in time-correlated misfit between measured drawdowns and best-fit Theis-evaluated drawdowns.

Nevertheless, most aquifer tests are undertaken in heterogeneous media. Furthermore, for many aquifer tests, at least some drawdown misfit can be attributed to the heterogeneous nature of the medium in which the test is undertaken. This is mostly ignored in real-world aquifer test data interpretation. For convenience, misfit is generally attributed to "measurement noise;" uncertainties in estimated $\underline{T}$ and $\underline{S}$ that are incurred by this misfit are calculated accordingly. (We do not address these uncertainties in this paper). We note that the above derivations of spatial averaging functions are not invalidated by heterogeneity-incurred misfit, for these derivations require no assumptions pertaining to misfit sources or misfit statistics; they only require that a least-squares objective function be minimized.

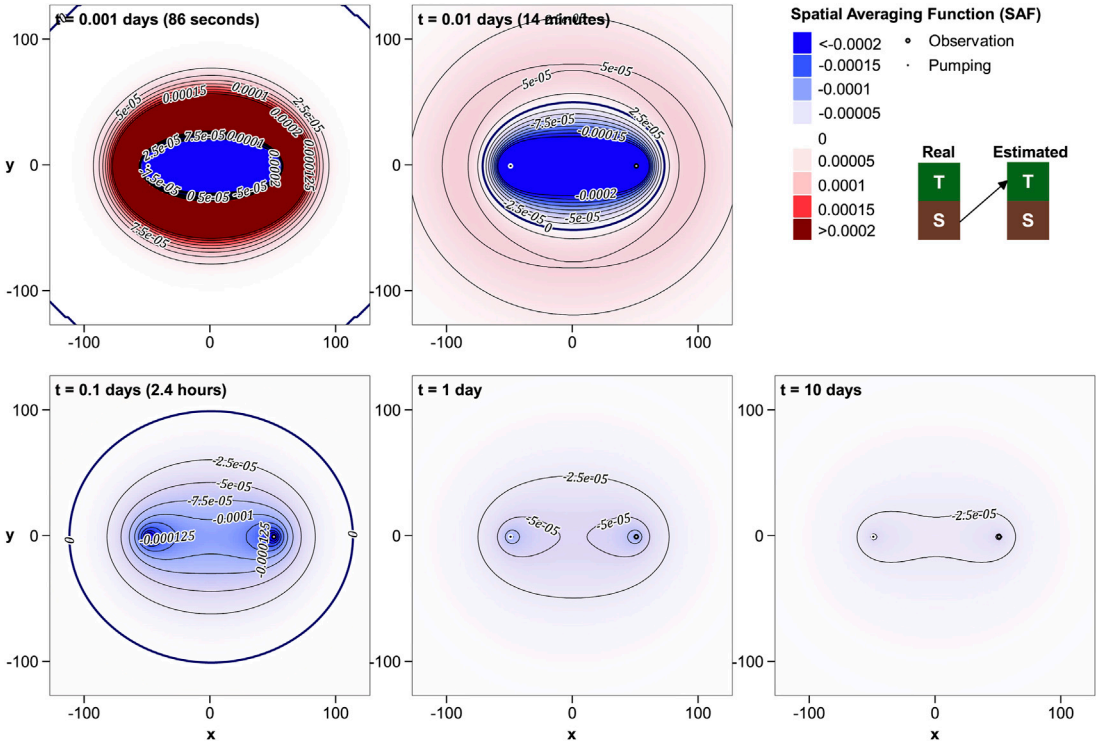

FIGURE 6
$R'_{\underline{T}-S}$ for separated wells at five different times.

3 An example

3.1 Specifications

We put the above theory to use by examining spatial averaging functions for two configurations of a pumping well and an observation well. In both cases, water is pumped at a rate of 5,000 m^3/day from an aquifer whose transmissivity is 100 m^2/day and whose storativity is 0.0001. (As stated above, because of the linearity assumption on which the above theory is based, the effects of aquifer heterogeneity on estimated $\underline{T}$ and $\underline{S}$ can be studied without actually introducing heterogeneity to the synthetic aquifer.) In the first case, the well separation is set to 100 m. In the second case it is set to 1 m. We refer to these as the "separated well case" and the "near-coincident well case" respectively. Non-integrable singularities in some Fréchet kernels prevent us from placing the pumping and observation wells at the same place, so we employ a small well separation as a proxy for coincident wells.

Drawdowns are sampled at a rate of 20 measurements per decade in time, starting at 0.001 days and finishing at 10 days. Drawdown measurement error is assumed to be random and independent, with a standard deviation of 0.05 m. The diagonal elements of the **Q** matrix of Eq. 13 are set to the inverse square of this, namely 400.0. (Note that the results presented below are invariant with multiplication of all elements of **Q** by a constant factor.)

Integration of spatial averaging kernels is undertaken using the midpoint rule over a uniform grid comprised of 2 m × 2 m square cells. Integration is required over only one quadrant of the *x*-*y* plane because of symmetry. The integration grid extends for 40 km in the *x* and *y* directions.

Note that the symmetry of this problem has another important implication. All of the results presented below remain the same if the pumping and observation wells are interchanged.

In the present study T_x, T_y, T and S are log-transformed. Hence relationships are sought between the logs of estimated $\underline{T}$ and $\underline{S}$ and the logs of real-world hydraulic properties; Fréchet integrals appearing in the previous section are modified accordingly. Log-transformation enhances linearity at the same time as it accommodates the wide range of values that these hydraulic properties can adopt. To simplify the following discussion, we mostly omit any reference to log transformation when describing spatial averaging functions; however the reader should keep their log-transformed status in mind.

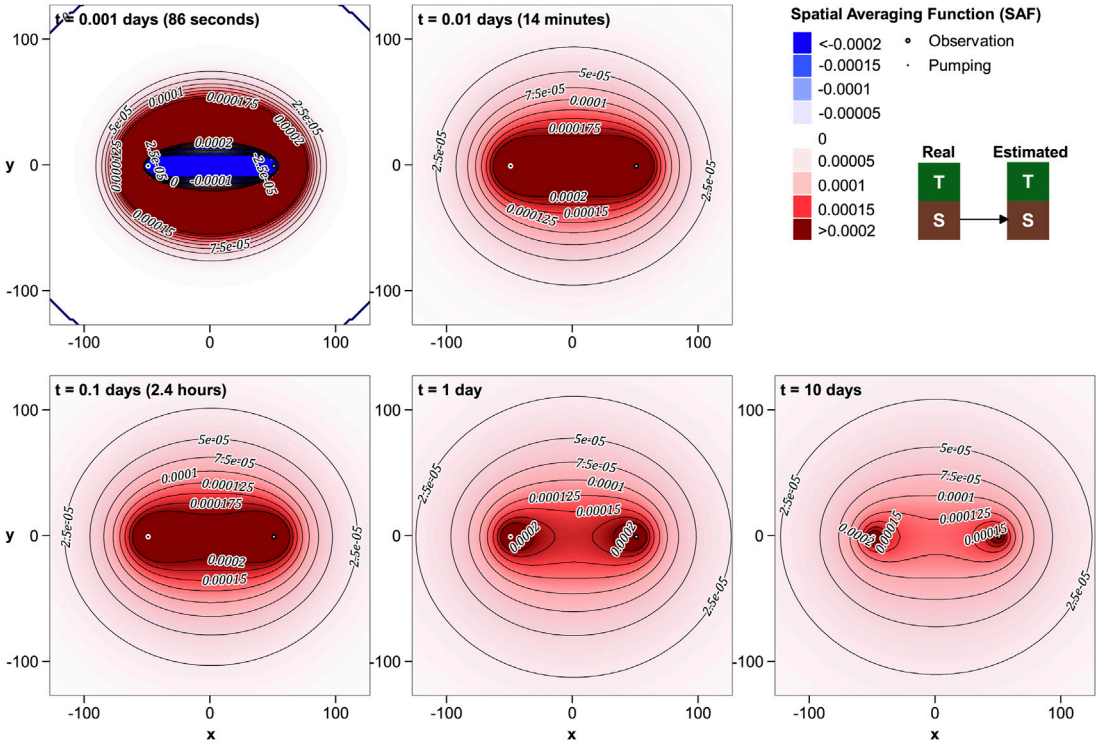

FIGURE 7
$R'_{\underline{S}-S}$ for separated wells at five different times.

Figure 2A shows drawdown plotted against the log (to base 10) of time for the separated and near-coincident well cases. Figure 2B plots the derivative of drawdown with respect to the natural log of global T and the natural log of global S for these same cases. The derivative of drawdown with respect to log T for separated wells changes from positive to negative after about 0.006 days (about 8 min). At very early times, an increase in transmissivity accelerates the propagation of drawdown from the extraction well to the observation well. However after that time, an increase in aquifer transmissivity induces less drawdown at the observation well for a given amount of flow. No such reversal occurs for the near-coincident well case.

3.2 Spatially integrated kernels

At any time during an aquifer test, the submatrices appearing in Eqs 21 and 23 can be computed in the manner described above. As such, they pertain to aquifer test interpretation that is based on drawdowns that are sampled up until that time. Each row of these submatrices comprises an integration kernel. In accordance with Eq. 19 each element in each row is multiplied by the corresponding element of $\mathbf{k}$; that is, it is multiplied by T, T_x, T_y or S pertaining to a point within the aquifer. These values are then summed as a proxy for spatial integration.

Numerical integration of these kernels on their own yields the following results at all times.

$$\int_A R'_{\underline{T}-T}(\boldsymbol{x})d\boldsymbol{x} = 1.0 \tag{24a}$$

$$\int_A R'_{\underline{T}-T_x}(\boldsymbol{x})d\boldsymbol{x} = 0.5 \tag{24b}$$

$$\int_A R'_{\underline{T}-T_y}(\boldsymbol{x})d\boldsymbol{x} = 0.5 \tag{24c}$$

$$\int_A R'_{\underline{T}-S}(\boldsymbol{x})d\boldsymbol{x} = 0.0 \tag{24d}$$

$$\int_A R'_{\underline{S}-T}(\boldsymbol{x})d\boldsymbol{x} = 0.0 \tag{24e}$$

$$\int_A R'_{\underline{S}-T_x}(\boldsymbol{x})d\boldsymbol{x} = -0.5 \tag{24f}$$

$$\int_A R'_{\underline{S}-T_y}(\boldsymbol{x})d\boldsymbol{x} = 0.5 \tag{24g}$$

$$\int_A R'_{\underline{S}-S}(\boldsymbol{x})d\boldsymbol{x} = 1.0 \tag{24h}$$

Collectively, Eqs 24a, 24d, 24e, 24h imply that the inverse problem is well-posed, for if the aquifer is homogeneous, values of $\underline{T}$ and $\underline{S}$ calculated using Eq. 18 are estimates of the true, global values of T and S.

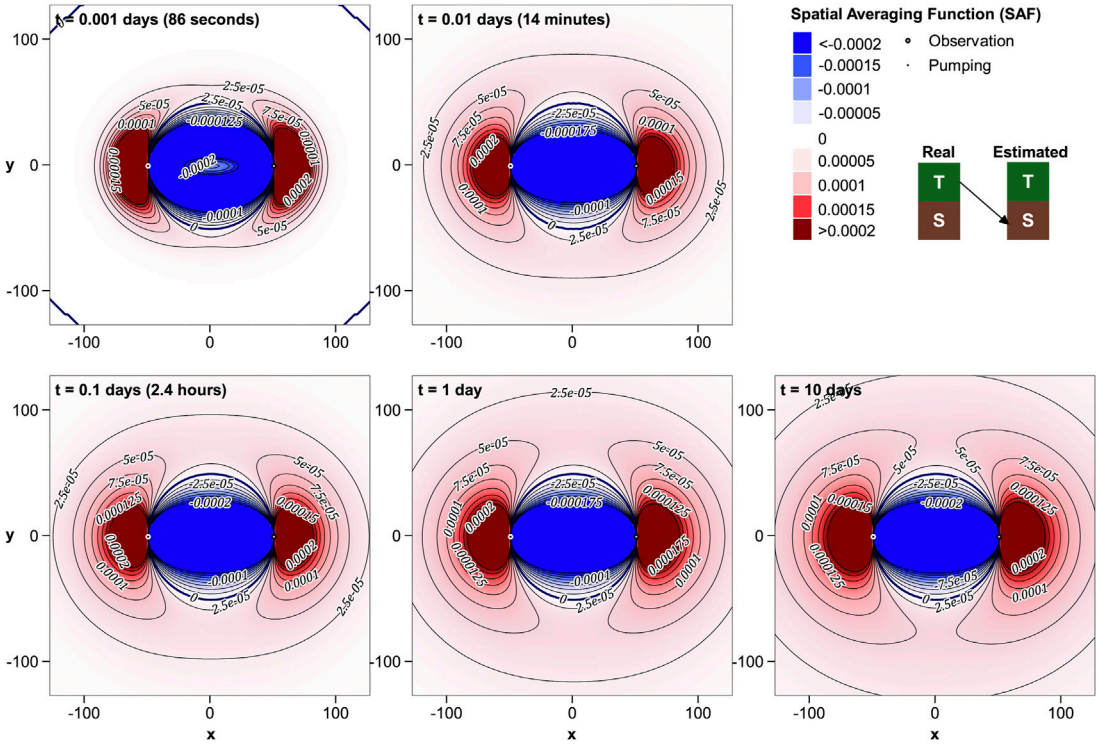

FIGURE 8
$R'_{\underline{S}-T}$ for separated wells at five different times.

Equations 24b, 24c, 24f, 24g are noteworthy. Suppose that T_x is multiplied by a factor c and that T_y is divided by this same factor. (Note that multiplication becomes addition in the log domain). This introduces horizontal anisotropy to the aquifer. According to the above equations, the estimated $\underline{T}$ is unchanged; it is therefore equal to the geometric mean of T_x and T_y. In contrast, estimated $\underline{S}$ is altered because the right sides of Eqs 24f and 24g have opposite signs; its value is decreased by a factor of c. However, if the real value of S throughout the aquifer is then multiplied by c, estimated $\underline{S}$ returns to its original value. Use of a variant of the Theis equation that accommodates aquifer anisotropy (Papadopulos, 1965) verifies that drawdowns at the observation well are unchanged under these conditions.

The integrals that are presented in Eqs 24a to 24h can assist in interpreting the kernel maps that are discussed below. For example, early-time values of $R'_{\underline{T}-T}$ for the separated-well case are significantly negative in some areas. These must be balanced by areas of $R'_{\underline{T}-T}$ positivity so that the negative contribution to the total integral is not only cancelled, but integrates to 1.0. These positive values may be spread out over large areas, and so may not be as obvious as intensely negative values when plotted in space. Similarly, areas of negative $R'_{\underline{S}-T}$ must be balanced by areas of positive $R'_{\underline{S}-T}$ so that $R'_{\underline{S}-T}$ spatially integrates to zero.

3.3 Maps of spatial averaging function for separated wells

This section provides maps of $R'_{\underline{X}-Y}$ where $\underline{X}$ is either $\underline{T}$ or $\underline{S}$ and Y is either T, T_x, T_y or S. Maps are presented for five different times. In each case, the map pertains to Theis-based interpretation of observation well drawdowns acquired up until that time. In all of these figures, red is indicative of positive values while blue indicates negative values. Shading is linear; the zero contour is highlighted. When viewing these plots, keep in mind that it is their spatial integral that matters, for this is what determines an estimated value of $\underline{T}$ or $\underline{S}$.

The Supplementary Material presents these same maps, but with logarithmic shading. This reveals the spatial characteristics of these functions in low-intensity areas that are distant from the pumping and observation wells. Because these areas are large, they make significant contributions to the spatial integrals though which $\underline{T}$ and $\underline{S}$ are calculated.

$R'_{\underline{T}-T}$ is mapped in Figure 3. Unsurprisingly, contributions of real T to estimated $\underline{T}$ expand outward from the pumping and observation wells with time, diminishing in magnitude, but covering a broader area. At larger times, contours of equal $R'_{\underline{T}-T}$ form ellipses with foci at

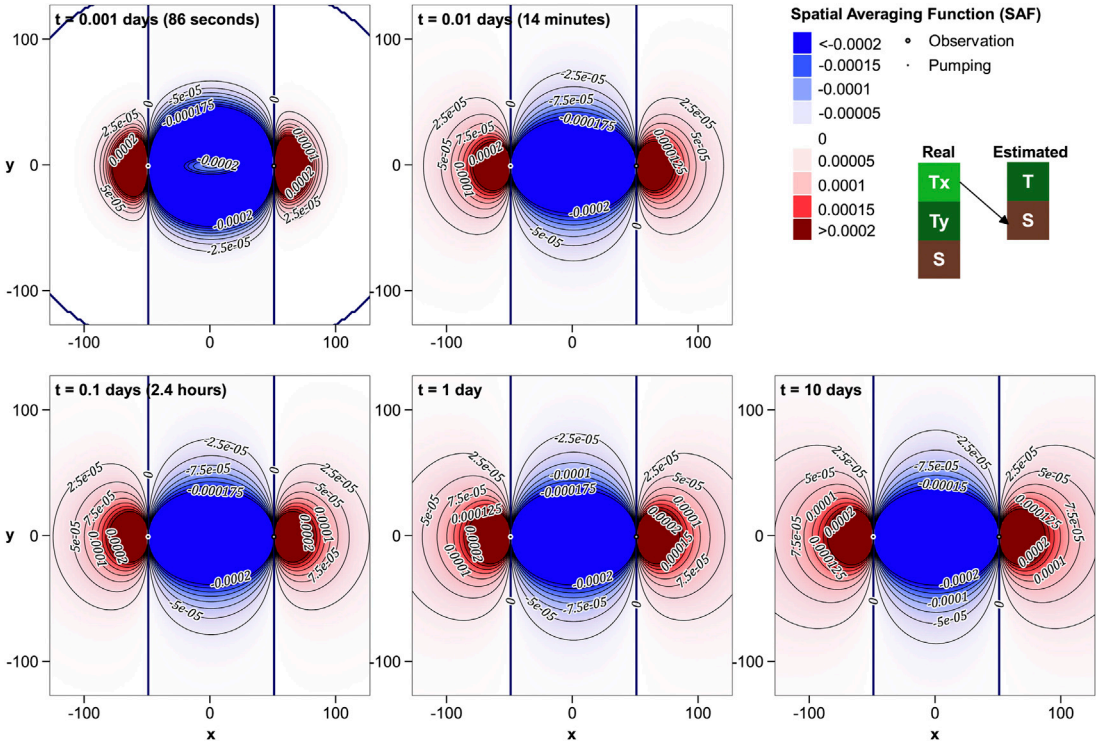

FIGURE 9
$R'_{\underline{S}-Tx}$ for separated wells at five different times.

the pumping and observation wells. In contrast, the early-time pattern is complex, with high values between the wells, and within lobes that extend beyond the wells. Small negative cusps lie to the north and south of the line that joins the wells. These negative cusps diminish in intensity with increasing time, but never completely disappear.

Figures 4, 5 show that contributions made by real-world T_x and T_y to Theis-estimated $\underline{T}$ are more complex than that made by T alone. (See also the plots in the Supplementary Material where logarithmic shading is employed.) Local T anisotropy close to and between the wells can strongly affect estimated $\underline{T}$ at very early times. However, contributions from these areas fade with time as smoother $R'_{\underline{T}-Tx}$ and $R'_{\underline{T}-Ty}$ kernels expand beyond the wells. At large times, estimated $\underline{T}$ reflects true T_x that prevails at large distances along the x axis and true T_y that prevails at large distances along the y axis. Reciprocally, T_y at large distances along the x axis and T_x at large distances along the y axis have little effect on estimated $\underline{T}$. That is to say, estimated $\underline{T}$ reflects components of directional real-world T that point towards the wells.

Figure 6 depicts the potential for contamination of estimated $\underline{T}$ by real-world S. At early times, real-world S between the extraction and observation wells can exert a considerable influence on estimated $\underline{T}$. Positive and negative contributions of S to $\underline{T}$ are strong, but collectively integrate to zero as outlined above.

The presence of a significant band of negatively-valued $R'_{\underline{T}\text{-}S}$ joining the pumping well to the observation well at early times is easily explained. Low storativity in this area hastens propagation of drawdown to the observation well; it therefore "looks like" high local T. This affects estimated $\underline{T}$ shortly after the commencement of pumping when the derivative of drawdown with respect to global S and global T are both positive; see Figure 2B. At later times, contributions of real-world S to estimated $\underline{T}$ become more diffuse. However an area of anomalous S between the pumping and observation wells is never quite "forgotten" by the Theis-based parameter estimation process.

Figure 7 maps the contribution to estimated $\underline{S}$ by real-world S. In contrast to spatial averaging functions that affect estimated $\underline{T}$, areas that contribute to estimated $\underline{S}$ tend to remain close to the pumping and observation wells even at late times. At very early times the pattern is complex; the wells are joined by a sliver of intensely negative S contribution to $\underline{S}$; this is surrounded by areas of intensely positive contributions of S to $\underline{S}$.

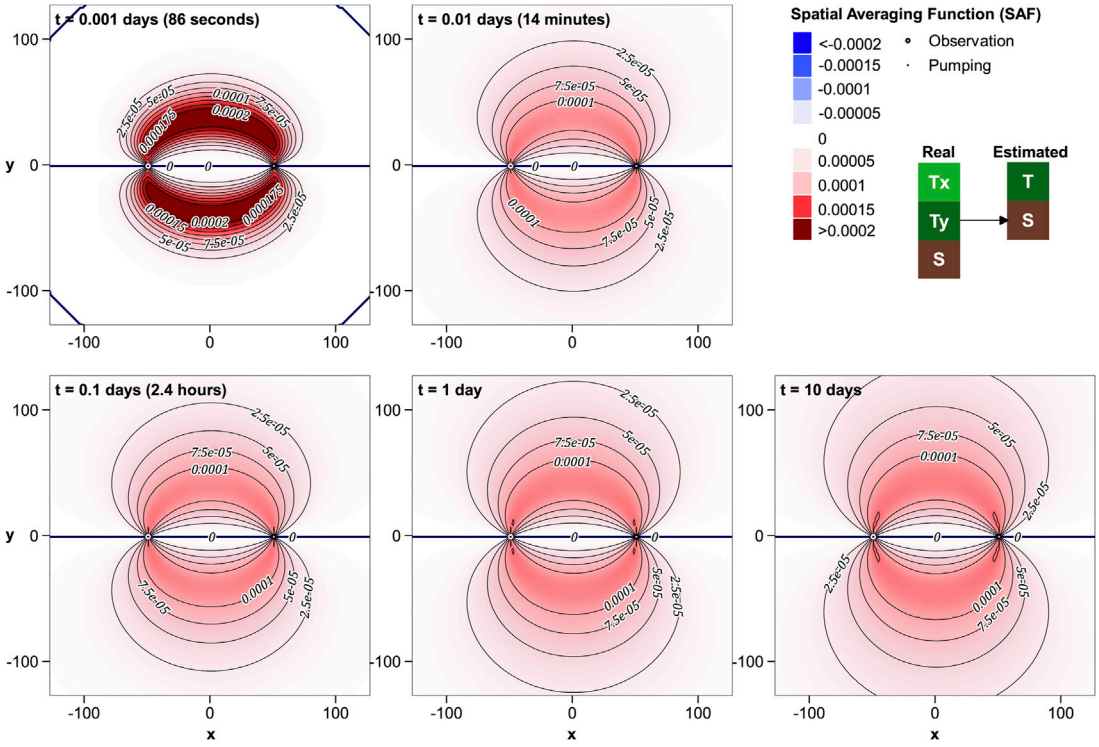

FIGURE 10
$R'_{\underline{S}-Ty}$ for separated wells at five different times.

The limited expansion of $R'_{\underline{S}-S}$ with time (and of other $R'_{\underline{S}-X}$ maps shown below) implies that estimates of $\underline{S}$ forthcoming from aquifer test analysis pertain to a smaller area than do estimates of $\underline{T}$. Furthermore, this disparity in area grows with the length of the test. These maps also imply that information pertaining to aquifer storativity ceases to be acquired after a relatively short time during a pumping test. This accords with identification by authors such as Kruseman and de Ridder (1990) of a pseudo-steady-state phase of drawdown development as time proceeds.

Figures 8–10 suggest a high potential for contamination of estimated $\underline{S}$ by between-well T, especially T_x. High values of T_x in this area can hasten propagation of drawdown from the pumping well to the observation well, thereby replicating the impact of low S. This effect is strong, and does not dissipate with time. In hydrogeological contexts where the natural variability of T is much greater than that of S, the potential for contamination of estimated $\underline{S}$ by between-well anomalies in T may be very high.

The impact of real-world T_y on estimated $\underline{S}$ is complex. Like that of T, it persists to late times. The near-well cusps of positive influence suggest that strategically-located areas of high T_y can gather water from distant areas that can slow the growth of drawdown in the observation well. This appears as high global $\underline{S}$ where these drawdowns are subjected to Theis interpretation.

3.4 Maps of spatial averaging function for near-coincident wells

As stated above, calculations for coincident wells are complicated by non-integrable singularities in Fréchet kernels. To avoid this problem, wells are placed 1 m apart. Furthermore, we do not provide maps of $R'_{\underline{S}-X}$ for the near-coincident case because $\underline{S}$ cannot be estimated unless a dedicated observation well is employed.

From Figures 11–14 it is apparent that maps of $R'_{\underline{T}-T}$ and $R'_{\underline{T}-S}$ are radially symmetric, while those of $R'_{\underline{T}-Tx}$ and $R'_{\underline{T}-Ty}$ are not. This is because anisotropy is the only thing that distinguishes one direction from another when the pumping and observation wells are coincident.

Figure 11 demonstrates that $R'_{\underline{T}-T}$ expands continuously with time, and is ubiquitously positive. Its rate of expansion is not quite as fast as for separated wells. $R'_{\underline{T}-Tx}$ and $R'_{\underline{T}-Ty}$ also maintain positivity over time and space. Their lobate shapes indicate that estimated $\underline{T}$ is influenced by T_y that prevails in the positive and negative y directions from the pumping well and by T_x that prevails in the positive and negative x directions from this well. These are the directions from which water flows towards the pumped well. Because x and y directions are arbitrary for coincident wells, a more general (but unsurprising) conclusion

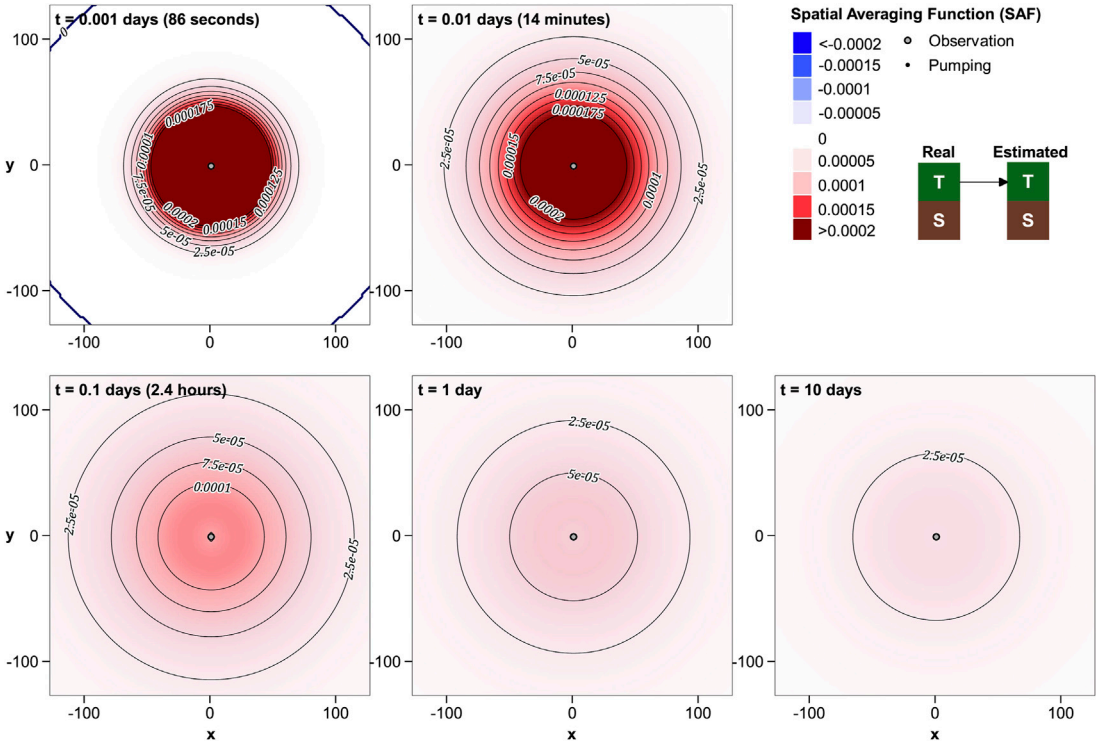

FIGURE 11
$R'_{\underline{T}-T}$ for near-coincident wells at five different times.

follows. It is that estimated $\underline{T}$ reflects the component of real-world T that points in the direction of the pumped well.

Figure 14 shows that near-well, real-world S can have a strong negative influence on estimated $\underline{T}$ at early times. Its influence is somewhat subdued at later times, but is never completely forgotten.

3.5 Radius of investigation

The radius of investigation of an aquifer test is discussed by a number of the authors that are cited in the introduction to this paper; see Bresciani et al. (2020) for a review. Different definitions highlight different aspects of drawdown propagation, and of responsiveness of drawdown to the presence of a distant barrier. The subject matter of the present paper suggests an additional definition, this being based on the area of aquifer that contributes to Theis-estimated $\underline{T}$.

As stated above, for separated wells contours of $R'_{\underline{T}-T}$ at large distances from the pumping and observation wells form ellipses with foci at these wells. These ellipses collapse to circles for coincident wells. We (somewhat arbitrarily) define the radius of investigation of an aquifer test as the length of the major semi-axis of an ellipse that encloses an area that contributes all but 10% to the estimated value of $\underline{T}$. Thus $R'_{\underline{T}-T}$ within the ellipse integrates to 0.9. (Note that the lengths of the semi-major and semi-minor axes of this ellipse are almost equal at large distances from the wells).

In Figure 15 the radius of investigation is plotted against time for both separated and near-coincident wells. It is apparent from this figure that $\underline{T}$ inferred from drawdowns that are measured in a separate observation well "feels" more of the surrounding real-world T than $\underline{T}$ that is inferred from drawdowns in a pumping well. Supposedly, this increased area of spatial averaging provides greater immunity from the effects of near-well anomalies in real-world S and T. However it renders estimated $\underline{T}$ more susceptible to the effects of system boundaries. It is of interest to note that the difference in radius of influence between the separated and near-coincident cases grows with time. For the parameters we chose for this example it is roughly equal to the well separation after 0.5 days, and grows to more than double this after 10 days.

4 Discussion

Work that is documented herein extends previous investigations into the relationship between aquifer-test-inferred hydraulic properties and real-world hydraulic

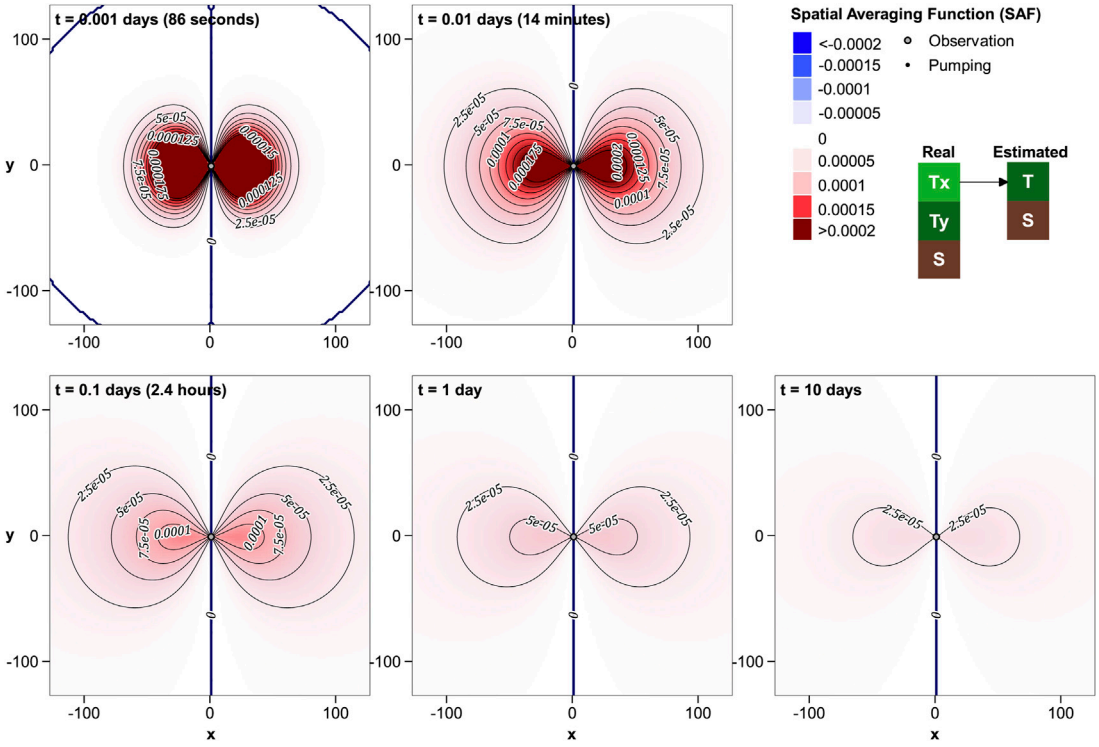

FIGURE 12
$R'_{\underline{T}-Tx}$ for near-coincident wells at five different times.

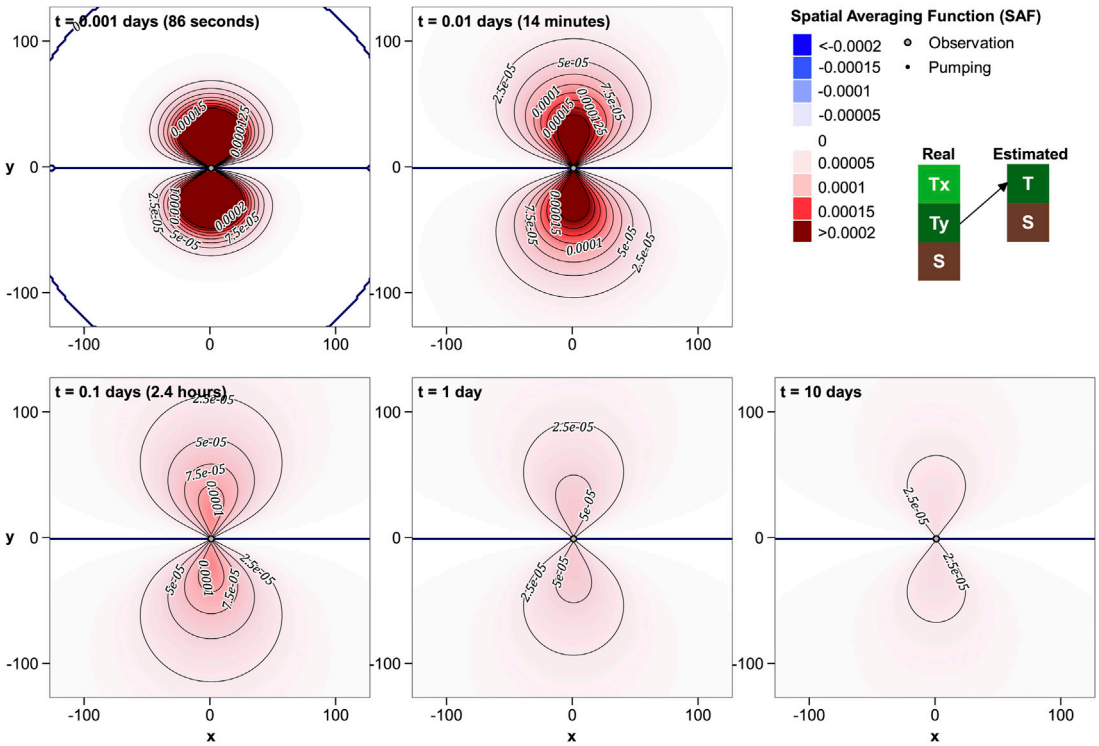

FIGURE 13
$R'_{\underline{T}-Ty}$ for near-coincident wells at five different times.

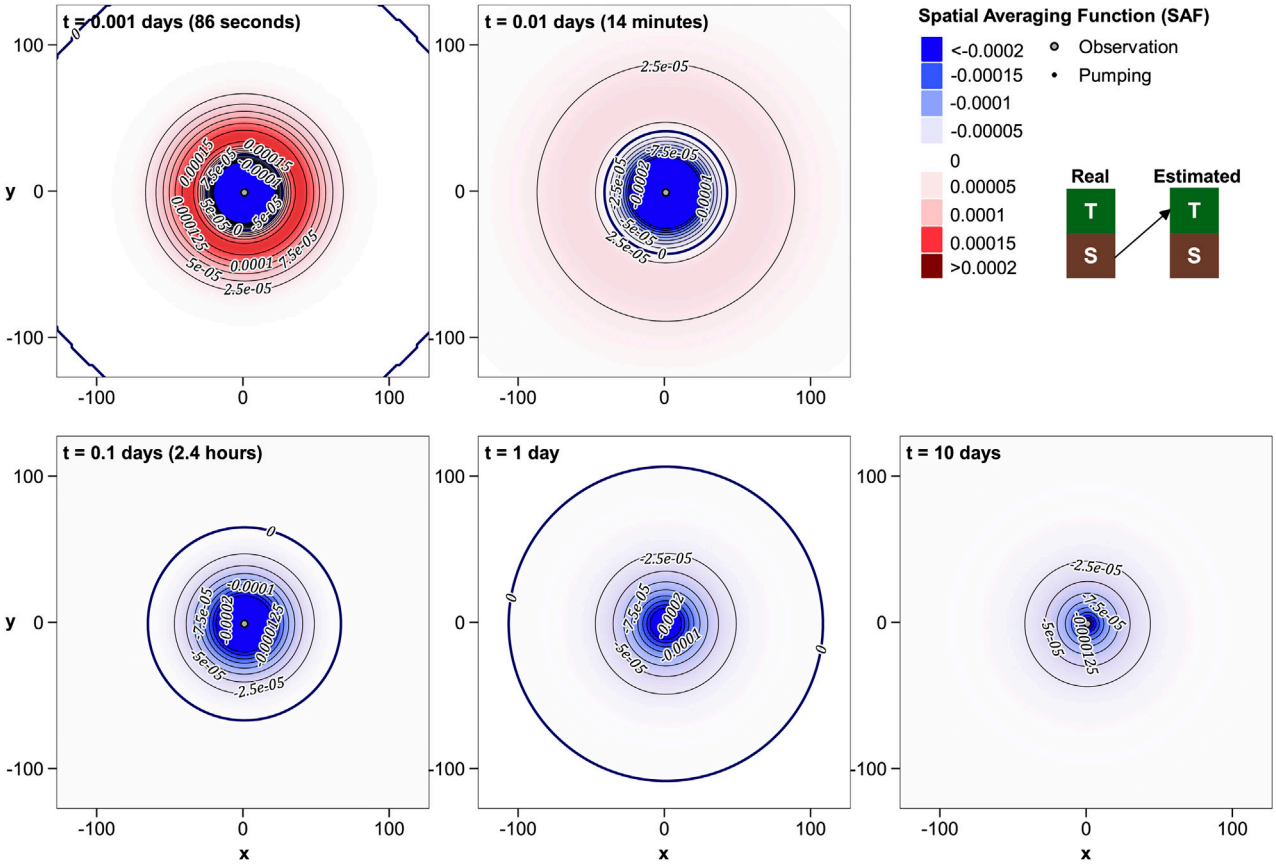

FIGURE 14
$R'_{\underline{T}-S}$ for near-coincident wells at five different times.

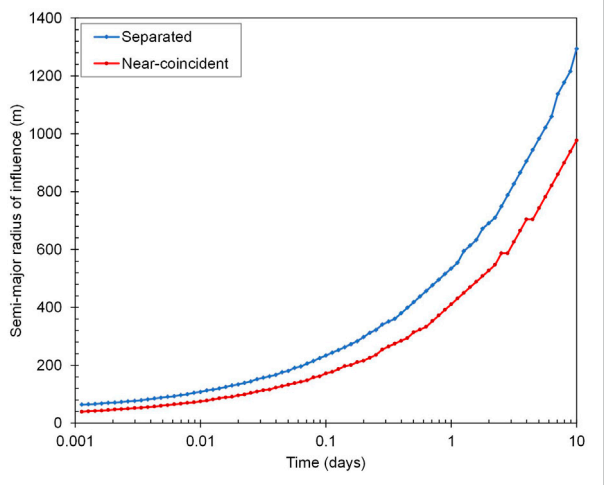

FIGURE 15
Radius of semi-major axis of ellipse of influence for separated and near-coincident wells.

properties. We have formulated spatial averaging functions that map real-world T, T_x, T_y and S to estimated $\underline{T}$ and $\underline{S}$ under the assumption that the latter are estimated by fitting a Theis curve to drawdown measurements. The analysis can be readily extended to accommodate the fitting of other types of curves to raw or processed drawdowns. These averaging functions are approximate, for their formulation assumes only small departures from hydraulic property uniformity. Nevertheless, the insights that they provide are highly instructive.

Values of $\underline{T}$ and $\underline{S}$ that emerge from interpretation of aquifer test data can be viewed as complex spatial averages of real-world T and S over an area that expands with time. These averaging functions cross parameter boundaries. Hence local anomalies in real-world T can influence the estimated value of $\underline{S}$ and *vice versa*.

The area over which real-world hydraulic properties are averaged to estimate $\underline{T}$ is greater than that over which they are averaged to estimate $\underline{S}$. The difference between these two areas grows with pumping time. Hence aquifer-test-estimated $\underline{S}$

tends to reflect real-world S in closer proximity to the pumping and observation wells than estimates of $\underline{T}$ reflect real-world T. At the same time, estimates of $\underline{S}$ are vulnerable to corruption by near-well anomalies in T. The reverse applies for pumping tests of short duration; that is, estimates of $\underline{T}$ can be corrupted by near-well anomalies in S. Importantly however, as the time over which drawdown data are acquired increases and the area over which T is averaged expands, opportunities for contamination of estimated $\underline{T}$ by S decrease. During this period the relationship between drawdown and the log of time approaches linearity, and direct estimates of $\underline{T}$ are available through Cooper-Jacob (1946) analysis of the slope of the drawdown line. It follows that estimates of $\underline{S}$ forthcoming from aquifer test analysis are often less reliable than those of $\underline{T}$, and may be somewhat dependent on the disposition of T between the pumping and observation wells. This is not a new conclusion; see, for example, Sanchez-Vila et al. (1999) and Trinchero et al. (2008).

Intuition may suggest that separation of pumping and observation wells yields insights into between-well T. Analyses that are presented herein show that this applies for only a very short time. Thereafter, significant contributions to interpreted $\underline{T}$ are made by material that is beyond both of the extraction and measurement wells. The longer an aquifer test proceeds, the less does the separation of the wells, or the material between them matter, and the more does the zone of expanding contribution of real T to inferred $\underline{T}$ expand into an area that surrounds both of these wells.

At no time during an aquifer test does estimated $\underline{T}$ reflect real-world T_x more than real-world T_y regardless of the offset direction of the observation well with respect to the pumping well. However, the manner in which distributed T_x and T_y are spatially averaged is very different. At very early times, estimated $\underline{T}$ is positively influenced by T_x along the line that joins the wells. However, it is also negatively influenced by both T_x and T_y to the north and south of this line. As time goes on, estimated $\underline{T}$ is much more reflective of the component of real-world T that points towards the midpoint of the wells than it is of the component of T in any other direction.

Once a certain amount of time has elapsed, the radius of investigation of a separated-well aquifer test becomes greater than that of a coincident-well aquifer test. The ratio of the two investigation radii continues to grow thereafter. The greater averaging area for separated wells protects estimated $\underline{T}$ from contamination by anomalies in near-well T and S. However, it renders it more vulnerable to the effects of hydrogeological boundaries.

We close the discussion by noting that this paper does not address uncertainties of estimated $\underline{T}$ and $\underline{S}$. It would not be a difficult matter to derive expressions for these uncertainties using the theory presented above. However this would require statistical characterisation of the spatial heterogeneity of subsurface T and S, including the scale of this heterogeneity. This is beyond the scope of the present paper. Furthermore, it can be argued that characterisation of the uncertainties of the complex spatial averages of T and S that are depicted herein may be of limited use to managers of a groundwater system. Of greater use are the uncertainties of arithmetic or geometric averages of system properties over user-specified areas, for example, circles or ellipses that circumscribe the pumping test wells, or the cells of a groundwater model grid that spans the area affected by pumping-induced drawdowns. These too can be calculated through a simple extension of the theory provided herein; this will be addressed in future work.

5 Conclusion

The relationships between aquifer-test-inferred transmissivity and storativity and those that prevail in a heterogeneous real world are complex, and sometimes non-intuitive, particularly at early pumping times.

Insights into these relationships provided by the present study can inform the design of an aquifer test. A matter of particular interest at some sites may be whether observation wells should be specially drilled so that pumping-induced drawdowns can be measured at one or a number of distances and directions from the pumped well. The incentive for such a design may be the gathering of information on near-well hydraulic property heterogeneity.

Results presented herein suggest that if insights into near-well hydraulic property heterogeneity are sought, then there is no need to pump for a long time, for this information emerges early in an aquifer test. They also suggest that considerable sophistication is required in interpreting multi-well drawdown data if this information is to be retrieved. This sophistication extends well beyond Theis-based analysis of drawdown in individual wells; the complex averaging functions that link real-world hydraulic properties to Theis-estimated $\underline{T}$ and $\underline{S}$ hide more than they reveal.

An issue of considerable importance is how estimates of transmissivity made through historical interpretation of aquifer test data should be used in parameterisation of a groundwater model. Before construction of a groundwater model, the outcomes of previous hydrogeological investigations that have been undertaken within its domain are generally reviewed. They often reveal that many aquifer tests have been conducted in the study area, some involving a single well, and some involving one or multiple observation wells. In most cases, drawdowns have been interpreted using Theis of Jacob-Cooper analysis.

The present study suggests that estimates of local transmissivity and storativity (particularly the latter) obtained in this way should be treated with caution. The spatial averaging

of local hydraulic properties that is implied in these estimates may preclude their direct transferral to proximal cells of a groundwater model. At the same time, these estimates should not be ignored. An advantage of using linear analysis to establish the relationship between estimated and real-world hydraulic properties, is that an extension of this analysis can provide estimates of error variance between hydraulic properties averaged over model cells and those obtained through aquifer test interpretation. Probabilistic parameterisation of proximal groundwater model cells can then follow. This is the subject of an ensuing paper.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

NM designed the analytical models, wrote software to generate the Frechet kernels/spatial averaging function and interpreted the results. He created all figures and wrote the majority of the paper. JD provided guidance on the concepts, linear algebra and Fortran functions to calculate the spatial averaging function. He assisted with interpreting the results and writing the paper. PH provided guidance on the analytical model, helped interpret the results, and assisted writing/reviewing the paper.

Funding

This study was financially supported by the Groundwater Modelling Decision-Support Initiative (GMDSI—https://gmdsi.org/).

Conflict of interest

Author JD was employed by Watermark Numerical Computing.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2022.1079287/full#supplementary-material

References

Aster, R. C., Borchers, B., and Thurber, C. H. (2019). *Parameter estimation and inverse problems*. 3rd. Amsterdam, Netherland: Springer.

Bresciani, E., Shandilya, R. N., Kang, P. K., and Lee, S. (2020). Well radius of influence and radius of investigation: What exactly are they and how to estimate them? *J. Hydrology* 583, 124646. doi:10.1016/j.jhydrol.2020.124646

Butler, J. J. (1988). Pumping tests in nonuniform aquifers - the radially symmetric case. *J. Hydrology* 101, 15–30. doi:10.1016/0022-1694(88)90025-x

Butler, J. J. (1990). The role of pumping tests in site characterization: Some theoretical considerations. *Ground Water* 28 (3), 394–402. doi:10.1111/j.1745-6584.1990.tb02269.x

Cooley, R., and Christensen, S. (2006). Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media. *Adv. Water Resour.* 29, 639–656. doi:10.1016/j.advwatres.2005.07.012

Cooley, R. L. (2004). *A theory for modeling ground-water flow in heterogenous media*. US Geological Survey Professional Paper, Washington, DC, USA, 1679, 220.

Cooper, H. H., and Jacob, C. E. (1946). A generalized graphical method for evaluating formation constants and summarizing well-field history. *Am. Geophys. Union* 27 (4), 526–534. doi:10.1029/tr027i004p00526

Copty, N. K., Trinchero, P., and Sanchez-Vila, X. (2011). Inferring spatial distribution of the radially integrated transmissivity from pumping tests in heterogeneous confined aquifers: Analysis of pumping tests in confined aquifers. *Water Resour. Res.* 47 (5). doi:10.1029/2010wr009877

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models*. Brisbane, Australia: Watermark Numerical Computing.

Draper, N. R., and Smith, H. (1998). *Applied regression analysis*. 3rd. Wiley-Interscience, Hoboken, NJ, USA.

Knight, J. H., and Kluitenberg, G. J. (2005). Some analytical solutions for sensitivity of well tests to variations in storativity and transmissivity. *Adv. Water Resour.* 28 (10), 1057–1075. doi:10.1016/j.advwatres.2004.08.018

Kruseman, G. P., and de Ridder, N. A. (1990). *Analysis and evaluation of pumping test data*. 2nd. Wageningen, Netherland: International Institute for Land Reclamation and Improvement.

Leven, C., and Dietrich, P. (2006). What information can we get from pumping tests?-comparing pumping test configurations using sensitivity coefficients. *J. Hydrology* 319 (1-4), 199–215. doi:10.1016/j.jhydrol.2005.06.030

Menke, W. (2018). "Geophysical data analysis," in *Discrete inverse theory*. 4th (Academic Press) Cambridge, MA, USA.

Moore, C., and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Adv. Water Resour.* 29 (4), 605–623. doi:10.1016/j.advwatres.2005.07.003

Oliver, D. S. (1990). The averaging process in permeability estimation from well-test data. *SPE Form. Eval.* 5 (3), 319–324. doi:10.2118/19845-pa

Oliver, D. S. (1993). The influence of nonuniform transmissivity and storativity on drawdown. *Water Resour. Res.* 29 (1), 169–178. doi:10.1029/92wr02061

Papadopulos, I. S. (1965). *Nonsteady flow to a well in an infinite anisotropic aquifer: Proceedings of the dubrovnik symposium on the hydrology of fractured rocks*. Wallingford, UK: International Association of Scientific Hydrology.

Sánchez-Vila, X., Meier, P. M., and Carrera, J. (1999). Pumping tests in heterogeneous aquifers: An analytical study of what can be obtained from their interpretation using Jacob's Method. *Water Resour. Res.* 35 (4), 943–952. doi:10.1029/1999wr900007

Theis, C. V. (1935). The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using groundwater storage. *Am. Geophys. Union Trans.* 16, 519–524. doi:10.1029/tr016i002p00519

Trinchero, P., Sanchez-Vila, X., and Fernandez-Garcia, D. (2008). Point-to-point connectivity, an abstract concept or a key issue for risk assessment studies? *Adv. Water Resour.* 31, 1742–1753. doi:10.1016/j.advwatres.2008.09.001

Zha, Y., Shi, L., Liang, Y., Michael Tso, C.-H., Zeng, W., and Zhang, Y. (2020). Analytical sensitivity map of head observations on heterogeneous hydraulic parameters via the sensitivity equation method. *J. Hydrology* 591, 125282. doi:10.1016/j.jhydrol.2020.125282

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/feart.2022.886304

OPEN ACCESS

EDITED BY
Michael Fienen,
United States Geological Survey,
United States

REVIEWED BY
Kevin Hayley,
Groundwater Solutions Pty., Australia
Jonathan P. Traylor,
United States Department of the Interior,
United States

*CORRESPONDENCE
Eduardo R. De Sousa,
edesousa@intera.com

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 28 February 2022
ACCEPTED 19 December 2022
PUBLISHED 10 January 2023

CITATION
De Sousa ER, Hipsey MR and Vogwill RIJ (2023), Data assimilation, sensitivity analysis and uncertainty quantification in semi-arid terminal catchments subject to long-term rainfall decline.
Front. Earth Sci. 10:886304.
doi: 10.3389/feart.2022.886304

Data assimilation, sensitivity analysis and uncertainty quantification in semi-arid terminal catchments subject to long-term rainfall decline

Eduardo R. De Sousa[1]*, Matthew R. Hipsey[2] and Ryan I. J. Vogwill[3]

[1]INTERA Inc., Perth, WA, Australia, [2]Centre for Water and Spatial Science, UWA School of Agriculture and Environment, The University of Western Australia, Perth, WA, Australia, [3]Hydrogeoenviro Pty Ltd., Perth, WA, Australia

Quantification of long-term hydrologic change in groundwater often requires the comparison of states pre- and post-change. The assessment of these changes in ungauged catchments using numerical models and other quantitative methods is particularly difficult from a conceptual point of view and due to parameter non-uniqueness and associated uncertainty of quantitative frameworks. In these contexts, the use of data assimilation, sensitivity analysis and uncertainty quantification techniques are critical to maximize the use of available data both in terms of conceptualization and quantification. This paper summarizes findings of a study undertaken in the Lake Muir-Unicup Natural Diversity Recovery Catchment (MUNDRC), a small-scale endorheic basin located in southwestern Australia that has been subject to a systematic decline in rainfall rates since 1970s. A combination of data assimilation techniques was applied to conceptual and numerical frameworks in order to understand and quantify impacts of rainfall decline on the catchment using a variety of metrics involving groundwater and lake levels, as well as fluxes between these compartments and mass balance components. Conceptualization was facilitated with the use of a novel data-driven method relating rainfall and groundwater responses running backwards in time, allowing the establishment of the likely baseline conditions prior to rainfall decline, estimation of net recharge rates and providing initial heads for the forward numerical modelling. Numerical model parameter and predictive uncertainties associated with data gaps were then minimized and quantified utilizing an Iterative Ensemble Smoother algorithm, while further refinement of conceptual model was made possible following results from sensitivity analysis, where major parameter controls on groundwater levels and other predictions of interest were quantified. The combination of methods can be considered as a template for other long-term catchment modelling studies that seek to constrain uncertainty in situations with sparse data availability.

KEYWORDS

groundwater modelling, data assimilation (DA), sensitivity analysis (SA), uncertainty quantification (UQ), endorheic basins

1 Introduction

Endorheic basins, also known as terminal catchments of internally drained basins, comprise a variety of geomorphic environments widely distributed across the globe. These environments with often distinct attributes are characterized by the lack of surface and groundwater discharge across their borders and have evaporation as the dominant outflow. Despite many of them being currently under pressure from climate change and anthropogenic activities, these basins are understudied when compared to traditional hillslope catchments. The unique surface and groundwater attributes (De Sousa, 2021) of these environments makes extrapolations and inferences from better studied hillslope catchments difficult. From a quantitative perspective, modelling efforts in these areas are not only difficult given the lack of numerical frameworks designed for surface-groundwater interactions in semi-arid settings (Jolly et al., 2008), but they also suffer from a lack of site-specific monitoring data.

The issue of data sparsity is a common theme in endorheic basins, evidenced by the fact that current literature relies heavily on indirect measurements and remote-sensing data (De Sousa, 2021). Data sparsity hinders the ability to establish a robust conceptualization and incurs large predictive uncertainty which is inherent to surface and groundwater models. The need to maximize the extraction of information from largely incomplete datasets and its use in conceptualization and numerical modelling is critical for the development of quantitative frameworks that are capable of accounting for hydrogeological/hydrological uncertainty and the ability of available data to constrain it.

The use of Data Assimilation (DA), Uncertainty Quantification (UQ) and Sensitivity Analysis (SA) techniques in hydrological modelling is an emerging field with great potential to support decision-making in catchments experiencing hydrological change, but to date many of these techniques have not yet been applied to endorheic basin studies. Challenges remain on how to apply them appropriately in situations where the observation data is less than ideal, such that they can output useful information relevant to inform our conceptualization and strategies for management (Thompson et al., 2015).

Research into DA was initially developed for the purpose of numerical weather prediction, and is often related to Kalman filter contexts, where the states of variables from numerical models are updated incrementally through time as new observation data becomes available. In this paper, we adopt a broader definition of DA, which relates to optimally combine observations with theory (usually as numerical models) to improve model integrity and the accuracy of predictions of interest (Asch et al., 2016). In this regard, DA techniques are used for several purposes, such as history matching and parameter optimization based on observed data, determination of initial conditions for a numerical forecast model, interpolation of sparse observation datasets using the physical knowledge of the system (i.e., numerical models), and reduction of predictive uncertainty of numerical models.

The use of UQ and SA techniques is often interrelated with DA techniques. While UQ tends to focus on quantifying and reducing parameter and predictive uncertainty due to lack of data or model defects, SA looks at the effect that model parameters have on outputs of interest (Pianosi et al., 2016). These techniques have the potential to support many of the questions that arise from investigation efforts in endorheic basins, from conceptualization to quantification, predictive modeling, and adaptive management (Figure 1), which are explored next.

1.1 Finding evidence of long-term groundwater trends and reconstruction of baseline conditions

The use of signal analysis techniques for processing of time series and extraction of useful information is an import area of signal processing and well-established techniques such as Fourier Transforms and Wavelets have been applied for decades (Maheshwari and Kumar, 2014). Studies focused on groundwater level time series analysis include works by Lafare et al. (2016) and Seeboonruang (2014).

In situations where long-term stressors are intermixed with short-term signals (such as seasonality), time decomposition techniques have the potential to untangle them. The Empirical Model Decomposition method developed by Huang et al. (1998) is a technique for processing of non-linear and non-stationary signals and/or time series, decomposing them into a number of zero-mean signals called Intrinsic Mode Functions (IMF) in an adaptive and fully data-driven way, from the assumption that any signal is composed by different IMF's and that each IMF represents a characteristic oscillation on a separated time scale. The EMD technique have been used in the hydrology field for identification of trends in lake levels (Wang et al., 2020) and groundwater forecasting (Gong et al., 2018).

Another challenge in the study of endorheic basins under long-term impacts is the establishment of pre-impact baseline conditions. For impact assessment in general, the definition of impacts often involves comparison of current and past hydrologic states. The absence of baseline data in these circumstances makes the definition of these impacts difficult both in terms of conceptualization and quantification. The Backward Water Table Fluctuation (BWTF) developed by De Sousa (2021) is a data-driven hindcasting technique based on rainfall and groundwater level fluctuations, enabling the reconstruction of baseline groundwater levels for periods where no monitoring data was available.

1.2 Data sparsity and uncertainty

Data sparsity in the endorheic basins reduces the reliability of investigation efforts, both in terms of identification of dominant processes (conceptualization) and predictive ability of quantitative frameworks.

More conventional applications of data assimilation involve the use of history-matching techniques to attempt the reduction of parameter uncertainty (and possibly predictive uncertainty). Excellent discussions on history-matching, data assimilation and their value in the reduction of uncertainty are presented by Nicols and Doherty (2020) and Gallagher and Doherty (2020). In these discussions, history matching is defined as the "act of tuning model parameters so that a model can reproduce past system behavior." Predictive uncertainty in hydrologic models is often expressed from a conceptual point of view using Bayes equation, where imposition of constraints on parameter values is obtained through history-matching. Where approximations of prior distributions are derived

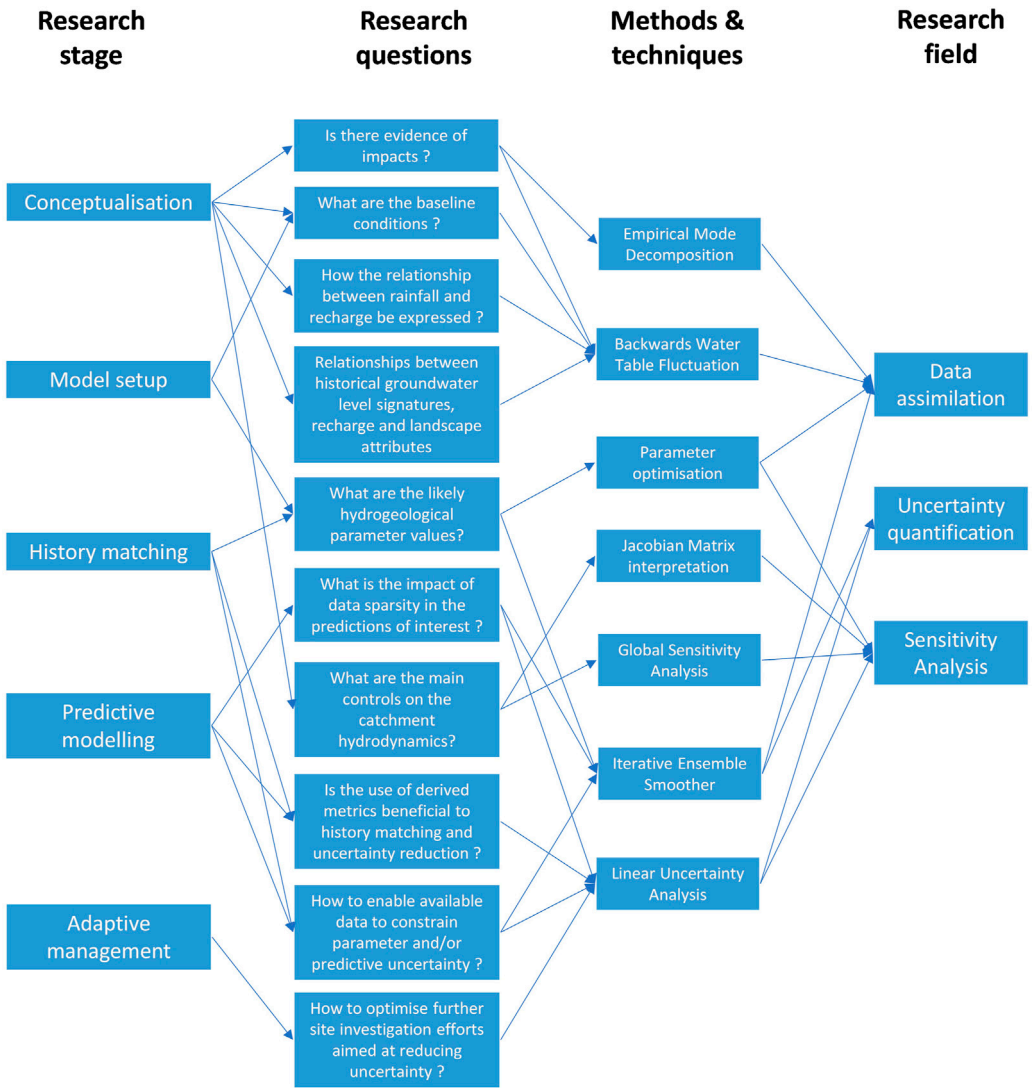

FIGURE 1
Example of DA, UQ and SA techniques, and their potential to support research in endorheic basins.

from the conceptual understanding and expert-knowledge of a system (also known as soft-data), history-matching against field measurements (hard-data) induce alterations to the prior parameter probability distributions and, consequently, predictive uncertainty. The resulting probability distribution from parameter sets that both conform with expert-knowledge and reproduce historical behavior approximates the posterior probability distribution.

As discussed in Nicols and Doherty (2020), mathematical expressions for posterior parameter and predictive probability distribution cannot be derived, however, they can be defined by sampling them. The PESTPP-IES iterative ensemble smoother (White, 2018) was designed to this end. Based on the algorithm described by Chen and Oliver (2013), PESTPP-IES uses ensemble realizations derived from an approximation of the prior parameter probability distribution and attempt to adjust them by the minimum amount required to match field observations and achieve history-matching as well as reduction in parameter uncertainty.

1.3 Maximizing the use of available data sets with derived metrics

In the "Concept-State-Process-System" (CSPS) framework introduced for the hierarchical assessment of aquatic ecosystem models, Hipsey et al. (2020) divides metrics used for history matching of model states into 3 major groups: 1—Direct comparison, where model results are compared with measured data at specific points in time and space; 2—Derived metrics describing model state, which do not involve a direct assessment of a state variable, but are derived from them (such as head differences, or ratios between variables); and 3—Metrics describing multi-scale variability in model state, used to describe how well the various scales of spatial or temporal variability are described in models.

The use of derived metrics involving groundwater head differences in space and time is not new in hydrogeology studies and is recommended by several authors (Hill and Tiedeman, 2007; Doherty et al., 2010). Nevertheless, studies demonstrating and

evaluating the value of derived metrics and how they contribute to reducing uncertainty is not often seen in literature.

1.4 Informing conceptualization and controls on catchment dynamics

Thompson et al. (2015) states that numerical models are important in understanding how complex catchment systems are responding to uncertain changes and while conceptual models usually guide the development of numerical models, iterative cycles between conceptualization and modelling results may be beneficial to refine conceptual understanding.

Saltelli et al. (2004) defines sensitivity analysis as "The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." In groundwater modelling practice, the sensitivity analysis is usually performed at the end of the modelling exercise, or as a by-product of parameter optimization methods used for history matching such as PEST (Doherty, 2015) and PESTPP (Welter et al., 2015). However, much can be gained from sensitivity analysis in terms of conceptualization if these techniques are employed in earlier stages of model development. Despite the fact that sensitivity can be anticipated to some extent by experienced modelers, rigorous analysis is useful to corroborate, indicate the need for new conceptualization, or diagnose non-linear behavior of models and/or numerical instabilities.

Algorithms based on the Gauss-Levemberg-Marquadt method such as PEST (Doherty, 2015, 2020), PESTPP (Welter et al., 2015) often require a tangent linear operator, also known as Jacobian matrix. These matrices contain partial derivatives of model outputs in respect to model parameters and are often required for history matching methods, sensitivity, and linear uncertainty analysis. Detailed analysis of these sensitive matrices can provide valuable insights on system functioning and dominant controls on catchment dynamics (as explored in the following sections). However, they are not often "dissected" and interpreted in context of model conceptualization.

While point-source sensitivities obtained from perturbation methods (such as those obtained by PEST and PESTPP) provide valuable insights on system and model behavior, there are situations where more robust sensitivity estimates are required. For these purposes, the use of global sensitivity analysis (GSA) may be useful. GSA methods characterize the effect of model parameter onto model outputs over a wide range of acceptable parameter values, covering larger portions of parameter space as opposed to point-source sensitivity. As a result, the behavior of model outputs that are non-linear and dependent on the combination of many parameters can be unraveled. Several methods for global sensitivity analysis with different degrees of computational effort and output results, as discussed in Saltelli et al. (2004, 2008).

1.5 Optimizing site investigation efforts

In data-scarce areas and resource-constrained investigations, it is important to collect data where it really matters. From a quantitative perspective, that means where and when an observation will promote the maximum reduction of predictive uncertainty. When a Jacobian sensitivity matrix is calculated for a parameter set that reasonably conforms with expert-knowledge and historical system behavior, it can be used for linear uncertainty analysis, also known as first-order second moment (FOSM) analysis. The theory behind linear uncertainty analysis is widely discussed in the literature (Moore and Doherty, 2006; James et al., 2009; Dausman et al., 2010; White et al., 2014) and it has been implemented in a number of model-independent software packages, including PEST, PESTPP and PyEMU (White et al., 2016).

This method provides an approximate mathematical characterization of prior and posterior probability distributions for parameters and predictions of interest (Nicols and Doherty, 2020). Furthermore, it can be used to demonstrate the value of history matching data (existing or not) in the reduction of parameter and predictive uncertainty. This enables the assessment of data worth not only for different data metrics, but also optimizing data acquisition efforts, by pre-empting its ability to constrain parameter and predictive uncertainty.

1.6 Study objectives and structure

The objective of this study is to apply and demonstrate the use of DA, UQ and SA techniques in the context of endorheic basins research, evaluating the ability of these methods to facilitate and enable conceptualization, quantification, and adaptive management measures. These techniques were applied during research undertaken at the Lake Muir-Unicup Natural Diversity Recovery Catchment (MUNDRC), a small scale semi-arid basin located in southwestern Australia, and subject to a systematic decline in rainfall rates over the past 50 years.

The application of the different techniques presented in this paper was not linear, in the sense they were not necessarily applied in the order they are presented. Multiple feedback loops between assessment of model results and conceptualization were undertaken, evolving the understanding of the site and robustness of quantitative assessments to the final form. The last part of this paper integrates the findings of all techniques and how they contributed to the research development.

2 Study site, conceptual and numerical framework

The area of investigation employed in this study is the Lake Muir-Unicup Natural Diversity Recovery Catchment (MUNDRC), located in southwestern Australia and listed under the Ramsar Convention as a Wetland of International Importance. This area consists of a complex system of lakes, swamps and flood plains, encompassing an area of 630 km^2 and is located 65 km from the coastline (Figure 2).

The conceptual model for the area, main hydrological drivers and effects from rainfall decline on surface and groundwater compartments are discussed in De Sousa (2021). Four hydrogeological units are associated with unconsolidated sediments and weathered portions of the crystalline basement. The combination of low relief, high specific yield, and flat lake bathymetry results in a relatively stagnant groundwater system with respect to horizontal flows along the lower plains surrounding Lake Muir. The flat topography also results in poor development of surface drainage lines, favoring infiltration processes over runoff, as well as the

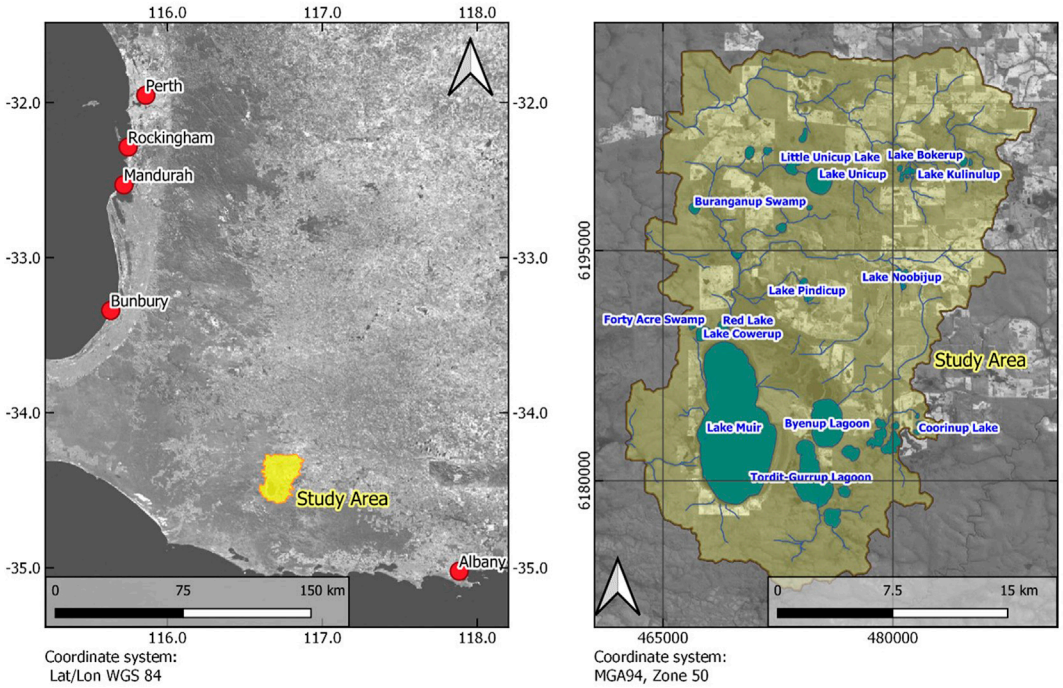

FIGURE 2
Location of study site, Lake Muir and nearby surface water compartments.

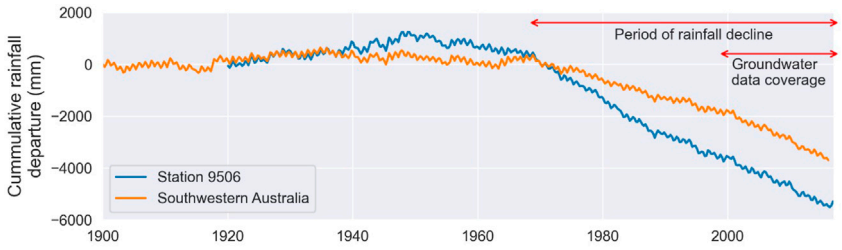

FIGURE 3
Cumulative rainfall departure for Station 9,506 and average southwestern Australia, using rainfall averages for the period of 1900–1970.

development of shallow groundwater tables with high correlation to the topography and significant seasonal oscillations.

Lake Muir is positioned in the lowest topographical area of the catchment and constitutes the largest groundwater discharge area, where water is constantly removed from the lake through evaporation. Water exchanges between the lake and adjacent aquifers is dynamic given the highly variable lake-aquifer interface areas resulting from the flat lakebed geometry.

Long-term rainfall records for the catchment show a systematic decrease in rainfall rates, particularly during the wet seasons. Hope and Foster (2005) analyzed winter rainfall rates in Western Australia for the period of 1925–2005 and identified an abrupt change in rainfall rates since 1970s.

Cumulative rainfall reductions for the MUNDRC have been undertaken using Accumulation Monthly Residual Rainfall (Ferdowsian et al., 2001) and are displayed in Figure 3, showing relatively small departures for the period from 1920–1970, with a pronounced negative departure from 1970 to present, showing a total deficit of 5,500 mm over 46 years. This reduction in rainfall results in smaller groundwater recharge and, consequently, reductions in groundwater levels and discharge volumes.

2.1 Numerical framework

A numerical framework for quantification of long-term impacts associated with rainfall decline on MUNDRC have been developed and described by De Sousa (2021). Aiming at representing the main hydrologic controls in the area and encapsulating both conceptualization and monitoring data, the framework consisted of a dynamically-coupled lake and groundwater model, accompanied by a rainfall-based groundwater recharge formulation.

The three-dimensional groundwater model was built using the finite element code FEFLOW (Diersch, 2014), coupled with a lake

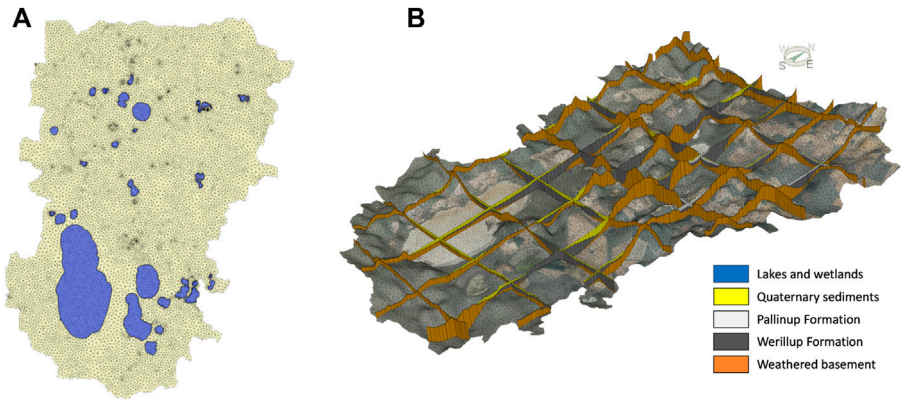

FIGURE 4
Groundwater model mesh and main surface water compartments **(A)**, and fence diagram illustrating the distribution of the main hydrogeological units **(B)**.

model component developed using FEFLOW's API. The model was defined based on catchment boundaries and geometry of the main hydrogeological units in the area (Figure 4).

The model was initially calibrated against lake and groundwater levels, as well as additional derived metrics discussed later in the section below, using piecewise-constant parameter zones, where hydrogeological parameters of the different aquifers were assumed to be homogeneous over their extent.

This numerical framework was upgraded in this paper by adopting a highly-parameterized approach using pilot points (Doherty et al., 2010). The use of pilot points not only allows for heterogeneity within the aquifers, but also builds the foundation for the DA, UQ and SA workflows presented in this paper. Pilot points were assigned covering the extent of each of the aquifer and recharge zones, using a predominantly regular grid with 1,500 m spacing (Figure 5). Each aquifer pilot point was assigned four parameters and two parameters were assigned for each recharge pilot point. Three additional parameters were also implemented to define initial lake level and multipliers for lake evaporation and rainfall, resulting in 3,295 parameters adjusted during history matching and analyzed in the uncertainty quantification and sensitivity analyses (Table 1). In order to facilitate the discussion, the following convention was used for parameter naming:

Ptype _Pzone _PPID

where Ptype is the type of parameter, Pzone relates to the original parameter zones defined De Sousa (2021) and PPID correspond to the pilot point number for the spatially distributed parameter groups. Parameters related to the lake model have the prefix "Lk_" and descriptor for parameter types and zones are presented in Tables 1, 2, respectively. Corresponding parent groups for each of the parameters are named with Ptype followed by Pzone.

2.2 History matching data, derived metrics, and predictions of interest

The DA, UQ and SA techniques employed in this study were based on historical lake and groundwater levels. These levels have been measured in Lake Muir and groundwater monitoring boreholes drilled as part of hydrogeological investigations by the former Western Australia Department of Land Management (New et al., 2004), and further expanded by former Department of Environment and Conservation (Grelet and Smith, 2009). These groundwater monitoring boreholes have been screened to target the different hydrogeological units in the area and are displayed in Figure 6, along with selected locations from which model results are presented and discussed.

The majority of the data have been collected on a monthly basis from early 2000s, while historical level measurements on Lake Muir have been conducted since 1980s, mostly during the wet seasons. In the context of impacts related to rainfall decline, the dataset coverage is relatively small, since the decline period started in the early 1970s.

The MUNDRC model used several direct and derived history-matching and predictive metrics, as summarized in Table 3. Direct metrics are defined here as values related to direct model outputs that do not require further post-processing (or in other words, raw output), while derived metrics are based on post-processing of model outputs (such as head-differences). The use of different metrics was three-fold: attempt to improve history matching, reduce parameter and predictive uncertainty, and to understand how they respond to parameter changes and contribute to the reduction of uncertainty.

Horizontal head differences between boreholes were added as observations, based on a Delaunay triangulation generated from the borehole locations. Head differences between borehole pairs defining each of the triangulation edges were use as observations, with quarterly snapshots generated for the period from 2000–2014. Seasonal head differences within each borehole were also included, in an attempt to inform the optimization process of groundwater level differences between wet and dry seasons. Groundwater level estimates for 1970 presented in De Sousa (2021) were included, together with the difference between these levels and the first record of each borehole, in an attempt to inform long-term changes.

In addition, predictive metrics have been added in for the sensitivity analysis workflows in the form of "virtual observations" (i.e., fake observations at prescribed locations in space-time), for sensitivity analysis. Virtual monthly groundwater level observations have been added for all boreholes, covering the period from 1970 to

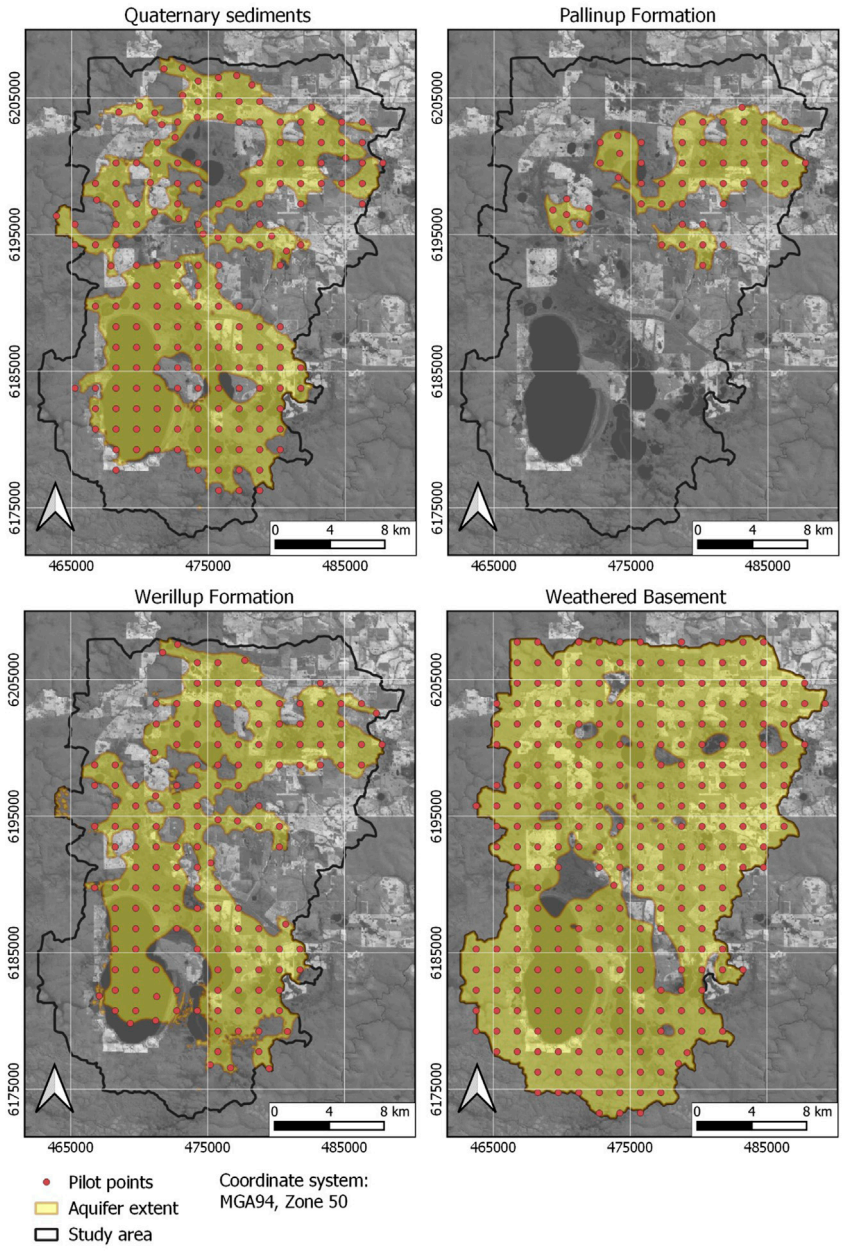

FIGURE 5
Distribution of pilot points for the different hydrogeological units.

2010 to observe whether sensitivity to groundwater levels vary over time. Virtual observations for lake levels were also included, in addition to monthly net recharge rates and exchange fluxes between Lake Muir and adjacent aquifers.

3 Data assimilation in conceptualization

3.1 Identifying rainfall decline trends in groundwater monitoring data

Given the strong correlation between rainfall rates and groundwater levels, it was expected that drawdown trends associated to rainfall decline would be strongly present in the historical data, however, the identification of long-term drawdowns in the available monitoring data is very subtle and difficult to undertake. Reasons for that include the relatively short monitoring period (15-year against 40–50 years of rainfall decline) and also that the consistent rainfall decline for such a long period may lead the catchment towards a new equilibrium state, with groundwater levels "adapting" to the lower recharge regime. Lastly, the high seasonality observed in groundwater levels add short-term variations in monitoring records masking subtler long-term drawdown trends.

Here, the EMD method have been applied to the groundwater time series from monitoring boreholes to isolate long-term drawdown terms from seasonal and higher frequency changes in groundwater levels.

TABLE 1 Parameter type descriptors used in parameter name definition.

Parameter symbol	Description	Spatially distributed	Units
Kh	Horizontal hydraulic conductivity	Yes	m/d
Va	Vertical anisotropy	Yes	(-)
Ss	Specific storage	Yes	m^{-1}
Sy	Specific yield	Yes	(-)
Ra	Rainfall fraction for recharge formulation	Yes	(-)
Ev	Constant outflow term for recharge formulation	Yes	(mm/day)
Lk_strt	Lake Muir starting level (1960)	No	mAHD
Lk_evap	Lake Muir evaporation multiplier	No	(-)
Lk_rain	Lake Muir rainfall multiplier	No	(-)

TABLE 2 Parameter group descriptors used in parameter name definitions.

Parameter group	Aquifer/recharge zone	Description
1	Aquifer	Quaternary sediments
2	Aquifer	Pallinup Formation
3	Aquifer	Werillup Formation
4	Aquifer	Weathered Basement
8	Recharge	Sedimentary aquifer outcrop zone
9	Recharge	Weathered basement outcrop zone

Groundwater time series from all monitoring boreholes were processed using the Python implementation of the EMD algorithm developed by Laszuk (2017). The original data and resulting IMF's for monitoring boreholes MU22A, MU45S and MU65S are presented in Figure 7. The EMD results show long-term decline terms in the last IMF of all three boreholes (IMF_6 for MU22A, IMF_5 for boreholes MU45S and MU65S), while seasonal signals for the three boreholes are clearly identified in IMF's 3, 1 and 2, respectively.

Another important aspect is that the magnitude of the long-term variations found by the EMD are much smaller than seasonal variations in the groundwater levels and higher-frequency IMF's, demonstrating the ability of the method to identify subtle drawdown patterns in areas under high seasonality effects.

3.2 Establishment of baseline groundwater levels and conceptual drawdown estimates

The assessment of environmental impacts often involves the comparison of current and/or past hydrologic states against states prior to the impact development. The absence of baseline data in these situations makes the assessment of these impacts extremely challenging, both on conceptual and quantitative levels.

In the MUNDRC, while groundwater monitoring data was available for a large number of boreholes spread across the catchment, the majority of groundwater level data was collected from the early 2000s approximately 30 years from the beginning of the rainfall decline. From that perspective, besides subtle drawdown trends observed in the EMD analyses, the premise that rainfall decline promoted groundwater drawdown in the catchment was merely conceptual.

Based on relationships between rainfall and groundwater responses, the Backwards Water Table Fluctuation (BWTF) method was developed. This technique enabled the reconstruction of groundwater levels in the MUNDRC prior to rainfall decline by running the calculations backwards in time and providing reverse hindcasts. Historical groundwater levels were estimated for each borehole in the catchment, utilizing a starting head (equating to the latest observation of each monitoring time series), rainfall fraction applied to rainfall historical time series, specific yield, and a constant outflow term. These parameters were calibrated against available data and ran backwards until 1970, prior to rainfall decline.

The pre-rainfall decline hindcasts obtained from this method provided not only preliminary drawdown estimates across the catchment, but also estimates on net recharge rates. Furthermore, the estimated groundwater levels were incorporated in the forward numerical framework (as discussed in history matching metrics), therefore enabling the history matching to reach for reasonable groundwater levels pre-rainfall decline and provide more robust estimates for groundwater level changes since 1970.

4 Uncertainty quantification and the role of history matching

4.1 Reducing uncertainty

The Iterative Ensemble Smoother implementation in PESTPP-IES have been utilized for history-matching aiming at 1)—reasonably

10.3389/feart.2022.886304

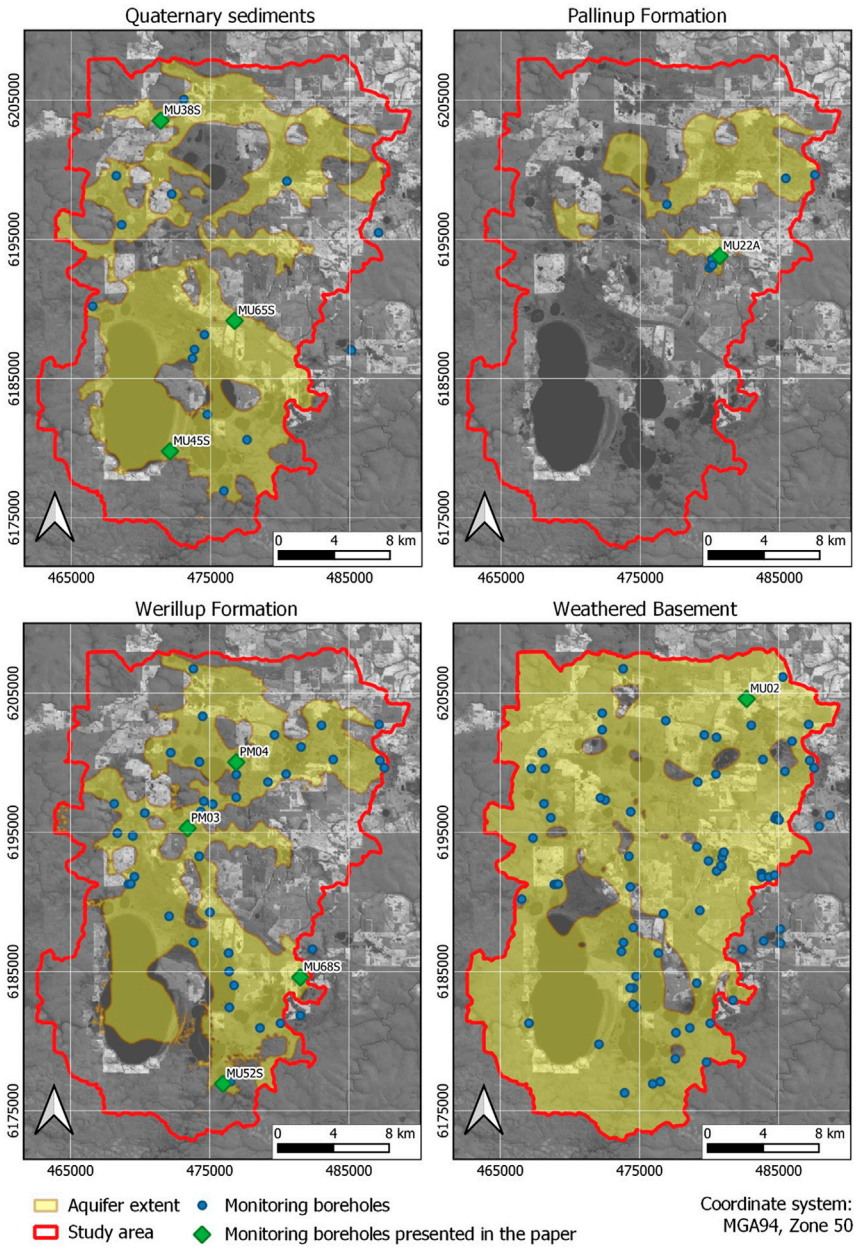

FIGURE 6
Location of groundwater monitoring boreholes screened in the different hydrogeological units.

representing historical behavior of the catchment, 2)—reduce parameter uncertainty, 3)—provide an ensemble of conceptually feasible parameter sets that are a reasonable approximation of posterior parameter distributions and 4)—enable the quantification of uncertainty of the different metrics and predictions of interest.

PESTPP-IES was used in conjunction with the numerical model described in the previous section and upgraded here with pilot-point parameterization. An ensemble of 150 realizations was constructed based on conceptual information of the site, which included likely parameter values, upper and lower bounds, and conceptual estimates of spatial correlation. The prior ensemble included a base realization, which served as a parameter means for the remaining realizations. This realization was based on the piecewise-zone calibration presented in De Sousa (2021).

Prior information was applied in the generation of the parameter sets through the definition of parameter means and boundaries as well as their spatial correlations. Upper and lower boundaries for aquifer parameter values were based on lithological descriptions and literature values (Reynolds and Marimuthu, 2007) of each of the aquifers. Parameter bounds for recharge outflow terms were set .7 to 1.2 while values between .5–.9 were assigned to rainfall fraction. The spatial correlation of parameters of the same type and group was defined using covariance matrices based on variograms defined for the area, under the assumption that spatial correlation is lost beyond the distance of 6 km.

TABLE 3 System observation metrics used in history matching, and predictive metrics used in the UQ and SA workflows.

Name	Description	Derived/ direct	History matching/ Sensitivity analysis	Temporal coverage	Number of sites	Number of observations
Heads	Historical groundwater levels at monitoring boreholes	Direct	History matching	1997–2017	103	17,735
BWTF_Heads	1970 Groundwater level estimates using the BWTF method	Derived	History matching	1970	57	57
HGrads	Horizontal historical head differences between boreholes	Derived	History matching	2000–2014	31[a]	17,272
BWTF_TGrads	Head differences between 1970 BWTF estimates and first observation of each borehole	Derived	History matching	1970–2008	57	57
T_Grads	Seasonal head differences in boreholes	Derived	History matching	2000–2018	107	1,655
Lake levels	Historical Lake Muir levels	Direct	History matching	1980–2010	1	58
Historical heads	Virtual monthly groundwater levels at monitoring bores	Direct	Sensitivity analysis	1970–2010	103	49,543
Historical lake levels	Virtual monthly levels for Lake Muir	Direct	Sensitivity analysis	1970–2017	1	576
Net recharge	Monthly net recharge rates for the entire model domain	Direct	Sensitivity analysis	1970–2017	2[b]	576
Lake Muir/GW fluxes	Exchange fluxes between Lake Muir and adjacent aquifers	Direct	Sensitivity analysis	1970–2017	1	576

[a]Number of quarterly time snapshots.
[b]Net recharge over the entire zones 8 (sedimentary aquifers) and 9 (weathered basement).

Details from the PESTPP-IES settings used in this calibration, including regularization settings and use of localization, are discussed in De Sousa (2021). The results from the history matching showed a reduction in the mean objective function value from 484,555 to 28,936 at the end of iteration 6. The standard deviation of the objective function values also reduces throughout the iterations, from a prior value of 583,308 to 641 at the end of optimization. Furthermore, the total number of runs required for the entire procedure (prior plus 6 iterations) was 1842, which is about half the number of adjustable parameters and demonstrates the efficiency of the algorithm. Despite similar objective functions obtained at the end of the IES optimization, the distribution of parameters from the different ensemble sets can be quite distinct, as illustrated by horizontal conductivity values in layer one displayed in Figure 8. Although the differences between realizations in this figure are not always apparent (since the color-scale span 4 orders of magnitude), the histograms for hydraulic conductivity displayed in Figure 10 show clearly that ranges often span over one order of magnitude.

To illustrate the reduction in uncertainty, simulated hydrographs from the prior and posterior ensemble were plotted against observed data and the results from the piecewise-constant zone calibration presented in De Sousa (2021) for selected monitoring locations and Lake Muir (Figure 9). When compared to the prior ensemble, the posterior realizations not only present a better fit with observed data, but also have a much narrower spread (therefore demonstrating the reduction of parameter uncertainty). In relation to the piecewise-constant calibration, posterior runs also present a significantly better fit, which is expected as the highly-parameterized form allows for heterogeneity within the parameter zones and adjust locally to each monitoring location. Lastly it can be observed that prior realizations are predominantly centered around the hydrographs from the piecewise-constant calibration run, which is expected as the parameters from this run were used as means for the generation of the prior ensemble set.

The reduction of uncertainty can also be observed when comparing parameter distributions from prior and posterior ensembles (Figure 10). In this figure, most sensitive parameters to groundwater level observations from selected hydrographs have been selected for plotting of histograms with prior and posterior distributions. It can be observed that there is an overall reduction in the spread of parameter values, and these reductions are particularly pronounced in highly sensitive parameters, such as those related to groundwater recharge. Parameters with low sensitivity such as vertical anisotropy shows little to no reduction in parameter uncertainty.

4.2 Quantification of predictive uncertainty

If the ensemble size utilized by the iterative ensemble smoother is of sufficient size, the execution of model runs using the posterior ensemble sets can be utilized collectively to define posterior probability distributions for predictions of interest.

The results from the posterior ensemble sets obtained using PESTPP-IES were used to assess uncertainty of predicted groundwater levels and other metrics as illustrated in Figures 11 and 12.

Despite the large number of locations and metrics, some common uncertainty patterns can be observed across different model results. For instance, uncertainty of simulated groundwater levels shows an increase during the period from 1980 to 1988 in many borehole locations such as MU68S and PM03. For these boreholes the maximum simulated drawdown occurs in this period, which

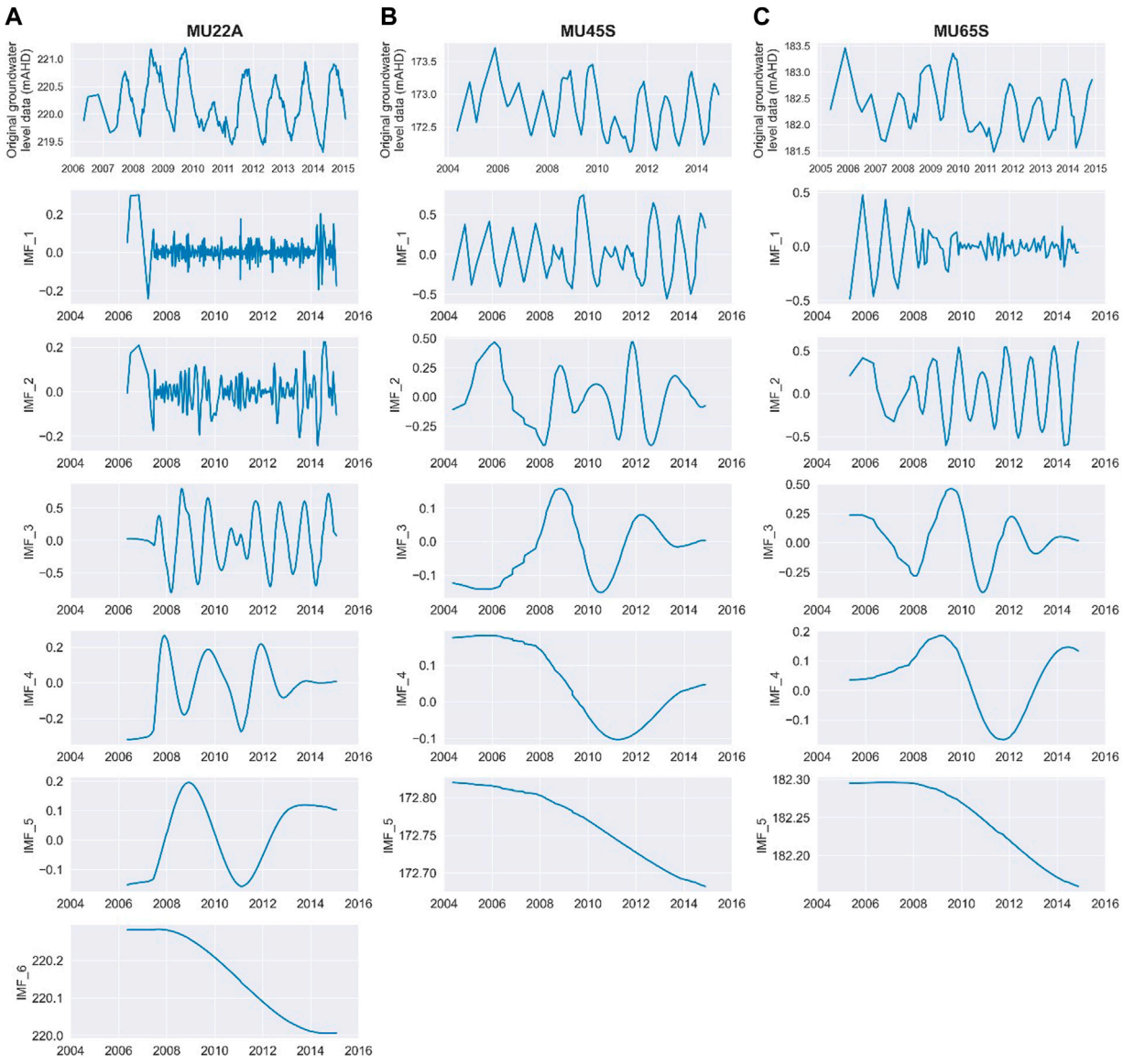

FIGURE 7
Historical groundwater levels and Intrinsic Mode Functions obtained from the EMD analysis for boreholes MU22A **(A)**, MU45S **(B)** and MU65S **(C)**.

suggest the degree of uncertainty is related to the magnitude of stresses being imposed in the catchment. The distribution of mean and standard deviation of simulated groundwater levels corroborates this hypothesis, with areas of larger standard deviation predominantly overlapping areas of larger mean drawdown.

The uncertainty around groundwater levels is relatively small when considering the absolute groundwater levels, with 95% confidence intervals under ±1 m from mean levels for the majority of monitoring boreholes. Nevertheless, when looked in terms of drawdown, these uncertainties can equate to 50%–100% of maximum simulated drawdown in some boreholes (PM-03, MU65S).

Uncertainty around the Lake Muir levels is relatively small, probably due to the fact the dominant fluxes in the lake are controlled by historical rainfall and evaporation time series (prescribed in the model) and only two parameters with corresponding multipliers, as relative contributions from groundwater into the lake inputs only account for approximately 30% (Figure 12).

Uncertainty of mass balance quantities displayed in the same figure provide some important insights. It can be noted that uncertainty over cumulative groundwater storage changes increase progressively through the entire simulated period, where uncertainty around rates such as net recharge and groundwater contributions to lake inflow remain relatively stable. Uncertainties around the period of 1970–1974 are slightly higher for net groundwater recharge, net balance for Lake Muir and relative contributions, as well as lake levels. This is the period where rainfall decline starts and it is possible that the sudden shift in rainfall rates produced larger stresses in the initial years and, consequently larger uncertainty.

5 Sensitivity analysis

We explored different sensitivity analysis techniques and their ability to contribute to the understanding of hydrologic processes

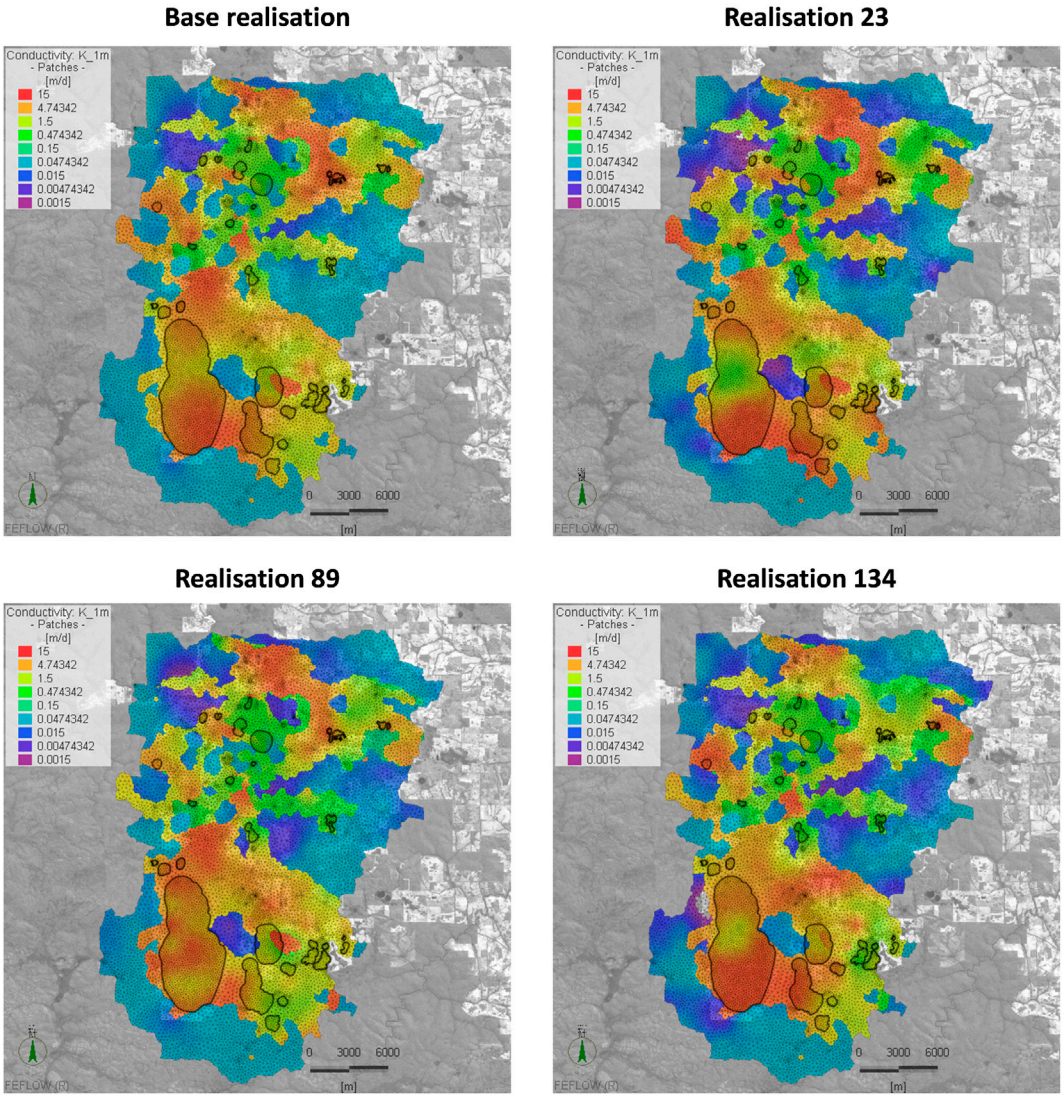

FIGURE 8
Calibrated horizontal conductivity values for layer 1 on selected posterior ensemble realizations.

occurring in the MUNDRC and terminal catchments in general. Local sensitivity analysis was employed to quantify point-source sensitivities (i.e., at a single location in parameter space) to understand spatial and temporal relationships, and global sensitivity methods were used to provide more robust sensitivity estimates and investigate broader controls on history matching and predictions of interest.

5.1 Improving understanding and conceptualization

A Jacobian matrix for the MUNDRC model has been generated for the parameter set with lowest residuals from the posterior ensemble obtained for the PESTPP-IES history matching work, using PEST-HP. The matrix was constructed considering all parameters (3,295) using a 3-point derivative approximation, in a total of 6,591 model runs. Analyzed inputs included observation and derived metrics used in the history-matching process, as well as virtual observations described in Section 2.2.

The comparison of groundwater level sensitivities to model parameters against the distance between monitoring point and pilot-point location is useful to establish distance-sensitivity relationships and estimate the "radius of influence" of certain parameters. Figure 13 shows plots of absolute sensitivity of groundwater levels in selected monitoring locations, where average sensitivity values for all groundwater levels in each location were calculated for all spatially distributed parameters (i.e., parameters from the pilot points). These plots show that in general, all parameters beyond 2–4 km from observation points show very low to no sensitivity, despite the maximum sensitivity of each parameter group (for example, recharge parameters show very high sensitivity for pilot points within 2–4 km, but the sensitivity is lost in parameters beyond that distance in the same way that storage parameters, which have much lower sensitivities).

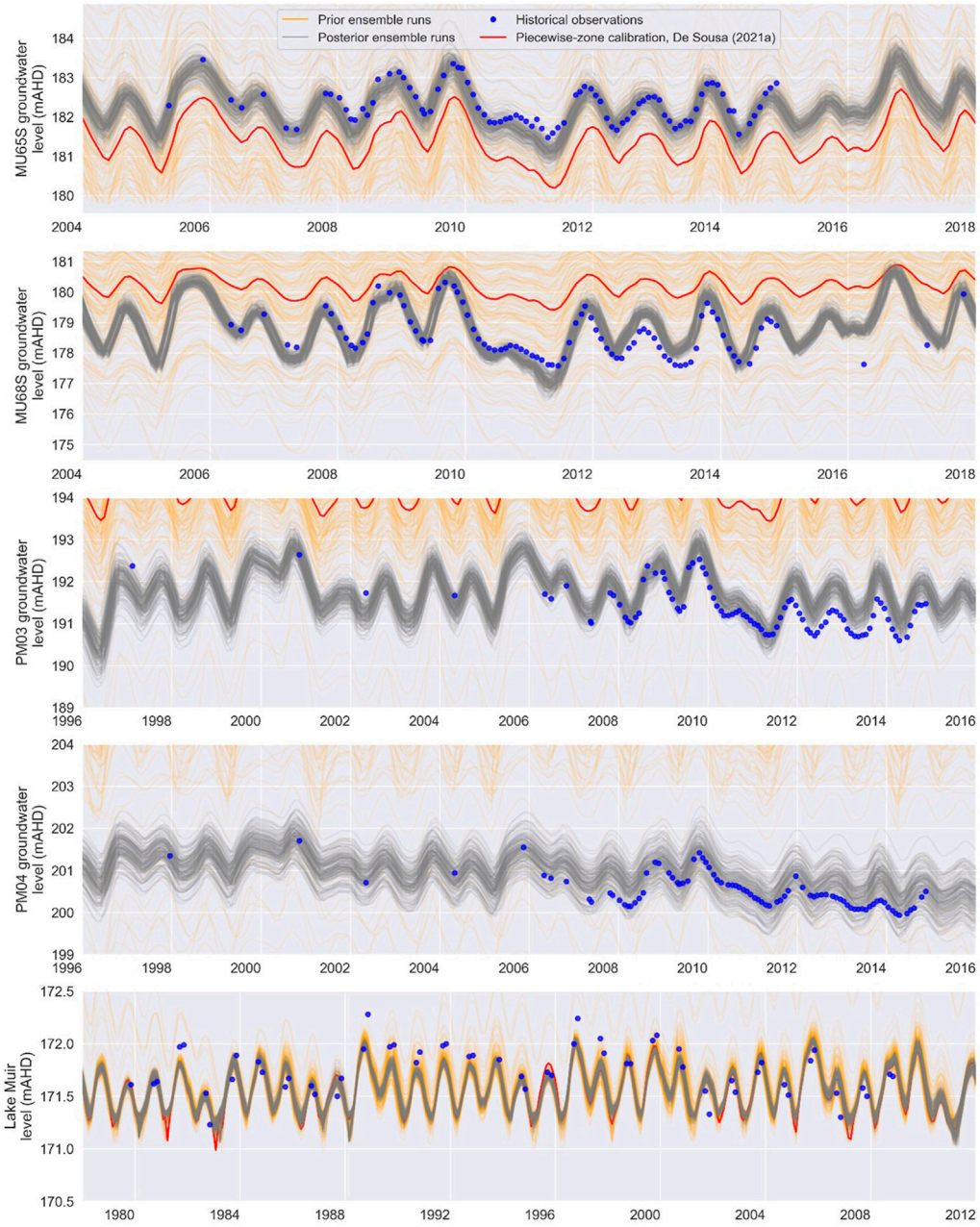

FIGURE 9
Simulated lake and groundwater levels prior and post-calibration at selected locations.

The distance-sensitivity plots also unravel relationships between maximum sensitivity on a parameter group basis, and their sensitivity noise, defined here as average groundwater level sensitivities to parameters located beyond the threshold distance. Ratios between maximum sensitivity and sensitivity noise are very high in parameter groups with high maximum sensitivity (such as Ev and Ra), but this ratio tends to degrade for parameter groups with low maximum sensitivity (for example, for groups Ss_3 and Sy_3 in monitoring borehole PM-03).

The use of virtual observations over the entire period between 1970 and 2010 showed that groundwater level sensitivity varies considerably over the simulated period. The assessment of sensitivities over time shows that sensitivity to recharge parameters increase over time (particularly in monitoring bores located away from lakes and other surface water compartments), suggesting that changes in net recharge rates have a cumulative effect on groundwater levels. In monitoring sites near lakes, this cumulative behavior is likely dampened by the model boundary conditions, as groundwater fluxes into these compartments adjust to the different recharge rates.

It can be observed that the sensitivities have a large influence from the rainfall signal, as sensitivity peaks from different parameters and location often align in periods of high or very low rainfall (such as years 2001 and 2006). This also demonstrates that sensitivity is influenced by the magnitude of hydrologic stress throughout the

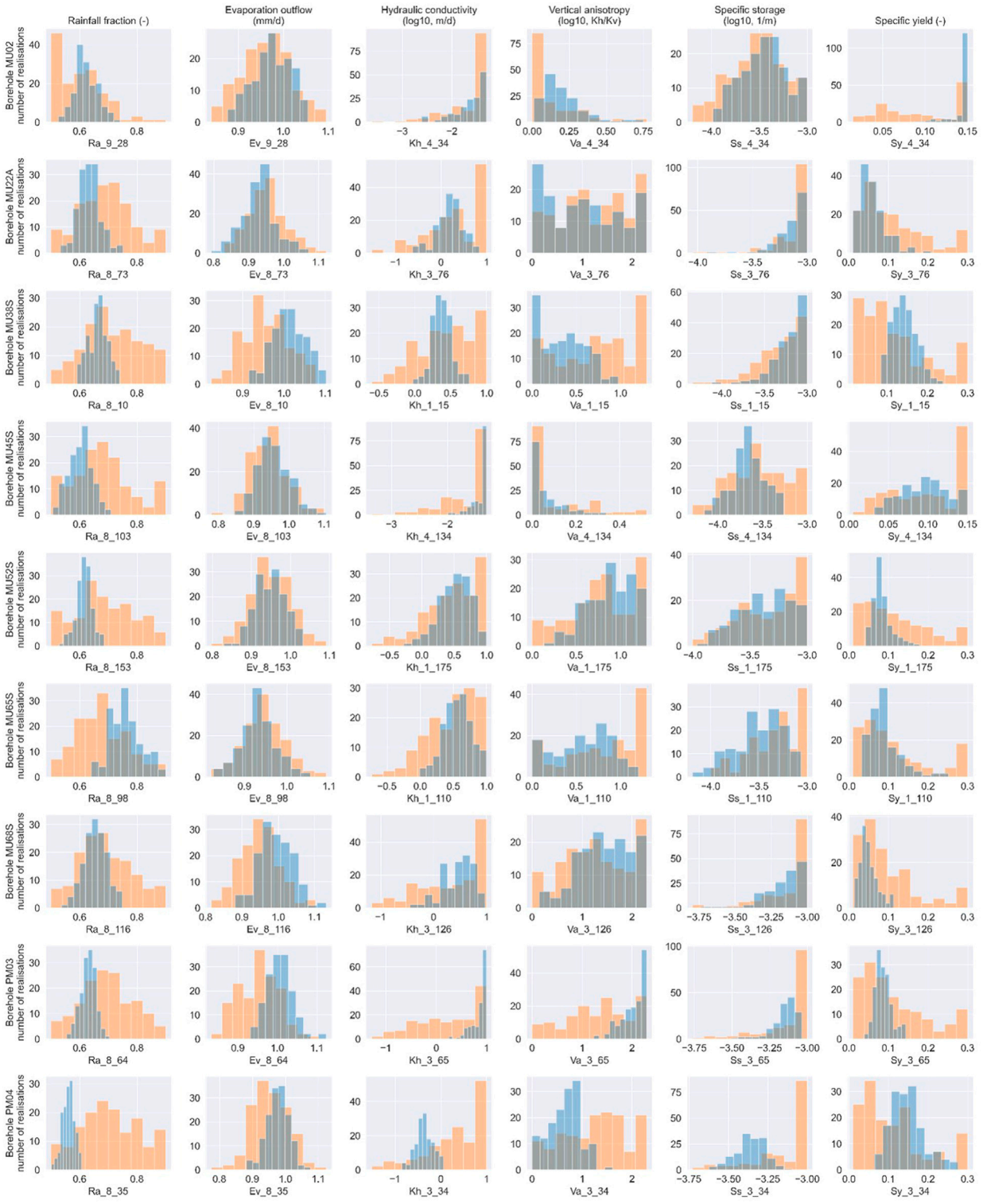

FIGURE 10
Parameter histograms of prior (orange) and posterior (blue) distributions of most sensitive parameters to observations in selected monitoring boreholes.

simulated period. Lastly the temporal behavior of storage parameter sensitivities is largely cyclical over the entire simulated period. To investigate whether these oscillations were associated with the seasonality observed in the catchment, groundwater level sensitivities to the parameters were grouped monthly and displayed as box and whiskers plots in Figure 14. These plots show that the groundwater level sensitivity is not only seasonal for storage parameters but, to a lesser degree, all other parameter types. High sensitivity peaks in these plots are normally in April-May, at the end of dry season, and low peaks are observed in September-October, at the

FIGURE 11
Simulated lake and groundwater level ensemble percentiles at selected locations.

end of wet season. This reflects not only the seasonality, but also suggest that the sensitivity is dependent on the hydrological state of the system, with sensitivities varying largely from periods of water surplus (wet season) and water deficit (dry season).

5.2 Assessing the value of history matching metrics and prioritizing site investigation efforts with linear uncertainty methods

Doherty and Hunt (2009) describe two statistics referred to as Identifiability and Relative Parameter Uncertainty Variance Reduction (RUVR). These statistics can be obtained from the Jacobian Matrix of a calibrated parameter set for any adjustable parameter and vary between 0 and 1, where the value of 0 means no reduction of uncertainty has been achieved through the history-matching process and the value of 1 indicates small parameter uncertainty in relation to the prior.

These analyses can be obtained considering history-matching observations that exist or not, and when applying different settings for observations (through weighting) and parameters (by fixing them or not) they can provide useful insights on the value of different observation groups, aggregate value of raw and derived metrics, and also inform site investigation efforts.

The several history-matching groups employed in the calibration of the MUNDRC model allowed a significant reduction in parameter uncertainties and also reasonable replication of past system behavior. However, the contribution of the different metrics to reduction in the uncertainty of different parameter groups was not clear. To investigate that, different linear analysis runs were done considering the entire

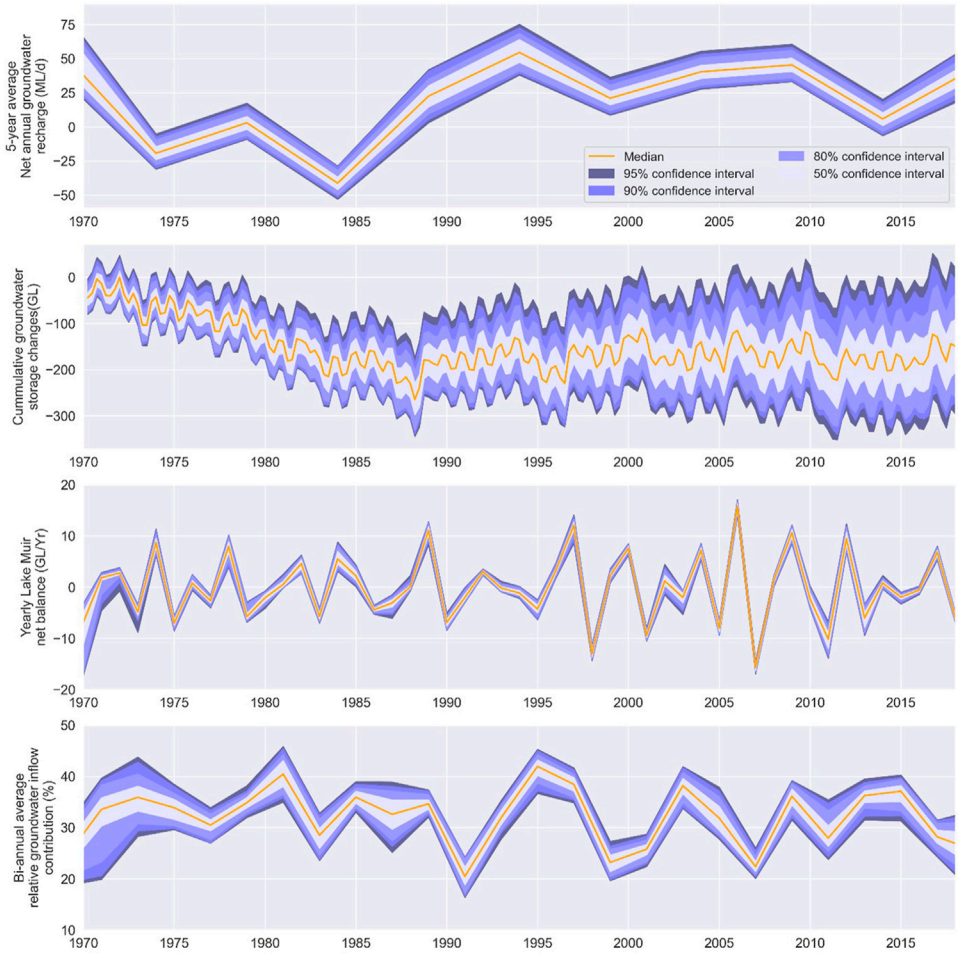

FIGURE 12
Simulated ensemble percentiles for different water balance predictive metrics.

history-matching data set and different observation groups individually as displayed in Figure 15, where values for identifiability and RUVR were averaged for the different parameter groups.

These results shows that 1) the sum of the values from individual observation groups is different from the values for the entire dataset, given the fact the these metrics are to some extent correlated in terms of sensitivity, 2) groundwater level observations shown overall the largest identifiability values as an individual group, 3) contributions from the derived metrics of horizontal and seasonal head differences to reduction of parameter uncertainty are most effective in the identification of storage parameters and 4) identifiability and RUVR values for lake values are predominantly controlled by lake observations, with subordinate contributions of groundwater levels. In the case of horizontal head differences, it also shows slightly higher values than the raw groundwater levels, showing that they have the same ability, if not more, of reducing parameter uncertainty.

In order to inform further investigation efforts in the area, additional linear analysis runs were undertaken by fixing the different parameter groups to understand what the reductions in the uncertainty of the remaining parameters would be if the fixed parameters were known. These results show clearly that largest benefits in terms of reducing parameter uncertainty would be from investigating recharge attributes (either infiltration rates, Ra, or evapotranspiration, Ev) as they would result in a minimum of 10% increase in identifiability values (with exception of lake parameters). The determination of unconfined storage (i.e., specific yield) would also be beneficial, as it controls the effective size of groundwater reservoir and the magnitude of head change associated with net recharge.

6 Discussion

6.1 The use of DA, UQ and SA techniques throughout the model development

While the DA, UQ and SA techniques are mostly used as accessories in conventional modelling practice, they have proven pivotal to the development of the numerical model and evolution of conceptual understanding of the MUNDRC. Where in earlier stages of model development the definition of aquifer geometry, boundary conditions and coupling with the lake model were reasonably straightforward, the initial iterations of history matching and conceptualization were hindered by the lack of baseline groundwater level data. At that stage, despite the clear reduction of

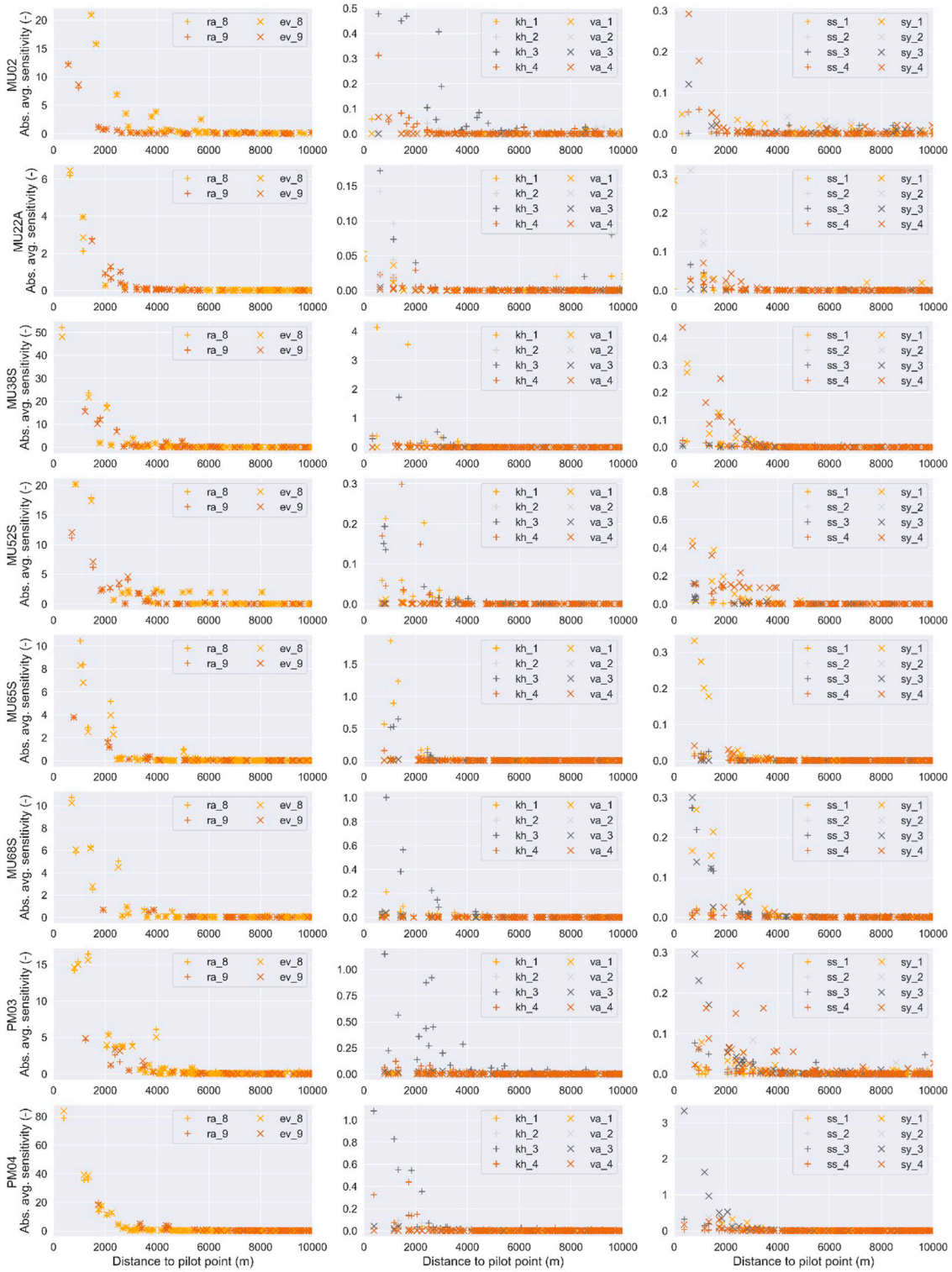

FIGURE 13
Scatter plots of absolute sensitivity values versus distance to pilot point for selected monitoring locations.

rainfall volumes, groundwater and lake level declines could not be clearly demonstrated.

The application of EMD on historical groundwater levels provided some evidence of longer-term drawdowns, by removing noise introduced by seasonal and higher-frequency variations, supporting the hypothesis that groundwater levels dropped as result of rainfall decline. While not employed in this study, the EMD's ability to decompose time series in different frequencies could enable new derived metrics for history matching by comparing simulated and observed IMF's,

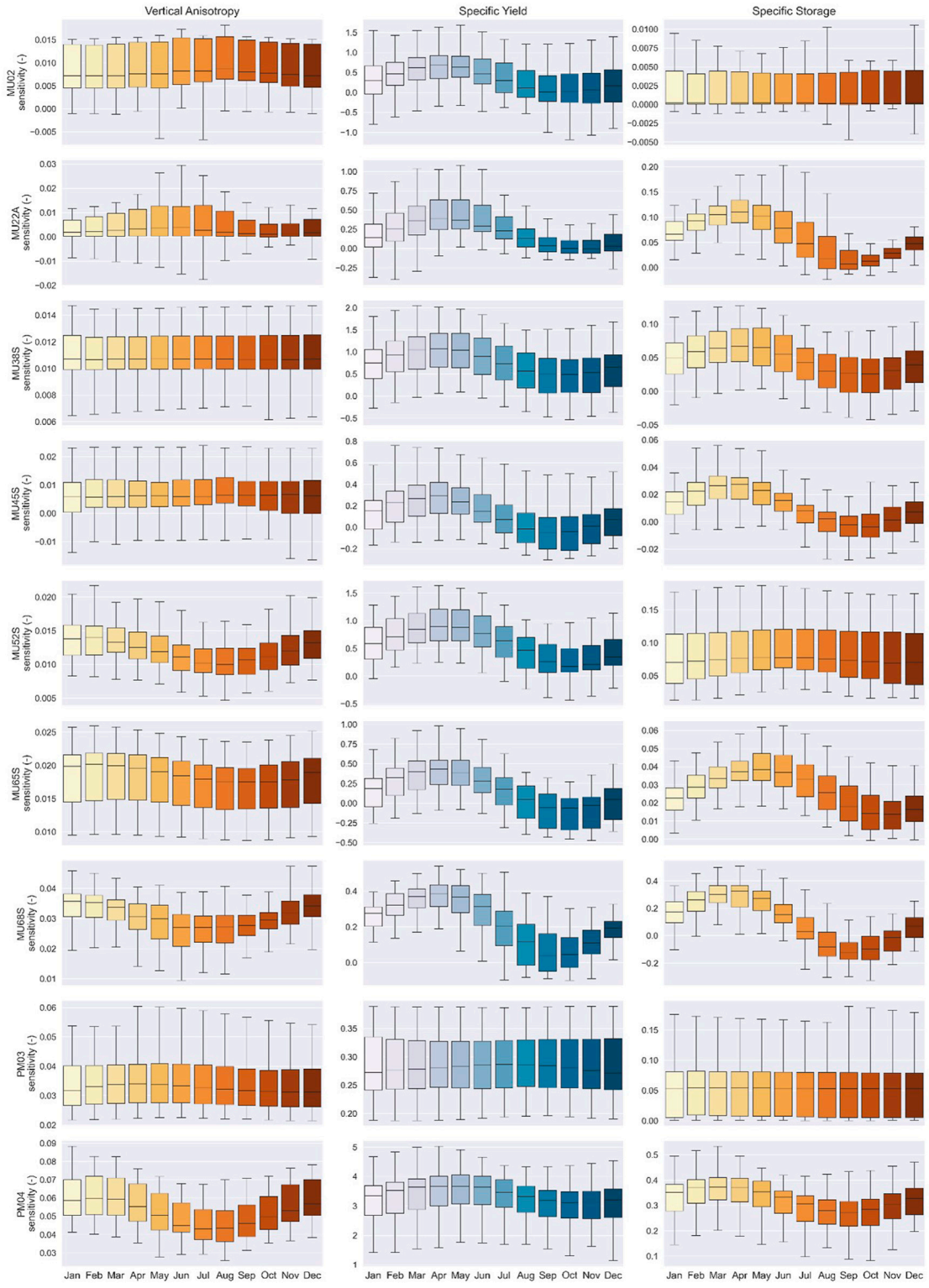

FIGURE 14
Monthly-grouped absolute sensitivity values for vertical anisotropy, specific yield and specific storage at selected locations.

similar to the "data transformation methods" described in Bennett et al. (2013) and the "Metrics describing multi-scale variability in model state" from Hipsey et al. (2020). These metrics potentially can contribute to history matching and robustness of models by highlighting aspects of model behavior that are not clear in the original time domain.

The high correlation between seasonality of groundwater levels and rainfall led to attempts of establishing relationships between

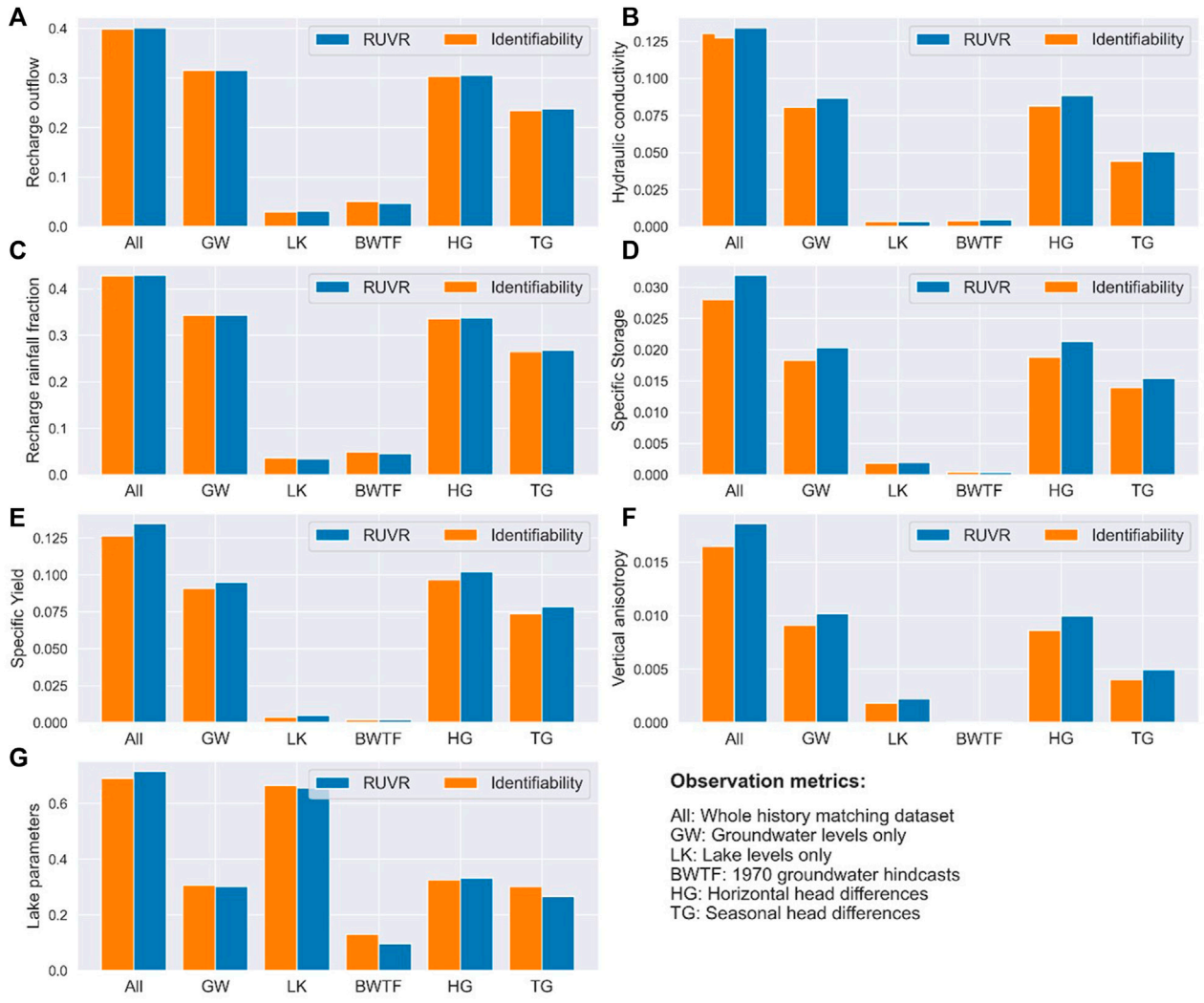

FIGURE 15
Average identifiability and RUVR values considering the entire history matching data set and individual metrics for **(A)** Recharge outflow term, **(B)** Horizontal hydraulic conductivity, **(C)** Recharge rainfall fraction, **(D)** Specific storage, **(E)** Specific yield, **(F)** Vertical anisotropy, and **(G)** Lake parameters.

rainfall rates, net groundwater recharge and groundwater levels. These attempts culminated in the development of the BWTF, which provided coarse estimates on net recharge but, most importantly, provided hindcast estimates of groundwater levels in 1970, prior to rainfall decline. The baseline groundwater level estimates from the BWTF analysis allowed for: 1—Reconstruction of groundwater levels prior to the rainfall decline (from a conceptual perspective), 2—Inclusion of these estimates in the history matching process, 3—Improvement of recharge implementation in the groundwater model by using a similar formulation and 4—Simulation of the whole trajectory from pre-rainfall decline towards present day.

Once the final model form was in place (1960–2018 simulated period with BWTF estimates in history matching, coupled lake model and improved recharge formulation), the history matching techniques assisted in reducing the uncertainty around aquifer parameters, which was particularly important since no aquifer test data was available. The piecewise-zone calibration presented in De Sousa (2021) provided reasonable average values for the entire catchment, and the highly-parameterized form presented in this paper allowed better representation of historical system behavior, representation of heterogeneity within each of the hydrogeological units and implemented the foundation for the UQ workflows.

The IES technique employed in history matching provided a quantitative assessment of parameter and predictive uncertainty constrained solely by conceptual expert-knowledge (i.e., prior) and allowed data assimilation in its more classic form, where prior model uncertainty has been reduced through the assimilation of site observations (i.e., history matching).

This paper has presented the final results of sensitivity analysis; nevertheless, several iterations have been undertaken throughout the development of the model, many of which contributed to the final form of conceptual and numerical model. The investigation of distance-sensitivity relationships provided insights on the area of influence of each parameter and this information can be helpful in the use of localization in iterative ensemble smoothers (Chen and Oliver, 2013). These relationships can also be used to prioritise site investigations, particularly if used in conjunction with linear uncertainty analysis.

Transient and seasonal trends of groundwater level sensitivity identified in the model demonstrate the value of using virtual observations over the entire simulated periods, even when corresponding field measurements are not available. Although these sensitivities cannot be used in the history matching process, they were useful to establish relationships between seasonality, aquifer net balance state (surplus, deficit or neutral) and sensitivity.

6.2 Distinct attributes of terminal catchments unveiled by these techniques

The results from the DA, UQ and SA techniques corroborated several attributes of the MUNDRC that are distinct of low-relief terminal catchments and unveiled new attributes that were expected by early conceptualization.

The sensitivity analysis of the Jacobian matrix showed the dominance of recharge parameters in terms of sensitivity and confirm the high influence of the interplay between rainfall infiltration and evapotranspiration in groundwater levels and catchment dynamics. Regarding groundwater levels, the conceptualization postulated that their response to rainfall events was rapid given the shallow groundwater table depths and relatively high hydraulic conductivity of sedimentary aquifers. The results from the BWTF and FEFLOW models agree with this hypothesis, as a good fit between simulated and groundwater levels was obtained for absolute values and seasonal oscillations without the use of a delay term in the recharge formulations. Furthermore, cluster analysis also demonstrated the relationships between land use and groundwater recharge, as well as its influence in associated groundwater level signatures.

The analysis of model results also led to some insights that were counter-intuitive and in disagreement with our early conceptualization. For instance, groundwater inflows into Lake Muir were expected to be higher during the dry season as the lake levels were at their lowest. Mass balance analysis of FEFLOW showed that while that is true in terms of relative contributions, the highest groundwater discharge rates occur during the wet season, where highest recharge rates replenish the aquifers increasing hydraulic gradients and consequently discharge rates. In another example, it was expected that groundwater discharge would occur predominantly through the base of the lake (assuming density effects on groundwater head distribution were negligible). Mass balance results at the lake nodes suggest that fluxes from lake to the aquifers occur through the base of the lake, groundwater discharge into Lake Muir occurs predominantly along the perimeter of the lake (De Sousa, 2021). Lastly, groundwater levels near surface water compartments are less sensitive to recharge and are to some extent regularized, in the sense that changes in recharge rates and groundwater level are compensated by adjusted flux rates between surface water bodies and adjacent aquifers.

Another new concept unveiled by the SA was the transient sensitivity of groundwater levels with regards to time. While this concept seems straight forward after the analysis of results, the concept of transient sensitivities has not, to the authors' knowledge, been demonstrated in literature.

The discussion presented in De Sousa (2021) suggests that Lake Muir is more resilient to rainfall decline the originally thought. The UQ works presented in this paper corroborate that and sensitivity analysis hinted at the underlying reasons. It was observed that sensitivity to rainfall multiplier (Lk_rain) was in general higher than sensitivity to the evaporation multiplier (Lk_evap), leading to the conclusion while evaporation rates are directly related to the lake area, rainfall rates are less susceptible to that as rainwater infiltrates the dry portions of the lake and ultimately is discharged there. The direct relation between lake area and evaporation volumes can be translated to lake level and evaporation volumes, therefore decline in lake levels caused by rainfall decline (both as direct rainfall and groundwater discharge) incur reduced evaporation rates, dampening the net lake losses. This is a mechanism that can be extrapolated to all lakes with shallow and flat bathymetry, but less likely to occur in lakes with steeper lake beds, as reductions in lake area (and evaporation) due to decline in lake level will be somewhat smaller.

6.3 Computational costs and limitations of the different techniques

The results presented in the previous sections demonstrate the value of DA, UQ and SA techniques in improving conceptual understanding and facilitating the quantification of impacts and catchment hydrologic processes. On the other hand, computation costs of each of these techniques may lead to a prioritization of efforts and cost/benefit assessment in a resource-constrained study. Furthermore, the use of these techniques needs to be undertaken cognizant of their limitations and computational costs, some of which are discussed here.

The use of the EMD method is computationally inexpensive and can be used in a batch fashion to process time series from multiple observation time series simultaneously, but the interpretation of the Intrinsic Mode Functions obtained from the EMD must be conducted with caution as they can be highly sensitive to the time series sampling frequency and are potentially problematic when time series have irregular observation intervals. This can be noticed on the analysis for borehole MU22A (Figure 7) which has a higher monitoring frequency and display an additional high frequency IMF when compared to the time series from boreholes MU45S and MU65S. It is possible that this issue could be minimized by resampling at regular intervals using interpolation methods and, despite not being tested, the application of EMD on the regular time series generated by the BWTF method may prove to be a better option than the raw monitoring data.

The IES method is an extremely powerful approach that allows history matching of highly-parameterized models with a very small number of runs, compared to number of parameters, enabling DA and UQ of large models that would previously be too expensive in terms of computational effort. While the IES method has shown good history matching results with small ensemble sizes (in particularly if localization is employed), questions remains whether these ensembles are sufficiently large to characterize the uncertainty, particularly in terms of probability distributions. A possible solution for that could potentially be to increase the ensemble size with parameter sets derived from the sampling of a posterior covariance matrix created based on the original ensemble parameter values. This procedure would continue to require a small number of runs for the history matching process and provide a larger ensemble size for UQ. Furthermore, the application of ensemble methods in groundwater modelling is relatively new and

more testing and use of this tool is required for definition of optimized settings such as ensemble size, localization matrices and so forth.

The DA and UQ applied in the MUNDRC model resulted in simulated groundwater and lake levels with relatively small uncertainty. It is important to emphasize, however, that the rainfall time series for the simulated period were actual values from historical records and "hard-wired" in the model lake and recharge components. Given that there is a strong relationship between groundwater level, sensitivity, and rainfall rates/events (as seen in the transient sensitivity plots), groundwater predictions in the future should account for uncertainty of rainfall time series inputs, and these can potentially promote larger predictive uncertainty.

7 Conclusion

The research presented in this paper illustrates the use of DA, UQ and SA in the study of terminal catchments, their value in the identification of particularities of hydrologic behavior in these settings and provide a blueprint for assessment of impacts associated to long-term rainfall decline in terminal catchments. On a conceptual level, main drivers of the groundwater and surface-groundwater interactions have been identified and corroborated by sensitivity analysis results. In terms of quantification and prediction, the developed numerical model coupling approaches and data assimilation tools used in the study provide a framework to estimate environmental impacts considering inherent hydrogeological and hydrological uncertainty, as well as the ability of monitoring data to constrain it. From a broader perspective, practicalities and lessons learned from the application of these techniques are lacking in literature, which is predominantly focused on theory and development of new techniques, and the paper also contributes to that regard.

Although several techniques have been explored in this study, it by no means exhaust the number of techniques available in the literature. Notable examples include the time-series analysis using transfer function noise modelling (Collenteur et al., 2019), evolutionary algorithms (Maier et al., 2014), time series clustering methods (Aghabozorgi et al., 2015) and ensemble machine learning techniques (Zounermat-Kermani et al., 2021).

This study has also shown that much can be gained through feedback loops between the application of these techniques (in particular SA) and conceptualization, as opposed to conventional use of UQ and SA at the end of model development. The conceptual model of MUNDRC have evolved substantially from its early inception through the multiple sensitivity analysis rounds until the final conceptual and numerical model form was achieved. Counter-intuitive findings from this process such as higher groundwater inflows to Lake Muir during the wet season, dominance of vertical dynamics of recharge and evapotranspiration over horizontal flows and potential surface runoff flows into the lake challenged the assumptions from initial conceptualization and resulted in a more robust final model form which conforms to expert-knowledge and was able to replicate historical system behavior.

The DA, UQ and SA techniques applied in the MUNDRC were undertaken with open-source software freely available on web which facilitated the model development significantly. Nevertheless, the implementation these techniques still remains an onerous task, particularly with regards to post-processing workflows. The adoption of these techniques in the broader modelling community will depend much on the development of tools to streamline these workflows and availability of educational resources, and initiatives such as the Groundwater Modelling Decision Support Initiative (GMDSI, https://gmdsi.org) are making a big impact in that direction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

Author contributions

The numerical framework and application of the different DA, UQ, and SA techniques to the models, as well as findings from the interpretation of model results and writing of the manuscript have been developed by ED, under supervision of the co-authors, who also provided feedback during the uncertainty and sensitivity analyses, contributed to the final format of the paper, editing and review of the manuscript, and also checked the scientific integrity of the research.

Funding

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Acknowledgments

The authors would like to acknowledge Jasmine Rutherford, Roger Hearn, and Margaret Smith (former Western Australia Department of Environment and Conservation) for providing the geology and groundwater monitoring data, as well as technical support and discussions about the MUNDRC hydrogeology.

Conflict of interest

Author ED was employed by the company INTERA Inc. Author RV was employed by the company Hydrogeoenviro Pty Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering – a decade review. *Inf. Syst.* 25, 16–38. doi:10.1016/j.is.2015.04.007

Asch, M., Bocquet, M., and Nodet, M. (2016). *Data assimilation: Methods, algorithms and applications*. Philadelphia, Pensylvania: Society for Industrial and Applied Mathematics, 311.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environ. Model. Softw.* 40, 1–20. doi:10.1016/j.envsoft.2012.09.011

Chen, Y., and Oliver, D. S. (2013). Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Comput. Geosci.* 17 (4), 689–703. doi:10.1007/s10596-013-9351-5

Collenteur, R. A., Bakker, M., Calje, R., Klop, S. A., and Schaars, F. (2019). Pastas: Open-Source software for the analysis of groundwater time series. *Groundwater* 57 (6), 877–885. doi:10.1111/gwat.12925

Dausman, A. M., Doherty, J., Langevin, C. D., and Sukop, M. C. (2010). Quantifying data worth toward reducing predictive uncertainty. *Ground Water* 48 (5), 729–740. doi:10.1111/j.1745-6584.2010.00679.x

De Sousa, E. R. (2021). Evaluation of long-term rainfall decline impacts on small-scale semi-arid endorheic basins and application to Lake Muir-Unicup Natural Diversity Recovery Catchment. PhD Thesis. Perth, Australia: School of Agriculture and Environment, University of Western Australia.

Diersch, H.-J. G. (2014). *FEFLOW finite element modeling of flow, mass and heat transport in porous and fractured media*. Berlin, Germany: Springer-Verlag, 996.

Doherty, J. (2015). *Calibration and Uncertainty Analysis for complex environmental models*. Brisbane: Watermark Numerical Computing, 236.

Doherty, J., Fienen, M., and Hunt, R. (2010). *Approaches to highly parameterized inversion: Pilot point theory, guidelines, and research directions*. Reston, Virginia: USGS Scientific Investigations. Technical report: Report 2010-5168.

Doherty, J., and Hunt, R. J. (2009). Two statistics for evaluating parameter identifiability and error reduction. *J. Hydrology* 366, 119–127. doi:10.1016/j.jhydrol.2008.12.018

Doherty, J. (2020). *PEST_HP - PEST for highly parallelized computing environments*. Brisbane, Australia: Watermark Numerical Computing.

Ferdowsian, R., Pannel, D. J., McCarron, C., Ryder, A., and Crossing, L. (2001). Explaining groundwater hydrographs: Separating atypical rainfall events from time trends. *Aust. J. Soil Res.* 39, 861. doi:10.1071/sr00037

Gallagher, M., and Doherty, J. (2020). *Water supply security for the township of biggenden: A GMDSI worked example report*. South Australia: National Centre for Groundwater Research and Training, Flinders University.

Gong, Y., Wang, Z., Xu, G., and Zhang, Z. (2018). A comparative study of groundwater level forecasting using data-driven models based on ensemble empirical mode decomposition. *Water* 10, 730. doi:10.3390/w10060730

Grelet, G., and Smith, M. G. (2009). *The Lake muir–unicup natural diversity Recovery catchment drilling Program: Completion report 2003–2006*. Perth, Western Australia: Department of Environment and Conservation, Bore Completion Report. (unpublished).

Hill, M. C., and Tiedeman, C. R. (2007). *Effective groundwater model calibration*. Hoboken NJ: Wiley.

Hipsey, M. R., Gal, G., Arhonditsis, G. B., Carey, C. C., Elliot, J. A., Frassl, M. A., et al. (2020). A system of metrics for the assessment and improvement of aquatic ecosystem models. *Environ. Model. Softw.* 128, 104697. doi:10.1016/j.envsoft.2020.104697

Hope, P., and Foster, I. (2005). *How our rainfall has changed – the south-west. Climate Note 5/05*. Perth, Western Australia: Indian Ocean Climate Initiative.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R.Soc. Lnd.* 454, 903–995. doi:10.1098/rspa.1998.0193

James, S. C., Doherty, J. E., and Eddebbarh, A. (2009). Practical post-calibration uncertainty analysis: Yucca Mountain, Nevada. *Ground Water* 47 (6), 851–869. doi:10.1111/j.1745-6584.2009.00626.x

Jolly, I. D., McEwan, K. L., and Holland, K. L. (2008). A review of groundwater-surface water interactions in arid/semi-arid wetlands and the consequences of salinity for wetland ecology. *Ecohydrology* 1, 43–58. doi:10.1002/eco.6

Lafare, A. E. A., Peach, D. W., and Hughes, A. G. (2016). Use of seasonal trend decomposition to understand groundwater behaviour in the Permo-Triassic Sandstone aquifer, Eden Valley, UK. *Hydrogeology J.* 24, 141–158. doi:10.1007/s10040-015-1309-3

Laszuk, D. (2017). *Python implementation of empirical Mode decomposition algorithm*. GitHub repository. Available at: https://github.com/laszukdawid/PyEMD. doi:10.5281/zenodo.5459184

Maheshwari, S., and Kumar, A. (2014). Empirical mode decomposition: Theory and applications. *Int. J. Electron. Electr. Eng.* 7 (8), 873–878.

Maier, H. R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L. S., Cunha, M. C., et al. (2014). Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. *Environ. Model. Softw.* 62, 271–299. doi:10.1016/j.envsoft.2014.09.013

Moore, C., and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Adv. Water Resour.* 29, 605–623. doi:10.1016/j.advwatres.2005.07.003

New, C. E. S., Smith, R. A., Hearn, R. W., and Wheeler, I. B. (2004). "Groundwater-lake interactions in the Lake muir-unicup Recovery catchment [online]," in Engineering Salinity Solutions: 1st National Salinity Engineering Conference 2004. Editors S. Dogramaci and A. Waterhouse (Barton, A.C.T: Engineers Australia), 460–465.

Nicols, C., and Doherty, J. (2020). *Exploring model defects using linear analysis: A GMDSI worked example report*. South Australia: National Centre for Groundwater Research and Training, Flinders University.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., et al. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environ. Model. Softw.* 79, 214–232. doi:10.1016/j.envsoft.2016.02.008

Reynolds, D. A., and Marimuthu, S. (2007). Deuterium composition and flow path analysis as additional calibration targets to calibrate groundwater flow simulation in a coastal wetlands system. *Hydrogeology J.* 15, 515–535. doi:10.1007/s10040-006-0113-5

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., GatelliSaisana, D. M., et al. (2008). *Global sensitivity analysis—the primer*. Chichester, West Sussex: John Wiley & Sons, 292.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice—a guide to assessing scientific models*. Chichester, West Sussex: John Wiley & Sons, 219.

Seeboonruang, U. (2014). An empirical decomposition of deep groundwater time series and possible link to climate variability. *Glob. NEST J.* 16 (1), 87–103.

Thompson, S. E., Sivapalan, M., Harman, C. J., Srinivasan, V., Hipsey, M. R., Reed, P., et al. (2015). Developing predictive insight into changing water systems: Use-inspired hydrologic science for the anthropocene. *Hydrology Earth Syst. Sci.* 17 (12), 5013–5039. doi:10.5194/hess-17-5013-2013

Wang, Z., Zhao, H., Sheng, Y., Geng, J., Wang, K., and Yang, H. (2020). Groundwater net discharge rates estimated from lake level change in Badain Jaran Desert, Northwest China. *Sci. China, Earth Sci.* 63 (5), 713–725. doi:10.1007/s11430-019-9533-8

Welter, D. E., White, J. T., Hunt, R. J., and Doherty, J. E. (2015). *Approaches in highly parameterized inversion— PEST++ Version 3, a Parameter ESTimation and uncertainty analysis software suite optimized for large environmental models*. Reston, Virginia: U.S. Geological Surbey, 54. U.S. Geological Survey Techniques and Methods, book 7, chap. C12.

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50 (2), 1152–1173. doi:10.1002/2013WR014767

White, J. T., Fienen, M. N., and Doherty, J. E. (2016). *pyEMU: a python framework for environmental model uncertainty analysis, version .01*. U.S. Geological Survey software release. doi:10.5066/F75D8Q01

Zounermat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *J. Hydrology* 598, 126266. doi:10.1016/j.jhydrol.2021.126266

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/feart.2023.1111065

OPEN ACCESS

EDITED BY
Shailesh Kumar Singh,
National Institute of Water and Atmospheric Research (NIWA), New Zealand

REVIEWED BY
Hans Jørgen Henriksen,
Geological Survey of Denmark and Greenland, Denmark
Jeremy Rohmer,
Bureau de Recherches Géologiques et Minières, France

*CORRESPONDENCE
Lee A. Chambers,
l.chambers@gns.cri.nz

SPECIALTY SECTION
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

RECEIVED 29 November 2022
ACCEPTED 15 February 2023
PUBLISHED 17 March 2023

CITATION
Chambers LA, Hemmings B, Cox SC, Moore C, Knowling MJ, Hayley K, Rekker J, Mourot FM, Glassey P and Levy R (2023), Quantifying uncertainty in the temporal disposition of groundwater inundation under sea level rise projections.
Front. Earth Sci. 11:1111065.
doi: 10.3389/feart.2023.1111065

Quantifying uncertainty in the temporal disposition of groundwater inundation under sea level rise projections

Lee A. Chambers[1]*, Brioch Hemmings[2], Simon C. Cox[1], Catherine Moore[1], Matthew J. Knowling[3], Kevin Hayley[4], Jens Rekker[5], Frédérique M. Mourot[2], Phil Glassey[1] and Richard Levy[1]

[1]GNS Science, Lower Hutt, New Zealand, [2]Wairakei Research Centre, GNS Science, Taupō, New Zealand, [3]School of Civil, Environmental and Mining Engineering, Faculty of Engineering, Computer and Mathematical Sciences, The University of Adelaide, Melbourne, VIC, Australia, [4]Groundwater Solutions Pty., Ltd., Melbourne, VIC, Australia, [5]Kōmanawa Solutions Ltd., Dunedin, Otago, New Zealand

Over the next century, coastal regions are under threat from projected rising sea levels and the potential emergence of groundwater at the land surface (groundwater inundation). The potential economic and social damages of this largely unseen, and often poorly characterised natural hazard are substantial. To support risk-based decision making in response to this emerging hazard, we present a Bayesian modelling framework (or workflow), which maps the spatial distribution of groundwater level uncertainty and inundation under Intergovernmental Panel on Climate Change (IPCC) projections of Sea Level Rise (SLR). Such probabilistic mapping assessments, which explicitly acknowledge the spatial uncertainty of groundwater flow model predictions, and the deep uncertainty of the IPCC-SLR projections themselves, remains challenging for coastal groundwater systems. Our study, therefore, presents a generalisable workflow to support decision makers, that we demonstrate for a case study of a low-lying coastal region in Aotearoa New Zealand. Our results provide posterior predictive distributions of groundwater levels to map susceptibility to the groundwater inundation hazard, according to exceedance of specified model top elevations. We also explore the value of history matching (model calibration) in the context of reducing predictive uncertainty, and the benefits of predicting changes (rather than absolute values) in relation to a decision threshold. The latter may have profound implications for the many at-risk coastal communities and ecosystems, which are typically data poor. We conclude that history matching can indeed increase the spatial confidence of posterior groundwater inundation predictions for the 2030-2050 timeframe.

KEYWORDS

sea level rise, groundwater inundation, MODFLOW, predictive uncertainty, iterative ensemble smoother, PEST++, data-assimilation

1 Introduction

Sea level observations (Jevrejeva et al., 2009; Vermeer and Rahmstorf, 2009) and projections (Kopp et al., 2014; Hall et al., 2016; IPCC, 2021) indicate alarming decade-to-century rises in global mean sea levels. Under high emissions scenarios, mean sea levels could exceed 1.0 m above 2000 levels by 2100 (IPCC, 2021). Globally, it now appears that we

are committed to 274 ± 68 mm of eustatic SLR, regardless of mitigation measures or climate change pathway (Box et al., 2022). Currently, mean sea levels are rising at rates of ~ 3–4 mm/year (Watson et al., 2015), and continued ocean warming, land-based ice melt (Yi et al., 2015; IPCC, 2021), and coastal subsidence (Nicholls et al., 2021) are expected to increase relative-SLR further.

SLR will have severe impacts on low-lying coastal regions. It is estimated that 267 million people live on coastal land <2 m above mean sea level (Hooijer and Verminnen, 2021). This number is projected to increase to ~1 billion by 2050 (Befus et al., 2020; Neumann et al., 2015). In these regions, SLR endangers coastal communities by increasing the frequency and severity of natural hazards, such as high-tide sea-water inundation (e.g., Cooper et al., 2013; Paulik et al., 2019), coastal erosion (e.g., Anderson et al., 2015) and surface water flooding (e.g., Sweet et al., 2014), whilst contributing to the permanent loss of land (e.g., Ramm et al., 2017; Ramm et al., 2018) and eventual displacement of communities (e.g., Nicholls et al., 2021).

Profound and often overlooked impacts of SLR include rising groundwater levels and the potential emergence of groundwater at the surface (that is, groundwater inundation). As sea levels rise, groundwater that is hydraulically connected to the sea will rise and eventually break out at the land surface. This could lead to groundwater inundation far inland, even before any sea-water inundation or surface water flooding occurs, potentially compounding such surface flooding (e.g., McCobb and Weiskel, 2003; Nicholls et al., 2007; Bjerklie et al., 2012; Goldsmith et al., 2015; Hoover et al., 2016; Befus et al., 2020).

These rising groundwater level and inundation projections represent additional and largely unseen natural hazards (e.g., Rotzoll and Fletcher, 2013) that are difficult to identify (e.g., McKenzie et al., 2010) and largely unrecognized by the general public (e.g., May, 2020). Typical flood defences may be prohibitively expensive or inappropriate (e.g., Yu et al., 2019), and may actually exacerbate rising groundwater levels and inundation (e.g., Cox et al., 2020).

Potential economic and social damages are substantial and include (but not limited to): road and property flooding (e.g., Abboud et al., 2018), reduced agricultural productivity (e.g., Barlow et al., 1996), reduced service life of roads and pavements (e.g., Knott et al., 2017), reduced capacity of waste and stormwater networks (e.g., Morris et al., 2018), wastewater treatment failure (e.g., Cox et al., 2020), and increased exposure of underground civil infrastructure (e.g., Macdonald et al., 2012), leading to foundation failures and corrosion (e.g., Colombo et al., 2018).

Given these potential impacts, groundwater inundation mapping will be an essential tool for supporting decisions on how to manage and communicate the impacts of SLR on coastal aquifer systems (e.g., Hoover et al., 2016; Merchán-Rivera et al., 2022). However, the subsurface is highly complex, and our ability to characterise this complexity is limited (e.g., Doherty and Moore, 2017). Furthermore, this hydrogeological uncertainty is confounded by the inherent "deep uncertainty" attached to the IPCC-SLR projections, themselves (e.g., IPCC, 2021). It is, therefore, impossible to reduce the uncertainty of SLR-related predictions to negligible levels.

However, through using numerical modelling techniques, it is possible to quantify spatial and temporal groundwater inundation susceptibility/risk, and to reduce this uncertainty to the extent that data allows. Such approaches should acknowledge the inherent spatial and temporal uncertainty of the simulated system (e.g., Merchán-Rivera et al., 2022), as well as the uncertainty of the aquifer stresses that may prevail in the future (e.g., SLR and/or climate variability). By characterising these system property and stresses probabilistically, we are then able to quantify the uncertainty of predictions in groundwater level rise and inundation. This is essential for facilitating risk-based decision-making (e.g., Freeze et al., 1990). Although some recent examples of groundwater inundation mapping exist (e.g., Hoover et al., 2016; Storlazzi et al., 2018; Habel et al., 2019; Befus et al., 2020; Merchán-Rivera et al., 2022), formal uncertainty quantification remains rare.

In this regard, Bayesian methods are considered some of the most rigorous approaches for decision-making under uncertainty (e.g., Caers, 2018). Industry standard tools for history matching (PEST and PEST++) can efficiently implement inversion-based algorithms for "highly-parametrised" models (e.g., 1000s of adjustable parameters) within a Bayesian framework (e.g., Doherty, 2015; White et al., 2020). This supports enhanced expression of uncertainty in system properties (e.g., heterogeneity), whilst providing greater potential for data assimilation from historical observations, and robust assessments of prediction uncertainty (e.g., Knowling et al., 2019).

This research adopts a Bayesian framework (or workflow), which is applied to estimate the spatial and temporal probability of groundwater inundation, under IPCC projections of relative-SLR. Specifically, the predictions of interest are a description of: 1) the transient progression of annual groundwater levels (heads) at specified times in the future as sea level changes, and 2) the total groundwater flux to the surface/wastewater drainage networks as sea level changes. Uncertainty accompanies all of these predictions, and this enables spatial mapping of the probability of groundwater inundation (groundwater emerging at the land surface).

This approach is novel in several ways. Firstly, a highly distributed parametrisation scheme allows the spatial detail and uncertainty of the predictions of interest to be estimated, and supports prediction uncertainty reduction, to the extent that the flow of information from available data allows. To our knowledge, the explicit application of temporal uncertainties in SLR projections, combined with spatially explicit uncertainty in groundwater flow model predictions, remains unexplored in the coastal groundwater modelling literature. Secondly, spatial and temporal estimates of drainage volumes provide an indication of what SLR mitigation measures may be required, for a range of SLR projections.

We demonstrate our approach for a real-world example to support the management of a low-lying coastal region (South Dunedin, Aotearoa New Zealand). Although local in scale, the framework is widely applicable and can be upscaled, or further developed for larger coastal regions where decision-support models are needed.

The paper is organised as follows. Section 2 introduces the case study problem, predictions of interest and the basis for the conceptual and numerical model. Section 3 describes the methodological detail required to implement our approach. Section 4 presents the results and discussion with conclusions following in Section 5. Reference is made to Supplementary

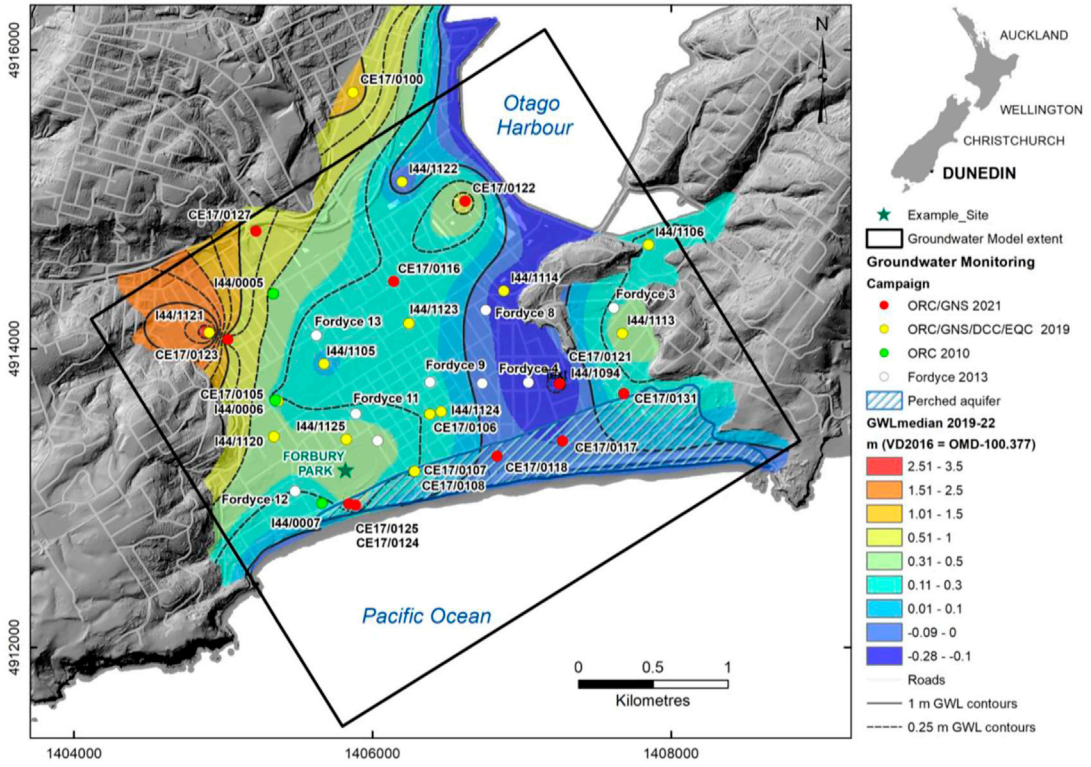

FIGURE 1
Groundwater monitoring sites located within groundwater model extent of South Dunedin. The piezometers used in this study (coloured by installation campaign) are shown with interpolated piezometric surface (updated from Cox et al., 2020). The star indicates the location of the Forbury Park "Example site," referenced in Section 4 of this paper. The blue shaded area is the interpreted extent of a perched aquifer in the sand dunes. The Otago Metric Datum (OMD) used in this study is equivalent to New Zealand Vertical Datum 2016 (NZVD2016) + 100.377 m.

Information (SI) throughout for further detail on the numerical model and our workflow.

2 Case study area

South Dunedin is approximately 6 km^2 and located behind sand dunes in the isthmus between the Dunedin hills in the west and the Otago Peninsula to the east (Figure 1). The coastal plain is typically <3 m above mean sea level and is now one of the most densely populated coastal urban centres in New Zealand, hosting many assets and critical infrastructure. Because of rising sea levels, the region is under threat from rising groundwater levels and inundation.

2.1 The groundwater emergence hazard

A shallow unconfined coastal groundwater system underlies South Dunedin. Groundwater levels are typically found <1 m below the ground surface and there is evidence of some hydraulic connection with the Pacific Ocean (Cox et al., 2020). The expected rise in groundwater levels resulting from SLR must be considered in future land-use and infrastructure planning in South Dunedin.

In the near term, SLR will compound interrelated hazards resulting from the complex interaction between shallow groundwater, buried civil infrastructure and surface waters (e.g., Bell et al., 2017). In the long term, SLR is expected to lead to the emergence of groundwater at the surface (groundwater inundation for the purpose of our research). Hence, central and local government, planners, engineers, and residents are amongst the many concerned by the extent of rising groundwater levels, and the inundation hazard (e.g., PCE, 2015).

2.2 Conceptual model

The latest geological and hydrogeological understanding of South Dunedin is described in detail by Glassey et al. (2003) and Cox et al. (2020) respectively. The current conceptual model of the groundwater system was based on these descriptions.

The topographically flat area represents a valley infilled with Quaternary sediments. The groundwater system flows within two sediment depositional units: 1) a younger Holocene unit comprising soft silt and clay of marine to estuarine origin, locally deposited during the post-glacial marine transgressions resulting from Holocene sealevel rise, overlying 2) a Pleistocene depositional unit comprising sands, silts, and some gravels, interpreted as alluvial deposits with hillslope deposits at the valley margins

(colluvium). These highly heterogeneous Pleistocene and Holocene sediments have a maximum depth of approximately 60 m. The contact between the Quaternary sediments and the underlying bedrock is relatively flat beneath most of South Dunedin, but has some (<40 m relief) paleo-topography (Glassey et al., 2003). Bedrock comprises either weak marine sedimentary rock (Caversham sandstone), or a variety of local interbedded igneous rocks (Dunedin Volcanic Group).

The groundwater system was treated conceptually as a single groundwater system for the purposes of this study, being bounded by basement rocks of the Dunedin Hills and Otago Peninsula, and the Pacific Ocean and Otago Harbour (Figure 1). The bedrock contact was treated as a no-flow boundary because recent investigations indicated negligible vertical hydraulic gradients (Cox et al., 2020), and limited vertical groundwater flow at the basin scale (Rekker, 2021).

In contrast to the underlying shallow unconfined groundwater system, a minor perched dune aquifer system to the south (Figure 1) demonstrates low electrical conductivity (i.e., relatively fresh composition) and the absence of a tidal signal (Cox et al., 2020). Unlike many other coastal areas in eastern New Zealand (e.g., Christchurch), there is no evidence to date which suggests any compartmentalisation by distinct inter-glacial aquitards, and the groundwater system lacks any deep groundwater at artesian pressures (e.g., Cox et al., 2021). Our conceptual model therefore assumes minimal "cross-boundary" interaction with other aquifers (e.g., the minor perched dune system to the south) and limited surface inflows from the surrounding catchments.

A streamline no-flow boundary was added along the northern boundary, separating the South Dunedin groundwater system from that of Harbourside, along a catchment and stormwater runoff boundary across the coastal plain (Figure 1). This assumption was justified because groundwater appears to flow approximately parallel to the boundary within coastal sediments.

2.2.1 Groundwater mass flow balance

As is typical of urban centres, surface hydrology is heavily modified and groundwater recharge is highly variable. The impervious land surface within the region causes approximately 60% of precipitation to be captured and routed *via* the stormwater network, mainly discharging to the Otago Harbour *via* the pumping station (Goldsmith and Hornblow, 2016). The remainder is available for groundwater recharge *via* pervious surfaces.

Potential groundwater recharge is relatively well constrained. A weather station within the modelled domain indicates an annual-average precipitation rate of 674 mm/yr between 1997 and 2021. The available stormwater pumping and precipitation data, combined with the imperviousness index for South Dunedin, indirectly leads to a recharge estimate of ~4,000 m^3/day.

Some of this groundwater then exits the system *via* infiltration to the ageing waste and stormwater networks, which is estimated at 2,000 m^3/day (Opus and URS, 2011a; Opus and URS, 2011b; Rekker, 2012; Fordyce, 2013; Cox et al., 2020). The spatial and temporal distribution of groundwater infiltration to the networks remains highly uncertain (Cox et al., 2020). The remainder leaves the system as submarine groundwater discharge. Offshore groundwater discharge and the geology which controls it, remains largely unknown. However, it is constrained by the difference in these mass balance recharge and drainage network estimates.

2.3 Groundwater level monitoring

There is a recent and extensive record of piezometric levels across South Dunedin (Figure 1). Automated meters currently record groundwater levels in 28 piezometers every 15 min within the modelled domain. These were installed over various field campaigns from 2010 to 2021 (see Cox et al., 2020 for a detailed description of the groundwater monitoring network and data coverage).

The interpolated median groundwater piezometric contours suggest that groundwater flows to the Pacific Ocean and Otago Harbour, as shown in Figure 1 (updated from Cox et al., 2020). Median groundwater levels are on average above mean sea-level, with the highest levels occurring in the north-western corner of the system.

Fluctuations in groundwater levels are nearly all restricted to <1 m in range, and dominated by short term variability linked to frontal rain systems, with some cyclicity at a 90–100 day period that reflects cumulative rainfall and recharge caused by the frequency of cyclonic storms (Cox et al., 2020). Any seasonal (e.g., summer vs. winter, or autumn vs. spring) cyclicity, or interannual variability over the decadal period of monitoring to date has been limited, making it relatively robust to use average levels for the steady-state approximation used for history matching (see Section 3).

The tidal range at the harbour/coast is approximately 1.7 m (see Supplementary Figure S1-1). Tidal fluctuations are recorded at some monitoring locations. For example, the groundwater level time series for piezometer I44/0007 (location shown in Figure 1) demonstrates a characteristic diminished amplitude and delayed arrival of the tidal signal (see Supplementary Figure S1-1). The groundwater time series at I44/0007 demonstrates a tidal range of approximately 0.3 m (a difference of 1.4 m at a distance of 120 m from the coast) with a lag in the peak of the tidal cycle of 131 min. This site is one of a few with a relatively strong tidal signal (Cox et al., 2020), but elsewhere hydraulic connection with the Pacific Ocean is still evident >1 km from the shore (see hydrographs for I44/0007 and CE17/0105 in Supplementary Figure S1-1, these piezometer locations are shown in Figure 1). Groundwater electrical conductivity and geochemistry suggest most of the groundwater is fresh and there is limited saline intrusion (<10% at 1 km from the shoreline, see Cox et al., 2020; Rekker, 2021).

2.4 Groundwater model

2.4.1 Model structure

The original numerical groundwater flow model was constructed by Rekker (2012) and modified for the purpose of this research (as described below). MODFLOW-NWT (Niswonger et al., 2011) was used to simulate constant-density groundwater flow under both steady-state and transient conditions. The finite-difference grid is a single-layer (representing Holocene and Pleistocene sediments) comprising

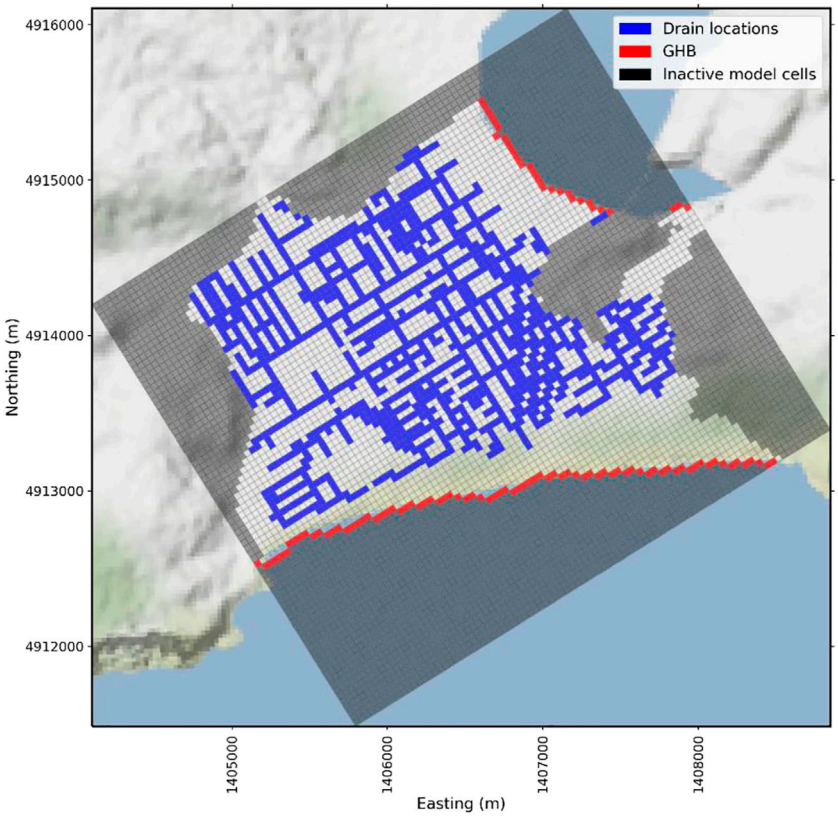

FIGURE 2
South Dunedin model extent showing model grid, boundary conditions and inactive model cells.

90 rows and 80 columns of uniform 40 × 40 m horizontal discretization. The boundary conditions and recharge array for the model are depicted in Figure 2.

The distribution of hydraulic properties was informed by Glassey et al. (2003). The model surface elevations were based on a digital elevation model informed by LiDAR data (1 m digital surface model pixels at specified vertical accuracy <0.2 m, 95% confidence) for South Dunedin (LINZ, 2021). We resampled the LiDAR data to obtain a regridded 40 × 40 m average for the model top elevation of the MODFLOW model domain. The original model bottom elevations estimated by Rekker (2012) from geophysical data, were maintained.

Recharge to the saturated zone is simulated using the MODFLOW recharge (RCH) package. The Otago Harbour and Pacific Ocean were simulated *via* the General-Head Boundary (GHB) package, and groundwater interaction with the stormwater and wastewater networks is simulated *via* the MODFLOW Drain (DRN) package (both head-dependent flux packages). The model bottom and other lateral boundaries are "no-flow" boundaries.

Hence, groundwater leaves the model domain as storm/wastewater flow (DRN package), or as submarine groundwater discharge (GHB package). The locations and invert elevations of the stormwater and wastewater networks was informed by city council GIS records. The stormwater network overlies the wastewater network. The surveyed sump elevations for the stormwater network were used in preference over the wastewater network to generate a network of drain locations and elevations within the model domain. That is, the storm and wastewater networks are not separated in our modelling approach. This representation of the storm and wastewater networks is adopted to account for the uncertainty in the conductance and elevation of both drainage networks in our modelling approach (see Section 3).

2.4.2 Temporal discretization

Simulations are divided into a steady-state "history" matching period, with stresses represented by long term average conditions for the period 2010–2020, and a transient "projection" period which simulates system response to IPCC-SLR projections for the period January 2010–January 2110. The density-corrected GHB stage for the history period is specified according to time-averaged tidal data for Port Otago obtained from the New Zealand Hydrographic Authority (Land Information New Zealand, LINZ).

Initial conditions for the transient projection period are established by the steady-state history matching period. The 100-year projection period that follows is discretised into annual stress-periods, and both simulation periods use the same time-invariant (static) properties of hydraulic conductivity, storage, GHB conductance, DRN conductance and DRN elevation. Time-variant properties of recharge and GHB stage are expressed for the projection period. Our approach then focuses on predicting groundwater levels and drain flows under changing GHB (rising sea

TABLE 1 Parameters and their distribution bounds. "Initial model value" refers to the native model parameter value (units also provided) to which the multiplier (or additive) parameter is applied.

	Unit	Parameterisation Method	Count	Style	Transform	Initial model Value	Lower bound	Upper bound
	Steady-state "history" matching period							
Horizontal K	m/day	Grid-based	3,610	mult	log	[a]Zonal	0.01	100
Horizontal K	m/day	Global	1	mult	log	[a]Zonal	0.01	100
Recharge	m/day	Grid-based	3,610	mult	log	1.47×10^{-3}	0.5	2
Recharge	m/day	Global	1	mult	log	1.47×10^{-3}	0.5	2
GHB cond. (South coast)	m^2/day	Grid-based	110	mult	log	800	0.01	100
GHB cond. (Harbourside)	m^2/day	Grid-based		mult	log	800	0.001	1,000
GHB cond	m^2/day	Global	1	mult	log	800	0.01	100
Drain elevation	m	Grid-based	1,259	add	none	[b]100.18	−0.5	0.5
Drain elevation	m	Global	1	add	none	[b]100.18	−0.5	0.5
Drain conductance	m^2/day	Grid-based	1,259	mult	log	2.6	0.1	10
Drain conductance	m^2/day	Global	1	mult	log	2.6	0.1	10
	Transient "projection" period							
Specific yield	—	Grid-based	3,610	mult	log	1.46×10^{-1}	0.5	2
Specific yield	—	Global	1	mult	log	1.46 × 10-[a]	0.5	2
Specific storage	—	Grid-based	3,610	mult	log	1×10^{-3}	0.01	100
Specific storage	—	Global	1	mult	log	1×10^{-3}	0.01	100
GHB stage	m	Global	1	mult	log	[c]Scenario	0.41	2.47
Temporal GHB	m	Global	100	mult	log	[c]Scenario	0.925	1.075
Temporal Recharge	m/day	Global	100	mult	log	1.47×10^{-3}	0.8	1.25

[a]Hydraulic conductivity is separated into four zones according to the identified lithology (Glassey et al., 2003).
[b]Drain invert elevations vary within the model domain. The estimated average invert elevation for the entire network is presented.
[c]Temporal GHB stage is dependent on the IPCC-SSP scenario.

levels) and recharge (climate variability) model boundary conditions, defined for the projection period.

3 Methodology

This section describes the methodological detail required to implement our approach, including the prediction specification, the development of the parameterisation scheme, history matching and uncertainty quantification. The scripted workflow is provided as a Jupyter Notebook to ensure transparency and reproducibility of the decision-support modelling described herein (Kluyver et al., 2016).

The workflow involves four main components: 1) early uncertainty quantification to assess prior parameter uncertainty and corresponding prediction uncertainty, to identify and resolve inadequacies in the conceptual model or numerical implementation, 2) history matching to condition model parameters that are pertinent to the predictions of interest, 3) Monte Carlo sampling of climate change and SLR parameters in the projection period to explore history matching informed predictive distributions of groundwater levels and, 4) the production of maps assessing the susceptibility to groundwater inundation, and quantification of drain flows under different SLR scenarios. We now describe our approach in detail.

3.1 Model parameterization

Model parameters were defined for both the history and the projection periods. During history matching the following parameters were adjusted: horizontal hydraulic conductivity, history period recharge, GHB conductance, drain conductance, and drain elevation. The additional parameters defined for the projection period comprised: specific yield, specific storage, temporal GHB stage, and temporal recharge (Table 1). Parameters added to the projection period remained unconditioned.

3.1.1 History matching parameters

The distribution of groundwater model hydraulic parameters, flux and head boundary conditions and recharge stresses are expressed through 9855 adjustable parameters for the steady-state history matching period (Table 1). Parameters are generally implemented as multi-scale multipliers which act upon initial model parameter values. Drain elevation parameters are represented as additive, rather than multiplier, parameters. For these, the parameters are applied as an addition or subtraction to the model drain invert elevation estimate.

Parameter operating scales reflect the expected scales of heterogeneity and uncertainty of model input values and are applied at the scale of geological model (global-scale) and the model cell (grid-scale) (e.g., White et al., 2020; Hemmings et al., 2020; McKenna et al., 2020). Initial parameter values, and the mean of their prior distributions, are one and zero, for multiplier and addition parameters, respectively (Table 1).

The prior parameter covariance matrix, from which the prior parameter realisations are drawn, is defined as a block-diagonal matrix. Diagonal elements of the prior parameter covariance matrix represent individual parameter variances, informed by prior, or "expert" knowledge of these model inputs (Table 1). Off-diagonal elements of the covariance matrix, were defined by geostructures built on exponential variograms with sills proportional to the prior parameter variances.

Upper and lower parameter bounds represent a six standard deviation envelope (±3 σ) around the mean of the distributions, equating to approximately a 99% confidence interval. An exponential variogram range of 1,200 m (range a = 400 m) was defined for spatially distributed parameters. However, to account for the anticipated high spatial variance in the (wastewater and stormwater) drainage infrastructure, the exponential variogram range for drain parameters (DRN package invert elevation and conductance) were reduced to 300 m (a = 100 m). Additionally, conservative prior uncertainties were assigned to abstract parameters representing boundary conditions of the structurally simple model (i.e., DRN and GHB conductance). This strategy was employed for uninformed prior uncertainties to avoid under-estimation of predictive uncertainty (e.g., Hugman and Doherty, 2022).

3.1.2 Projection period parameters

An additional 7,423 adjustable parameters were defined for the transient projection period (i.e., 17,276 parameters in total) to represent IPCC projection uncertainty (Table 1). IPCC projections for South Dunedin indicate minimal changes to annual average rainfall rates (e.g., Mourot et al., 2022). However, to represent interannual recharge variability and its uncertainty over the projection period, additional, independent (i.e., no temporal covariance) annual recharge multipliers were included in the analysis.

The initial model input recharge parameter values were estimated from long-term, annual average conditions for the steady-state history period (i.e., a 10-year timeframe). Upper and lower bounds for the temporal recharge multiplier (projection period) were informed by the variance of the 10-year moving average of historic annual rainfall rates. This was based on local long-term New Zealand MetService data for the period 1960–2021

TABLE 2 Relative-SLR projections for South Dunedin (https://searise.takiwa.co/) showing median (p50), 17th percentile (p17) and 83rd percentile projections for the SSP-8.5 (medium confidence) scenario. The realized ensemble of SLR projections drawn from this scenario are shown in Figure 3.

Scenario	Year	p17 (m)	p50 (m)	p83 (m)
SSP5-8.5 (medium confidence)	2030	0.08	0.11	0.14
SSP5-8.5 (medium confidence)	2050	0.20	0.25	0.32
SSP5-8.5 (medium confidence)	2070	0.34	0.43	0.56
SSP5-8.5 (medium confidence)	2100	0.64	0.81	1.06

(Table 1). As a consequence, the model is focussed towards predicting the transient progression of long-term annual conditions of groundwater levels, but not short-term (events-based) fluctuations that may be important for managing individual rain-event flood risk.

In contrast to groundwater recharge, projected rises in sea levels are significant, but also highly uncertain during the 21st century and beyond. The modelling workflow uses improved, location specific SLR projections provided by the NZ SeaRise: Te Tai Pari O Aotearoa Endeavour programme. These projections, which can be accessed through https://searise.takiwa.co/, include the effects of vertical land movement for every 2 km of the coast of Aotearoa New Zealand to the year 2,300. Here, to follow coastal planning recommendations specific to New Zealand (MfE, 2017), we focus on SLR projections associated with the IPCC Shared Socioeconomic Pathway (SSP) medium confidence, high emmissions scenario SSP5-8.5. However, the workflow is rapid and easily adaptable to explore any of the SLR scenarios, so we present an additional scenario in the Supplementary Information.

SLR projection uncertainty was propagated through the groundwater model to the predictions of interest according to the defined uncertainty interval for the IPCC-SSP scenario (SSP5-8.5 medium confidence; inferred from Table 2, where p17-p83 is assumed to encompass 2 σ). This SLR scenario uncertainty is represented through the variance on a global (spatially and temporally constant) multiplier, which acts on the median SLR projection timeseries (implemented through the GHB stage) applied across all stress periods. For SSP5-8.5 (Table 2), the variance of this global multiplier, with a mean of 1.0, was defined as 0.12 (standard deviation of 0.34). Also note, the potential range of the forcing applied to the GHB stage increases into the future as the uncertainty of the SLR scenario increases (i.e., heteroscedasticity). The resulting sampled projection period realisations of SLR for the SSP5-8.5 (medium confidence) scenario are illustrated in Figure 3.

Inter-annual variability and uncertainty for each individual SLR realization is defined through annual multipliers sampled within a ±3 σ range of 0.925–1.075, and covariance defined through a temporal exponential variogram with a range of 15 years (range a = 5 years). This choice was informed by a variogram analysis of the detrended annual average sea level recorded at the Green Island tide gauge (Bell et al., 2022).

Appending SLR parameters to the model parameter covariances supports drawing realisations for the projection period, thus allowing the ensemble of realisations to characterise the embedded deep uncertainty of future SLR projections (e.g., Kopp

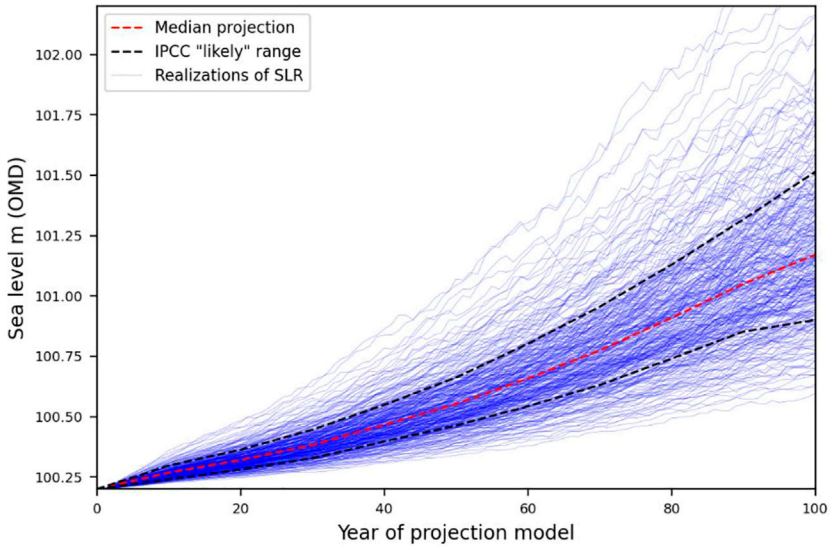

FIGURE 3
Realizations of sea level rise attached to the GHB stage of the 100-year projection model (2010–2110). The plot shows 300 realizations of sea level rise for the IPCC SSP5-8.5 (medium confidence) scenario.

et al., 2019), and their impact on the decision-critical prediction. To the best of the authors' knowledge, the explicit application of IPCC-SSP SLR scenario uncertainty to probabilistic groundwater flow model predictions, remains unexplored.

3.2 History matching, uncertainty quantification, and predictions

A prior-based Monte-Carlo uncertainty analysis was used to assess the credibility of the model structure and the prior parameter probability distributions, *via* observations of prior-data conflict (e.g., Egidi et al., 2022). History matching was then used to derive the posterior parameter ensemble, based on observations from the "history" period, using the iterative Ensemble Smoother (iES) in PEST++ (White, 2018). We then analysed the extent to which history matching (to the available data) was able to refine the distributions of parameter values, their combinations, and the corresponding predictions of interest.

Predictive scenarios, which include additional SLR and recharge uncertainty in the 2010–2110 projection period, were then simulated. This was achieved by combining the posterior parameter ensemble (for the history period) with additional unconditioned parameters relating only to the 100-year projection period (i.e., temporal GHB stage and recharge parameters, see Table 1). The resulting history matching informed parameter ensemble represents the conditioned uncertainty of groundwater levels in response to IPCC-SSP scenarios. These spatially distributed groundwater level predictions can then be used to map the potential SLR-driven groundwater inundation hazard in South Dunedin, supporting risk-based decision making.

TABLE 3 Measurement error (standard deviation, σ_m).

Observation group		Count	σ_m
Less-than inequality constraints		3,525	0.4
Waste/stormwater exchange flux		1	500
Groundwater levels	Long-term measurements (>2 years)	16	0.15
	Short-term measurements (<2 years)	12	0.25

3.2.1 Geostatistical draws, observations, and weights

An ensemble of 300 parameter realisations, providing a representation of prior parameter uncertainty, were drawn by Monte-Carlo multi-variate Gaussian sampling of the prior parameter covariance matrix, and then conditioned through history matching. These 300 parameter realizations were ultimately propagated through to the SLR scenario projection period.

The choice of the number of realisations (to propagate through the analysis) is a trade-off between minimising computational burden of the history matching process, whilst endeavouring to sufficiently capture prediction uncertainty, and accommodate the dimensionality of the solution space (e.g., Knowling et al., 2019; White et al., 2020; Hunt et al., 2021). To ensure that 300 realisations appropriately captured the prediction uncertainty, we preformed a convergence analysis, focussing on four prediciton locations of interest. The results of this convergence analysis are shown in Supplementary Figure S3-5. The converenge analysis indicates that 300 realisations effectively captures the prior prediction

distribution behaviour (ensemble mean, standard deviation and 95th percentile) represented by 1,000 realisations.

The history matching dataset comprised of long-term averages of system observations. Groundwater level observations were separated into two groups relating to the duration of the piezometer dataset (Table 3). An additional estimate of the annual average total groundwater-waste/stormwater exchange flux of 2,000 m^3/day was included as a target observation for history matching (Opus and URS, 2011a; Opus and URS, 2011b; Rekker, 2012; Fordyce, 2014).

Given the spatial sparsity of groundwater level measurements, it was beneficial to include observations which represent physically "realistic" constraints on simulated groundwater levels for the history matching period. This was implemented through the "less-than" inequality constraint (White, 2018). Less-than inequality observations contribute to the objective function only when the simulated value exceeds the observation value.

For our purposes, less-than observations were defined for simulated heads in every model cell. The observation value was set according to the model top elevation in the corresponding cell. This effectively implements a history matching constraint, which enforces the condition that long-term average groundwater levels should fall below the model top elevation (e.g., White, 2018; White et al., 2019; Fienen et al., 2021).

Initial observation weights were defined to reflect the estimated observation error. Weights were then re-adjusted to direct parameter upgrades towards objective function components that were considered most relevant to the decision-support objective (e.g., Doherty and Welter, 2010; Fienen et al., 2022). In particular, because groundwater level observations used for history matching are well "aligned" with the decision-critical prediction, these were assigned a greater weight (e.g., Dausman et al., 2010; Knowling et al., 2019; Fienen et al., 2020). This was achieved by scaling the inequality and groundwater-waste/stormwater flux observation group weights by 1×10^{-1}.

4 Results and discussion

The main outputs of this research are hazard informed maps for decision support. We therefore begin our examination of the results with this aspect of the study. We then discuss projected drainage volumes. This is followed by examining the value of history matching (e.g., Doherty and Moore, 2017), and the use of 'difference from a baseline,' or comparative outcomes of model predictions as an alternative approach, when investigating the SLR-driven groundwater hazard (e.g., Sepúlveda and Doherty, 2015).

The IPCC AR6 report introduced the SSP scenarios, which are representative of a broad range of plausible societal and climatic futures (IPCC, 2021). As detailed above, the focus of this research is the presented framework/workflow, so we mainly discuss results for the recommended SSP5-8.5 (medium confidence) high emissions scenario (MfE, 2017; see Table 2; Figure 3). However, we also briefly discuss and compare results for the SSP2-4.5 (medium confidence) scenario, which is included in the Supplementary Information.

4.1 Projected probability of groundwater inundation

To explore predictive uncertainty under IPCC projections of SLR, the estimated probability (and thus susceptibility) to groundwater inundation was based on a history matched (posterior) parameter ensemble. This posterior was derived using a Monte-Carlo representation of parameter uncertainty, that was propagated to the SLR projection period. The resulting probability of groundwater inundation was estimated from the posterior groundwater level distributions at every model cell, and collating the number of occasions that groundwater levels exceeded the model top elevation (i.e., exceedance probability, Figure 5). For the purposes of our research, susceptibility to inundation and probability are on the same general scale: highly susceptible areas correspond to the highest probabilities of groundwater levels exceeding the model top, and *vice versa*.

Using the SSP5-8.5 projection, the model simulated groundwater levels prior to 2030, indicate that the simulated probability of groundwater inundation is generally low across the South Dunedin model domain. This is likely associated with the low to moderate SLR projection and relatively constrained SLR uncertainty for this timeframe (Figure 3). By 2030, regions of higher groundwater inundation probability begin to emerge (Figure 4A). These regions become more defined by 2050 (Figure 4B) and are broadly constrained to three zones, in low-lying areas, surrounding the example site, I44/0006 and I44/1113 (Figure 1). This is consistent with reports of depths to groundwater of <0.5 m below the ground surface in these areas (Cox et al., 2020). It is not surprising, therefore, that these low-lying regions would be susceptible to inundation for low to moderate rises in sea level.

Under increasing (and accelerating) SLR for the 2070–2100 timeframe, the spatial extent of the more susceptible areas continues to increase (Figures 4C, D). As expected, the regions with the highest inundation probabilities are dominated by the same low-lying open areas, especially where there is an absence of drainage in the model (e.g., in the region of I44/0006, Figure 4). However, many additional zones do appear susceptible to the inundation hazard, despite being >1 m above sea level and a considerable distance inshore (e.g., in the region of I44/0005, Figure 4).

These same broad trends are apparent for the SSP2-4.5 (medium confidence) scenario. Although, the simulated probability and spatial extent of groundwater inundation is slightly diminished for the 2070–2100 timeframe (see Supplementary Figure S4-1). We attribute this reduction in susceptibility to the lower SLR projection attached to the model boundary condition for the SSP2-4.5 (medium confidence) scenario (see Supplementary Figure S4-2). Importantly, the lower likelihood high SLR realisations captured by our modelling approach leads to elevated probabilities of groundwater inundation for this timeframe. Our results suggest that even for the more optimistic SSP2-4.5 emissions pathway, significant susceptibility to the groundwater indundation hazard remains.

The zones most prone to groundwater inundation are not correlated with the distance from the Pacific Ocean or Otago Harbour boundary conditions (Figure 4). We hypothesize that this may be related to the increased hydraulic conductivity of the

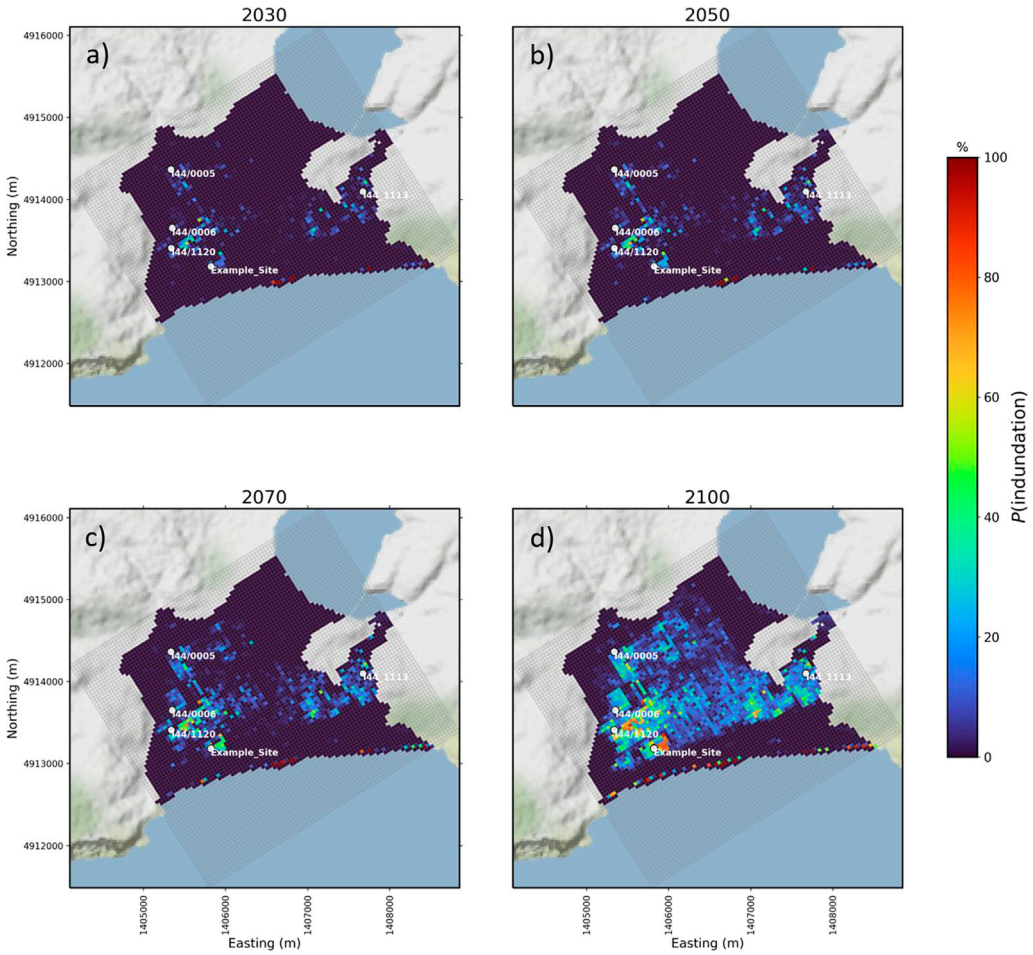

FIGURE 4
The projected SLR-driven probability of groundwater inundation for 2030, 2050, 2070 and 2100 based on the IPCC SSP5-8.5 (medium confidence) scenario (see Figure 3 for realizations of relative-SLR attached to GHB stage in the model). The model top elevation is based on a Digital Elevation Model (DEM) informed by recent LiDAR data (LINZ, 2021).

sediments and low topographic relief of these areas. Our results imply that groundwater emergence at a considerable distance inshore may occur before, or even compound overland flooding (e.g., Befus et al., 2020; Plane et al., 2019). This has implications for adaptation strategies that focus solely on overland flooding. Ignoring the effects of SLR-driven groundwater level rises may significantly underestimate the spatial extent and timing of surface water flooding (e.g., Anderson et al., 2018).

The presented inundation probabilities are all relative to the model top elevation (Figure 4), estimated from a mean aggregation of the LiDAR data (Section 2). It is acknowledged that the uncertainty of the LiDAR data, and how it is aggregated to the model top, has not been explicitly addressed in this study. A strong correlation is likely to exist between model surface elevation and simulated water levels. We believe that our framing of water level predictions as relative to the model top will help mitigate potential elevation and aggregation errors. However, caution should be exercised when attempting to assess inundation probabilities at scales less than the model grid resolution. Small-scale topographic features within a model cell may be characterised by higher (or lower) inundation probabilities than those predicted, at the model grid scale relative to the model top. The impact of elevation and aggregation uncertainty on predictions of groundwater inundation at a finer scale could be addressed in future work.

4.2 Simulated drain flows

The projected SLR-driven probability and spatial extent of inundation (Figure 4) is mitigated by the interaction between rising groundwater levels and the waste/stormwater drainage networks. This mitigating effect is controlled by the relative elevation of groundwater levels as sea levels rise, and also the spatial conductance of the drainage networks (an abstract numerical representation of the complex interaction between groundwater and the drainage networks, Table 1).

This effect is demonstrated by the total flux of groundwater discharging to the drainage networks represented in the model of South Dunedin, which is projected to increase substantially

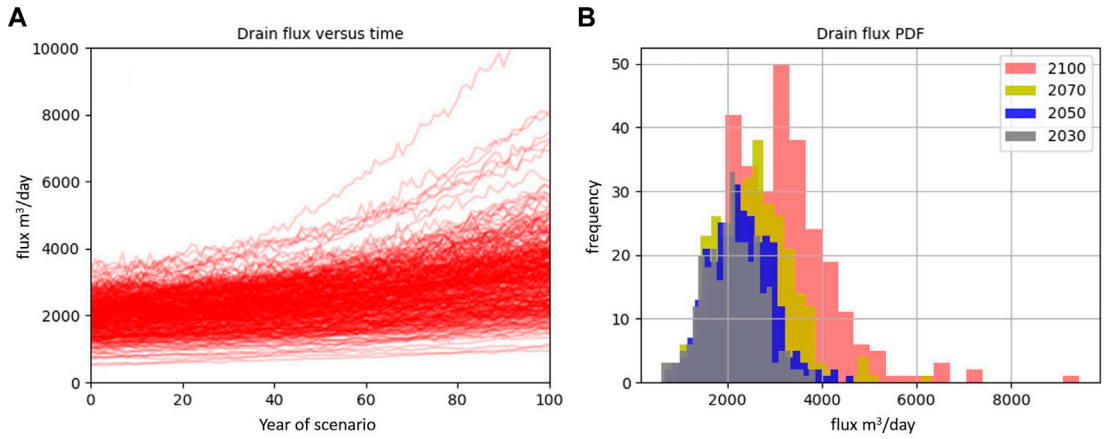

FIGURE 5
Plots showing **(A)** time-series of individual realizations of total drain fluxes, and **(B)** Probability Density Functions (PDFs) of projected total drain fluxes as groundwater levels change [IPCC SSP5-8.5 (medium confidence) scenario]. These plots show the estimated total groundwater flux to the waste/stormwater networks represented in the model of South Dunedin.

(Figure 5). As with the groundwater level predictions, the uncertainty of the total drain flux prediction also increases over the duration of the projection scenario (Figure 5). For example, in 2030, the mean and standard deviation of drainage flows are 2,150 and 494 m^3/day, respectively. In 2100, this increases to 2,835 and 718 m^3/day, respectively (a 32% increase in projected mean drainage flows).

Drain conductance and elevation are expressed as (nested) uncertain parameters in the numerical modelling workflow presented herein, but the history matching results indicate that the available data provides little information for condition of these parameters, especially in a spatial sense (see Supplementary Figure S3-7). Significant uncertainty persists for these posterior predictions. Additional monitoring, data collection and refinement of the estimated spatial (and temporal) fluxes to the existing drainage network may help reduce the uncertainty of these (and other) parameters, and thereby help to reduce the uncertainty of both drain flux and groundwater inundation predictions.

Notwithstanding the large uncertainty of these predictions, our results are consistent with other recent studies (e.g., Habel et al., 2017; Befus et al., 2020), which suggest that drainage may offset the impacts of SLR and emergent groundwaters. However, the planned renewal of the waste and stormwater networks in the 2020–2050 timeframe (Goldsmith and Hornblow, 2016) may limit, or even reduce the capacity of the drainage networks to accept infiltrating groundwater.

This has profound implications for decision-makers in South Dunedin, since our approach conservatively assumes that the drainage system will be available to accommodate SLR-driven groundwater level rises (Figure 5). This tenuous (linear) assumption may significantly overestimate the hydraulic response of the waste/stormwater networks for conditions that may prevail in the future. Decision-makers should therefore consider potential future limitations, or reductions of drainage flows in future management scenarios, to avoid underestimation of the potential groundwater inundation hazard.

The projected SLR-driven increase in the base flux to the waste/stormwater networks (Figure 5) will also be an important consideration. Increased "dry condition" flows to the drainage networks might limit their capacity for their primary function (removal of wastewater and stormwater, e.g., Morris et al., 2018). This has significant implications for managing event flows, since increases in the base flux may compound rises in groundwater levels in response to these events. Where the flows to these networks requires treatment and pumping to discharge, as is the case in South Dunedin, treatment facilities will likely receive higher loads at significant extra cost with ramifications for facility downtime and failure (e.g., Cox et al., 2020).

4.3 The value of history matching

The simulated outputs from the prior-based Monte Carlo uncertainty quantification displayed minimal prior-data conflict (PDC) in relation to the predictions of interest (groundwater levels; see Supplementary Figure S3-9). That is, prior simulated output distributions generally encompass the values of system observations. However, the prior uncertainty of simulated outputs was significant and contributed to predictions of relatively high probability of inundation (during the history period), across the model domain (Supplementary Figure S3-1). This high uncertainty in simulated outputs of management interest, the availability of aligned observations, and the general lack of PDC provided a defensible basis for undertaking history matching.

Six iterations using the iES algorithm were used to history match simulated outputs to historical observations (Section 3.2). This required a total of 3,240 model runs. The match to long-term average groundwater levels and total drain flux improved significantly in the first two iterations and levelled off following the fourth (Supplementary Table S3-1). After history matching, the posterior simulated groundwater level distributions generally encompass their respective observation within the defined observation error. The prior and posterior Probability Density

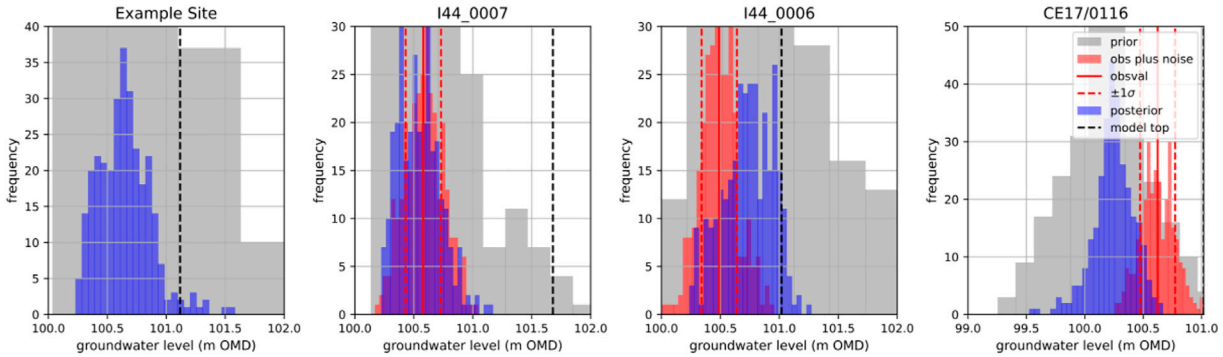

FIGURE 6
Histograms (PDFs) for selected observations showing prior Monte Carlo, posterior and observation plus noise iES distributions. Blue histograms show the distribution of model outputs and red histograms show the realizations of the observation value, which is based on the observed long-term (mean) groundwater level and supplied standard deviation (i.e., σ_m). Note, the unweighted example site for which no observation exists (Example Site). Note also, the x-range is truncated to focus on the posterior model outputs (history period).

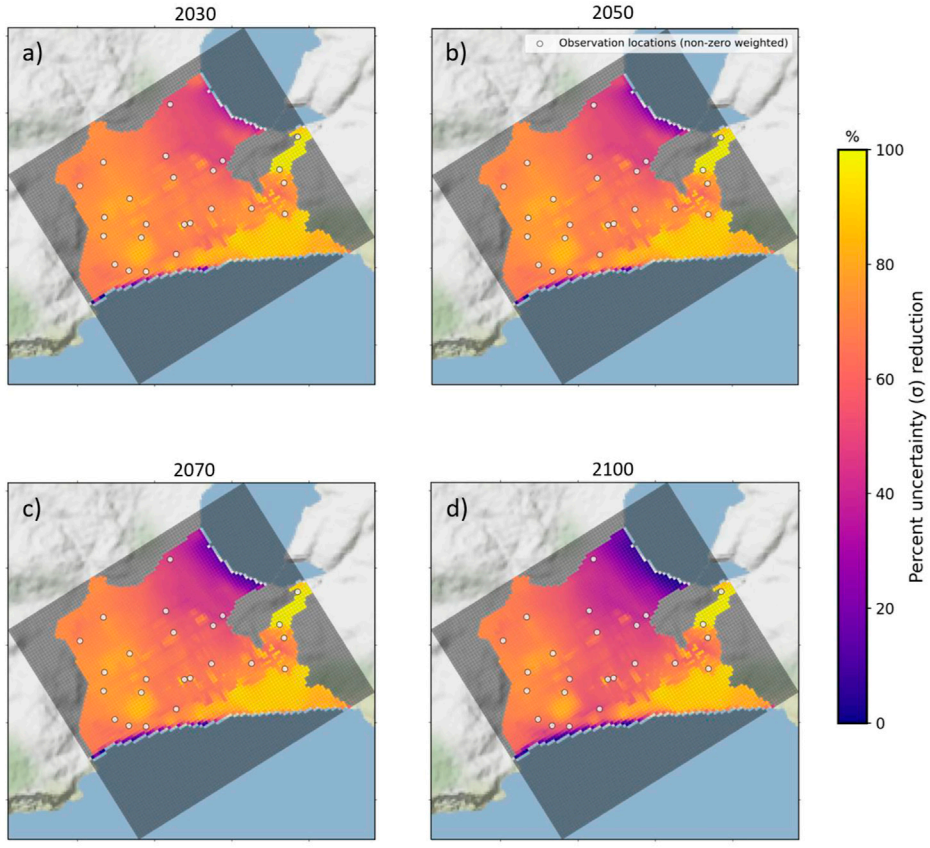

FIGURE 7
Percent uncertainty change for posterior *versus* prior distributions of groundwater level predictions for **(A)** 2030, **(B)** 2050, **(C)** 2070 and **(D)** 2100. Observation locations used for history matching are also shown (i.e., non-zero weighted groundwater level observations).

Functions (PDFs) for three observation locations, and an additional example site, are provided in Figure 6 (distributions for all observations are shown in Supplementary Figure S3-9).

As expected, history matching to several thousand observations (groundwater levels, inequality constraints and total groundwater flux) significantly reduced the uncertainty of simulated groundwater

level predictions, as indicated by the widths of the respective prior and posteriors PDFs in Figure 6 and Supplementary Figure S3-9. Through history matching, the simulated probability of groundwater inundation for the history period, was reduced to 0% across most of the model domain (see Supplementary Figure S3-8). We mainly attribute this improvement to the conditioning of horizontal hydraulic conductivity and drain conductance parameters through the assimilation of the information contained within the observation dataset (Supplementary Figure S3-4).

The history matching process outcome, i.e., the posterior parameter ensemble, can be considered to be effective, since the parameter ensemble was updated by the assimilation of information from observation data. It can also be concluded that these data were suitable for reducing the uncertainty of parameters to which the predictions were sensitive (Supplementary Figure S3-6).

However, results from both the history and projection periods, depict a high spatial and temporal variation in the uncertainty reduction of the groundwater level simulated output that results from history matching (Figure 7 and Supplementary Figure S3-6). The spatial distribution of observation data, the updated impervious surface recharge model (Supplementary Figure S2-1), and simulated drainage clearly plays an important role in the spatial distribution of uncertainty reduction. For example, generally, the largest uncertainty reductions occur over pervious surfaces where the observation density is high, and where there is absence of drainage in the model (e.g., to the southwest of the model domain).

Uncertainty reductions are generally high (>60%) for the history period (Supplementary Figure S3-6) and for the projected 2030–2050 timeframe (Figures 7A, B). As discussed, the conditioning of parameters to historical observations propagated this uncertainty reduction to the projection period groundwater level predictions. Before 2050, our results suggest that steady-state only history matching can indeed reduce the uncertainty of groundwater level predictions, despite the intractable nature of the uncertainty inherited from the IPCC projections of SLR (e.g., Kopp et al., 2019).

Generally, however, posterior prediction uncertainty increases substantially for the 2070–2100 timeframe (Figures 7C, D). Spatially, the history matching constrained uncertainty increases are mainly isolated to locations where drainage is represented in the model, and to the northeast of the model domain where groundwater level observation data is sparse (i.e., near the harbour boundary condition). For the groundwater level prediction, we mainly attribute this loss in spatial confidence to the large uncertainty of the drainage parameters, and the uncertainty inherited from the IPCC-SLR projection, which increases precipitously for the 2070–2100 timeframe (see, e.g., Figure 3).

In this context, groundwater inundation assessments typically rely on the use of a single deterministic realisation of SLR (e.g., median or p83 scenario, see Table 2). Unfortunately, these approaches eschew the deep uncertainty attached to the IPCC-SLR projections themselves (e.g., Kopp et al., 2019), and do not allow expert knowledge to be considered through weighting the likelihood of SLR over the full range of scenario projections (e.g., Purvis et al., 2008). We have therefore presented a consistent methodology to explore the full range of SLR projections and their impact on the decision-critical predictions (see, e.g., Figure 7).

4.4 Predictions of relative change (i.e., differences)

The availability of appropriate groundwater monitoring datasets, particularly at the spatial density and duration of the results presented herein, is relatively rare, especially compared to the global number of at-risk, coastal communities and ecosytems (e.g., Neumann et al., 2015; Hooijer and Verminnen, 2021). For predictions of absolute groundwater levels and inundation, a lack of monitoring data may limit the potential for history matching to condition (and reduce) model parameter and corresponding prediction uncertainty. Our results suggest that the prior uncertainty of these absolute predictions may be too high to provide any meaningful information in terms of robust decision-support (see Supplementary Figure S4-2).

A considered reframing of the projection simulations, to predict the relative changes of model predictions (i.e., differences in projected groundwater levels, Figure 8) may, in practice, be a better approach for communities that do not have dense monitoring networks. Such an approach should reduce the impact of model structural errors on predictive uncertainty, and may also help to mitigate the contribution to uncertainty inherited from the prior parameter distributions and structural defects of the groundwater model (e.g., Sepúlveda and Doherty, 2015).

The results presented in Figure 4 display the distribution of changes in groundwater levels relative to an arbitrary "decision threshold" (e.g., Knowling et al., 2019; White et al., 2019), which in this instance is emergence over the model top (or land surface) estimated from LiDAR data. Clearly, for relative-type predictions, such a decision threshold is not available. An alternative is to define a threshold based on an anticipated impactful change in groundwater levels. For example, Figure 8 uses a decision threshold of a 0.25 m increase in groundwater levels (for the same sites presented in Figure 6). Probabilistic mapping of simulated outputs against this (or multiple) relative decision thresholds is also possible.

For predictions of relative change, the projected prior *versus* posterior probability of groundwater levels exceeding the decision threshold appears relatively low for the 2030–2050 timeframe (Figure 8). Similar to the predictions of absolute values, there is then a marked increase in the probability of groundwater levels exceeding the difference threshold for the 2070–2100 timeframe.

However, in contrast to predictions of absolute values, there is a surprising lack of discrepancy between the prior and posterior difference projections (Figure 8). This is consistent with the accepted logic that models are more suitable predictors of relative change, rather than absolutes (e.g., Sepúlveda and Doherty, 2015), which also aligns with conclusions drawn from a number of other recent studies (e.g., Knowling et al., 2019; White et al., 2020).

Our results indicate that the workflow deployed here for South Dunedin could be modified and deployed with reasonable utility, even in settings with limited (or unreliable data), by curtailing, or forgoing the history matching step, and exploring predictions in a relative sense. It may then be possible to delineate areas that are more susceptible to SLR-driven groundwater level rises, or demonstrate the merits of one management strategy *versus* another. This has implications for the way in which a numerical model is used for decision-support, and the type of information that decision makers may wish to obtain from numerical models.

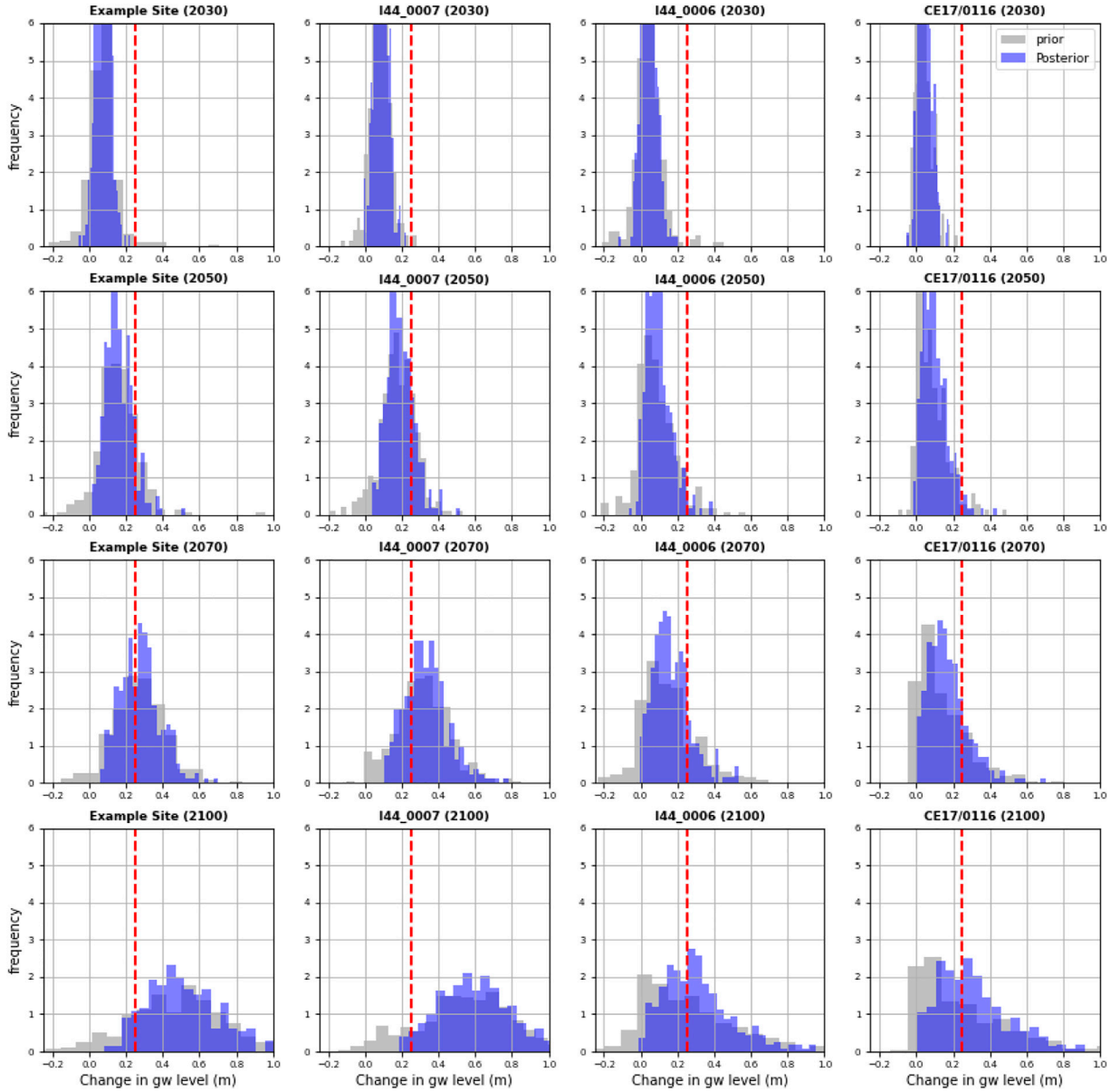

FIGURE 8
Prior (grey) *versus* posterior (blue) distributions for the projected change in groundwater levels (m) at the selected sites for the SSP5-8.5 (medium confidence) scenario. An arbitrary decision threshold of 0.25 m is also illustrated (red dashed line). The projected change in groundwater levels is calculated from the difference between year 0 and the given year of the projection model (for each individual realization).

4.5 Further considerations and recommendations

The history matching informed predictive distributions of groundwater levels presented herein supports quantification of the uncertainties in groundwater level rise and inundation, for stresses that may prevail in the future. It is acknowledged that the modelling workflow does not capture all of the potential contributing sources to predictive uncertainty. We therefore adopted a highly parameterised approach and defined broad prior parameter uncertainties to provide some protection against prediction uncertainty underestimation. Although, some uncertainties relating to error in model structure and conceptualisation (e.g., Wagener et al., 2021) may persist, unaccounted for.

Consequently, caution should be exercised in the application of these results. As discussed in Section 4.1, it may be inappropriate to apply these results at spatial scales finer than the model grid resolution. Similarly, for temporal scales, the model projections represent the long-term progression of annual conditions to estimate a general "annual" exposure to a hazard, or change in exposure to a hazard. Detailed hazard, vulnerability and damage thresholds also commonly encompass short-term fluctuations and events (e.g., Paulik et al., 2019). The direct application of these

results to temporal scales that are finer than the model temporal resolution is also likely to be inappropriate. Nevertheless, the modelling workflow and results presented herein may serve as a basis for making downscaled (both temporally and spatially) predicitions.

A real strength is the scripted nature of our workflow, which facilitates such (follow up) investigations, whilst supporting the incorporation of model revisions and exploration of alternative management (e.g., SSP or drainage) scenarios, in a way that is rapid, reproducible and transparent. The workflow could easily be extended to implement dataworth analyses to establish the value of existing and yet to be collected monitoring data. Or, for example, the cost of exploring (or foregoing) transient history matching in terms of reducing predictive uncertainty at finer temporal scale, in an events based models (e.g., Moore and Doherty, 2005).

In this context, it is recommended that future research should explore predictions of an episodic nature, such as improving model-based predictions of groundwater levels in response to individual storm-surge or rainfall events (e.g., a rainfall event with a return period of 10 years). It would then be possible to begin to address the fundamental question of how these events interact with rising sea levels and a changing climate.

5 Conclusion

The potential for a spatial and temporal detailed map of groundwater inundation probabilities, and corresponding drainage volumes that may be required to mitigate SLR, was investigated in this study. While the mapping of groundwater inundation is discussed in Morris et al. (2018) and others, projecting this mapping into a risk framework has been missing from the literature. The distributed properties that support the risk maps of groundwater inundation in response to SLR extends the recent work of Merchán-Rivera et al. (2022), which also applied a Bayesian framework to the creation of risk maps, but used spatially lumped hydraulic properties. The spatially distributed hydraulic properties adopted in this work enabled a detailed delineation of areas that is not possible using a spatially lumped parameterisation scheme. The Bayesian methodology adopted supports a regional scale delineation of the distribution of groundwater inundation projections.

Our approach has attempted to equip decision-makers with all the necessary information to distinguish where the probability for groundwater inundation is relatively high, and where it is relatively low. This approach also includes providing a level of confidence that a proposed decision threshold will be exceeded, which may necessitate the implementation of a (potentially costly) management strategy. However, knowledge of actual thresholds for damage, and therefore asset vulnerability, appears to be missing from the literature. In this regard, the tolerable probability of groundwater inundation, and how this translates more broadly into risk, remains for decision-makers to determine.

Previous studies on groundwater responses to SLR have focussed on groundwater flooding areas, or the movement of the fresh-salt water interface. However, the mitigation of groundwater flooding, at least initially, is likely to involve consideration of the additional flows that drainage networks may be required to accomodate. This study extends previous work by explicitly focussing on the likelihood of relative increases in drainage flows, given its importance as a management consideration.

The uncertainty of the SLR projections represents a small contribution to the uncertainty of the groundwater flooding probabilities for predictions within the next few decades. As the projections extend further into the future, however, the SLR uncertainty begins to dominate the uncertainty of the groundwater flooding predictions. This highlights the necessity of exploring model uncertainty in the context of the prediction being made (Doherty, 2015). For near-time predictions, history matching appears to reduce the uncertainty of groundwater level rises, whereas the same cannot be said for predictions in the distant future.

Also demonstrated was the relative value of history matching when formulating predictions as a difference from a baseline, rather than the absolute value of a prediction. For the specific predictions and history matching dataset combination explored, the worth of history matching was doubtful when casting the prediction as a difference from a baseline. Whereas history matching was useful if the absolute magnitude of the groundwater level was of concern. This issue was also explored in different contexts in Knowling et al. (2019), Hemmings et al. (2021), Moore and Doherty (2005) and others.

Finally, we note that the analysis described in this paper was supported by a scripted workflow (e.g., White et al., 2020). The combination of a spatially and temporally distributed parameterisation scheme, history matching and uncertainty quantification over a regional scale is complex. This scripted workflow provides a transparent record of the many (unavoidable subjective) decisions made during our modelling process, whilst supporting similar analyses that could easily extend the scripted workflow provided in the Supplementary Information.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

LC and BH were responsible for coding of the workflow and analysis of the results. SC and CM supported analysis of the results. MK and KH supported the development of the workflow. FM, PG and JR supported the development of the numerical model. All authors contributed to the writing of the manuscript.

Funding

This research was supported by the NZ SeaRise Endeavour programme, funded by the Ministry of Business, Innovation and Employment (MBIE) contract RTVU1705. This research programme was conducted in collaboration with GNS Science, Victoria University of Wellington (VUW) and NIWA. The case study example development was supported by Otago Regional Council (ORC) and Dunedin City Council (DCC) who monitor groundwater and are cognisant of the many issues impacting their community's future.

Acknowledgments

We would also like to thank the helpful and constructive reviews provided by two reviewers. These reviews substantially improved the manscript.

Conflict of interest

Author KH was employed by Groundwater Solutions Pty., Ltd. and Author JR was employed by Kōmanawa Solutions Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2023.1111065/full#supplementary-material

References

Abboud, J. M., Ryan, M. C., and Osborn, G. D. (2018). Groundwater flooding in a river-connected alluvial aquifer. *J. Flood Risk Manag.* 11 (4), e12334. doi:10.1111/jfr3.12334

Anderson, T. R., Fletcher, C. H., Barbee, M. M., Frazer, L. N., and Romine, B. M. (2015). Doubling of coastal erosion under rising sea level by mid-century in Hawaii. *Nat. Hazards* 78 (1), 75–103. doi:10.1007/s11069-015-1698-6

Anderson, T. R., Fletcher, C. H., Barbee, M. M., Romine, B. M., Lemmo, S., and Delevaux, J. (2018). Modeling multiple sea level rise stresses reveals up to twice the land at risk compared to strictly passive flooding methods. *Sci. Rep.* 8 (1), 1–14. doi:10.1038/s41598-018-32658-x

Barlow, P. M., Wagner, B. J., and Belitz, K. (1996). Pumping strategies for management of a shallow water table: The value of the simulation-optimization approach. *Groundwater* 34 (2), 305–317. doi:10.1111/j.1745-6584.1996.tb01890.x

Befus, K. M., Barnard, P. L., Hoover, D. J., Finzi Hart, J. A., and Voss, C. I. (2020). Increasing threat of coastal groundwater hazards from sea-level rise in California. *Nat. Clim. Change* 10 (10), 946–952.

Bell, R., Hannah, J., and Andrews, C. (2022). *Update to 2020 of the annual mean sea level series and trends around New Zealand*. Hamilton, New Zealand: Prepared for Ministry for the Environment. August 2022 (NIWA Client Report No: 2021236HN). NIWA. Available at: https://environment.govt.nz/publications/update-to-2020-of-the-annual-mean-sea-level-series-and-trends-around-new-zealand.

Bell, R., Lawrence, J., Allan, S., Blackett, P., and Stephens, S. (2017). *Coastal hazards and climate change: Guidance for local government*. Wellington (NZ): Ministry for the Environment, 279.

Bjerklie, D. M., Mullaney, J. R., Stone, J. R., Skinner, B. J., and Ramlow, M. A. (2012). "Preliminary investigation of the effects of sea-level rise on groundwater levels in New Haven, Connecticut," in *Open-File Report 2012-1025. US Geological Survey* (Reston, Virginia: The US Geological Survey).

Box, J. E., Hubbard, A., Bahr, D. B., Colgan, W. T., Fettweis, X., Mankoff, K. D., et al. (2022). Greenland ice sheet climate disequilibrium and committed sea-level rise. *Nat. Clim. Chang.* 12, 808–813. doi:10.1038/s41558-022-01441-2

Caers, J. (2018). "Bayesianism in the geosciences," in *Handbook of mathematical geosciences* (Fifty Years of IAMG), 527–566. doi:10.1007/978-3-319-78999-6_27

Colombo, L., Gattinoni, P., and Scesi, L. (2018). Stochastic modelling of groundwater flow for hazard assessment along the underground infrastructures in Milan (northern Italy). *Tunn. Undergr. Space Technol.* 79, 110–120. doi:10.1016/j.tust.2018.05.007

Cooper, H. M., Fletcher, C. H., Chen, Q., and Barbee, M. M. (2013). Sea-level rise vulnerability mapping for adaptation decisions using LiDAR DEMs. *Prog. Phys. Geogr.* 37 (6), 745–766. doi:10.1177/0309133313496835

Cox, S. C., Ettema, M. H. J., Mager, S. M., Glassey, P. J., Hornblow, S., and Yeo, S. (2020). *Dunedin groundwater monitoring and spatial observations. Lower Hutt (NZ)*. Avalon, Wellington, New Zealand: GNS Science GNS Science, 86. Report 2020/11. doi:10.21420/AVAJ-EE81

Cox, S. C., van Ballegooy, S., Rutter, H. K., Harte, D. S., Holden, C., Gulley, A. K., et al. (2021). Can artesian groundwater and earthquake-induced aquifer leakage exacerbate the manifestation of liquefaction? *Eng. Geol.* 281, 105982. doi:10.1016/j.enggeo.2020.105982

Dausman, A. M., Doherty, J., Langevin, C. D., and Sukop, M. C. (2010). Quantifying data worth toward reducing predictive uncertainty *Groundwater* 48 (5), 729–740.

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models*, 227pp. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., and Moore, C. (2017). "Simple is beautiful," in *Moore C. 2019. Groundwater modelling uncertainty – implications for decision making. Summary report of the national groundwater modelling uncertainty workshop, 10 July 2017*. Editors H. Middlemis, G. Walker, L. Peeters, S. Richardson, and P. Hayes (Sydney, Australia: Flinders University, National Centre for Groundwater Research and Training). doi:10.25957/5ca5641defe56

Doherty, J., and Welter, D. (2010). A short exploration of structural noise. *Water Resour. Res.* 46 (5).

Egidi, L., Pauli, F., and Torelli, N. (2022)., 28. Avoiding prior–data conflict in regression models via mixture priors. *Can. J. Stat.* 50 (2), 491–510.

Fienen, M. N., Corson-Dosch, N. T., White, J. T., Leaf, A. T., and Hunt, R. J. (2022). Risk-based wellhead protection decision support: A repeatable workflow approach. *Groundwater* 60 (1), 71–86.

Freeze, R. A., Massmann, J., Smith, L., Sperling, T., James, B., and Fordyce, E. (1990)., 28. Hydrogeological decision analysis: 1. A framework groundwater 5Groundwater dynamics of a shallow coastal aquifer [MAppSc thesis]. Dunedin (NZ): University of Otago. 155p, 738–766.

Fordyce, E. (2014). *Groundwater dynamics of a shallow coastal aquifer*. (Doctoral dissertation, University of Otago). 155.

Glassey, P., Barrell, D., Forsyth, J., and Macleod, R. (2003). The geology of Dunedin, New Zealand, and the management of geological hazards. *Quat. Int.* 103, 23–40. doi:10.1016/S1040-6182(02)00139-8

Goldsmith, M., and Hornblow, S. (2016). *The natural hazards of South Dunedin*. Dunedin, New Zealand: Otago Regional Council Report, Otago Regional Council, 69.

Goldsmith, M., Payan, J-L., MorrisValentine, R. C., MacLean, S., Xiaofeng, L., Vaitupu, N., et al. (2015). *Coastal Otago flooding flood event 3 june 2015*. Dunedin (NZ): Otago Regional Council, 56p.

Habel, S., Fletcher, C. H., Rotzoll, K., and El-Kadi, A. I. (2017). Development of a model to simulate groundwater inundation induced by sea-level rise and high tides in Honolulu, Hawaii. *Water Res.* 114, 122–134. doi:10.1016/j.watres.2017.02.035

Habel, S., Fletcher, C. H., Rotzoll, K., El-Kadi, A. I., and Oki, D. S. (2019). Comparison of a simple hydrostatic and a data-intensive 3D numerical modeling method of simulating sea-level rise induced groundwater inundation for Honolulu, Hawai'i, USA. *Environ. Res. Commu-nications* 1 (4), 041005. Available at:. doi:10.1088/2515-7620/ab21fe

Hall, J. A., Gill, S., Obeysekera, J., Sweet, W., Knuuti, K., and Marburger, J. (2016). *Regional Sea level scenarios for coastal risk management: Managing the uncertainty of future Sea Level change and extreme water levels for department of defense coastal sites worldwide*. Alexandria: U.S. Department of Defense, Strategic Environmental Research and Development Program.

Hemmings, B., Knowling, M. J., and Moore, C. R. (2020). Early uncertainty quantification for an improved decision support modeling workflow: A streamflow reliability and water quality example. *Front. Earth Sci.* 8, 565613. doi:10.3389/feart.2020.565613

Hooijer, A., and Verminnen, R. (2021). Global LIDAR land elevation data reveal greatest sea-level rise vulnerability in the tropics. *Nat. Commun.* 12, 3592–3597. doi:10.1038/s41467-021-23810-9

Hoover, D. J., Odigie, K. O., Swarzenski, P. W., and Barnard, P. L. (2016). Sea-level rise and coastal groundwater inundation and shoaling at select sites in California, USA. *J. Hydrology Regional Stud.* 11, 234–249. doi:10.1016/j.ejrh.2015.12.055

Hugman, R., and Doherty, J. (2022). Complex or simple-does a model have to be one or the other? *Front. Earth Sci* 10, 1–12. doi:10.3389/feart.2022.867379

Hunt, R. J., White, J. T., Duncan, L., Haugh, C., and Doherty, J. (2021). Evaluating lower computational burden approaches for calibration of large environmental models. *Ground Water* 59 (6), 788–798. doi:10.1111/gwat.13106

Jevrejeva, S., Grinsted, A., and Moore, J. C. (2009). Anthropogenic forcing dominates sea-level rise since 1850. *Geophys. Res. Lett.* 36, L20706. doi:10.1029/2009gl040216

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., et al. (2016). "Jupyter notebooks—a publishing format for reproducible computational workflows," in *Positioning and power in academic publishing: Players, agents and agendas*. Editors F. Loizides and B. Schmidt (Amsterdam: IOS Press) 87–90.

Knott, J. F., Elshaer, M., Daniel, J. S., Jacobs, J. M., and Kirshen, P. (2017). Assessing the effects of rising groundwater from sea level Rise on the Service life of pavements in coastal road infrastructure Transport. *Res. Rec. J. Transp. Res. Board.* No. 11.

Knowling, M., White, J., and Moore, C. (2019). Role of model parameterization in risk-based decision support: An empirical exploration. *Adv. Water Resour.* 128, 59–73. doi:10.1016/j.advwatres.2019.04.010

Kopp, R. E., Gilmore, E. A., Little, C. M., Lorenzo-Trueba, J., Ramenzoni, V. C., and Sweet, W. V. (2019). Usable science for managing the risks of sea-level rise. *Earth's Future* 7 (1), 1235–1269. doi:10.1029/2018EF001145

Kopp, R. E., Horton, R. M., Little, C. M., Mitrovica, J. X., Oppenheimer, M., Rasmussen, D. J., et al. (2014). Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites. *Earth's Future* 2, 383–406. doi:10.1002/2014EF000239

Land Information New Zealand (2021). *Otago – Dunedin and mosgiel LiDAR 1m DSM*. Available at: https://data.linz.govt.nz/layer/107710-otago-dunedin-and-mosgiel-lidar-1m-dem-2021/.

Ministry for the Environment (2017). "Coastal hazards and climate change: Guidance for local government," in *Prepared for the Ministry for the environment by Bell RG*. Editors J. Lawrence, S. Allan, P. Blackett, and S. A. Stephens (Wellington: Ministry for the Environment).

Macdonald, D., Dixon, A., Newell, A., and Hallaways, A. (2012). Groundwater flooding within an urbanised flood plain: Groundwater flooding within urbanised flood plain. *J. Flood Risk Manage.* 5, 68–80. doi:10.1111/j.1753-318X.2011.01127.x

IPCC (2021). "Climate change 2021: The physical science basis," in *Contribution of working group I to the sixth assessment report of the intergovernmental Panel on climate change*. Editors V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press). In press. doi:10.1017/9781009157896

May, C. (2020). Rising groundwater and sea-level rise. *Nat. Clim. Change* 10, 889–890. doi:10.1038/s41558-020-0886-x

McCobb, T. D., and Weiskel, P. K. (2003). Long-term hydrologic monitoring protocol for coastal ecosystems. *U.S. Geol. Surv. Open-File Rep.* 02-497, 94p.

McKenna, S. A., Akhriev, A., Echeverría Ciaurri, D., and Zhuk, S. (2020). Efficient uncertainty quantification of reservoir properties for parameter estimation and production forecasting. *Math. Geosci.* 52 (2), 233–251. doi:10.1007/s11004-019-09810-y

McKenzie, A. A., Rutter, H. K., and Hulbert, A. G. (2010). The use of elevation models to predict areas at risk of groundwater flooding. *Geol. Soc. Spec. Publ.* 345 (1), 75–79.

Merchán-Rivera, P., Geist, A., Disse, M., Huang, J., and Chiogna, G. (2022). A Bayesian framework to assess and create risk maps of groundwater flooding. *J. Hydrol.* 610, 127797. doi:10.1016/j.jhydrol.2022.127797

Moore, C., and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resour. Res.* 41 (5). doi:10.1029/2004wr003501

Morris, S. E., Cobby, D., Zaidman, M., and Fisher, K. (2018). Modelling and mapping groundwater flooding at the ground surface in Chalk catchments: Modelling and mapping groundwater flooding. *J. Flood Risk Manage.* 11, S251–S268. doi:10.1111/jfr3.12201

Mourot, F. M., Westerhoff, R. S., White, P. A., and Cameron, S. G. (2022). Climate change and New Zealand's groundwater resources: A methodology to support adaptation. *J. Hydrol. Reg. Stud.* 40, 101053.

Neumann, B., Vafeidis, A. T., Zimmermann, J., and Nicholls, R. J. (2015). Future coastal population growth and exposure to sea-level rise and coastal flooding - a global assessment. *PLoS One* 10, e0118571. doi:10.1371/journal.pone.0118571

Nicholls, R. J., Wong, P. P., Burkett, V. R., Codignotto, J. O., Hay, J. E., McLean, R. F., et al. (2007). "Coastal systems and low-lying areas," in *Climate change 2007. Impacts, adaptation and vulnerability. Contribution ofworking group II to the fourth assessment report of the intergovernmental panel on climate change*. Editors M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. Van Der Linden, and C. E. Hanson (Cambridge, UK: Cambridge University Press), 315–356.

Nicholls, R. J., Lincke, D., Hinkel, J., Brown, S., Vafeidis, A. T., Meyssignac, B., et al. (2021). A global analysis of subsidence, relative sea-level change and coastal flood exposure. *Nat. Clim. Chang.* 11, 338–342. doi:10.1038/s41558-021-00993-z

Niswonger, R. G., Panday, S., and Ibaraki, M. (2011). MODFLOW-NWT, a Newton formulation for MODFLOW-2005. US Geological Survey Techniques and Methods 6 (A37), 44.

Opus International Consultants Ltd and URS New Zealand Ltd (2011a). *Integrated catchment management plans 2010-2060: Phase 1 wastewater system – final report. 113 p. + 9 appendices*. Prepared for Dunedin City Council.

Opus International Consultants Ltd and URS New Zealand Ltd (2011b). *Integrated catchment management plans 2010-2060: Phase 2 wastewater – model build and hydraulic system performance report*. Prepared for Dunedin City Council, 84. + 7 appendices.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9 (1), 62–66.

Parliamentary Commissioner for the Environment (2015). *Preparing New Zealand for rising seas: Certainty and uncertainty*. Wellington (NZ): Parliamentary Commissioner for the Environment, 92.

Paulik, R., Stephens, S., Wadwha, S., Bell, R., Popovich, B., and Robinson, B. (2019). *Coastal flooding exposure under future sea-level rise for New Zealand*. Wellington, NZ: National Institute of Water & Atmospheric Research. NIWA Client Report 2019119WN, prepared for The Deep South Science Challenge, 2019.

Plane, E., Hill, K., and May, C. (2019). A rapid assessment method to identify potential groundwater flooding hotspots as sea levels rise in coastal cities. *Water* 11 (11), 2228. doi:10.3390/w11112228

Purvis, M. J., Bates, P. D., and Hayes, C. M. (2008). A probabilistic methodology to estimate future coastal flood risk due to sea level rise. *Coast. Eng.* 55 (12), 1062–1073. doi:10.1016/j.coastaleng.2008.04.008

Ramm, T. D., Watson, C. S., and White, C. J. (2018). Strategic adaptation pathway planning to manage sea-level rise and changing coastal flood risk. *Environ. Sci. Policy* 87, 92–101. doi:10.1016/j.envsci.2018.06.001

Ramm, T. D., White, C. J., Chan, A. H. C., and Watson, C. S. (2017). A review of methodologies applied in Australian practice to evaluate long-term coastal adaptation options. *Clim. Risk Manag.* 17, 35–51. doi:10.1016/j.crm.2017.06.005

Rekker, J. (2012). *The South Dunedin coastal aquifer & effect of sea level fluctuations*. Dunedin, New Zealand: Otago Regional Council, 25.

Rekker, J. (2021). *Tonga Park – aquifer test report oct 2021*. Dunedin, New Zealand: Otago Regional Council, 40. Otago Regional Council Report.

Rotzoll, K., and Fletcher, C. (2013). Assessment of groundwater inundation as a consequence of sea-level rise. *Nat. Clim. Change* 3, 477–481. doi:10.1038/nclimate1725

Sepúlveda, N., and Doherty, J. (2015). Uncertainty analysis of a groundwater flow model in east-Central Florida. *Groundwater* 53 (3), 464–474. doi:10.1111/gwat.12232

Storlazzi, C. D., Gingerich, S. B., Cheriton, O. M., Swarzenski, P. W., Quataert, E., Voss, C. I., et al. (2018). Most atolls will Be uninhabitable by the mid-21st century because of Sea-level rise exacerbating wave-driven flooding. *Sci. Adv.* 4 (4), eaap9741. doi:10.1126/sciadv.aap9741

Sweet, W., Park, J., Marra, J., Zervas, C., and Gill, S. (2014). *Sea level rise and nuisance flood frequency changes around the United States*. Washington, DC: National Oceanic and Atmospheric Administration (NOAA), Technical Report.

Vermeer, M., and Rahmstorf, S. (2009). Global sea level linked to global temperature. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21527–21532. doi:10.1073/pnas.0907765106

Wagener, T., Gleeson, T., Coxon, G., Hartmann, A., Howden, N., Pianosi, F., et al. (2021). On doing hydrology with dragons: Realizing the value of perceptual models and knowledge accumulation. *Wiley Interdiscip. Rev. Water* 8 (6), e1550. doi:10.1002/wat2.1550

Watson, C. S., White, N. J., Church, J. A., King, M. A., Burgette, R. J., and Legresy, B. (2015). Unabated global mean sea-level rise over the satellite altimeter era. *Nat. Clim. Change* 5 (6), 565–568. doi:10.1038/nclimate2635

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020). Toward reproducible environmental modeling for decision support: A worked example. *Front. Earth Sci.* 8, 50. doi:10.3389/feart.2020.00050

White, J. T., Knowling, M. J., and Moore, C. R. (2019). Consequences of model simplification in risk-based decision making: An analysis of groundwater-model vertical discretization. *Groundwater* 58, 695–709. doi:10.1111/gwat.12957

Yi, S., Sun, W., Heki, K., and Qian, A. (2015). An increase in the rate of global mean sea level rise since 2010. *Geophys. Res. Lett.* 42 (10), 3998–4006. doi:10.1002/2015GL063902

Yu, X., Moraetis, D., Nikolaidis, N. P., Li, B., Duffy, C., and Liu, B. (2019). A coupled surface-subsurface hydrologic model to assess groundwater flood risk spatially and temporally. *Environ. Modell. Softw.* 114, 129–139. doi:10.1016/j.envsoft.2019.01.008

frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 04 July 2023
DOI 10.3389/feart.2023.1168609

OPEN ACCESS

EDITED BY
Michael Fienen,
United States Geological Survey,
United States

REVIEWED BY
Kevin Hiscock,
University of East Anglia, United Kingdom
Laura Schachter,
United States Geological Survey,
United States
Santosh Murlidhar Pingale,
National Institute of Hydrology (Roorkee),
India

*CORRESPONDENCE
Nikolas Benavides Höglund,
nikolas.hoglund@geol.lu.se

RECEIVED 17 February 2023
ACCEPTED 22 June 2023
PUBLISHED 04 July 2023

CITATION
Benavides Höglund N, Sparrenbom C and Hugman R (2023), A probabilistic assessment of surface water-groundwater exchange flux at a PCE contaminated site using groundwater modelling.
Front. Earth Sci. 11:1168609.
doi: 10.3389/feart.2023.1168609

A probabilistic assessment of surface water-groundwater exchange flux at a PCE contaminated site using groundwater modelling

Nikolas Benavides Höglund[1]*, Charlotte Sparrenbom[1] and Rui Hugman[2]

[1]Department of Geology, Lund University, Lund, Sweden, [2]INTERA, Perth, WA, Australia

Polluted groundwater discharge at a chlorinated solvent contaminated site in Hagfors, Sweden, is affecting a nearby stream flowing through a sparsely populated area. Because of difficulties related to source zone remediation, decision makers recently changed the short-term site management objective to mitigating discharge of polluted groundwater to the stream. To help formulating targeted remediation strategies pertaining to the new objective, we developed a groundwater numerical decision-support model. To facilitate reproducibility, the modelling workflow was scripted. The model was designed to quantify and reduce the uncertainty of surface water-groundwater (SW-GW) exchange fluxes for the studied period (2016–2020) through the use of history-matching. In addition to classical observations, thermal anomalies detected in fiber optic distributed temperature sensing (FO-DTS) measurements were used to inform the model of groundwater discharge. After assessing SW-GW exchange fluxes, we used measurements of surface water chemistry to provide a probabilistic estimation of mass influx and spatio-temporal distributions of contaminated groundwater discharge. Results show 1) SW-GW exchange fluxes are likely to be significantly larger than previously estimated, and 2) prior estimations of mass influx are located near the center of the posterior probability distribution. Based on this, we recommend decision makers to focus remediation action on specific segments of the stream.

KEYWORDS

groundwater modelling, surface water-groundwater interaction, uncertainty quantification, groundwater contamination, tetrachloroethylene, PCE, distributed temperature sensing

1 Introduction

High quality freshwater is not an endless resource, and for that reason, we have a responsibility to limit the effects of past and current human actions on future water quality and quantity. Old sins of industrial malpractice and alike lurk in the underground and contaminated sites constitutes a global problem at a local scale (Schmoll et al., 2006).

As groundwater move through the subsurface, nearby surface waters are at risk of contamination through transport and discharge of polluted groundwater, ultimately putting public health at risk through exposure. Exploring surface water-groundwater (SW-GW)

exchange behavior at these sites is essential to discover locations of polluted discharge and formulate targeted remediation strategies to ensure good enough water quality for future needs at a reasonable cost.

There are several methods available for estimating SW-GW exchange flux, and their suitability typically vary depending on the scale of the investigation. For small scale estimations, direct measurements using seepage meters can be utilized (e.g., Rosenberry, 2008). For larger scales, groundwater and surface-water stage monitoring networks can be interpreted to estimate SW-GW exchange using Darcy's Law (Woessner, 2020). Indirect methods for inferring exchange fluxes include inference from temperature measurements (Andersson, 2005), and other geophysical and geochemical tracers, such as electrical conductivity (EC) and stable and radioactive isotopes (Cook, 2013). For characterization in high detail at small to medium scale (up to 30 km of cable length), fiber optic distributed temperature sensing (FO-DTS) (Selker et al., 2006) has shown to be a promising method (Briggs et al., 2012).

Numerical models can be used in a variety of contexts where SW-GW exchange affect a prediction of interest. Lately, the use of groundwater numerical models as a means to quantify SW-GW exchange fluxes has gained prominence (Ntona et al., 2022). However, to be useful in a decision-making context, models should be able to quantify (and ideally reduce) the uncertainty of simulated predictions (Caers, 2011). This is typically done by assimilating measurements of field data (also known as observations), such as hydraulic head and streamflow rates, into the model in a process known as history-matching (Doherty and Simmons, 2013). In a review of the different types of observations frequently occurring in groundwater and surface-water modelling literature, Schilling et al. (2019) found that including at least one unconventional observation type is typically beneficial in terms of reducing predictive uncertainty. This is because classical observations can sometimes be poor in information pertaining to SW-GW exchange behavior (Schilling et al., 2019). Doherty and Moore (2020) recommend developers of decision support models to focus on the ability of a model to provide receptacles for decision critical information, rather than on its ability to simulate environmental processes. This can typically be achieved by adopting a highly parameterized approach to modelling (White et al., 2020). Wöhling et al. (2018) and Partington et al. (2020) constitute recent examples where highly parameterized models were used to assimilate unconventional observation types for assessing SW-GW exchange fluxes. Wöhling et al. (2018) found that integrating field observations with "soft" information in site-specific expert knowledge could enhance the plausibility of the calibrated model. Partington et al. (2020) examined the worth of classical and unconventional observation data (Radon-222, Carbon-14 and EC) in terms of reducing SW-GW exchange flux predictive uncertainty, and found Radon-222 and EC to be of particular value during low- and regular streamflow conditions.

Tetrachloroethylene (also known as perchloroethylene, henceforth referred to as PCE) is a chlorinated solvent (a volatile organic compound, VOC) primarily used in dry cleaning and metal degreasing and exposure is highly suspected to cause cancer in humans (Guha et al., 2012; Barul et al., 2017). Chlorinated solvents are denser than water, and are often referred to as dense non-aqueous phase liquids (DNAPLs) and a common groundwater contaminant that typically form large plumes (up to several kilometers in length) when dissolved in flowing groundwater (Pankow and Cherry, 1996). Methods for estimating mass flux and discharge of VOCs from groundwater to surface water typically rely on some variation of control plane (i.e., cross section multi-level sampling orthogonal to the direction of groundwater flow), where plume discharge is defined as the amount of contaminant mass migrating through the control plane per unit of time (Pankow and Cherry., 1996; Guilbeault et al., 2005; Chapman et al., 2007). In a recent study, Nickels et al. (2023) used point-scale streambed measurements of hydraulic parameters and VOC concentration to quantify VOC discharge from groundwater to surface water in high detail at small scale.

In this study, we develop a highly parameterized groundwater numerical model to characterize and assess the SW-GW exchange fluxes of an ecologically sensitive stream, affected by PCE-polluted groundwater outflowing from a nearby chlorinated solvent contaminated site. The aim is to locate and quantify the amount and seasonal variation of groundwater discharge that occur adjacent and downstream of the site. Using surface-water chemistry samples, we then calculate probabilistic estimates of PCE mass influx to the stream, thereby providing decision makers with suggestions for targeted remediation. In order to reduce and quantify predictive uncertainties, we assimilate a combination of classical and unconventional observation types, including FO-DTS thermal anomalies and site-specific knowledge during history-matching. To increase transparency and facilitate reproducibility, model development is performed and documented using open source tools and environments.

2 The Hagfors contaminated site

Hagfors is a town in Värmland Province, southwestern Sweden. South of the town center, an industrial scale dry-cleaning facility (Figure 1) was in operation from the 1970s to the early 1990s, providing dry-cleaning services for the Swedish Armed Forces (Nilsen and Jepsen, 2005; SEPA, 2007). During this period, a large but unknown amount (estimated to 50 tonnes or perhaps more) of PCE was spilled and leaked into the ground, forming at least two point sources (Nilsen, 2013). Because the former dry-cleaning facility (the site) was operating on behalf of the Swedish state, responsibility for remediating the contamination was first designated to the county administrative board and later transferred to the Geological Survey of Sweden (SGU).

The site is situated on Geijersholmsåsen, a glaciofluvial deposit superposing crystalline bedrock extending in the NE-SW direction. It mainly consists of sand and varies between 10 and 30 m in thickness (Gustafsson, 2017). Depth to the water table varies from approximately 12 m near the source zones, to less than 1 m south of the site where a ravine cuts through the sediment. The aquifer is considered unconfined in the study area and transitioning into partially confined near Lake Värmullen where silt and clay covers coarser sediment. Creek Örbäcken, approximately 4 m wide and half a meter deep, flows through a drainage canal around the site from the north to east, before flowing into the ravine south of the

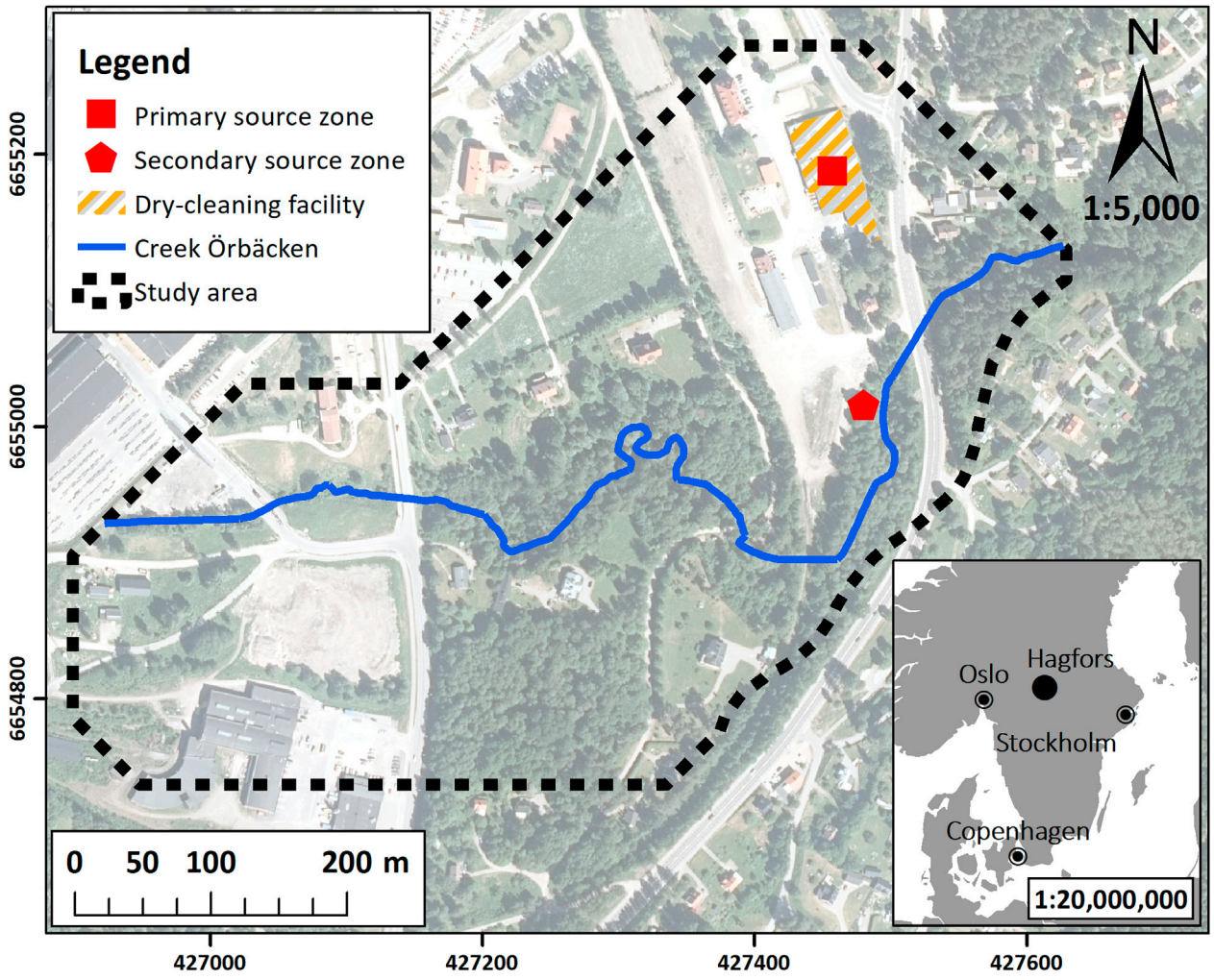

FIGURE 1
Map of the study area showing contamination source zone locations, creek Örbäcken and the extent of the model domain (units are in meters according to the Swedish national reference system, SWEREF99TM). The general direction of groundwater flow (and the direction of flow in the creek) is from the northeast towards the southwest.

site. Here, in the transition zone between two vertically stacked hydrogeological units, a natural degradation zone is located (Åkesson et al., 2021). The creek eventually feeds into Lake Värmullen *c.* one and a half kilometers west of the site. Earlier investigations have shown PCE concentrations exceeding Swedish drinking water guidelines values (of 10 µg/L, Swedish National Food Agency, 2001) in samples collected from the creek adjacent to the site and down towards the mouth of the lake (Nilsen, 2013).

Since the contaminant was discovered in the 1990s, multiple remediation campaigns of different scale have been undertaken. In 1996, the site was treated using soil vapor extraction, resulting in the removal of 1.5–2 tonnes of PCE from the primary source zone (Nilsen, 2003). Between 2003 and 2004, the site was treated using thermal remediation (steam injection treatment), resulting in the removal of an additional 5 tonnes of PCE from the primary source zone (Nilsen, 2003; Nilsen and Jepsen, 2005). Although large quantities had by then been removed, Nilsen (2013) estimated that there still remained between 20 and 30 tonnes of PCE in the primary source zone, and an additional 10 tonnes of PCE in the secondary source zone.

Creek Örbäcken (the creek) is the primary source of exposure to PCE for people in the area, as it flows through a sparsely populated area. It is also a conduit for rapid transport of PCE to Lake Värmullen. No drinking water wells are known within the area. In 2015, SGU changed the strategic objective from primary source zone treatment to mitigating influx of PCE to the creek (Larsson, 2020a). Yearly PCE mass influx to the creek has previously been estimated to 130 kg using control plane based calculation (Nilsen, 2013) and to 121 kg by computing the arithmetic mean of surface water concentrations multiplied by streamflow rates sampled and measured from December 2018 to February 2020 (Larsson, 2020b). In 2018 and 2019, *in situ* pilot nano zero-valent iron (nZVI) injection tests were performed in the plume emanating from the primary source zone, approximately 300 m southwest of its source (Larsson, 2021). The purpose was to evaluate the potential of a permeable reactive barrier solution for mitigating groundwater

influx to the stream. Unfortunately, results of the campaign indicated no reduction of PCE.

2.1 Previous modelling work of the site

Prior to the remediation efforts presented above, a number of site investigations were performed. As part of the investigative work, to date three different environmental models have been developed.

Andersson (2012), developed a three-dimensional steady-state model in order to 'study flow patterns within different parts of the groundwater reservoir and to get a better idea of flows and transport times to the surface water recipient' (creek Örbäcken). The model was developed using Visual Modflow 2011.1, a graphical user interface (GUI) to MODFLOW, and consisted of thirteen hydraulic conductivity (K) zones across four layers. The river (RIV) package was used to calculate SW-GW exchange fluxes in seven zones along the creek, and particle tracking was used to estimate advective transport times from both source zones to the creek. The results indicated a loss of *c.* 262 m^3 per day in groundwater recharge in the upstream section of the creek, and a gain of *c.* 879 m^3 per day in groundwater discharge in the downstream section. Transport times were estimated to between 250–400 days from the primary source zone and *c.* 60 days from the secondary source zone. The model was history matched using manual regularization (i.e., "trial and error") by means of adjusting K in the thirteen zones. However, at least five history-matching targets were omitted due to poor fit with field-data in locations of complex geology (Andersson, 2012). Andersson (2012) noted that the model suffered from numerical instability and suggested that a smaller model with higher resolution could improve the fit to data around the area of complex geology. Predictive uncertainties were not explored.

Havn (2018) developed two steady-state MODFLOW models of the site; a 'homogenous' and a 'heterogeneous' version, using the GMS 10.3 GUI. The reason for developing a homogeneous model was to 'understand the overall picture of the catchment' (Havn, 2018). To facilitate visualization in three dimensions, it was constructed using eight homogenous layers. The heterogeneous model consisted of sixteen layers and was developed to 'simulate and estimate pollution' from the site. It consisted of five adjustable parameters, including hydraulic conductivity (assigned on a layer-by-layer basis) and stream conductance. Both versions of the model were subject to history matching using two approaches; manual regularization and 'automated calibration' (Havn, 2018) using the PEST software. Both approaches, however, lead to large residuals (hydraulic head error exceeding 1.5 m) considering the size of the study area and density of available data. Nevertheless, a solute transport model was developed to run using results from the flow model. Havn (2018) concluded that the model was not able to quantify the scale of pollution and suggested that a higher model resolution could lead to improvements in model capability. A parameter sensitivity analysis was conducted, but predictive uncertainties were not explored.

Korsgaard (2018) developed a 2-dimensional steady-state model using the GUI Visual Modflow Premium 4.6 Classic. The numerical model was discretized as a 100-m-long cross section along the plume emanating from the secondary source zone, reaching across the creek. The primary purpose of the model was to test different remediation scenarios for reducing flux of contaminated groundwater into the creek, including "dig-and-dump" and "pump-and-treat". A secondary purpose was to estimate the daily volume of contaminated groundwater expected to be collected for remediation treatment. The model consisted of 26 layers with local refinement near the creek. The layers were divided into five K-zones subject to manual parameter adjustment. Seven remediation scenarios were evaluated using groundwater flow-, particle tracking and solute transport simulation. However, no history-matching was performed, and predictive uncertainties were not explored.

3 Model scope

To provide decision makers with information relevant to the current CSM-objective (mitigation of contaminated discharge to the creek), a numerical model was developed to explore SW-GW exchange behavior in the study area (Figure 1). After considering available observation data and computing power, a subjective decision was made to limit the studied period to between the years 2016–2020. To capture seasonal variability in SW-GW exchange fluxes, we choose to history-match field data and simulate SW-GW exchange fluxes under transient conditions. To provide decision makers with as much detail as the selected approach is capable of delivering, the prediction of interest is cell-by-cell SW-GW exchange fluxes on a weekly temporal resolution during the studied period. To increase data assimilation capability, and to reduce risk of numerical instability, we opt for a single-layer model designed around parametrical complexity rather than around structural complexity. This way, parametrical heterogeneity may form as needed, and the model run-time is kept low, which is desirable in a history-matching context (Doherty and Moore, 2020; Hugman and Doherty, 2021). To reduce and quantify predictive uncertainties, we leverage tools of the PEST (Doherty, 2020a) and PEST++ (White, 2018) software suites.

The model architecture and workflow is described in further detail below.

4 Materials and methods

The data used in this study was collected on site as well as downloaded from Swedish authority databases. Streamflow measurements, stream stage measurements, fiber-optic distributed temperature sensing (FO-DTS) and the bulk of hydraulic head measurements were collected by environmental consultant firms Nirás AB (Sebök, 2016; Larsson, 2017; Larsson, 2020b) and Sweco AB (Nilsen, 2013) and supplied to the MIRACHL research group on behalf of SGU. Complementary measurements were collected by the MIRACHL research group during two fieldwork campaigns in the springs of 2017 (see Åkesson et al., 2021) and 2019. Where needed, previously georeferenced data was converted to conform to the Swedish national reference system SWEREF99TM.

Preprocessing of data and model development was performed using the Jupyter Notebook interactive computing platform

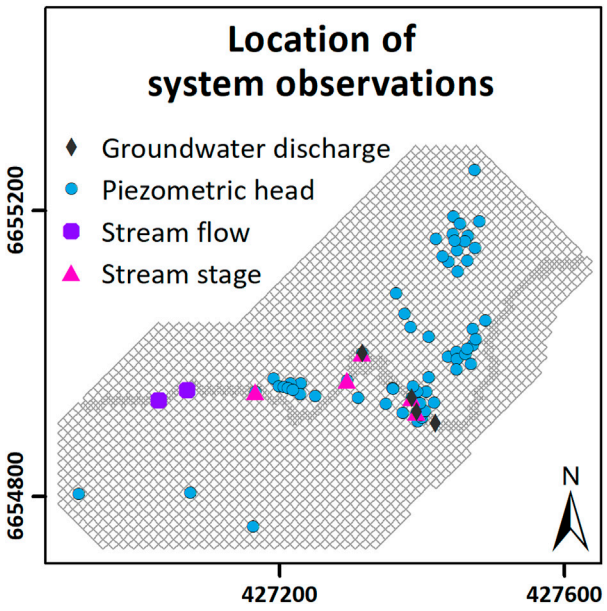

FIGURE 2
Map showing numerical model grid and type and location of history-matching targets. The model grid is refined near monitoring wells and the creek (units are in meters according to the Swedish national reference system, SWEREF99TM).

(Kluyver et al., 2016), following a recent example by White et al. (2020b) on facilitating model reproducibility. The notebooks, which include the work up until the point of history-matching, can be accessed from a Github repository (see Data Availability Statement).

The datasets used in the study are now presented below, followed by a description of the model architecture, model development and history-matching process.

4.1 Datasets and data preparation

Geospatial point cloud data of two types; Lidar and borehole logs, were used to define the spatial extent and topography of the upper and lower model boundary. Because the Lidar data (Lantmäteriet, 2016) was sampled using a relatively high resolution (2x2 m cell size), the dataset was curated to avoid propagating misleading altitudes at bridge crossings before being interpolated to the model grid.

Borehole data collected at the site (Nilsen, 2013; Larsson, 2017) with confirmed or assumed contact with crystalline bedrock, as well as regional borehole data downloaded from the SGU Wells Archive (SGU, 2015) was used to interpolate the extent of the lower model surface.

Daily precipitation data for the Gustavsfors A weather station, located approximately 15 km northeast of Hagfors, was downloaded from the SMHI Open Data Database (SMHI, 2021). Monthly computed evapotranspiration was downloaded for SMHI catchment area 64808 (eastern Hagfors) using the S-HYPE application (Strömqvist et al., 2012), and curated into mean daily evapotranspiration.

Hydraulic head measurements were collected from 63 single- and multilevel wells across the site using both piezometers and manual level meters (Larsson, 2020b). Measurements sampled using piezometers were typically recorded every fourth hour and was resampled into daily averages.

Stream stage, as well as the difference between groundwater head and stream stage (*head-stage* differences), were measured at five locations along the creek. Measurements were collected using a dual piezometer system where a primary piezometer was installed inside a monitoring well recording the groundwater level, and a secondary piezometer was installed on the outside of the monitoring well recording surface water hydraulic pressure (Larsson, 2017; Larsson, 2020b).

Streamflow was recorded in two different gages located approximately 40 m apart. Data collected with Gage-1 spans the full studied period (2016–2020). However, because streamflow recorded by Gage-1 was suspected to be affected by uncertainties inherent in the sampling methodology, a second gage (Gage-2) was installed in 2018 (Larsson, 2020b).

Fiber-optic distributed temperature sensing (FO-DTS) was used to measure surface water temperatures along three sections of the creek in December 2015 (Sebök, 2016). The discrepancy between surface water-groundwater temperatures were approximately 5 °C, and three warm-water anomalies were detected, indicating influx of groundwater at these positions.

Locations were hydraulic head, stream stage, streamflow and FO-DTS measurements were collected are shown on Figure 2.

4.2 Model architecture and development

The model employed in this study is a composite model (the model), consisting of preprocessing software, numerical solvers and postprocessing software open to the public domain. In addition, complementary preprocessing scripts were developed (see 4.2.1) in order to improve site-specific history-matching capability. Model settings, input files and scripts were written and prepared using the Jupyter Notebook environment.

Groundwater flow is simulated with MODFLOW6 (MF6) (Langevin et al., 2022). The MF6 model has a single layer with local refinement around streams and monitoring wells. The model has two stress periods. The first stress period is a steady-state period implemented to acquire representative heads, stream stages and streamflow rates for the beginning of the second stress period; a transient period ranging from December 2015 to December 2019. General head boundaries (GHB) are placed along the boundary of the model domain representing inflows (NW), outflows (SW) and lateral boundaries (SE and NW) of the glaciofluvial aquifer (Figure 3). The Streamflow Routing (SFR) package of MF6 was used to simulate streamflow, stream stage and surface water-groundwater exchange flux in the creek. Setup and configuration of MF6 and its input packages was performed using the python package Flopy (Bakker et al., 2016).

Five instances of the lumped parameter recharge model, LUMPREM2 (Doherty J., 2021), were prepared using the python package Lumpyrem (Hugman, 2021), based on daily rainfall and evapotranspiration data presented above (4.1). One instance was used to compute groundwater recharge for use as input by the

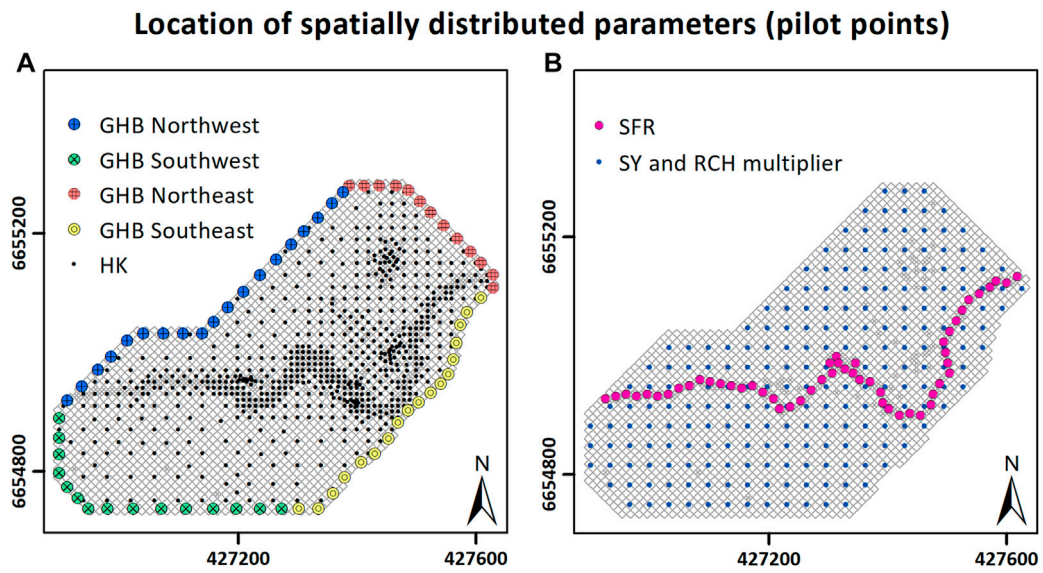

FIGURE 3
Maps showing pilot-point locations of **(A)** the general head boundaries (GHB) and hydraulic conductivity (HK) and **(B)**, parameters associated with the streamflow routing (SFR) package, specific yield (SY) and groundwater recharge (RCH) multiplier. Units are in meters according to the Swedish national reference system, SWEREF99TM.

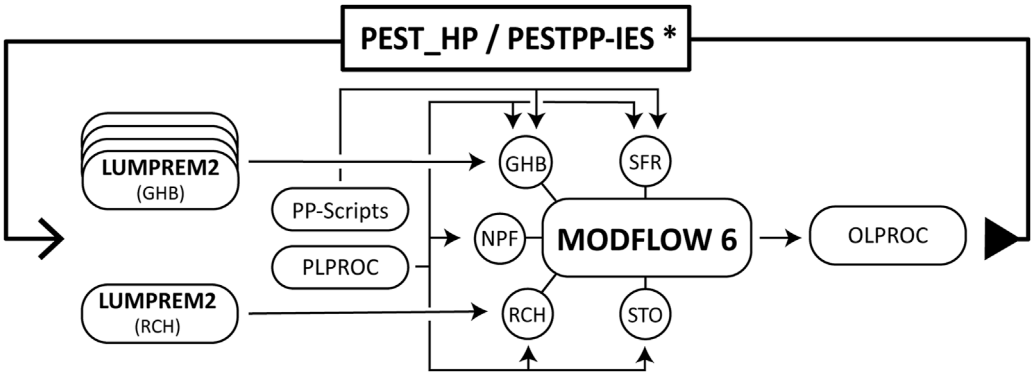

FIGURE 4
Flowchart showing the composite model architecture and flow of information during history-matching (* and uncertainty quantification using PESTPP-IES). The acronyms GHB, SFR, NPF, RCH and STO refer to the MF6 packages General-Head Boundary, Streamflow Routing, Node Property Flow, Recharge, and Storage.

MF6 Recharge package (RCH). The remaining instances were used to compute time-varying boundary head elevation to the GHBs.

The model architecture is illustrated in Figure 4.

4.2.1 Model parameterization

The model is parameterized using 1709 adjustable parameter values. Pilot points were used to allow spatial variation of physical parameters (Figure 3), including hydraulic conductivity, specific yield, boundary conductance, and spatial variables of the SFR package representing the creek. Hydraulic conductivity pilot points were placed with higher density near the creek and around monitoring wells where the model grid is refined. Covariance matrices taking pilot point density into consideration were created using the PPCOV_SVA and MKPPSTAT utilities (Doherty, 2020b) of the PEST suite. They were applied to constrain parameter covariance and encourage PEST to spread parameter heterogeneity. A temporal covariance matrix was also created for constraining upstream inflow into the starting cell of the SFR package.

During history-matching, writing of parameter values to model input files was done using the model preprocessing software PLPROC (Doherty, 2021c) and three scripts written in the

Python language. The three scripts (PP-Scripts in Figure 4) were developed to complement functionality difficult to implement through PLPROC for writing parameter values to the SFR and GHB package of MF6.

4.2.2 Observation targets and feature engineering

History-matching targets include measurements of hydraulic head, streamflow and stream stage collected during the studied period. OLPROC (Doherty, 2021b) was used to time-interpolate model outputs to field measurement time. In addition, OLPROC was also used to feature-engineer existing datasets into datasets of temporal measurement differences for use as observations. Inequality observations (also known as "one-way observations") (Doherty, 2020a; White et al., 2020a) were used in this study to inform PEST of groundwater influx into the creek at four locations indicated by thermal anomalies in FO-DTS data. Inequality observations were also used to inform PEST that all cells belonging to the SFR-package (which is used to represent the creek) should have an outflow between each cell and its downstream neighbor cell (i.e., the creek should never dry out).

In total 172,059 observations, divided into 15 groups, were used as history-matching targets. Weights were assigned with equal importance to each type of observation during calibration with PEST_HP (Doherty, 2020a). For uncertainty quantification with PESTPP-IES (White, 2018), realizations of measurement noise were generated by changing observation weights to reflect the inverse of the standard deviation of measurement noise for each observation group.

4.2.3 History-matching and uncertainty quantification

History-matching was performed in a three-stage process. First, a standalone instance of the LUMPREM2 model was matched against historical groundwater measurements in a single monitoring well (NI15-O48) using PEST (Doherty, 2018) with Tikhonov (preferred value) regularization. The parameter values that emerged through this process was selected as the initial parameter values of the five LUMPREM2 instances used in the composite model. Secondly, the composite model was history-matched using PEST_HP on Lunarc Aurora, Lund University's high performance computing (HPC) cluster. After eight iterations PEST_HP was terminated using early stopping to reduce risk of overfitting. Finally, the model was redeployed to the HPC to undergo history-matching and uncertainty quantification using PESTPP-IES. The available computing power allowed for a large ensemble (500 realizations) to be generated and used in the uncertainty quantification process. After five iterations no further meaningful reduction of the objective function was recorded. To reduce risk of underestimating predictive uncertainty, model output obtained from the third iteration of history-matching and uncertainty quantification are presented as the results of this study.

4.3 Estimating groundwater PCE discharge

Because the model is instructed to compute streamflow in addition to SW-GW exchange fluxes, we can use measurements of surface water chemistry collected during the studied period to provide estimations of contaminant mass (PCE) influx. In order to estimate the PCE mass influx required for a given sample, we make the following assumptions; 1) groundwater discharge in the study area is the only source of measured PCE in the creek, 2) the difference in streamflow at the sample location between the sample date and model output date is negligible (the temporal discrepancy between model output dates and sample dates vary between 0–4 days), 3) there is no intra-day variability in PCE concentration at the sample location (i.e., the measured concentration is representative for the full sample date), and 4) the sampled surface water in a cell, from which we obtain measurements of concentration, will be composed of an unknown ratio of groundwater discharge to surface water from upstream of the sample location.

Streamflow through a model cell representing a sample location in the creek can be described as:

$$Q_T = Q_{gw} + Q_{sw}$$

where Q_T is the total streamflow, Q_{gw} is streamflow fed by groundwater discharge and Q_{sw} is non-groundwater fed streamflow (all flows are in [m^3/d]). Upstream groundwater discharge contributing to the total streamflow through a cell can be calculated as:

$$Q_{gw} = \sum_{i=0}^{n} Q_{gw}^{i} \left[Q_{gw}^{i} > 0 \right]$$

where n is the number of upstream cells and Q_{gw}^{i} is groundwater discharge [m^3/d] in upstream cells with positive discharge. The index begins at zero to include discharge occurring in the cell representing the sample location. The ratio of groundwater discharge to total streamflow is given by:

$$Q_{gw}^{r} = \frac{Q_{gw}}{Q_T}$$

As groundwater is discharged into the creek, the groundwater PCE concentrations are diluted by surface water. Using the ratio of groundwater discharge to total streamflow (Q_{gw}^{r}), we can infer the concentration of PCE in upstream groundwater discharge required for a given surface water sample:

$$C_{gw} = \frac{C_M}{Q_{gw}^{r}}$$

where C_{gw} is the concentration of the upstream groundwater discharge and C_M is the measured concentration of the water sample (concentrations are in [kg/m^3]). Using the concentration of the upstream groundwater discharge (C_{gw}), we can infer the mass of PCE [kg] discharged into the stream for a given surface water sample:

$$PCE_m = Q_{gw} \cdot C_{gw}$$

By calculating PCE_m for each member of the model ensemble, the prior and posterior uncertainty in streamflow and SW-GW exchange fluxes is taken into consideration and a probabilistic estimation of the PCE influx is provided. There are 631 surface water chemistry samples collected from different locations in the creek during the studied period for which a PCE-influx estimation is provided.

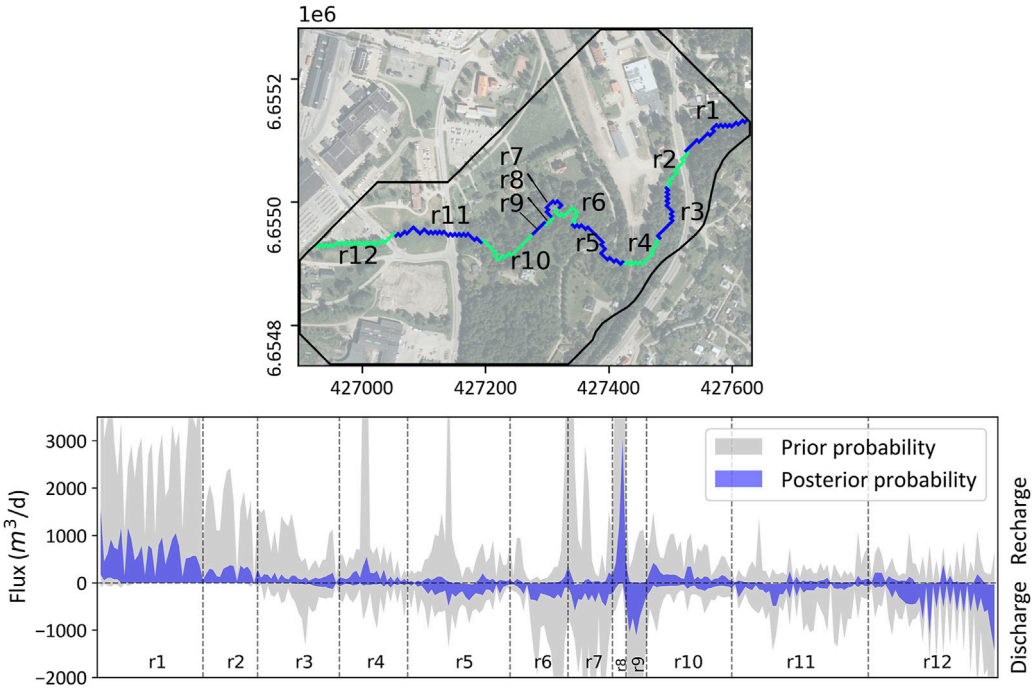

FIGURE 5
Prior and posterior SW-GW exchange flux uncertainty. The creek is divided into reaches (segments) based on SW-GW exchange behavior. Map units are in meters according to the Swedish national reference system, SWEREF99TM.

5 Results

5.1 Simulated SW-GW exchange fluxes

The model was instructed to calculate SW-GW exchange fluxes once per week during the studied period (2016–2020). Of the 500 initial realizations, 485 realizations resulted in convergence. History-matching reduced uncertainties in simulated SW-GW exchange fluxes in all sections of the creek. Upon inspecting *a posteriori* model results, we have divided the creek into twelve reaches (segments), based on their SW-GW exchange behavior, as shown in Figure 5. Predictive uncertainties remain fairly high in the first reach, but decrease significantly in the second reach, which is located adjacent to monitoring wells from which data was included as history-matching targets. Predictive uncertainties also remain fairly high in reaches eight and nine, which represents the final part of a small meander bend.

In general, the creek is contributing to groundwater recharge in the first four reaches. Mean simulated recharge in this section is calculated to *c*. 7204 m^3/d but is associated with a considerable variability during the studied period ($\sigma \approx 3070$ m3/d). Reaches five through seven represent three segments where groundwater is discharged to the creek. Mean simulated discharge in this section is calculated to c. 3102 m3/d ($\sigma \approx 1679$ m3/d). In reaches seven through nine, which represent a relatively small meandering section of the creek, the posterior uncertainty remain relatively high. Considered at the mean of the posterior ensemble, this section contribute with a slight mean groundwater discharge of c. 224 m3/d ($\sigma \approx 608$ m3/d). Simulated SW-GW exchange behavior in reach ten is considered neutral with a very small mean groundwater recharge of c. 5 m3/d ($\sigma \approx 427$ m3/d). The final two reaches, reach eleven and twelve, contribute with a mean groundwater discharge of c. 5328 m3/d ($\sigma \approx 2574$ m^3/d).

Temporal variability in simulated SW-GW exchange fluxes is shown in Figure 6. Locations where groundwater is recharged from the creek is described as losing conditions, and *vice versa*. As shown, temporal variability in SW-GW exchange behavior remain fairly static for the studied period. The most pronounced variability can be observed during the months of April and May of 2018 and 2019 in reaches five through twelve, indicating less discharge to the creek compared to the same period of the two preceding years. This is also, in general, the periods where predictive uncertainties pertaining to temporal variability are the highest.

5.2 PCE mass influx estimation

Using equations 1 to 4, measured concentrations of surface-water PCE was used to infer the groundwater discharge PCE concentrations, for each of the 631 samples. Eq. (5) was then used to compute PCE mass influx by multiplying simulated groundwater discharge with inferred groundwater PCE concentrations. This was done for each member of the model ensembles, resulting in 293,789 computations of prior and posterior daily PCE mass influx estimations respectively. The results are shown in Figure 7, and is color coded by surface water

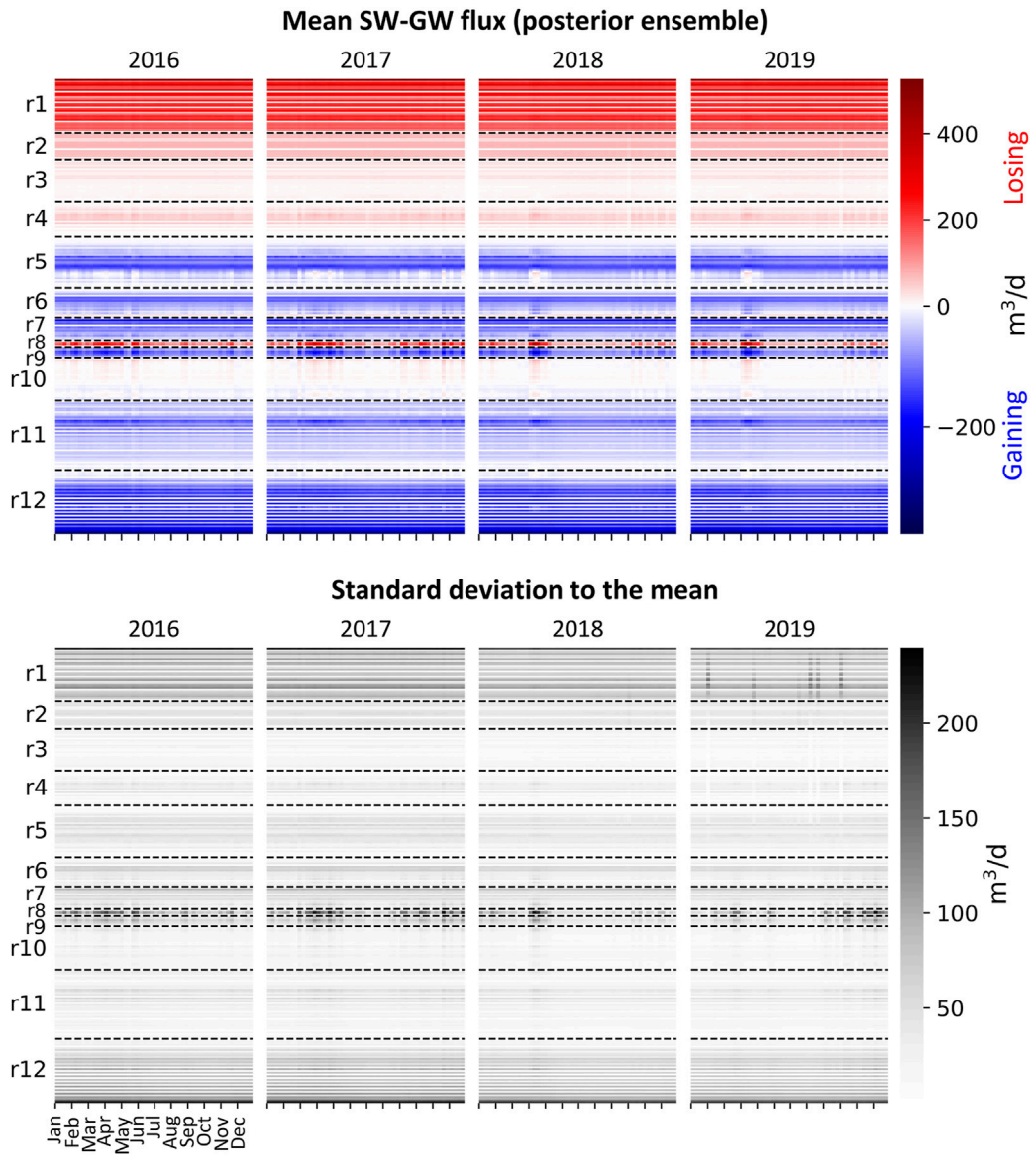

FIGURE 6
Heatmaps showing temporal variability in SW-GW exchange flux during the studied period in weekly output. The upper row (colorized) show mean of the posterior ensemble and the lower (grayscale) row show posterior ensemble standard deviation of the mean.

flow regime, which we categorized as low flow (<25th percentile), regular flow (between 25th and 75th percentile) and high flow (>75th percentile). In general, computed upstream PCE influx increase in the downstream direction and is highest during periods of high flow, which can be observed in the upper plot of Figure 7. Locations of influx (groundwater discharge) for the three flow regimes, and their respective uncertainty, is shown on the bottom plot. The computed upstream PCE influx follow a log-normal distribution, with a $\log_{10}$ geometric mean of *c.* 0.55 kg/d. Uncertainties in computed mass influx are described using the 5th and 95th percentiles and are based on posterior uncertainties in SW-GW exchange fluxes and streamflow, as well as laboratory measurement uncertainties pertaining to the chemistry samples.

6 Discussion

6.1 Model workflow challenges and opportunities

Challenges arising during construction of the model were mainly associated with the SFR package of MF6. Implementing the SFR package in a history-matching context require extra careful consideration in comparison to many of the commonly used MODFLOW packages. This is because SFR does not allow a parameter known as *reach streambed top elevation* (*rtp*) to increase in the downstream direction. If this requirement is not met MODFLOW will return an error and history-matching will be terminated prematurely. In our case, which will likely also be the

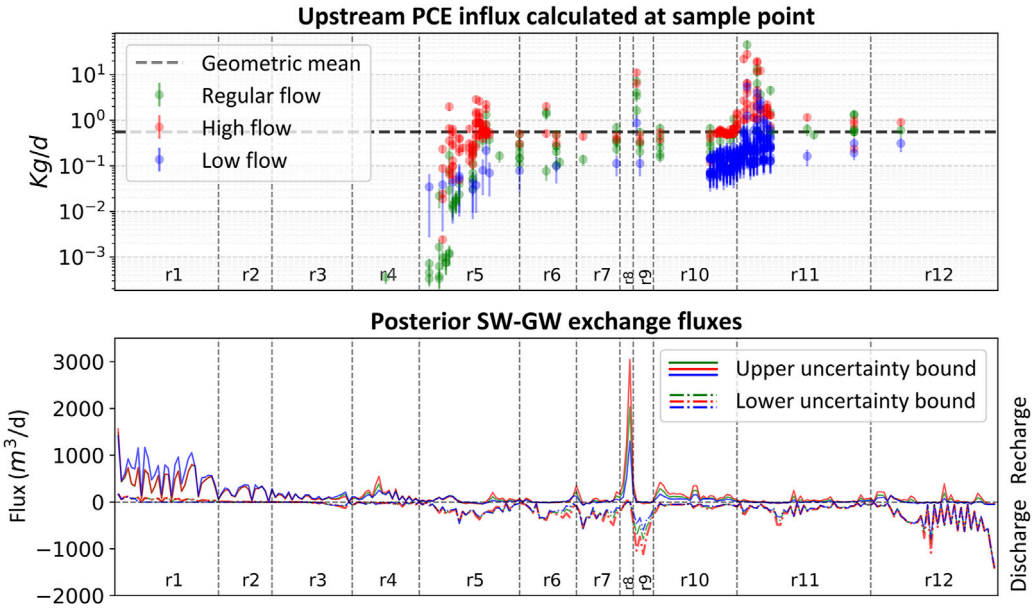

FIGURE 7
Upper plot showing estimated PCE mass influx per sample under regular, high and low streamflow. The geometric mean is 0.55 kg/d. High peaks in mass influx tend to occur in areas characterized by groundwater discharge. Bottom plot showing upper and lower uncertainty bounds of posterior SW-GW exchange fluxes under different flow regimes. Notable differences in predictive uncertainty can be observed in reaches one, eight and nine.

case for many others utilizing the SFR package, initial *rtp* values were obtained by sampling a digital elevation map (DEM). However, undulating topography in the DEM yielded invalid input for a portion of the SFR cells. Leaf et al. (2021) created SFRmaker, a Python package designed to automate the workflow of implementing the SFR package and curate valid input. However, at the time of writing this paper, SFRmaker only supports structured grids. Because the model in this study utilizes an unstructured grid, we implemented a solution inspired by SFRmaker to ensure that the requirement described above is met:

$$|rtp_i| = \begin{cases} x_i, & if\ i = 1 \\ x_i, & if\ x_i < x_{i-1} \\ x_{i-1}, & if\ x_i \geq x_{i-1} \end{cases}$$

where x_i is the sampled elevation at the center point of the i-th cell of the SFR package.

Another potential issue related to the SFR package in the context of history-matching is the drying out of streamflow cells. During early iterations, we discovered that PEST_HP sought solutions of minimum error variance that included dry streamflow cells for small parts of the creek. Because we know from extensive site investigation, as well as from measured data, that the creek does not dry out, this presented a problem. In order to address this issue, we implemented inequality observations to instruct PEST_HP and PESTPP-IES to seek solutions where the streamflow between a cell and its downstream neighbor was positive.

Fiber-optic distributed temperature sensing (FO-DTS) data presents an interesting opportunity in terms of data assimilation. To the best of our knowledge, this is the first time FO-DTS thermal anomalies are used in the context of history-matching. In order for groundwater discharge to be detected as a thermal anomaly in FO-DTS data, groundwater and surface water must be of different temperatures. Because of this, usability of FO-DTS may vary according to site and season. In addition, length of the FO cable presents a practical constraint on its usability, making it better suited for use in models representing smaller sites. In this study, thermal anomalies were implemented in the form of inequality observations of groundwater discharge, meaning that PEST considers the residual of an observation as zero when discharge is greater than 0 m^3/d. Future work where thermal anomalies are assimilated during history-matching could explore the use of less conservative inequality constraints, or, even the use of thermal anomalies as regular, numerical observations. Longer time series would also facilitate studies on data worth under a variety of predictions where SW-GW interaction play a role.

6.2 Simulated SW-GW exchange fluxes and estimations of PCE mass influx

Predictive uncertainties pertaining to SW-GW exchange fluxes were reduced significantly as a result of history-matching, enabling high resolution spatiotemporal characterization as shown in Figure 6. Predictive uncertainties remain relatively high in the first and last (12th) reach, as well as in reach eight and nine, located centrally in the study area. Uncertainties pertaining to the first and last reach can likely be explained by the absence of history-matching targets in these sections of the creek. Reaches eight and nine represent the final part of a small meander bend. As shown in the bottom plot of Figure 7, uncertainty in this section is particularly sensitive to variability in streamflow compared to other sections of the creek. One possible cause for this could be

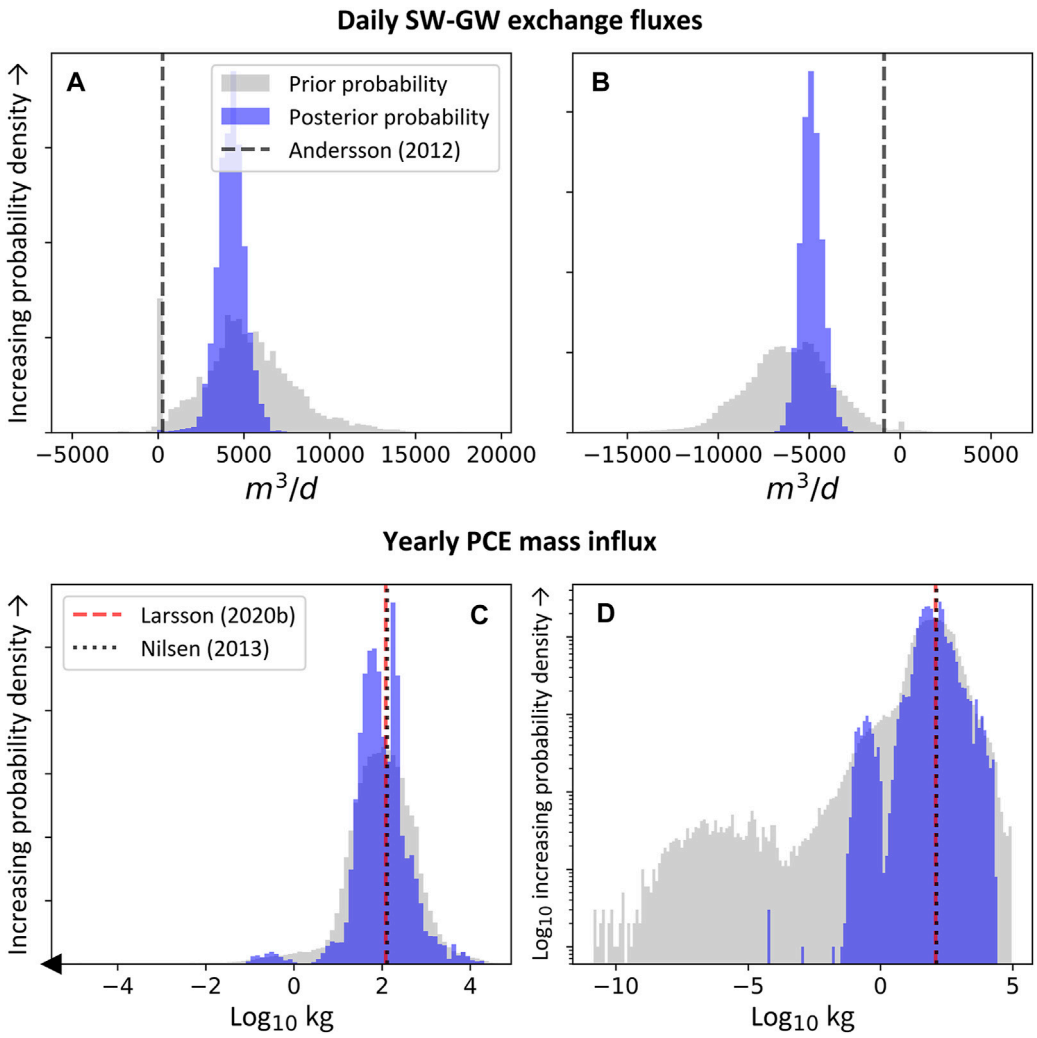

FIGURE 8
Comparisons between results in this study (represented as prior and posterior probability distributions in gray and blue) and results obtained in earlier studies of the Hagfors contaminated site. Upper row showing comparisons with results in Andersson (2012) of **(A)** simulated SW-GW exchange flux of the upstream section of the creek characterized by recharge (approximately corresponding to reaches two to four), and **(B)** SW-GW exchange flux of reaches five to eleven characterized by discharge. Bottom row showing comparisons with previous estimates of PCE mass influx where **(C)** show yearly PCE mass influx using a truncated *x*-axis to enhance visibility of the posterior probability distribution, and **(D)** yearly PCE mass influx with a logarithmic *y*-axis to highlight reduction of predictive uncertainty achieved during history-matching.

the model's inability to allow for temporal variability of spatial parameters pertaining to the creek. For example, during times of heavy rains and melting snow (conditions that cause high streamflow rates), the stream width is expected to widen as water levels rise and progressively cover the point bar deposits. Although stream width is considered adjustable parameters in this study, it is not configured to allow for temporal variability.

Seasonal variability in SW-GW exchange fluxes is relatively low, with the exception of reaches five through eleven in April and May of 2018 and 2019 (Figure 6). In 2018, northern Europe was affected by an extreme drought that persisted into 2019 (Bakke et al., 2020). This is a plausible explanation for the anomalous behavior in SW-GW exchange fluxes observed in the creek during this period, which indicate greater groundwater recharge and lesser groundwater discharge than normal.

As shown in the upper plot of Figure 7, uncertainty in PCE mass influx is greater during times of low flow. This is because posterior uncertainties in streamflow are greater during periods of low flow (mean coefficient of variation, mCV ≈0.23), compared to periods of regular flow (mCV ≈0.15) or high flow (mCV ≈0.08).

The estimations of PCE mass influx were based on a list of assumptions (4.3). Although we can be fairly certain that the first assumption holds true (no alternative sources to measured PCE other than groundwater discharge in the studied area), the second and third assumptions require discussion. The second assumption, namely, that difference in streamflow between sample date and model output date (varying between zero to 4 days) is negligible is a simplified assumption, as, for example, bursts of heavy rainfall may momentarily impact streamflow rates and levels. The third assumption, that there is no intra-day variability in PCE

concentration at the sample locations, also represent a simplified assumption. PCE is a dense non-aqueous phase liquid (DNAPL), a hydrophobic compound known to generate extensive variability in mass discharge (Guilbeault et al., 2005) not only temporally, but also spatially. Larsson (2020b) suggest that incomplete mixing of the surface water in the creek may lead to great variability in measured concentrations depending on whether the sample was collected centrally or near the banks. Both assumptions discussed above are associated with uncertainties that are unquantified pertaining to the mass influx estimations. However, because the samples on which the estimations are based, were collected during periods of varying streamflow, as well as spatially varying along both axes of the creek, it could be argued that the estimations of mass influx, when looked at as a distribution, take this uncertainty somewhat into consideration.

6.3 Earlier studies and new findings

As discussed earlier (2.1), three models were previously developed of this site. Because particle tracking and solute transport modelling was outside the scope of this study, direct comparisons with Havn (2018) and Korsgaard (2018) are difficult to make at this point.

Andersson (2012) discretized the creek into seven zones and calculated total SW-GW exchange fluxes for each zone. Anderson (2012) found that groundwater recharge was occurring in the first zone, and discharge was occurring in the remaining six zones. As we have shown (Figures 5–7), SW-GW exchange behavior in the creek is complex, especially in the meandering sections where spatiotemporal variations can be expected to be large. Therefore, we cannot exclude the possibility of groundwater recharge occurring further downstream if we are to consider the distribution of posterior uncertainty. By selecting a subsection of the creek overlapping the two studies, we found the section of upstream recharge to be significantly longer (*c.* 28 percent), and more importantly, that SW-GW exchange fluxes were significantly larger than previously estimated (Figure 8, upper row).

Unfortunately, the model was not equally successful in reducing predictive uncertainties pertaining to PCE mass influx. By multiplying results obtained through Equation (5) by 365, we can compare our estimates with prior estimations (Nilsen, 2013; Larsson, 2020b) of mass influx (Figure 8, bottom row) per year. As shown, both prior estimations are located near the center of the posterior probability distribution. Interestingly, we find two peaks in the posterior probability distribution. To determine whether the bimodal distribution was caused by the model or the chemistry input dataset, we tracked each realization (thereby also tracking each set of model parameters) contributing to the results of the left and right peak respectively. However, we found no meaningful variability in model contribution between the two peaks (i.e., all realizations contributed to both peaks). Measured concentrations in the creek, however, appear with a bimodal distribution, and is therefore the only logical contributing factor for the shape of the mass influx estimations as shown on Figure 8. As Larsson (2020b) suggested, incomplete mixing of surface water is a likely explanation for why the surface water chemistry dataset appear bimodal.

Uncertainty in PCE mass influx, as estimated in this study, can originate from uncertainty in streamflow, uncertainty in SW-GW exchange fluxes and in uncertainties related to surface water chemistry measurements. Uncertainties pertaining to the first two origins were reduced significantly during history-matching, but uncertainties related to surface water chemistry persist. Nevertheless, an uncertainty range has been quantified.

7 Conclusion

With the current CSM objective being set on mitigating influx of PCE to creek Örbäcken, a prediction relevant to the objective (characterization of SW-GW exchange fluxes) was selected for this study. The ensuing model workflow and architecture was designed to facilitate data assimilation of prediction pertinent information in historical measurements through history-matching. By adopting a single-layer approach with local refinement near the creek and around monitoring wells for which historical measurements were available, the model run time could be constrained. This was important, because history-matching requires many model runs. We used a highly parameterized model, which allowed for parametrical heterogeneity to evolve where needed during history-matching. In addition to classical types of observations, we also assimilated thermal anomalies in FO-DTS measurements as locations of groundwater discharge, through the use of inequality observations. Challenges pertaining to implementing the SFR package in a history-matching context and suggestions for how to overcome them was discussed. Predictive uncertainties were reduced and explored using the iterative ensemble smoother of the PEST++ suite.

As a result, we were able to characterize SW-GW exchange fluxes in the creek in high spatiotemporal resolution, showing locations of (and quantifying) contaminated groundwater discharge. Seasonal variability pertaining to SW-GW exchange fluxes was found to be low, with the exception of an unusual drought event that occurred during 2018–2019. We also found that SW-GW exchange fluxes are likely to be significantly larger than previously estimated. Using surface water chemistry measurements, we estimated PCE mass influx and found estimations in two earlier studies to be located near the center of the posterior probability distribution. The uncertainty pertaining to PCE mass influx was only reduced slightly, but has now been quantified.

Our recommendation for decision makers, with regards to the current CSM objective, is to focus remediation action toward reaches 5–7, 9, 11 and 12, according to modelling results.

Data availability statement

Modeling workflow up until the point of history-matching is documented in a collection of Jupyter Notebooks. They can be accessed from the following GitHub repository: https://github.com/nikobenho/hagfors_gwm. Requests to access the history-matched model should be directed to nikolas.hoglund@geol.lu.se.

Author contributions

NH, CS, and RH conceived the idea. NH analysed the data. NH (with support and guidance from RH) set up the modelling environment and undertook all simulations and their postprocessing. NH and CS wrote the paper. RH reviewed the manuscript, provided critical feedback and helped shape the research. All authors contributed to the article and approved the submitted version.

Funding

Funding for the work was provided by Formas, The Swedish Research Council for Environment (ref. 2016-20099 and 2016-00808), SBUF (ref. 13336), ÅForsk (ref. 14-332), SGU, NCC, and Sven Tyréns Stiftelse and Lund University.

Acknowledgments

Ulf Winnberg and Kristin Forsberg at SGU is thanked for fruitful discussion concerning site strategy and encouragement. Nicklas Larsson and Gro Lilbæk at Nirás AB is acknowledged for providing data crucial for the implementation and realization of this study. Professor John Doherty (Flinders University, Watermark Numerical Computing) is thanked for providing the Linux version of the PEST suite, as well as for providing compilation instructions. The data handling/computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Aurora and LU Local partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Marcos Acebes at SNIC support is acknowledged for assistance concerning technical and implementation aspects.

Conflict of interest

Author RH was employed by INTERA.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Åkesson, S., Sparrenbom, C. J., Paul, C. J., Jansson, R., and Holmstrand, H. (2021). Characterizing natural degradation of tetrachloroethene (PCE) using a multidisciplinary approach. *Ambio* 50. doi:10.1007/s13280-020-01418-5

Anderson, M. P. (2005). Heat as a ground water tracer. *Groundwater* 43, 951–968. doi:10.1111/j.1745-6584.2005.00052.x

Andersson, S. (2012). *Modellering av grundvattenflöden vid f.d Hagforstvätten med Visual MODFLOW (in Swedish)*. Report No: 133.1178.000. Karlstad, Sweden: Sweco Environment AB.

Bakke, S. J., Ionita, M., and Tallaksen, L, M. (2020). The 2018 northern European hydrological drought and its drivers in a historical perspective. *Hydrology Earth Syst. Sci.* 24, 5621–5653. doi:10.5194/hess-2020-239

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting MODFLOW model development using Python and FloPy. *Groundwater* 54, 733–739. doi:10.1111/gwat.12413

Barul, C., Fayossé, A., Carton, M., Pilorget, C., Woronoff, A-S., Stücker, I., et al. (2017). Occupational exposure to chlorinated solvents and risk of head and neck cancer in men: A population-based case-control study in France. *Environ. Health* 16, 77. doi:10.1186/s12940-017-0286-5

Briggs, M. A., Lautz, L. K., and McKenzie, J. M. (2012). A comparison of fibre-optic distributed temperature sensing to traditional methods of evaluating groundwater inflow to streams. *Hydrol. Process.* 26, 1277–1290. doi:10.1002/hyp.8200

Caers, Jef. (2011). *Modeling uncertainty in the Earth sciences*. Chichester, UK: John Wiley and Sons.

Chapman, S. W., Parker, B. L., Cherry, J. A., Aravena, R., and Hunkeler, D. (2007). Groundwater–surface water interaction and its role on TCE groundwater plume attenuation. *J. Contam. Hydrology* 91, 203–232. doi:10.1016/j.jconhyd.2006.10.006

Cook, P. G. (2013). Estimating groundwater discharge to rivers from river chemistry surveys: Groundwater discharge to rivers. *Hydrol. Process.* 27, 3694–3707. doi:10.1002/hyp.9493

Doherty, J. E. (2021c). *PLPROC – a parameter list processor*. Brisbane: Watermark Numerical Computing.

Doherty, J. E. (2020b). *Groundwater data utilities – Part B: Program descriptions*. Brisbane: Watermark Numerical Computing.

Doherty, J. E. (2021b). *OLPROC – an observation list processor for use with PEST and PEST++*. Brisbane: Watermark Numerical Computing.

Doherty, J. E. (2018). *Pest – model-independent parameter estimation – user manual Part 1: PEST, SENSAN and global optimisers*. Brisbane: Watermark Numerical Computing.

Doherty, J. E. (2020a). *PEST_HP – PEST for highly parallelized computing environments*. Brisbane: Watermark Numerical Computing.

Doherty, J., and Simmons, C. T. (2013). Groundwater modelling in decision support: Reflections on a unified conceptual framework. *Hydrogeol. J.* 21, 1531–1537. doi:10.1007/s10040-013-1027-7

Doherty, J. (2021a). *Version 2 of the LUMPREM groundwater recharge model*. Brisbane: Watermark Numerical Computing.

Guha, N., Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., et al. (2012). Carcinogenicity of trichloroethylene, tetrachloroethylene, some other chlorinated solvents, and their metabolites. *Lancet Oncol.* 13, 1192–1193. doi:10.1016/s1470-2045(12)70485-0

Guilbeault, M. A., Parker, L. B., and Cherry, J. A. (2005). Mass and flux distributions from DNAPL zones in sandy aquifers. *Groundwater* 43, 70–86. doi:10.1111/j.1745-6584.2005.tb02287.x

Gustafsson, M. (2017). *Grundvattenmagasinet Hagfors (in Swedish)*. Report No. 250 500 029. Uppsala, Sweden: Sveriges Geologiska Undersökning SGU.

Havn, H. H. (2018). Master's thesis. Copenhagen, Denmark: Copenhagen University. Groundwater flow and transport of chlorinated solvents at Hagfors, Sweden: Modelling historic impact and possible remediation of pollution of Örbäcken Creek

Hugman, R. (2021). Lumpyrem. Available at: https://github.com/rhugman/lumpyrem.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., et al. (2016). "Jupyter notebooks – A publishing format for reproducible computational workflows," in *Positioning and power in academic publishing: Players, agents and agendas*. Editors Fernando Loizides and Birgit Scmidt (Göttingen, Germany: IOS Press).

Korsgaard, A. (2018). Hagforstvätten. Uppställning av och beräkning med grundvattenmodell (in Swedish). Report No. 215147-22. Malmö, Sweden: NIRÅS.

Langevin, C. D., Hughes, J. D., Banta, E. R., Provost, A. M., and Panday, S. (2022). *MODFLOW 6 modular hydrologic model*. Reston, Virginia, USA: United States Geological Survey.

Lantmäteriet (Swedish mapping, cadastral and land registration authority) (2016). GSD-Höjddata, 608 grid 2+. Available at: https://maps.slu.se/get/ (Accessed May 21, 2021).

Larsson, N. (2020a). *F.d. FFV tvätteri Hagfors – handlingsplan*. Malmö, Sweden: NIRÀS.

Larsson, N. (2021). *Hagforstvätten Pilotförsök – resultat av miljökontroll vid permeabel reaktiv barriär i utströmningsområde 3 (in Swedish). Project No. 32400316-010*. Malmö, Sweden: NIRÀS.

Larsson, N. (2020b). *Hagforstvätten – justering av miljökontrollprogram för Örbäcken*. Malmö, Sweden: NIRÀS.

Larsson, N. (2017). Project No. 5000515. Sweden: Malmö: NIRÁS. Hagforstvätten – resultatrapport från undersökningar dec 2013 – jan 2017 (in Swedish)

Leaf, A. T., Fienen, M. N., and Reeves, H. W. (2021). SFRmaker and linesink-maker: Rapid construction of streamflow routing networks from hydrography data. *Groundwater* 59, 761–771. doi:10.1111/gwat.13095

Nickels, J. L., Genereux, D. P., and Knappe, D. R. U. (2023). Improved Darcian streambed measurements to quantify flux and mass discharge of volatile organic compounds from a contaminated aquifer to an urban stream. *Journal of Contaminant Hydrology* 253, 104124. doi:10.1016/j.jconhyd.2022.104124

Nilsen, J. (2013). Hagforstvätten, huvudstudie (in Sweden). Report No. 133.1178 000. Karlstad, Sweden: Sweco Environment AB

Nilsen, J. (2003). Hagforstvätten. Lägesrapport kall sanering av perkloretylen (in Swedish). Report No. 154.4157 000. Karlstad, Sweden: Sweco VBB AB.

Nilsen, J., and Jepsen, J. D. (2005). Hagforstvätten. Termisk insitu-sanering av perkloretylen, avslutande rapport (in Swedish). Report No. 233.4157 000. Karlstad, Sweden: Sweco VBB AB.

Ntona, M. M., Busico, G., Mastrocicco, M., and Kazakis, N. (2022). Modeling groundwater and surface water interaction: An overview of current status and future challenges. *Sci. Total Environ.* 846, 157355. doi:10.1016/j.scitotenv.2022.157355

Pankow, J. F., and Cherry, J. A. (1996). *Dense chlorinated solvents and other DNAPLs in groundwater: History, behavior, and remediation*. Portland, Oregon: Waterloo Press.

Partington, D., Knowling, M. J., Simmons, C. T., Cook, P. G., Xie, Y., Iwanaga, T., et al. (2020). Worth of hydraulic and water chemistry observation data in terms of the reliability of surface water-groundwater exchange flux predictions under varied flow conditions. *J. Hydrology* 590, 125441. doi:10.1016/j.jhydrol.2020.125441

Schilling, O. S., Cook, P. G., and Brunner, P. (2019). Beyond classical observations in hydrogeology: The advantages of including exchange flux, temperature, tracer concentration, residence time, and soil moisture observations in groundwater model calibration. *Rev. Geophys.* 57, 146–182. doi:10.1029/2018RG000619

Schmoll, O., Howard, G., Chilton, J., and Chorus, I. (2006). *Protecting groundwater for health: Managing the quality of drinking-water sources*. London, UK: IWA Publishing.

Sebök, E. (2016). *Temperaturmålinger i Örbäcken, Hagfors, Sverige (in Danish)*.

Selker, J. S., The´venaz, L., Huwald, H., Mallet, A., Luxemburg, W., van de Giesen, N., et al. (2006). Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resour. Res.* 42, W12202. doi:10.1029/2006WR005326

SEPA (Swedish Environmental Protection Agency) (2007). Klorerade lösningsmedel – identifiering och val av efterbehandlingsmetod (in Swedish). Report No. 5663. Stockholm, Sweden: Naturvårdsverket

SMHI (2021). Weather station data for Gustavsfors A. Available at: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer/.

Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C., and Lindström, G. (2012). Water and nutrient predictions in ungauged basins: Set-up and evaluation of a model at the national scale. *Hydrological Sci. J.* 57, 229–247. doi:10.1080/02626667.2011.637497

Swedish National Food Agency (2001). *Slvfs 2001:30: The Swedish National Food Agency's regulations on drinking water*. Stockholm, Sweden: Swedish National Food Agency.

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020a). Toward reproducible environmental modeling for decision support: A worked example. *Front. Earth Sci.* 8. doi:10.3389/feart.2020.00050

White, J. T., Hunt, R. J., Fienen, M. N., and Doherty, J. E. (2020b). Approaches to highly parameterized inversion: PEST++ version 5, a software suite for parameter estimation, uncertainty analysis, management optimization and sensitivity analysis. *U.S. Geol. Surv.* Tech. Methods *7C26*. doi:10.3133/tm7c26

Woessner, W. W. (2020). *Groundwater-surface water exchange*. Ontario, Canada: The Groundwater Project.

Wöhling, T., Gosses, M. J., Wilson, S. R., and Davidson, P. (2018). Quantifying River-groundwater interactions of New Zealand's gravel-bed rivers: The wairau plain. *Groundwater* 54, 647–666. doi:10.1111/gwat.12625

Frontiers in Earth Science

Investigates the processes operating within the major spheres of our planet

Advances our understanding across the earth sciences, providing a theoretical background for better use of our planet's resources and equipping us to face major environmental challenges.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact