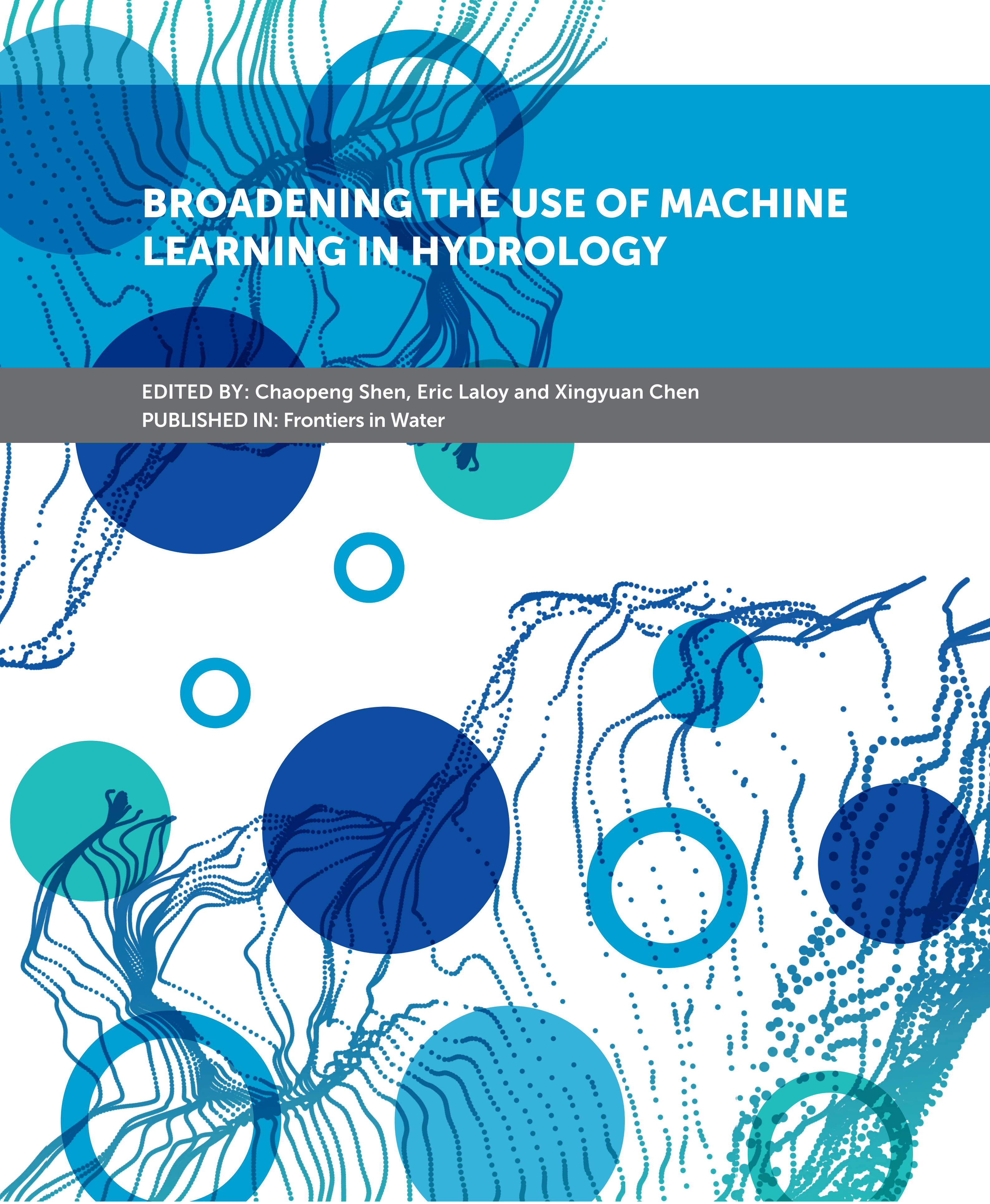


BROADENING THE USE OF MACHINE LEARNING IN HYDROLOGY

EDITED BY: Chaopeng Shen, Eric Laloy and Xingyuan Chen
PUBLISHED IN: Frontiers in Water

frontiers Research Topics

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

ISSN 1664-8714
ISBN 978-2-88966-982-0
DOI 10.3389/978-2-88966-982-0

BROADENING THE USE OF MACHINE LEARNING IN HYDROLOGY

Topic Editors:
Chaopeng Shen, The Pennsylvania State University, United States
Eric Laloy, Belgian Nuclear Research Centre, Belgium
Xingyuan Chen, Pacific Northwest National Laboratory, United States

Citation: Shen, C., Laloy, E., Chen, X., eds. (2021). Broadening the Use of Machine Learning in Hydrology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-982-0

Table of Contents

frontiers in Water

EDITORIAL
published: 11 May 2021
doi: 10.3389/frwa.2021.681023

Editorial: Broadening the Use of Machine Learning in Hydrology

Chaopeng Shen[1]*, *Xingyuan Chen*[2] *and Eric Laloy*[3]

[1] *Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, United States,* [2] *Earth System Measurements & Data, Pacific Northwest National Laboratory, Richland, WA, United States,* [3] *Institute for Environment, Health and Safety, Belgian Nuclear Research Centre, Mol, Belgium*

Keywords: machine learning, hydrology, water cycle, integration, snow, groundwater, streamflow (runoff), precipitation

OPEN ACCESS

Edited and reviewed by:
Harrie-Jan Hendricks Franssen,
Julich-Forschungszentrum,
Helmholtz-Verband Deutscher
Forschungszentren (HZ), Germany

***Correspondence:**
Chaopeng Shen
cshen@engr.psu.edu

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *15 March 2021*
Accepted: *01 April 2021*
Published: *11 May 2021*

Citation:
Shen C, Chen X and Laloy E (2021)
Editorial: Broadening the Use of
Machine Learning in Hydrology.
Front. Water 3:681023.
doi: 10.3389/frwa.2021.681023

Editorial on the Research Topic

Broadening the Use of Machine Learning in Hydrology

The introduction of deep learning (DL) (LeCun et al., 2015) into hydrology around 2016–2018 (Tao et al., 2016; Laloy et al., 2017, 2018; Shen, 2018; Shen et al., 2018), especially the use of long short-term memory (LSTM) as a dynamical modeling tool for soil moisture and streamflow (Fang et al., 2017; Kratzert et al., 2019), has ignited a surge in machine learning applications across all domains of hydrology. At the core, machine learning is a set of tools that allow us to build and train models that extract and reproduce the spatial and temporal patterns in the datasets they encounter. In particular, the central philosophy of DL has been to minimize the intervention of the human experts in feature design and to facilitate maximal extraction of information from data (Goodfellow et al., 2016). Improved prediction quality in hydrologic machine learning (ML) models has been achieved not by infusing process-based assumptions into the models, but by conducting extensive training of the models with large quantities of a priori data. It has been argued by Nearing et al. (2020) that there could be significantly more information in large-scale hydrological data sets than hydrologists have been able to translate into theory or process-based models. The hydrology community is poised to fully explore the power in the vast amount of data using machine learning in various subdomains of hydrology.

In this Research Topic, we sought to broaden the use of machine learning (ML) in hydrology rather than emphasizing the depth of a specific topic. We sought applications of machine learning in both data-rich and data-scarce settings. We are highly encouraged to see the diversity and breadth covered by the resulting collection of published papers, which have almost covered the entire water cycle. A variety of machine learning techniques have been adapted to address various challenges existing in predicting the hydrologic cycle, ranging from a dynamical modeling tool to event localization, and from information extraction to a hypothesis generator. In the following section, we briefly go over some editor-identified highlights of the papers.

Precipitation, as *the beginning of the hydrologic cycle*, is a major source of uncertainty, and most satellite products are still too coarse for water management purposes, making precipitation downscaling a high-stakes activity. Sun and Tang employed an attention-based, deep convolutional neural network (AU-Net) to downscale coarse-resolution satellite-based precipitation data products to 1 km resolution (learning from gauge-based precipitation data products), with the help of auxiliary predictors including elevation, vegetation index, and air temperature. Novel to hydrology, authors employed an attention mechanism that extracts multiscale features by fusing gauged data. However, there are often missing values in gauged precipitation data due to various instrumentation and data quality issues. Mital et al. developed a new sequential imputation algorithm based on a Random Forest technique for interpolating the missing values in spatio-temporal daily precipitation records. They found that, for reliable imputation, having a few strongly correlated references is more effective than having a larger number of weakly correlated references.

Snow is an important precipitation component that is even more difficult to measure (*in-situ* or remotely) than rainfall. Meyal et al. wrote one of the first papers to simulate a snow water equivalent (SWE) using LSTM, leveraging climatic and SWE data from five Snow Telemetry (SNOTEL) stations. They reported Nash Sutcliffe efficiency coefficient (NSE) values ranging from 0.85 to 0.96. The authors build an automated prediction system with online data ingestion. This standard application demonstrated the plausibility of using LSTM for large-scale operational SWE modeling. With only five training sites, however, it remains to be seen if the model can be applied to larger scales.

Streamflow is an important and human-relevant component of the hydrologic cycle. Duan et al. employed a temporal convolutional neural network (TCNN), a one-dimensional dilated convolutional unit with sequential or causal connections, for long-term streamflow projection in California. By comparing the performance of TCNN against other machine learning approaches including the LSTM, Duan et al. not only showed that TCNN excelled at capturing high flows, but also qualitatively demonstrated that TCNN yielded physically plausible estimations of streamflow in responding to precipitation under future extreme climate scenarios beyond the historic records (e.g., under high temperature and quadrupled precipitation), showing that causal convolutions could enhance the stability of ML models when extrapolated outside of their trained conditions.

While still dealing with surface water, Oppel and Mewes present a slightly different application that used machine learning to localize events. They compared several machine learning approaches ranging from support vector machines to extreme learning machines to identify the beginning and end of multiple flood events along with their associated volumes from hydrographs. They also demonstrated that the ML methods afford additional benefits in facilitating the automation of the workflow, which can lead to increased scalability for practical operations.

With the groundwater of the hydrological cycle, Sahu et al. trained a Multilayer Perceptron (MLP) model to predict three-point observations of groundwater levels using temperature, precipitation, river discharge, and past groundwater data as inputs. The authors conducted a sensitivity analysis of features' importance and observed that providing all available inputs to their MLP models was not necessarily the optimal choice. They also found that MLPs trained solely on temperature and historical groundwater level measurements as features were unreliable at all locations, which alluded to the dynamical linkage between surface hydrology and groundwater. Future sensitivity analysis will likely be accompanied by uncertainty estimates to ensure the robustness of the analysis. We also note more effort should be focused on finding ways to generalize these types of models outside of locations with data included in the training set. Groundwater flow problems, due to their lack of observation, the three-dimensional nature of the problem, and strong heterogeneity, are difficult to formulate into uniform learnable problems.

Diving deep into the subsurface environment, Generative Adversarial Networks (GAN) are becoming an alternative to Multiple-point Statistics (MPS) techniques to generate stochastic subsurface fields from training images. An open issue for all the training image-based simulation techniques (including GAN and MPS) is to generate consistent 3D field realizations when only 2D training data sets are available. This is especially relevant to groundwater hydrology for which it is difficult, if not impossible, to collect exhaustive and accurate data about the 3D subsurface distribution of rock types (or physical properties). Coiffier et al. introduced a novel approach termed Dimension Augmenter GAN (DiAGAN) that enables GANs to generate 3D fields from 2D examples. The method is simple to implement as it introduces a random cut sampling step between the generator and the discriminator of a standard GAN. Numerical experiments show that for complex binary subsurface media, the proposed approach is efficient and provides results of similar quality as those obtained by a state-of-the-art MPS method.

Around the world, many aspects of urban water systems, e.g., water supply, discharge, and stormwater management, require upgrades to adapt to the challenges of global change and urban growth. We expect there will be a substantial surge in applications of ML in urban water systems to improve their efficiency and transform them into smart cities. Allen-Dumas et al. wrote a thorough review that synthesized ways in which ML techniques have been applied to different parts of the urban water system in order to address multiple water hazards. They discussed ML applications in monitoring, early warning, prediction of urban water hazards (floods, drought, water contamination, soil erosion, and sediment transport), multi-hazard risks (compound risks), selection of best management practices, etc. They argued that by weaving together multiple ML methods for different risks, we can eventually arrive at a comprehensive watershed-to-community planning workflow for smart-city management of urban water resources.

In agreement with the general trend in the field of hydrology, the abovementioned papers have covered most components of the hydrologic cycle. Outside of this Research Topic, machine learning has been applied to soil moisture (Fang et al., 2019), soil data extraction (Chaney et al., 2019), hydrology-influenced water quality variables including in-stream water temperature (Rahmani et al., 2020) and dissolved oxygen (Zhi et al., 2021), human water management through reservoirs (Yang et al., 2019; Ouyang et al., 2021), subsurface reactive transport (Laloy and Jacques, 2019; He et al., 2020), and vadose zone hydrology (Bandai and Ghezzehei, 2021), among others. ML is not only applicable in data-rich regions but can also be leveraged by data-scarce regions (Feng et al., 2021; Ma et al., 2021). DL-native methods for uncertainty quantification have also emerged (Zhu et al., 2019; Fang et al., 2020). What is still missing to date includes vegetation hydraulics, glaciers, preferential flow, hyporheic exchange, and regional groundwater recharge, though this list is incomplete. We believe these components will be covered by machine learning approaches in the future.

While the broadening of ML has been, to some extent, achieved, one can also notice some limitations and unrealized potential. First, most of the abovementioned use cases are siloed to one variable, e.g., streamflow or precipitation. Second, many of the presented examples are built on small datasets, which means that instead of having learned universally-applicable physical laws, they were locally-fitted models based on the measurement sites in question. The implications of these limitations are that the models are not transferable outside the training region, their potential prediction failures are not yet sufficiently tested, and the information from one observed variable cannot influence the other variables.

There are many angles from which one can overcome the limitations. From a purely data-driven perspective, multi-task learning could allow multiple variables to interact and inform each other. A multiphysics land surface model can be trained to simultaneously predict multiple physical variables in the context of multi-task learning, which is known to improve all tasks. This is because many tasks can use shared representations and are thus constrained by multiple targets at the same time (Caruana, 1997). Alternatively, one may seek to organically tie in physical processes with machine learning, allowing known physical laws such as the mass balance and the law of flow to serve as the connective tissue between different model components. While there is a substantial amount of effort in the direction of knowledge-guided machine learning (Read et al., 2019), there are certainly many different paths toward the goal of integrating physics with machine learning. Outside of this Research Topic, there are methods for parameter learning (Tsai et al., 2020a) and physics-informed neural networks (He et al., 2020; Tartakovsky et al., 2020).

One of such pathways, perhaps a niche one, was documented in (Tsai et al., 2020b). This paper used machine learning to generate articulable hypotheses about which physical factor between soil texture, soil thickness, and slope *caused* water storage and streamflow to be linked in a certain way in a basin, and tested them using a physically-based model. While machine learning is very powerful, due to data limitations and factor covariation, it often cannot distinguish between causal or associative relationships, and what it found are therefore merely *hypotheses*. To test these competing hypotheses, Tsai et al. configured a physically-based hydrology model, PAWS+CLM (Shen and Phanikumar, 2010; Shen et al., 2013; Niu et al., 2017; Ji et al., 2019) to represent these hypotheses, e.g., they increased soil thickness or changed soil texture in one of the synthetic simulations and checked if the storage-streamflow relationships changed in agreement with the hypothesized effect as a result. The outcome of the process-based model can in fact be merged with the machine learning hypotheses in a Bayesian and algorithmic way, which implies this avenue can in fact be autonomously executed. While this paradigm is not expected to become popular any time soon, it does suggest physical models provide unique information that can fill in the gaps (in this case, assessment for a causal relationship) for machine learning methods.

Multiple pathways exist for ML to help to make advances in hydrology: (1) incorporating physics in ML models; (2) improving the interpretability of ML models; (3) developing coupled, physics-informed neural networks; (4) quantifying and propagating uncertainty in model results; (5) developing publicly available benchmark training data sets that can be used to aid and test new ML methods; and (6) building a community computational platform to allow sharing of ML pipelines with easy access to pre-trained ML models (e.g., similar to Model Zoo, https://modelzoo.co/), standardized application-ready datasets, interoperable process-based models, and supercomputing and/or cloud computing resources. Generating public benchmark training data sets (similar to ImageNet, http://www.image-net.org/) that researchers can use to build better ML models is the key to advancing applications of ML in Earth science domains (Dramsch, 2020; Maskey et al., 2020). There is a unique opportunity here to enhance the use of the new generation of remote sensing products that capture components of the water cycle (precipitation, snow, soil moisture, evapotranspiration, groundwater, and runoff), as well as coupled carbon and nutrient cycle components, with increasing spatial and temporal resolutions. Training data may also be generated from process-based models. Leveraging open-source resources from federal agencies is necessary for the success of such extensive and expensive effort. For example, NASA's Earth Sciences Data Systems (ESDS) have generated high-quality training data sets that are open and easily accessible. NOAA, USGS, and other federal agencies have been maintaining extensive observation networks and are developing a large number of integrated Earth system models. Standardized data management practices would significantly increase data usability, and we call for significant investment to support community efforts that address these challenges.

AUTHOR CONTRIBUTIONS

CS, XC, and EL edited the Research Topic and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

CS was sponsored by the U.S. National Science Foundation Award OAC #1940190. XC was supported by the ExaSheds project funded by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Government.

ACKNOWLEDGMENTS

As editors, we thank all authors and reviewers who contributed to this Research Topic. We also thank Frontiers for facilitating this process.

REFERENCES

Bandai, T., and Ghezzehei, T. A. (2021). Physics-informed neural networks with monotonicity constraints for richardson-richards equation: estimation of constitutive relationships and soil water flux density from volumetric water content measurements. *Water Resour. Res*. 57:e2020WR027642. doi: 10.1029/2020wr027642

Caruana, R. (1997). Multitask learning. *Mach. Learn*. 28, 41–75. https://doi.org/10/d3gsgj

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous united states. *Water Resour. Res*. 55, 2916–2938. https://doi.org/10/ggj68b

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *arXiv:2006.13311*. 61, 1–55. doi: 10.1016/bs.agph.2020.08.002

Fang, K., Kifer, D., Lawson, K., and Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resour. Res*. 56:e2020WR028095. doi: 10.1029/2020wr028095

Fang, K., Pan, M., and Shen, C. (2019). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Trans. Geosci. Remote Sensing*. 57, 2221–2233. https://doi.org/10/gghp3v

Fang, K., Shen, C., Kifer, D., and Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophys. Res. Lett*. 44, 11030–11039. https://doi.org/10/gcr7mq

Feng, D., Lawson, K., and Shen, C. (2021). Prediction in ungauged regions with sparse flow duration curves and input-selection ensemble modeling. *arXiv*. http://arxiv.org/abs/2011.13380

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press. Available online at: https://www.deeplearningbook.org/

He, Q., Barajas-Solano, D., Tartakovsky, G., and Tartakovsky, A. M. (2020). Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. *Adv. Water Resour*. 141:103610. doi: 10.1016/j.advwatres.2020.103610

Ji, X., Lesack, L. F. W., Melack, J. M., Wang, S., Riley, W. J., and Shen, C. (2019). Seasonal and interannual patterns and controls of hydrological fluxes in an amazon floodplain lake with a surface-subsurface process model. *Water Resour. Res*. 55, 3056–3075. https://doi.org/10/gghp4s

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci*. 23, 5089–5110. https://doi.org/10/gghmz4

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resour. Res*. 54, 381–406. https://doi.org/10/gdbxmz

Laloy, E., Hérault, R., Lee, J., Jacques, D., and Linde, N. (2017). Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour*. 110, 387–405. https://doi.org/10/gcqftj

Laloy, E., and Jacques, D. (2019). Emulation of CPU-demanding reactive transport models: a comparison of gaussian processes, polynomial chaos expansion, and deep neural networks. *Comput. Geosci*. 23, 1193–1215. doi: 10.1007/s10596-019-09875-y

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*. 521, 436–444. https://doi.org/10/bmqp

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resour. Res*. e2020WR028600. doi: 10.1029/2020wr028600

Maskey, M., Alemohammad, H., Murphy, K. J., and Ramachandran, R. (2020). Advancing AI for earth science: a data systems perspective. *Eos* 101. doi: 10.1029/2020EO151245

Nearing, G., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J., et al. (2020). What role does hydrological science play in the age of machine learning? *arXiv* doi: 10.31223/osf.io/3sx6g

Niu, J., Shen, C., Chambers, J., Melack, J. M., and Riley, W. J. (2017). Interannual variation in hydrologic budgets in an Amazonian watershed with a coupled subsurface—Land surface process model. *J. Hydromet*. 18, 2597–2617. https://doi.org/10/gcrcf8

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., and Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: a demonstration of effectiveness and a delineation of challenges. *arXiv*. arxiv:2101.04423

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C. (2020). 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

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res*. 54, 8558–8593. https://doi.org/10/gd8cqb

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS Opinions: incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci*. 22, 5639–5656. doi: 10.5194/hess-22-5639-2018

Shen, C., Niu, J., and Phanikumar, M. S. (2013). Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface—Land surface processes model. *Water Resour. Res*. 49, 2552–2572. https://doi.org/10/f5gcrx

Shen, C., and Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resour*. 33, 1524–1541. https://doi.org/10/c4r8k5

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., and Ihler, A. (2016). A deep neural network modeling framework to reduce bias in satellite precipitation products. *J. Hydromet*. 17, 931–945. https://doi.org/10/ggj7gh

Tartakovsky, A. M., Marrero, C. O., Perdikaris, P., Tartakovsky, G. D., and Barajas-Solano, D. (2020). Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. *Water Resour. Res*. 56:e2019WR026731. doi: 10.1029/2019wr026731

Tsai, W.-P., Pan, M., Lawson, K., Liu, J., Feng, D., and Shen, C. (2020a). From parameter calibration to parameter learning: revolutionizing large-scale geoscientific modeling with big data. *arXiv:2007.15751*. http://arxiv.org/abs/2007.15751

Tsai, W. P., Fang, K., Ji, X., Lawson, K., and Shen, C. (2020b). Revealing causal controls of storage-streamflow relationships with a data-centric bayesian framework combining machine learning and process-based modeling. *Front. Water* 2:583000. doi: 10.3389/frwa.2020.583000

Yang, S., Yang, D., Chen, J., and Zhao, B. (2019). Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model. *J. Hydrol*. 579:124229. https://doi.org/10/ggj668

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., et al. (2021). From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol*. 55, 2357–2368. doi: 10.1021/acs.est.0c06783

Zhu, Y., Zabaras, N., Koutsourelakis, P.-S., and Perdikaris, P. (2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comp. Phys*. 394, 56–81. https://doi.org/10/ggddhn

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 in Water

ORIGINAL RESEARCH
published: 28 July 2020
doi: 10.3389/frwa.2020.00018

On the Automation of Flood Event Separation From Continuous Time Series

Henning Oppel[1,2]* *and Benjamin Mewes*[2]

[1] *Center for Environmental System Research, Kassel University, Kassel, Germany,* [2] *Institute of Hydrologic Engineering and Water Management, Ruhr-University Bochum, Bochum, Germany*

Can machine learning effectively lower the effort necessary to extract important information from raw data for hydrological research questions? On the example of a typical water-management task, the extraction of direct runoff flood events from continuous hydrographs, we demonstrate how machine learning can be used to automate the application of expert knowledge to big data sets and extract the relevant information. In particular, we tested seven different algorithms to detect event beginning and end solely from a given excerpt from the continuous hydrograph. First, the number of required data points within the excerpts as well as the amount of training data has been determined. In a local application, we were able to show that all applied Machine learning algorithms were capable to reproduce manually defined event boundaries. Automatically delineated events were afflicted with a relative duration error of 20 and 5% event volume. Moreover, we could show that hydrograph separation patterns could easily be learned by the algorithms and are regionally and trans-regionally transferable without significant performance loss. Hence, the training data sets can be very small and trained algorithms can be applied to new catchments lacking training data. The results showed the great potential of machine learning to extract relevant information efficiently and, hence, lower the effort for data preprocessing for water management studies. Moreover, the transferability of trained algorithms to other catchments is a clear advantage to common methods.

Keywords: flood event separation, information extraction, time series, automation, data preprocessing

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Jie Niu,
Jinan University, China
Meysam Salarijazi,
Gorgan University of Agricultural
Sciences and Natural Resources, Iran

****Correspondence:***
Henning Oppel
henning.oppel@rub.de

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *26 March 2020*
Accepted: *17 June 2020*
Published: *28 July 2020*

Citation:
Oppel H and Mewes B (2020) On the
Automation of Flood Event Separation
From Continuous Time Series.
Front. Water 2:18.
doi: 10.3389/frwa.2020.00018

1. INTRODUCTION

Machine-learning has proven its capability in a vast range of applications, especially in those cases when a certain pattern has to be revealed from a huge data archive in order to reproduce it afterwards. Water management tasks require these capabilities in various steps. Natural and anthropocentric processes have to be reproduced in order to model future events and behaviors (Mount et al., 2016). Hence, machine learning (ML) has been applied in a broad range of applications, like streamflow simulation (Shortridge et al., 2016), the interpretation of remote sensing images (Mountrakis et al., 2011), modeling of evapotranspiration (Tabari et al., 2012), rainfall forecasting (Yu et al., 2017), process analysis (Oppel and Schumann, 2020), and many more. However, all water related tasks require pre-processed data. Pre-processing is in this case defined as the extraction of the relevant information from raw data. A typical example is the need for direct runoff flood events that have to be extracted from continuous time series of discharge.

This kind of information can be used for flood event research, training of hydrological models for flood forecasting, design tasks, etc. Despite its relevance and expense, there is no single accepted method to efficiently automate this problem.

Especially the separation of rain fed direct runoff from the base flow, i.e., discharge from deeper soil layers and groundwater with higher transit times, has been subject to much scientific work. This might be due to the fact that rain fed direct runoff events are especially relevant for flood security (Fischer, 2018). The most accurate way to separate direct and base flow runoff in order to define flood events is to use tracer based methods (Klaus and McDonnell, 2013; Weiler et al., 2017). However, tracer data are only rarely available and are not collected on a continual basis. Hence, their application is limited to very few case studies and is not suitable for automated information extraction especially for long time series.

There are three main groups of methods to extract flood events from continuous time series: graphical methods, digital filtering and recession based methods. Graphical approaches (Hall, 1968; Maidment, 1993) are well-established in the water management community, yet they rely on assumptions and experience of the user (Mei and Anagnostou, 2015). Moreover, these types of methods cannot be applied to large data sets and do not allow for automation. Digital filtering techniques overcame this drawback. These methods use a one- (Lyne and Hollick, 1979), two- (Su, 1995; Eckhardt, 2005) or three-parametric (Eckhardt, 2005) base equation to reproduce the long wave response of a hydrograph. The calculated response is treated as the baseflow, the residual of baseflow and hydrograph is treated as the direct runoff. The intersections of baseflow and direct runoff curves can be treated as beginning and end of individual events. These methods are especially applicable to extract information from long time series and allow for automation, like Merz et al. (2006), Merz and Blöschl (2009), and Su (1995). Gonzales et al. (2009) and Zhang et al. (2017) stated that digital filtering techniques, especially the three-parametric filter (Eckhardt, 2005), delivers superior results to all other methods. However, they also pointed out that these methods require local calibration.

The calibration process limits the application of a digital filter to its fitted catchment. Moreover, the missing physical reasoning of the parameters introduced parameter uncertainty to the process (Furey and Gupta, 2001; Blume et al., 2007; Stewart, 2015). Recession based methods try to overcome the lack of physical reasoning (Tallaksen, 1995; Hammond and Han, 2006; Mei and Anagnostou, 2015; Dahak and Boutaghane, 2019). They either rely on a linear (Blume et al., 2007) or non-linear (Wittenberg and Aksoy, 2010) connection between storage and the active process that defines the hydrograph. Other methods try to estimate the parameters of digital filter from the recession curves (Collischonn and Fan, 2012; Mei and Anagnostou, 2015; Stewart, 2015). The drawback of these approaches is the missing automation. Stewart (2015) analyzed several recession curves and their connections to the separation of direct runoff and base flow. Although a connection between direct runoff and base flow was identified, they also found that recession analysis relying on streamflow data solely can be misleading. Under different conditions of the catchment different processes are active, and hence, the connection between storage and runoff changes. Beside this process uncertainty most methods require calibration just like digital filtering techniques and cannot be transferred to other basins.

As already pointed out, the common methods either lack a way to automate them or they require local calibration. Either way, the effort to extract the relevant information is high. Another drawback is that especially the physically based methods search for the true separation of direct runoff and base flow. But, in some cases this might not be the target of a separation. For example: if the task is to evaluate just the first peak of each flood event, no common method can adopt to that target. The power of ML algorithms to detect patterns and to reproduce them in further application could be a solution to this topic. Thiesen et al. (2019) demonstrated that data-driven approaches with different predicors can be applied to the task of hydrograph separation. They found that models using discharge as predictors returned the best results. Although their automated flood event separation performed well, they required a large amount of training data which is limiting the applicability of their approach. Thiesen et al. (2019) estimated a label (flood event / no flood event) for each time step of the continuous time series and, hence, searched for the true separation of direct runoff and base flow. As stated before, this might not be applicable in all cases. Therefore we assumed that the event, i.e., the time stamp of the flood peak is known, but the time of event beginning and end are unknown.

In the first part of the study we assessed which part of a flood hydrograph is relevant to determine the begin and end of the event. Based on a training set generated by expert knowledge we analyzed how many points from a hydrograph excerpt are needed to estimate the event boundaries. Moreover, we analyzed which machine learning algorithms are suitable for this type of problem and how many training data is required to automate the separation process. A major shortcoming of common methods is the local bound applicability. Therefore, we tested if trained algorithms could successfully be applied in new catchments on a regional and trans-regional scale.

2. MATERIALS AND METHODS

In this section we will shortly introduce the case study basins of the Upper Main and the Regen. In the subsequent section, the ML algorithms and their settings will be presented. This section is completed with the introduction of the entropy concept and the performance criteria used to evaluate the ML-algorithms.

2.1. Data

For this study, continuous time series from 15 gauges in south-east Germany have been used. Five gauges from the basin of the Upper Main have been used for local application and the tests on required training data and predictors. Additional five gauges from the Upper Main basin and five other gauges from the Regen basin have been used for regional and trans-regional validation of the trained algorithms solely. The time series had an hourly temporal resolution and covered the time span from 2001 to 2007 in the Upper Main basin, 1999 to 2012 in the Regen basin.

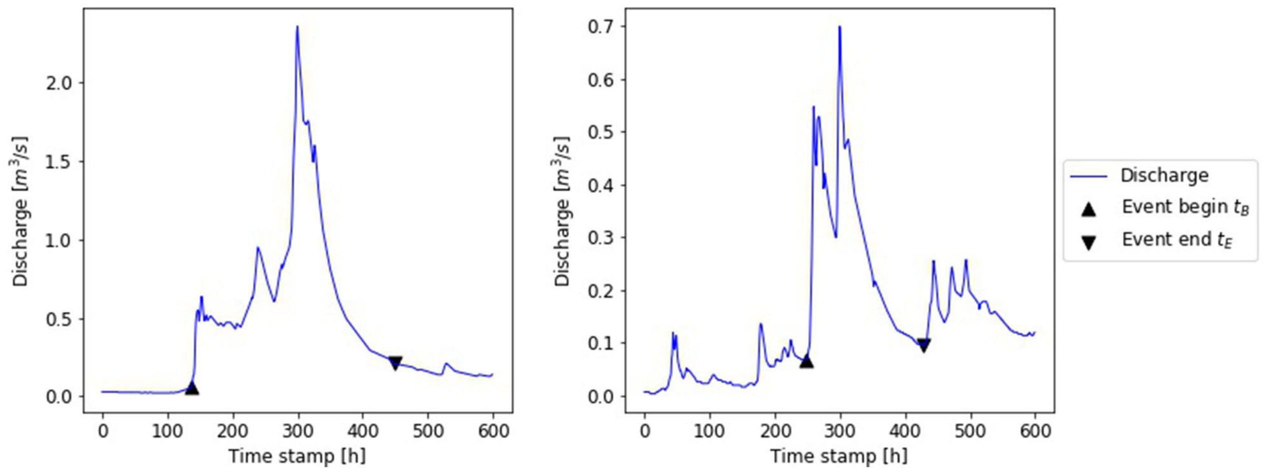


FIGURE 1 | Flood events observed at gauge Friedersdorf with manual defined markers of event begin t_B and end t_E to capture direct runoff.

We assumed that users know what kind of flood events they are interested in and just needs to automate the process of separation (Moreover, the process of peak identification can be automated with a peak-over-threshold (POT) method). Hence, we defined the time stamps of five highest discharge peaks per year as the events of interest. The number of five events per year has been chosen to create a large data basis while maintaining the focus on floods. To create a training and validation data set beginning t_B and end t_E of each event have been defined manually. Due to the focus on flood events our strategy for manual flood separation was to capture begin and end of direct runoff. Although precipitation data was available, we excluded it on purpose to focus on the hydrographs. The begin of the direct runoff t_B was defined as the first significant increase of discharge prior to the peak. The end of direct runoff t_E was defined as either the last change of slope of the recession curve starting from the peak before the next rise, or the last ordinate of the recession curve before the next event (compare **Figure 1**). Target variables t_B and t_E were defined as difference between the time stamp of the peak and the time stamp of the events begin/end.

As the spatial arrangement of the chosen gauges shows (**Figure 2**), training and validation gauges have been selected to cover similar relationships of neighboring and nested catchments. Additionally, the training and validation sets have been compiled to cover the same ranges of catchments area. Each set comprises small catchments with an area between 10 and 100 km^2 and large catchments with an area between 100 and 1,400 km^2 (compare **Table 1**).

The transferability of trained ML algorithms was analyzed by using a regional model strategy. The ML algorithms were trained with the data from the five gauges from the *Training* data set, defined in **Table 1**. All *Training* gauges were located in the upper Main basin. For validation the trained algorithms were used to estimate t_B and t_E for flood events observed at gauges from the regional and trans-regional data set (compare **Table 1**).

2.2. Machine Learning Algorithms

The No-free-Lunch-Theorem pays its tribute to the plethora of available ML-algorithms and reduces the problem of choice to an optimization problem: If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems. A certain algorithm is more or less suitable for a specific problem (Wolpert and Macready, 1997). Accordingly, several approaches have to be taken into account in parallel. Additionally, Elshorbagy et al. (2010a,b) found that a single algorithm is not able to cover the whole range of hydrologic variability. Hence, they recommended to use an ensemble of algorithms for water related tasks. In order to assess which type of algorithm is suitable to the addressed task of this study we used seven different algorithms (provided by Pedregosa et al., 2011 as Python package *scikit-learn*), representing five different algorithm structures.

Artificial neuronal networks (ANN) are the most commonly applied ML algorithms which is also true for hydrological applications (Minns and Hall, 2005; Solomatine and Ostfeld, 2008). The structure of an ANN is inspired by the structure of the human brain (Goodfellow et al., 2016). Multiple input features are connected through multiple neurons on a variable number of hidden layers with the output of the network. The output neuron represents the target variable of the regression (or classification) task. The hidden layers of the ANN define the level of abstraction of the problem. The more layers, the more abstraction is given to the input features (Alpaydin, 2010). Because this study addressed the topic of pattern recognition in hydrographs with a, to this point, unknown degree of abstraction, ANNs with different numbers of hidden layers have been applied. Specifically, an ANN with a single and an ANN with two hidden layers have been applied. The number of neurons per layer has been adjusted during the training process. Both regressors were based on the multi-layer perceptron and used a stochastic gradient descent for optimization (Goodfellow et al., 2016). Additionally, an Extreme Learning Machine (ELM) was added

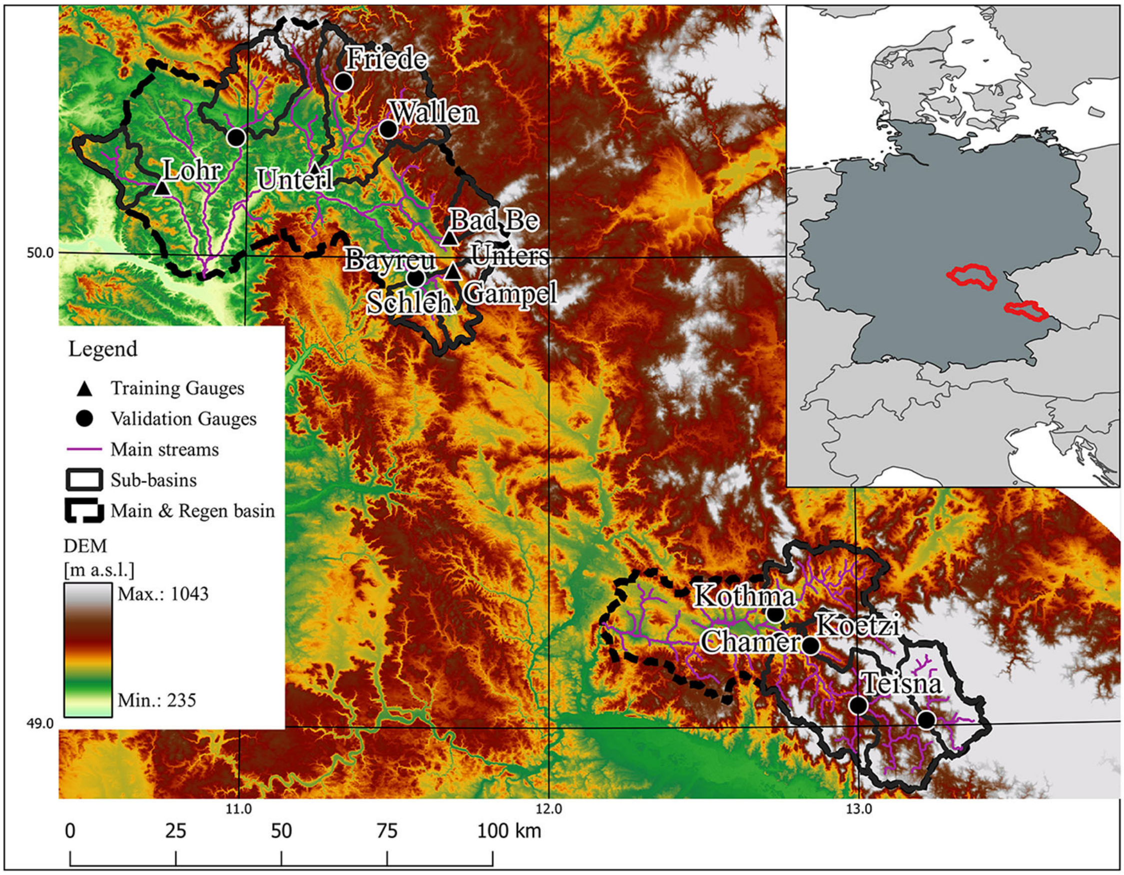


FIGURE 2 | Case study basins of the Upper Main (upper left) and Regen (lower right) in south-east Germany. Five gauges (triangle) have been used for local application and as training data for (trans-) regional application (circle).

TABLE 1 | Gauges and catchment areas in the case study regions.

Training		Reg. Validation		Tr.Validation	
Gauge	**Area [km^2]**	**Gauge**	**Area [km^2]**	**Gauge**	**Area [km^2]**
Bad Berneck	99.7	Bayreuth	340.3	Chamerau	1356.5
Gampelmuehle	62.2	Coburg	346.3	Kothmaissling	405
Lohr	165.3	Friedersdorf	11.1	Koetzing	224.4
Unterlangenstadt	713.9	Schlehenmuehle	70.95	Teisnach	626.6
Untersteinach	73.5	Wallenfels	96.45	Zwiesel	293.4

to the group of used algorithms. The ELM is a special type of ANN (Guang-Bin Huang et al., 2004) that was designed for a faster learning process. In a *classic* ANN each connection between neurons is assigned with a weight that is updated in the optimization process. An ELM has fixed weights for the connection between hidden layer and the output neuron. Only the remaining connections are optimized during the training process. Due to this simplification, the ELM learns faster while regression outputs remain stable (Guang-Bin Huang et al., 2004).

The three types of neuronal networks are accompanied by 4 other algorithms. As a representative for the similarity-based algorithms the K-nearest-neighbor (KNN) algorithms has been applied (Kelleher et al., 2015). Here, no model in the common sense is trained. For regression, the KNN uses the predictors to define similarity between the elements of a new data set and the known cases of the training data. The output is then defined as the average of the *k*-nearest elements. In this study, *k* was defined iteratively during the training process within a range of [5;10]. Parameter values for *k* outside the specified range were tested, but rarely proved to be a better alternative. In order to accelerate the training of the KNN, the comparatively small parameter space was chosen. A Support Vector Machine (SVM) algorithm, an error-based approach, has also been used in this study. The SVM fits a M-dimensional regression model to the given problem, where M can be greater than the dimension of the original feature space. To maintain a reasonable computation time, the SVM focuses on data points outside a certain margin around the regression line, the so-called support vectors (Cortes and Vapnik, 1995). Another type of ML-algorithm included in this study was a Classification and Regression Tree (CART). Regression trees are built node per node with a successive reduction of regression error between the estimates and the true values. CART-regressors have been used as base estimators for a Random Forest (RF) that has been used additionally in this study. The RFs consisted of 1,000 regression trees, each trained with a randomly chosen subset of the given training data. The average of all regression results is returned as estimate of the RF. We applied the RF due to its common application in hydrological studies (Yu et al., 2017; Addor et al., 2018; Oppel and Schumann, 2020). Moreover, the use of an ensemble regressor accounts for the recommendations of Elshorbagy et al. (2010a). Details on implementation are provided by Pedregosa et al. (2011).

The applied algorithms face several inherent problems and advantages, so the right choice of a suiting algorithm depends on the available data and the problem to be solved. SVMs, for example, work perfectly if the margin of the separating vector is small. Thus, they tend to overfit if that is not the case in the data they are trained on. Moreover, the choice of the internal kernel is not trivial and has impact on the results and the training behavior. CART trees are very comprehensible models and quickly converging models, but tend to overfit, so the remaining degrees of freedom have to be considered for an interpretation of CART results. RF on the other hand reduce to vulnerability of overfitting, yet the build less comprehensible outcomes due to the large number of possible model trees. ANNs are robust against overfitting, but require more data to converge in complex situations than the other approaches. ELM inherits the advantages and problems of ANNs and SVM. KNN converge very quickly and are often a suitable method. Nevertheless, the general ability of KNN for ML prediction requires information on internal structure of the data and its internal clustering of groups.

2.3. Shannon Entropy

The entropy concept, introduced by Shannon (1948), is the underlying concept of information theory (Cover and Thomas, 2006). Shannon's entropy concept is used to determine the information content within a given data set. Entropy H is calculated for a discrete random variable X with possible values $x_1, \ldots, x_n$ by:

$$H = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{1}$$

where $P(x_i)$ is the probability that X takes exactly the value x_i. The basis b of the log-function can take any value, but is usually set to $b = 2$, which gives H the unit *bit*. As Equation (1) shows, the entropy value is a measure for uncertainty of the considered variable. If all samples drawn from X would take the same value, the probability of this value would be 1 and hence the entropy would be equal to 0.0, because one would be absolutely certain about the outcome of new samples drawn from X. The entropy increases to a value of 1.0 if the sample would be equally distributed on two outcomes (Kelleher et al., 2015). The higher the entropy, the wider the histogram of X is spread.

The problem with Equation (1) is that it can only be applied to discrete data. Unfortunately most hydrological relevant data is continuous. This was also the case in this study, because the ordinates of the hydrograph are intended to determine the events temporal boundaries. Gong et al. (2014) showed that the use of frequency histograms, which is also refereed as *Bin Counting*, is a feasible and reliable approach to represent the continuous as a discrete distribution function. To apply *Bin Counting* the width of bins has to be determined. Scott (1979) proposed the following estimator for the optimal bin-width h^*:

$$h^* = 3.49\sigma N^{-1/3} \tag{2}$$

where σ is the standard deviation of the data and N is the number of samples. We followed the recommendations of Scott (1979) and Gong et al. (2014) and used *Bin Counting* to calculate the entropy of the predictor and target variables.

2.4. Performance Criteria

Estimation errors manifest as differences between estimated and manually defined time stamps of event begin and end, resulting in different event metrics duration and volume. The deviations of these metrics were used to define the performance criteria. First, the mean volume reproduction MVR was defined as follows:

$$MVR = \frac{\sum_{i-1}^{N} \frac{V_{Est;i}}{V_{Man;i}}}{N} \tag{3}$$

where N is the number of considered events and V is the estimated (*Est*) or manual defined (*Man*) event volume. The MVR is defined within $[0; +\text{inf}]$ with an optimal value of 1. The second metric accounts for the duration of the event. For each event two sets of time stamps are available: set M containing all time stamps of the manually separated event, and D containing all time stamps of the estimated event. Time stamps within both sets are correctly ascertained time stamps by the ML-algorithm. This set I can be expressed as the intersection of both sets $I = D \cap M$. Temporal coverage of an estimated event has been calculated as the ratio of the cardinalities of I and M, i.e., the ratio of correctly ascertained time stamps and the true number of event time stamps:

$$COV = \frac{|D \cap M|}{|M|} = \frac{|I|}{|M|} \tag{4}$$

Temporal coverage COV is defined on $[0; 1]$ with an optimal value of 1. Note that COV only accounts for errors of time stamps, not the actual event duration. An estimate of event boundaries that sets event begin and end wrong, but outside of the true event boundaries, has a coverage equal to 1. However, the error will be accompanied by an MVR greater than one. The combined evaluation of COV and MVR reveals that the time stamps were set outside the true event boundaries.

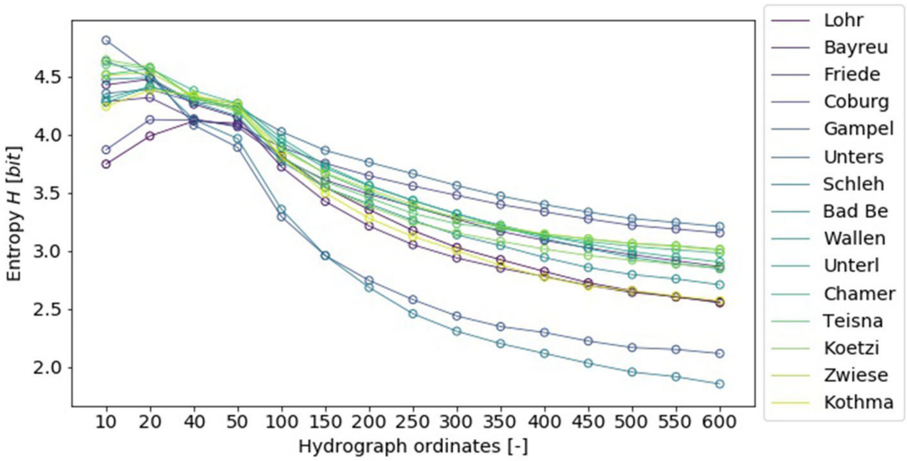


FIGURE 3 | Entropy values of the data sets *H* considering a varying number of ordinates from hydrographs (number of ordinates considered on the abscissa).

3. RESULTS

In this section, the analysis regarding the automation of the flood event hydrograph separation will be presented. Section 3.1 presents the selection of the ML-predictors, i.e., the number of hydrograph ordinates necessary to predict the event boundaries. This is followed by the results of the local application of the ML-algorithms (section 3.2).

3.1. Predictor Selection

As predictors for the estimation of event boundaries (time stamps of beginning and end of a flood event), we intended to use the ordinates of the hydrograph itself. Therefore, we had to determine the required amount of ordinates to achieve satisfactory results, while keeping the amount of predictors as low as possible to minimize the training effort of the ML-algorithms.

In other words we wanted to focus on the necessary hydrograph components to determine flood begin and end. In course of the graphical, manual separation (section 2.1) we observed that we mainly paid attention to the shape of the hydrograph in comparison to its closer hydrological context for our decisions. Transferring this to the numerical data of the hydrographs (Q) means that the set of hydrograph ordinate with the highest uncertainty about Q conveyed the highest amount of relevant information to the separation process. In order to determine the length of these sets we performed an entropy analysis for different lengths of sets (see below). We used the entropy metric to evaluate the information content H, because its values quantifies the uncertainty of a data set (compare section 2.3).

Although H calculated separately for the predictor and the target variable set allowed us to compare the information within the data, they do not tell us if these information coincide. The common approach to quantify the shared information content of data sets is to use the mutual information (MI) (Sharma and Mehrotra, 2014). The MI-value concept evaluates the joint probability distribution of two (or more) data sets and evaluates the information obtained from the predictor data set about the target data set. Due to the high dimension of our predictor data set (number of hydrograph ordinates between 10 and 600), the joint probability distributions could not be estimated. Hence, the concept of MI was not applicable. Hence, we relied on H calculated for target and predictor sets separately, to evaluate the predictor data sets. We assumed that an entropy value of the predictor set similar to the entropy of the target variable set is a necessary but not sufficient condition for an optimal predictor.

First, we calculated the entropy of the target variables for manually separated events, the time stamps of event beginning and end. Equation (1) and (2) were applied to all available data sets. We obtained average entropy values of $H_A = 1.55$ *bit* for the event beginnings and $H_E = 2.15$ *bit* for event ends. The standard deviation of H_A and H_E between the considered sub-basins was $\sigma(H_A) = 0.15$ and $\sigma(H_E) = 0.39$. The entropy values showed that the position of the flood beginning (in relation to the peak) is afflicted with less uncertainty than the end of the flood. A result that is in concordance with our experience from the manual flood separation.

In the second step of the analysis we calculated the entropy for different predictor data sets. The first data set evaluated consisted of 10 hydrograph ordinates, half of the ordinates prior to the peak the other half succeeded the peak. The amount of ordinates was increased incrementally up to 600 ordinates. The obtained entropy values showed that the data sets contained the highest entropy if only a few ordinates were used (**Figure 3**). Data sets with 10–50 ordinates, regardless of the sub-basin, showed an entropy value of $H \geq 3.7$ *bit* which is equal to the sum of H_A and H_E. With an increasing amount of data the entropy values decreased significantly. With 500 data points considered, the entropy values lowered to a range of $[2.0, 3.5]$ *bit* and did not change any further with increasing data points.

In order to evaluate our assumption of the connection between equal entropy values and predictive performance, a test with different ML-algorithms in all sub-basins was carried out. Each data set was split randomly into training data (50%)

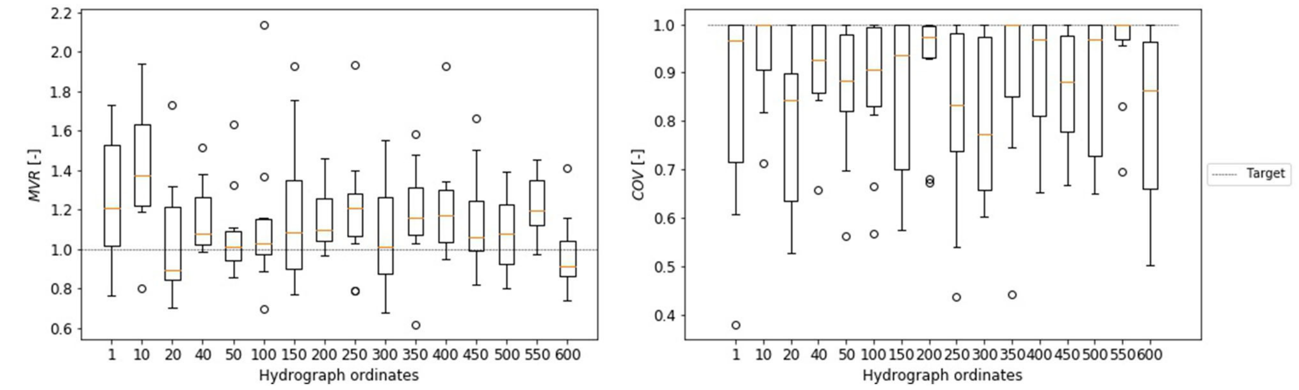


FIGURE 4 | Dependence of the number of hydrograph ordinates and the mean volume reproduction (MVR) and temporal coverage (COV) of automatically separated flood events. Application of a trained RF in sub-basin Bad-Berneck.

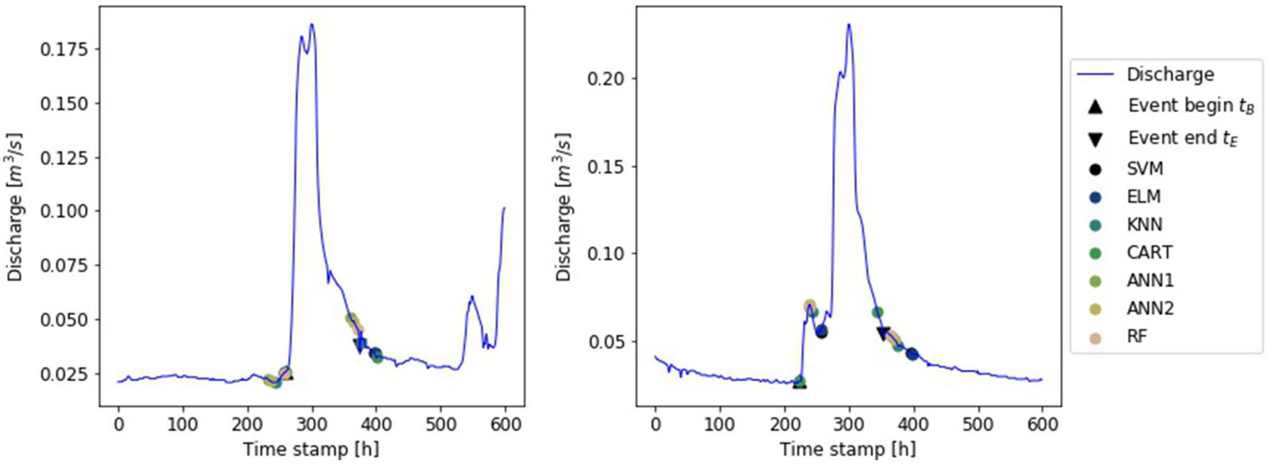


FIGURE 5 | Hydrographs for two flood events at gauge Lohr with manually and automated defined markers of event begin t_B and end t_E.

and validation data (50%). Each ML-algorithm was trained and validated with the MVR (Equation 3) and COV (Equation 4). To minimize uncertainty due to the choice of training data, the evaluation was repeated 10-times for each data set. The obtained results were comparable in all applications. For the majority of catchments the best MVR-results (median and variance) were achieved with 40 or 50 ordinates used as predictors and the optimal COV with 50 ordinates. As an example the results of RF application in catchment Bad Berneck are shown in **Figure 4** (Results for all other algorithms and catchments can be found in the **Supplementary Material**).

Based on the experimental results and the evaluation of the entropy values we chose 40 ordinates, 20 prior to the peak and 20 succeeding the peak, as the predictor data set for the following analysis.

3.2. Automated Flood Hydrograph Separation

For each data set from the catchments marked as *training gauges* in **Table 1** and **Figure 2**, we tested if flood hydrograph separation could be automated by means of ML. Like in the previous section, we randomly chose 50% of the available flood event data for training of the algorithms. Their performance was validated with withheld data from the respective gauge. Again, the procedure has been repeated to lower the uncertainty due to the randomly chosen subsets. In this case, 500 iterations were performed. For each event of the validation data, t_B and t_E were estimated with all available ML-algorithms (**Figure 5**).

The results showed that the ML-algorithms were able to perform the required automation task. However, they tended to overestimate the volume of the events (**Figure 6**), while the temporal coverage was met in the most cases (**Figure 7**). Only the ANN1 and ANN2 did not match the temporal extend of the events. The combination of a COV lower than 1 and MVR greater than 1 (compare **Figure 5**, right panel) showed that one time stamp was set too close to the peak, while the other was set too far from to peak. Giving a low coverage of the event and high volume error. In these cases it was the event start that was set too close to the peak and the end was set too far. A different behavior is visible in the results of the ELM and KNN. While the COV is

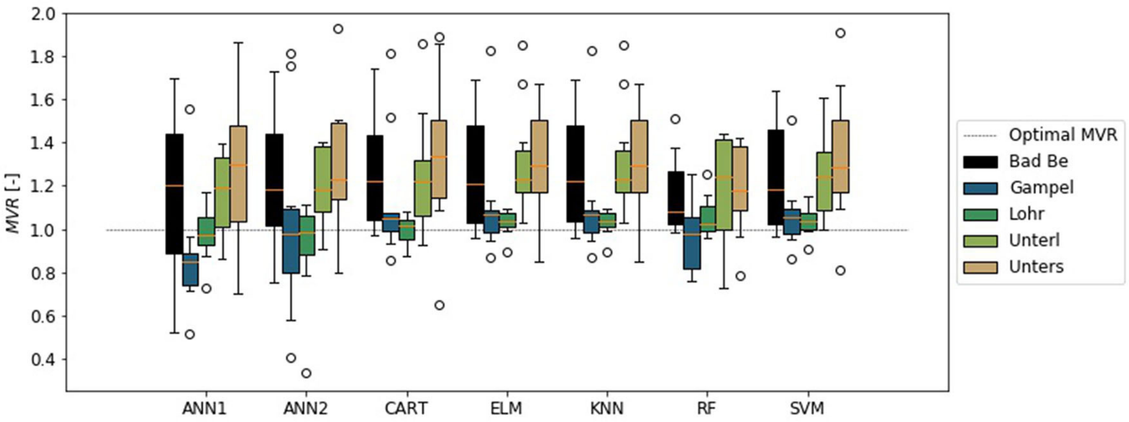


FIGURE 6 | Mean volume reproduction (MVR) of the validation flood events in local application of trained artificial neuronal network with 1- (ANN1) and 2-hidden layers (ANN2), regression tree (CART), extreme learning machine (ELM), k-nearest-neighbor (KNN), random forest (RF), and support vector machine (SVM).

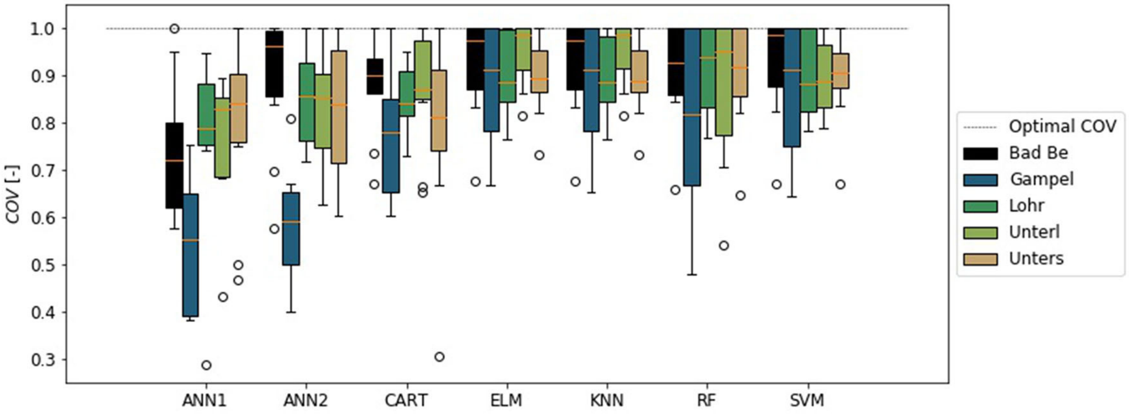


FIGURE 7 | Temporal coverage (COV) of the validation flood event duration in local application of trained artificial neuronal network with 1- (ANN1) and 2-hidden layers (ANN2), regression tree (CART), extreme learning machine (ELM), k-nearest-neighbor (KNN), random forest (RF), and support vector machine (SVM).

close to 1, the MVR shows an average overestimation of event volume of 20%. This shows that ELM and KNN separated too long flood events. The best results were obtained with the RF and the SVM.

The results also showed regional dependence of the model error. Independent from the chosen algorithm, *Bad Berneck* showed the highest volume errors, while *Gampelmuehle* showed the highest COV errors. It is striking that *Gampelmuehle* on the other hand showed one of the lowest volume errors, and *Bad Berneck* the lowest COV errors. Contrary to that, the three remaining basins showed comparable results for both criteria. An explanation for this observation lies within the response time of these catchments. In comparison to the other hydrographs, they are significantly more flashier and the duration of the flood events is significantly shorter.

4. DISCUSSIONS

The presented results showed that ML is in general capable to automate the considered task. But several choices, like the amount of training data have to be discussed and the transferability of trained algorithms has to be tested. This sections provides discussions on these topics.

4.1. Training Data

The results showed that all algorithms could be used in local application to automate the task of flood event separation from continuous time series. Yet, the true benefit of the automation is unclear, because we randomly selected the size of the training data set. A true benefit for automation would be a minimal requirement of training data, because this would minimize the manual effort for separation. The results in section 3.2 showed that we could at least half the manual effort. But how many manually separated flood events are really necessary to train the algorithms?

To answer these questions, an iterative analysis has been performed. First, 25% of the available flood events were randomly chosen as validation data set and removed from the data pool. In the succeeding steps a variable amount of training data was chosen from this pool to train the ML-algorithms.

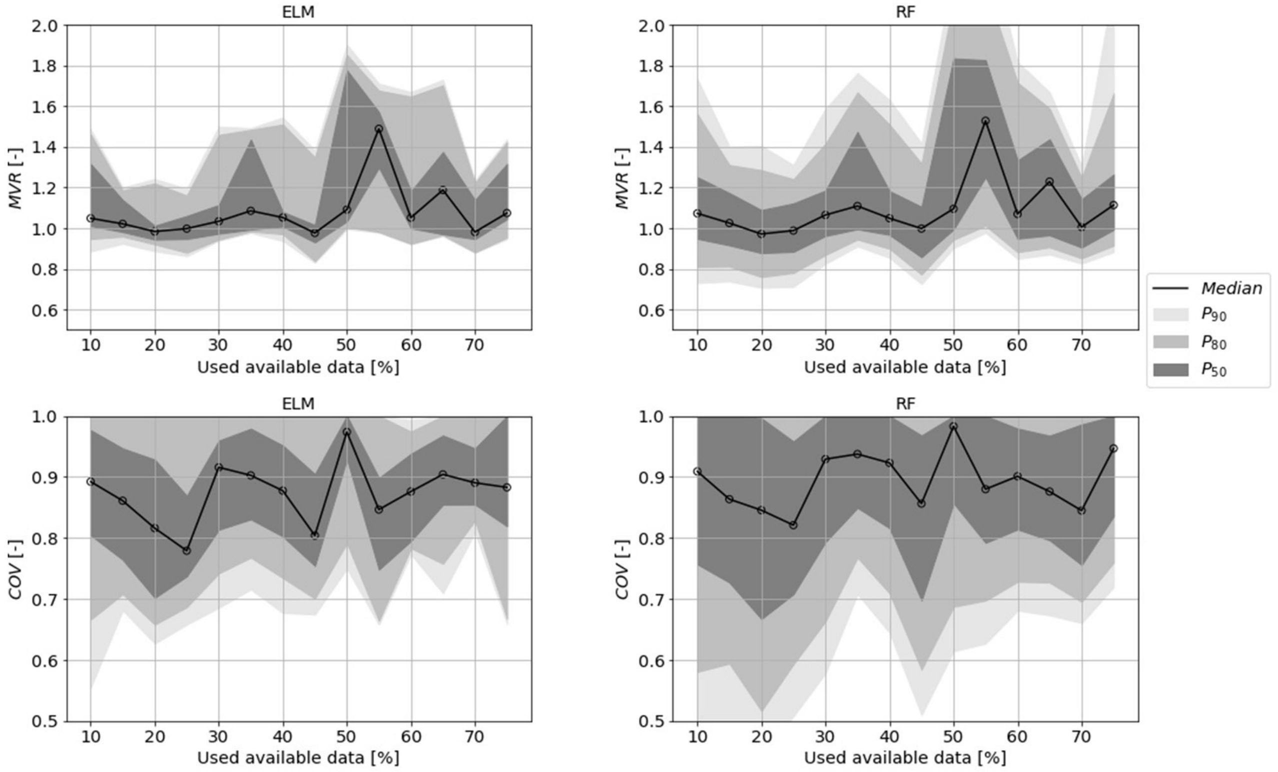


FIGURE 8 | Dependence of mean volume reproduction (MVR)/temporal coverage (COV) and size of the training data set of a random forest (RF) and an extreme learning machine (ELM). Uncertainty belts drawn in gray scales for different probabilities (50, 80, 90%). The amount of training data has been raised incrementally to train the algorithms and were validated in each step with the same data set, containing 25% of the available data.

In each step, the trained algorithms were validated with the same validation data set. In order to minimize uncertainty due to the randomly chosen data sets, this procedure was repeated 500 times.

The results showed that the required amount of training data was surprisingly low for all algorithms. The median MVR reached the optimum of $MVR = 1.0$ with the lowest amount of uncertainty with only 20–30% percent used training data (**Figure 8**, full plot with all ML-algorithms in the **Supplementary Material**). This was true for all ML-algorithms used in this study. The results for the COV criterion were similar to these findings. But in contrast to the MVR criterion, the uncertainty decreased slightly with increasing training data. The combined evaluation of MVR and COV showed different types of estimation errors. With a small data set the duration of the separated flood events is afflicted with higher uncertainty, while the true volume of the event is more likely to be met and vice versa for larger data sets. However, the orders of magnitude differ. The certainty of event duration does not increase to the same extent as the uncertainty of the volume increases.

Note that in this study only 20 events per sub-basin were available, which means that a training data set of 4–5 manually separated flood events was a sufficient training data set for the automation of the task.

4.2. Transferability

In this section we present the results of the conducted test on the ability to transfer the trained algorithms to other catchments. First, a regional transfer has been tested. Here, we used the data sets from the local application (sections 3.2, 4.1) to train the ML-algorithms and validated their performance at five new gauges in the same basin, i.e., regional neighborhood (**Figure 2**). Likewise to the procedure in section 4.1, we analyzed the impact of training data on the performance. Here, we had a total of 117 flood events for training and validated with the individual data sets from the new five catchments.

The performance of the ANN1 and ANN2 stabilized at 30–40% of used data for both criteria (**Figure 9**). Estimates from both algorithms reached a median $MVR \approx 1.05$ and a median $COV \leq 0.8$. A similar performance was achieved with the ELM and the KNN, only that the obtained COV values were larger than 0.8. Additionally, the ELM and KNN showed faster learning than all other algorithms. Results stabilized at approx. 5% of used training data. Further changes in median performances and the uncertainty belts with increasing training data were insignificant. The only algorithm that showed constant improvement, i.e., a reduction of the uncertainty belt, was the RF. However, this improvement was accompanied by a steady increase of volume. With all available data used for training, the volume was overestimated by approx. 10%. The concept of

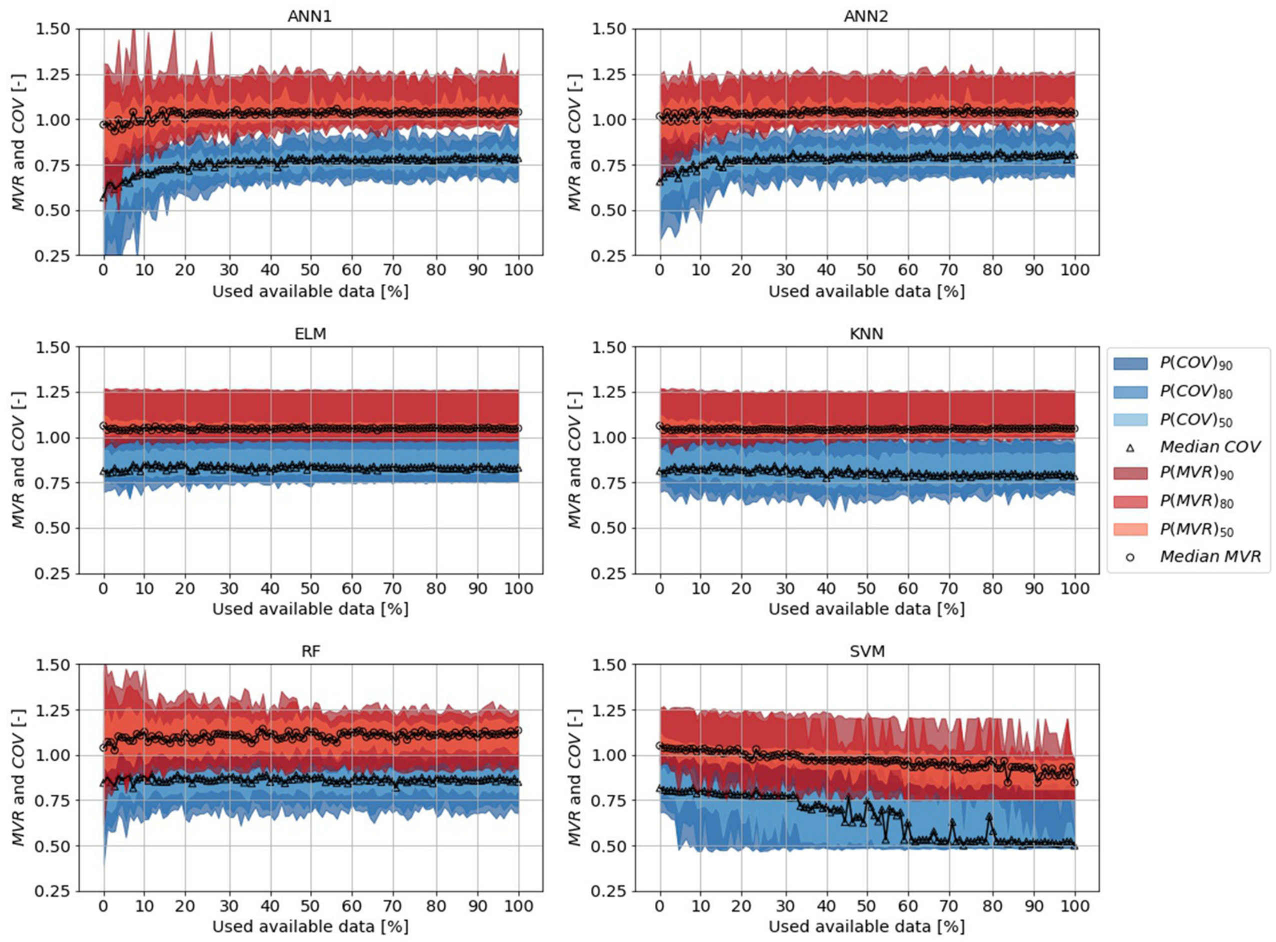


FIGURE 9 | Dependence of mean volume reproduction (MVR)/temporal coverage (COV) and size of the training data set for different ML-algorithms in regional application. Uncertainty belts with different probabilities (50, 80, 90%) drawn for MVR in red scales, for COV in blue scales. The amount of training data has been raised incrementally to train the algorithms. Validation was performed on data sets in regional neighborhood of the training data sources in the Main basin.

support vectors, as used in the SVM, proved to be not useful in this case. Recall that the support vector defines a range around the M-dimensional regression *"line"* and all data points falling within the defined range are excluded from the optimization. This focus on the outliners of the problem resulted in the inferior performance of the SVM (**Figure 9**). Note that the results of the CART algorithm are not shown in **Figure 9**, because the results are similar to the results of the KNN, but with a median $MVR = 1.1$ and median $COV = 0.75$.

In summary, the results showed that even with a small data set automated hydrograph separation could be performed in regional application. Neural network estimators (ELM, ANN1 and ANN2) and similarity-based estimators (KNN) performed best. Flood event duration estimates were afflicted with median bias of 20%. However, this mismatch of event duration did not result in a significant volume error (5% overestimation with ELM & KNN). Our results showed that a training data set of 35 manually separated flood events was needed to train ANNs, only the ELM and KNN should be used with less available data.

Based on this results, we asked if the algorithms could be applied to catchments of another basin, i.e., if the trained algorithms could be used in a trans-regional application. Likewise to the regional application, trained algorithms were used to estimate the time stamps of event begin and end of the floods events, but in this case for catchments in the Regen basin (**Figure 2**). The results of the trans-regional applications approved our findings of the regional application (**Figure 10**). Again, the ANN1 and ANN2 required 30-40% of the data to reach stable results. ELM and KNN, again, required less training data. Contrary to the regional application, the median MVR of the RF converged toward the optimum value of 1.0 with increasing data. Again, a training data set of approx. 35 flood events was sufficient to automate the task of hydrograph separation, even in a trans-regional application.

4.3. Hydrograph Similarity

Our results showed that we could successfully apply an ELM or KNN trained with data from five basins in the Upper Main

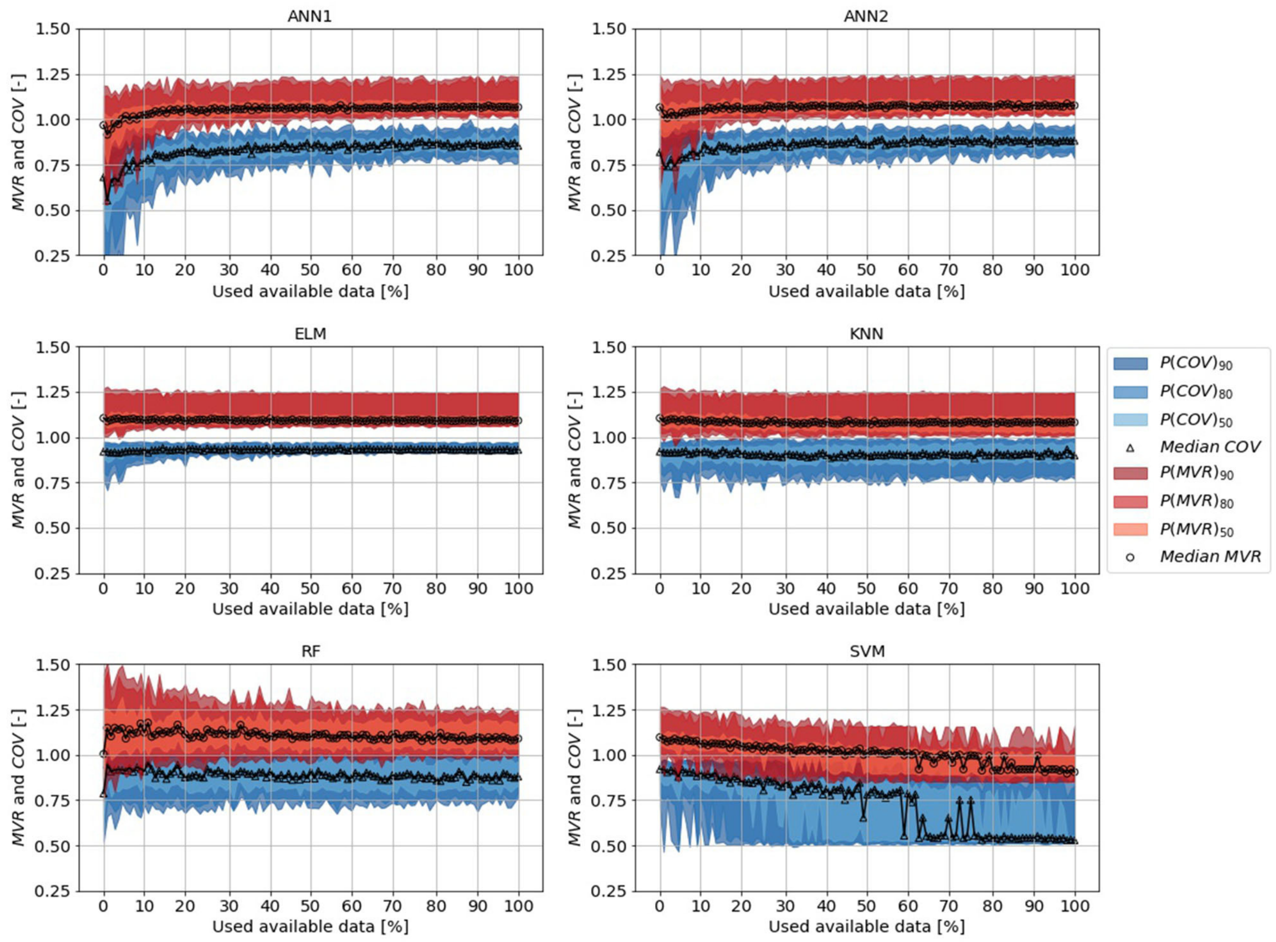


FIGURE 10 | Dependence of mean volume reproduction (MVR)/temporal coverage (COV) and size of the training data set for different ML-algorithms in trans-regional application. Uncertainty belts with different probabilities (50, 80, 90%) drawn for MVR in red scales, for COV in blue scales. The amount of training data has been raised incrementally to train the algorithms. Validation was performed on data sets in the Regen basin.

to other sub-basins within the same catchment *and* in another catchment. This brought up the question: why did it work? A trained, i.e., calibrated model can only be applied to other data without significant performance decrease if the patterns, i.e., variance, within the new data matches the training data. In the previous analysis we proved that our trained models could be applied without performance decrease. Hence, we made the hypothesis that the hydrographs within the training and validation data set, i.e., their variance was similar. As stated in section 3.1 the entropy concept is a good tool to assess the information, i.e., the variance within data sets. Hence, we analyzed the entropy of the training and validation data sets in order to test our hypothesis.

Although entropy quantifies the amount of information, it cannot assess the actual information and is, hence, not applicable to evaluate the equivalence of two data sets. But, if redundant information is added to a data set its entropy value decreases (compare section 2.3). We exploited this behavior of the entropy metric to assess the information equivalence of the training and validation data sets.

We incrementally enlarged a merged data set comprising hydrographs from the training data and one of the validation data sets (regional/transregional). In each step we added a single hydrograph to the data set and calculated the entropy value (Equation 1). First we added all training hydrographs, then we added the validation hydrographs. In order to assess the uncertainty of H, due to data availability we repeated this procedure 500 times, in each iteration only used 50% of the available data (randomly selected).

The results of this analysis supported our hypothesis (**Figure 11**). We found that H increased very quickly with only 2 or 3 data sets (actual position of H_{Max} depending of selected hydrographs). After that H decreased, with some variance in its development due to data selection. Although variance was visible, $H_{Max} < 2.5$ $[bit]$ was never exceeded with the additional validation data, neither with the regional nor with the trans-regional data set. Note that H_{Max} in this analysis was lower than the entropy values in section 3.1, because normalized hydrographs have been used to assess the information given to the ML-algorithms.

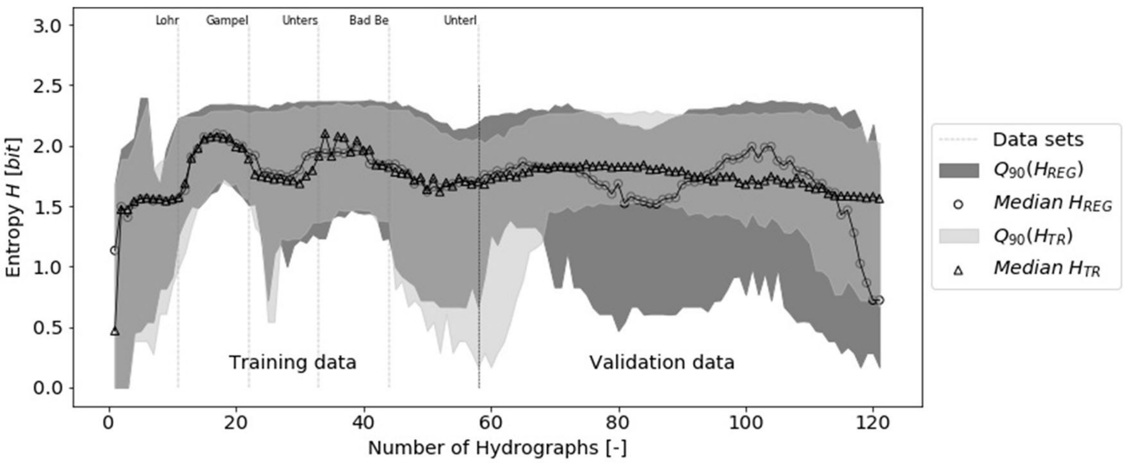


FIGURE 11 | Development of entropy *H* for merged training and validation (regional *REG*/trans-regional *TR*) data sets. Median (black lines) and 90%-uncertainty belts calculated by randomly adding 50% of the available hydrographs per sub-basin to the merged data set.

5. CONCLUSIONS

In this article we demonstrated how machine learning can be used to automate the task of hydrograph separation from continuous time series. As predictor for the used ML-algorithm we used the ordinates of hydrograph, solely. This minimized the effort for data pre-processing. An analysis of entropy values and numerical experiments showed that only a short excerpt of the hydrograph (40 values, 20 prior, and another 20 succeeding the flood peak) were required for hourly discharge data.

Seven different ML-algorithms were trained with manually separated flood events and were applied locally, regionally and trans-regionally. All applications showed that machine learning was able to extract the relevant information (flood event duration and volume). In the local application, i.e., application of the trained algorithms to the same catchment, RF and SVM showed the best results. However, in regional and trans-regional application, i.e., application to other catchments than the training data source, estimators based on artificial neuronal networks (ELM, ANN with 1 hidden layer) and similarity based estimator (KNN) performed best.

Moreover, we demonstrated that the application of ML minimizes the effort for manual data pre-processing. For local application, data sets containing only 4–5 manually separated events were sufficient to transfer the experts knowledge to the algorithms. For a transfer of the trained algorithms to other catchments lacking training data, the manual effort increased slightly. In our applications, 35 events from 5 gauges, i.e., 7 events per gauge transferred the required amount of information to the ML-algorithm.

A striking observation was that the performance of flood event separation was comparable in local, regional and trans-regional application. With an assessment of information equivalence in the training and validation data sets we demonstrated that the variance of our predictors necessary to be applied to other data sets, could be covered with our training data set. The result of the analysis not only supported our hypothesis about information equivalence, but also provided an explanation why our approach to automation of event separation had a quicker learning process than other approaches like Thiesen et al. (2019). We excluded the majority of natural variance within the continuous time with the focus on the events we are interested in (via POT-method). From the time-stamp returned by POT we used the 40-discharge ordinates around the peak as predictors for the estimation of event beginning and end. With this procedure we focused the ML-algorithms on the shape of the flood event and trained it to identify its begin and end. Our results proved that this approach delivered good results and requires a minimum amount of manual work for training.

However, we have to focus on this topic in future works. We excluded the transfer to other climatic conditions and we excluded the impact of biased data. With additional data, taking more catchments into account, we want to test the application of trained algorithms to a wider range of possible applications than presented in this study. Moreover, more numerical experiments have to be carried out to evaluate the impact of the training data and choices made by the user, for example the chosen separation target. In this study we tried to separate the full flood event. However, other users might be interested in other tasks. Although our results are promising in this respect, further tests must be carried out.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

This study was developed and conducted by both authors (HO and BM). HO provided the main text body of this publication which was streamlined by BM.

FUNDING

The financial support of the German Federal Ministry of Education and Research (BMBF) in terms of the project Wasserressourcen als bedeutende Faktoren der Energiewende auf lokaler und globaler Ebene (WANDEL), a sub-project of the Globale Ressource Wasser (GRoW) joint project initiative (Funding number: O2WGR1430A) for HO is gratefully acknowledged.

ACKNOWLEDGMENTS

The authors would like to thank the Bavarian Ministry of the Environment for providing the Data used in this study. We also like to thank Svenja Fischer for her feedback that helped to improve this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2020.00018/full#supplementary-material

REFERENCES

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., and Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resour. Res.* 54, 8792–8812. doi: 10.1029/2018WR022606

Alpaydin, E. (2010). *Introduction to Machine Learning. Adaptive Computation and Machine Learning, 2nd Edn*. Cambridge, MA: MIT Press.

Blume, T., Zehe, E., and Axel, B. (2007). Rainfall–runoff response, event-based runoff coefficients and hydrograph separation. *Hydrol. Sci. J.* 52, 843–862. doi: 10.1623/hysj.52.5.843

Collischonn, W., and Fan, F. M. (2012). Defining parameters for Eckhardts digital baseflow filter. *Hydrol. Process.* 27, 2614–2622. doi: 10.1002/hyp.9391

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edn*. Hoboken, NJ: Wiley-Interscience.

Dahak, A., and Boutaghane, H. (2019). Identification of flow components with the trigonometric hydrograph separation method: a case study from Madjez Ressoul catchment, Algeria. *Arab. J. Geosci.* 12:463. doi: 10.1007/s12517-019-4616-5

Eckhardt, K. (2005). How to construct recursive digital filters for baseflow separation. *Hydrol. Process.* 19, 507–515. doi: 10.1002/hyp.5675

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P. (2010a). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: concepts and methodology. *Hydrol. Earth Syst. Sci.* 14, 1931–1941. doi: 10.5194/hess-14-1931-2010

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P. (2010b). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: application. *Hydrol. Earth Syst. Sci.* 14, 1943–1961. doi: 10.5194/hess-14-1943-2010

Fischer, S. (2018). A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons. *J. Appl. Stat.* 45, 2831–2847. doi: 10.1080/02664763.2018.1441385

Furey, P. R., and Gupta, V. K. (2001). A physically based filter for separating base flow from streamflow time series. *Water Resour. Res.* 37, 2709–2722. doi: 10.1029/2001WR000243

Gong, W., Yang, D., Gupta, H. V., and Nearing, G. (2014). Estimating information entropy for hydrological data: one-dimensional case. *Water Resour. Res.* 50, 5003–5018. doi: 10.1002/2014WR015874

Gonzales, A., Nonner, J., Heijkers, J., and Uhlenbrook, S. (2009). Comparison of different base flow separation methods in a lowland catchment. *Hydrol. Earth Syst. Sci.* 13, 2055–2068. doi: 10.5194/hess-13-2055-2009

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Hall, F. R. (1968). Base-flow recessions-a review. *Water Resour. Res.* 4, 973–983. doi: 10.1029/WR004i005p00973

Hammond, M., and Han, D. (2006). Recession curve estimation for storm event separations. *J. Hydrol.* 330, 573–585. doi: 10.1016/j.jhydrol.2006.04.027

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2004). "Extreme learning machine: a new learning scheme of feedforward neural networks, in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, Vol. 2, 985–990 (Budapest). doi: 10.1109/IJCNN.2004.1380068

Kelleher, J. D., MacNamee, B., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA; London: MIT Press.

Klaus, J., and McDonnell, J. J. (2013). Hydrograph separation using stable isotopes: review and evaluation. *J. Hydrol.* 505, 47–64. doi: 10.1016/j.jhydrol.2013.09.006

Lyne, V., and Hollick, M. (1979). "Stochastic time-variable rainfall-runoff modelling, in *Institute of Engineers Australia National Conference* (Barton, ACT: Institute of Engineers Australia), 89–93.

Maidment, D. R., editor (1993). *Handbook of Hydrology*. New York, NY: McGraw-Hill.

Mei, Y., and Anagnostou, E. N. (2015). A hydrograph separation method based on information from rainfall and runoff records. *J. Hydrol.* 523, 636–649. doi: 10.1016/j.jhydrol.2015.01.083

Merz, R., and Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resour. Res.* 45. doi: 10.1029/2008WR007163

Merz, R., Blöschl, G., and Parajka, J. (2006). Spatio-temporal variability of event runoff coefficients. *J. Hydrol.* 331, 591–604. doi: 10.1016/j.jhydrol.2006.06.008

Minns, A. W., and Hall, M. J. (2005). "Artifical neuronal network concepts in hydrology, in *Encyclopedia of Hydrological Sciences, Vol. 1*, ed M. G. Anderson (Chichester: Wiley), 307–319. doi: 10.1002/0470848944.hsa018

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., et al. (2016). Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei science plan. *Hydrol. Sci. J.* 8, 1–17. doi: 10.1080/02626667.2016.1159683

Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: a review. *ISPRS J. Photogrammetry Remote Sens.* 66, 247–259. doi: 10.1016/j.isprsjprs.2010.11.001

Oppel, H., and Schumann, A. (2020). Machine learning based identification of dominant controls on runoff dynamics. *Hydrol. Process.* 34, 1–16. doi: 10.1002/hyp.13740

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.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* 66, 605–610. doi: 10.1093/biomet/66.3.605

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

Sharma, A., and Mehrotra, R. (2014). An information theoretic alternative to model a natural system using observational information alone. *Water Resour. Res.* 50, 650–660. doi: 10.1002/2013WR013845

Shortridge, J. E., Guikema, S. D., and Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrol. Earth Syst. Sci.* 20, 2611–2628. doi: 10.5194/hess-20-2611-2016

Solomatine, D. P., and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *J. Hydroinform.* 10, 3–22. doi: 10.2166/hydro.2008.015

Stewart, M. K. (2015). Promising new baseflow separation and recession analysis methods applied to streamflow at Glendhu catchment, New Zealand. *Hydrol. Earth Syst. Sci*. 19, 2587–2603. doi: 10.5194/hess-19-2587-2015

Su, N. (1995). The unit hydrograph model for hydrograph separation. *Environ. Int*. 21, 509–515. doi: 10.1016/0160-4120(95)00050-U

Tabari, H., Kisi, O., Ezani, A., and Talaee, P. H. (2012). SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. *J. Hydrol*. 444–445:78–89. doi: 10.1016/j.jhydrol.2012.04.007

Tallaksen, L. (1995). A review of baseflow recession analysis. *J. Hydrol*. 165, 349–370. doi: 10.1016/0022-1694(94)02540-R

Thiesen, S., Darscheid, P., and Ehret, U. (2019). Identifying rainfall-runoff events in discharge time series: a data-driven method based on information theory. *Hydrol. Earth Syst. Sci*. 23, 1015–1034. doi: 10.5194/hess-23-1015-2019

Weiler, M., Seibert, J., and Stahl, K. (2017). Magic components-why quantifying rain, snowmelt, and icemelt in river discharge is not easy. *Hydrol. Process*. 32, 160–166. doi: 10.1002/hyp.11361

Wittenberg, H., and Aksoy, H. (2010). Groundwater intrusion into leaky sewer systems. *Water Sci. Technol*. 62, 92–98. doi: 10.2166/wst.2010.287

Wolpert, D. H., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput*. 1, 67–82. doi: 10.1109/4235.585893

Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., and Tseng, H.-W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *J. Hydrol*. 552, 92–104. doi: 10.1016/j.jhydrol.2017.06.020

Zhang, J., Zhang, Y., Song, J., and Cheng, L. (2017). Evaluating relative merits of four baseflow separation methods in eastern Australia. *J. Hydrol*. 549, 252–263. doi: 10.1016/j.jhydrol.2017.04.004

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.

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/frwa.2020.00020

Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests

Utkarsh Mital[1]*, *Dipankar Dwivedi*[1], *James B. Brown*[2], *Boris Faybishenko*[1], *Scott L. Painter*[3] and *Carl I. Steefel*[1]

[1] Energy Geosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [2] Environmental Genomics and System Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [3] Climate Change Science Institute and Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States

Meteorological records, including precipitation, commonly have missing values. Accurate imputation of missing precipitation values is challenging, however, because precipitation exhibits a high degree of spatial and temporal variability. Data-driven spatial interpolation of meteorological records is an increasingly popular approach in which missing values at a target station are imputed using synchronous data from reference stations. The success of spatial interpolation depends on whether precipitation records at the target station are strongly correlated with precipitation records at reference stations. However, the need for reference stations to have complete datasets implies that stations with incomplete records, even though strongly correlated with the target station, are excluded. To address this limitation, we develop a new sequential imputation algorithm for imputing missing values in spatio-temporal daily precipitation records. We demonstrate the benefits of sequential imputation by incorporating it within a spatial interpolation based on a Random Forest technique. Results show that for reliable imputation, having a few strongly correlated references is more effective than having a larger number of weakly correlated references. Further, we observe that sequential imputation becomes more beneficial as the number of stations with incomplete records increases. Overall, we present a new approach for imputing missing precipitation data which may also apply to other meteorological variables.

Keywords: precipitation, hydrology and water, imputation, sequential imputation, machine learning, Random Forest

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Andreas Panagopoulos,
Institute of Soil and Water Resources
(ISWR), Greece
Luk J. M. Peeters,
CSIRO Land and Water, Australia

****Correspondence:***
Utkarsh Mital
umital@lbl.gov

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *15 April 2020*
Accepted: *26 June 2020*
Published: *07 August 2020*

Citation:
Mital U, Dwivedi D, Brown JB,
Faybishenko B, Painter SL and
Steefel CI (2020) Sequential
Imputation of Missing Spatio-Temporal
Precipitation Data Using Random
Forests. Front. Water 2:20.
doi: 10.3389/frwa.2020.00020

INTRODUCTION

Precipitation is an important component of the ecohydrological cycle and plays a crucial role in driving the Earth's climate. It serves as an input for various ecohydrological models to determine snowpack, infiltration, surface-water flow, groundwater recharge, and transport of chemicals, sediments, nutrients, and pesticides (Devi et al., 2015). Numerical modeling of surface flow typically requires a complete time series of precipitation along with other meteorological records (e.g., temperature, relative humidity, solar radiation) as inputs for simulations (Dwivedi et al., 2017, 2018; Hubbard et al., 2018, 2020; Zachara et al., 2020). However, meteorological records often have missing values for various reasons, such as due to malfunctioning of equipment, network interruptions, and natural hazards (Varadharajan et al., 2019). Missing values need to be reconstructed or imputed accurately to ensure that estimates of statistical properties, such as mean and co-variance, are consistent and unbiased (Schneider, 2001) because inaccurate estimates can hurt the accuracy of ecohydrological models. Reconstructing an incomplete daily precipitation time series is especially difficult since it exhibits a high degree of spatial and temporal variability (Simolo et al., 2010).

Past efforts for imputing missing values of a precipitation time series fall under two broad categories: autoregression of univariate time series and spatial interpolation of precipitation records. Autoregressive methods are self-contained and impute missing values by using data from the same time series that is being filled. Simple applications could involve using a mean value of the time series, or using data from 1 or several days before and after the date of missing data (Acock and Pachepsky, 2000). More sophisticated versions of autoregressive approaches implement stochastic methods and machine learning (Box and Jenkins, 1976; Adhikari and Agrawal, 2013). To illustrate some recent studies, Gao et al. (2018) highlighted methods to explicitly model the autocorrelation and heteroscedasticity (or changing variance over time) of hydrological time series (such as precipitation, discharge, and groundwater levels). They proposed the use of autoregressive moving average models and autoregressive conditional heteroscedasticity models. Chuan et al. (2019) combined a probabilistic principal component analysis model and an expectation-maximization algorithm, which enabled them to obtain probabilistic estimates of missing precipitation values. Gorshenin et al. (2019) used a pattern-based methodology to classify dry and wet days, then filled in precipitation for wet days using machine learning approaches (such as k-nearest neighbors, expectation-maximization, support vector machines, and random forests). However, an overarching limitation of autoregressive methods is the need for the imputed variable to show a high temporal autocorrelation, which is not necessarily valid for precipitation (Simolo et al., 2010). Therefore, such methods have limited applicability when it comes to reconstructing a precipitation time series.

Spatial interpolation methods, on the other hand, impute missing values at the target station by taking weighted averages of synchronous data, i.e., data at the same time, from reference stations (typically neighboring stations). The success of these methods relies on the existence of strong correlations among precipitation patterns between the target and reference stations. The two most prominent approaches are inverse-distance weighting (Shepard, 1968) and normal-ratio methods (Paulhus and Kohler, 1952). The inverse-distance weighting assumes the weights to be proportional to the distance from the target, while the normal-ratio method assumes the weights to be proportional to the ratio of average annual precipitation at the target and reference stations. Another prominent interpolation approach is based on kriging or gaussian processes, which assigns weights by accounting for spatial correlations within data (Oliver and Webster, 2015). Teegavarapu and Chandramouli (2005) proposed several improvements to weighting methods and also introduced the coefficient of correlation weighting method—here the weights are proportional to the coefficient of correlation with the target. Recent studies have proposed new weighting schemes using more sophisticated frameworks (e.g., Morales Martínez et al., 2019; Teegavarapu, 2020). In parallel, studies have also been conducted to account for various uncertainties in imputation. For example, Ramos-Calzado et al. (2008) proposed a weighting method to account for measurement uncertainties in a precipitation time series. Lo Presti et al. (2010) proposed a methodology to approximate each missing value by a distribution of values where each value in the distribution is obtained via a univariate regression with each of the reference stations. Simolo et al. (2010) pointed out that weighting approaches have a tendency to overestimate the number of rainy days and to underestimate heavy precipitation events. They addressed this issue by proposing a spatial interpolation procedure that systematically preserved the probability distribution, long-term statistics, and timing of precipitation events.

A critical review of the literature shows that, in general, spatial interpolation techniques have two fundamental shortcomings: (i) how to optimally select neighbors, i.e., reference stations, and (ii) how to assign weights to selected stations. While selecting reference stations is typically done using statistical correlation measures, assigning weights to selected stations is currently an ongoing area of research. The methods reviewed so far are based on the idea of specifying a functional form of the weighting relationships. The appropriate functional form may vary from one region to another depending on the prevalent patterns of precipitation as influenced by local topographic and convective effects. Using a functional form that is either inappropriate or too simple could distort the statistical properties of the datasets (such as mean and covariance). Some researchers have proposed to address these shortcomings by using Bayesian approaches (e.g., Yozgatligil et al., 2013; Chen et al., 2019; Jahan et al., 2019). These fall under the broad category of expectation-maximization and data augmentation algorithms, thus yielding a probability distribution for each missing value.

An alternative approach for imputing missing data is the application of data-driven or machine learning (ML) methods which are becoming increasingly prominent for imputing using spatial interpolation. These methods do not need a functional form to be specified *a priori* and can learn a multi-variate relationship between the target station and reference stations using available datasets. Studies have found that the performance of ML methods tends to be superior to that of traditional weighting methods (e.g., Teegavarapu and Chandramouli, 2005; Hasanpour Kashani and Dinpashoh, 2012; Londhe et al., 2015). In addition, studies have been conducted to identify an optimal architecture for ML-based methods (Coulibaly and Evora, 2007; Kim and Pachepsky, 2010). In this work, we use a Random Forests (RF) method. The RF is an ensemble learning method which reduces associated bias and variance, making predictions less prone to overfitting. In addition, a recent study showed that RF-based imputation is generally robust, and performance improves with increasing correlation between the target and references (Tang and Ishwaran, 2017).

Regardless of the imputation technique, an inherent limitation of spatial interpolation algorithms is the need for reference stations to have complete records during the time-period of interest. This limitation is critical for ML algorithms where incomplete records preclude data-driven learning of multi-variate relationships. The success of spatial interpolation, therefore, depends on whether precipitation at the target station is highly correlated with precipitation at stations with complete records. A station with an incomplete record is typically excluded from the analysis even though that station may have a high correlation with the target station. In this work, we hypothesize that stations with incomplete records contain information that can improve spatial interpolation if they are included in the analysis. We propose a new algorithm, namely sequential imputation, that leverages incomplete records to impute missing values. In this approach, stations that are imputed first are also included as reference stations for imputing subsequent stations. We implement this algorithm in the context of imputing missing daily values of precipitation and demonstrate its benefits by incorporating it in an RF-based spatial interpolation.

In what follows, we start by describing our study area and data sources and follow this with a brief introduction to the Random Forests (RF) method. We then describe all our numerical experiments, starting with a baseline imputation that helps evaluate the performance of sequential imputation. This is followed by a description of the sequential imputation algorithm, along with an outline of different scenarios to evaluate sequential imputation. We compare the results of sequential imputation with a non-sequential imputation in which incomplete records are not leveraged for subsequent imputations. Finally, we discuss the implications of our results and provide some concluding thoughts.

METHODOLOGY

Study Area and Data Sources

We conducted this study using data from the Upper Colorado Water Resource Region (UCWRR), which is one of 21 major water resource regions classified by the United States Geological Survey to divide and sub-divide the United States into successively smaller catchment areas. The UCWRR is the principal source of water in the southwestern United States and includes eight subregions, 60 sub-basins, 523 watersheds, and 3,179 sub-watersheds. Several agencies have active weather monitoring stations in UCWRR. For our study, we considered the weather stations maintained by the Natural Resources Conservation Service (NRCS). **Table 1** summarizes the various networks that comprise the NRCS database.

TABLE 1 | Summary of NRCS stations in UCWRR.

Network	# Of stations	# Of complete records	# Of incomplete records
SNOTEL	134	94	40
SCAN	12	1	11
ACIS	5	2	3
SNOLITE	1	0	1
All	152	97	55

SNOTEL: Snowpack Telemetry; SCAN: Soil Climate Analysis Network; ACIS: Applied Climate Information System; SNOLITE: SNOTEL with Iridium Satellite System.

Figure 1 shows the spatial distribution of NRCS stations in UCWRR. Ninety-seven stations have complete records which primarily belong to the Snowpack Telemetry (SNOTEL) network. We considered data spanning the 10-year window from 2008 to 2017. Over this period, NRCS had 152 active stations in UCWRR which report daily precipitation data. For this study, our dataset is restricted to the 97 stations with complete records. We downloaded the data through the NRCS Interactive Map and Report Generator[1] (accessed Jan 16, 2020).

Spatial Interpolation Method: Random Forests (RF)

RF is an ML-method based on an ensemble or aggregation of decision-trees (Breiman, 2001). A decision-tree is a flowchart-like structure that recursively partitions the input feature space into smaller subspaces (**Figure 2**). Recursion is carried out till the subspaces are small enough to fit simple linear models on them In regression problems, the decision rules for partitioning are determined such that the mean-squared error between the tree output and the observed output is minimized. The RF model trains each decision-tree on a different set of data points obtained by sampling the training data with replacement (or bootstrapping). Furthermore, each tree may also consider a different subset of input features selected randomly. The final output of the random forest is obtained by aggregating (or ensembling) the results of all decision trees. For regression problems, aggregation is done by taking the mean. **Figure 2** shows a schematic of an RF regressor.

The ensemble nature of RF leads to several benefits (Breiman, 2001; Louppe, 2015). First, it makes RF less prone to overfitting, despite the susceptibility of individual trees to overfitting (Segal, 2004). For regression problems, overfitting refers to low values of mean-squared error on training data, and high values of mean-squared error on test data. Second, it enables an evaluation of the relative importance of a variable (which, in this work, refers to a reference station) for predicting the output. This is typically done by determining how often a variable is used for partitioning the input feature space, across all trees. Third, the ensemble nature of RF makes it possible to not set aside a test set. Since the input for each decision tree is obtained by bootstrapping, the unsampled data can be used to estimate the generalization error. In addition, RF does not require extensive hyperparameter tuning compared to other ML approaches (Ahmad et al., 2017).

In this study, we implement RF using Python's *scikit-learn* module (Pedregosa et al., 2011). Precipitation data from reference stations acts as input, and precipitation data at the target station is specified as the output. Unlike typical spatial interpolation approaches, we do not specify distances between the reference and target stations. Distances are static variables and their influence on dynamic precipitation relationships gets learnt as a constant bias, regardless of whether they are explicitly specified or not.

[1] https://www.wcc.nrcs.usda.gov

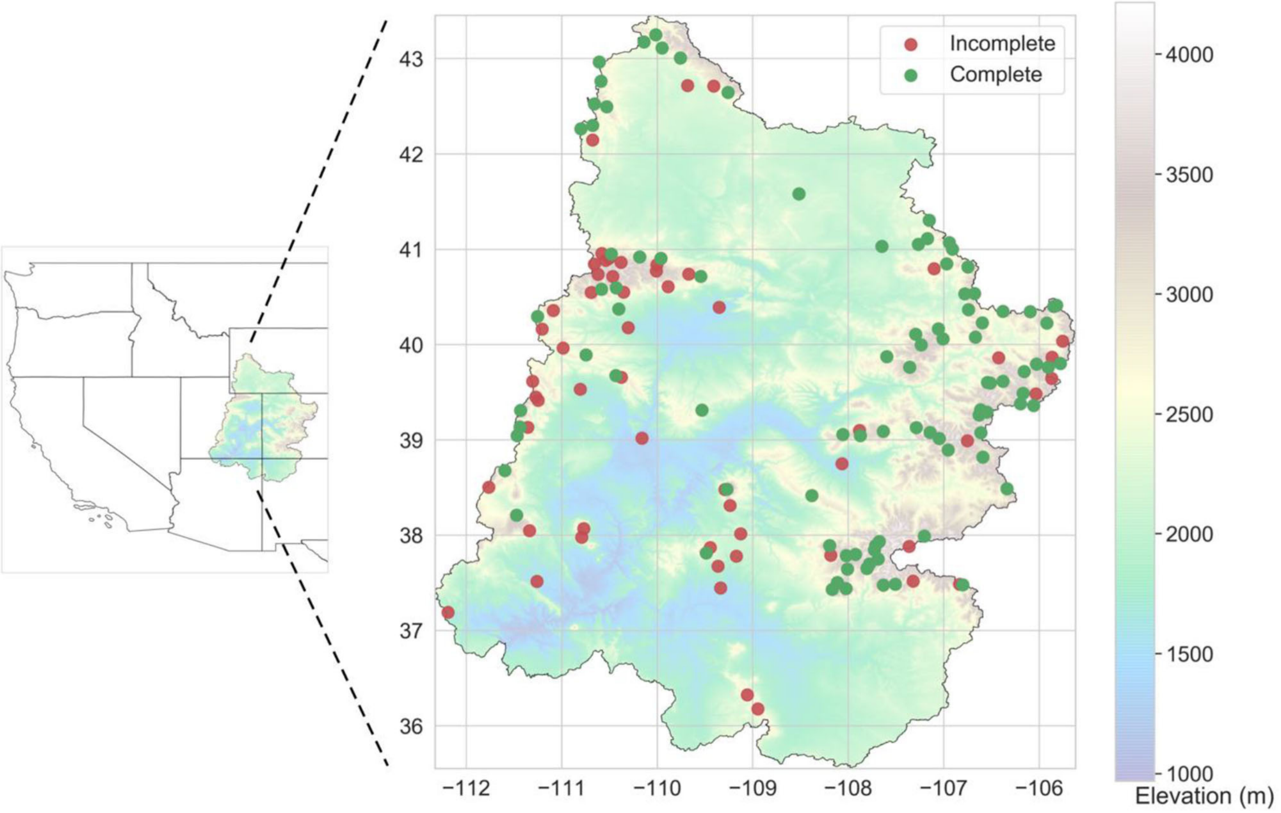


FIGURE 1 | Spatial extent of UCWRR, along with the layout of stations in the NRCS database (comprising of 97 complete and 55 incomplete records).

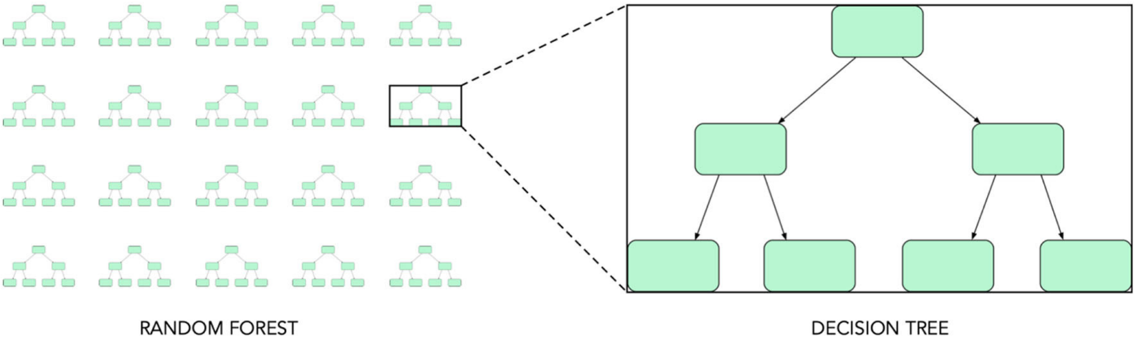


FIGURE 2 | Schematic of a Random Forest regressor, adapted from Stockman et al. (2019).

Overview of Numerical Experiments

To investigate if stations with incomplete records contain information that can improve spatial interpolation, we designed three sets of numerical experiments: baseline, sequential, and non-sequential imputation. In baseline imputation, each station in our dataset is modeled using the remaining stations as reference stations. This represents an upper bound on the performance of sequential imputation when we have multiple stations with incomplete records. The baseline imputation provides statistics to help evaluate the performance of sequential imputation. In sequential imputation, a subset of stations in our dataset is marked as artificially incomplete. For each station in the artificially incomplete subset, 20% of the values are randomly marked as "missing." The missing values are imputed by leveraging other artificially incomplete stations in the subset, in addition to using stations outside the subset. Finally, in non-sequential imputation, the same artificially incomplete subset as sequential imputation is considered, and missing values are imputed using just the stations that are outside the subset. We describe the three sets of numerical experiments in detail in sections Numerical Experiments: Baseline Imputation and Numerical Experiments: Sequential and Non-sequential Imputation. Before describing each of these experiments, it would be instructive to discuss our performance criterion for evaluating imputation.

Evaluating Imputation: Nash-Sutcliffe Efficiency (NSE)

We evaluated the overall performance of imputation by computing the Nash-Sutcliffe Efficiency (*NSE*) on test data given by

$$NSE = 1 - \frac{\sum_{i=1}^{N} \left(y_i^o - y_i^m\right)^2}{\sum_{i=1}^{N} \left(y_i^o - \overline{y^o}\right)^2} \tag{1}$$

where N is the size of the test set, y_i^o is i-th observed value, y_i^m is the corresponding modeled value, and $\overline{y^o}$ is the mean of all observed values in the test set.

The *NSE* is a normalized statistical measure that determines the relative magnitude of the residual variance (or noise) of a model when compared to the measured data variance. It is dimensionless and ranges from $-\infty$ to 1. An *NSE* value equal to 1 implies that the modeled (in our case, imputed) values perfectly match the observations; an *NSE* value equal to 0 implies that the modeled values are only as good as the mean of observations; and a negative *NSE* value implies that the mean of observations is a better predictor than modeled values. Positive *NSE* values are desirable, and higher values imply greater accuracy of the (imputation) model.

Two other common statistical measures for evaluating the overall accuracy of prediction are Pearson's product-moment correlation coefficient *R*, and the Kolmogorov-Smirnov statistic. While the former evaluates the timing and shape of the modeled time series, the latter evaluates its cumulative distribution. Gupta et al. (2009) decomposed the *NSE* into three distinctive components representing the correlation, bias, and a measure of relative variability in the modeled and observed values. They showed that *NSE* relates to the ability of a model to reproduce the mean and variance of the hydrological observations, as well as the timing and shape of the time series. For these reasons, the use of *NSE* was preferred over other statistical measures to evaluate the accuracy of imputation.

We also evaluated the performance of sequential imputation for predicting dry events and extreme wet events. This is because spatial interpolation approaches tend to overpredict the number of dry events and underestimate the intensity of extreme wet events (Simolo et al., 2010; Teegavarapu, 2020). A common practice is to consider a day as a dry event if the daily precipitation does not exceed a threshold of 1 mm (Hertig et al., 2019). We considered a threshold of 2.54 mm since that is the resolution of our dataset. We considered a day as an extreme wet event if the daily precipitation exceeded the 95th percentile of the entire precipitation record for a given station (Zhai et al., 2005; Hertig et al., 2019). To evaluate prediction accuracy for dry events, we computed the percentage error, or the percentage of days that were correctly modeled as dry days. To evaluate prediction accuracy for extreme wet events, we computed *NSE* values exclusively for days that exceeded the 95th percentile of daily precipitation values; this enabled us to evaluate the predicted magnitude. In what follows, we use the acronym *NSEE* to denote *NSE* for extreme events.

Numerical Experiments: Baseline Imputation

For our first set of numerical experiments, we conducted baseline imputations where each station in our dataset is modeled using the remaining stations as reference stations. Our dataset consists of 97 stations with complete records (as outlined in **Figure 1** and **Table 1**). This set of numerical experiments is a test of the RF-based imputation method and provides an upper bound on the performance of the sequential imputation algorithm discussed in the section Sequential Imputation Algorithm. More importantly, it provides estimates of the variance for modeling each station, which will be used to evaluate the performance of the sequential imputation algorithm. Specifically, each station in our dataset was considered, in turn, to be a target station (or model output), with the rest of the stations acting as references (or input features). For each target station, 80% of the data were randomly selected for training, and the remaining 20% were used for testing. The test set effectively acted as missing data to be imputed. We conducted this exercise 15 times for each station. Prior to these runs, we also conducted an independent set of baseline runs to tune the hyperparameters of RF.

Sequential Imputation Algorithm

ML-based spatial interpolation learns multi-variate relationships between the reference stations and the target station. Studies have noted that for imputation results to be reliable, data at reference stations should be strongly correlated to data at the target station (e.g., Teegavarapu and Chandramouli, 2005; Yozgatligil et al., 2013). However, ML-based spatial interpolation excludes stations that have incomplete records, even though they may be strongly correlated with the target station. Here, we develop a technique (i.e., sequential imputation) where stations that are imputed first are used as reference stations for imputing subsequent stations. In what follows, we refer to a station with a complete record as a "complete station," and a station with an incomplete record as an "incomplete station." The sequential imputation algorithm involves the following steps:

1. Add all complete stations to the list of reference stations.
2. Calculate correlations between incomplete stations and reference stations.
3. Pick the incomplete station having the highest aggregate correlation with reference stations.
4. Impute missing values for the station picked in Step 3, using all the reference stations.
5. Add the imputed station to the list of reference stations.
6. Repeat steps 2–4 till missing values of all the stations are imputed.

In this study, correlation refers to Pearson's product-moment correlation coefficient, hereafter denoted by *R*. We chose this measure for its simplicity. Step 3 requires calculating an aggregate correlation of each incomplete station with the reference stations. This step assumes that the incomplete station having the highest aggregate correlation with reference stations will have the most accurate imputation. We will verify this assumption in the Results section. To determine an appropriate aggregate correlation measure for Step 3, we implemented the following procedure:

i. Compute correlations of a target station with each of the reference stations.
ii. Sort the correlation values in descending order (highest to lowest).
iii. Calculate the cumulative sum of the sorted correlations. Denote each partial sum as S_i, where subscript i refers to the first i sorted correlations.

i varies from 1 to N, and N is the number of reference stations in the dataset. Each S_i is an aggregate measure of correlation between a target station and the reference stations. For instance, S_2 refers to the sum of first two sorted correlations, S_3 refers to the sum of first three sorted correlations, and so on. We computed values of S_i for all the 97 stations in our dataset and compared their values with *NSE* determined from baseline imputations. The S_i having the highest correlation with *NSE* was picked to quantify aggregate correlation (for Step 3 of sequential imputation). For practical applications, the above procedure to determine an appropriate aggregate correlation may be implemented using non-sequential imputations. Note that other aggregate measures may be envisioned (e.g., mutual information, spearman's correlation), but we sought to pick one that is relatively simple to keep our focus on the sequential imputation approach.

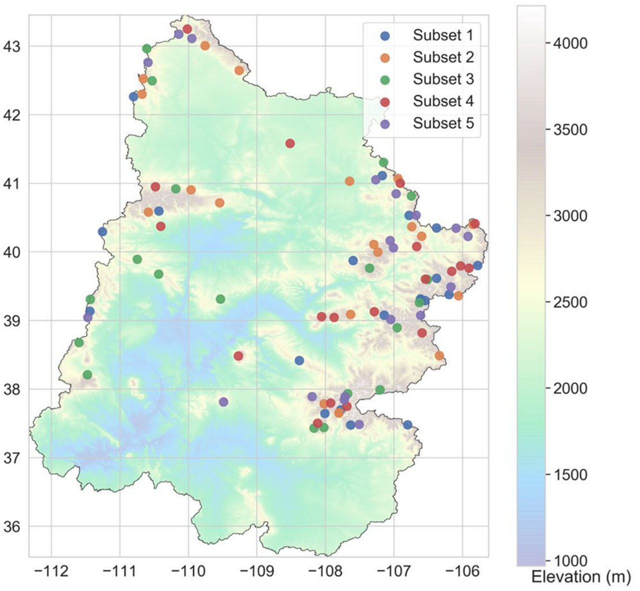


FIGURE 3 | Division of complete stations (see **Figure 1**) into five subsets.

TABLE 2 | Scenarios for sequential and non-sequential imputation.

	Artificially incomplete subsets	Complete subsets
Scenario 1	2, 3, 4, 5	1
Scenario 2	3, 4, 5	1, 2
Scenario 3	4, 5	1, 2, 3
Scenario 4	5	1, 2, 3, 4

Numerical Experiments: Sequential and Non-sequential Imputation

To investigate the benefits of sequential imputation, we divided our dataset of 97 complete stations into five (almost) evenly sized subsets and labeled them 1 through 5, as shown in **Figure 3**. The division into subsets was random. We then considered four different scenarios, each of which marked certain subsets as artificially incomplete. These are shown in **Table 2**.

Precipitation records typically have missing values resulting from random mechanisms such as malfunctioning of equipment, network interruptions, and natural hazards. In other words, the probability that a precipitation value is missing does not depend on the value of precipitation itself. These random mechanisms also assume that the location or physiography of a weather station has no bearing on whether its record is complete or incomplete. This *missing at random* mechanism (Schafer and Graham, 2002) is reflected in our decision to create subsets randomly, and enables us to evaluate the sequential imputation approach in a more generic setting.

Figures 4A–D shows the division of our dataset into complete and artificially incomplete subsets for each of the scenarios listed in **Table 2**. Scenario 1 had 77 out of 97 records marked as artificially incomplete. Each subsequent scenario had fewer records marked as artificially incomplete, culminating with Scenario 4 which had only 19 such records. These scenarios were designed to investigate how the proportion of incomplete records affects imputation. We expected sequential imputation to be more beneficial as the proportion of incomplete records increased in the dataset.

The stations belonging to the artificially incomplete subsets had 20% of their data marked as missing. Previous studies on imputation have considered two broad mechanisms for marking missing values. One approach involves marking missing values randomly (e.g., Teegavarapu and Chandramouli, 2005; Kim and Pachepsky, 2010), while the other approach assumes that missing values form continuous gaps in time (e.g., Simolo et al., 2010; Yozgatligil et al., 2013). Since spatial interpolation assumes no temporal autocorrelation and is agnostic to the timestamp of the data, the mechanism for marking missing values is not relevant. For simplicity, we assumed that values were missing completely at random. The missing values were imputed using sequential and non-sequential imputations; both these imputations were compared and enabled us to highlight the benefits of sequential imputation. Specifically, we calculated *NSE* corresponding to both sequential and non-sequential runs and computed the change (or increase) Δ in *NSE* for each station as follows:

$$\Delta NSE = NSE_{\text{sequential}} - NSE_{\text{non-sequential}} \qquad (2)$$

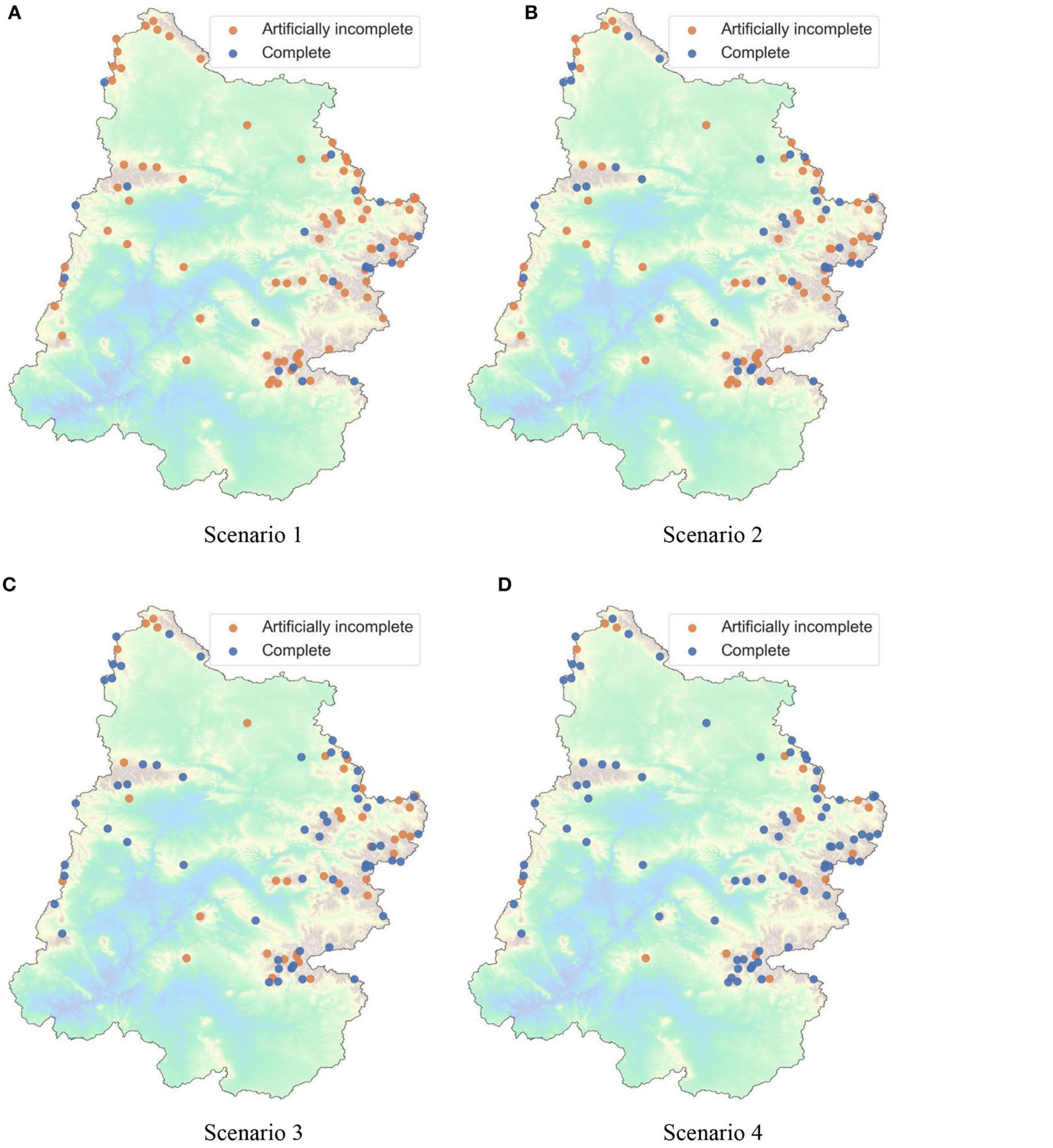


FIGURE 4 | Artificially incomplete and complete datasets for different scenarios of sequential and non-sequential imputation. Note: Colormap for elevation is same as **Figures 1**, **3**. **(A)** Scenario 1. **(B)** Scenario 2. **(C)** Scenario 3. **(D)** Scenario 4.

To evaluate improvement in prediction of extreme wet events, *NSE* in **Equation 2** was replaced by *NSEE*. To evaluate improvement in prediction of dry days, we computed the percentage error (i.e., the percentage of days that were correctly modeled as dry days) corresponding to both sequential and non-sequential runs. We then computed the change (or decrease) Δ in percentage error (*PE*) as follows:

$$\Delta PE = PE_{\text{non-sequential}} - PE_{\text{sequential}} \tag{3}$$

RESULTS

Baseline Imputation

We performed baseline imputation to estimate statistics to evaluate the performance of the sequential imputation algorithm. **Figures 5A–C** show results of baseline imputations on missing data for all stations. Each station was modeled 15 times, with different splits of training and testing (missing) data, and the accuracy of each model for imputation was quantified by computing *NSE* on test data. This provided us with a distribution

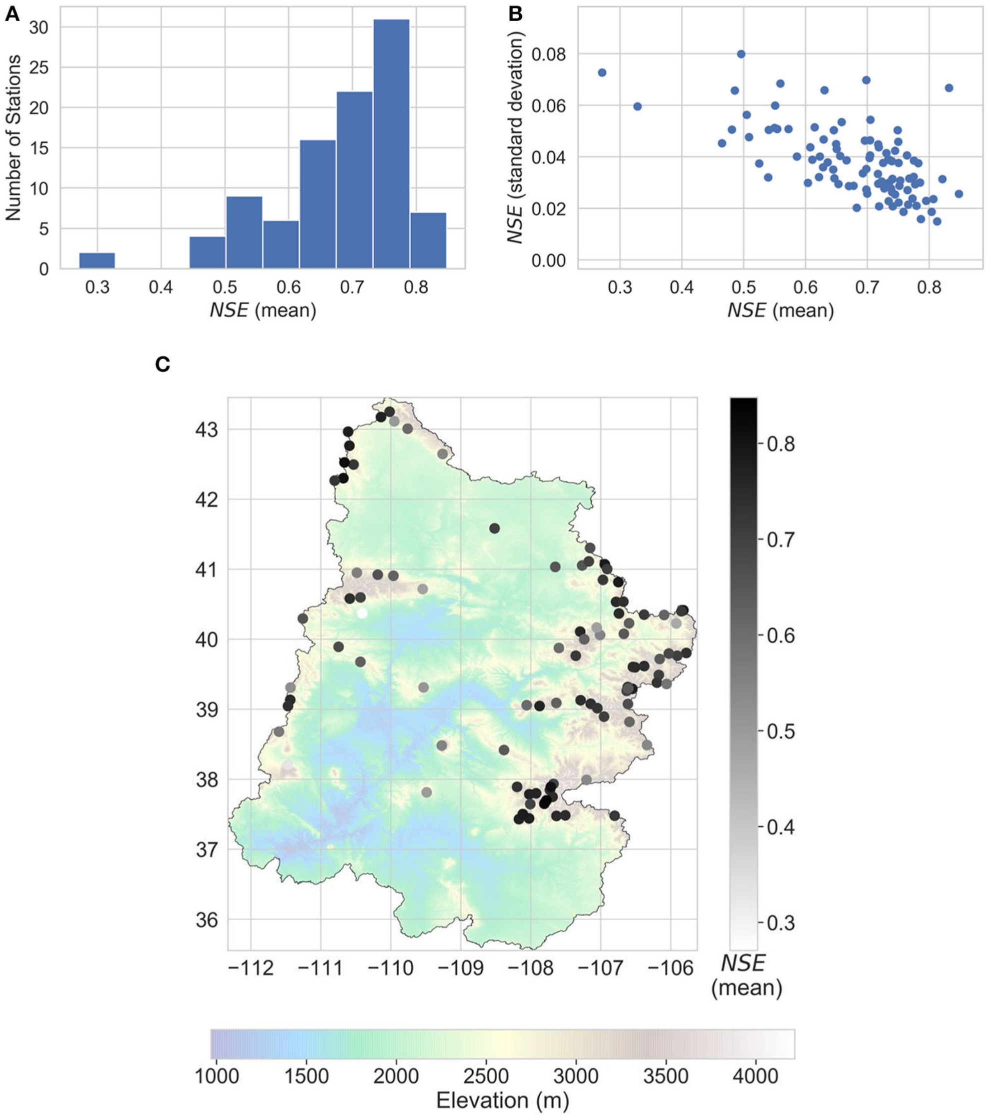


FIGURE 5 | Results of baseline imputations on missing data. **(A)** Distribution of mean *NSE* (μ_s), **(B)** scatter of mean (μ_s), and standard deviation (σ_s) of *NSE*, **(C)** geospatial distribution of mean *NSE* (μ_s).

of *NSE* values (instead of just one value) for reconstructing each station, from which we estimated the mean μ and standard deviation σ of *NSE* for each station. For clarity, we denote the mean and standard deviation of a particular station *s*, by μ_s and σ_s, respectively. **Figure 5A** compiles the μ_s for all the stations and shows them as a histogram. Approximately 95% of the stations have a mean *NSE* >0.5, and approximately two-thirds of the stations have a mean *NSE* >0.65. **Figure 5B** compiles the μ_s and σ_s for all stations and shows them as a scatter plot. We see that for each station, the *NSE* values have a small standard deviation relative to their mean. **Figure 5C** shows the geospatial distribution of μ_s.

Figure 6 shows sample scatter plots of true and predicted precipitation on test data using baseline imputations. The dotted line shows the 45-degree line which corresponds to a perfect match (i.e., *NSE* = 1) between true and predicted values. Note that our dataset has a resolution of 0.1 inch or 2.54 mm, which results in visible jumps in the abscissa (or "true values"). Subfigure (a) corresponds to a relatively high value of *NSE* (~0.8), and subfigure (b) corresponds to a relatively low value of *NSE* (~0.5). We see from these plots that for a high value of *NSE*, the relative scatter is smaller and closer to the dotted line.

Aggregate Correlation Between Target Incomplete Stations and Reference Stations

To identify an appropriate aggregate correlation measure for sequential imputation, we analyzed results of baseline imputations. Specifically, we computed values of S_i for all the target stations (i.e., S_i^s) and compared their values with the corresponding μ_s. Since strong correlations with reference stations lead to more accurate imputation, we expect S_i to be positively correlated with μ, regardless of the value of i. As defined in the section Sequential Imputation Algorithm, S_i for a target station is the sum of first i sorted correlations with reference stations. For clarity, we denote S_i^s to refer to S_i for a particular target station s. **Figure 7A** shows a scatter plot of S_2^s and μ_s for all the stations in our dataset (as outlined in **Figure 1** and **Table 1**). The correlation coefficient was 0.95. Similarly, we computed correlations between S_i^s and μ_s for all values of i [denoted as Corr(μ_s, S_i^s)], and plotted them in **Figure 7B**. These results show that the correlation between S_i^s and μ_s is higher for lower values of i. On the basis of **Figure 7**, we used S_2 as the similarity measure for sequential imputation. For practical applications, an appropriate

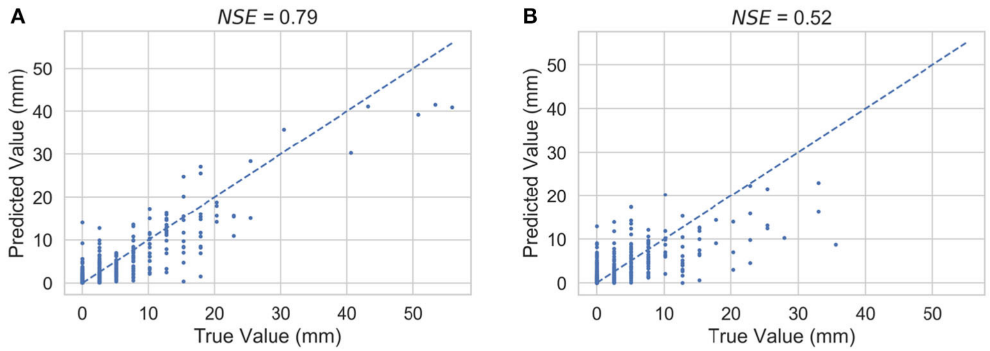


FIGURE 6 | Sample scatter plots of true and predicted precipitation on test data using baseline imputations: **(A)** *NSE* = 0.79, **(B)** *NSE* = 0.52. Note that the jumps in true values are due to the coarse resolution (of 2.54 mm) of the dataset.

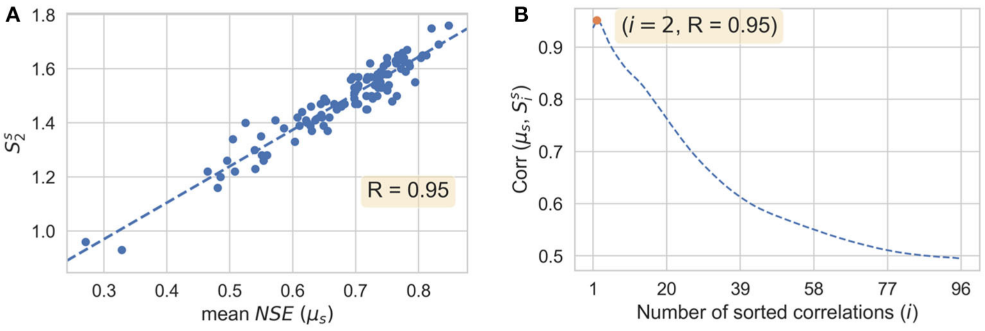


FIGURE 7 | Quantifying similarity between target and reference stations: **(A)** scatter plot between S_2^s and μ_s (mean value of *NSE*) with a linear fit, **(B)** correlations between μ_s and S_i^s as a function of *i* (annotation: maximum value of the correlation).

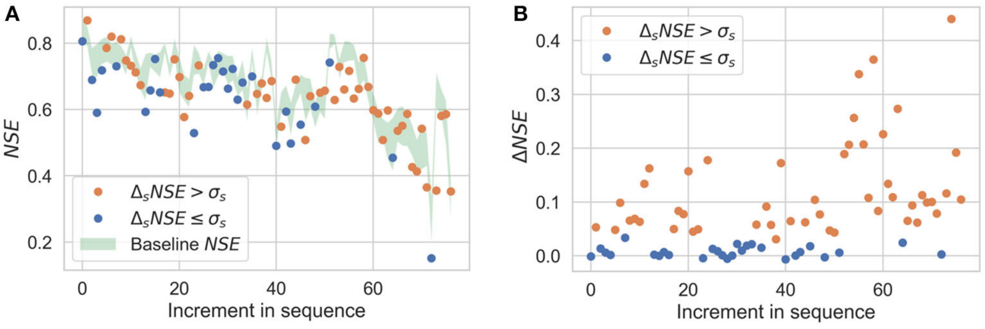


FIGURE 8 | Sequential imputation results for Scenario 1. **(A)** *NSE* obtained during sequential imputation plotted as a function of increment in sequence and superimposed over baseline *NSE* values, **(B)** Change Δ*NSE* for each increment in sequence, when compared to a non-sequential imputation. Orange dots are considered significant improvements (i.e., $\Delta_s NSE > \sigma_s$).

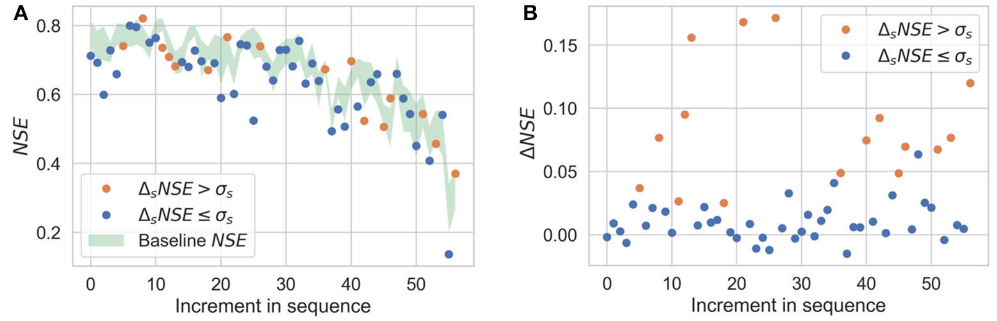


FIGURE 9 | Sequential imputation results for Scenario 2; captions of **(A,B)** are same as in **Figure 8**.

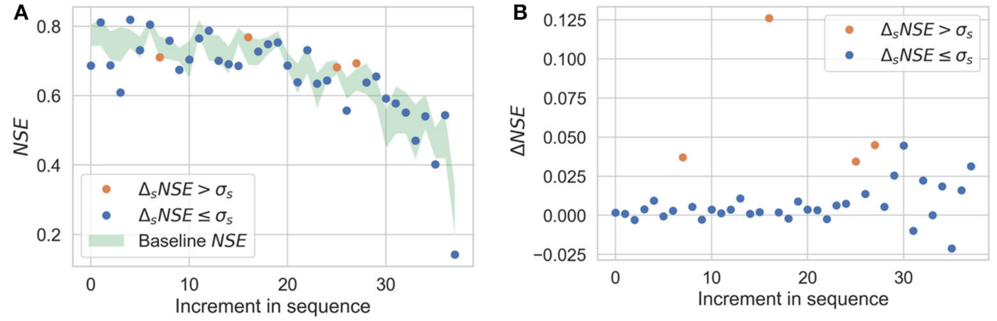


FIGURE 10 | Sequential imputation results for Scenario 3; captions of **(A,B)** are same as in **Figure 8**.

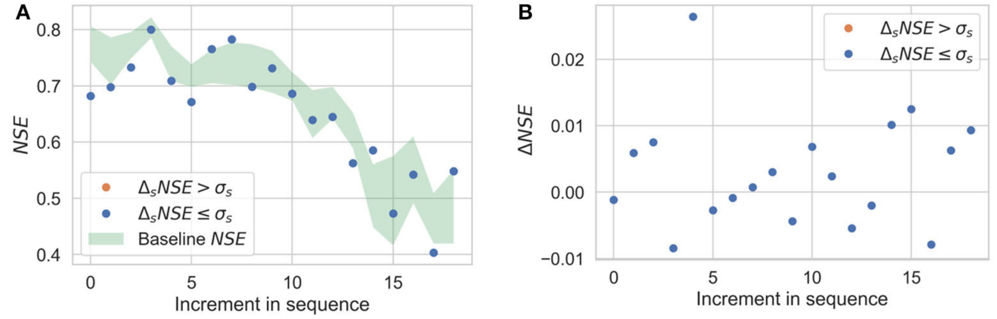


FIGURE 11 | Sequential imputation results for Scenario 4; captions of **(A,B)** are same as in **Figure 8**.

similarity measure may be determined by analyzing results of non-sequential imputations.

Sequential Imputation

To implement the sequential imputation algorithm, the artificially incomplete subsets in each of the four scenarios were reconstructed using sequential and non-sequential imputation (see section Numerical Experiments: Sequential and Non-sequential Imputation). For a given station, sequential imputation was considered to have made a significant improvement if the corresponding $\Delta_s NSE$ (i.e., ΔNSE for station s computed using Equation 2) was greater than σ_s estimated from baseline runs. This was done to ensure that the change in *NSE* during sequential imputation may not be attributed to noise.

Figures 8A–11A show the results of sequential imputation for Scenarios 1–4, respectively, with values of *NSE* for each station corresponding to sequential imputation. The values are plotted in the order of sequential imputation and are superimposed over the baseline values of *NSE*. The baseline *NSE* curve is centered at its mean and the thickness represents its standard deviation (as shown in **Figure 5B**). The baseline curve provides an upper bound on the performance of the sequential imputation algorithm. **Figures 8B–11B** show change in NSE for each increment in sequence, when compared to a non-sequential imputation.

Results for the scenarios are summarized in **Table 3**.

Figure 12 shows scatter plots of true and predicted precipitation on test data for a station that showed significant improvement during sequential imputation in Scenario 1. Subfigure (a) shows the scatter for non-sequential imputation, and subfigure (b) shows the scatter for sequential imputation. The dotted line shows the 45-degree line which corresponds to a perfect match (i.e., $NSE = 1$) between true and predicted values. Recall that our dataset has a resolution of 0.1 inch or 2.54 mm, which results in visible jumps in the abscissa (or "true values").

Figures 13, **14** show the results of sequential imputation for predicting dry [subfigures (a)] and extreme wet [subfigures (b)] events for Scenarios 1, 2. The values are plotted in the order of sequential imputation and denote the change in *PE* or *NSEE* during sequential imputation when compared to a non-sequential imputation. The Δ values are color-coded according to results of **Figures 8–11**. The results for Scenarios 3, 4 are not shown for the sake of brevity.

TABLE 3 | Summary of results for Scenarios 1–4 for sequential and non-sequential imputation.

	# Of imputed stations	# Of stations where $\Delta_s NSE > \sigma_s$
Scenario 1	77	49
Scenario 2	57	16
Scenario 3	38	4
Scenario 4	19	0

DISCUSSION

Figure 5A shows the mean *NSE* (μ_s) for all the stations as a histogram. As noted earlier, approximately 95% of the stations have $\mu_s > 0.5$, and approximately two-thirds of the stations have a $\mu_s > 0.65$. Moriasi et al. (2007) reviewed over twenty studies related to watershed modeling and recommended that for a monthly time step, models can be judged as "satisfactory" if *NSE* is >0.5; a lower threshold was recommended for daily time steps. Therefore, our spatial interpolation technique for imputing missing values can be considered to be effective.

The geospatial distribution of mean *NSE* in **Figure 5C** suggests that lower values of *NSE* tend to arise when there is a lower density of reference stations in close proximity. This is because distant stations tend to experience dissimilar precipitation patterns than the target station, making them less likely to be reliable predictors of precipitation at the target

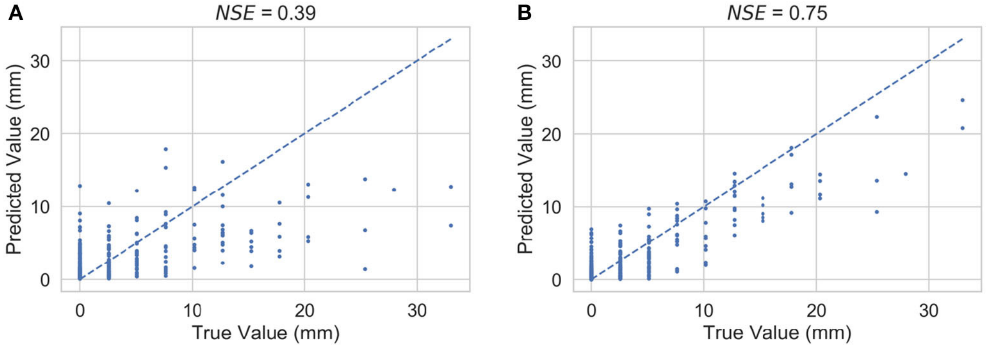


FIGURE 12 | Scatter plots of true and predicted precipitation on test data for a station that showed significant improvement during sequential imputation in Scenario 1: **(A)** non-sequential imputation, **(B)** sequential imputation. Jumps in true values are due to the coarse resolution (of 2.54 mm) of the dataset.

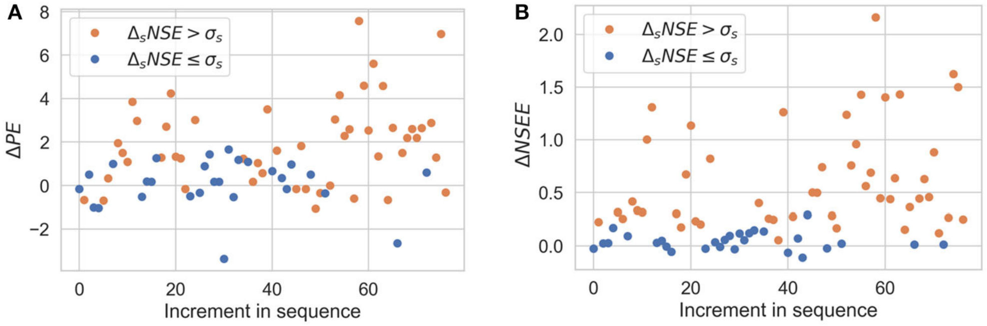


FIGURE 13 | Sequential imputation results for predicting dry **(A)** and extreme wet **(B)** events for Scenario 1. Orange dots correspond to significant improvements in overall predictions as shown in **Figure 8**.

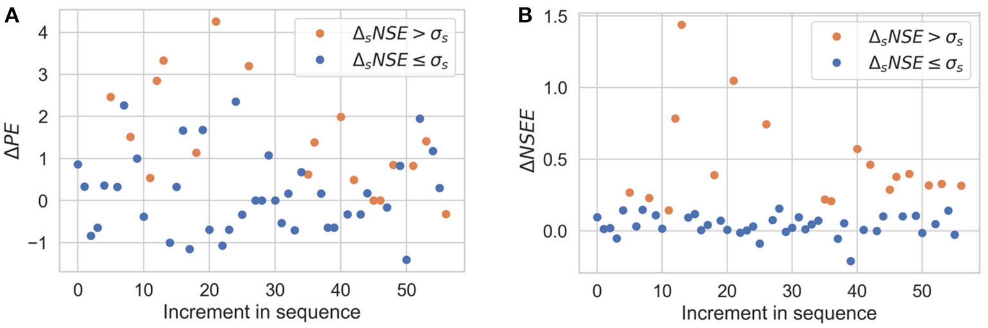


FIGURE 14 | Sequential imputation results for predicting dry **(A)** and extreme wet **(B)** events for Scenario 2. Orange dots correspond to significant improvements in overall predictions as shown in **Figure 9**.

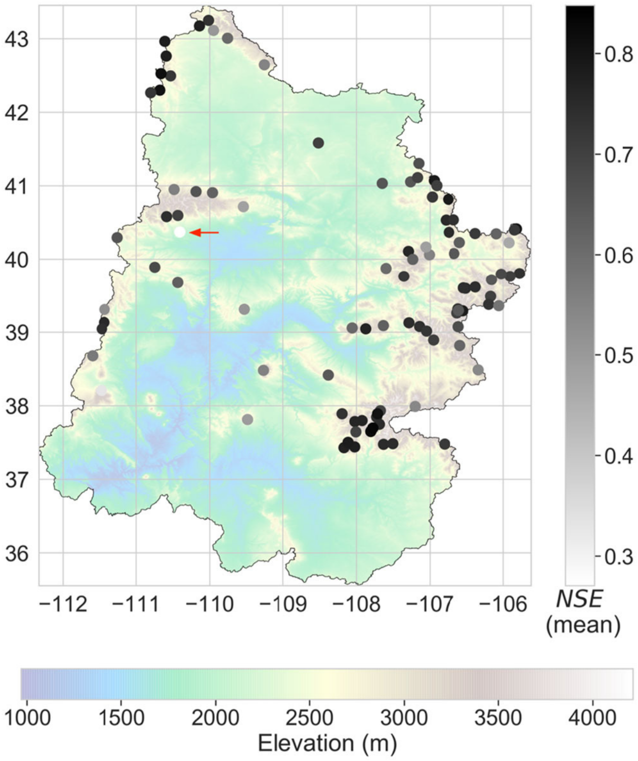


FIGURE 15 | Geospatial distribution of mean *NSE* (μ_s) with a red arrow marking a station that has a low *NSE*.

station. This observation is why the inverse-distance weighting method is popular.

Although proximity of reference stations may be considered necessary for accurate imputation of precipitation values, it is not sufficient (e.g., Teegavarapu and Chandramouli, 2005). We show an example of this in **Figure 15**, which is a modified version of **Figure 5C** with an arrow marking a station. The marked station has a low *NSE* despite having reference stations that exist in close proximity. This is because the reference stations closest to it have significantly different values of elevation (for reference, the marked station has an elevation of 2,113 m, while the closest station has an elevation of 3,085 m). For accurate spatial interpolation at a target location, the reference stations should have physiographic similarity with the target. Factors influencing physiographic similarity are location, elevation, coastal proximity, topographic facet orientation, vertical atmospheric layer, topographic position, and orographic effectiveness of the terrain (Daly et al., 2008). Note that it is not known a priori how these different factors interact with each other and subsequently influence the physiographic properties of target and reference stations. Selecting reference stations based on predefined physiographic criteria may result in an unintentional exclusion of stations that have a high correlation with the target station. Overall, any predefined physiographic criterion will lack the flexibility in selecting stations and may not result in the best imputation performance.

Figure 6 shows sample scatter plots of true and predicted precipitation on test data using baseline imputations. We see from these plots that for a high value of *NSE*, the relative scatter is smaller. In addition, we can also observe that even for a high value of *NSE*, there is a tendency to overpredict the number of dry days and underestimate the intensity of extreme wet events. For subfigure (a), the 95th percentile threshold is at 15.24 mm, and for subfigure (b), it is at 12.7 mm. Recall that we define events beyond the 95th percentile threshold as extreme wet events.

Figures 8–11 demonstrate the benefits of sequential imputations when compared with non-sequential imputations. In what follows, we will use the phrase "incomplete station" to refer to an artificially incomplete station. **Figures 8–11** show that as the proportion of incomplete stations increases, there is a higher percentage of stations benefitting from sequential imputation. ΔNSE values that correspond to significant improvements (i.e., $\Delta_s NSE > \sigma_s$) tend to be higher than those that do not. A value of ΔNSE that does not correspond to a significant improvement (i.e., $\Delta_s NSE \leq \sigma_s$) implies that the previously imputed stations do not add extra information for spatial interpolation. This can be for two reasons: (i) the previously imputed stations are weakly correlated to the target station, or (ii) the previously imputed stations show strong correlations with the target station, but also show strong correlations with stations already in the complete subset. The

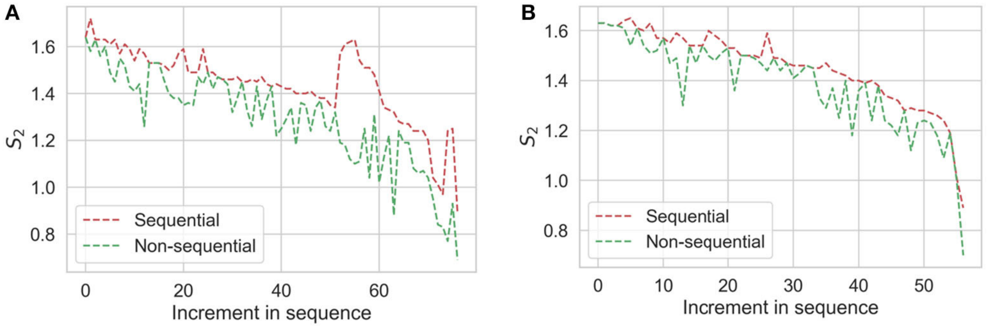


FIGURE 16 | Comparisons between S_2 for sequential and non-sequential imputations: **(A)** Scenario 1 and **(B)** Scenario 2.

second reason could happen if there is a cluster of stations that have similar physiography and experience similar precipitation patterns. Sequential imputation of stations in a cluster may not add new information if other stations in the cluster already have complete records. For instance, consider Scenario 4 where the proportion of incomplete stations is small and sequential imputation does not provide any benefits. **Figure 4D** shows that the incomplete stations in Scenario 4 are either isolated (and could be weakly correlated to other incomplete stations) or are a part of a cluster with multiple complete records. **Figures 3**, **4** show that the stations in our dataset tend to form clusters; these figures help us understand why we observe a smaller percentage of stations benefitting from sequential imputation as the proportion of incomplete stations decreases. The clustering tendency implies that when there is a small subset of incomplete stations, there is a high probability that previously imputed stations do not add any extra information for spatial information.

Figure 12 shows scatter plots of true and predicted precipitation on test data for a station that showed significant improvement during sequential imputation in Scenario 1. As noted for **Figure 6** as well, these plots help visualize that as the *NSE* value increases during sequential imputation, the relative scatter decreases demonstrating improved spatial interpolation. **Figures 13**, **14** demonstrate that the benefits of sequential imputation also carry over to predicting dry events and extreme events despite the underlying limitations of spatial interpolation as noted in the section Evaluating Imputation: Nash Sutcliffe Efficiency (NSE). We observe a general trend that the improvements (or values of Δ) tend to be higher for stations that correspond to significant overall improvements (i.e., $\Delta_s NSE > \ \sigma_s$) as discussed above.

Results for aggregate correlations (**Figure 7B**) show that the correlation between S_i (i.e., partial sum of first i sorted correlations) and *NSE* is high for lower values of i, and gets progressively weaker as i increases. This implies that for reliable imputation, having a few references that are strongly correlated is more important than having many references that are weakly correlated. This highlights why sequential imputation is a powerful technique, since leveraging even one incomplete station that is highly correlated to the target station can make a significant improvement. We illustrate this further in **Figure 16**, where we show values of S_2 for all stations at the time of sequential imputation in Scenarios 1 and 2. As expected, values of S_2 during sequential imputation are higher than those during non-sequential imputation, which is consistent with improved imputations.

It is important to note that stations imputed earlier during sequential imputation tend to have a higher *NSE*, indicating a more reliable imputation. *NSE* values tend to decrease along the imputation sequence. This is primarily a consequence of the order in which we pick stations for sequential imputation. Stations that are imputed earlier in the sequence have a higher aggregate correlation with reference datasets, implying that missing data would be modeled with greater accuracy. This can be verified by observing the trend of the baseline *NSE* curve in **Figures 8A–11A**, which also shows a reduction in *NSE* values along the imputation sequence. Stations that are imputed later in the sequence will tend to have a lower value of *NSE* because they have a lower baseline *NSE* to begin with; they could still exhibit significant improvements during sequential imputation when compared to non-sequential imputation (as shown in **Figures 8B–10B**).

Finally, we note that the performance of sequential imputation could be negatively impacted if the data gaps among stations occur synchronously. In particular, this could happen if a station earlier in the sequence was poorly imputed and has a high correlation with a station imputed later in the sequence. However, the proposed sequential approach can still be implemented, and this approach will outperform or equally match the non-sequential approach.

CONCLUSIONS

Spatial interpolation algorithms typically require reference stations that have complete records; therefore, stations with missing data or incomplete records are not used. This limitation is critical for machine learning algorithms where incomplete records preclude data-driven learning of multivariate relationships. In this study, we proposed a new algorithm, called the sequential imputation algorithm, for imputing missing time-series precipitation data. We hypothesized that stations with incomplete records contain information that can be used toward improving spatial interpolation. We confirmed this hypothesis by using the sequential imputation algorithm which was incorporated within a spatial interpolation method based on Random Forests.

We demonstrated the benefits of sequential imputation as compared to non-sequential imputation. Specifically, we showed that sequential imputation helps leverage other incomplete records for more reliable imputation. We observed that as the proportion of stations with incomplete records increases, there is a higher percentage of stations benefitting from sequential imputation. On the other hand, if the proportion of stations with incomplete records is small, there is a high probability that sequential imputation does not add any extra information for spatial information. We also observed that the benefits of sequential imputation carry over to improved predictions of dry events and extreme events. Finally, results showed that for reliable imputation, having a few strongly correlated references is more important than having many references that are weakly correlated. This highlights why sequential imputation is a powerful technique, since including even one incomplete station that is highly correlated to the target station can make a significant improvement in imputation.

Although we demonstrated sequential imputation using Random Forests, it can be implemented using other ML-based and spatial interpolation methods found in the literature. Furthermore, we presented a new but generic algorithm for imputing missing records in daily precipitation time-series that is potentially applicable to other meteorological variables as well.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.wcc.nrcs.usda.gov.

AUTHOR CONTRIBUTIONS

UM and DD conceived and designed the study. UM acquired the data, developed the new algorithm, conducted all the numerical experiments, and analyzed the results. DD and JB provided input on methods and statistical analysis. BF provided input on data acquisition and time series analysis. DD helped analyze the results. SP and CS provided input on the conception of the study and were in charge of overall direction and planning. UM took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

FUNDING

This work was funded by the ExaSheds project, which was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Systems Sciences Division, Data Management Program, under Award Number DE-AC02-05CH11231.

REFERENCES

Acock, M. C., and Pachepsky, Y. A. (2000). Estimating missing weather data for agricultural simulations using group method of data handling. *J. Appl. Meteorol.* 39, 1176–1184. doi: 10.1175/1520-0450(2000)039<1176:EMWDFA>2.0.CO;2

Adhikari, R., and Agrawal, R. K. (2013). *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrücken: LAP LAMBERT Academic Publishing.

Ahmad, M. W., Mourshed, M., and Rezgui, Y. (2017). Trees vs. neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build.* 147, 77–89. doi: 10.1016/j.enbuild.2017.04.038

Box, G. E., and Jenkins, G. M. (1976). *Time Series Analysis. Forecasting and control. Holden-Day Series in Time Series Analysis*. San Francisco, CA: Holden-Day.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, L., Xu, J., Wang, G., and Shen, Z. (2019). Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. *J. Hydrol.* 572, 449–460. doi: 10.1016/j.jhydrol.2019.03.025

Chuan, Z. L., Deni, S. M., Fam, S.-F., and Ismail, N. (2019). The effectiveness of a probabilistic principal component analysis model and expectation maximisation algorithm in treating missing daily rainfall data. *Asia-Pac. J. Atmos. Sci*. 56, 119–129. doi: 10.1007/s13143-019-00135-8

Coulibaly, P., and Evora, N. D. (2007). Comparison of neural network methods for infilling missing daily weather records. *J. Hydrol.* 341, 27–41. doi: 10.1016/j.jhydrol.2007.04.020

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.* 28, 2031–2064. doi: 10.1002/joc.1688

Devi, G. K., Ganasri, B. P., and Dwarakish, G. S. (2015). A review on hydrological models. *Aquat. Proced.* 4, 1001–1007. doi: 10.1016/j.aqpro.2015.02.126

Dwivedi, D., Arora, B., Steefel, C. I., Dafflon, B., and Versteeg, R. (2018). Hot spots and hot moments of nitrogen in a riparian corridor. *Water Resour. Res.* 54, 205–222. doi: 10.1002/2017WR022346

Dwivedi, D., Steefel, I. C., Arora, B., and Bisht, G. (2017). Impact of intra-meander hyporheic flow on nitrogen cycling. *Proced. Earth Planet. Sci.* 17, 404–407. doi: 10.1016/j.proeps.2016.12.102

Gao, Y., Merz, C., Lischeid, G., and Schneider, M. (2018). A review on missing hydrological data processing. *Environ. Earth Sci.* 77:47. doi: 10.1007/s12665-018-7228-6

Gorshenin, A., Lebedeva, M., Lukina, S., and Yakovleva, A. (2019). "Application of machine learning algorithms to handle missing values in precipitation data," in *Distributed Computer and Communication Networks*, eds V. M. Vishnevskiy, K. E. Samouylov, and D. V. Kozyrev (Cham: Springer International Publishing), 563–577.

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.* 377, 80–91. doi: 10.1016/j.jhydrol.2009.08.003

Hasanpour Kashani, M., and Dinpashoh, Y. (2012). Evaluation of efficiency of different estimation methods for missing climatological data. *Stoch. Environ. Res. Risk Assess.* 26, 59–71. doi: 10.1007/s00477-011-0536-y

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., et al. (2019). Comparison of statistical downscaling methods with respect to extreme events over Europe: validation results from the perfect predictor experiment of the COST Action VALUE. *Int. J. Climatol.* 39, 3846–3867. doi: 10.1002/joc.5469

Hubbard, S. S., Varadharajan, C., Wu, Y., Wainwright, H., and Dwivedi, D. (2020). Emerging technologies and radical collaboration to advance predictive understanding of watershed hydro-biogeochemistry. *Hydrol. Process.* 34, 3175–3182. doi: 10.1002/hyp.13807

Hubbard, S. S., Williams, K. H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., et al. (2018). The East River, Colorado, Watershed: a mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. *Vadose Zone J.* 17, 1–25. doi: 10.2136/vzj2018.03.0061

Jahan, F., Sinha, N. C., Rahman, M. M., Rahman, M. M., Mondal, M., and Islam M. A. (2019). Comparison of missing value estimation techniques in rainfall data of Bangladesh. *Theor. Appl. Climatol.* 136, 1115–1131. doi: 10.1007/s00704-018-2537-y

Kim, J.-W., and Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *J. Hydrol.* 394, 305–314. doi: 10.1016/j.jhydrol.2010.09.005

Lo Presti, R., Barca, E., and Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environ. Monit. Assess.* 160, 1–22. doi: 10.1007/s10661-008-0653-3

Londhe, S., Dixit, P., Shah, S., and Narkhede, S. (2015). Infilling of missing daily rainfall records using artificial neural network. *ISH J. Hydraul. Eng.* 21, 255–264. doi: 10.1080/09715010.2015.1016126

Louppe, G. (2015). *Understanding random forests: from theory to practice* (Ph.D. dissertation). University of Liège, Liège, Belgium.

Morales Martínez, J. L., Horta -Rangel, F. A., Segovia-Domínguez, I., Robles Morua, A., and Hernández, J. H. (2019). Analysis of a new spatial interpolation weighting method to estimate missing data applied to rainfall records. *Atmósfera* 32, 237–259. doi: 10.20937/ATM.2019.32.03.06

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. T*rans. ASABE* 50, 885–900. doi: 10.13031/2013.23153

Oliver, M. A., and Webster, R. (2015). *Basic Steps in Geostatistics: The Variogram and Kriging*. Cham: SpringerBriefs in Agriculture, Springer International Publishing.

Paulhus, J. L. H., and Kohler, M. A. (1952). Interpolation of missing precipitation records. *Mon. Weather Rev.* 80, 129–133.

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.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., and Pita-López, M. F. (2008). A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *Int. J. Climatol.* 28, 1525–1534. doi: 10.1002/joc.1657

Schafer, J. L., and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177. doi: 10.1037/1082-989X.7.2.147

Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* 14, 853–871. doi: 10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *UCSF Center Bioinform. Mol. Biostat.* 15. Available online at: https://archive.ics.uci.edu/ml/about.html

Shepard, D. (1968). "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM National Conference* (New York City, NY: ACM Press), 517–524.

Simolo, C., Brunetti, M., Maugeri, M., and Nanni, T. (2010). Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol.* 30, 1564–1576. doi: 10.1002/joc.1992

Stockman, M., Dwivedi, D., Gentz, R., and Peisert, S. (2019). Detecting control system misbehavior by fingerprinting programmable logic controller functionality. *Int. J. Crit. Infrastruct. Prot.* 26:100306. doi: 10.1016/j.ijcip.2019.100306

Tang, F., and Ishwaran, H. (2017). Random forest missing data algorithms. *Stat. Anal. Data Min. ASA Data Sci. J.* 10, 363–377. doi: 10.1002/sam.11348

Teegavarapu, R. S. V. (2020). Precipitation imputation with probability space-based weighting methods. *J. Hydrol.* 581:124447. doi: 10.1016/j.jhydrol.2019.124447

Teegavarapu, R. S. V., and Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* 312, 191–206. doi: 10.1016/j.jhydrol.2005.02.015

Varadharajan, C., Faybishenko, B., Henderson, A., Henderson, M., Hendrix, V. C., Hubbard, S. S., et al. (2019). Challenges in building an end-to-end system for acquisition, management, and integration of diverse data from sensor networks in watersheds: lessons from a mountainous community observatory in East River, Colorado. *IEEE Access* 7, 182796–182813. doi: 10.1109/ACCESS.2019.2957793

Yozgatligil, C., Aslan, S., Iyigun, C., and Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theor. Appl. Climatol.* 112, 143–167. doi: 10.1007/s00704-012-0723-x

Zachara, J. M., Chen, X., Song, X., Shuai, P., Murray, C., and Resch, C. T. (2020). Kilometer-scale hydrologic exchange flows in a gravel bed river corridor and their implications to solute migration. *Water Resour. Res.* 56:e2019WR025258. doi: 10.1029/2019WR025258

Zhai, P., Zhang, X., Wan, H., and Pan, X. (2005). Trends in total precipitation and frequency of daily precipitation extremes over China. *J. Clim.* 18, 1096–1108. doi: 10.1175/JCLI-3318.1

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.

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/frwa.2020.00028

Using Convolutional Neural Networks for Streamflow Projection in California

Shiheng Duan, Paul Ullrich and Lele Shu*

Atmospheric Science Graduate Group, University of California, Davis, Davis, CA, United States

In this study, a novel temporal convolutional neural network (TCNN) model is developed for long-term streamflow projection in California within the Catchment Attributes for Large-Sample Studies (CAMELS) watershed regions. The TCNN model consists of several convolution blocks and causal convolution is used as physical constraint. The ensemble performance of the model is first compared with other machine learning models for streamflow prediction. The model is further assessed through comparison with reduced models and using different hyperparameters, with results suggesting that this model correctly ascertains the physical relationship between input variables and streamflow. The stability of the model and its behavior in the extrapolated regime is assessed through an idealized extreme test with quadruple precipitation and 5°C higher temperature. Future streamflow projections are then developed using daily high-resolution Localized Constructed Analogs dataset (LOCA). To understand the importance of the nonlinear machine learning approach, we estimate the degree of nonlinearity in the streamflow response among input variables. Our work shows the ability and potential for TCNNs to perform future hydrology projections.

Keywords: machine learning, temporal convolutional neural network, model sensitivity, streamflow projection, projection analysis

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Jie Niu,
Jinan University, China
Kuai Fang,
Pennsylvania State University (PSU),
United States

****Correspondence:***
Shiheng Duan
shiduan@ucdavis.edu

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *14 May 2020*
Accepted: *04 August 2020*
Published: *16 September 2020*

Citation:
Duan S, Ullrich P and Shu L (2020)
Using Convolutional Neural Networks
for Streamflow Projection in California.
Front. Water 2:28.
doi: 10.3389/frwa.2020.00028

1. INTRODUCTION

Streamflow is an undeniably important hydrologic quantity for agriculture, society and ecosystems. While historical records of streamflow have been indispensable in informing us of the probability associated with particular flow conditions, it is unclear to what degree these predictions are valid under future meteorological conditions in light of climate change. Failure to correctly predict reservoir inputs has the potential to lead to reservoir failure, such as was witnessed recently with the Oroville reservoir spillway collapse (White et al., 2019). Long-term projections of streamflow that capture the climatology of streamflow within each watershed are further useful for informing water management strategy. Models for streamflow prediction and projection can be generally divided into two categories: physically-based models and data-driven models (Shen, 2018). Since physically-based hydrological models typically require significant computational expense and extensive calibration of land surface characteristics, machine learning (ML) models are being increasingly employed for streamflow prediction, especially Artificial Neural Networks (ANNs) (Gao et al., 2010; Noori and Kalin, 2016; Atieh et al., 2017; Peng et al., 2017), Support Vector Machines (SVMs) (Kisi and Cimen, 2011; Huang et al., 2014), and recurrent networks like Long-Short Term Memory (LSTM) (Feng et al., 2019; Kratzert et al., 2019; Le et al., 2019; Yan et al., 2019).

Instead of directly simulating physical processes, ML models mimic the physical rules from historical datasets to develop a functional relationship between inputs and outputs. The learning process largely consists of repeated matrix algebra to adjust the weights in the models, which makes it amenable to acceleration by graphics processing units (GPUs). Further, because ML is broadly applicable across a variety of industries and fields, significant investments have been made in the software supporting its use. Compared with physically-based models, ML models are generally faster to train and can operate with essentially any predictors (Kratzert et al., 2019). However, the structure of the model and predictor selection are important since they determine the model performance. The general principle governing these models is to build a simple, easy to train model with all the necessary predictors—while avoiding redundant predictors—and ensuring the relationships being clear and direct.

Significant research on ML data-driven models for streamflow has been directed toward data preprocessing, with the purpose being to reduce the number of degrees of freedom in the input dataset and so make any underlying patterns or relationships easier to be identified by ML algorithms. Streamflow at a single gauge station is a fairly traditional 1D time-series dataset, but one that is composed of different components at a variety of frequencies. Consequently Kisi and Cimen (2011) used the discrete wavelet transform (DWT) with SVM for monthly streamflow prediction. The DWT was used to decompose streamflow into high-frequency and low-frequency components, referred to as the "details" and "approximation" in their study, respectively. The approximation, which is the low-frequency component, acts as the baseflow while other high-frequency details represent the variation with shorter period. Their results demonstrated preprocessing with DWT increased the prediction accuracy compared with a model leveraging the raw series. Analogously, Peng et al. (2017) employed the empirical wavelet transform (EWT). Unlike DWT, the EWT decomposition consisted of only three modes, which were used for an ANN model and a residual component. Huang et al. (2014) introduced the empirical mode decomposition (EMD) method for streamflow preprocessing. They decomposed the original data into five intrinsic mode functions and a residual. Instead of removing the residual, they retained it and excluded the high-frequency intrinsic mode function, producing better performance compared with the model using only the original data series. Although these preprocessing steps can simplify the streamflow series and increase performance, they also introduce additional hyperparameters and uncertainty into the model which may impact model robustness.

ML model research has also focused on limiting the choice of predictors—including both input variables and time window size—so as to reduce the number of inputs (Rasouli et al., 2012). Since traditional ML models do not generally incorporate comprehensive physical relationships, ML model developers can focus on only the predictors that explain the most output variability. For streamflow, the most common predictors are precipitation (P) and streamflow (Q) over some historical time period. However, other predictors have been explored as well, informed by our understanding of the system's physical drivers; for instance, Rasouli et al. (2012) investigated several climate indices as predictors, and demonstrated that these can be beneficial for prediction with long lead times up to 7 days. If one only uses precipitation and historical streamflow as predictors, the 1-day lag streamflow prediction problem can be expressed as: Identify a function f so that the predicted daily time series

$$\hat{Q}_t = f(P_{t-N}, P_{t-N+1}, \ldots, P_t; Q_{t-N}, Q_{t-N+1}, \ldots, Q_{t-1}) \quad (1)$$

satisfies $\hat{Q}_t \approx Q_t$ (measured under some prescribed metric). Here the subscript represents the time index and N represents the number of historical time points used for prediction. N must typically be large enough to incorporate all historical information relevant to prediction of streamflow at present, but large values of N can lead to increased model complexity which can in turn reduce performance. The value of N is thus usually decided by calculating the autocorrelation or partial correlation; Yaseen et al. (2016) used this approach for monthly streamflow prediction, eventually deciding on a time lag of 5 months.

A common feature in early data-driven streamflow prediction models is that the input variables were independent of time when fed into the ANNs or SVMs. For example, there were no connections within each layer of dense ANNs, and consequently the network could not "remember" past states. Under such architectures, temporal features in the predictors that may be vital for time series prediction might be neglected. To deal with this problem, some recurrent ML models have been adapted to recognize time dependent features (Le et al., 2019; Yan et al., 2019). Among such models, the most commonly used network (at present) is the LSTM. Kratzert et al. (2019) used LSTM and Catchment Attributes for Large-Sample Studies dataset (CAMELS) to predict streamflow over CONUS. Their results demonstrated that the LSTM model is capable of extracting temporal features and the results from the ML model can then be used to interpret the physical characteristics of different basins. Feng et al. (2019) added the previous flow rate as data integration, which improves the prediction accuracy of LSTM model. They also employed a convolution data integration method, although the resulting model did not outperform feeding observations directly into LSTM model.

Although there are many ML prediction models, not all can be directly employed for long-term projection. Under future climate change scenarios driven by increased greenhouse gas concentrations, the U.S. West is expected to experience more precipitation and higher surface temperature (Huang and Ullrich, 2017; Ullrich et al., 2018). It is similarly expected that the resultant streamflow patterns will also change. In ML models, since the model is developed and trained with a prescribed training dataset, it is generally expected that the target variable is the same in both training and testing sets. In the real world, however, the statistical properties of the target variable may be changing in time (for instance, under climate change). Under such scenarios, the prediction model may be inconsistent with future projection data, a problem referred to as concept drift (Tsymbal, 2004). Although streamflow can be used in a predictive model framework such as Equation (1) (an initial-boundary value problem), a simple substitution of $\hat{Q}$ for Q to produce a projection model can lead to errors in streamflow that accumulate over time, potentially biasing the projection. Consequently, projection models must be more heavily constrained to external forcing data, which can restrict the selection of ML model. In the context of projection, Koirala et al. (2014) used the Catchment-based Macro-scale Floodplain Model (CaMa-Flood) with runoff from CMIP5 models as input to derive streamflow under different climate scenarios. Gao et al. (2010) used an ANN and ECHAM5/MPI-OM model output to derive monthly projection for Huaihe River Basin. These studies demonstrated the potential for ML in streamflow projection.

In the present work, we document the development and validation of a ML-based modeling system for estimating future daily streamflow in California under climate change. After intercomparison among various ML models, a general Temporal Convolutional Neural Network (TCNN) is selected as our candidate system. Although CNNs have not been typically employed for streamflow prediction and projection—being more widely known for image processing—recent work has shown that they exhibit comparable performance to recurrent networks for time series problems (Bai et al., 2018). Consequently our study aims to further establish that TCNNs are competitive for streamflow forecasting with only atmospheric forcing data. Model sensitivities to input variables and time window size are investigated to develop optimal configurations for each basin. With the ML-based streamflow model in hand, future streamflow projections are constructed through the end of the twenty-first century using statistically downscaled LOCA meteorology as input. To the best of the authors' knowledge, this is the first work to assess TCNNs for streamflow projection with only atmospheric forcing data. The comprehensive study of the model's sensitivity to covariates and time window size are further novelties of this study. Although this work identifies a strategy for production of future streamflow projections, future work is needed to validate the methodology against physical constraints and investigate the impacts these changes may convey.

The remainder of the paper is structured as follows: section 2 provides technical details about our study, including descriptions of the data sources and the ML model structures. Section 3 explores prediction and projection across ML models, examines the sensitivity of the TCNN to input variables and time window size, and assesses the linearity of the problem. Insights from these future streamflow projections are presented in section 4. Conclusions follow in section 5.

2. DATA AND MODELS

2.1. CAMELS

The Catchment Attributes for Large-Sample Studies (CAMELS) dataset provides the hydrologic data for this study (Newman et al., 2014). The CAMELS dataset contains gauge streamflow data and forcing data for 671 basins that feature minimal human disturbance and at least 20 years of data over CONUS. The forcing data is provided as a basin average from NLDAS, Daymet, and Maurer, and includes precipitation, day length, solar radiation, and temperature. The streamflow time series data is obtained from USGS gauge stations. The dataset covers 40 watersheds in California, which we have downselected to the 20 watersheds without missing values for this study. **Figure 1** shows the location, HUC8 identifier, and name of these watersheds. Based on the location of these watersheds, we divided them into five categories and there are the corresponding abbreviations: NC for Northern California (Basin 11381500, 11451100, 11475560, 11522500, and 11528700), SN for Sierra Nevada (Basin 10343500, 11264500, 11266500, and 11284400), SC for Southern California (Basin 10258500, 10259000, and 10259200), CC for Central Coast (Basin 11141280, 11143000, 11148900, 11224500, and 11253310) and BA for the Bay Area (Basin 11162500, 11176400, and 11180500). In our model, streamflow is normalized by the basin area to avoid discrepancies in the magnitude of the streamflow. The data period is from January 1st, 1980 through December 31st, 2014. In total, we select 10,000 daily samples for training (approximately 27 years) and leave the remainder of the dataset for testing. These training samples are consecutive from the beginning of the time series.

2.2. LOCA Downscaled Meteorology

For future streamflow projection, the Localized Constructed Analogs (LOCA) dataset (Pierce et al., 2014) is employed. This dataset provides the three necessary input variables for this study, namely precipitation, solar radiation, and near-surface temperature. LOCA is a downscaled dataset ensemble with 6 kilometer resolution over North America from central Mexico through Southern Canada. Among all available LOCA datasets, we downselect four global climate model for this study, which are HadGEM2-ES, CNRM-CM5, CanESM2, and MIROC5 under RCP8.5. These models agree with the four models chosen by California's Climate Action Team Research Working Group as priority models for research contributing to California's Fourth Climate Change Assessment (Pierce et al., 2018). The climatology of these models can be described as warm/dry (HadGEM2-ES), cool/wet (CNRM-CM5), and average (CanESM2). Finally, MIROC5 was selected because it is the most unlike the other three. Since all the basins have irregular shapes, TempestRemap (Ullrich and Taylor, 2015; Ullrich et al., 2016) is used to conservatively regrid the LOCA data to obtain basin-mean forcing data. Because of the uncertainty from both the climate model output and the downscaling process, the historical LOCA data and CAMELS data have some significant disagreements, especially in the values of solar radiation. Specifically, LOCA tends to overestimate the solar radiation compared with NLDAS, as seen in **Tables S1–S5**. To avoid issues related to this systematic difference, the LOCA data was linearly transformed based on the historical forcing data to match the mean and variance of observations. The same transformation was also applied on the projection forcing data. Specifically, for a given daily input X_{LOCA}, either historical or projection, we denote the transformed value as X_{trans}, where μ and σ represent the corresponding mean and standard deviation from the historical period:

$$X_{\text{trans}} = \frac{X_{\text{LOCA}} - \mu_{\text{LOCA\_hist}}}{\sigma_{\text{LOCA\_hist}}} \times \sigma_{\text{NLDAS\_hist}} + \mu_{\text{NLDAS\_hist}} \quad (2)$$

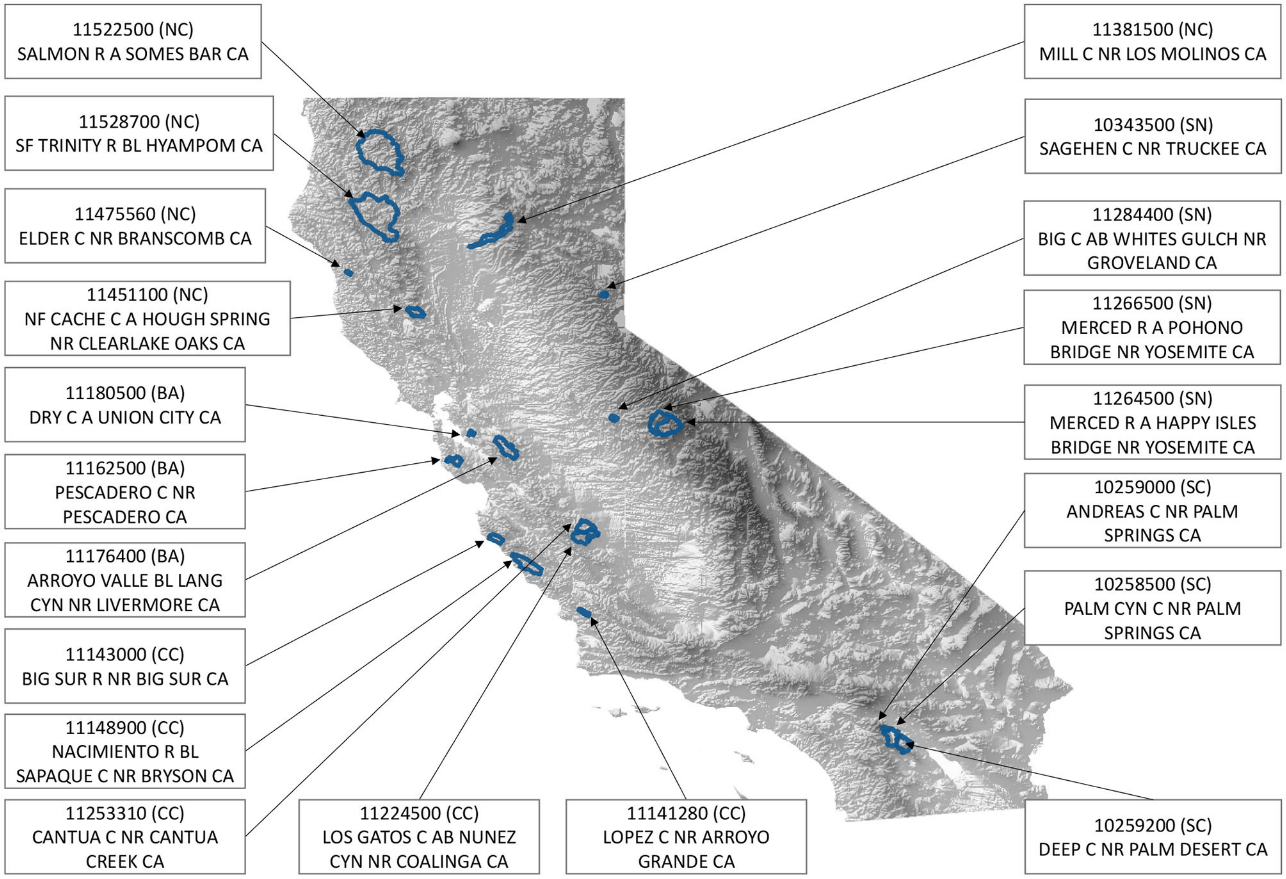


FIGURE 1 | A topographic plot of California and the 20 watershed regions considered in this study.

The values of μ and σ for NLDAS and the climate model ensemble can be found in **Tables S1–S5**.

Although LOCA provides historical daily atmospheric forcing data, it is not suitable for model training since it is generated from several climate models via an statistical downscaling method. The climate models produce a simulated climatology which is only constrained to the real world through prescribed atmospheric greenhouse gas concentrations, so there is effectively no relationship between LOCA and observed gage-based streamflow measurements. This is also the reason why we only analyze the climatology of flow rate in section 4, and do not directly compare the time series of streamflow.

2.3. Model Predictors and Target

As mentioned earlier, the input variables (predictors) for our streamflow models are precipitation, temperature, and solar radiation. By default, the input time window size is set it to 365 days (although this is explored later in the text). In general, the length of the input time window needs to be long enough to capture the relevant physical relationships between input variables and streamflow. For each of our ML models, the target variable is streamflow on the last day in the time window. In other words, our objective is to determine the function f in the following equation:

$$Q_t = f(P_{t-N+1}, P_{t-N+2}, \ldots, P_t; T_{t-N+1}, T_{t-N+2}, \ldots, T_t; S_{t-N+1}, S_{t-N+2}, \ldots, S_t) \quad (3)$$

where Q denotes streamflow, P the precipitation, T the temperature, and S the solar radiation. Note that this equation is only provided for the reader to better understand the relationship between streamflow and the independent quantities. The actual functional relationship will vary based on the model architecture. The subscript denotes the corresponding daily value for that particular quantity, and N denotes the input time window size. The input and output variables are all normalized before feeding them into the models via

$$X_i = \frac{x_i - \mu_x}{\sigma(x)}. \quad (4)$$

Here X_i and x_i are the *ith* normalized and original variable, and μ and σ stand for mean and standard deviation of that variable. With the normalized variables having zero mean and unit variance, the specific units and range of the inputs will not influence the model. In turn, the normalization procedure is expected to improve the model performance (e.g., Shanker et al., 1996).

2.4. Machine Learning (ML) Models

Four machine learning model architectures have been investigated and compared with a baseline linear regression model. For the predictive simulations, model performance is quantified by the Nash-Sutcliffe model efficiency (NSE) coefficient (Nash and Sutcliffe, 1970), which is defined as:

$$\mathrm{NSE} = 1 - \frac{\sum (Q_m^t - Q_o^t)^2}{\sum (Q_o^t - \bar{Q}_o)^2} \tag{5}$$

where Q_m^t denotes the predicted flow at time t, Q_o^t the observed flow at time t, and $\bar{Q}_o$ the mean observed flow. Here the observed quantities refer to output streamflow from USGS gauge stations. Larger NSE values indicate better performance. Since the NSE is proportional to the square of the difference between model and observations, it tends to put greater emphasis on high flow periods. To maximize NSE, we set $1-\mathrm{NSE}$ as the loss function for our models—that is, the quantity to be minimized during training process. For each model, training is performed separately on each basin but with the same model architecture.

Before training these networks, we first need to set the hyperparameters, which are tuning factors in the model architectures and training process. Common hyperparameters include the number of layers, optimizer, and number of epochs: The number of layers is important to the specific model architecture; the optimizer refers to the gradient descent algorithm used in the training process; and the number of epochs refers to the number of times that the model is trained on the entire training set. The Adam optimizer is used with 0.0005 as the learning rate. We trained each model for 150 epochs with the batch size set to 512. These training configurations are set based on the training loss function, which ensures the loss decreases and stabilizes at a low value. Although the hyperparameters are important for overall model performance (Bergstra and Bengio, 2012), in this work we hold the optimizer and the number of epochs the same for all models. This study does not investigate differences that may arise through more fine tuning of these hyperparameters for specific models—indeed a comprehensive investigation of the optimal hyperparameters for each model is beyond our current computational capability. The remainder of this section describes the architecture of the models investigated in this study.

2.4.1. Linear Regression

Linear regression refers to the simple linear regression model that only incorporates first-order terms from Equation (3). This precludes nonlinear relationships between days in the time series of input variables or between different input variables. As mentioned earlier, the simple linear regression model will be our baseline for assessing the ML models.

2.4.2. Artificial Neural Network (ANN)

An ANN is a traditional neural network composed of dense neural layers (Hassoun et al., 1995). It is an all-connected network without interactions within each single layer. According to the universal approximation theorem, with enough hidden units and depth among the hidden layers, an ANN can simulate any nonlinear relationship (Csáji et al., 2001; Lu et al., 2017). For time series data, however, recurrent neural networks such as GRU and LSTM normally outperform ANNs because of their ability to capture temporal features. In our work, the ANN model has two hidden layers with 100 hidden units and a "ReLU" activation function in each layer. The set of hyperparameters is set based on our coarse tuning for all interested basins. This ANN model is a nonlinear model without temporal features, which is the baseline for the following GRU, LSTM, and CNN models. ANNs have been previously investigated for streamflow prediction in Kisi and Kerem Cigizoglu (2007), and they compared the performance from different ANN models. Noori and Kalin (2016) used an ANN coupled model and it was found that ANN can help improve the streamflow prediction when coupled with physically-based SWAT model.

2.4.3. Gated Recurrent Units (GRU)

As mentioned earlier, recurrent neural networks (RNNs) are typically used to deal with time series and related quantities. However, under simple recurrent designs the gradient will often vanish or explode during the training process (Bengio et al., 1994). As introduced by Cho et al. (2014), GRUs are a typical gated recurrent neural network whose design can help to avoid gradient vanishing for recurrent networks. There are two gates in a GRU cell, referred to as the update gate u and relevant gate r. A general GRU cell is depicted in **Figure S1**, followed by a set of equations defining the GRU cell. There can be several hidden units in a GRU layer and the number of hidden units is the number of features in cell states.

Similar with the ANN model, a GRU model consists of several layers, with each layer containing several hidden units. In our work, we analyze a three-layer GRU model connected with a dense layer. The GRU layers are set to extract temporal features and the final dense layer is for output. Each GRU layer has 50 hidden units; this number is from coarse hyperparameter searches. The stacked layer design provides sufficient complexity to fit the streamflow data, and provides a similar stacked architecture to compare with the TCNN model.

2.4.4. Long-Short Term Memory (LSTM)

The LSTM model is another example of a gated network, and one which has been increasingly explored in recent years for streamflow forecasting (Kratzert et al., 2019). The primary difference between the LSTM and GRU models is that the LSTM features three gates—the update gate u, the forget gate f and the output gate o. A typical LSTM cell and the corresponding formulas are shown in **Figure S2**. Like our GRU model, the LSTM model has three LSTM layers and one dense layer. Within each LSTM layer, there are 50 hidden units. The stacked layer design ensures complexity to fit the streamflow data and the same number of cells with GRU can help compare the different model performance.

2.4.5. Temporal Convolutional Neural Network (TCNN)

CNNs remain a widely used model for image processing and analysis because of their ability to extract and decompose features (Gu et al., 2018). The typical input of the CNN model is an image with width, length, and color channels. In our study of streamflow, which is one-dimensional data, the input shape is the number of variables times the input time window size. A typical CNN is comprised of convolutional layers and dense layers. Some CNNs will further add pooling layers between convolutional layers to reduce the dimensionality of the problem and extract important features. But it has also been argued that with sufficiently large convolutional layers, the network can perform well only using convolutional operations (Springenberg et al., 2014). Thus, for simplicity we have only used convolutional layers in our CNN.

A typical CNN architecture used for time-series data is the Temporal Convolutional Neural Network (TCNN) (Lea et al., 2017). Compared with the more well-known CNN for images, TCNNs consist of a one-dimensional network using dilated causal convolutions to keep temporal causation and residual blocks for deeper networks. Bai et al. (2018) tested TCNNs and LSTMs with different time series problems, and argued that TCNNs are better in terms of accuracy and speed for problems of similar complexity. Thus, to compare with our three-layer GRU and LSTM models, we will assess a three-block TCNN with residual connections. Each block has two convolution layers and one residual connection. Dilation rates are set to 1, 6, 12, respectively and kernel size is fixed at 7 for all convolution layers. The number of filters are set to 40, 20, 20 for each block. In the final block, the reception field is large enough to cover the entirety of the time window. Also, with stacked causal convolution blocks, the input information will be concentrated within the last few neurons. To reduce redundancy and avoid overfitting, a slice layer is set after the final TCNN block to only keep last 20 neurons. The TCNN model flow chart and an illustration of dilated causal convolution are shown in **Figure S3**.

2.5. Ensemble Runs

Unlike the linear model, which has an exact analytical solution, all the neural networks use gradient-based method to optimize the loss function. Since the networks allow for local minima, different initial weights can potentially produce different models with different performance. Thus one needs to be careful to avoid drawing conclusions on the relative performance of each model that are merely a byproduct of the initial weights. In order to eliminate this effect, we run each model 15 times to get an ensemble distribution of NSE values. Thus our results and conclusions are based on the statistical distribution of model performance across the ensemble.

Throughout this study we make use of boxplots for assessing comparative performance between ensembles. As shown in Krzywinski and Altman (2014), comparative performance is intuitive from the boxplot—namely, if the median for one model is above the interquartile range of another, we are confident that it is the better model. However, if the median from the second model lies within the interquartile range of the first model, performance could be the result of randomness in the training process, making it difficult to determine the better model.

3. RESULTS

In this section we first compare the various ML models discussed in section 2.4 to demonstrate the competitive performance of the TCNN. The TCNN is then examined in light of stability under extreme forcing, its sensitivity to choice of input variables across basins, and sensitivity to time window size. A physical interpretation of the observed model sensitivity is also discussed here.

3.1. Model Intercomparison

Figure 2 shows the ensemble prediction results for each basin among the four ML models, plus the linear regression model. The linear regression model performs the worst among available models in almost all basins, in testament to the nonlinearity of the prediction problem. The ANN model tends to achieve a higher NSE value than the linear regression model for almost all basins, but in terms of NSE the ANN is still inferior to the recurrent networks and the TCNN, especially for basins where the NSE values for the recurrent networks and the TCNN are over 0.6, such as 11475560(NC) and 11522500(NC). In these basins, the relatively low NSE scores from the ANN indicate that there are some temporal features that the ANN cannot capture but which are important for streamflow prediction. Nonetheless, for some basins, the ANN outperforms the LSTM. There are two possible reasons this may occur. Firstly, it could be that temporal features are not important for these basins, a hypothesis that is supported by the observation that the ANN tends to also be better than GRU [e.g., basins 11176400(BA) and 11224500(CC)]. On the other hand, the TCNN doesn't have a recurrent architecture so it can effectively ignore the temporal features and mimic the ANN. This could suggest that LSTMs may not be as generalizable as TCNNs. Another possible reason is that the LSTM hyperparameter set is suboptimal for these basins – assessing this possibility may require a more comprehensive basin-dependent hyperparameter search.

Among models with temporal features (TCNN, LSTM, GRU), the TCNN exhibits the best average performance. The average NSE over all basins and all ensemble runs is 0.40 for LSTM, 0.44 for GRU, and 0.55 for TCNN. The average NSE value for the best run over all basins is 0.58 for LSTM, 0.58 for GRU, and 0.65 for TCNN. For those basins where LSTM and GRU achieve the highest NSE values, the performance of the TCNN model is competitive—for example, in basins 11141280(CC) and 11284400(SN) the NSE values among the different neural networks are all higher than 0.5. For basins where neural networks do not perform well, such as basins 10259200(SC), 11176400(BA), and 11253310(CC), the TCNN is nonetheless the best among the different neural networks. Notably, the recurrent networks can achieve high NSE values for some basins while performing poorly for other basins. That is, their performance varies substantially among different basins. The TCNN, however, is more stable among all basins: The standard deviation of the ensembles over all basins is 0.47 for LSTM, 0.38 for GRU, and 0.23 for TCNN. Although the choice of hyperparameters is important in these results, the wider spread in the NSE value indicates that for streamflow prediction the recurrent networks have more local

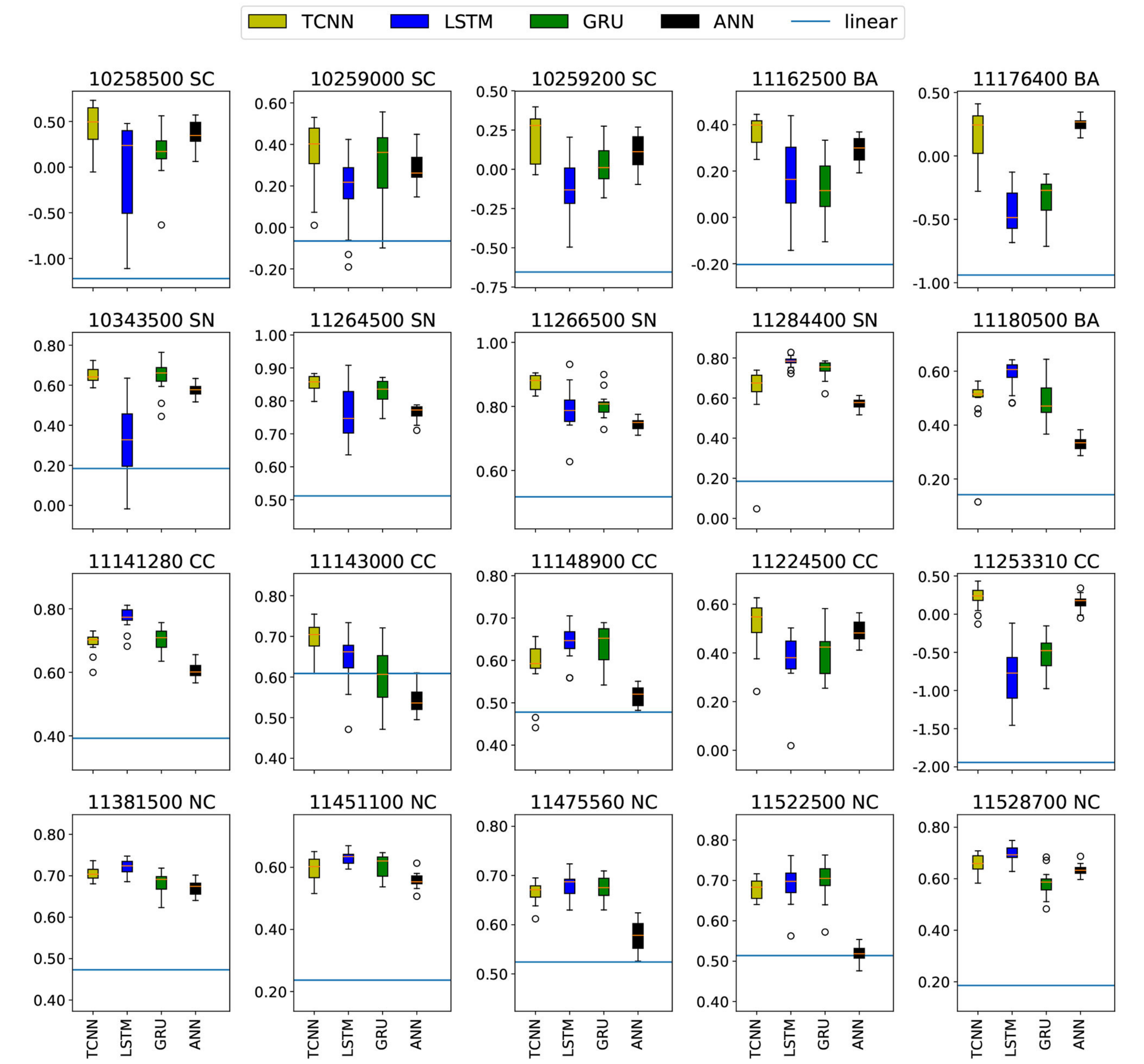


FIGURE 2 | Ensemble prediction comparison for all basins with different models. The boxplots denote results over each ensemble of 15 model runs for the ML models. The straight line denotes the linear regression result. For basin 11224500 CC the linear regression model produced a NSE of −2.47.

minima over the optimization space, and consequently must be trained many times to find a globally optimal configuration.

The stacked recurrent networks here are chosen to compare with the stacked TCNN model. A one-layer LSTM model, such as the one used in Kratzert et al. (2019), is also investigated. We employ 256 hidden units for the LSTM to match (Kratzert et al., 2019). Similarly a one-layer GRU model with 256 hidden units is also compared. With this configuration, the average NSE over all basins and all ensemble runs does improve to 0.47 for the one-layer LSTM, but degrades to 0.34 for the one-layer GRU. The standard deviation is 0.40 and 0.74 for the one-layer LSTM and GRU, respectively. When comparing the average NSE for the best run, the one-layer LSTM and GRU achieve 0.61 and 0.56, respectively. We again tested another one-layer LSTM model with 370 hidden units so as to match the number of free parameters within the TCNN model. The average NSE over all ensemble runs is 0.44, and 0.58 for the best run. The standard deviation is 0.43. A comprehensive comparison can be found in **Table 1**. **Figure S4** shows the ensemble prediction comparison of the TCNN model with one-layer recurrent networks. Compared

TABLE 1 | Mean and standard deviation for ensemble prediction comparison with different models.

	Mean NSE for all ensemble runs	Mean NSE for the best run	Standard deviation
TCNN	0.55	0.65	0.23
Stacked LSTM	0.40	0.58	0.47
Stacked GRU	0.44	0.58	0.38
One-layer GRU (256)	0.34	0.56	0.74
One-layer LSTM (256)	0.47	0.61	0.40
One-layer LSTM (370)	0.44	0.58	0.43

The number in parentheses denotes the number of hidden units.

with the one-layer recurrent networks, the TCNN model still exhibits slightly better performance with lower variation.

Besides evaluating the NSE value for the whole prediction period, we also examined the model performance for high flow and low flow days. High flow (low flow) days are defined as days when the observed flow rate is higher (lower) than the $95th$ ($5th$) percentile over all days. Since the low flow series for some basins is zero throughout, NSE cannot be used to assess performance. Instead, we use mean squared error (MSE) to quantify the performance, given by

$$\mathrm{MSE} = \sum (Q_m^t - Q_o^t)^2. \tag{6}$$

The MSE spread for all basins over the ensemble can be found in **Figures S5**, **S6**. Whereas the TCNN tends to perform well during high flow periods, the LSTM does exhibit better performance in low flow periods. For the high flow period, when average MSE is compared over all the ensemble runs, the TCNN achieves the best performance on 12 basins compared with 4 basins for LSTM. When comparing the minimum MSE for the high flow period, the TCNN is the best model for 10 basins compared with 6 basins for LSTM. For the low flow period, when using the average MSE, TCNN is the best model for 2 basins compared with 10 basins for LSTM; when assessing minimum MSE value, the LSTM is superior in 16 basins. The reason for this behavior is likely a simple consequence of the chosen hyperparameters of the model; further optimization will likely result in incremental improvements to both the TCNN and LSTM. Notably, the purpose for the comparisons in this section are not to show the TCNN is better than the LSTM, since such a proof would require us to effectively test all possible architectures and hyperparameters. Instead, these results demonstrate that the TCNN can achieve comparable performance to other commonly used models.

In addition to assessing model performance, training time also merits comparison among the different models. The average training time for one basin with the ANN on a single RTX 2080Ti is 11 s, 77 s for TCNN, 149 s for stacked GRU, and 150 s for stacked LSTM. For the one-layer LSTM model, it takes 220 s for 256 hidden units and 380 s for 370 hidden units. Hence, for this particular configuration the TCNN model is the fastest among the models with temporal features, although only by a factor of two.

Based on the results presented in this section, TCNN is chosen as our candidate network for prediction and projection of streamflow. The remainder of this paper now focuses assessing and explaining the performance of the TCNN.

3.2. Model Stability Under Extreme Climatological Forcing

One of the biggest challenges for ML projection is concept drift (also known as non-stationarity). Under climate change, it is widely accepted that the statistical properties of the input predictors and output streamflow will change through time. Although surface temperatures are expected to increase almost everywhere, in parts of California these increases are also accompanied by an increase in total precipitation of about 1.2% per decade (Ullrich et al., 2018). It is further expected that the input variance will increase in conjunction with more frequent extreme precipitation and temperature events (Swain et al., 2018). However, because the TCNN model is trained on historical data, the end-of-century inputs may incur extrapolation, which has the potential to produce unphysical results such as negative flow. To test whether the TCNN model is able to produce physically reasonable results even when inputs are not within the range of the training data, an idealized test is devised to stress the model far beyond the long-term range of possible inputs. Specifically, the model was executed with quadruple precipitation and a temperature increase of 5 degrees Celsius from the training set. Only one simulation was performed for each basin, using the TCNN model with highest NSE value from the ensemble run.

The extreme scenario investigated here is unrealistic even in light of climate change. However, if the ML model were not stable, this extreme scenario, far outside the realm of the training data, should cause the model to "blow up" or generate negative flow rates. However, if our model can still produce acceptable results under such an extreme scenario, we have greater confidence that it will generate reasonable projection results under the RCP8.5 scenario. Results from a single representative basin are depicted in **Figure 3**. Although only one basin in shown here, the results are analogous in other basins (not shown). As expected, the projected streamflow is generally much larger than historical, with much higher flood peaks. In addition, the high flow period is longer under this test as a result of precipitation accumulation, and low flow periods produce consistently higher streamflow. The regression line from the scatter plot is $Q_p = 4.001 \times Q_h$ ($R^2 = 0.74$), where Q_p is

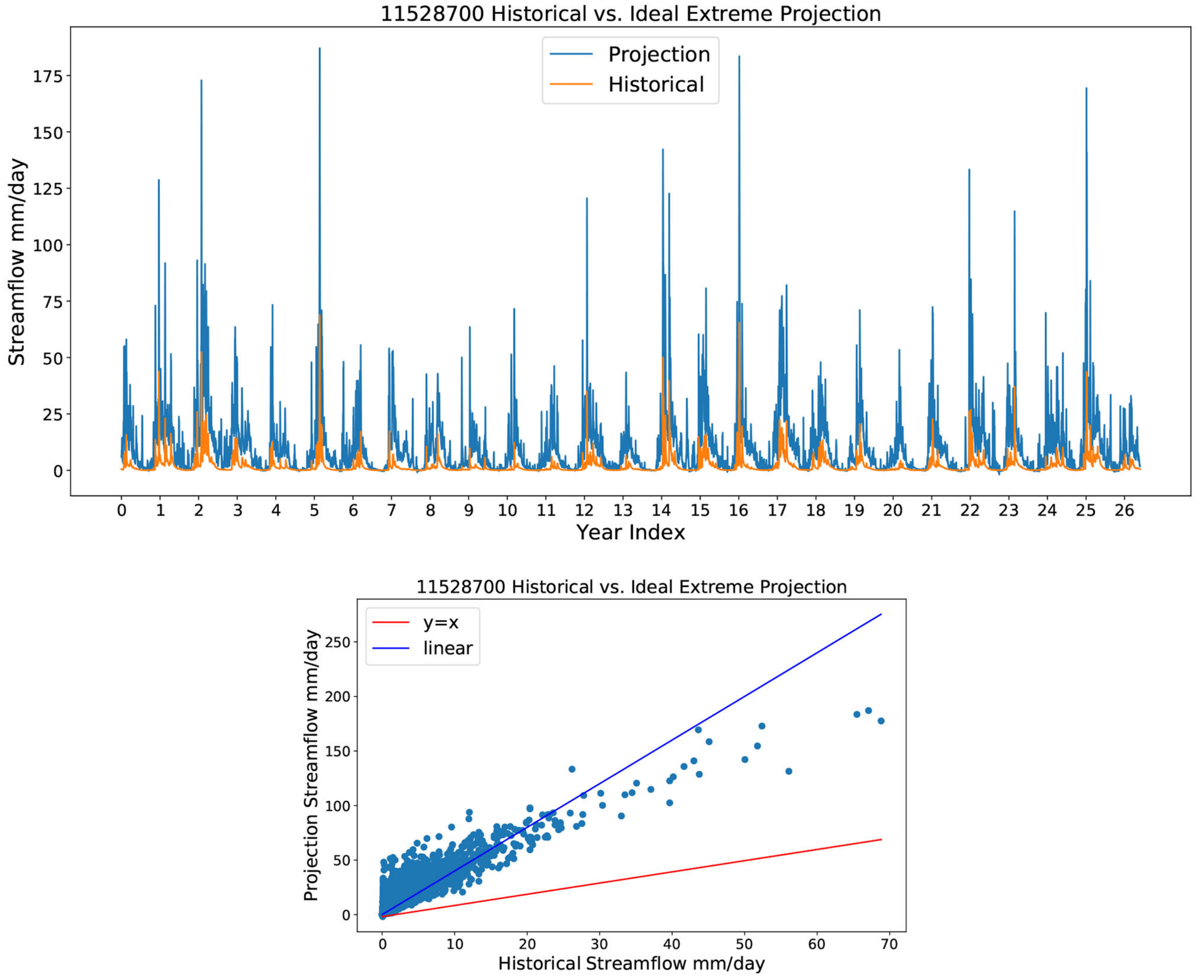


FIGURE 3 | A depiction of the streamflow response under idealized extreme forcing showing **(top)** time series of flow and **(bottom)** historical streamflow vs. projected streamflow.

projection streamflow and Q_h is the historical streamflow. Thus the 4× increase in precipitation produces approximately a 4× increase in streamflow. However, this simple linear factor appears to underestimate flows on the low flow days and overestimate flows during high flow days, again indicative of nonlinearity in the streamflow dynamics.

3.3. Model Sensitivity to Input Variables

As discussed earlier, the input variables for our full model are precipitation, temperature, and solar radiation. Although input fields beyond precipitation can improve model performance by capturing significant physical relationships, they also increase the complexity of the model, potentially leading to a wider spread among trained models. To test the importance of these variables for streamflow prediction three reduced models were compared, consisting of precipitation solely (p), precipitation and temperature (pt), and precipitation and solar radiation (ps). When comparing the performance of reduced models and the full model, the 15-model ensemble was again used to avoid noise from the initial state.

The overall performance of ps and pt models is again assessed using box plots of NSE values. **Figure 4** shows the result of the ensemble comparison. It is apparent that for some basins temperature boosts predictability, while for others solar radiation is more important. There are only three basins where the best pst model is better than the best ps or pt model [11264500(SN), 11266500(SN), and 11381500(NC)], and in each of these cases the improvement with all three variables is modest. In each basin, the dominant variable does reflect the geographic features of the basin. Basins where temperature significantly improves performance are 10343500 (SN), 11264500 (SN), 11266500 (SN), 11451100 (NC), 11522500 (NC), 11176400 (BA), and 11224500 (CC) which include three Sierra Nevada basins, two Northern California basins, and one Bay Area basin. Basins where solar radiation improves performance are 10258500 (SC), 10259000 (SC), 11143000 (CC), 11253310 (CC), 11180500 (BA), 11284400 (SN), and 11475560 (NC)—except for the last two, these are located in coastal areas or in the inland desert of Southern California. These results suggest that, to a close approximately, we can divide the basins into three categories using these reduced models: those where temperature is important (generally in mountainous regions), those where solar radiation is important (generally near coastlines), and those where temperature and solar radiation offer no significant benefit to predictability.

The physical explanation underlying the performance of the reduced models is related to the climatological properties of these different basins. For instance, in the basins of the Sierra Nevadas and Northern California, accumulation and melt of wintertime snowpack generally plays an important role in driving streamflow. However, the inclusion of temperature in these mountainous regions does not necessarily guarantee a performance improvement. For instance, temperature does not improve the model of 11528700 (NC), where snow is a major driver for streamflow; nonetheless, the inclusion of temperature also does not significantly degrade performance. Further, in basins where temperature improves performance we also generally see that inclusion of solar radiation does provide some improvement over the models only using precipitation—this suggests that the ML model is potentially identifying the relationship between solar radiation and temperature, or is instead using solar radiation to estimate snow melt rates.

The physical processes driving streamflow in the coastal basins are significantly different than those of the mountains. Namely, coastal basins do not experience significant temperature variations as a result of temperature regulation by the ocean. Further, because the ocean provides a ready source of moisture, air remains close to saturation. In accordance with the Penman-Monteith equation, evaporation from these basins will be driven primarily by radiative forcing, in agreement with our results. Among the central coast basins, the one exception that shows improved performance with temperature, but no significant improvement from solar radiation is 11224500 (CC). Although this basin is on the Central Coast, it is far from the coastline and so subject to larger temperature swings and lower relative humidity. The relatively high-altitude coastal ranges in this basin do produce occasional snow accumulation, but it is unlikely that snow dynamics plays a role here.

For those basins where inclusion of solar radiation and temperature produce worse model performance (i.e., the three Southern California basins), we hypothesize that the ML model is either identifying non-existent physical relationships between these variables and streamflow in the training data, or that the increased model complexity is making it more difficult for the model to converge to an optimal configuration. The truth is likely a combination of both of these factors, as for all three SC basins the "best performing" pst model is not significantly worse than the median p-only model, but is clearly worse than the best p-only model.

In conclusion, the reduced models explored here are helpful for giving insight into the processes that are most relevant for each basin, and thus the relevant causative relationships. Here snowpack dynamics and coastal meteorology have emerged as two obvious geographical features important for determining model behavior. Given this behavior agrees with our physical understanding of the system, we have further evidence to suggest that the models are behaving credibly.

3.4. Model Sensitivity to Time Window Size

The input time window size is an important hyperparameter for our model, and one that is intrinsically connected to the physical processes driving streamflow. However, a time window that is too large can reduce model performance and slow training time. Some past studies set the time window size based on the results from a purely statistical analysis of autocorrelation or partial correlation (Yaseen et al., 2016; Peng et al., 2017). In this study, we estimate the time window size from an understanding of the physical properties of each region. For streamflow prediction and projection, the response time for precipitation, groundwater and snowpack can range from several hours to months, and a proper time window size should capture all necessary features and avoid redundant information. The seasonality of the streamflow varies regionally and depends on the climatic characteristics and the contribution of snow/ice, and anthropologenic interventions. An investigation of monthly global steamflow (Dettinger and Diaz, 2000) indicated that lags between the peak precipitation and peak steamflow peaks up to 11 months, while 0–3 months was the typical value. In this study we explore 100, 180, and 365 days as different window sizes. The 365-day window corresponds to an entire water year, and so should capture all potential physical processes except for long-term withdrawals or variations in groundwater. The 100-day window captures a typical season length and the 180-day window is in between these two. **Figure 5** shows the ensemble performance results comparing models with different time window sizes.

What stands out in **Figure 5** is the monotonic tendencies in most basins. There are increasing tendencies with the time window size for basins 11224500 (CC), 10343500 (SN), 11264500 (SN), and 11266500 (SN), while 11162500 (BA), 11176400 (BA), 11253310 (CC), 11475560 (NC), and 11528700 (NC) show decreasing tendencies. An increasing tendency implies the presence of slow processes governing streamflow, whereas a decreasing tendency implies upstream processes are fast and there is no significant benefit in using a larger window size. In fact, we can again classify basins into two categories by their monotonic tendencies. Similar with the previous interpretation of different predictors, these results are likely to be related to physical factors, especially snowpack—particularly because of its long response time. In general, the basins with increasing tendencies are in mountainous area like Sierra Nevada and the Coastal Ranges while basins with decreasing tendencies are in Northern California, the Bay Area, and the Central Coast, which are closer to the Pacific. Mountainous areas tend to have more snowpack due to their higher elevation and thus streamflow there is more likely influenced by snowpack. For coastal areas, snowpack does not play a role in streamflow dynamics, and since the temperature is more stable relative to inland areas, the impact from snowpack will also be weaker than that in inland basins. Therefore, snowpack should be the primary factor driving the

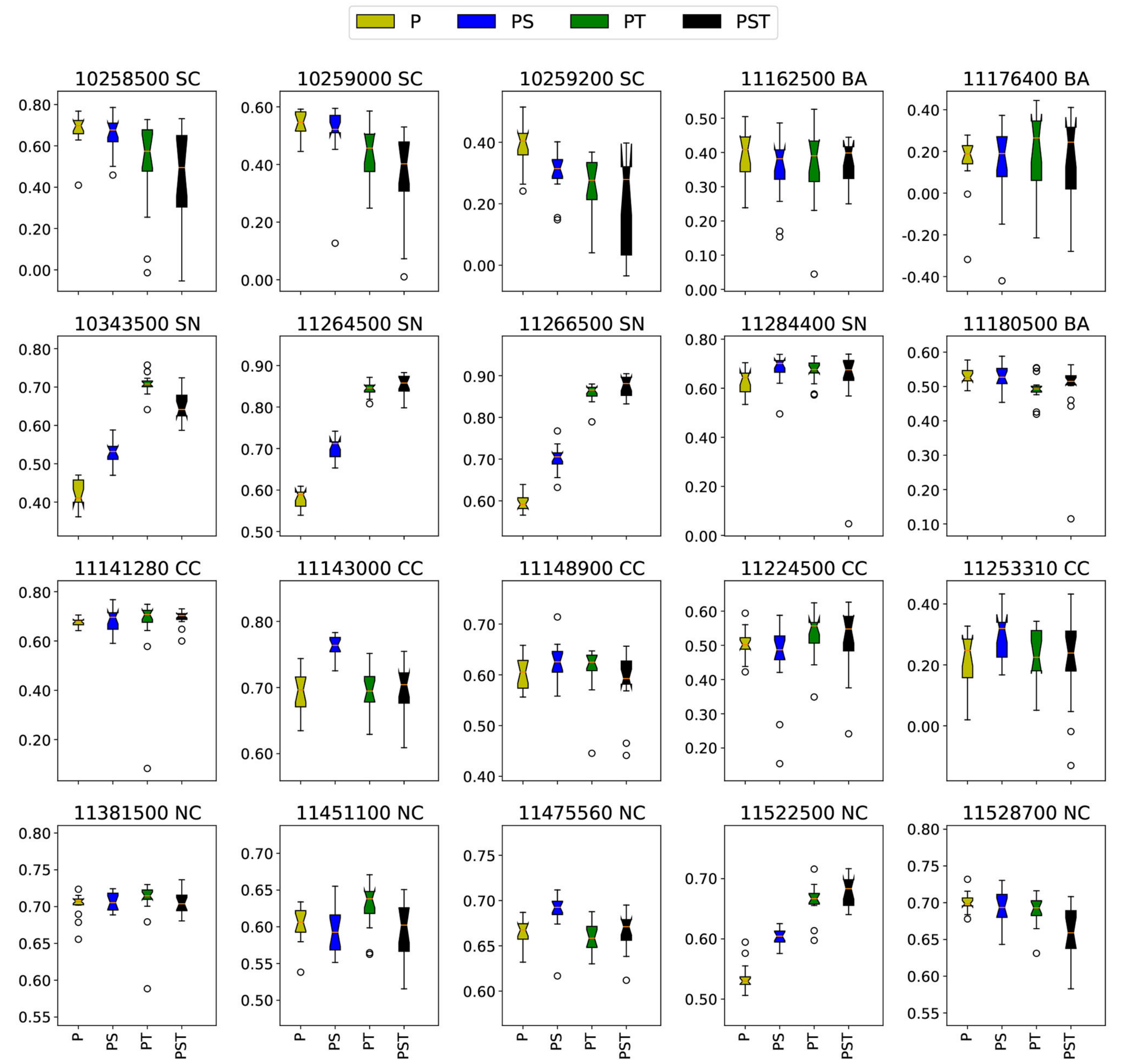


FIGURE 4 | Ensemble prediction comparison for all basins with different reduced TCNN models.

direction of the tendency. Another factor not explored here that may affect the tendency is the groundwater response time—this may play a role in central coast basins such as 11143000 (CC) and 11224500 (CC), which respond positively to increased time window size.

4. PROJECTED STREAMFLOW

The best models from the ensemble run for each basin are now employed with remapped and rescaled LOCA data to produce our projection dataset. As described in section 2.2, we first apply TempestRemap to obtain mean forcing data for irregular basins from the gridded LOCA product. Then both future and historical forcing from LOCA are rescaled (bias corrected) based on the historical observations before being used to drive the ML model. **Tables S6–S8** show the mean daily precipitation, temperature and solar radiation from NLDAS and the four climate models employed. **Figures S7–S9** also show the climatological daily mean of these variables. Generally CanESM2, CNRMCM5, and HadGEM2ES suggest a future wetter climate with more precipitation, while MIROC5 tends to produce similar

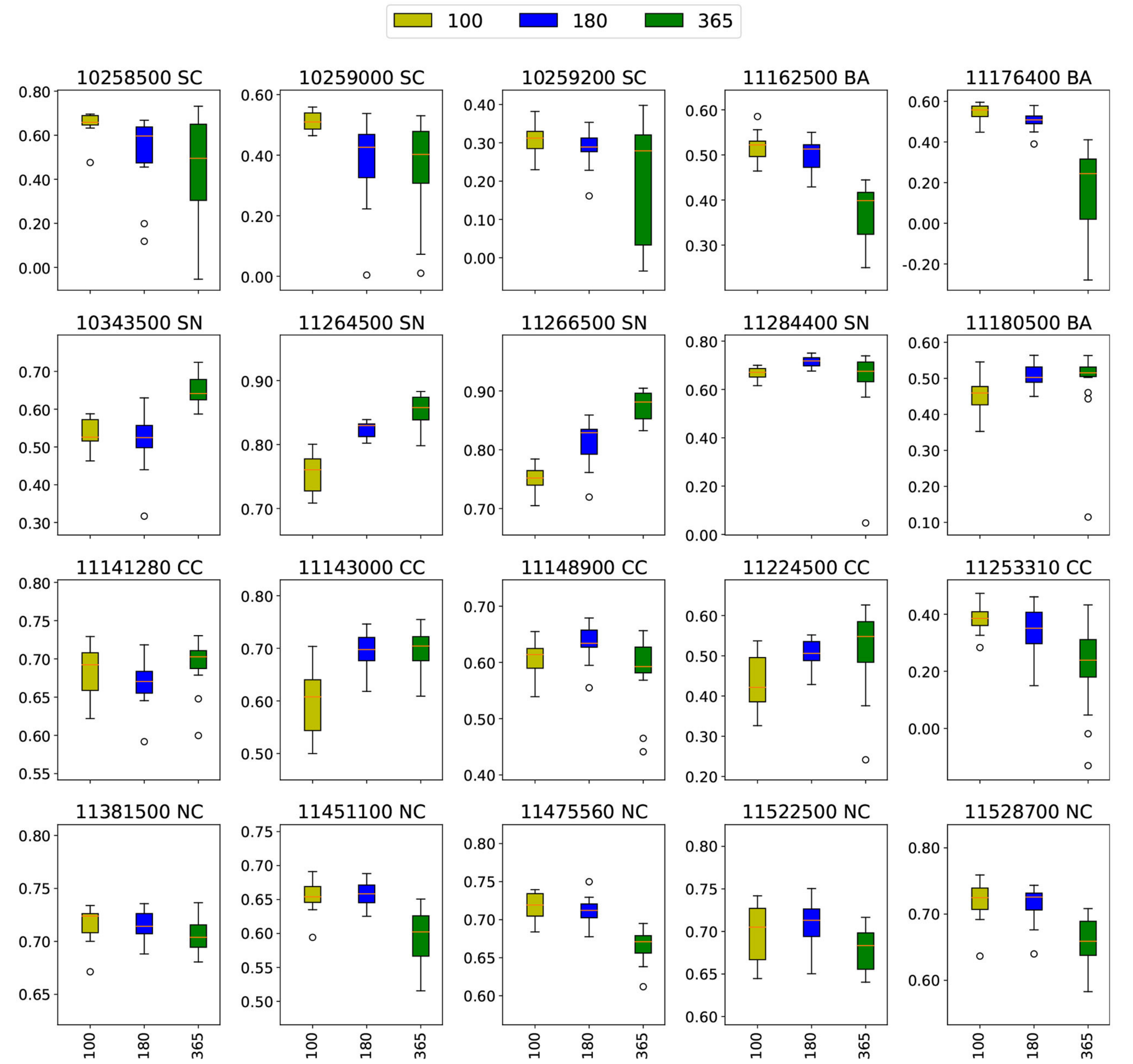


FIGURE 5 | Ensemble prediction comparison for all basins with different window sizes.

or less precipitation for these basins. Essentially all the basins are projected to experience higher daily temperatures, but the change in solar radiation is small. **Figures 6**, **7** show climatological daily streamflow with historical and under the future projection with RCP8.5 forcing. The daily streamflow projection dataset produced in this manner is available at Duan et al. (2020) with the units of millimeters per day. Within the database, each file has the name as the format of "nnnnnnnn-model-scenario.csv." The first eight digits are HUC8 identifiers for each basin, followed by the climate model name, and then the scenario (either "hist" or "RCP8.5").

4.1. Analysis of the Projected Streamflow

Since the historical forcing from different climate models are corrected to match observations (as discussed in section 2.2), historical streamflow exhibits nearly the same pattern and magnitude with forcings from different climate models (**Figure 6**). Compared with USGS observation, the flows tend to match fairly well except in a few SC and SN basins, where a clear magnitude difference at the flow peak emerges. For 10258500(SC) and 10259200(SC), even with the NLDAS forcing data the TCNN underestimates the peak, so we can conclude that the TCNN simply does not identify a relationship

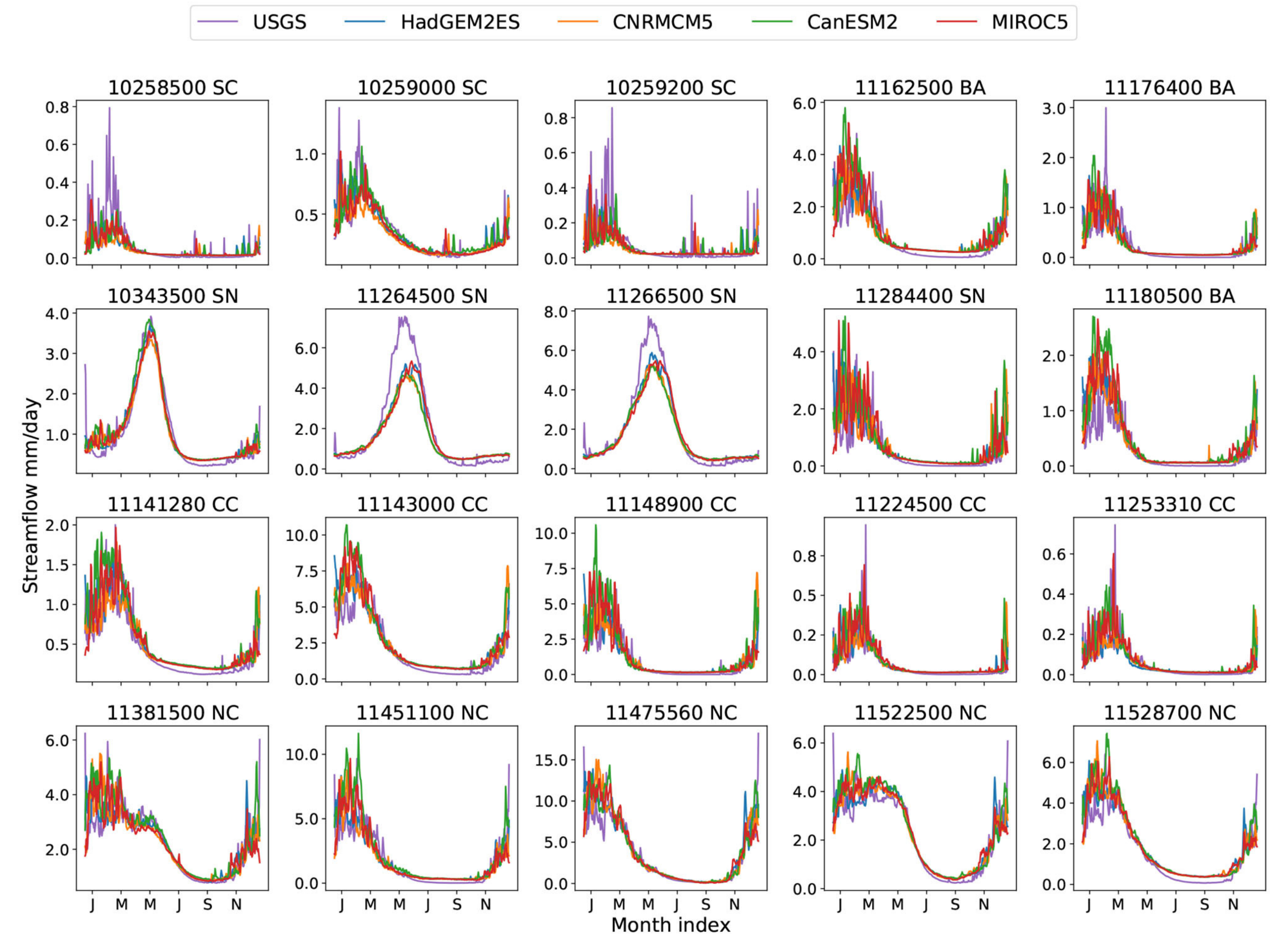


FIGURE 6 | Historical climatological daily streamflow from USGS and four climate models.

between forcing and streamflow during these high flow events. Looking at the SN basins, **Figure 2** shows that the TCNN model achieves an NSE score around 0.9 for 11264500(SN) and 11266500(SN), so in this case the differences are likely due to differences between the forcing from NLDAS vs. LOCA. Namely, we can deduce that for these basins the Gaussian bias correction (2) still produces a forcing which is still somewhat inconsistent with historical forcing. For 11264500(SN) and 11266500(SN) the primary source of this error appears to be wintertime and springtime temperatures, which are intimately connected to precipitation phase and snowpack melt rate; when the LOCA temperatures and radiation are replaced with NLDAS temperatures and radiation (while retaining the LOCA precipitation) the correct streamflow curves are recovered (**Figure S13**).

To assess the magnitude of future change, we examine the projected flow duration curve (FDC) vs. the historical FDC from the same climate model. **Figure 8**, **Figures S14–S16** show the projected future and historical FDCs with four different climate models. When the projected streamflow curve is above the historical curve, the ML model indicates that higher streamflow rates become more probable. It is perhaps not surprising that since precipitation increases across almost all basins, almost all of the basins show increasing streamflow. The projections also generally indicate that the peak flow rate will be higher, potentially indicative of an increased probability of flooding (although the degree to which this is possible is a subject for future investigation). Note that the multimodel CMIP5 ensemble does produce some disagreement: For instance, under the MIROC5 projection, the FDC curves for historical and projection match closely for the most basins. As noted earlier, the MIROC5 model is considered the most unlike the other CMIP5 models in this investigation, tending to produce precipitation amounts that are relatively constant over time.

Although most basins see an increase in flow rate, basins 10343500 (SN), 11264500 (SN), and 11266500 (SN) are notable exceptions. For these three basins, the future FDC curves are sometimes below the historical curves (this is even more obvious with MIROC5 forcing). For basins 11264500 (SN) and 1266500 (SN) lower flow rates become more probable

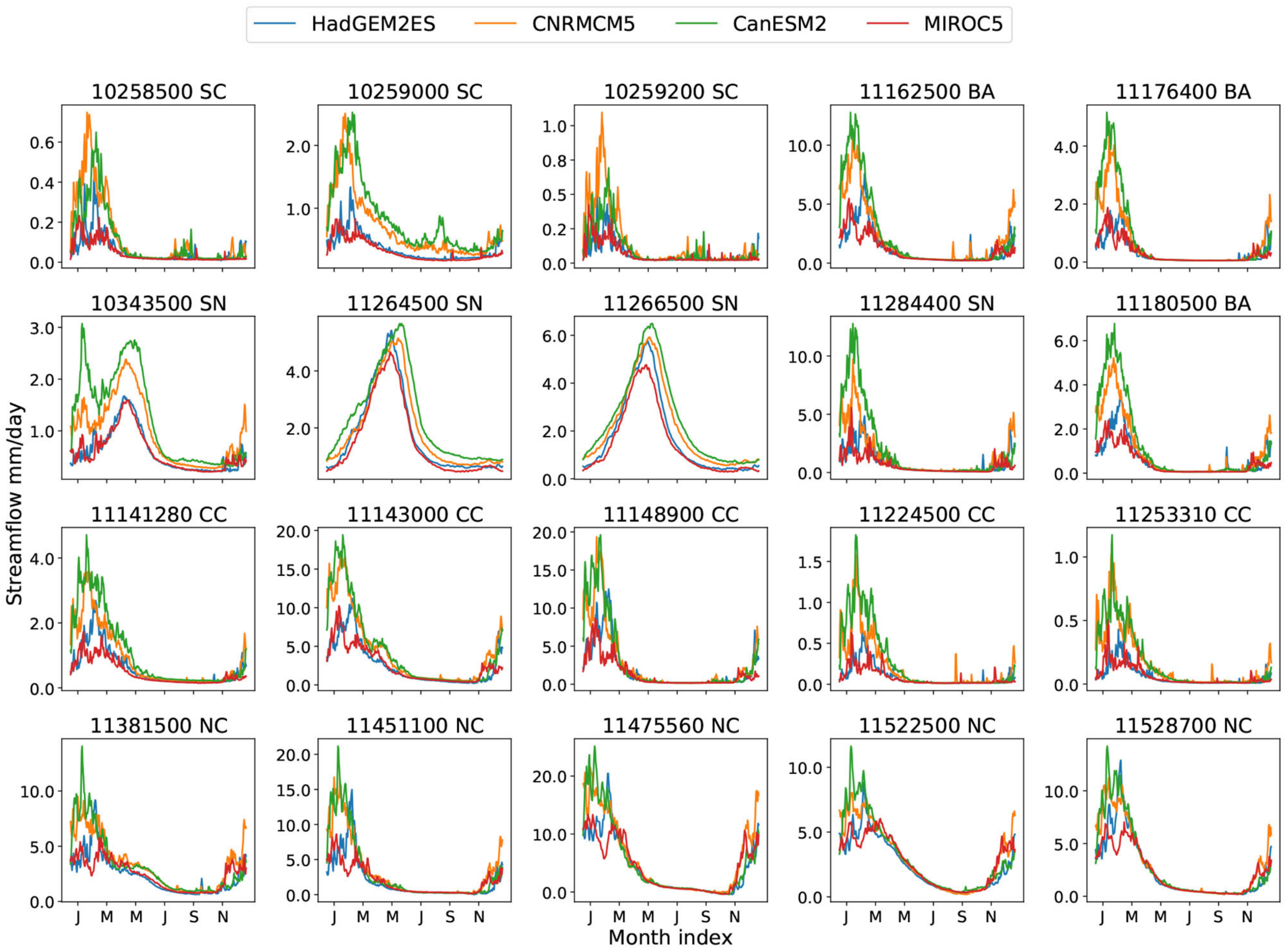


FIGURE 7 | Projection climatological daily streamflow from four climate models.

but the maximum flow rate decreases. These three basins are all in the Sierra Nevada area—10343500 (SN) in the Tahoe National Forest and the other two in Yosemite. Examining **Figures 6**, **7**, these three basins exhibit significant differences in the character of their flow compared with other basins. Namely, the climatological streamflow for these basins shows a peak in late Spring and Summer, while other basins are peaked in the winter season. Since we have shown earlier that streamflow in these basins are driven by snow dynamics, differences in streamflow are likely due to the impact of a slow snowmelt process. Notably, this is in accord with our previous discussion in sections 3.3, 3.4, where these basins are temperature dominant and benefit from longer time window sizes. These projection results lend further evidence to the claim that streamflow in these basins is highly dependent on snowmelt.

The change in the peak flow timing for each basin was also investigated. The peak time is defined as the day of maximal flow rate for the year, measured in days since the beginning of a water year (set to October 1*st* in our study). **Figure 9** shows the peak time for each basin in historical and projection years with MIROC5 forcing. Peak timing figures with forcings from other climate models can be found in **Figures S17–S19**. Although there is generally no significant change in peak timing for most basins, the Sierra Nevada basins are again outliers. Namely, there is a statistically significant shift to earlier peak times in these snowpack-dominated basins. Although it is not always the case for all the climate models, the projected lead of peak time associated with decrease of streamflow in the future again captures the unique hydrology dynamics in the Sierra Nevadas.

4.2. Understanding Nonlinearity in the Projection

To better understand the nonlinearity of the streamflow response to forcing under climate change, we consider a decomposition of the response according to its predictors. Specifically, the impact of precipitation alone on the projected streamflow can be isolated by holding the temperature and solar radiation at historical values while using the future projected precipitation. An analogous approach can then be employed for temperature and solar radiation. By then

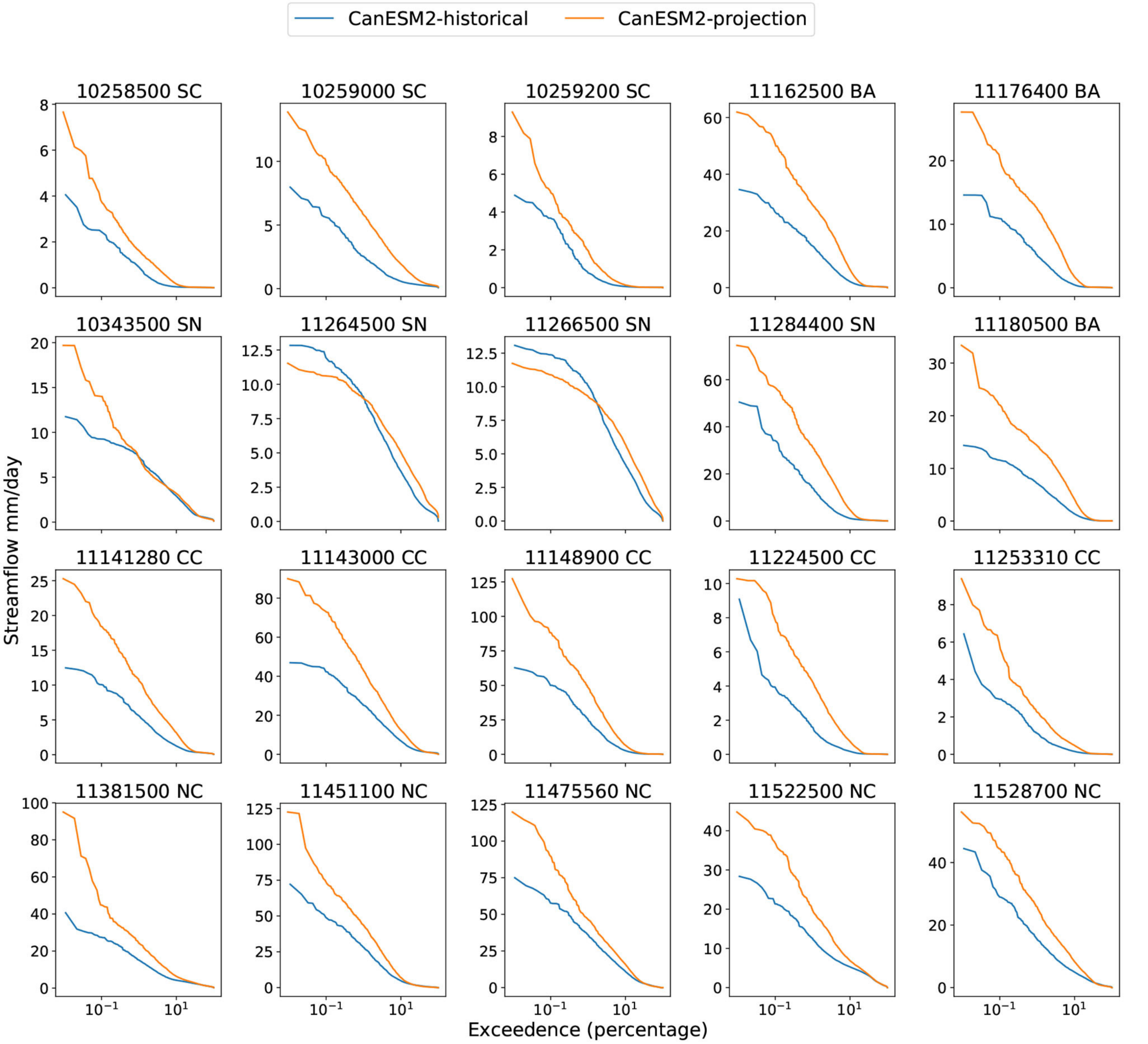


FIGURE 8 | Flow duration curve with CanESM2 forcing over both historical and future (projection) periods.

subtracting the historical streamflow time series from each of these streamflow projections, we obtain ΔQ_p, ΔQ_t, ΔQ_s, the change in streamflow from precipitation alone, temperature alone, and solar radiation alone. These are contrasted against ΔQ_{pts}, which denotes the change in streamflow from all three factors. From the first-order Taylor series expansion we then have

$$\begin{aligned}\Delta Q_{pts} &= \Delta Q_p + \Delta Q_t + \Delta Q_s + r \\ &= \Delta Q_{linear} + r\end{aligned} \tag{7}$$

for some residual r that captures the influence of high-order terms. The linear response is defined as summation of three individual responses. To reduce noise from daily variations in streamflow, the monthly averaged streamflow is used for comparison. In **Figure 10**, we plot ΔQ_{pts} vs. ΔQ_{linear}, with the R^2 value in the title. A fully linear response would be expected to lay along the $y = x$ line.

As seen in **Figure 10**, almost all basins show a nearly linear response to the input variables, except for basin 10343500(SN), 11264500 (SN), and 11266400(SN)—all in the Sierra Nevada mountains. From our discussion in sections 3.3, 3.4, these SN basins are temperature dominated and require a longer time window size to correctly capture streamflow, indicating the interplay between precipitation and temperature in governing snow processes.

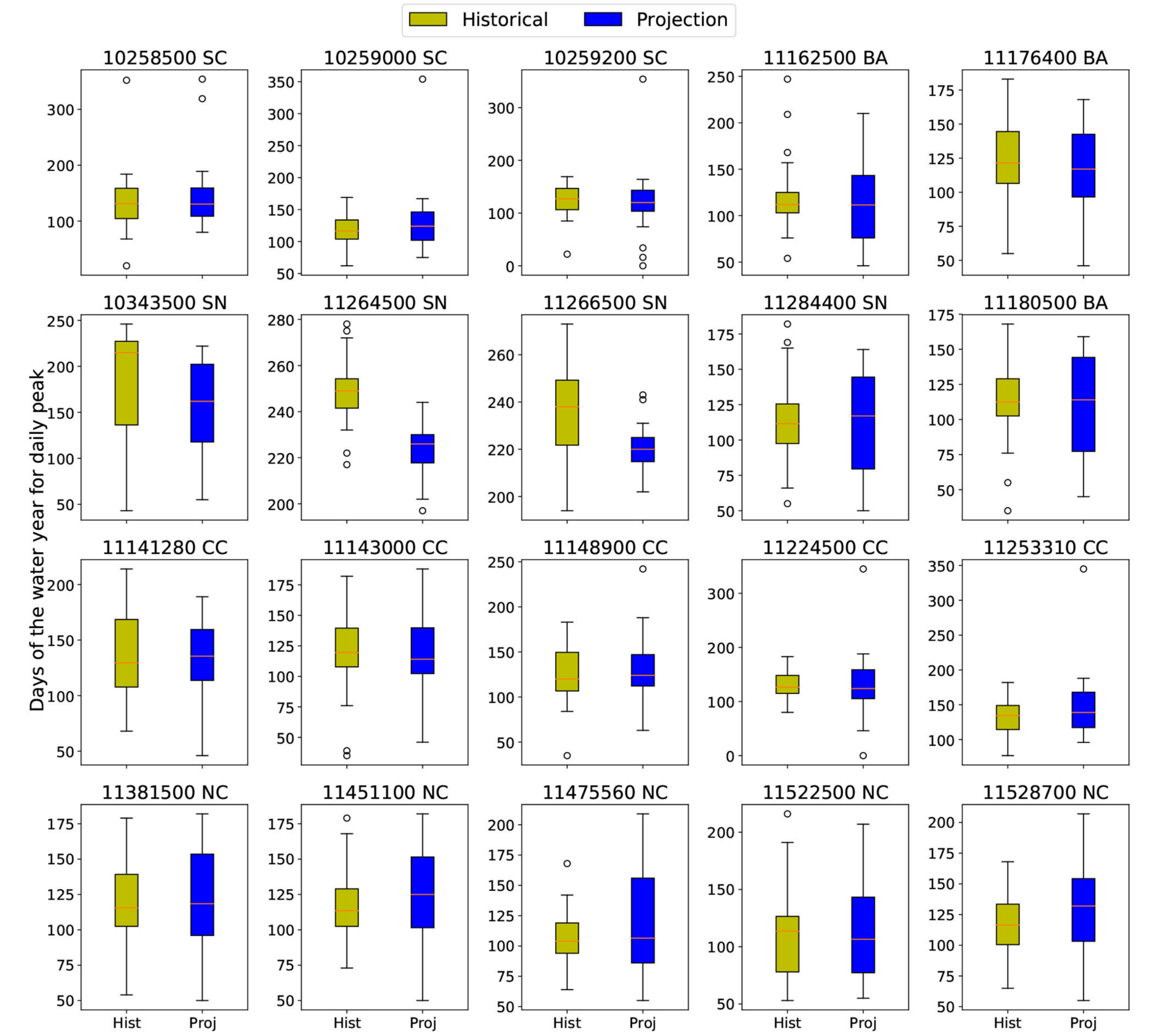


FIGURE 9 | Day of peak flow for each basin with MIROC5 forcing.

5. CONCLUSIONS AND FUTURE WORK

In this study, we have designed and analyzed a general temporal convolutional neural network for streamflow projection in California. Causal convolution is used to maintain physical causation. The input consists of precipitation, temperature, and solar radiation over a particular past window size. In prediction mode, the TCNN model is compared with other commonly used ML models based on ensemble performance so as to eliminate random effects from initializing the training. The results of this intercomparison indicate there are some important temporal features that ANNs struggle to capture, in contrast to TCNNs and other recurrent neural networks (LSTMs and GRUs). Compared with other recurrent networks, the TCNN model is faster and more stable under training. Overall, the TCNN produces better agreement both on average and in the high-flow regime, whereas the LSTM was better in the low-flow regime. Like these other networks, the TCNN model can also be generalized to other basins while maintaining the same architecture.

To demonstrate model stability under extreme forcing, an idealized test with quadruple precipitation and 5 Celsius higher temperature is implemented to verify whether the model produces reasonable results when tested with data outside the training regime. A qualitative analysis and linear regression of

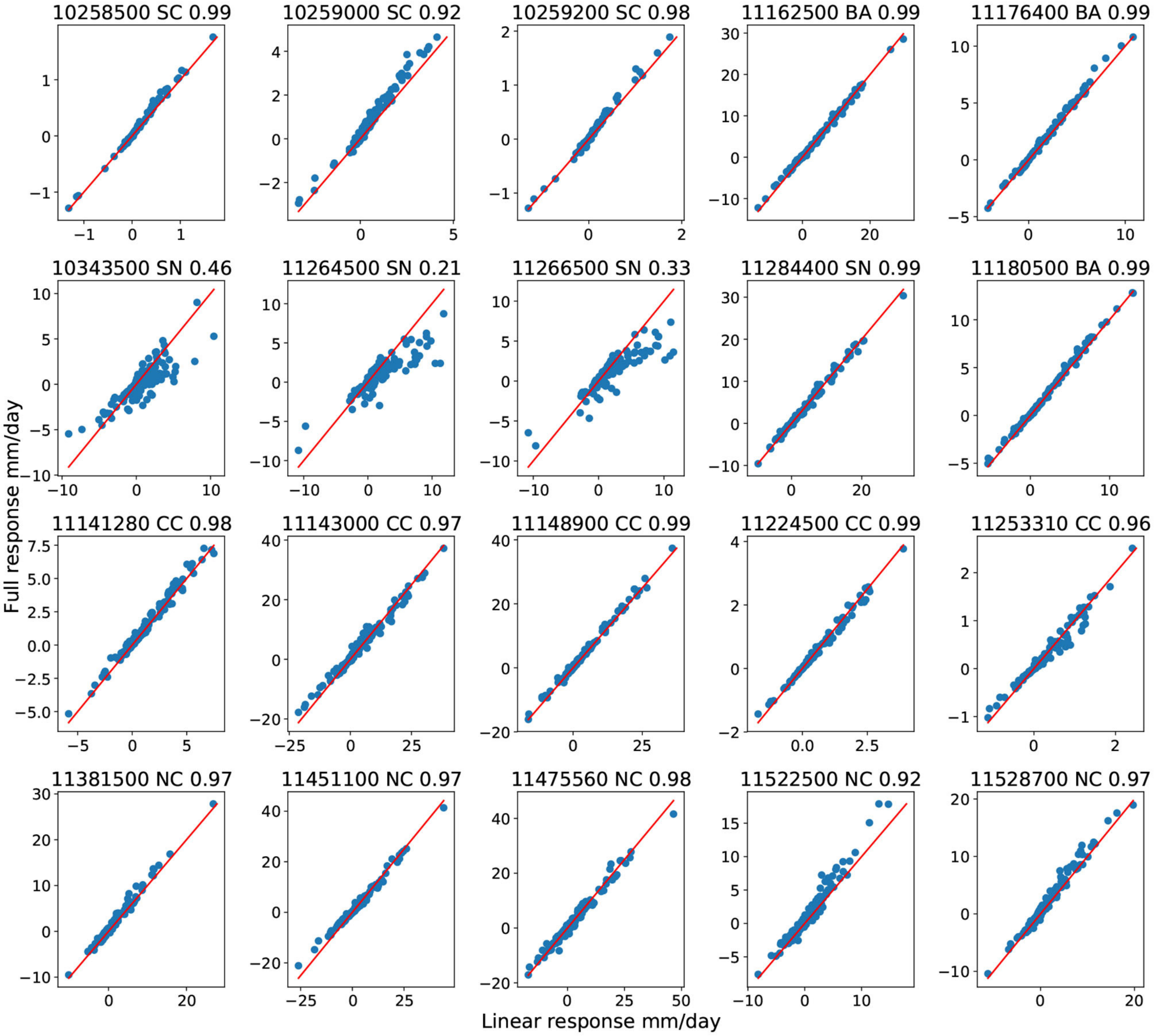


FIGURE 10 | Full response and linear response for all basins with CanESM2 forcing.

projected streamflow against historical precipitation suggests our model produces physically acceptable results for projection.

We have also observed that the TCNN model can build different functional relationships for different basins, as demonstrated through the examination of reduced models, models with different time window sizes, and the nonlinear response of the model to input variables. With this understanding of the "under the hood" workings of the ML model, we can distinguish different geographic features across basins. This classification ability suggests our model can simulate physical processes with causal convolution as a constraint. In regions where snowpack is relevant, we conclude that temperature should be included as a model covariate; whereas in coastal regions, solar radiation should be included. Including both variables was not observed to significantly improve model performance in any basin. Also, in regions where snowpack is relevant, a longer time window size is desirable for model performance (here we tested a 365-day window), whereas in other regions a shorter time window of 100-days produced better results.

Under the RCP8.5 scenario, the nonlinearity of the streamflow response was examined by decomposing the response into three modes by the predictors. By inspecting the linear response and full response, we observed that most basins exhibit a linear response from precipitation, temperature, and solar radiation, except for the basins in Sierra Nevada. The nonlinearity is likely associated with snowpack, which is a physical feature that is sensitive to both precipitation and temperature.

Model results for future projections and historical hindcasts were compared to understand the changing character of the streamflow. Generally streamflow in most basins increases through the end of the century, except for the Sierra Nevada basins. Peak flow time remained statistically indistinguishable among most basins, except the Sierra Nevada basins which showed a shift to earlier dates under some models. These results further indicate that the snow dynamics in the Sierra Nevada is important for correctly capturing streamflow in these basins.

The idealized test here mainly deals with the problem of model stability under extrapolation. In terms of ensuring the model produces physically plausible results under extreme forcings, we need to compare with a physically based model with the same extreme forcing. This problem has been saved for our future work. Also, to better understand the ML model and ensure its credibility for producing future projections, we intend to next cross-validate our projection datasets with a physically-based model over the same time period. Model credibility can also be enhanced through alternative designs that explicitly include physically-based conservation laws. For instance, subsurface flow or evaporation are not produced as outputs, and so validation of the water budget is impossible. With a more complicated design, ML models could predict streamflow, evaporation and groundwater, and be constrained via an appropriate physically-based conservation law. Such constraints would further enable physical interpretation of the model results. Finally, we wish to determine if the TCNN can be used to interpolate predictors to higher temporal resolution, for use (for instance) in physically-based models. The ML model could also be used to examine model performance when the strict causation is relaxed (namely, if future streamflow could provide a better estimate of present streamflow).

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 in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SD, PU, and LS designed the model. SD and PU designed the experiments and wrote the manuscript. SD carried out the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by California Energy Commission grant Advanced Statistical-Dynamical Downscaling Methods and Products for California Electrical System project (award no. EPC-16-063) and the U.S. Department of Energy Regional and Global Climate Modeling Program (RGCM) An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US project (award no. DE-SC0016605).

ACKNOWLEDGMENTS

We wish to thank David Pierce and Sam Iacobellis for their help with LOCA dataset and Andrew Newman for his help with CAMELS dataset. We would like to thank MetroIT in UC Davis for the help with the GPU cluster.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2020.00028/full#supplementary-material

REFERENCES

Atieh, M., Taylor, G., Sattar, A. M., and Gharabaghi, B. (2017). Prediction of flow duration curves for ungauged basins. *J. Hydrol.* 545, 383–394. doi: 10.1016/j.jhydrol.2016.12.048

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13, 281–305.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. doi: 10.3115/v1/D14-1179

Csáji, B. C. et al. (2001). *Approximation With Artificial Neural Networks*. Faculty of Sciences, Etvs Lornd University, Hungary.

Dettinger, M. D., and Diaz, H. F. (2000). Global characteristics of stream flow seasonality and variability. *J. Hydrometeorol.* 1, 289–310. doi: 10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2

Duan, S., Ullrich, P., and Shu, L. (2020). *California streamflow projection dataset.* Zenodo. doi: 10.5281/zenodo.3823273

Feng, D., Fang, K., and Shen, C. (2019). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *arXiv preprint arXiv:1912.08949*. doi: 10.1029/2019WR026793

Gao, C., Gemmer, M., Zeng, X., Liu, B., Su, B., and Wen, Y. (2010). Projected streamflow in the Huaihe river basin (2010-2100) using artificial neural network. *Stochast. Environ. Res. Risk Assess.* 24, 685–697. doi: 10.1007/s00477-009-0355-6

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recogn.* 77,354–377. doi: 10.1016/j.patcog.2017.10.013

Hassoun, M. H. et al. (1995). *Fundamentals of Artificial Neural Networks*. MIT Press.

Huang, S., Chang, J., Huang, Q., and Chen, Y. (2014). Monthly streamflow prediction using modified emd-based support vector machine. *J. Hydrol.* 511, 764–775. doi: 10.1016/j.jhydrol.2014.01.062

Huang, X., and Ullrich, P. A. (2017). The changing character of twenty-first-century precipitation over the western united states in the variable-resolution CESM. *J. Clim*. 30, 7555–7575. doi: 10.1175/JCLI-D-16-0673.1

Kisi, O., and Cimen, M. (2011). A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *J. Hydrol*. 399, 132–140. doi: 10.1016/j.jhydrol.2010.12.041

Kisi, O., and Kerem Cigizoglu, H. (2007). Comparison of different ANN techniques in river flow prediction. *Civil Eng. Environ. Syst*. 24, 211–231. doi: 10.1080/10286600600888565

Koirala, S., Hirabayashi, Y., Mahendran, R., and Kanae, S. (2014). Global assessment of agreement among streamflow projections using CMIP5 model outputs. *Environ. Res. Lett*. 9:064017. doi: 10.1088/1748-9326/9/6/064017

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *arXiv preprint arXiv:1907.08456*. doi: 10.5194/hess-2019-368

Krzywinski, M., and Altman, N. (2014). Visualizing samples with box plots: use box plots to illustrate the spread and differences of samples. *Nat. Methods* 11, 119–121. doi: 10.1038/nmeth.2813

Le, X.-H., Ho, H. V., Lee, G., and Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* 11:1387. doi: 10.3390/w11071387

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (2017). "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 156–165. doi: 10.1109/CVPR.2017.113

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). "The expressive power of neural networks: a view from the width," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 6231–6239.

Nash, J. E., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I-a discussion of principles. *J. Hydrol*. 10, 282–290. doi: 10.1016/0022-1694(70)90255-6

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., and Blodgett, D. (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR.

Noori, N., and Kalin, L. (2016). Coupling swat and ann models for enhanced daily streamflow prediction. *J. Hydrol*. 533, 141–151. doi: 10.1016/j.jhydrol.2015.11.050

Peng, T., Zhou, J., Zhang, C., and Fu, W. (2017). Streamflow forecasting using empirical wavelet transform and artificial neural networks. *Water* 9:406. doi: 10.3390/w9060406

Pierce, D. W., Cayan, D. R., and Thrasher, B. L. (2014). Statistical downscaling using localized constructed analogs (loca). *J. Hydrometeorol*. 15, 2558–2585. doi: 10.1175/JHM-D-14-0082.1

Pierce, D. W., Kalansky, J. F., and Cayan, D. R. (2018). *Climate, drought, and sea level rise scenarios for California's fourth climate change assessment*. Technical report, Technical Report CCCA4-CEC-2018-006, California Energy Commission.

Rasouli, K., Hsieh, W. W., and Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol*. 414, 284–293. doi: 10.1016/j.jhydrol.2011.10.039

Shanker, M., Hu, M. Y., and Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega* 24, 385–397. doi: 10.1016/0305-0483(96)00010-2

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

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: the all convolutional net. *arXiv preprint arXiv:1412.6806*.

Swain, D. L., Langenbrunner, B., Neelin, J. D., and Hall, A. (2018). Increasing precipitation volatility in twenty-first-century California. *Nat. Clim. Change* 8, 427–433. doi: 10.1038/s41558-018-0140-y

Tsymbal, A. (2004). *The Problem of Concept Drift: Definitions and Related Work*. Computer Science Department, Trinity College Dublin.

Ullrich, P. A., Devendran, D., and Johansen, H. (2016). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part II. *Mon. Weather Rev*. 144, 1529–1549. doi: 10.1175/MWR-D-15-0301.1

Ullrich, P. A., and Taylor, M. A. (2015). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I. *Mon. Weather Rev*. 143, 2419–2440. doi: 10.1175/MWR-D-14-00343.1

Ullrich, P. A., Xu, Z., Rhoades, A. M., Dettinger, M. D., Mount, J. F., Jones, A. D., et al. (2018). California's drought of the future: a midcentury recreation of the exceptional conditions of 2012-2017. *Earth's Fut*. 6, 1568–1587. doi: 10.1029/2018EF001007

White, A. B., Moore, B. J., Gottas, D. J., and Neiman, P. J. (2019). Winter storm conditions leading to excessive runoff above California's Oroville Dam during January and February 2017. *Bull. Am. Meteorol. Soc*. 100, 55–70. doi: 10.1175/BAMS-D-18-0091.1

Yan, L., Feng, J., and Hang, T. (2019). "Small watershed stream-flow forecasting based on LSTM," in *International Conference on Ubiquitous Information Management and Communication* (Phuket: Springer), 1006–1014. doi: 10.1007/978-3-030-19063-7_79

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., et al. (2016). Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. *J. Hydrol*. 542, 603–614. doi: 10.1016/j.jhydrol.2016.09.035

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 in Water

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/frwa.2020.560598

3D Geological Image Synthesis From 2D Examples Using Generative Adversarial Networks

Guillaume Coiffier[1]*, *Philippe Renard*[2] and *Sylvain Lefebvre*[3]

[1] Centre for Hydrogeology and Geothermics, École Normale Supérieure de Lyon, Lyon, France, [2] Computer Science Department, University of Neuchâtel, Neuchâtel, Switzerland, [3] Université de Lorraine, CNRS, INRIA, Loria, France

Generative Adversarial Networks (GAN) are becoming an alternative to Multiple-point Statistics (MPS) techniques to generate stochastic fields from training images. But a difficulty for all the training image based techniques (including GAN and MPS) is to generate 3D fields when only 2D training data sets are available. In this paper, we introduce a novel approach called Dimension Augmenter GAN (DiAGAN) enabling GANs to generate 3D fields from 2D examples. The method is simple to implement and is based on the introduction of a random cut sampling step between the generator and the discriminator of a standard GAN. Numerical experiments show that the proposed approach provides an efficient solution to this long lasting problem.

Keywords: geology, heterogeneity, stochastic model, groundwater, generative adversarial network, deep learning

OPEN ACCESS

Edited by:
Eric Laloy,
Belgian Nuclear Research Centre, Belgium

Reviewed by:
Kirill M. Gerke,
Russian Academy of Sciences (RAS), Russia
Zhaoyang Larry Jin,
Stanford University, United States
Ahmed H. Elsheikh,
Heriot-Watt University, United Kingdom

***Correspondence:**
Guillaume Coiffier
guillaume.coiffier@inria.fr

Specialty section:
This article was submitted to Water and Hydrocomplexity, a section of the journal Frontiers in Water

Received: 09 May 2020
Accepted: 17 August 2020
Published: 28 October 2020

Citation:
Coiffier G, Renard P and Lefebvre S (2020) 3D Geological Image Synthesis From 2D Examples Using Generative Adversarial Networks. Front. Water 2:560598. doi: 10.3389/frwa.2020.560598

1. INTRODUCTION

For a wide range of problems in the hydrological sciences, there is a need to employ stochastic models to generate spatial fields. These fields can represent for example climatic data, or physical parameters of the atmosphere, ground or underground.

In that framework, multiple-points statistics and the concept of training image became very popular in the last 10 years (Journel and Zhang, 2006; Hu and Chugunova, 2008; Mariethoz and Caers, 2014; Linde et al., 2015). The key idea in these approaches is to use an exhaustively mapped example (the training image) of the type of spatial patterns that are expected to occur for a given variable and at a given scale. The training image is then used to train a spatial statistics non-parametric model. The main advantage of that approach is that it allows to transfer information about spatial patterns coming from external information such as an analog site or data set to constrain the stochastic model. The range of applications is very broad and includes subsurface hydrophysical parameters (Mariethoz et al., 2010; Barfod et al., 2018), rainfall simulation (Oriani et al., 2014), bedrock topography below glaciers (Zuo et al., 2020), soil properties (Meerschman et al., 2014), landforms attributes (Vannametee et al., 2014), etc.

More recently, Deep Learning algorithms and especially Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) have sparked a very strong interest thanks to their ability to generate stochastic fields showing a high degree of similarity with the training data sets (Chan and Elsheikh, 2017; Mosser et al., 2017a,b; Laloy et al., 2018). While this is very similar in principle to Multiple Point Statistics (MPS), a key feature of GAN is that they can be trained to represent a mapping from a low dimensional space to the manifold that supports the training data set. This feature allows GAN to represent complex fields via a low dimension vector of continuous values. Such a parametrization can then be used for example in the context of inverse modeling (Laloy et al., 2018). Although they need a long training time, GAN also appear to be usually faster than previous MPS methods at generating a set of realizations.

When applied to underground hydrology, and especially for uncertainty analysis or to solve inverse problems, a difficulty with the MPS or GAN approaches is to obtain 3D training images or examples. In practice, it is often very difficult, if not impossible, to collect exhaustive and accurate data about the three dimensional distribution of rock types (or physical properties) at depth. To circumvent that problem, in many applications, the 3D training images are derived from other types of models such as object based or process based models (de Marsily et al., 2005). But this is not always satisfying because the object based models imply additional assumptions and are not directly based on data when these are available in 2D.

Indeed, many two-dimensional data sets are available. They are much easier to collect than the 3D data sets using for example: remote sensing techniques, direct observation on outcrops, or microscopic data acquisition on thin sections of rocks. Furthermore, geologists are used to draw conceptual cross sections of typical structures. Therefore, simulating 3D stochastic fields from 2D examples is of high practical importance, and different techniques have been developed to solve that problem. In the framework of MPS, the approaches are often based on probability aggregation techniques or use some successive 2D simulation techniques (Okabe and Blunt, 2007; Comunian et al., 2012; Kessler et al., 2013; Cordua et al., 2016; Chen et al., 2018). In rock physics, the simulation of 3D porous media from 2D sections has been addressed in numerous articles (e.g., Adler et al., 1990; Yeong and Torquato, 1998; Karsanina and Gerke, 2018). Simulated annealing is often used in this context and allows to generate random fields which reproduce specific morphological features such as correlation functions or connectivity curves derived from the 2D data sets (e.g., Gerke and Karsanina, 2015; Lemmens et al., 2019). Deep Learning techniques have been used to accelerate MPS simulations in this framework (Feng et al., 2018). The 3D synthesis of textures from 2D examples has also been a main research topic in the field of computer graphics as shown in the review paper of Wei et al. (2009). These optimization methods achieve good results in the unstructured or weakly structured cases, but fail to capture long-range correlations. Works around GANs aiming at inferring a fixed 3D structured shape out of 2D projections has also been conducted (Gadelha et al., 2017, 2019), demonstrating the ability of these algorithms to infer an approximation of the three-dimensionnal shape of a deterministic object (like a plane or chair) out of a set of 2D views from a 3D scene. Here our problem is slightly different since we aim at inferring the statistical distribution of stochastic geological structures, which can have long range correlations, from a limited set of cross-sections through the domain.

In this paper, we introduce a novel approach based on GAN, called Dimension Augmenter GAN (DiAGAN). It allows to generate 3D fields from 2D training images, with sufficient resemblance and variability for geostatistical applications. The paper introduces first the general principle of GAN, then it describes the DiAGAN approach and illustrates it with a few examples.

2. METHODOLOGY

2.1. Generative Adversarial Networks

In the broadest sense, Machine Learning consists of designing algorithms that automatically find complex mappings from a given input data set $\mathcal{X}$ to a given target data set $\mathcal{Y}$. In practice, $\mathcal{X}$ is often infinite and intractable, and we only have access to a finite set of training examples with their associated mapped values $\mathcal{D} = \{(x_1, y_1), ...(x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$. Training algorithms then aim at finding the parametrized function f_ω that minimizes a loss function $\mathcal{L}$ defined over $\mathcal{Y} \times \mathcal{Y}$:

$$f_\omega = \arg\min_{\eta} \sum_{x,y \in \mathcal{D}} \mathcal{L}(f_\eta(x), y)$$

with the hope of also minimizing the loss over $\mathcal{X}$, which would lead to good performances on new unknown data.

In Deep Learning more specifically, the parameterized function f_ω takes the form of a deep neural network: a structure made of an alternation of parameterized linear transformations called layers and non-linear component-wise activation functions. Neural networks are designed to be differentiable, so that their parameters can be optimized (trained) through gradient descent algorithms. Deep neural networks in their various forms had a deep impact in many domains in computer science with state of the art performance, including image recognition, natural language processing, data classification or artificial intelligence.

Technically, as long as a correct loss function is defined, that is to say a function that is differentiable and reaches an optimum when the desired mapping is achieved, any transformation from $\mathcal{X}$ to $\mathcal{Y}$ can be learned, provided enough training examples in $\mathcal{D}$ are given. In the context of procedural image synthesis, one difficulty is to define a loss function that assesses the quality of the generated samples in relation to the training image. Classical distances on the space of images, like the pixel-wise L_2 norm indeed fail to capture the notion of resemblance between two images (note that in the following text we use the word image both for standard 2D images and 3D grids made of voxels). One has to define a loss function that takes into account multi-scale features and must be robust for instance to the fact that certain geological objects or patterns can be placed anywhere in the image, as their shape and frequency needs to be similar with the training image while their location is not fixed.

Generative Adversarial Networks (GAN) are a family of Deep Learning algorithms designed to tackle this problem. One of the key ideas here is that comparing two images can be done using a neural network f_ω that takes an image x as an input, and computes a numerical score $f_\omega(x)$ such that the higher the score, the more confident the network is of being fed with an image from the dataset it was trained with. Such a network f_ω, called a *critic*, can be plugged on the output of an image generator g_θ, in order to give relevant numerical feedback.

More formally, the problem of image synthesis is the following. Given the set of all possible images $\mathcal{X}$ of a certain size, we have access to a finite subset of training images (TI) $\mathcal{D}$. We suppose that images from $\mathcal{D}$ were sampled from a probability distribution $\mathbb{P}_{data}$ over $\mathcal{X}$. From there, the goal is to be able to sample any image following the same probability distribution $\mathbb{P}_{data}$. Because neural networks are essentially mappings, this is reduced in the case of GANs to finding a mapping from a latent space with a known distribution $\mathbb{P}_z$ to the support space of $\mathbb{P}_{data}$, under the constraint that $\mathbb{P}_\theta = g_\theta(\mathbb{P}_z)$ is as close as possible from $\mathbb{P}_{data}$. Since $\mathbb{P}_{data}$ and $\mathbb{P}_\theta$ are intractable and can only be sampled from, the generator g_θ and the critic f_ω have to be trained simultaneously in an adversarial fashion. Intuitively, one can see this process as if on the one hand, the generator tries to fool the critic by minimizing the distance between generated examples and real training examples. On the other hand, the critic is trained to separate training images from fake ones, thus maximizing the distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{data}$.

In the original GAN algorithm, Goodfellow et al. (2014) considered the Kullback-Leibler (KL) divergence as a distance between probability distributions. We chose to follow more recent works using the Wasserstein-1 distance, or Earth-mover (EM) distance (Arjovsky et al., 2017; Gulrajani et al., 2017), in an algorithm called the Wasserstein GAN (WGAN). This ultimately lead us to a two player zero-sum game using the following objective function (Gulrajani et al., 2017):

$$\min_\theta \max_\omega \underbrace{\mathbb{E}_{x\sim\mathbb{P}_{data}}[f_\omega(x)] - \mathbb{E}_{z\sim\mathbb{P}_z}[f_\omega(g_\theta(z))]}_{\text{EM distance}} - \lambda \underbrace{\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(||\nabla_{\hat{x}} f_\omega(\hat{x})||_2 - 1)^2]}_{\text{gradient penalty term}} \tag{1}$$

Equation (1) is directly used as a loss function during the training phase of a WGAN, during which theoretical expectations of probability distributions are replaced by statistical means over sampled examples. From the generator's point of view, the parameters are optimized to minimize $-\mathbb{E}_{z\sim\mathbb{P}_z}[f_\omega(g_\theta(z))]$, while the critic's parameters are optimized with relation to the whole expression. In other words, g_θ is trained to obtain the greatest possible score from the critic, whereas the critic optimizes both terms of Equation (1). The first term boils down to maximizing the EM distance between TIs, which are associated to high values of score, and generated images, associated to low values. However, this expression of the EM distance term only holds for f_ω being a Lipschitz function. The second term, called gradient penalty, proposed by Gulrajani et al. (2017), is a way to enforce this constraint on f_ω. In Equation (1), $\mathbb{P}_{\hat{x}}$ is a mix of $\mathbb{P}_\theta$ and $\mathbb{P}_{data}$. More specifically, $\hat{x} \sim \mathbb{P}_{\hat{x}}$ means that the variable $\hat{x}$ was sampled from the distribution $\varepsilon\mathbb{P}_{data} + (1-\varepsilon)\mathbb{P}_\theta$ with ε being uniformly chosen in [0;1].

For the noise distribution $\mathbb{P}_z$, we consider vectors of $\mathbb{R}^d$ where each coordinate is sampled independently from a standard normal distribution. Given the parameters of the generator, such a latent vector contains the whole information about the generated image, thus offering a compact representation. For generating images of size 64x64x64, we determined experimentally that d=256 was sufficient to represent the whole distribution.

2.2. From 2D to 3D

Going from 2D training examples to 3D realizations with a GAN is not straightforward, as the dimensionality of generated and training images should match in order for them to be fed to the same critic network. Most GAN algorithms directly feed the generated image inside the critic, but we propose to add an intermediate step to transform generated images, in order for them to match the TIs. Our training data consists of a set of triplets (C_x,C_y,C_z), where C_i is a two-dimensional image representing a typical cut perpendicular to axis *i*. The generator aims at generating 3D images which cuts along axis *i* resembles C_i. Note that the cuts are provided as independent images. They are not crossing each other at specific locations and they may not be completely compatible as will discuss in one of the examples.

To make the critic compare cuts, we incorporated a random cut sampler between the generator and the critic (**Figure 1**). This sampler extracts from a 3D generated example a triple of cuts (C'_x,C'_y,C'_z), with C'_i being chosen uniformly among all possible cuts along the corresponding axis. The discriminator then proceeds in comparing the patterns in the two triplets of cuts. The comparison is done for each direction but it does not account for the compatibility of the patterns along the intersection of the cuts since this information is not available in the training data. Depending on the type of symmetry and on available data, a similar technique can be used to account for one single set of cuts when the 3D patterns display similar structures along the *x*, *y*, and *z* directions, or instead a set of cuts along two axis when examples are available only along these two directions.

The sampler select the positions of the cuts randomly in a uniform distribution along each axis of the 3D domain, with only one cut per dimension. This mean that most of a generated image in the 3D domain will not be fed into the critic. However, since this sampling is random and thanks to the continuity of the GAN, this strategy proved to be rather efficient as we will illustrate with a few examples. Other approaches that may involve additional computations can be envisioned, this will be discussed in section 4.

In summary, the whole approach is stochastic and assumes that the 2D images that have been given in input can represent a cut anywhere in the domain. DiAGAN aims at reproducing these patterns in a stochastic manner, and assuming a spatial stationarity of the statistical process. In particular, it will not check precisely and in a deterministic manner the compatibility of the cuts at their intersection.

2.3. Neural Network Architecture

Following architectures proposed by Radford et al. (2015), we use convolutional neural networks for both our generator and critic. Throughout our experiments, we determined that precise architecture tuning was not necessary to get satisfactory results, as the WGAN is quite robust to those kind of hyperparameters. The main principles we used was to alternate convolutionnal layers, eventually normalization and upscaling, and non-linear activation.

Training images are normalized so that their values lay in [0;1]. Latent vectors *z* are sampled from a normal distribution of zero mean and standard deviation of 0.5. *z*

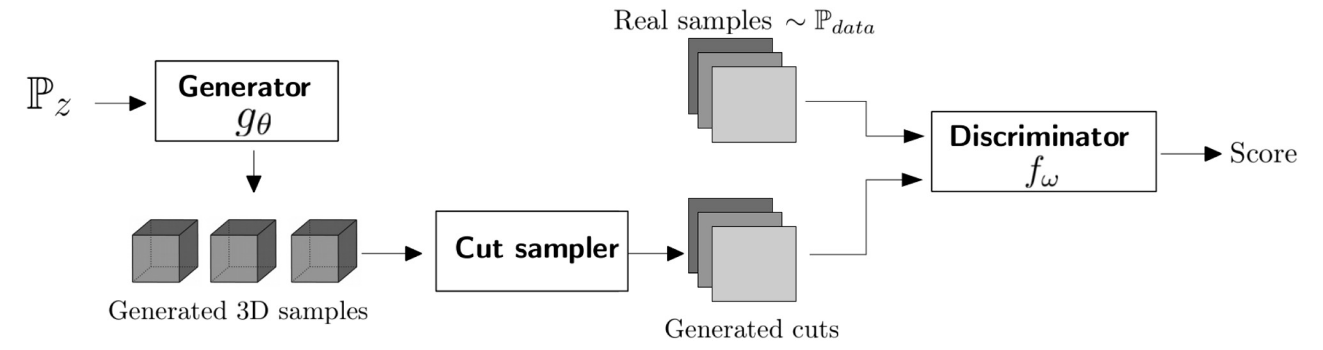


FIGURE 1 | The overall organization of DiAGAN.

is then reshaped into a 3-dimensional tensor that is fed into the convolutions. One can use regular convolution coupled with an upscaling function (like a trilinear interpolation), or a transposed convolution with a stride parameter greater than 1. The number of convolutional feature maps is decreasing with the depth, being divided by two at each layer.

We use the ReLU activation function ($x \mapsto max(0, x)$) for every internal layers, and a sigmoid $x \mapsto 1/(1 + exp(-x))$ for the final activation, in order to project the values back in the interval [0;1]. Alternatively, one could consider a normalization in [-1;1], Leaky ReLU function of parameter α=0.2 ($x \mapsto ReLU(x) - \alpha ReLU(-x)$), and an hyberbolic tangent as the final activation.

The discriminator is composed of convolutional layers with 2D kernels. After the cut sampling step, we are left with 3 images of size NxN stacked into a 1xNx3N tensor, or a 3xNxN. 2D convolutions are applied to this tensor, alternating with max pooling operations. This halves the size of the tensor while the number of feature maps in the convolution is doubled. Alternatively again, one could replace the poolings by strides in the convolution. Activation functions are also ReLU of Leaky ReLU. At the end, a global max pooling operation is applied, to retrieve one numerical value for each feature map. The obtained vector of features is aggregated into a single numerical value by a dense layer.

We use instance normalization layers (Ulyanov et al., 2016) or batch normalization (Ioffe and Szegedy, 2015) after each convolution to guarantee the stability of the computation. The two methods performed equally well. Those normalization layers are present in the generator before each activation layers, but not in the critic, as they mess with the gradient penalty term (Gulrajani et al., 2017).

Table 1 summarizes the architecture of the convolutional networks that we use for DiAGAN. But note that the overall method is pretty robust, we tested several other architectures that worked also reasonably well. We expect that many other architectures could give good results.

Both generator and discriminator were trained using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$.

TABLE 1 | Basic architecture used for DiAGAN with images of size 64 × 64 × 64 and a normalization of the TIs in [0;1].

Generator	**Critic**
input = Noise (256,)	input = Cuts (3,64,64)
Linear (256 → 4096)	**Conv2D** (3 → 8), kernel (3,3)
Reshape 4096 → (1, 16, 16, 16)	**ReLU**
Conv3D (1 → 128), kernel (3,3,3)	**Conv2D** (8 → 16), kernel (3,3)
InstanceNorm, **ReLU**	**ReLU**, **MaxPooling** (2,2)
Upscale x2	**Conv2D** (16 → 32), kernel (3,3)
Conv3D (128 → 64), kernel (3,3,3)	**ReLU**, **MaxPooling** (2,2)
InstanceNorm, **ReLU**	**Conv2D** (32 → 64), kernel (3,3)
Upscale x2	**ReLU**, **MaxPooling** (2,2)
Conv3D (64 → 32), kernel (3,3,3)	**Conv2D** (3 → 8), kernel (3,3)
InstanceNorm, **ReLU**	**ReLU**, **GlobalMaxPooling**
Conv3D (32 → 1), kernel (3,3,3)	**Linear** (64 → 1)
Sigmoid	
Output = (64,64,64)	Output = (1,)

2.4. Quantitative Analysis of the Results

To assess the quality of the results, and compare the simulation with the reference when it is possible, we compute the indicator variograms and connectivity functions, as well as the Frechet Inception Distance (FID) between generated and real examples.

The indicator variogram $\gamma(h)$ is defined as follows.

$$\gamma(h) = \frac{1}{2}\mathbb{E}\left[\left(I(x) - I(x+h)\right)^2\right], \tag{2}$$

with $I(x)$ being the indicator function of the facies 1. Since the cases that we investigate in this paper are binary, the indicator variogram is identical for facies 1 and 0.

The connectivity functions $\tau_i(h)$ describes the probability that two pixels located at a distance h belong to the same connected component, knowing that the first pixel is within the facies i.

$$\tau_i(h) = \mathbb{P}\left[x \Longleftrightarrow x + h | I(x) = i\right] \tag{3}$$

$\tau_i(h)$ is computed for facies 0 and 1 separately. The connectivity functions are known to be different when computed on 3D or 2D fields. Indeed, the probability of having a connection between two points is generally higher in 3D for the same type of statistical configurations because the number of possible paths between the two points is larger in 3D. In the following of the paper, we compute the connectivity functions in 3D when we have a 3D reference. When the reference contained only 2D sections, we compare the connectivity functions computed only on 2D sections.

For the indicator variogram and the connectivity functions, we will compare the curves computed on individual realizations belonging either to the training set and to the set of simulations. The simulations are stochastic and therefore there will be some variability between these curves. To ease the comparison, we plot the mean values (as a function of distance) for the training sets and simulated sets. Because the number of simulations is limited (as well as the number of training examples), these mean curves are subject to estimation errors (standard error on the mean). Similarly to visualize the range of variations of the curves, we use the standard deviation of the curves computed on the reference set and superpose individual curves for the simulations. The standard deviation estimated on the reference is subject to an approximation error as well. These errors are not displayed on the graphs for the sake of visibility, but they have been studied for the first three cases.

The Frechet Inception Distance (FID), introduced by Heusel et al. (2017), is a heuristic measure of the distance between the distributions of generated sample and training images. Training image samples x and generated samples $\hat{x}$ are fed into an InceptionV3 neural network (Salimans et al., 2016) trained for a classification task on the ImageNet dataset. This network outputs feature vectors y and $\hat{y}$. Denoting by (m, C) [resp. $(\hat{m}, \hat{C})$] the mean vector and covariance matrix over the samples y (resp. $\hat{y}$), the FID score is then defined as:

$$FID = ||m - \hat{m}||_2^2 + Tr(C + \hat{C} - 2(C\hat{C})^{1/2}) \tag{4}$$

In DiAGAN, the FID gives an insight on the convergence and the relative quality of generated samples.

The variograms and connectivity functions, $\gamma(h)$, $\tau(h)$ and the FID are computed and plotted along the three main directions X, Y, and Z.

3. RESULTS

DiAGAN is implemented in both Pytorch and Tensorflow, two of the most popular Deep Learning librairies in Python. Both implementation lead to similar results.

3.1. Data Sets

To illustrate the proposed methodology, we consider six examples. The training images are shown in **Figure 2**. In all those examples, we consider only binary cases even if the method can be applied to discrete problems with more lithologies or even continuous problems. Outputs of our GANs, that are intrinsically continuous are thresholded to obtain binary data. Voxels with positive values are set to facies 1, while voxels 0 stay the same. Note that for most of the data sets the grid used for the simulation domain is smaller than the size of the training data sets. Therefore the size of the objects may look different when comparing a simulation with the training data visually.

In general, a GAN requires a large training data set. However, in most applications in geosciences only a few training images are available or can be drawn manually by a geologist. Insufficient variability in the training data set prevents a GAN to train correctly. To tackle this problem, we use here large input images that were sub-sampled to generate a large number of smaller images. This fits our context well, since the challenge we seek to resolve is to synthesize plausible volume data from 2D inputs.

The datasets used in this paper are of two types.

On the one hand, three-dimensionnal data sets are represented in **Figures 2A–D**. They are used to test the method and allow to make a visual and quantitative comparison between the known 3D structure and the simulations obtained by our procedure using only 2D cuts through these volumes as training data.

- **Figure 2A** is a procedurally generated set of packed spheres. The grid has a size of 300 × 300 × 300 voxels. The spheres have diameters taken from a uniform distribution between 8 and 12 pixels and do not intersect each other. The global proportion of voxels occupied by the spheres is around 20%. This synthetic data set constitutes a benchmark even if it is far from a real-life geological application.
- **Figure 2B** is taken from the literature (Mariethoz and Kelly, 2011), it shows a stack of folded geological layers that were generated using an MPS method based on invariant distances and a rotation field[1]. The grid has a size of 126 × 126 × 120 voxels.
- **Figure 2C** is a rendering of a CT scan of the sandpack F42A obtained from Imperial College, London[2]. The grid has a size of 300 × 300 × 300 voxels.
- Finally, **Figure 2D** represents a geological reservoir at a kilometer scale that contains a set of fluvial channels. The image has been generated for this paper using the Tetris object based algorithm implemented in Ar2GEMS (Boucher et al., 2010). The grid has a size of 126 × 126 × 64 voxels.

On the other hand, training images depicted in **Figures 2E,F** are purely two dimensional. Assessing the quality of the output for these examples is more difficult because the 3D ground truth is not available and may not even exist. However, these cases are important to demonstrates the generalization capabilities of DiAGAN.

- The training image displayed in **Figure 2E1** was taken from Laloy et al. (2018). It has a size of 1,000 × 1,000 pixels. This image was inspired by the training image of 2D channels

[1] http://trainingimages.org/training-images-library.html

[2] https://www.imperial.ac.uk/earth-science/research/research-groups/perm/research/pore-scale-modelling/micro-ct-images-and-networks/sand-pack-f42a/

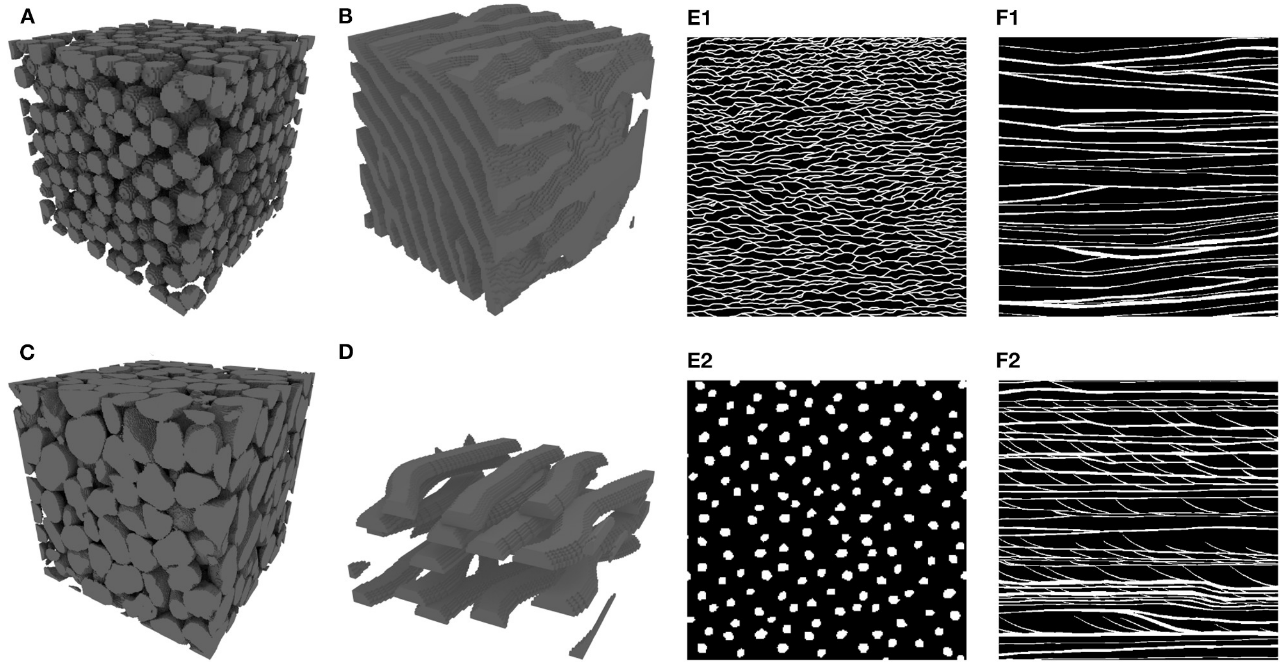


FIGURE 2 | The six training data sets used in this study. Training images **(A–D)** are 3D examples fed as cuts, while TI **(E)** and **(F)** are cuts with no 3D ground truth. Details about the dimensions of the data sets and the data sources are given in the text. In all these data sets, the facies 0 is represented either in black or transparent, the facies 1 is represented either in gray for the 3D blocks and in white for the 2D images.

created and used by Strebelle (2002) and extended by Laloy et al. (2018) to improve the diversity of the data set during the training phase. Since the original image has been heavily used as a benchmark in the MPS literature, we will refer to it as the Strebellian channels. It represents a map view of channel and matrix oriented along the (X, Y) and (X, Z) planes. The image displayed in **Figure 2E2** was created manually for this paper and used to represent a vertical cross section along the (Y, Z) plane displaying roughly circular objects sections through the channels. It has a size of 600 × 746 pixels. With these two training images as input data, the aim is to simulate 3D channels or conduits propagating along the X direction. We took great care to ensure matching scales along the two images.

- Finally, **Figures 2F1,F2** correspond to two perpendicular vertical sections (of about 30 × 30m) that have been mapped by Huysmans and Dassargues (2011) through the Brussels sand deposit. Both images have a size of 600 × 600 pixels. These data were used previously in Comunian et al. (2012). Image (f1) is a cut perpendicular to the X axis and (f2) is perpendicular to the Y axis. The horizontal cut is not available in that case. Notice how the two images differ: while layers perpendicular to the X axis are roughly horizontal and loosely connected, the one perpendicular to the Y axis presents some cross-bedding.

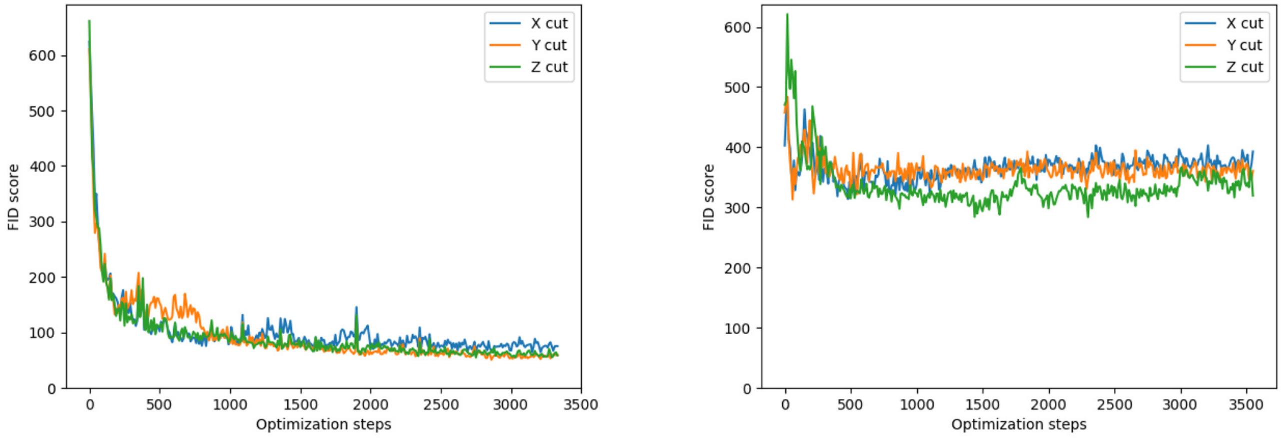
3.2. Realizations From 3D Datasets

DiAGAN has been first applied to generate 3D volumes of 64 × 64 × 64 voxels from sections taken inside the 3D examples. In real practical cases, these full 3D data sets would not be available. The aim here is to test the methodology in a situation where the 3D structure is known and can be used as a reference. While only 2D data taken from the 3D structure are used for training, the analysis of the resulting 3D simulation is compared to the original 3D data sets allowing to check if the simulation was correct. In particular, variograms and connectivity functions are computed in 3D.

The training times for these cases take from a few hours to a whole day on a rather old Tesla K40C GPU with 12 Gb of RAM. Once done, the generator is able to produce a simulation in less than a second.

Figure 3 a shows the evolution of the FID score during the training phase for the simulation of the 3D random packing of spheres. The FID score is dropping rapidly in the initial phase of training. This evolution is demonstrating the convergence of the method.

Figures 4, **5** show that the simulations of the spheres and sand grains display the correct order of magnitudes for the size of the objects as compared to the 3D references. The variograms and connectivity functions are reasonably well-reproduced. The trends are correct even if some minor differences are visible for example between the mean curves for the variograms. The computation of the standard error on the mean values for these curves show that the difference is not due to statistical variations but to slight differences between the set of simulations generated with DiAGAN and the set of training data resulting from some model errors. The same remark holds for the connectivity curves.



Balls dataset (a)

Strebelle’s channels dataset (e1 & e2)

FIGURE 3 | Evolution of the Frechet Inception Distance during training for two datasets. The FID score in approximated using 100 samples for both the TI and the generated images. A random cut along axes X, Y, and Z was taken from each samples to be fed inside the InceptionV3 model.

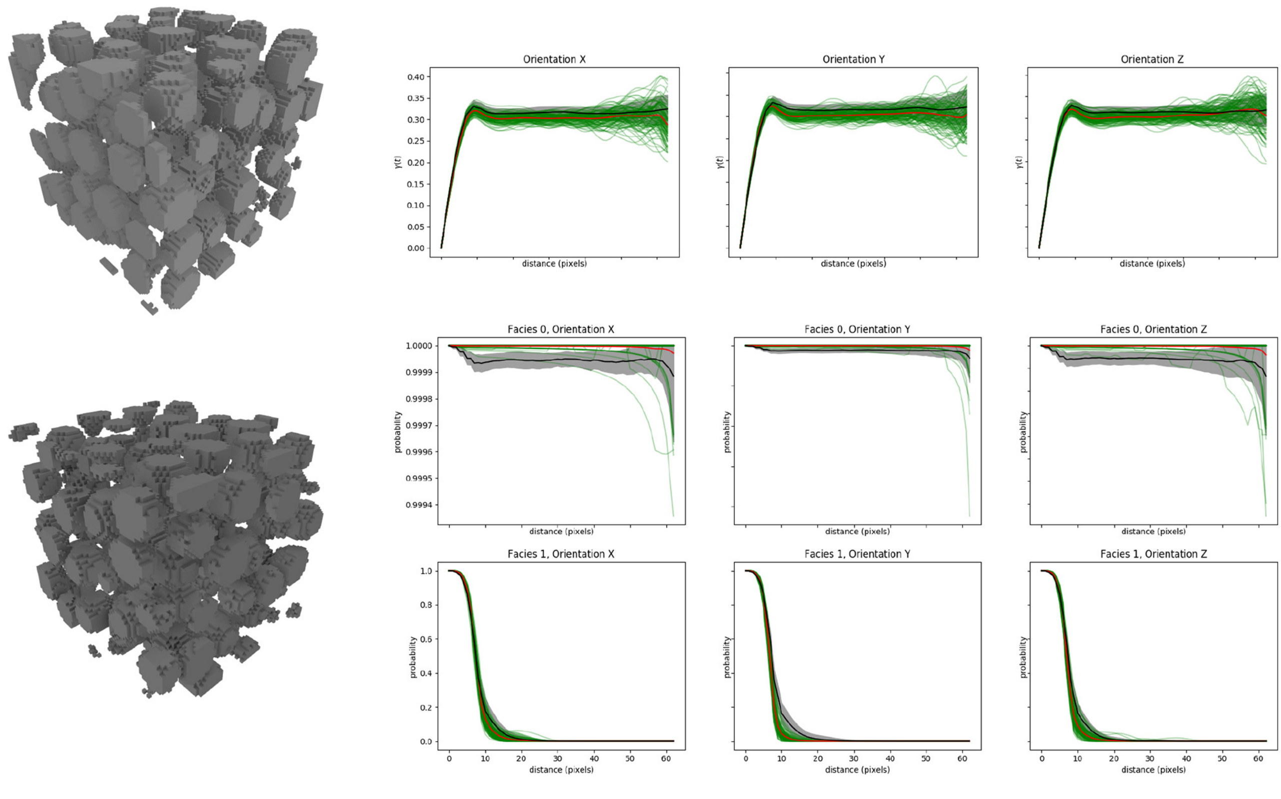


FIGURE 4 | Results of DiAGAN on the ball dataset (a). Note for the visual comparison that the simulated domain has a size of 64x64x64 voxels while the training data set covers a larger area. The three upper curves present the variogram of 100 DiAGAN realizations (green) and their mean (red) along the three axis. The black curve is the mean of observations in the TI for samples of the same size, while the gray area is this mean plus or minus the observed standard deviation. Middle and bottom curves present the connectivity curve of the two facies of the image along the three axes, with the same color conventions.

Figure 5 and the statistical analysis of the standard error for the mean and for the standard deviation shows that the f42a case is well-simulated with DiAGAN. The mean curves for the variograms and the connectivity functions and the variability around the mean are well-reproduced for that case (differences within standard errors).

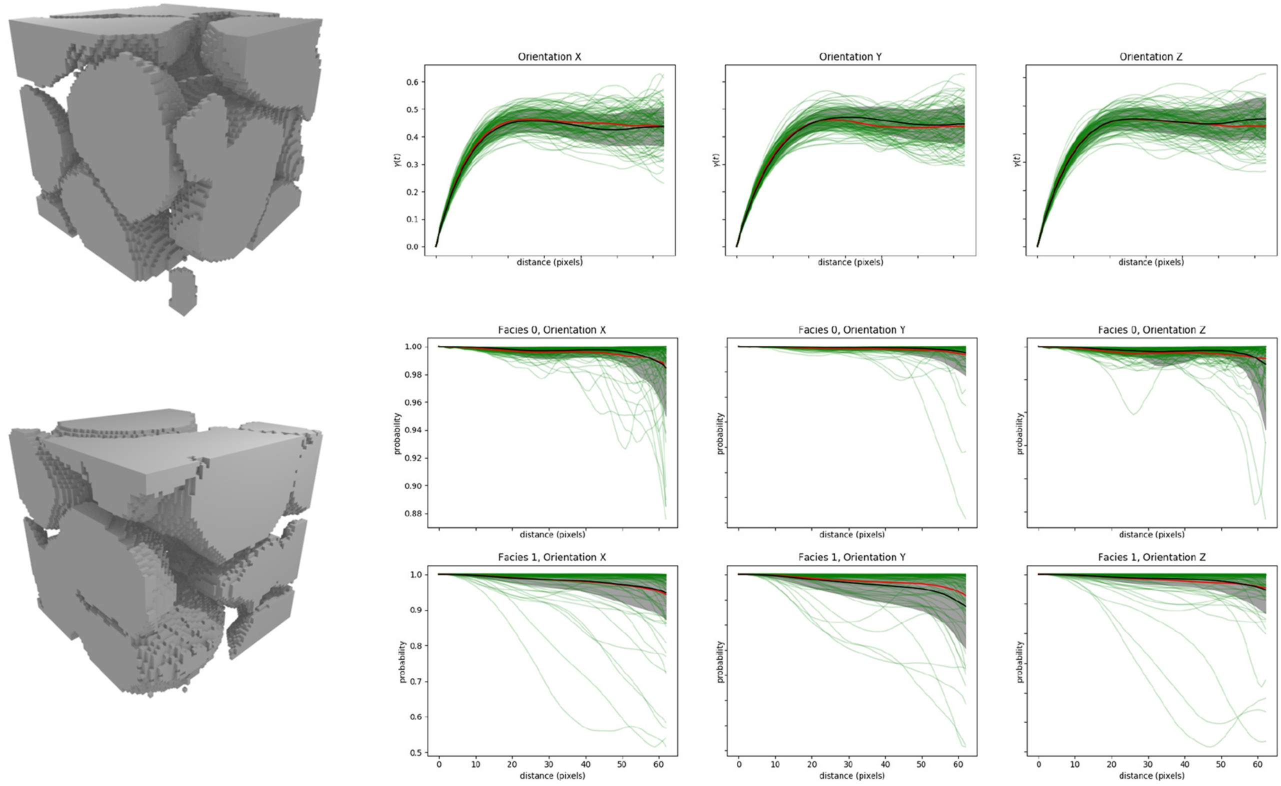


FIGURE 5 | Results of DiAGAN on the sand grain dataset (c). Note for the visual comparison that the simulated domain has a size of 64 × 64 × 64 voxels while the training data set covers a larger area of 300 × 300 × 300 voxels. The three upper curves present the variogram of 100 DiAGAN realizations (green) and their mean (red) along the three axis. The black curve is the mean of observations in the TI for samples of the same size, while the gray area is this mean plus or minus the observed standard deviation. Middle and bottom curves present the connectivity curve of the two facies of the image along the three axes, with the same color conventions.

The visual inspection of an ensemble of cross sections (**Figure 6**) through the simulations allows to compare the training image and the simulations with DiAGAN. We observe that the size of the balls or sand grains are similar. The variability of the shapes and the position of the objects appear to be well-reproduced as well. A further comparison with results from a standard 3D to 3D GAN showed that for these cases that are relatively simple and isotropic, there were no significant differences in the quality of the results.

The two other 3D examples that are considered here are more difficult because they consider large objects traversing the whole domain. The data sets contain less repetitions and identifying the underlying statistical distribution from these images is therefore more difficult than for the two previous cases. In addition, for the folded layers dataset (**Figure 2B**) the orientation of the layers varies inside the training image and therefore there is a non-stationarity in the training data. The results from DiAGAN for this example (**Figure 7**) show less variability in the orientations or in the variograms, but a correct visual reconstruction of the layers. Continuous fold layers that were solid in the TI present holes or imperfections in the generated examples. We therefore observe a reduction of the sill of the variogram and a connectivity that remains high instead of fluctuating for facies 0, while often dropping faster than in the TI for facies 1. The differences between the statistical estimates of the mean and standard deviations are obviously important indicating that the DiAGAN model in this case does not capture all the structure of the examples.

Finally, for the channelized reservoir case (**Figure 2D**), DiAGAN simulated the channels from the 2D sections pretty well (**Figure 8**), although the output is noisier than the reference. The structure and form of the TI is reproduced with satisfactory accuracy as well as the directional variograms and connectivity functions.

To conclude that part, these 4 examples show that DiAGAN can generate 3D realizations that are close to the 3D references from 2D examples. There are some differences but one has to remember that the simulation of 3D structures from only 2D cross-sections is a problem that is more difficult than the generation of 3D simulations from 3D examples because only a part of the information is provided to the algorithm. It is therefore not surprising that there is a quality loss. If 3D training data is available, a more traditional 3D to 3D GAN should be used to obtain the best quality.

3.3. Realizations From 2D Datasets

We are now considering some examples corresponding to the real potential application of DiAGAN. Only 2D sections are

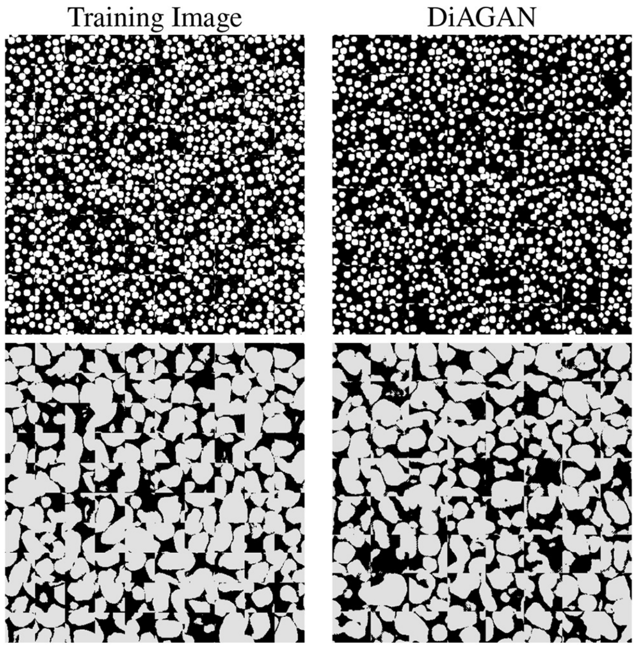


FIGURE 6 | Visual comparison of cuts taken from the TI and DiAGAN generated samples, for the balls dataset (a) and the sand grains dataset (c). Cuts are tiled together for visualization.

available and the 3D ground truth is absent. The evaluation of DiAGAN's output in these cases is more challenging. As written above, this situation is more difficult than the case in which a set of 3D examples are provided. The algorithm has to compensate the lack of information about the 3D geometries by assuming (implicitly in the case of GAN) some type of regularities or symmetries. This problem was discussed previously in Comunian et al. (2012) for example. It is therefore expected that 3D simulations based only on 2D examples should be of lower quality than 3D simulations based on 3D examples when the geometries of the geological objects are complex. For a rather simple case presenting a high degree of symmetry (like the balls presented in the previous section), the loss of information is moderate and therefore it is easier to reconstruct the 3D objects.

If we consider first the case based on the 2D Strebellian channels taken as a training image in the (X, Y) and (X, Z) planes (**Figure 2E1**), DiAGAN correctly simulates a three dimensional network of conduits having a circular section in the (Y, Z) plane. The vertical and horizontal connections between the conduits that are visible in **Figure 9** are due to the fact that we use the same training image along the (X, Y) and (X, Z) planes. The sinuosity of the channels in the horizontal plane must also be reproduced in the vertical plane and DiAGAN finds a reasonable solution by creating the network of conduits.

In **Figure 10**, we present a set of cuts taken from generated samples. The cuts along the (X, Y) and (X, Z) planes resembles the channels from the TI, but they may be broken when a channel moves out of the cut to ensure the 3D continuity that we see in **Figure 9**. The perpendicular cuts along the (Y, Z) plane are much more isotropic but they depart from the target training image made of disks (**Figure 2E2**). We think that this discrepancy is due to the fact that the cuts in different directions are not perfectly compatible. Those images have been drawn separately, and despite the effort that we made to respect similar sizes for the object in the common direction, nothing ensures that a 3D geometry with these sections can really exist. DiAGAN is however capable of obtaining a compromise and a reasonable solution in this situation. Finally, we note that the convergence of the FID criteria for this problem is slower and more difficult to reach (see **Figure 3**) because of the issue described above, i.e., the incompatibility between the cuts along the X and Y axis.

For the Brussels sands deposit, the situation is easier since the training images have been acquired from an existing 3D structure. In this case, DiAGAN generates some roughly horizontal layers with cross beddings but only oriented along the Y direction as it was indicated in the training data set (**Figure 11**). The simulations are slightly noisy but the cuts sampled in the 3D simulations (**Figure 12**) show very well that the cross bedding occurs only in the Y axis plane. The quality of the results is pretty similar to what was obtained earlier with MPS techniques (see Figure 18 in Comunian et al., 2012).

4. DISCUSSION AND CONCLUSION

Generative adversarial networks represent a new and really different method to generate random fields having a predefined spatial structure (prior distribution). This has already been shown and experimented by several authors (Laloy et al., 2017, 2018; Mosser et al., 2017a).

The main novel idea presented in this paper is to introduce a cut sampler in the GAN process between the generator and the discriminator. Our numerical experiments show that this very simple idea makes it possible to reconstruct 3D parameter fields from a series of 2D examples. This is the main contribution of the paper since this was impossible with the methods cited above. We have tested the idea on a series of simple situations and the results are of comparable quality with those obtained with an MPS method previously published (Comunian et al., 2012). Our experiments demonstrate the feasibility of the approach. It is also to notice that although all experiments generated 64x64x64 images, which is a good trade-off between complexity and memory usage, an algorithm like DiAGAN is able to produce images of any size without retraining, by simply providing a latent noise vector of greater or smaller dimension. This is the main interest of fully convolutional architectures for both the generator and the discriminator (see **Table 1**). Note that if 3D examples are available, the traditional GANs are expected to generate better simulations because they will account for the complete 3D information. The idea is not to replace existing GAN implementations with DiAGAN. The quality of the simulations obtained for example by Laloy et al. (2018) or Zhang et al. (2019) are excellent. The idea here is to show how these techniques may be slightly modified to generate 3D realizations when only 2D examples are available as it often occurs in practice.

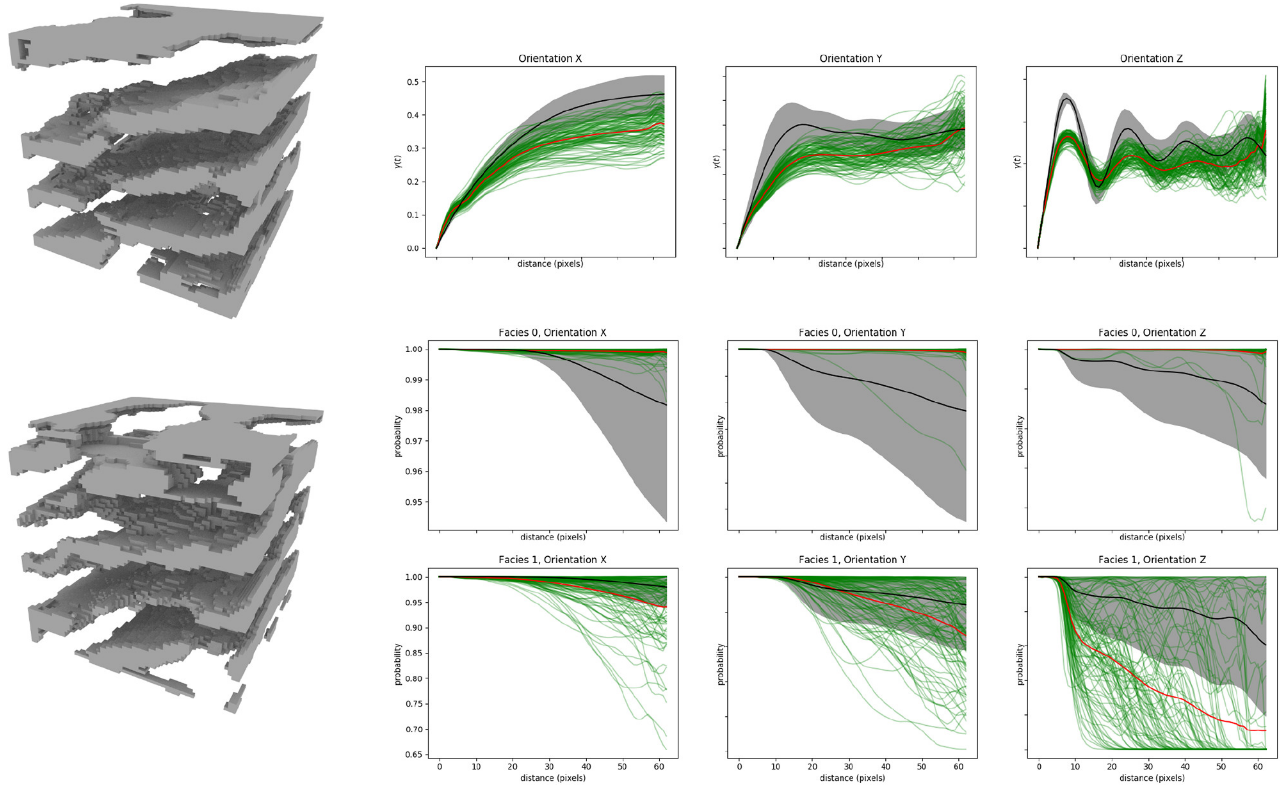


FIGURE 7 | Results of DiAGAN on the categorical fold dataset (b). The three upper curves present the variogram of 100 DiAGAN realizations (green) and their mean (red) along the three axis. The black curve is the mean of observations in the TI for samples of the same size, while the gray area is this mean plus or minus the observed standard deviation. Middle and bottom curves present the connectivity curve of the two facies of the image along the three axes, with the same color conventions.

The DiAGAN method could be further improved or extended. One question that was not explored in this work is the effect of using one single random cut per direction or more cuts. The argument to use one cut only per direction was to keep the algorithm as efficient as possible. We have seen in the numerical experiments that the random cut allowed to obtain results of rather good quality. More cuts may improve the quality but would imply more computations and would slow down the method. Further research could investigate if this is worth doing or not.

Another point that could be improved is the fact that the input of the critic is stacked. It forces the cuts to have the same size, and thus to have cubic realizations. In order to have realizations of any shape, it would be pretty straightforward to have different critics for the different orientations. This could also be relevant from a quality point of view, since these parts will be independently trained to identify different patterns. It is therefore possible that this approach would improve the quality of non-symmetrical examples (for example the channels).

On top of this, while DiAGAN demonstrated satisfactory results on various architectures, it is very likely that the algorithm could benefit from recent and future state of the art techniques in Deep Learning, like more efficient neural network architectures. This could improve both the quality of the outputs and the training time.

Finally, at the moment DiAGAN is not conditional. For geoscience applications, this is a requirement. Some methods have already been developed to condition the GANs to hard data (Zhang et al., 2019). In the future similar techniques should also be implemented in DiAGAN to make it applicable for real applications. What is less clear at the moment is how non-stationarities in patterns may be controlled. In traditional MPS, we can force some trends and describe rather precisely how the probabilities of finding different facies may vary in space as a function of some geological knowledge. This has still to be investigated for the GANs.

As compared to the MPS approach, the main advantage of DiAGAN is the possibility to use the latent input space of Gaussian vectors. Indeed, generating a sample using DiAGAN consists in feeding the generator with a latent vector and applying all the layers. This is very efficiently done on modern computers and GPUs. One can expect a speed-up of several order of magnitude compared to more traditional multiple-point statistics. The fact that the latent space is continuous, differentiable, and that it is fast to generate realizations makes this

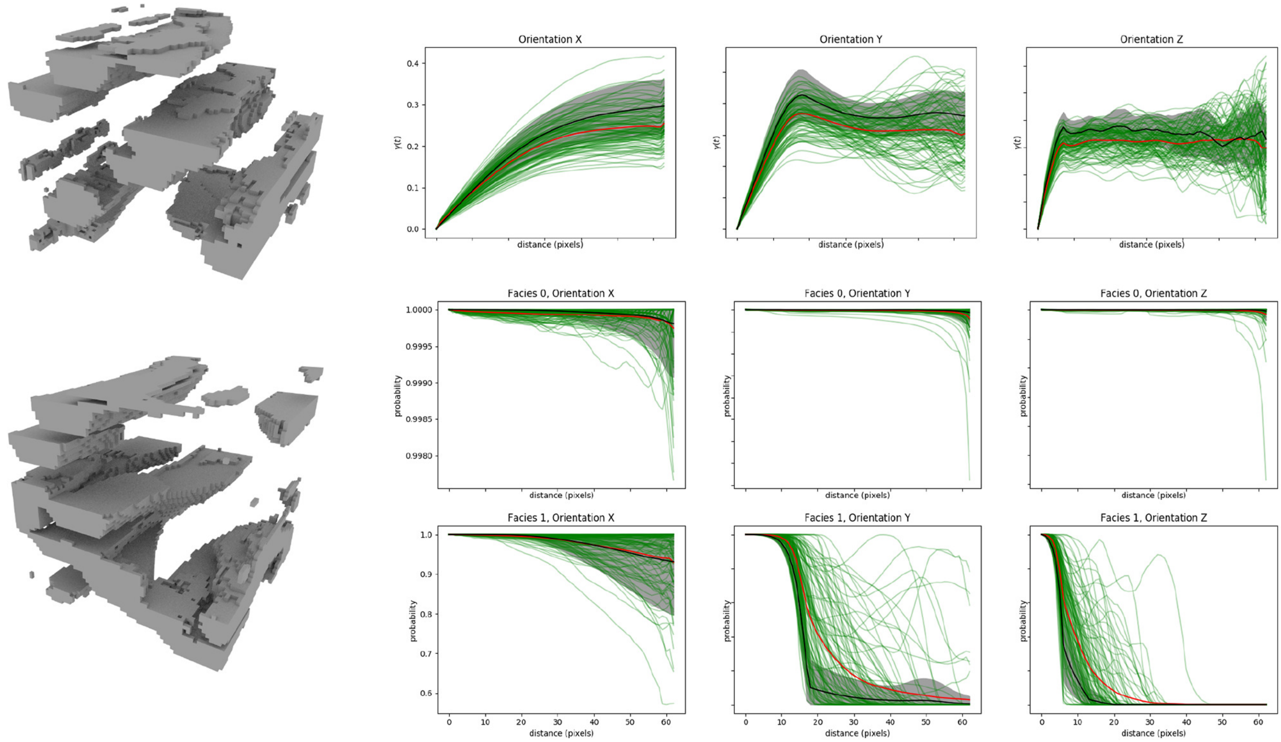


FIGURE 8 | Results of DiAGAN on the procedurally generated channel dataset (d). The three upper curves present the variogram of 100 DiAGAN realizations (green) and their mean (red) along the three axis. The black curve is the mean of observations in the TI for samples of the same size, while the gray area is this mean plus or minus the observed standard deviation. Middle and bottom curves present the connectivity curve of the two facies of the image along the three axes, with the same color conventions.

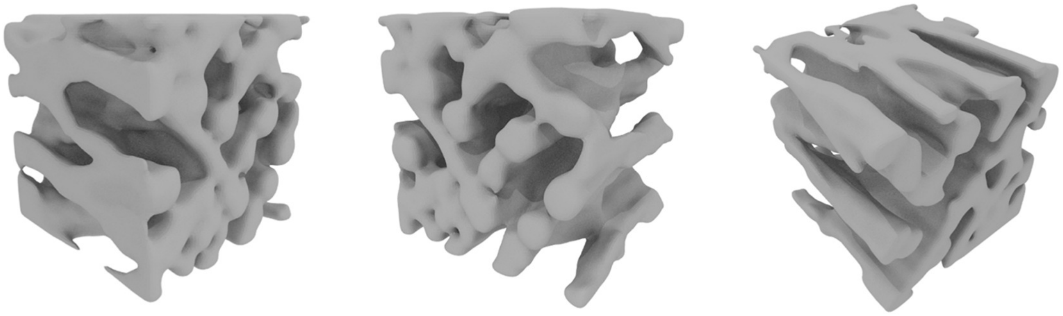


FIGURE 9 | Results of DiAGAN for the 2D channel dataset (Training Images E1 and E2 in **Figure 2**). Images were obtained by applying a density filter on the original voxel data.

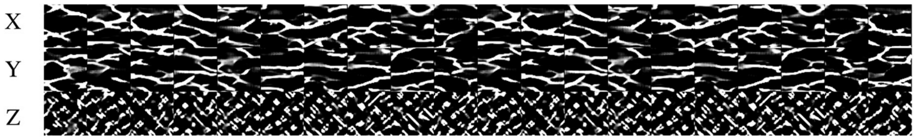


FIGURE 10 | Sample of cuts taken from DiAGAN for the 2D channel dataset (Training Images E1 and E2 in **Figure 2**).

approach potentially very efficient for inverse problem solving. This path has been explored recently for example by Mosser et al. (2019), Laloy et al. (2019) or Liu et al. (2019). However, one remaining issue is that the inverse problem will involve the computation of a forward model using the fields generated with the GANs, and if these fields are discrete (like those studied

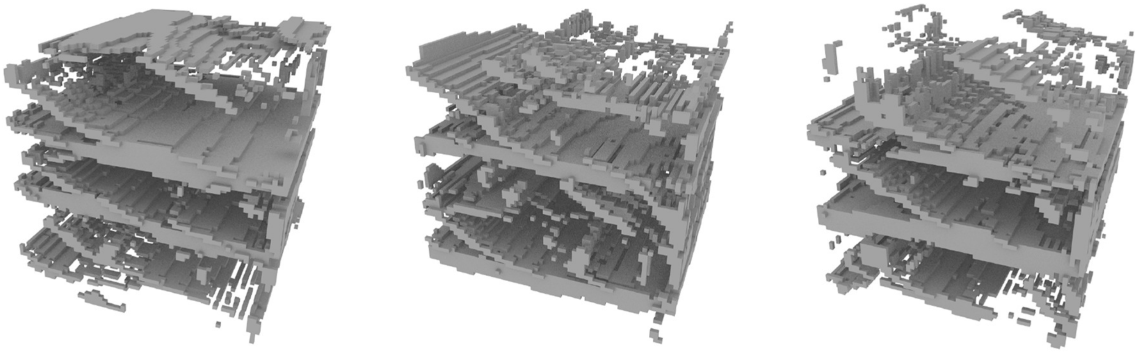


FIGURE 11 | Results of DiAGAN on the Brussel's sand deposit dataset (Training Images F1 and F2 in **Figure 2**).

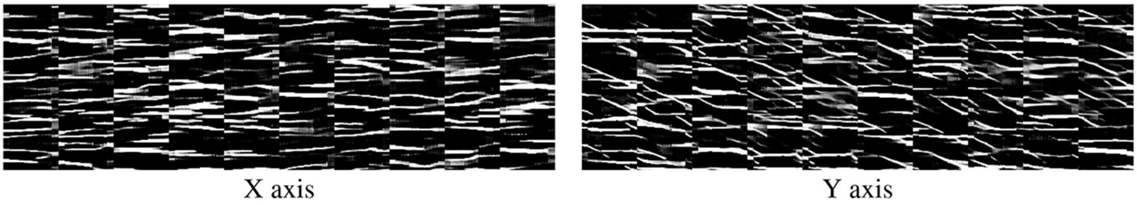


FIGURE 12 | Sample of cuts taken from DiAGAN for Brussel's sand deposit TI (Training Images F1 and F2 in **Figure 2**).

in this paper), it is possible that the response of the forward model may become discontinuous and not differentiable, posing an issue in the inverse problem formulation. This has still to be explored to better identify the domains of application of these techniques.

Back to the computing time aspects, one has also to remember that training time is still long and can last up to several days of computation. This means that for the moment, if a reasonable number of simulations are needed (several hundreds for example), MPS is still faster. Of course, it depends on the dimension of the problem, the complexity of the patterns to simulate, and the size of the training data set.

As of today, Generative Adversarial Networks represent a very interesting alternative to classical geostatistics clearly worth exploring. Their strength are different from the current state of the art methods, which make them a good complementary method.

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/randlab/diaGAN.

AUTHOR CONTRIBUTIONS

All the coding and numerical experiments were conducted by GC, the idea of DiAGAN emerged from discussions between GC, PR, and SL, and the project was supervised by PR. The writing and editing of the paper was done by GC, PR, and SL.

FUNDING

PR was partly funded by the SNF project Phenix (200020_182600/1) during that work.

ACKNOWLEDGMENTS

The authors of the paper would like to thank Przemyslaw Juda for his support and providing the codes for the quality assessment of the realizations, Dr. Boris Galvan and Prof. Stephen A. Miller for having provided access to the GPU cluster of their team, and Prof. Marijke Huysmans for having shared with us the Brussel sands data set. The DiAGAN code and the data sets are accessible online via the github repository: https://github.com/randlab/diaGAN. Note that two implementations are proposed, one using pytorch and one using tensorflow. They provide similar results while they are implemented slightly differently. This illustrates that the method is rather robust.

REFERENCES

Adler, P., Jacquin, C. G., and Quiblier, J. (1990). Flow in simulated porous media. *Int. J. Multiphase Flow* 16, 691–712. doi: 10.1016/0301-9322(90)90025-E

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Barfod, A. A., Straubhaar, J., Høyer, A.-S., Hoffimann, J., Christiansen, A. V., Møler, I., et al. (2018). Hydrostratigraphic modelling using multiple-point statistics and airborne transient electromagnetic methods. *Hydrol. Earth Syst. Sci.* 22, 3351–3373. doi: 10.5194/hess-22-3351-2018

Boucher, A., Gupta, R., Caers, J., and Satija, A. (2010). "Tetris: a training image generator for SGEMs," *Paper Presented at Proceedings of 23nd SCRF Annual Affiliates Meeting* (Palo Alto, CA: Stanford University).

Chan, S., and Elsheikh, A. H. (2017). Parametrization and generation of geological models with generative adversarial networks. *arXiv preprint arXiv:1708.01810*.

Chen, Q., Mariethoz, G., Liu, G., Comunian, A., and Ma, X. (2018). Locality-based 3-d multiple-point statistics reconstruction using 2-d geological cross sections. *Hydrol. Earth Syst. Sci*. 22, 6547–6566. doi: 10.5194/hess-22-6547-2018

Comunian, A., Renard, P., and Straubhaar, J. (2012). 3D multiple-point statistics simulation using 2D training images. *Comput. Geosci*. 40, 49–65. doi: 10.1016/j.cageo.2011.07.009

Cordua, K. S., Hansen, T. M., Gulbrandsen, M. L., Barnes, C., and Mosegaard, K. (2016). Mixed-point geostatistical simulation: a combination of two- and multiple-point geostatistics. *Geophys. Res. Lett*. 43, 9030–9037. doi: 10.1002/2016GL070348

de Marsily, G., Delay, F., Goncalves, J., Renard, P., Teles, V., and Violette, S. (2005). Dealing with spatial heterogeneity. *Hydrogeol. J.* 13, 161–183. doi: 10.1007/s10040-004-0432-3

Feng, J., Teng, Q., He, X., and Wu, X. (2018). Accelerating multi-point statistics reconstruction method for porous media via deep learning. *Acta Mater*. 159, 296–308. doi: 10.1016/j.actamat.2018.08.026

Gadelha, M., Maji, S., and Wang, R. (2017). "3D shape induction from 2D views of multiple objects," in *2017 International Conference on 3D Vision (3DV)* (Qingdao: IEEE), 402–411. doi: 10.1109/3DV.2017.00053

Gadelha, M., Rai, A., Maji, S., and Wang, R. (2019). Inferring 3D shapes from image collections using adversarial networks. *arXiv preprint arXiv:1906.04910*. doi: 10.1007/s11263-020-01335-w

Gerke, K. M., and Karsanina, M. V. (2015). Improving stochastic reconstructions by weighting correlation functions in an objective function. *Europhys. Lett*. 111:56002. doi: 10.1209/0295-5075/111/56002

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Montreal, QC: Neural Information Processing Systems Foundation, Inc.), 2672-2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems,* eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Neural Information Processing Systems Foundation, Inc), 5769-5779.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems,* eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Neural Information Processing Systems Foundation, Inc), 6626–6637.

Hu, L., and Chugunova, T. (2008). Multiple-point geostatistics for modeling subsurface heterogeneity: a comprehensive review. *Water Resour. Res*. 44. doi: 10.1029/2008WR006993

Huysmans, M., and Dassargues, A. (2011). Direct multiple-point geostatistical simulation of edge properties for modeling thin irregularly shaped surfaces. *Math. Geosci*. 43:521. doi: 10.1007/s11004-011-9336-7

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Journel, A., and Zhang, T. (2006). The necessity of a multiple-point prior model. *Math. Geol*. 38, 591–610. doi: 10.1007/s11004-006-9031-2

Karsanina, M. V., and Gerke, K. M. (2018). Hierarchical optimization: fast and robust multiscale stochastic reconstructions with rescaled correlation functions. *Phys. Rev. Lett*. 121:265501. doi: 10.1103/PhysRevLett.121.265501

Kessler, T. C., Comunian, A., Oriani, F., Renard, P., Nilsson, B., Klint, K. E., et al. (2013). Modeling fine-scale geological heterogeneity-examples of sand lenses in tills. *Groundwater* 51, 692–705. doi: 10.1111/j.1745-6584.2012.01015.x

Kingma, D. P., and Ba, J. (2014). ADAM: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resour. Res*. 54, 381–406. doi: 10.1002/2017WR022148

Laloy, E., Hérault, R., Lee, J., Jacques, D., and Linde, N. (2017). Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour*. 110, 387–405. doi: 10.1016/j.advwatres.2017.09.029

Laloy, E., Linde, N., Ruffino, C., Hérault, R., Gasso, G., and Jacques, D. (2019). Gradient-based deterministic inversion of geophysical data with generative adversarial networks: is it feasible? *Comput. Geosci*. 110:104333. doi: 10.1016/j.cageo.2019.104333

Lemmens, L., Rogiers, B., Jacques, D., Huysmans, M., Swennen, R., Urai, J. L., et al. (2019). Nested multiresolution hierarchical simulated annealing algorithm for porous media reconstruction. *Phys. Rev. E* 100:053316. doi: 10.1103/PhysRevE.100.053316

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

Liu, Y., Sun, W., and Durlofsky, L. J. (2019). A deep-learning-based geological parameterization for history matching complex models. *Math. Geosci*. 51, 725–766. doi: 10.1007/s11004-019-09794-9

Mariethoz, G., and Caers, J. (2014). *Multiple-Point Geostatistics: Stochastic Modeling With Training Images*. Chichester: John Wiley & Sons. doi: 10.1002/9781118662953

Mariethoz, G., and Kelly, B. F. (2011). Modeling complex geological structures with elementary training images and transform-invariant distances. *Water Resour. Res*. 47. doi: 10.1029/2011WR010412

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

Meerschman, E., Van Meirvenne, M., Mariethoz, G., Islam, M. M., De Smedt, P., Van De Vijver, E., et al. (2014). Using bivariate multiple-point statistics and proximal soil sensor data to map fossil ice-wedge polygons. *Geoderma* 213, 571–577. doi: 10.1016/j.geoderma.2013.01.016

Mosser, L., Dubrule, O., and Blunt, M. J. (2017a). Reconstruction of three-dimensional porous media using generative adversarial neural networks. *Phys. Rev. E* 96:043309. doi: 10.1103/PhysRevE.96.043309

Mosser, L., Dubrule, O., and Blunt, M. J. (2017b). Stochastic reconstruction of an oolitic limestone by generative adversarial networks. *Trans. Porous Media* 125, 1–23. doi: 10.1007/s11242-018-1039-9

Mosser, L., Dubrule, O., and Blunt, M. J. (2019). Deepflow: history matching in the space of deep generative models. *CoRR abs/1905.05749*.

Okabe, H., and Blunt, M. J. (2007). Pore space reconstruction of vuggy carbonates using microtomography and multiple-point statistics. *Water Resour. Res*. 43. doi: 10.1029/2006WR005680

Oriani, F., Straubhaar, J., Renard, P., and Mariethoz, G. (2014). Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrol. Earth Syst. Sci.* 18, 3015–3031. doi: 10.5194/hess-18-3015-2014

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Barcelona: Neural Information Processing Systems Foundation, Inc), 2234-2242.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol*. 34, 1–21. doi: 10.1023/A:1014009426274

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Vannametee, E., Babel, L., Hendriks, M., Schuur, J., De Jong, S., Bierkens, M., et al. (2014). Semi-automated mapping of landforms using multiple point geostatistics. *Geomorphology* 221, 298–319. doi: 10.1016/j.geomorph.2014.05.032

Wei, L.-Y., Lefebvre, S., Kwatra, V., and Turk, G. (2009). "State of the art in example-based texture synthesis," in *Eurographics 2009, State of the Art Report, EG-STAR* (Munich: Eurographics Association), 93–117.

Yeong, C., and Torquato, S. (1998). Reconstructing random media. *Phys. Rev. E* 57:495. doi: 10.1103/PhysRevE.57.495

Zhang, T.-F., Tilke, P., Dupont, E., Zhu, L.-C., Liang, L., and Bailey, W. (2019). Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks. *Petrol. Sci.* 16, 541–549. doi: 10.1007/s12182-019-0328-4

Zuo, C., Yin, Z., Pan, Z., MacKie, E. J., and Caers, J. (2020). A tree-based direct sampling method for stochastic surface and subsurface hydrological modeling. *Water Resour. Res.* 56:e2019WR026130. doi: 10.1029/2019WR026130

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 in Water

ORIGINAL RESEARCH
published: 19 November 2020
doi: 10.3389/frwa.2020.574917

Automated Cloud Based Long Short-Term Memory Neural Network Based SWE Prediction

Alireza Yekta Meyal[1]*, *Roelof Versteeg*[1], *Erek Alper*[1], *Doug Johnson*[1], *Anastasia Rodzianko*[1], *Maya Franklin*[2] *and Haruko Wainwright*[2]

[1] Subsurface Insights, Hanover, NH, United States, [2] Lawrence Berkeley National Laboratory, Berkeley, CA, United States

Snow derived water is a critical component of the US water supply. Measurements of the Snow Water Equivalent (SWE) and associated predictions of peak SWE and snowmelt onset are essential inputs for water management efforts. This paper aims to develop an integrated framework for real-time data ingestion, estimation, prediction and visualization of SWE based on daily snow datasets. In particular, we develop a data-driven approach for estimating and predicting SWE dynamics using the Long Short-Term Memory neural network (LSTM) method. Our approach uses historical datasets (precipitation, air temperature, SWE, and snow thickness) collected at NRCS Snow Telemetry (SNOTEL) stations to train the LSTM network and current year data to predict SWE behavior. The performance of our prediction was compared for different prediction dates and prediction training datasets. Our results suggest that the proposed LSTM network can be an efficient tool for forecasting the SWE timeseries, as well as Peak SWE and snowmelt timing. Results showed that the window size impacts the model performance (where the Nash Sutcliffe efficiency (NSE) ranged from 0.96 to 0.85 and the Rooted Mean Square Error (RMSE) ranged from 0.038 to 0.07) with an optimum number that should be calibrated for different stations and climate conditions. In addition, by implementing the LSTM prediction capability in a cloud based site-monitoring platform, we automate model-data integration. By making the data accessible through a graphical web interface and an underlying API which exposes both training and prediction capabilities. The associated results can be made easily accessible to a broad range of stakeholders.

Keywords: SWE, LSTM, prediction, real-time web based interface, forecasting, model-data integration, neural network

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Jie Niu,
Jinan University, China
Xi Chen,
University of Cincinnati, United States

****Correspondence:***
Alireza Yekta Meyal
alireza.meyal@subsurfaceinsights.com

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *22 June 2020*
Accepted: *27 October 2020*
Published: *19 November 2020*

Citation:
Meyal AY, Versteeg R, Alper E, Johnson D, Rodzianko A, Franklin M and Wainwright H (2020) Automated Cloud Based Long Short-Term Memory Neural Network Based SWE Prediction. Front. Water 2:574917. doi: 10.3389/frwa.2020.574917

INTRODUCTION

Accurate estimation and prediction of snow water equivalent (SWE) in mountain watersheds has been a longstanding challenge (Bair et al., 2018), while, it is a key metric used by hydrologists and water managers to assess water resources in snow-dominated catchments or basins (Bales et al., 2006; Painter et al., 2016). SWE is defined as the equivalent amount of water if the snow mass is completely melted. SWE is one of the main parameters used in accurate prediction of snowmelt runoff and snowpack and water supply forecasting (Schneider and Molotch, 2016). Consequently there is substantial interest in forecasting seasonal SWE dynamics, including parameters such as peak SWE and snowmelt timing onset (Odei et al., 2009). This snowmelt timing is critical for ecological processes in snow-dominated regions, controlling plant dynamics, net ecosystem exchanges, and soil carbon (Harte et al., 2015; Sloat et al., 2015; Wainwright et al., 2020). Snowmelt timing also drives peak flow timing during which significant nutrient export occurs from the catchments (Carroll et al., 2018). In recognition of its value for water resource prediction SWE and associated measurements (temperature, precipitation, windspeed and direction, and snow thickness) are measured across the west area by the U.S. Natural Resource Conservation Service's (NRCS) through over 800 automated data collection stations known as SNOTEL (SNOw TELemetry) stations, as well as by airborne observations (Painter et al., 2016). Stations are typically located in small clearings in evergreen forests. Data from these stations is transferred multiple times a day to a central database, from where the data is publicly accessible through web interfaces and software APIs. Each SNOTEL station has a long record of historical data, often more than 30 years, encompassing a variety of metrological conditions at each site. This results in typically more than 10,000 data points. In addition, snow accumulation and melting is a highly heterogeneous process affected by a complex terrain or regional scale atmospheric forcing which support us to use deep learning method for SWE forecasting.

There is a long standing interest in the use of probabilistic forecasting and Artificial Neural Networks (ANN) such as recurrent neural networks (RNNs) (Kumar et al., 2004) for hydrology and SWE forecasting (Huang and Cressie, 1996; Winstral et al., 2019; Magnusson et al., 2020). More recently deep learning methods such as the long short-term memory network (LSTM) have demonstrated a significant promise in hydrological time series analysis and forecasting (Xiang et al., 2020), such as soil moisture modeling (Fang et al., 2017), monthly water-table depth predictions (Zhang et al., 2018), and daily or hourly rainfall-runoff modeling (Hu et al., 2018; Kratzert et al., 2018; Le et al., 2019; Fan et al., 2020).

In this paper, we develop the integrated framework of real-time ingestion, estimation/prediction and visualization (webinterface) of the snow dynamics based on SNOTEL generated time-series data. In particular, we demonstrate the feasibility of using LSTM trained on for predicting future SWE dynamics. Although we use a few selected SNOTEL stations, our framework is general, and hence, it can be used for other stations across the US. In addition, the feasibility of automating and exposing this capability through a webinterface and an underlying API is demonstrated. It includes quality control, flagging and interpolation, which is often a bottleneck of applying deep learning to environmental datasets. We believe that this framework makes the predictions and deep learning easily accessible to different interested parties for public use or stakeholder use.

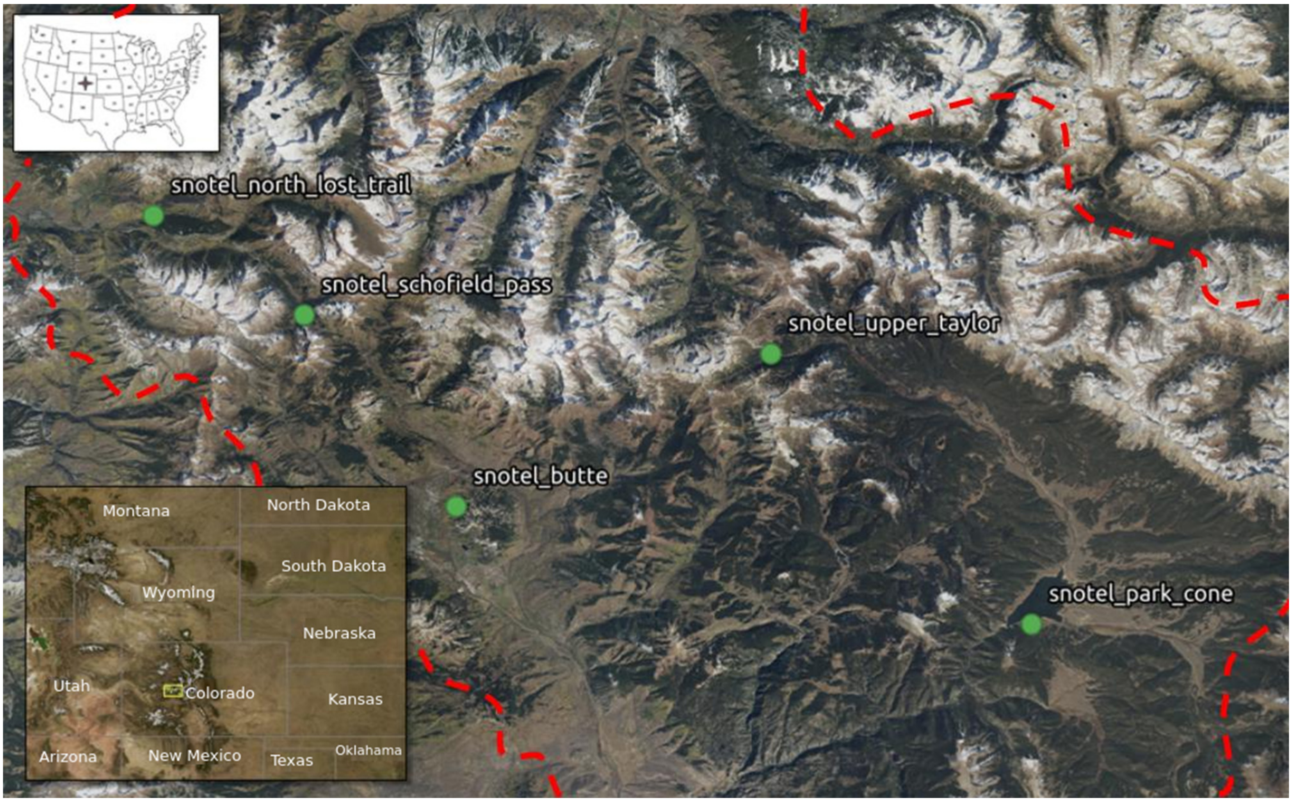


FIGURE 1 | Location and names of the five SNOTEL Stations used in this study (red dash line: East River watershed boundary).

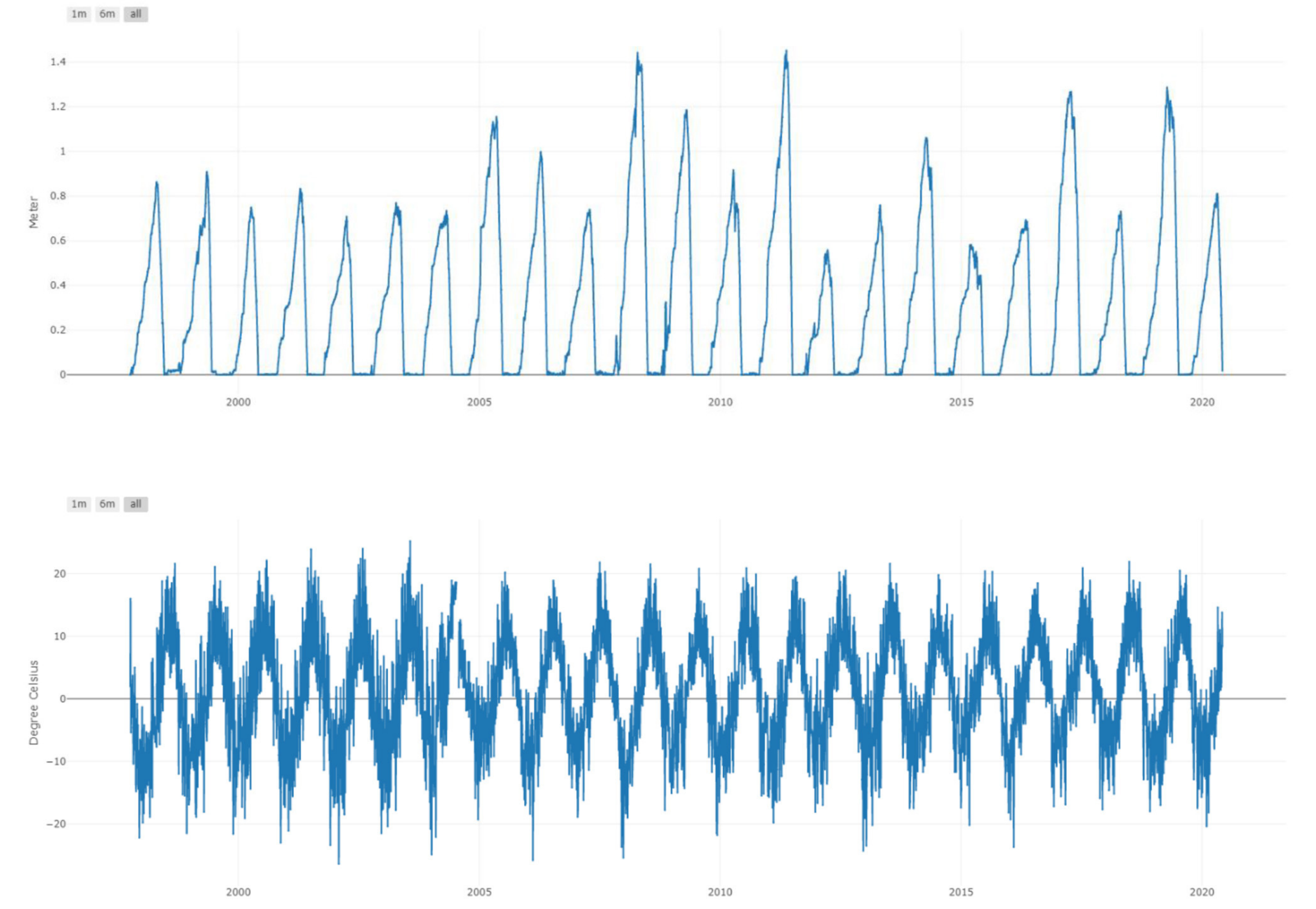


FIGURE 2 | Example of data used in this study. Top: SWE data from the Schofield station. Bottom: Temperature at the Schofield station.

MATERIALS AND METHODS

Study Area

In this paper, we focus on five SNOTEL stations, which are located in Western Colorado within the central Rocky Mountains (**Figure 1**). Our interest in this area is associated with work done by several of the authors on the multiyear, multi-institution Department of Energy funded research effort (the Lawrence Berkeley National Lab (LBNL) Watershed Science Focus Area (SFA) (Hubbard et al., 2018), which focuses on the East River Watershed located near Crested Butte, Colorado. The East River watershed measures ∼300 km^2. It includes montane to alpine ecosystems with an elevation ranging from 2,500 m to 4,000 m. The vegetation in this watershed is diverse, including mixed conifer forest, aspen forest, and open meadows (Harte et al., 2015). Streamflow is dominated by snowmelt in spring and summer (Markstrom et al., 2012). This watershed is a typical headwater catchment in the Colorado River Basin. As the Colorado River Basin provides 75% of the water demand for 40 million people in seven states and two countries (Deems et al., 2013) an understanding of hydro-biogeochemical processes in these headwaters catchments is of obvious value and interest.

Data

Data types collected at SWE stations varies across stations. As far as we can ascertain, all stations collect SWE, snow thickness, precipitation and air temperature. Many stations also collect other environmental parameters such as wind speed and wind direction, air pressure and incoming broadband solar radiation. While in general data quality and continuity is high, there are instances (especially in the data from the early 2000s where SNOTEL data is not continuous, is noisy or has outliers. In this paper we have not included gapfilling or noise elimination strategies as we limited our prediction to use data from five selected stations for the last 10 years which are of good quality (**Figure 2**). However, robust error handling strategies need to be built in for real life applications.

Methods

Project Data Ingestion and Exposure Through Cloud Based API

We have developed robust capabilities to automatically retrieve, normalize (variable names, units, and timestamps), ingest and link heterogeneous data (hydrological, geochemical, geophysical, microbiological, and remote sensing) from numerous public and project specific data sources. These data are stored in project specific relational databases. The datamodel underlying these databases is a substantially modified version of the ODM2 (Observation Data Model version 2) datamodel (Horsburgh et al., 2016).

Data in the database is accessible to both through a web interface (which allows both data visualization in a variety of manners and data download) and a rich API. While the public data hosted in our database can be obtained by users themselves through APIs provided by different organizations, our architecture allows users both to access and locate uniform data across the project site using a single API call as well as provides advanced visualization and analytical capabilities.

An example of the capabilities of this interface is shown in **Figure 3**, which shows the SWE for the Butte SNOTEL Station for different water years (defined by the USGS as the period between October 1st of 1 year and September 30th of the next). Thus, the water year 2020 runs from 10-1-2019 until 9-30-2020.

Long Short-Term Memory Network

The prediction of time series behavior is of interest for a wide range of applications. Numerous statistical and machine learning approaches exist to predict time series behavior (Fawaz et al., 2019). When dealing with long-term dependencies, traditional feed forward Artificial neural networks (ANNs) are limited (Bengio et al., 1994; Fang et al., 2017). However, the Long Short-Term Memory (LSTM) network method is well-suited for long term dependencies (Hochreiter and Schmidhuber, 1997; Fan et al., 2020). LSTM is a deep neural network (DNN) method which has been successfully applied in various fields (Sahoo et al., 2019) for especially for time sequence prediction problems.

LSTM is well-suited to classify, process, and predict time series given time lags of unknown duration. It can be trained on and

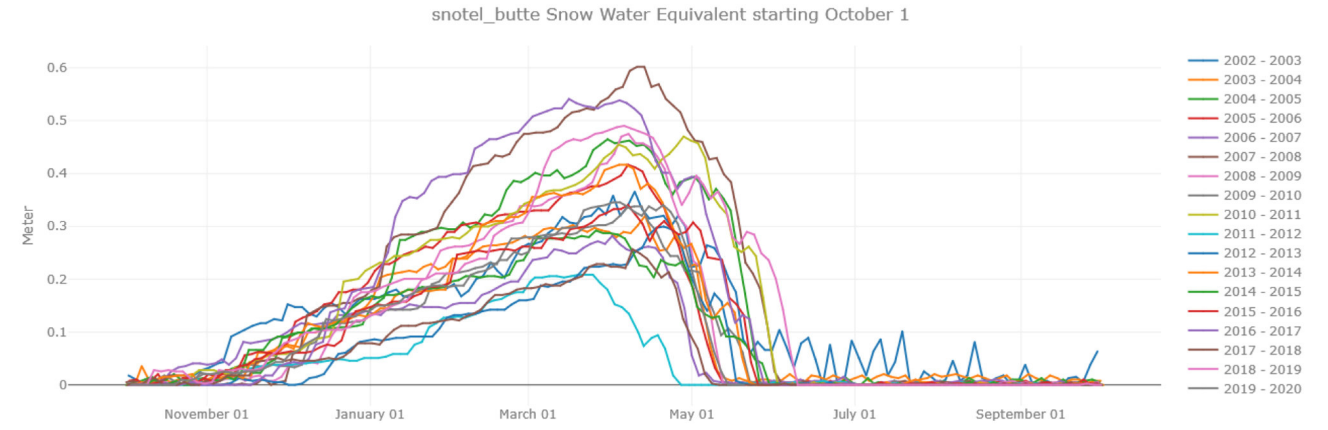


FIGURE 3 | Graph of year over year SWE measured at the Butte SNOTEL station generated by the web interface of our cloud based data management system.

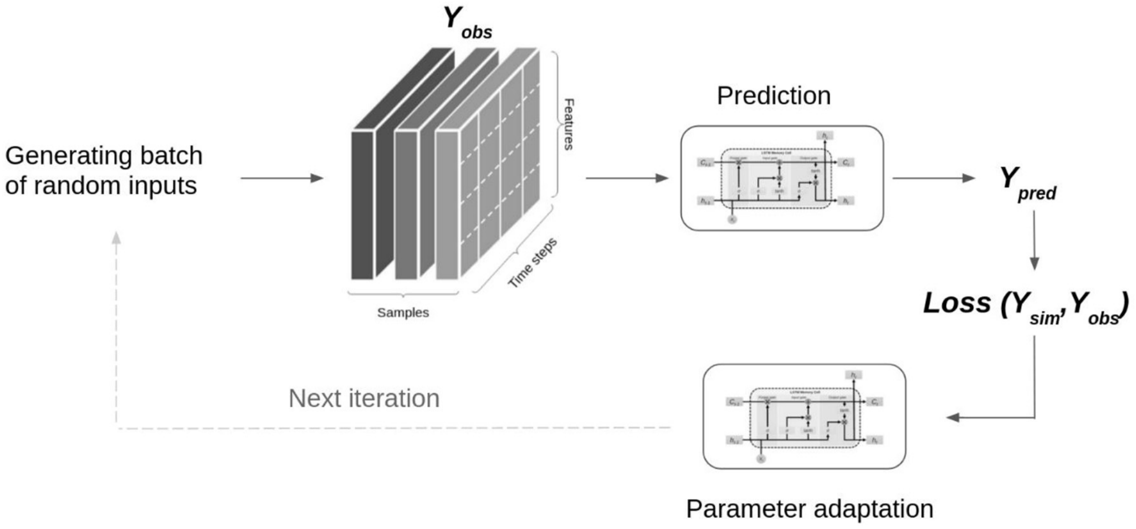


FIGURE 4 | One step iteration of the training process in the LSTM approach.

deal with long sequences and does not rely on a pre-specified window lagged observation as input (Kratzert et al., 2018). In addition, LSTM is well-suited to deal with time series prediction problems with multiple input variables (Le et al., 2019).

In order to use LSTM, we first need to train and calibrate a model. Once this is done the model can be used to predict future values. **Figure 4** shows a schematic of the one-step iteration in the LSTM training/calibration procedure (Fan et al., 2020).

A random batch of input data, consisting of several independent training samples (depicted by the gray colors) that is used in each step. Thus, the input to every LSTM prediction layer is three dimensional, with the three dimensions being samples (or sequences), time steps and features (observations at a timestep). As can be seen in **Figure 4** each training sample consists of several days (timesteps) of look-back data and one target value (Y) to predict. Therefore, the number of samples refer to the number of observations fed into the LSTM network. The number of timesteps or lookback, describes the time window (past data) needed by the LSTM. In each LSTM training iteration step, some of the available training data is used to update some model parameters such as the weights, biases, and learnable network parameters. This update is done in such a way that the loss function is reduced. The loss function is computed from the observed training samples and the network's predictions. In this study, we used the mean-square-error as the loss function for parameter optimization (Kratzert et al., 2018). The gradient descent optimization algorithm is used to reduce the loss function which is equivalent to the unexplained fraction of variance (Xiang et al., 2020).

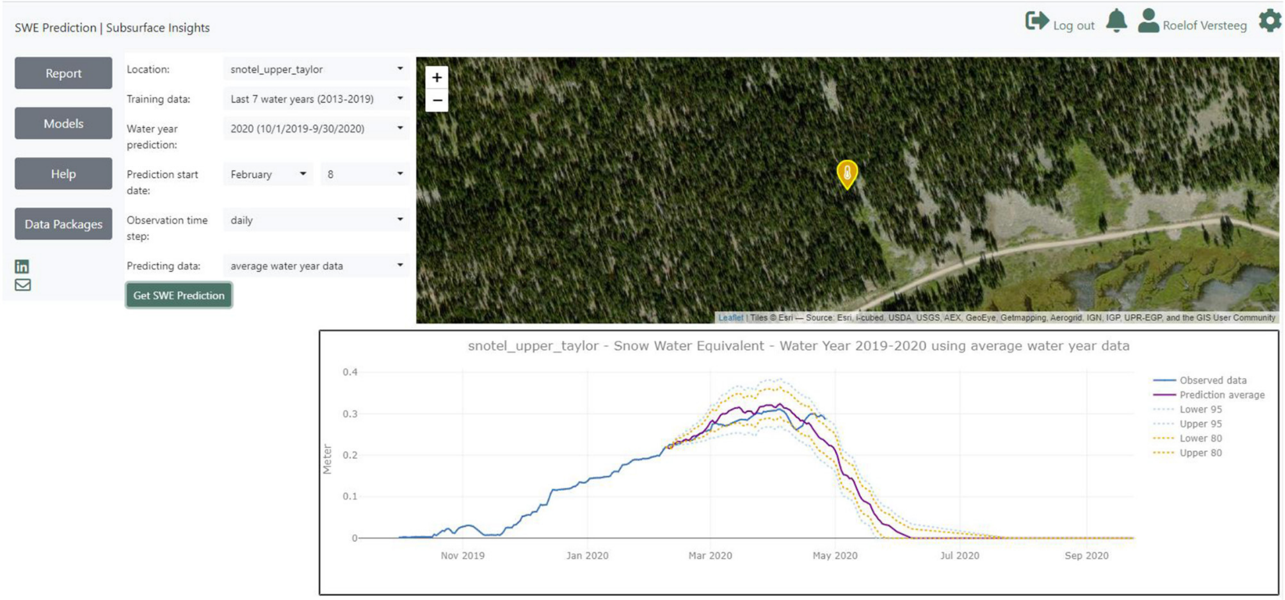
In building a LSTM, after normalizing the raw data, the dataset is first made suitable for a supervised learning problem by splitting into test and training data and by formatting the data in the right input format. Following common practices 60% of data is used for training and 40% is used for validation. After the model is trained and validated, the model can then be used to generates predictions for the future values. The model performance can be evaluated by using testing datasets. Forecasting uncertainties can be represented with confidence intervals. These confidence intervals give us an interval within which we expect the real value to lie with a specified probability that uses standard deviation and mean values of previous observations and current real data. The range of confidence intervals communicates our confidence in the uncertainty associated with the forecast. The confidence intervals are calculated by standard deviation, percentage multiplier and forecast distribution. The percentage multiplier depends on the coverage probability as shown in **Table 1** (Hyndman et al., 2018).

TABLE 1 | Multipliers to be used for confidence intervals.

Percentage	Multiplier
70	1.04
80	1.28
90	1.64
95	1.96

Application of LSTM to SWE Time-Series Analysis and Prediction

In our prediction problem (and in the software implementation), we assume that we have the SWE time-series up to a specific (generally the current) date in a specific (generally the current) water year, and aim to predict the future SWE from this date for


SWE Prediction | Subsurface Insights

Location: snotel_upper_taylor

Training data: Last 7 water years (2013-2019)

Water year prediction: 2020 (10/1/2019-9/30/2020)

Prediction start date: February 8

Observation time step: daily

Predicting data: average water year data

Log out
Roelof Versteeg
Report
Models
Help
Data Packages
Get SWE Prediction
+
−
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community
snotel_upper_taylor - Snow Water Equivalent - Water Year 2019-2020 using average water year data
Meter
0.4
0.3
0.2
0.1
0
Nov 2019
Jan 2020
Mar 2020
May 2020
Jul 2020
Sep 2020
Observed data
Prediction average
Lower 95
Upper 95
Lower 80
Upper 80

FIGURE 5 | Web interface to the prediction API.

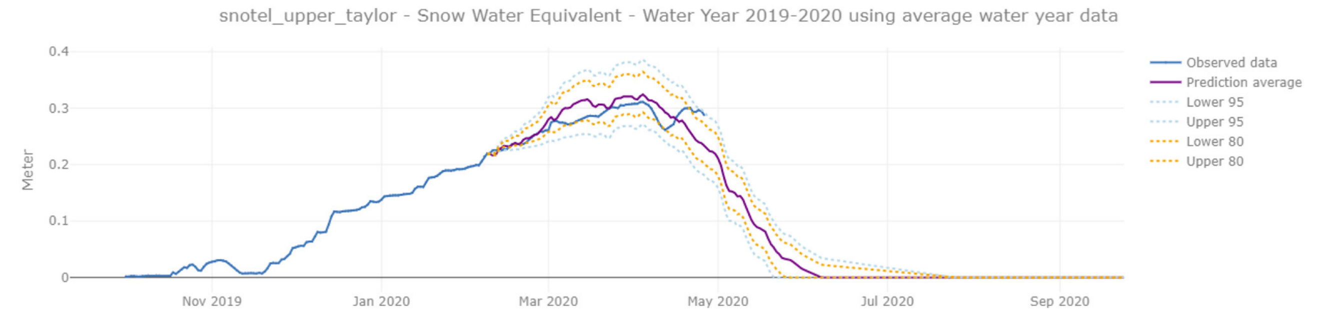


FIGURE 6 | Detail of **Figure 5** showing prediction for a start date of February 7. The prediction was made in Note that our prediction matches actual data (available through the end of April) quite well.

the remainder of the water year based on the historical datasets. As our architecture pulls in new SWE data on a daily basis this prediction is quasi real-time, and in general most interest will be in using our prediction in this mode. However, by allowing the flexibility of providing dates in the past our code allows for performance assessment.

For SWE timeseries forecasting we use the method described above, which is implemented as a python code which uses the Tensorflow (Abadi et al., 2016) and Keras (Chollet et al., 2015) libraries. This code is exposed through an API. The API can be accessed directly programmatically or through a web interface which provides a visual interface to the API (**Figure 5**). Parameters passed to the API include which SNOTEL location to forecast for, how many years of historic data to use for training, what type of snow years (below average, average, or above average) prediction data to use, and for which date we should predict for.

Once our code receives the parameters it first retrieves the raw datasets (which includes SWE, precipitation, snow thickness, and air temperature) needed for prediction through a call to the data API. These are long sequences of thousands of observations for data in previous water years. These sequences are split into samples which are reshaped for the LSTM model. The size of these samples is called the window size (Fan et al., 2020) and has impact on the forecast accuracy.

The reshaped data is used to train the LSTM network. The supervised learning problem is framed as predicting the SWE at a specific day given SWE and associated data (precipitation, snow thickness, and air temperature) up to that day. In our analysis we used training datasets with between 5 and 10 years of recent SNOTEL data, but our code is able to deal with different lengths of data to train the LSTM network.

Once the network is trained, we can use it to make predictions about SWE for a specific water year and date within this year. For this prediction, the model needs the history of SWE over the past months and days in the current water year until the prediction start date. The model then predicts SWE for the remainder of that water year. For the prediction data we allow users to select any of snow years worth of data (e.g. "below average," "average," "above average" years) to accommodate different kinds of snow years.

It should be noted that in this study, the proposed LSTM method was tested for several SNOTEL stations in the one watershed (East River watershed), but as automated, it can be used for any other stations in other watershed, hence, the model and the system are not dependent on any dataset and station. As presented in **Figure 5**, any location (station related to any watershed) can be selected for SWE prediction.

Model Evaluation Criteria

To evaluate forecasting performance, we can use different statistical criteria. The ones we use include the Nash Sutcliffe model efficiency coefficient (NSE) (Nash and Sutcliffe, 1970) and Rooted Mean Square Error (RMSE) which are a widely used performance evaluation method for hydrological modeling (Krause et al., 2005; Arnold et al., 2012). Both of these compare predicted values with observed values. The NSE evaluates the model performance to predict testing data different from the mean and gives the proportion of the initial variance accounted for by the model (Nash and Sutcliffe, 1970). The RMSE is used to evaluate how closely the predicted values match the observed values, based on the relative range of the data.

$$NSE = 1 - \frac{\sum_{i=1}^{n} \left(Y_i^o - Y_i^p\right)^2}{\sum_{i=1}^{n} \left(Y_i^o - Y'\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i^o - Y_i^p\right)^2}{n}} \tag{2}$$

where Y_i^o, Y_i^p and Y'epresent the observed, predicted and the average observed data at time i respectively. NSE ranges from $-\infty$ to 1, and the value close to 1 is equivalent the better model performance (Arnold et al., 2012). In general, a lower RMSE represents a higher accuracy and a better fit.

RESULTS

Forecasts and Performance

The method described above generates a site-specific LSTM model which can be used to predict SWE. This model can be trained using different datasets (e.g., the last 10 years of data),

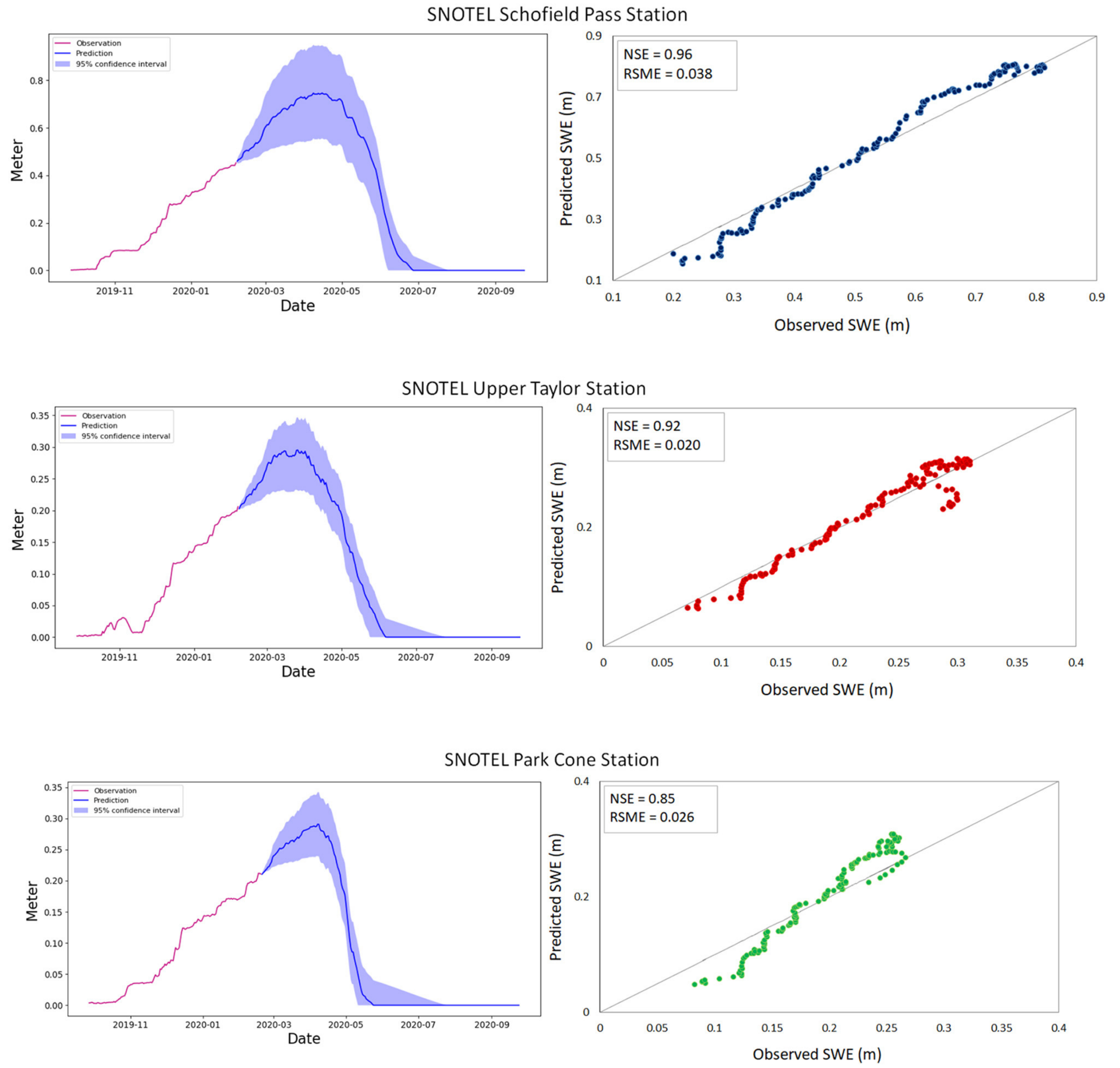


FIGURE 7 | SWE observed and predicted for the current year (2020) with their performance, for SNOTEL Schofield Pass station, SNOTEL Upper Taylor station, and SNOTEL Park Cone station.

and be used to predict SWE dynamics in different types of years (low snow, medium snow, high snow years). The model can use any specified start date in the past to evaluate the performance of the approach in SWE forecasting.

We evaluated SWE forecast performance, obtained from LSTM model, by considering 3 month forecasting for different SNOTEL stations (Schofield Pass, Upper Taylor, and Park Cone). The observation data obtained from stations were available until May 1, 2020 and 3 month before this time is February 1, 2020 that was the starting date to forecast. Given these conditions, the performance of the model can be evaluated with the observed data from the stations and the predicted SWE data from the LSTM model. However, the model can use any dates in the past to forecast, hence, it can be invoked programmatically makes it easy to evaluate performance and to use it for scenario modeling. An example of this prediction is shown in **Figures 5**, **6**.

Figure 7 shows the SWE prediction for the selected stations. In addition, the uncertainty in the prediction and the match between predicted and observed data for each station are presented in the associated graph for each station. The predicted data is obtained from the LSTM model and current water year data. All the graphs illustrate the both peak SWE and snowmelt timing captured within the confidence interval and therefore the performance is consistent among these three locations. The model and the results are validated by applying criteria such as NSE value, RMSE value. The LSTM model has a narrower range of RMSE between 0.026 m and 0.03 m relative to the Upper Taylor and Schofield Pass station. The value of NSE is also improved from 0.85 (Park Cone) to 0.96 (Schofield Pass). The results illustrate equally good performance. However, since the LSTM model is highly dependent on the meteorological variables such as rainfall, the model with smoother observed data is able to capture more precisely the peak of snowmelt timing and SWE forecast as well. As mentioned, all three stations shown in **Figure 7** have acceptable results, although the model performance at Park Cone Station is not as good as the other two stations, this is due to meteorological data (rainfall) which is smoother at the other two stations.

We can also evaluate the prediction behavior for different days by changing the starting date. An example of this is shown in **Figure 8** which shows the SWE prediction for the SNOTEL Schofield Pass station for different days in the past

TABLE 2 | Prediction performance for key metrics (peak SWE and Snow melt timing) for different start dates.

	From	Peak SWE (m)	Snow melt timing	Performance (NSE)
Observation	–	0.81	04-21-2020	–
Forecasting	December	1.07	04-08-2020	2.23
	January	0.98	04-08-2020	1.21
	February	0.78	04-07-2020	0.96
	March	0.77	04-07-2020	0.95
	April	0.76	04-08-2020	0.94

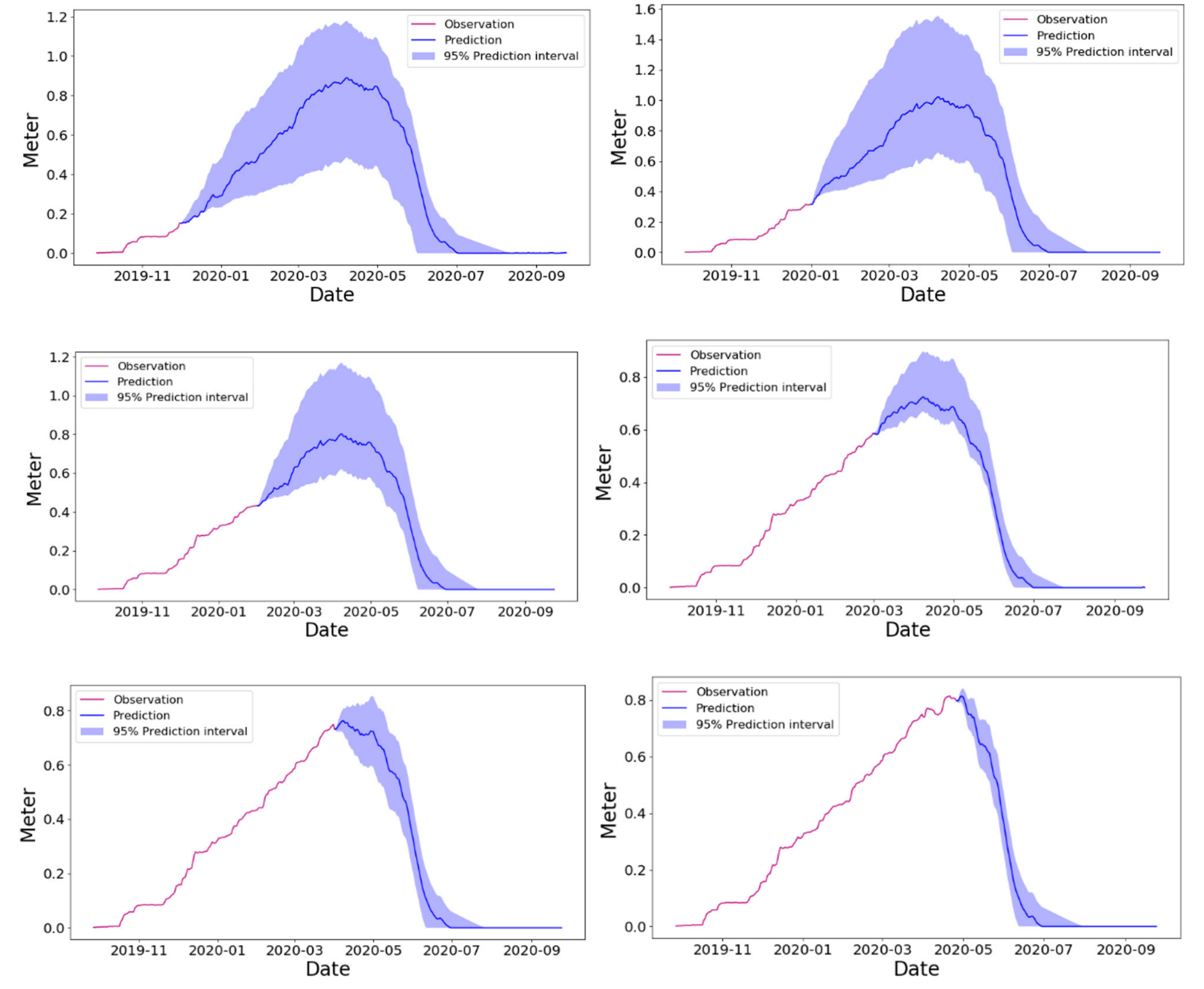


FIGURE 8 | Forecasting from the different past month for the water year 2020.

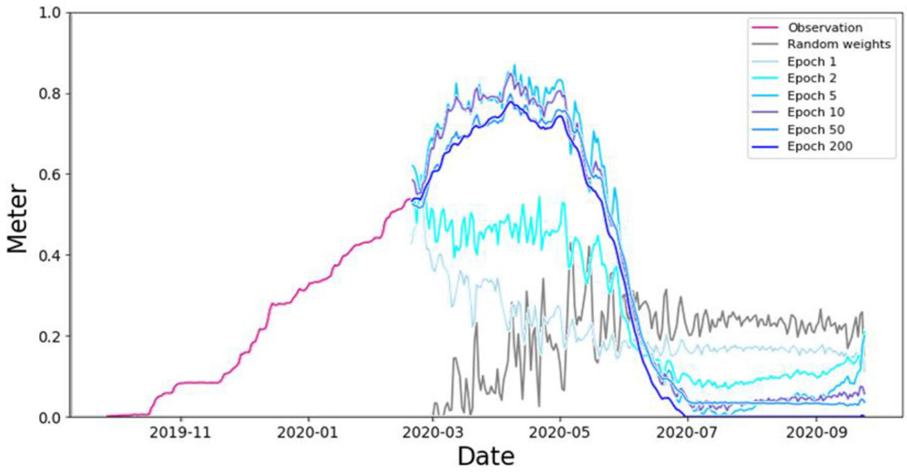


FIGURE 9 | Improvement of SWE prediction during the learning process of the LSTM as the number of epochs increases.

(from December 1, 2019, to March 1, 2020). This demonstrates the model's ability to predict SWE at any time of the water year. As is expected the confidence interval becomes narrower reduces over time as we get closer to the end of the year (**Table 2**). In all the cases, the SWE time series are contained within the interval, which validates our methodology.

Model Parameter Effect

There are multiple model parameters which we can vary in the LSTM network. These include the epoch and the number of historic years we use as training data. In NN applications, the epoch is one cycle through the full training set in which model parameters are updated.

Figure 9 illustrates the LSTM learning process for different numbers of training epochs. It shows how the training network improves from the initial state from scratch (where it has random weights) as we go to 200 epochs.

Effect of the Number of Years Used to Train Our Dataset

As mentioned before (and as should be intuitively clear), the number of years of data we use to train our dataset has an impact on the model performance, and it is important to evaluate and analyze this impact. We can evaluate the effect of the number of years used on the model performance (which is represented by Loss function or NSE). We compared 4 different lengths: 2, 5, 10, and 15 years. The comparison results are shown in **Figure 10**.

The statistical results for the overall performances of LSTM models for both length of training data are listed in **Table 3**. As shown in **Table 3**, the prediction performed well for the 10 years length (2010–2020), with average NSEs of 0.96 and RSME 0.038. Although we initially expected that increasing the number of years used to train our model (15 years for our model 2005–2020) would have better performance, the 10 years window size performed better. This behavior can be observed in **Figure 10**, in the Predicted-Observed graph. The blue dots that represent 10 year window size (2010–2020) show a better performance than the red dots that represent the 15 year window size. This may be associated with a shift in system behavior—which would be better expressed in recent data than in older data. In addition, it should be noticed that the confidence interval becomes narrower when the window size increases (**Figure 10**).

We can combine the results shown in **Figures 9–11** which shows the loss function behavior for different lengths of training data. As shown in **Figure 11**, the Loss function of the LSTM model decreased (or NSE increased) when the length of training data increases. It should be noticed also that when training with longer dataset the curve of the loss function becomes smoother. However, as shown in **Table 3** and **Figure 10** adding more years of data beyond 10 years does not increase performance. It is interesting to consider why this is, and while a detailed analysis of this falls outside the scope of this paper it could be because SWE characteristics have changed over the last 10 years. If this is the case, more recent SWE behavior would be a better predictor of current behavior than SWE behavior of 15 or 20 years back.

LSTM Automation

In this study, we have presented a step-by-step workflow on the SWE prediction by obtaining the metrological data form stations, training the model with the different windows size, checking the confidence interval, plotting the results and calculating the performance of the prediction. However, all these process and capability can be automated at different levels. First by automatically creating trained networks for any SNOTEL site, using API-able approach and creating daily updated predictions using new data for every day by rerunning the prediction. Finally, the predicted results can be delivered to interested end users. This delivery can be either done through an API or through a web interface as shown in **Figures 5**, **6**. Due to the flexibility of the API, the effect of using different training datasets can be rapidly compared.

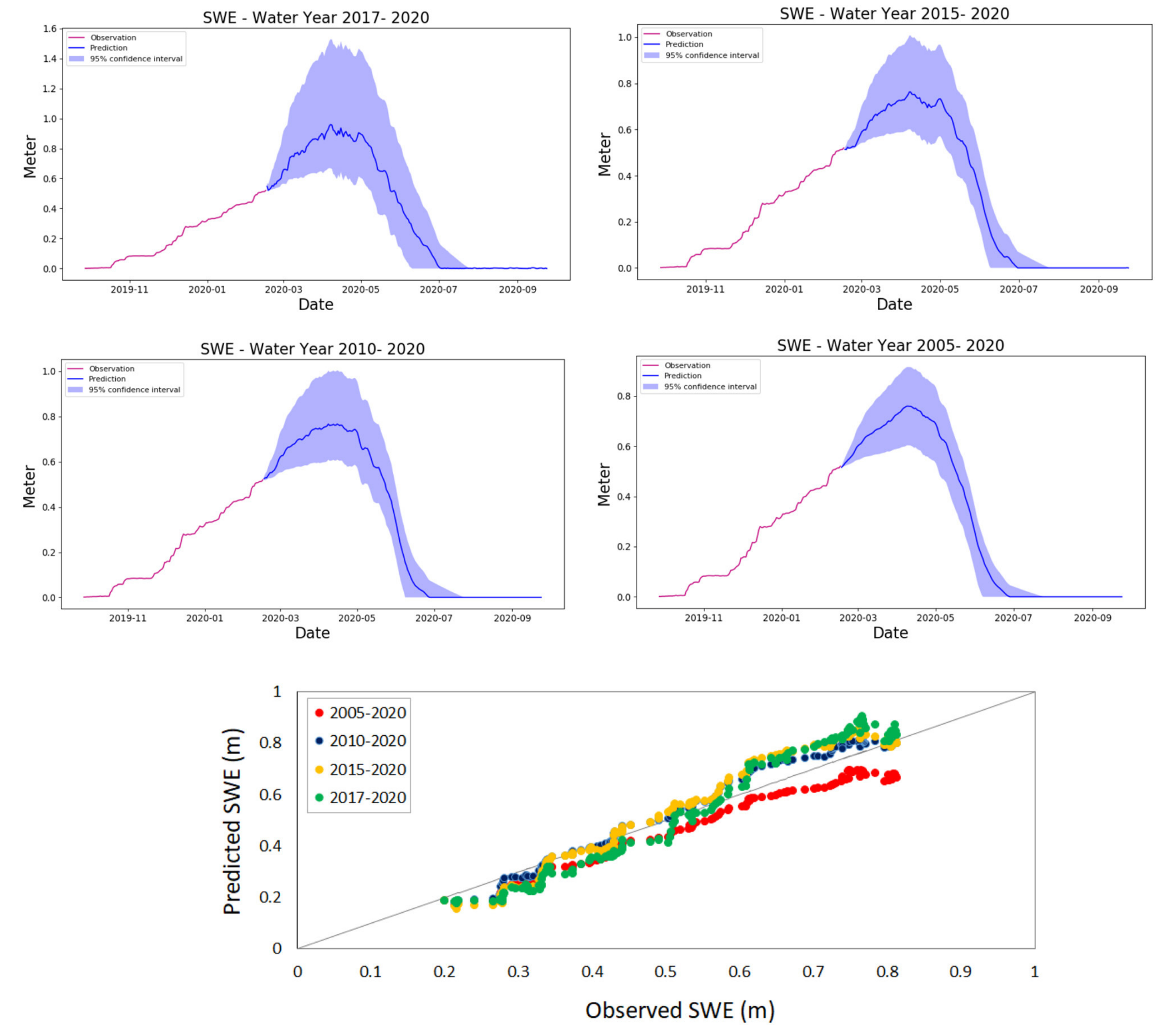


FIGURE 10 | SWE predicted for the different length of years of training data (2, 5, 10, and 15 years).

TABLE 3 | Statistics of LSTM model for SWE prediction on the SNOTEL Schofield Pass Station for different window size.

	2005–2020	2010–2020	2015–2020	2017–2020
NSE	0.88	0.96	0.89	0.85
RSME	0.063	0.038	0.062	0.07

DISCUSSION

In this study, we demonstrated that LSTM networks can be trained to accurately predict SWE behavior for different NRCS SNOTEL stations. Prediction accuracy and performance were analyzed for different epoch number and length of training data. Our results demonstrate that training data length affects the model performance. While 7–10 years of training data length seems to be suitable for the sites we examined this number should be determined for different stations and climate conditions.

There are multiple other efforts which have focused on SWE. This includes the work by Guan et al. (2013) which retrospectively estimates SWE distribution by using the blended method. Similarly (Fassnacht et al., 2003) applied inverse weighted distance and regression techniques to evaluate SWE across the entire Colorado River. Bair et al. (2018), used different machine learning techniques (bagged regression trees and feed-forward neural networks). Schneider and Molotch (2016) used regression techniques to estimate the spatial distribution of

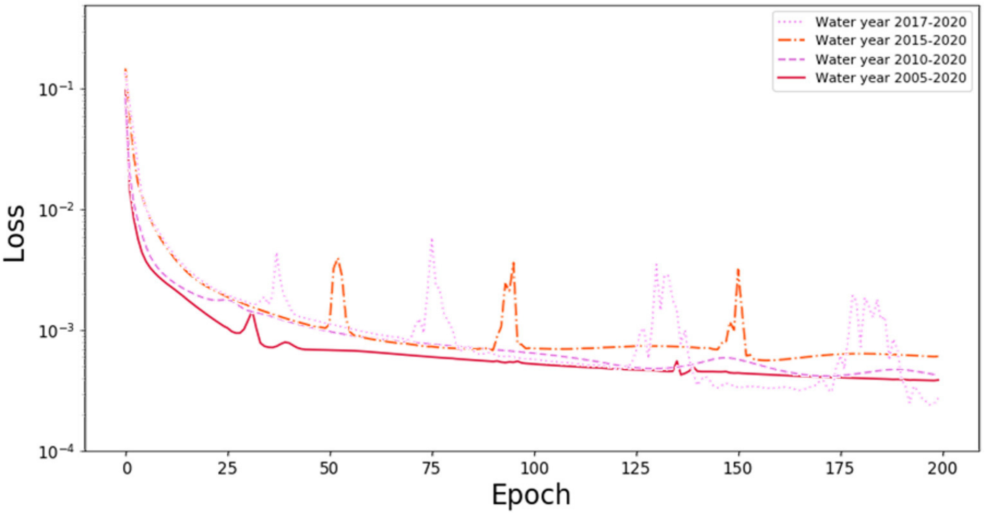


FIGURE 11 | Loss function after 200 epochs for predictions using different water years (2, 5, 10, and 15 years).

SWE for the Upper Colorado River basin weekly from January to June 2001–2012. Leisenring and Moradkhani (2011) compare common sequential data assimilation methods, the ensemble Kalman filter (EnKF), the ensemble square root filter (EnSRF), and four variants of the particle filter (PF), to explain. These efforts differ from ours in that we provide a forecast for the water year. In addition, in this study, we analyze presented the impact of the training data set on the forecast accuracy of LSTM. This analysis complements the work by other groups which used the LSTM method to runoff prediction such as Kratzert et al. (2018) and Zhang et al. (2018).

We demonstrated the feasibility of automated model/data coupling and model generation, with the model accessible through the API and through a web interface. We expect that this ability will be of interest to multiple stakeholders. One limitation of the current study is that the current prediction effort uses single station data. We are currently exploring how we can extend this prediction by integrating multiple SNOTEL stations and satellite data on watershed snow coverage to give watershed-wide SWE and water predictions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: data can be accessed at: https://www.wcc.nrcs.usda.gov/.

AUTHOR CONTRIBUTIONS

AM implemented and tested the LSTM algorithm and applied it to the Snotel data. RV designed and enhanced the data model and provided method validation. EA implemented the data ingestion pipeline for the Snotel data. DJ designed and implemented the overall backend and supported API implementation. AR developed the webinterface. MF and HW developed an initial implantation of the LSTM method on SWE data. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under SBIR Award DE-SC0018447 to Subsurface Insights (Cloud Based Watershed And Terrestrial Ecosystem Data Management, Integration And Analytics), and under Award Number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory as part of the Watershed Function Scientific Focus Area.

ACKNOWLEDGMENTS

This research benefited from numerous discussions on SWE with Rosemary Carroll, Ken Williams, and the overall LBNL SFA team.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]*. arXiv:1603.04467.

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

Bair, E. H., Abreu Calfa, A., Rittger, K., and Dozier, J. (2018). Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan. *Cryosphere* 12, 1579–1594. doi: 10.5194/tc-12-1579-2018

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., and Dozier, J. (2006). Mountain hydrology of the western United States. *Water Resourc. Res.* 42:W08432. doi: 10.1029/2005WR004387

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Carroll, R. W., Bearup, L. A., Brown, W., Dong, W., Bill, M., and Willlams, K. H. (2018). Factors controlling seasonal groundwater and solute flux from snow-dominated basins. *Hydrol. Processes* 32, 2187–2202. doi: 10.1002/hyp.13151

Chollet, F., et al. (2015). *Keras*. Available online at: https://github.com/fchollet/keras

Deems, J. S., Painter, T. H., Barsugli, J. J., Belnap, J., and Udall, B. (2013). Combined impacts of current and future dust deposition and regional warming on Colorado River Basin snow dynamics and hydrology. *Hydrol. Earth Syst. Sci.* 17, 4401–4413. doi: 10.5194/hess-17-4401-2013

Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J., and Jiang, J. (2020). Comparison of long short term memory networks and the hydrological model in runoff simulation. *Water* 12:175. doi: 10.3390/w12010175

Fang, K., Shen, C., Kifer, D., and Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental US using a deep learning neural network. *Geophys. Res. Lett. 44,* 11.030–11.039. doi: 10.1002/2017GL075619

Fassnacht, S. R., Dressler, K. A., and Bales, R. C. (2003). Snow water equivalent interpolation for the Colorado River Basin from snow telemetry (SNOTEL) data. *Water Resourc. Res.* 39:1208. doi: 10.1029/2002WR001512

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining Knowl. Discov.* 333, 917–963. doi: 10.1007/s10618-019-00619-1

Guan, B., Molotch, N. P., Waliser, D. E., Jepsen, S. M., Painter, T. H., and Dozier, J. (2013). Snow water equivalent in the Sierra Nevada: blending snow sensor observations with snowmelt model simulations. *Water Resources Res.* 49, 5029–5046. doi: 10.1002/wrcr.20387

Harte, J., Saleska, S. R., and Levy, C. (2015). Convergent ecosystem responses to 23-year ambient and manipulated warming link advancing snowmelt and shrub encroachment to transient and long-term climate–soil carbon feedback. *Glob. Change Biol.* 21, 2349–2356. doi: 10.1111/gcb.12831

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., et al. (2016). Observations data model 2: a community information model for spatially discrete earth observations. *Environ. Modell. Softw.* 79, 55–74. doi: 10.1016/j.envsoft.2016.01.010

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., and Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* 10:1543. doi: 10.3390/w10111543

Huang, H.-C., and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Comput. Stat. Data Anal.* 22, 159–175. doi: 10.1016/0167-9473(95)00047-X

Hubbard, S. S., Williams, K. H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., et al. (2018). The East River, Colorado, watershed: a mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. *Vadose Zone J.* 17, 1–25. doi: 10.2136/vzj2018.03.0061

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2018). *forecast: Forecasting Functions for Time Series and Linear Models, 2018.* Software, R package.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022. doi: 10.5194/hess-22-6005-2018

Krause, P., Boyle, D., and Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* 5, 89–97. doi: 10.5194/adgeo-5-89-2005

Kumar, D. N., Raju, K. S., and Sathish, T. (2004). River flow forecasting using recurrent neural networks. *Water Resour. Manage.* 18, 143–161. doi: 10.1023/B:WARM.0000024727.94701.12

Le, X.-H., Ho, H. V., Lee, G., and Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* 11:1387. doi: 10.3390/w11071387

Leisenring, M., and Moradkhani, H. (2011). Snow water equivalent prediction using Bayesian data assimilation methods. *Stochastic Environ. Res. Risk Assess.* 25, 253–270. doi: 10.1007/s00477-010-0445-5

Magnusson, J., Nævdal, G., Matt, F., Burkhart, J. F., and Winstral, A. (2020). Improving hydropower inflow forecasts by assimilating snow data. *Hydrol. Res.* 51, 226–237. doi: 10.2166/nh.2020.025

Markstrom, S. L., Hay, L. E., Ward-Garrison, C. D., Risley, J. C., Battaglin, W. A., Bjerklie, D. M., et al. (2012). *Integrated Watershed-Scale Response to Climate Change for Selected Basins Across the United States. U.S. Geological Survey.* Scientific Investigations Report 2011-5077, 143

Nash, J. E., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—a discussion of principles. *J. Hydrol.* 10, 282–290. doi: 10.1016/0022-1694(70)90255-6

Odei, J. B., Hooten, M. B., and Jin, J. (2009). *Inter-Annual Modeling and Seasonal Forecasting of Intermountain Snowpack Dynamics.* 870–878.

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., et al. (2016). The airborne snow observatory: fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sens. Environ.* 184, 139–152. doi: 10.1016/j.rse.2016.06.018

Sahoo, B. B., Jha, R., Singh, A., and Kumar, D. (2019). Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophys.* 67, 1471–1481. doi: 10.1007/s11600-019-00330-1

Schneider, D., and Molotch, N. P. (2016). Real-time estimation of snow water equivalent in the U pper C olorado R iver B asin using MODIS-based SWE reconstructions and SNOTEL data. *Water Resour. Res.* 52, 7892–7910. doi: 10.1002/2016WR019067

Sloat, L. L., Henderson, A. N., Lamanna, C., and Enquist, B. J. (2015). The effect of the foresummer drought on carbon exchange in subalpine meadows. *Ecosystems* 18, 533–545. doi: 10.1007/s10021-015-9845-1

Wainwright, H. M., Steefel, C., Trutner, S. D., Henderson, A. N., Nikolopoulos, E. I., Wilmer, C. F., et al. (2020). Satellite-derived foresummer drought sensitivity of plant productivity in Rocky Mountain headwater catchments: spatial heterogeneity and geological-geomorphological control. *Environ. Res. Lett.* 15:084018. doi: 10.1088/1748-9326/ab8fd0

Winstral, A., Magnusson, J., Schirmer, M., and Jonas, T. (2019). The bias-detecting ensemble: a new and efficient technique for dynamically incorporating observations into physics-based, multilayer snow models. *Water Resour. Res.* 55, 613–631. doi: 10.1029/2018WR024521

Xiang, Z., Yan, J., and Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resour. Res.* 56:e2019WR025326. doi: 10.1029/2019WR025326

Zhang, J., Zhu, Y., Zhang, X., Ye, M., and Yang, J. (2018). Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* 561, 918–929. doi: 10.1016/j.jhydrol.2018.04.065

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.

ORIGINAL RESEARCH
published: 19 November 2020
doi: 10.3389/frwa.2020.573034

Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network

Reetik Kumar Sahu[1], *Juliane Müller*[1]*, *Jangho Park*[1], *Charuleka Varadharajan*[2], *Bhavna Arora*[2], *Boris Faybishenko*[2] and *Deborah Agarwal*[3]

[1] *Lawrence Berkeley National Laboratory, Computational Research Division, Center for Computational Sciences and Engineering, Berkeley, CA, United States,* [2] *Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area, Berkeley, CA, United States,* [3] *Lawrence Berkeley National Laboratory, Data Science and Technology, Computational Research Division, Berkeley, CA, United States*

With the growing use of machine learning (ML) techniques in hydrological applications, there is a need to analyze the robustness, performance, and reliability of predictions made with these ML models. In this paper we analyze the accuracy and variability of groundwater level predictions obtained from a Multilayer Perceptron (MLP) model with optimized hyperparameters for different amounts and types of available training data. The MLP model is trained on point observations of features like groundwater levels, temperature, precipitation, and river flow in various combinations, for different periods and temporal resolutions. We analyze the sensitivity of the MLP predictions at three different test locations in California, United States and derive recommendations for training features to obtain accurate predictions. We show that the use of all available features and data for training the MLP does not necessarily ensure the best predictive performance at all locations. More specifically, river flow and precipitation data are important training features for some, but not all locations. However, we find that predictions made with MLPs that are trained solely on temperature and historical groundwater level measurements as features, without additional hydrological information, are unreliable at all locations.

Keywords: machine learning, groundwater level prediction, feature selection, sensitivty analysis, hyperparameter optimization

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Jie Niu,
Jinan University, China
Wei Shao,
Nanjing University of Information
Science and Technology, China

****Correspondence:***
Juliane Müller
julianemueller@lbl.gov

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *15 June 2020*
Accepted: *12 October 2020*
Published: *19 November 2020*

Citation:
Sahu RK, Müller J, Park J, Varadharajan C, Arora B, Faybishenko B and Agarwal D (2020) Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network. Front. Water 2:573034. doi: 10.3389/frwa.2020.573034

INTRODUCTION

Groundwater is an important source of freshwater, accounting for almost 38% of the global irrigation demand (Siebert et al., 2010). With growing economies and increasing food demand, the stress on freshwater aquifers has increased in places like North America and Asia (Aeschbach-Hertig and Gleeson, 2012). This situation is further aggravated by increased climate variability. In California, USA, groundwater provides nearly 40% of the water used by the state's cities and farms. Many of the state's groundwater basins have experienced long-term overdraft due to withdrawal rates exceeding recharge rates. The negative impacts of long-term overdraft include higher energy requirements for pumping water from deeper wells, land subsidence, reduced river flow, and impaired water quality (especially in coastal aquifers due to saltwater intrusion). Thus, in 2014, following a series of droughts, the Sustainable Groundwater Management Act (SGMA) was passed, requiring local agencies to sustainably manage groundwater and minimize undesirable results (DWR, 2020). This in turn requires decision makers access to accurate, reliable, and timely predictions of groundwater levels.

Traditionally, groundwater depths and other water budget components such as runoff and soil moisture are estimated using mechanistic multi-scale, multi-physics simulation models such as MODFLOW, PARFLOW, HydroGeoSphere, and TOUGH (Xu et al., 2011; Steefel et al., 2015; Langevin et al., 2017). These models capture physical processes of mass, momentum, and energy transfer through partial differential equations and require extensive characterization of hydrostratigraphic properties and accurate boundary conditions, including recharge sources, climate variability and changes in water use (Sahoo et al., 2017). Such information is not always known a priori, and some parameters can only be determined by solving an inverse problem (Arora et al., 2011), which itself requires running simulation models repeatedly until their values have been determined, thereby substantially increasing the computational costs (Arora et al., 2012). In addition, running the high-fidelity simulation models at high resolution requires high performance computing resources. Therefore, it is difficult for groundwater sustainability agencies and policy makers to use these simulations to guide water management decisions.

With the improvement of sensor technologies and data systems, an unprecedented amount of environmental data are being collected, through established long-term monitoring networks, including river flow, groundwater level, water quality, temperature, and precipitation (Rode et al., 2016). This has resulted in an increased interest in applying ML methods for hydrological applications (Deka, 2014; Shen, 2018) such as river flow forecasts (Lin et al., 2006; Rasouli et al., 2012; Deo and Sahin, 2016; Kratzert et al., 2018); water quality estimation and prediction (Ahmad et al., 2010; Najah et al., 2013; Xu and Liu, 2013), and water demand forecasts (Ghiassi et al., 2008; Herrera et al., 2010; Adamowski et al., 2012; Tiwari and Adamowski, 2013).

Deep learning (DL) models can be trained to approximate the behavior of a complex system, such as a groundwater basin, in a computationally inexpensive way while making highly accurate predictions. DL techniques can utilize the climate and hydrogeology data to capture the relationships between groundwater levels and other dependent features such as nearby river flow, precipitation and temperature. Recent advances in ML have enabled making groundwater predictions by using purely data-driven models (Taormina et al., 2012; Moosavi et al., 2013; Sahoo et al., 2017; Müller et al., 2020). ML techniques have been used for both prediction and optimization purposes including modeling of groundwater levels and or quality, optimization of groundwater well design, pumping rate, and location (Banerjee et al., 2011; Gaur et al., 2013). As an example, Sahoo et al. (2017) utilized a hybrid feedforward neural network (FNN) to model groundwater level changes in the High Plains aquifer, United States, using both *in-situ* and remote measurements with model simulations of different input features (climate and anthropogenic). Their DL models were trained on monthly data spanning over 33 years. Emamgholizadeh et al. (2014) built a groundwater prediction model using an FNN model built from 9 years of monthly data that included rainfall recharge, pumping rate and irrigated return flow at the Bastam Plain, Iran. The FNN model showed the highest accuracy when built with a lag time of 2 months giving a prediction error of about 3% of difference between observed maximum and minimum levels. Guzman et al. (2017) utilized a dynamic form of a Recurrent Neural Network (RNN) model to predict groundwater levels in the Mississippi River Valley Alluvial aquifer, United States. Eight years of daily historical input time series including precipitation and groundwater levels were used to forecast groundwater levels for up to 3 months. Their results showed that models generated with 100 lag days provided the most accurate prediction of groundwater levels. Adamowski and Chan (2011) coupled discrete wavelet transforms (WA) and artificial neural networks (ANN) to predict groundwater levels using monthly average precipitation, temperature, and groundwater level at two sites in the Chateauguay watershed in Quebec, Canada. Their WA-ANN models performed better than standard autoregressive integrated moving average (ARIMA) time series models.

All of these prior studies involved building the DL model to predict groundwater levels at a single well. In contrast, Mohanty et al. (2015) built an FNN model to predict weekly groundwater levels simultaneously at 18 different locations in the Mahanadi Delta, India. The input features in this study included weekly values of precipitation, pumping from tubewells, and the river stage. The DL model could predict groundwater levels up to 4 weeks of lead time with a prediction error of about 8% of the annual groundwater-level change. Our previous study (Müller et al., 2020) compared results from a variety of DL methods including multilayer perceptron (MLP), RNN, long short term memory (LSTM), and 1D-convolutional neural network (CNN) designed with our hyperparameter optimization approach for both single- and multi-well groundwater level predictions in California, and were able to attain prediction accuracies of 6–20%, depending on the DL model.

Each of the referenced applications utilize different ML models and architecture under different scenarios such as multi-point vs. single-point sites, with data of varying temporal resolutions (hourly, daily, weekly, and monthly). Despite these differences, and the constraints imposed by data availability, all of these models have similar ranges for prediction accuracies. This raises the following questions: What is the right DL model to use? How should the parameters of the model be tuned? What data should we use to build an accurate prediction model? Most importantly, in order to use DL models effectively to make reliable future groundwater predictions in a computationally inexpensive manner, we must first understand which input features are necessary and sufficient. Additionally, these prior studies only report results from a single optimized neural network, and they do not address the inherent stochasticity that arises during training when using stochastic gradient descent (Amari, 1993). Thus, when training the DL model for the same architecture multiple times, we obtain different performances, and therefore different future predictions. In order to ensure the reliability of the DL model predictions, we must report confidence intervals, as well as average, best, and worst-case predictions. These uncertainty estimates will enable water managers to analyze and explore a wide spectrum of sustainable management practices and to identify those that are the most robust for all scenarios.

To address this critical need, we conduct a critical analysis of the sensitivity of DL model predictions to the choice of input features used to train a model. In particular, we compare the sensitivity of groundwater predictions to different choices of input features including groundwater levels, temperature, precipitation, and river flow. This kind of analysis will extend our understanding of the applicability of ML techniques for hydrological predictions and provide guidance on how to build accurate and reliable models. These DL models can potentially enable water managers to better prepare and sustainably manage water resources in the face of future climate variability.

The remainder of this article is organized as follows. In section Description of Numerical Study, we provide details of the setup for our numerical experiments (including the data collection, processing, and model framework), and present their results in section Numerical Results. In section Discussion, we discuss the results of the numerical experiments in the context of applying ML techniques to groundwater and outline potential future research directions. Finally, in section Conclusion, we present the conclusions of our study.

DESCRIPTION OF NUMERICAL STUDY

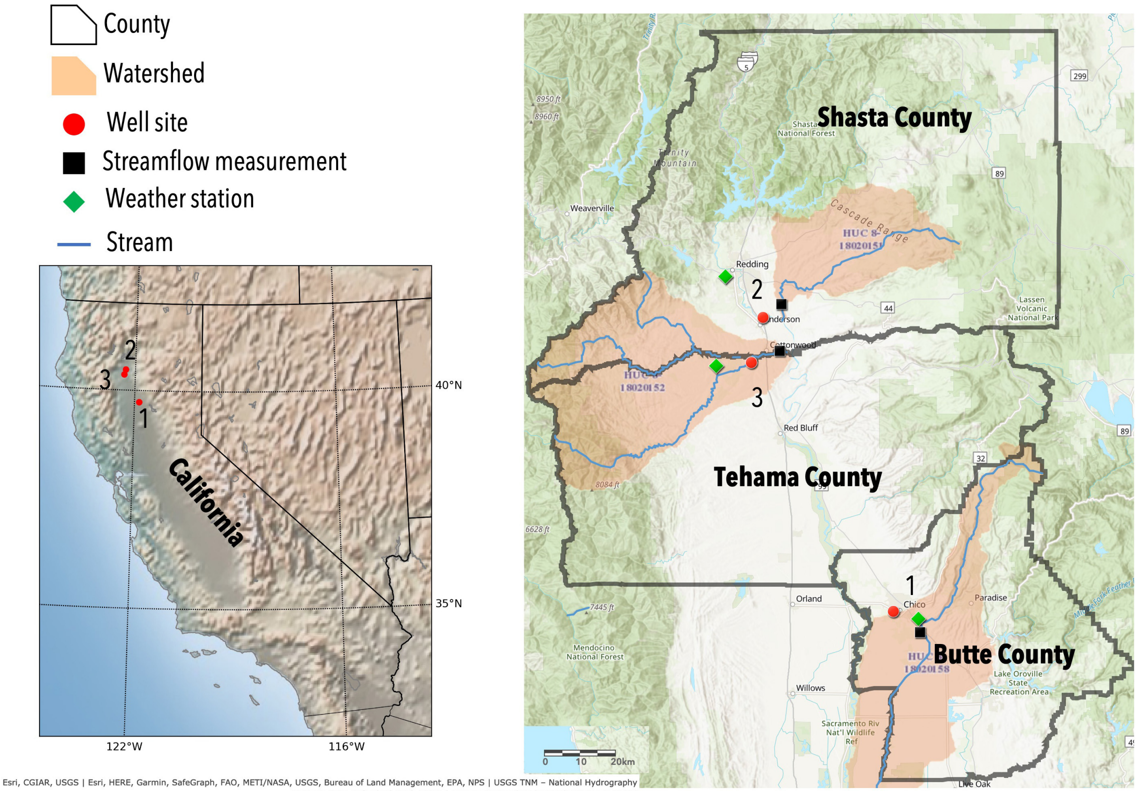
In this section, we describe the setup of our numerical experiments, including the data we used, model selection and hyperparameters, our experimental setup for sensitivity analysis, and the method for computing confidence intervals.

Data Collection and Preparation

We focused our study on wells in three different locations in Northern California, United States in Butte County, Shasta County, and Tehama County with different hydrostratigraphy and land use (**Figure 1**). Moreover, they represent different SGMA basin prioritization categories (high, medium, and low respectively), which are determined by historical groundwater trends (DWR, 2020). We primarily chose these well locations since they had relatively long-term daily observations that were publicly available. We briefly describe the sites below.

The Butte County Well Site

The majority of Butte county is located in the Sacramento Valley groundwater basin which is filled with sediments from marine and terrestrial environments. The groundwater well in this study (22N01E28J001M) is a dedicated monitoring well of depth 200 m and screened at 140–170 m. The well site is located in the Vina subbasin of Butte county, which covers 750 sq. km of the



Esri, CGIAR, USGS | Esri, HERE, Garmin, SafeGraph, FAO, METI/NASA, USGS, Bureau of Land Management, EPA, NPS | USGS TNM – National Hydrography Dataset. Data Refreshed April, 2020. | source: National Hydrography Dataset: USGS

FIGURE 1 | Location of the three well sites in California: Butte, Shasta, and Tehama County. The red dots show the location of the observation wells. Weather station (green diamond) and river flow monitoring station (black square) are located close to the well site. The map was created using ArcGIS® software by Esri.

northern portion of Butte county. This subbasin is categorized as a high priority basin under the 2019 SGMA basin prioritization report (DWR, 2020), showing an immediate need to mitigate the groundwater depletion therein. The aquifer system includes stream channel and alluvial fan deposits, and deposits of the Modesto and Tuscan formations (DWR, 2004). Groundwater is a major water source for about 150 sq. km of irrigated land in the basin. Out of the total county wide freshwater withdrawal, **about 94% is attributed to groundwater pumping** for different uses (Dieter et al., 2018) while the rest is from surface water withdrawals. The nearest discharge monitoring station (Butte Creek Durham) measures the daily discharge rate at the Butte Creek which is about 8 km from the well. The Butte Creek and the much larger Feather Creek are the main sources for surface water diversion in the county (Butte County Department of Water and Resource Conservation, 2016). Temperature and precipitation data were obtained from the Chico weather station located 7 km from the well.

The Shasta County Well Site

The groundwater well in Shasta County is an observation well (30N04W10H005M) of depth 49 m and screened at 33–48 m. It is located in the Anderson subbasin which is a part of the Redding Groundwater Basin covering an area of about 400 sq. km. This subbasin is one of the primary agricultural regions in the county, is categorized as a medium priority basin according to the SGMA guidelines (DWR, 2020). Eighty to ninety percent of the basin's precipitation typically occurs from November to April. The aquifer system is comprised of continental deposits of late Tertiary to Quaternary age. The Quaternary deposits include Holocene alluvium and Pleistocene Modesto and Riverbank formations (California Department of Water Resources, 2004). The nature of surface water-groundwater interaction across the basin is complex, both spatially and temporally, but in most areas shallow groundwater levels lead to groundwater discharge to surface streams. During pronounced drought conditions, groundwater levels may decline to a level such that streams that formerly gained river flow from groundwater discharge now recharge the groundwater system through streambed infiltration. Major water supplies in this region are provided by surface storage reservoirs (Bureau of Reclamation, 2011). Agricultural, industrial, and municipal groundwater users in the basin pump primarily from deeper continental deposits, whereas domestic groundwater users generally pump from shallower deposits. **Groundwater withdrawals contribute about 54% of the total county wide freshwater withdrawal** from different sources (Dieter et al., 2018). Although this well is closest to the Sacramento River, the nearest discharge monitoring station is located in Cow Creek, which feeds into the Sacramento River and is about 5 km away. Since the nearest discharge station in the Sacramento River was located 20 km upstream, we chose to use the discharge observations from the Cow Creek station, as the closest approximation of discharge trends and seasonality that determines surface water influence on groundwater behavior. The temperature and precipitation data were obtained from the Redding Fire station located 15 km from the well.

The Tehama County Well Site

The groundwater well in Tehama County is an observation well (29N04W20A002M) of depth 137 m, with a screen at 109–131 m depth. It is located in the Bowman subbasin which is categorized as a low priority basin. This subbasin, is also a part of the Redding groundwater basin covering 495 sq. km in the north central portion of the county. The aquifer system of the Bowman subbasin is comprised of continental deposits of late Tertiary to Quaternary age. The Quaternary deposits include Holocene alluvium (thickness ranging from 0 to 10 m) and Pleistocene Modesto and Riverbank Formations (thickness ranging from 0 to 15 m). The Tertiary deposits include the Pliocene Tehama Formation (thickness may reach up to 150 m) and Tuscan Formations (thickness may reach up to 750 m) (Ayres and Brown, 2008). The Bowman subbasin is primarily a rural area where groundwater is used for agriculture, domestic, and municipal purposes. Groundwater sources represent the majority of supply, followed by local surface water. During an average water year, Tehama County does not experience any water shortages since the water supply is generally higher than the water demand. **Groundwater contributes about 37% of the county's total freshwater withdrawal** (Dieter et al., 2018). The observation well is located 0.6 km from the Cottonwood Creek. However, the closest river flow monitoring station is located about 8 km from the test site. The weather data was obtained from the Davis Ranch station located 10 km from the well.

Input Features, Data Sources, and Preprocessing Methodology

For our DL model, we identified features that we expect to directly or indirectly impact groundwater levels including temperature (T), precipitation (P), and river flow (i.e., discharge; Q). Daily historical observations of these variables from 2010 to 2018 are used together with groundwater level measurements (G) to train the neural network models (**Table 1**). In addition, we use the week of the year of the measurements' timestamps as a training feature, which naturally represents the inherent seasonality in the dataset.

The observation wells indicate regional drawdowns due to groundwater extraction through pumping activities (for agriculture, urban use, or other). However, pumping data are not reported in California, and are not publicly available. Higher pumping rates are observed during summer months when the temperature is high with infrequent and small precipitation events and low surface water availability. Low precipitation years therefore lead to higher depletion rates, whereas wet years show lower depletion rates (**Figure 2**). Our assumption is that the ML model can capture the interaction between groundwater level and pumping through other proxy hydrological or climate variables that typically drive pumping (precipitation, temperature, or river flow).

All the datasets are processed for quality assurance and quality control (QA/QC), including gap-filling (also called as "missing value imputation") and normalization. The QA/QC helps to remove erroneous values or outliers (unrealistic values) in the measurements due to faulty sensors or equipment

TABLE 1 | Data sources of historical observations.

County	Groundwater well station code	Average depth to groundwater level from surface (meters)	$\Delta_{\max} = \mathrm{GWL}_{\max}^{obs} - \mathrm{GWL}_{\min}^{obs}$ (meters)	Weather station code	River flow station code
Butte	22N01E28J001M	16.3	9.4	Chico (CHI)	Butte Creek Durham (BCD)
Shasta	30N04W10H005M	6.2	4.6	Redding Fire Station (RFS)	Cow Creek (COW)
Tehama	29N04W20A002M	15.6	4.2	Davis Ranch (DVR)	Cottonwood Creek (COT)

CNRA, California Natural Resources Agency; CDEC, California Data Exchange Center. Observations were obtained from CNRA and CDEC.

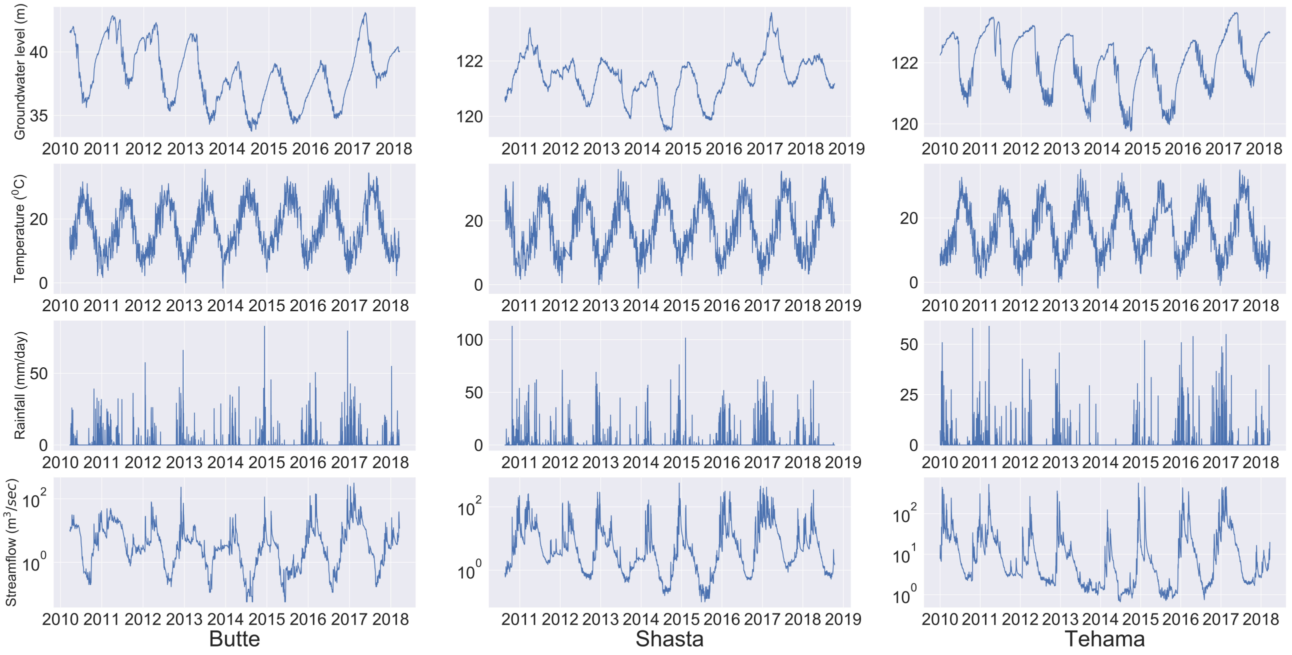


FIGURE 2 | Timeseries of all features at the three well sites at a daily frequency from 2010 to 2018. The top panel at each site shows the groundwater level (meters above mean sea level). The second panel shows the temperature (°C), the third panel shows the precipitation (mm), and the bottom panel of sites shows the river flow (m3/sec).

failures, and the data gaps are then filled by using time series imputation techniques. We imputed the missing values using the *imputeTS* package (Moritz and Bartz-Beielstein, 2017) in R. This package is used for univariate time series imputation. We use the na.seadec (Seasonally Decomposed Missing Value Imputation) function with the application of the "kalman" algorithm, of the *imputeTS* package, which is well-suited for gap filling of time series exhibiting seasonality. Using this approach, the seasonal component is first removed, missing data are imputed in the general trend and then the seasonal component and the general trend are combined to generate a gap-filled uninterrupted time series. The missing values of each of the features in the datasets contribute to at most 2% of total length of the time series. The ratios of missing data to the total period at each monitoring station are provided in **Supplementary Table 1**.

Since our input features have significantly different ranges of absolute values, we scale each dataset to the range [0, 1]. This ensures that during the learning process (iterative weight adjustment) a percentage change in the weighted input sample is reflected with a similar percentage change at the nodes of the output layer (Kanellopoulos and Wilkinson, 1997). To this end, we use the minimum and the maximum values of each dataset. For temperature, precipitation, and river flow data, these lower, and upper limits are known and the task is unambiguous. Since the observed values for the river discharge have a huge variation (orders of magnitudes difference between summer and winter due to the lack of precipitation in California in the summer months), we log-normalized the values to attenuate the effect of high values that would occur in a uniform scaling. For the groundwater levels, determining the minimum and maximum levels is more difficult as the water table depths reached unprecedented lows during the $2012-2016$ drought. Fixing the lower bound at the historically observed minimum value is unreliable, because future droughts may cause the lowest observed groundwater level to further decrease. A similar argument can be made for the maximum groundwater levels, which are expected to increase in particular for heavily overdrafted basins as sustainable groundwater management practices are being implemented. Thus, in this study, we set the minimum groundwater level to the lowest historically observed value less 15% and the maximum level to the highest historically observed value plus 15%. Given these lower and upper bounds, we then scale the groundwater data to [0, 1]. Note that scaling the input data values does not force the predicted values to remain within the lower and upper limits used for scaling.

Neural Network Model and Hyperparameter Tuning

In this study we implement an MLP type of neural network to build the groundwater prediction model. The MLP is a feedforward type of neural network with different hyperparameters that need to be adjusted before its training. The MLP was chosen as it was the best performing model in terms of accuracy and compute time, based on comparison with CNN, RNN, LSTM neural networks (Müller et al., 2020).

The choice of hyperparameters reflect the complexity of the MLP model. Hand tuning, grid and random sampling are the most widely used methods for choosing the hyperparameters of DL models (Bergstra and Bengio, 2012). Hand tuning is time consuming, it does not scale well to large search spaces, and it does not usually lead to the optimal hyperparameters. Thus, we use an automated hyperparameter optimization (HPO) method to find the best DL model hyperparameters.

We follow (Müller et al., 2020) to formulate a bilevel optimization problem:

$$\min_{\theta,\ \mathbf{w}^*} \ell\left(\theta,\ \mathbf{w}^*;\ \mathcal{D}_{val}\right) \tag{1}$$

$$s.t.\ \theta \in \Omega \tag{2}$$

$$\mathbf{w}^* \in \arg\min{}_{w \in \mathcal{W}} L(\mathbf{w}; \theta, \mathcal{D}_{train}) \tag{3}$$

where θ are the hyperparameters in the search space Ω; $\mathbf{w}$ are the weights and biases associated with each node in the MLP, $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$ are the training and validation datasets, respectively. The search space Ω is a product of finite sets of integer values. At the upper-level optimization problem (Equation 1), the optimizer selects a set of hyperparameters θ (the model architecture). Given θ, the lower-level problem (Equation 3) is solved with RMSprop, in which we find optimal weights $\mathbf{w}^*$ that minimize the loss function L for the training data. Once we obtain $\mathbf{w}^*$, we can then evaluate the upper-level objective function l that reflects how good a choice θ is. Based on the outcome for l, the optimizer at the upper-level selects the next set of hyperparameters for which the lower-level problem is solved, and so on until convergence at the upper-level is achieved. For solving the upper-level optimization problem, we use a derivative-free optimization algorithm that uses radial basis function surrogate models, see Müller et al. (2020) for further details. Since a stochastic optimizer is used to solve the lower-level problem (Equation 3), the performance of the MLP for a given architecture θ depends on the random number seed of the stochastic optimizer. Therefore, in order to obtain an approximated expected performance for a given MLP architecture, we solve the lower-level problem five times and average the results.

In our study, we search for the hyperparameters in a 6-dimensional search space, $\Omega = \prod_{h=1}^{6} \theta_h$:

- Number of layers: $\theta_1 \in \{ 1, 2, \ldots, 6 \}$
- Number of nodes per layer: $\theta_2 \in \{ 5, 10, \ldots, 50 \}$
- Number of lags: $\theta_3 \in \{ 30, 35, \ldots, 365 \}$
- Dropout rate: $\theta_4 \in \{ 0.1, 0.2, \ldots, 0.5 \}$
- Batch size: $\theta_5 \in \{ 50, 55, \ldots, 200 \}$
- Epochs: $\theta_6 \in \{ 50, 100, \ldots, 500 \}$

and we map these numbers to consecutive integers for optimization. Thus, if we used complete enumeration to find the optimal MLP architecture, we would have to train $6,120,000$ different MLPs, which is impractical for real-world decision-support applications. In the "upper level optimization," we iteratively test only 50 different MLP neural network hyperparameters (50 different hyperparameter sets that describe the network architecture). This was sufficient to achieve convergence at the test site (Müller et al., 2020). To handle the lagged temporal relationship between variables, we use the concept of Time-Delayed Neural Network (Waibel et al., 1989). A consecutive set of observations is used as one input instead of one observation. We call this amount of historical data the *lag*. Lag is one of the most important hyperparameters in a feedforward neural network for handling time series data (Zhang, 2003).

We divide the observations of all the features into training ($\mathcal{D}_{train}(\theta)$), cross-validation, ($\mathcal{D}_{val}(\theta)$, finding the optimal hyperparameters), and testing data ($\mathcal{D}_{test}(\theta)$), with a 50–25–25% split, except when indicated otherwise. The MLP models are trained to predict the groundwater level for the next time step (e.g., day or month). This output is computed based on the values of all features at the current time step and several previous time steps (equal to the lag number). For example, with a lag of 4-time steps, measurements of all features including the groundwater level from the past 4-time steps are used along with the current time step's data, to predict the groundwater level at the next time step. The MLP's output is a single groundwater level value for the next time step. When training the MLP for a given set of hyperparameters, observed values of all features are used to optimize the weights in the MLP model. During cross-validation and testing, the observed values of only temperature, precipitation, river flow, and week of year are used as drivers for making groundwater level predictions. To make groundwater level predictions over several time steps, the predicted groundwater level from the previous timestep is recursively incorporated with the observed values of temperature, precipitation and river flow to make the new input sample. Using the recursive approach during the testing and validation period, we test the capability of the MLP model to make multi-month predictions of groundwater level using projections of future meteorological or hydrological features. This can potentially enable decision support for sustainable groundwater management in the long run.

In this study, we use backpropagation (Rumelhart et al., 1986) to train the MLP. Hyperparameters such as activation functions and the optimization method used in training the MLP are fixed (Rectified linear unit (Nair and Hinton, 2010) and RMSprop (Tieleman and Hinton, 2012), respectively). We conducted our numerical experiments with python (version 3.7) on Ubuntu 16.04 with Intel® Xeon(R) CPU E3-1245 v6 @ 3.70GHz ×8, and 31.2 GiB memory. We use the Keras package (Chollet, 2016) with the TensorFlow (Abadi et al., 2016) backend for our deep learning architectures.

DL Model Ensembles to Quantify Prediction Accuracy and Variability

Given that a DL model training involves a stochastic optimizer, we cannot infer prediction accuracy from a single DL model trial. Thus, we train the model multiple times ($N_e = 20$ trials) for the same DL model architecture and the same inputs to gain insights into the inherent prediction variability. Each trial generates a future groundwater level prediction of N_t time steps and a corresponding error between the predicted and the observed values for all time steps of the testing period. The accuracy of a trial i is quantified by the RMSE (δ_i) of the groundwater prediction (G^{pred}), which is computed in Equation (4). The average of the error (δ_i) generated across the trials gives the mean prediction error (δ) of the MLP model (Equation 5).

$$\delta_i = \sqrt{\frac{\sum_{j=k}^{N_t+k-1}\left(G_{i,j}^{pred} - G_j^{obs}\right)^2}{N_t}} \quad for\ i \in \{1, 2, 3, \ldots, N_e\} \tag{4}$$

$$\delta = \frac{1}{N_e}\sum_{i=1}^{N_e}\delta_i \tag{5}$$

where $G_{i,j}^{pred}$ is the groundwater prediction made at the j^{th} time step for the i^{th} trial, G_j^{obs} is the corresponding observed groundwater level at the j^{th} timestep. The testing period of N_t time steps starts from time step k in the dataset and runs until N_t+k-1 time step. In order to quantify the prediction variability, at each time step j, we compute the standard deviation (σ_j) of the ensemble over the N_e trials (Equation 6). We compute the standard deviation as follows:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N_e}\left(G_{i,j}^{pred} - \overline{G}_j^{pred}\right)^2}{N_e}}, \quad where\ \overline{G}_j^{pred} = \frac{1}{N_e}\sum_{i=1}^{N_e} G_{i,j}^{pred}\ and\ j = k, \ldots, k + N_k - 1 \tag{6}$$

TABLE 2 | Combinations of input features for training the DL model.

Scenario label	Input features
G-T-P-Q-4-d	Groundwater, Temperature, Precipitation, River flow, week of year
G-P-Q-4-d	Groundwater, Precipitation, River flow, week of year
G-T-P-4-d	Groundwater, Temperature, Precipitation, week of year
G-T-Q-4-d	Groundwater, Temperature, River flow, week of year
G-P-4-d	Groundwater, Precipitation, week of year
G-Q-4-d	Groundwater, River flow, week of year
G-T-4-d	Groundwater, Temperature, week of year
G-T-P-Q-2-d	Groundwater, Temperature, Precipitation, River flow, week of year
G-T-P-Q-4-m	Groundwater, Temperature, Precipitation, River flow, month of year

We assume that groundwater levels are always available during the training period. G, groundwater; T, temperature; P, precipitation; Q, river discharge; 4 indicates 4 years of training; 2 indicates 2 years of training; d, daily resolution; m, monthly resolution.

To compute the overall prediction variability (S) of a DL model architecture, the average of the standard deviations σ_j; $j = k, \ldots, k+N_t-1$ is computed as indicated in Equation (7). Lower values of S means the DL model architecture is more robust to the stochasticity in the training.

$$S = \frac{1}{N_t}\sum_{j=k}^{k+N_t-1}\sigma_j \tag{7}$$

Sensitivity Analysis of DL Model Predictions

In our numerical study we examine how different combinations of input features and the length of training time series affect the prediction accuracy of the DL model. These combinations represent potential settings in different watersheds where different amounts and types of data are collected by local agencies. This study enables us to identify input data that are necessary and sufficient for making accurate predictions of future groundwater levels. It also allows us to gain insights into "how much accuracy we lose" when certain data are not available. We examine eight different input feature scenarios for training the MLP model (**Table 2**). For example, the experiment labeled G-T-P-4-d indicates that groundwater, temperature, precipitation, and week of year are used as input features. The number 4 indicates that we used 4 years of historical observations as training data and *d* indicates a daily data resolution for validation and testing. Using data at the monthly resolution (indicated by *m*) means that the number of training data points is reduced by 97%. In this scenario, we also replace the week of year feature with the month of the year.

We designed the numerical experiments such that they address the following questions: (1) Which input features are sufficient to predict the groundwater level accurately? (2) Is there a minimum amount of data necessary to build a reasonably accurate prediction model? (3) How robust are the DL model predictions given different input feature combinations?

In order to answer these questions, we optimized and trained the MLP for each experiment shown in **Table 2**. We cannot expect the same DL model to perform well for all experiments, because the lack of certain input features potentially requires different model architecture, and if not adjusted, using too complex models may lead to data overfitting. For each experiment, we solve the bi-level optimization approach described in the section Neural Network Model and Hyperparameter Tuning to find the best model architecture. We solve the lower level problem five times to obtain an average model performance. For the optimal hyperparameter choice, we train the MLP network $N_e = 20$ times, each time generating a different MLP model. Using the resulting model ensemble, we obtain N_e replications of future groundwater predictions, which allow us to compute the statistics of the DL model performance, to quantify the prediction variability, and analyze the sensitivity of the model predictions to the input data.

NUMERICAL RESULTS

In this section we describe the results of our numerical experiments and provide a discussion of their implications on our guiding questions.

Sensitivity Analysis of Prediction Errors

We compare the future predictions obtained with our optimized and trained MLPs when using the different input feature scenarios described in **Table 2**. We make predictions for a time frame that was not used during optimization or training of the MLP (2 years unless otherwise specified), to assess the ability of the models to extrapolate beyond their training time frame.

We find that for all sites the mean prediction error ($\bar{\delta}$) ranges from 0.4 to 3.7 m (**Table 3**). Ideally, a good predictive model has low prediction errors and low variability. From the numerical results, we observe that for the Butte well, we achieve the lowest mean prediction error when training the model on G-P-4-d scenario and the lowest prediction variability in the G-T-P-Q-4-d scenario. For both Shasta and Tehama wells, we find multiple scenarios that give the same lowest values of prediction error and prediction variability.

The MLP model that is optimized and trained on all input features (Groundwater, Temperature, Precipitation and River flow) performs reasonably well across all the locations, showing similar values of normalized error of about 0.1 or 10% *of* $\Delta_{\max}$ (**Figure 3**), where $\Delta_{\max}$ is the difference between the observed maximum and minimum groundwater levels. We use this scenario (G-T-P-Q-4-d) as the base scenario in the following per-site analysis to understand the sensitivity of different input features.

TABLE 3 | MLP predictive performance at all three sites and for nine scenarios.

Scenario label	Butte		Shasta		Tehama	
	$\bar{\delta}$ (m)	S (m)	$\bar{\delta}$ (m)	S (m)	$\bar{\delta}$ (m)	S (m)
G-T-P-Q-4-d	1.1	0.5	0.4	0.3	0.4	0.2
G-P-Q-4-d	1.0	0.6	0.5	0.1	0.5	0.2
G-T-P-4-d	1.1	0.6	0.6	0.4	0.4	0.2
G-T-Q-4-d	1.3	1.0	0.4	0.3	0.4	0.2
G-P-4-d	0.8	0.6	0.6	0.1	0.5	0.4
G-Q-4-d	1.2	0.6	0.5	0.1	0.4	0.2
G-T-4-d	3.7	2.4	1.4	1.1	0.9	0.3
G-T-P-Q-2-d	1.1	0.6	0.5	0.3	0.5	0.3
G-T-P-Q-4-m	1.5	0.8	0.8	0.3	0.4	0.3

$\bar{\delta}$ indicates the mean prediction error as the difference between the model prediction and true groundwater level. S indicates the overall prediction variability (rounded to first decimal place). Low values are better. All values are computed over 20 trials as described in section DL Model Ensembles to Quantify Prediction Accuracy and Variability.

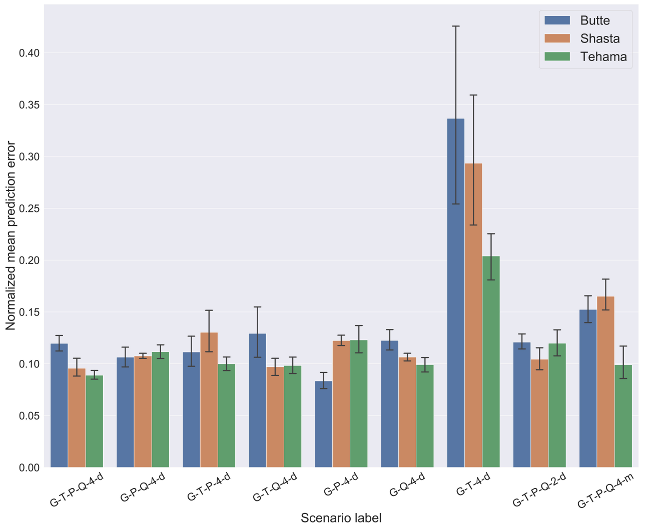


FIGURE 3 | Barplot comparing the normalized mean prediction error and their standard deviation at all well-sites for all experiments. The normalized values are obtained by dividing the error by the difference between the maximum and minimum observed groundwater levels ($\Delta_{\max}$). This normalization helps us to compare model performance across different locations. Lower bars indicate smaller prediction errors and therefore better model performance. The errors were computed by comparing the true and predicted groundwater levels in the testing dataset.

Butte Site

The Butte site is highly sensitive to the precipitation data (the prediction error increases significantly compared to the base case scenario when we remove precipitation as an input feature). In fact, the MLP model trained only on groundwater and precipitation provides the lowest prediction error. A comparison of the prediction errors of the G-P-Q-4-d and G-Q-4-d scenarios with the base scenario reveals that the precipitation events in the past are most likely to impact the future groundwater availability at this site. As the groundwater table at this site is fairly deep (16 m below ground surface), we postulate that river flow likely does not directly impact the groundwater level at the well site. Instead, given the high proportion of water use being groundwater at this site, the water table are likely driven by pumping and dependent on the amount of rainfall received over the past year.

Shasta Site

Based on simulations with different input feature scenarios, we observe that the Shasta site is most sensitive to the river flow feature. This can be seen by the error differences between the scenarios G-T-P-Q-4-d and G-T-P-4-d. Precipitation is the second most important feature. Although the river flow feature is generated using Cow Creek discharge rates, given our scaling and normalization procedure, we assume it is representative of the discharge fluctuations in the Sacramento River (which is closer to the well site). The sensitivity to river flow can be attributed to the shallow depth to groundwater (about 6 m), and short distance from the river, suggesting possible hydraulic connectivity.

Tehama Site

The MLP model trained on the base case scenario gives the lowest prediction error. We observe relatively small changes in prediction accuracy when input features such as river flow and precipitation are individually removed, showing equal input feature sensitivity. The MLP model trained only on groundwater and river flow (G-Q-4-d) also gives the same prediction performance as the base scenario. However, this is not observed when the MLP model is trained only on groundwater and precipitation (G-P-4-d), or groundwater and temperature (G-T-4-d). This indicates that the river flow carries more groundwater-relevant information, followed by precipitation in this region. This is consistent with the relatively low reliance on groundwater for water use in this region.

In all cases, the predictions are the worst when all input features except for groundwater and temperature are removed. In the following sections we only present summarized findings from the numerical experiments. The groundwater level prediction results of the individual scenarios at each of the test sites are provided in **Supplementary Section 3**.

Stochasticity in Training and Associated Prediction Variability

In order to better illustrate the stochasticity associated with the training process, we train an MLP with the same architecture but with different random number seeds. This results in a slightly different model for each run. For example, three different trials resulted in three different accuracies, with some trials yielding much more accurate outcomes than others (**Figure 4**). Therefore, we should not base decisions for groundwater management on a single trial with an MLP.

At the Butte well site the prediction variability (*S*) resulting from the stochasticity in training shows the highest sensitivity to the precipitation feature followed by the river flow feature (**Figure 5**). By comparing the different scenarios across all well sites, we find that the MLP models trained on groundwater and temperature features (G-T-4-d) have a wider spread in the predictions.

We illustrate the increasing variability of the MLP predictions when excluding necessary input features in **Figures 6**, **7**. The prediction ensemble generated with all input features (G-T-P-Q-4-d) at the Butte site is able to predict the groundwater levels

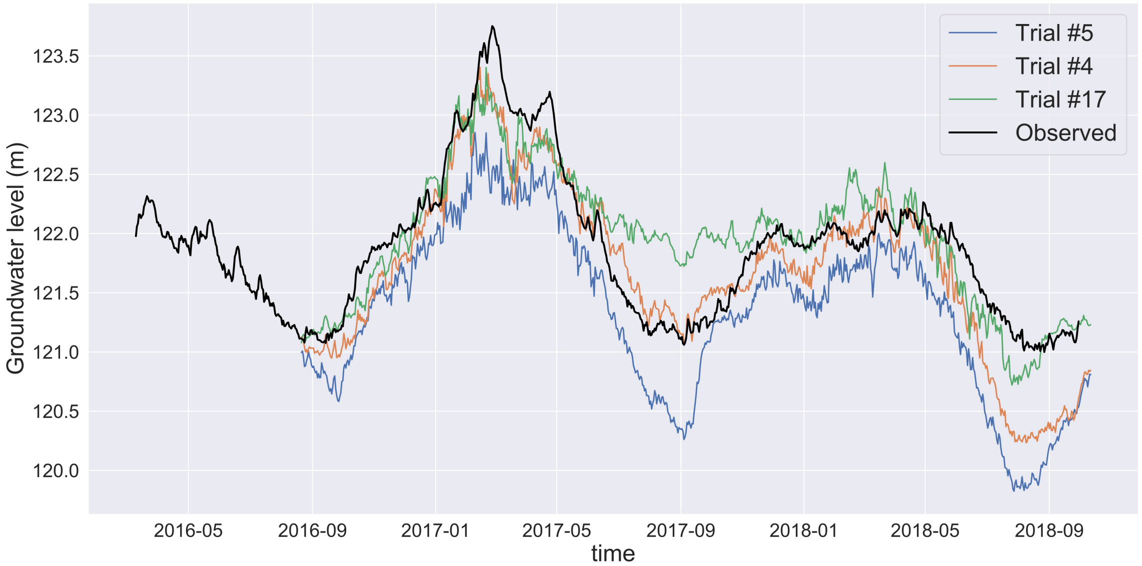


FIGURE 4 | MLP prediction for three trials with the same input features and hyperparameters at Shasta with different random seeds. The black time series shows the observed groundwater level and the other colors represent predictions from three ensemble members. The stochasticity in the training leads to different MLP models and corresponding different predictions. The prediction indicated by the Trial #5 (blue line) shows a large decrease of the groundwater level and has the highest prediction error. Trial #4 (orange line) is closest to the observed groundwater levels (truth) while Trial #17 (green line) shows a more optimistic future with less groundwater depletion.

over the 2 years with good accuracy (**Figure 6**). On the other hand, the predictions at Butte site when using only groundwater and temperature data for training the MLP have low prediction accuracy and high prediction variability (**Figure 7**). Although the model is still able to capture the seasonality of the groundwater levels, the differences between the observed and the mean of the predicted groundwater levels are large. We conducted a similar analysis for Shasta and Tehama (see **Supplementary Sections 3.1** and **3.7**). Note, however, that low prediction variability does not automatically imply high prediction accuracy, and thus both variability and prediction accuracy must be considered.

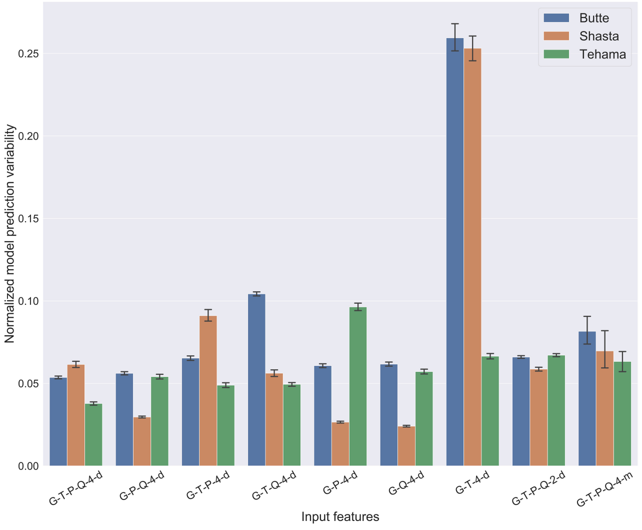


FIGURE 5 | Barplot comparing normalized model prediction variability (*S*). Lower *S* values indicate more robust MLP models that make more reliable predictions, while higher values indicate a higher variability in future predictions. The model prediction variability *S* is also normalized in the same way as the mean prediction error ($\bar{\delta}$). The standard deviation of *S*, represented by the black line on each bar indicates the variation in the ensemble prediction spread across all time steps.

Analysis of Monthly vs. Daily Training Data

Climate model data and groundwater observations are often available at a monthly temporal resolution rather than at a daily frequency. Therefore, we examined the effect of using lower-resolution data for training the MLP model, by averaging the daily values for each month. Using monthly data means that, for the same date range, the number of available training points is significantly lower: the total amount of data points is reduced by about 97% ($\approx$ 100 monthly vs. $\approx$ 2, 900 daily). The groundwater predictions at Butte trained on monthly data are much smoother and the daily groundwater drawdowns (high frequency oscillations that we observe in the daily data) are not present (**Figure 8**). The predictions show that the MLP is still able to capture the seasonality in the data (lower groundwater levels in the summer, higher levels in the winter). When compared to the corresponding daily frequency model at Butte, G-T-P-Q-4-d (**Figure 6**), we observe that the prediction errors are higher and the model does not pick up on the larger amounts of water that are available during the wet years (2017 and 2018). The prediction variability is also relatively low, indicating that for monthly predictions, the stochasticity that arises from training the models is lower, perhaps due to overfitting. At Shasta, the MLP model trained on monthly data shows a lower prediction accuracy and higher prediction variability than the base scenario. At Tehama, the MLP model built on monthly data shows similar prediction accuracy, but a higher prediction variability in comparison to the daily frequency base scenario suggesting

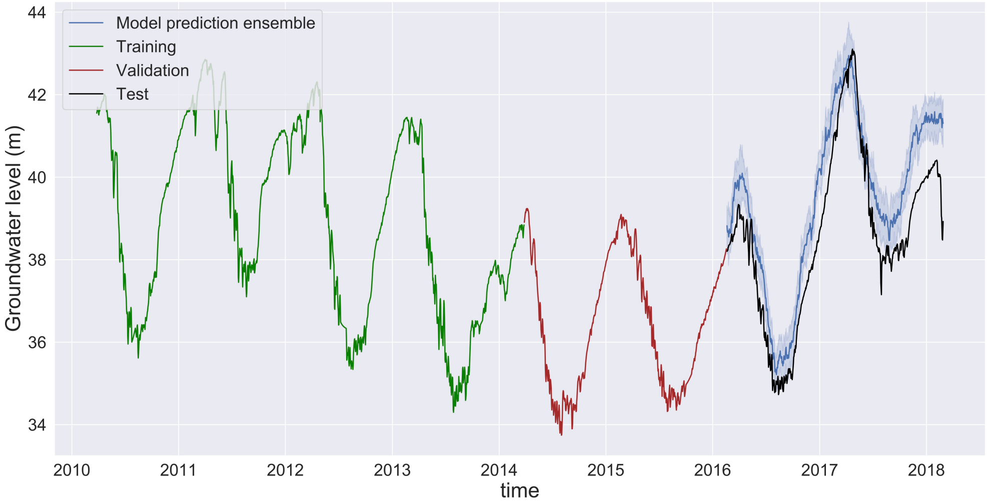


FIGURE 6 | Groundwater level prediction at Butte with input features: groundwater level, temperature, precipitation, and river flow (G-T-P-Q-4-d). The small differences between the predicted groundwater levels (ensemble mean, dark blue) and the observed levels (black) indicate a high prediction accuracy. The narrow blue band around the mean prediction indicates higher reliability in model prediction, and thus low prediction variability. Predictions made by MLP models in Shasta and Tehama are provided in **Supplementary Figures 1, 2**.

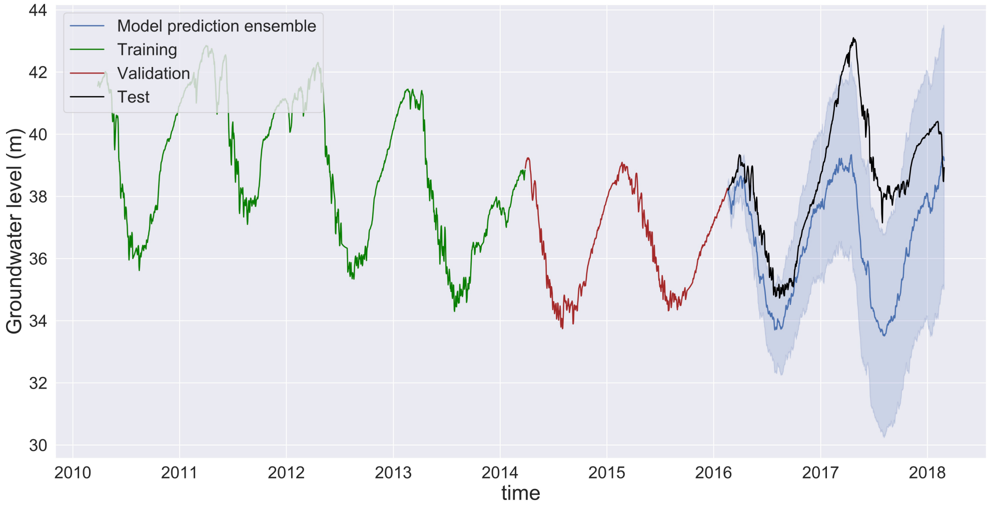


FIGURE 7 | Groundwater level prediction at Butte with input features: groundwater and temperature. Large differences between the ensemble mean (dark blue) and observed (black) show low prediction accuracy. The high variability of the predictions of individual ensemble members (blue band) shows that the MLP model is not reliable. Predictions made by MLP models in Shasta and Tehama are provided in **Supplementary Figures 18, 19**.

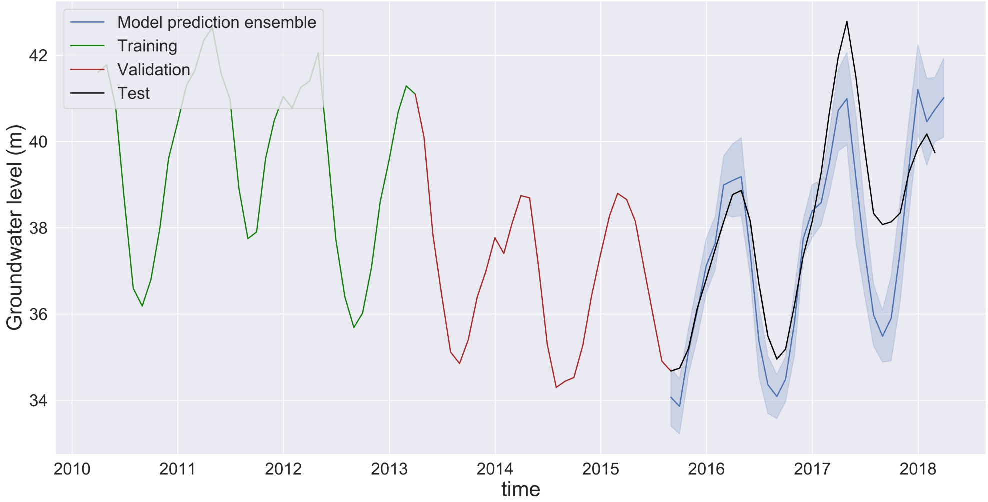


FIGURE 8 | Groundwater level prediction at Butte using monthly averaged data for all features (groundwater, temperature, precipitation, and river flow). Although the ensemble spread is narrow (the model predictions are reliable), the prediction error is high, indicating a lack of sufficiently numerous training data points.

a lack of sufficient training data to build a robust model (see **Supplementary Figure 24**).

Choice of Optimal Lag Hyperparameter

The lag hyperparameter helps the MLP model capture long-term dependencies between groundwater and other features. As mentioned previously the input to the MLP model is a lagged time series data at each timestep (section Neural Network Model and Hyperparameter Tuning). A lag number of 30 indicates that 30 days of past observations of the features are required to make the next-day groundwater level prediction. The lag parameter is a hyperparameter that is automatically optimized. **Table 4** lists the optimal lags that lead to the best groundwater level predictions. At the Butte site, we observe that most input feature combinations require a lag $>$ 300 days. When using monthly data, the optimal lag is 23 months ($\approx$2 years). At Shasta, the optimal lag values are $>$ 70 days; at Tehama, the optimal lag values are $>$ 260 days. The results indicate that the optimal lag is dependent on the specific experimental conditions and cannot be generalized to be the same across different scenarios and well sites. An incorrect lag can be detrimental to the model's predictive performance. Values of the other hyperparameters chosen are presented in **Supplementary Tables 2–4**.

Sensitivity Analysis of the Prediction Performance to the Length of the Training Time series Data

Analyzing the sensitivity of the MLP's predictive performance to the length of the training data addresses two questions. First, we will examine if the predictive performance of the MLP model is reduced by using a smaller training set. Second, by using a shorter time series for HPO and training, we can assess the accuracy of groundwater predictions for a longer time period. We experiment with using only 2 years of data for training and 2 years of validation, thus testing the MLP's prediction accuracy over 4 years. At the three sites, our MLP models are still able to predict the groundwater levels fairly accurately compared to the base scenario (**Figure 3**).

TABLE 4 | Optimal lag hyperparameter chosen in the hyperparameter optimization process at all three sites and for each scenario.

Scenario label	Butte	Shasta	Tehama
G-T-P-Q-4-d	335	70*	315*
G-P-Q-4-d	350	260	200
G-T-P-4-d	350	95	290
G-T-Q-4-d	355	100	260
G-P-4-d	335*	250	305
G-Q-4-d	355	230	285
G-T-4-d	150	45	355
G-T-P-Q-2-d	305	150	170
G-T-P-Q-4-m	23	3	21

(*) indicates the best performing input feature scenario at each site.

At the Butte site, the overall prediction accuracy is the same as the base scenario with a slightly higher prediction variability (**Figure 9**). The seasonality in the groundwater levels (less water in the summer and more in the winter) is captured well. The groundwater predictions are close to the true values for the first 1.5 years of prediction (2014–2015), but in the subsequent years the model predictions fail to accurately capture the highs and lows. The errors of the groundwater predictions accumulate over time, due to how we make next-day predictions [use the previous [lag] days of groundwater level data, and at some point, we start making predictions based on predictions and thus the errors accumulate]. At the Tehama site (see **Supplementary Figure 22**), the MLP model makes accurate predictions for the first 2 years (2014 and 2015) and subsequently we observe that the MLP predictions fail to capture the highs and the lows. This may either be related to error accumulation or a missing feature, such as snow pack or pumping data. A similar result holds for the Shasta well: the MLP is able to capture the seasonal behavior of the groundwater levels, but as we make predictions over multiple years, the prediction inaccuracies increase (see **Supplementary Figure 21**).

DISCUSSION

Future Prediction Using MLP Models

With a suitable choice of input features (e.g., G, T, P, and Q), MLP models can reliably predict groundwater levels for up to 1 year and possibly longer at a daily frequency. This is observed at all sites despite the differences in the contribution of groundwater to the county's water budget. In addition, models built exclusively with meteorological variables using temperature, precipitation and groundwater as input features (G-T-P-4-d) also show a good prediction accuracy of about 85–90%. Long-term forecasts of these meteorological variables generated from

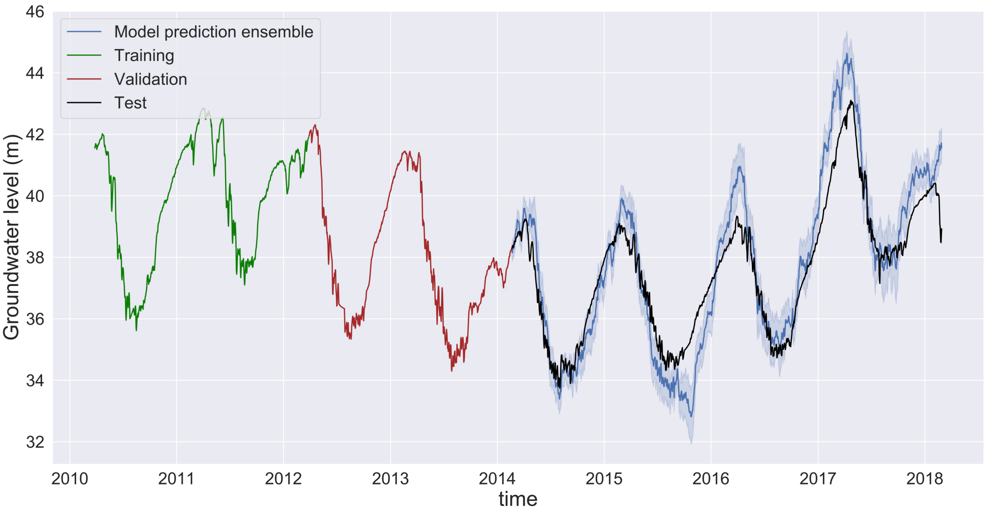


FIGURE 9 | Groundwater level prediction at Butte with input features: groundwater level, temperature, precipitation, and river flow when using only 2 years of data each for hyperparameter optimization process and training. The MLP is able to capture the seasonality of the groundwater levels, and it reflects well the groundwater levels during the drought years and the wet years.

different weather models can potentially be used to predict future groundwater levels. This can help derive sustainable groundwater management strategies.

Impact of Data Availability

A major challenge in this study was the selection of well sites and monitoring stations that adequate measurement for training, and located in near proximity. For example, at the Shasta site, we would ideally use the discharge rate of the Sacramento River rather than the Cow Creek in the MLP model. But we did not find such a monitoring station near the well site. On the other hand, it is also difficult to find groundwater wells with a long period of measurements close to river flow or weather monitoring stations. Experiments with MLP models trained on monthly averaged data and the analysis of optimal lag hyperparameter chosen at the three sites (for different scenarios) also suggest that access to a longer time range of data can help build better prediction models. This recurring issue of site selection currently makes DL techniques inapplicable in the majority of watersheds in California.

Models built from monthly frequency data show a higher prediction error than the daily frequency-based model and are unreliable for making long term (multi-year) predictions. We found that daily data were unavailable for most of the sites in California. In fact, out of the 3,907 monitoring wells in the state, only 387 had daily measurements through California statewide groundwater elevation monitoring (CASGEM) network, and most of the high-resolution datasets were only available for wells in northern California, in mostly low-priority basins. Prediction accuracies can be improved with access to higher-resolution daily data, or longer monthly datasets (spanning decades). Additionally, our current analysis is performed in the absence of pumping data, which is not publicly available. Yet pumping is a critical component of groundwater budget, and in several places the primary driver of groundwater table depths. Access to such data can potentially better equip our current DL models with human behavior and improve management strategies. The potential advantage of using additional data for obtaining more accurate predictions may lead to investments into more *in-situ* or remote measurement infrastructure. Based on our current results, we recommend using more than 2 years of daily data for training.

Impact of Training Stochasticity on Prediction Results Matters

In addition to the prediction accuracy, we find that it is also important to measure the prediction variability of the MLP, which is due to stochasticity in the training process. The Keras tool used in the study generated different weight optimized MLP models for the same set of hyperparameters and training data. We cannot analyze future predictions or derive water management strategies based on a single training trial. We recommend training a DL model of a given architecture multiple times, as the stochasticity of the optimizer used during the training leads to multiple prediction models that are consistent with the training data. The resulting model ensembles allow us to assess the model's prediction reliability. Thus, in addition to potential uncertainty in the data collected we also need to take into account the variability in the training process. Our study showed that models trained on groundwater, temperature, precipitation, and temperature data (G-T-P-Q-4-d) yield the lowest prediction variability, whereas models trained only on groundwater and temperature data have the highest prediction variability. Note however, that low variability does not necessarily mean high prediction accuracy, and thus both metrics need to be taken into account when assessing the quality of the DL model predictions. In a future study, one can tackle this problem from a bi-objective perspective in which the prediction accuracy is maximized and the variability is minimized simultaneously.

Automated HPO Framework for Future DL Applications

A key innovation in this study is the use of an HPO framework to test different model architecture for making prediction models. The setup of our study, and the HPO is general enough to be applicable to any other type of neural network (e.g., CNN and LSTM). The sensitivity analysis requires conducting multiple experiments testing different input feature combinations and our results indicate that each experiment requires a different combination of hyperparameters. Hand tuning the model architectures for each experiment can be a cumbersome process especially when the number of features is large. The HPO framework used in this study automates this process and ensures the best model architecture (within the given bounds). We can also potentially incorporate the choice of input feature into the framework as a decision variable. The HPO formulation will then choose the best combination of input features and its best architecture simultaneously.

Multi-Well MLP Models

The current analysis has been conducted for single groundwater well sites only, which does not reflect the overall health of a groundwater aquifer. Thus, a spatially distributed parameter sensitivity analysis across multiple groundwater well sites and climatic parameters may reflect a more realistic behavior of a groundwater aquifer and human use. Our previous study (Müller et al., 2020) successfully built DL models to simultaneously predict daily groundwater level at three locations in Butte county. However, we saw that when we use an average prediction error metric to measure the prediction performance across the three wells, only two wells have accurate predictions. Thus, one remedy could be a reformulation of the objective function by introducing weights that reflect the importance of each well to ensure optimal prediction performance across all wells. Although training can be compute-intensive, once trained and optimized, DL models are a more viable option for performing multi-scenario analyses than high-fidelity simulation models, because the required computational time to make future predictions is orders of magnitude lower. Our multi-scenario analysis can readily be used by groundwater managers who have access to historical groundwater and local weather data.

CONCLUSION

With the increased deployment of ML tools in hydrological sciences, there is a need to understand the sensitivity of their prediction performance to different input features. Groundwater level timeseries are highly non-linear and non-stationary, making them difficult to model with standard ARIMA models. DL models offer a promising alternative for capturing the complex interactions between features such as groundwater levels, river flow, temperature, and precipitation.

In our study, we were able to accurately predict groundwater levels at three different groundwater well locations (Butte, Shasta, and Tehama) in California using an MLP model. Additionally, we conducted a sensitivity analysis using multiple different feature combination scenarios and compared the accuracy and reliability of the resulting predictions. Our analysis shows that models trained on groundwater, temperature, river flow and precipitation data (G-T-P-Q-4-d) lead to the best predictive performance at two of the three sites, while models trained without hydrological features and based only on past groundwater and temperature data consistently showed the lowest prediction accuracy at all locations. The best predictive models are shown to reliably predict groundwater levels at least 1 year into the future. The MLP prediction performance is also affected by the data's temporal resolution and the length of the training period. The MLP models trained with only 2 years (rather than four) of data still gave reasonable accuracy and indicate the potential capability for long-term predictions. In addition to accuracy, we find that it is also important to measure the prediction variability caused by the stochasticity in the training process. The MLP model architectures for different choices of input features, training length and temporal frequency which were obtained using a hyperparameter optimization framework indicate that the optimal combination is location-specific. These results indicate that DL models are a good choice for modeling groundwater levels, contingent on the availability of adequately long time-series of prior groundwater levels and some hydrological variables (precipitation or river flow at the minimum).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

RS, JM, CV, BA, and BF conceived the presented idea. RS and JP carried out the numerical experiment. RS wrote the manuscript with support from other authors. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

FUNDING

This work was supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2020.573034/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]* arXiv:1603.04467.

Adamowski, J., and Chan, H. F. (2011). A wavelet neural network conjunction model for groundwater level forecasting. *J. Hydrol.* 407, 28–40. doi: 10.1016/j.jhydrol.2011.06.013

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., and Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res*. 48. doi: 10.1029/2010WR009945

Aeschbach-Hertig, W., and Gleeson, T. (2012). Regional strategies for the accelerating global problem of groundwater depletion. *Nat. Geosci.* 5, 853–861. doi: 10.1038/ngeo1617

Ahmad, S., Kalra, A., and Stephen, H. (2010). Estimating soil moisture using remote sensing data: a machine learning approach. *Adv. Water Resour*. 33, 69–80. doi: 10.1016/j.advwatres.2009.10.008

Amari, S. I. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing* 5, 185–196. doi: 10.1016/0925-2312(93)90006-O

Arora, B., Mohanty, B. P., and McGuire, J. T. (2011). Inverse estimation of parameters for multidomain flow models in soil columns with different macropore densities. *Water Resour. Res.* 47:2010WR009451. doi: 10.1029/2010WR009451

Arora, B., Mohanty, B. P., and McGuire, J. T. (2012). Uncertainty in dual permeability model parameters for structured soils. *Water Resour. Res.* 48:W01524. doi: 10.1029/2011WR010500

Ayres, J., and Brown, C. (2008). *Tehama County AB-3030 Groundwater Management Plan*. Retrieved from: http://www.tehamacountypublicworks.ca.gov/flood/groundwater/bowman.pdf (accessed April 20, 2020).

Banerjee, P., Singh, V. S., Chatttopadhyay, K., Chandra, P. C., and Singh, B. (2011). Artificial neural network model as a potential alternative for groundwater salinity forecasting. *J. Hydrol.* 398, 212–220. doi: 10.1016/j.jhydrol.2010.12.016

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Machine Learn. Res.* 13, 281–305. Available online at: https://dl.acm.org/doi/abs/10.5555/2188385.2188395

Bureau of Reclamation, Mid Pacific Region Sacramento California US department of Interior, and Anderson-Cottonwood Irrigation District (2011). *Anderson-Cottonwood Irrigation DistrictIntegrated Regional Water Management Program – Groundwater Production Element Project*. Available online at: https://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/california_waterfix/exhibits/docs/CSPA%20et%20al/aqua_60.pdf (accessed April 20, 2020).

Butte County Department of Water and Resource Conservation. (2016). *Butte County Water Inventory and Analysis*. Available online at: https://www.buttecounty.net/wrcdocs/Reports/I%26A/2016WI%26AFINAL.pdf (accessed April 20, 2020).

California Department of Water Resources (2004). Anderson Subbasin Hydrogeology. Retrieved from: https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/Groundwater-Management/Bulletin-118/Files/2003-Basin-Descriptions/5_006_03_AndersonSubbasin.pdf (accessed April 20, 2020).

Chollet, F. (2016). *Keras Deep Learning Library. Code*:Available online at: https://github.com/fchollet (accessed April 20, 2020). Documentation: http://keras.io (accessed April 20, 2020).

Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. *Appl. Soft Comput.* 19, 372–386. doi: 10.1016/j.asoc.2014.02.002

Deo, R. C., and Sahin, M. (2016). An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. *Environ. Monitor. Assess.* 188:90. doi: 10.1007/s10661-016-5094-9

Dieter, C. A., Maupin, M. A., Caldwell, R. R., Harris, M. A., Ivahnenko, T. I., Lovelace, J. K., et al. (2018). *Estimated Use of Water in the United State in 2015*. Reston, VA: U.S. Geological Survey Circular. doi: 10.3133/cir1441

DWR, California Department of Water Resources. (2004). *Vina Subbasin Hydrogeology*. Retrieved from: https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/Groundwater-Management/Bulletin-118/Files/2003-Basin-Descriptions/5_021_57_VinaSubbasin.pdf (accessed April 20, 2020).

DWR, California Department of Water Resources. (2020). *Basin Prioritization*. Retrieved from: https://water.ca.gov/Programs/Groundwater-Management/Basin-Prioritization (accessed April 20, 2020).

Emamgholizadeh, S., Moslemi, K., and Karami, G. (2014). Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resour. Manag.* 28, 5433–5446. doi: 10.1007/s11269-014-0810-0

Gaur, S., Ch, S., Graillot, D., Chahar, B. R., and Kumar, D. N. (2013). Application of artificial neural networks and particle swarm optimization for the management of groundwater resources. *Water Resour. Manag.* 27, 927–941. doi: 10.1007/s11269-012-0226-7

Ghiassi, M., Zimbra, D. K., and Saidane, H. (2008). Urban water demand forecasting with a dynamic artificial neural network model. *J. Water Resour. Plann. Manag.* 134, 138–146. doi: 10.1061/(ASCE)0733-9496(2008)134:2(138)

Guzman, S. M., Paz, J. O., and Tagert, M. L. M. (2017). The use of NARX neural networks to forecast daily groundwater levels. *Water Resour. Manag.* 31, 1591–1603. doi: 10.1007/s11269-017-1598-5

Herrera, M., Torgo, L., Izquierdo, J., and Pérez-Garćia, R. (2010). Predictive models for forecasting hourly urban water demand. *J. Hydrol.* 387, 141–150. doi: 10.1016/j.jhydrol.2010.04.005

Kanellopoulos, I., and Wilkinson, G. G. (1997). Strategies and best practice for neural network image classification. *Int. J. Remote Sens.* 18, 711–725. doi: 10.1080/014311697218719

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022. doi: 10.5194/hess-22-6005-2018

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*. Reston, VA: U.S. Geological Survey. doi: 10.3133/tm6A55

Lin, J. Y., Cheng, C. T., and Chau, K. W. (2006). Using support vector machines for long-term discharge prediction. *Hydrol. Sci. J.* 51, 599–612. doi: 10.1623/hysj.51.4.599

Mohanty, S., Jha, M. K., Raul, S. K., Panda, R. K., and Sudheer, K. P. (2015). Using artificial neural network approach for simultaneous forecasting of weekly groundwater levels at multiple sites. *Water Resour. Manag.* 29, 5521–5532. doi: 10.1007/s11269-015-1132-6

Moosavi, V., Vafakhah, M., Shirmohammadi, B., and Behnia, N. (2013). A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. *Water Resour. Manag.* 27, 1301–1321. doi: 10.1007/s11269-012-0239-2

Moritz, S., and Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. *R J.* 9, 207–218. doi: 10.32614/RJ-2017-009

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B., et al. (2020). Surrogate optimization of deep neural networks for groundwater predictions. *J. Glob. Optim*. doi: 10.1007/s10898-020-00912-0

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa. 807–814.

Najah, A., El-Shafie, A., Karim, O. A., and El-Shafie, A. H. (2013). Application of artificial neural networks for water quality prediction. *Neural. Comput. Appl.* 22, 187–201. doi: 10.1007/s00521-012-0940-3

Rasouli, K., Hsieh, W. W., and Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.* 10:198. doi: 10.1016/j.jhydrol.2011.10.039

Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., et al. (2016). Sensors in the stream: the high-frequency wave of the present. *Environ. Sci. Technol.* 50, 10297–10307. doi: 10.1021/acs.est.6b02155

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sahoo, S., Russo, T. A., Elliott, J., and Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resour. Res.* 53, 3878–3895. doi: 10.1002/2016WR019933

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

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

Steefel, C. I., Appelo, C. A. J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., et al. (2015). Reactive transport codes for subsurface environmental simulation. *Computat. Geosci*. 19, 445–478. doi: 10.1007/s10596-014-9443-x

Taormina, R., Chau, K., and Sethi, R. (2012). Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Eng. Appl. Artif. Intell.* 25, 1670–1676. doi: 10.1016/j.engappai.2012.02.009

Tieleman, T., and Hinton, G. (2012). Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* 4, 26–31.

Tiwari, M. K., and Adamowski, J. (2013). Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. *Water Resour. Res.* 49, 6486–6507. doi: 10.1002/wrcr.20517

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoustics Speech Signal Process.* 37, 328–339. doi: 10.1109/29.21701

Xu, L., and Liu, S. (2013). Study of short-term water quality prediction model based on wavelet neural network. *Math. Comput. Modell.* 58, 807–813. doi: 10.1016/j.mcm.2012.12.023

Xu, T., Spycher, N., Sonnenthal, E., Zhang, G., Zheng, L., and Pruess, K. (2011). {TOUGHREACT} Version 2.0: A simulator for subsurface reactive transport under non-isothermal multiphase flow conditions. *Comput. Geosci.* 37, 763–774. doi: 10.1016/j.cageo.2010.10.007

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50, 159–175. doi: 10.1016/S0925-2312(01)00702-0

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
in Water

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/frwa.2020.536743

Downscaling Satellite and Reanalysis Precipitation Products Using Attention-Based Deep Convolutional Neural Nets

Alexander Y. Sun [1]* *and Guoqiang Tang* [2,3]

[1] Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, United States, [2] Coldwater Laboratory, University of Saskatchewan, Canmore, AB, Canada, [3] Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK, Canada

High-quality and high-resolution precipitation products are critically important to many hydrological applications. Advances in satellite remote sensing instruments and data retrieval algorithms continue to improve the quality of the operational precipitation products. However, most satellite products existing today are still too coarse to be ingested for local water management and planning purposes. Recent advances in deep learning algorithms enable the fusion of multi-source, high-dimensional data for statistical learning. In this study, we investigated the efficacy of an attention-based, deep convolutional neural network (AU-Net) for learning spatial and temporal mappings from coarse-resolution to fine-resolution precipitation products. The skills of AU-Net models, developed using combinations of static and dynamic predictors, were evaluated over a 3 × 3° study area in Central Texas, U.S., a region known for its complex precipitation patterns and low predictability. Three coarse-resolution satellite/reanalysis precipitation products, ERA5-Land (0.1°), TRMM (0.25°), and IMERG (0.1°), are used as part of the inputs, while the predictand is the 1-km PRISM data. Auxiliary predictors include elevation, vegetation index, and air temperature. The study period includes 18 years of data (2001–2018) at the monthly scale for training, validation, and testing. Results show that the trained AU-Net models achieve different degrees of success in downscaling the baseline coarse-resolution products, depending on the total precipitation, the accuracy of large-scale patterns captured by the baseline products, and the amount of information transferable from predictors. Higher precipitation rate tends to affect AU-Net model performance negatively. Use of the attention mechanism in the AU-Net models allows for infilling of multiscale features and generation of sharper images. Correction using gauge data, if there is any, can further improve the results significantly.

Keywords: PRISM, TRMM, deep learning, convolutional neural net, global precipitation measurement (GPM) satellite, precipitation downscaling, attention-based U-net

OPEN ACCESS

Edited by:
Xingyuan Chen,
Pacific Northwest National Laboratory (DOE), United States

Reviewed by:
Tiantian Yang,
University of Oklahoma, United States
Hui Lu,
Tsinghua University, China

****Correspondence:***
Alexander Y. Sun
alex.sun@beg.utexas.edu

Specialty section:
This article was submitted to Water and Hydrocomplexity, a section of the journal Frontiers in Water

Received: *20 February 2020*
Accepted: *02 November 2020*
Published: *26 November 2020*

Citation:
Sun AY and Tang G (2020) Downscaling Satellite and Reanalysis Precipitation Products Using Attention-Based Deep Convolutional Neural Nets. Front. Water 2:536743. doi: 10.3389/frwa.2020.536743

1. INTRODUCTION

Precipitation is a primary driver of water and energy cycle (Trenberth et al., 2007), providing essential inputs to many water, food, and energy applications including, but not limited to, global and regional climate variability assessments, land surface-atmosphere interactions, natural hazard prevention, crop yield management, hydrological forecasting, and surface and groundwater resources planning (Hong et al., 2007; Seneviratne et al., 2010; Becker et al., 2013; Schewe et al., 2014). To a large degree, the effectiveness of many disaster response and water resources management decisions hinge on the quantity and quality, as well as the spatial and temporal resolution of precipitation products. Currently available precipitation products may be classified into ground-based, satellite-based, reanalysis, and hybrid multi-source/multi-sensor products.

Ground-based products are derived from rain gauges and weather radar. However, the spatial coverage of rain gauge networks is often limited, also varying significantly across different countries owing to temporal sampling resolutions, periods of operation, data latency, and data access (Kidd et al., 2017). At the global scale, ground-based products are only available at a relatively coarse resolution (≥0.5°) and updated rather infrequently (Sun Q. et al., 2018). High-resolution gridded products are only available in a few developed counties that have extensive gauge network coverage. For example, in the U.S., the Parameter-elevation Regressions on Independent Slopes Model (PRISM) gauge-based product (4-km resolution, 1895–present), developed by the Oregon State University (Daly et al., 1997), is widely used for operational planning and validation of satellite products. Similarly, the Stage IV radar-based, gauge-adjusted precipitation data (4 km, 2002–present) available from the National Center for Environmental Prediction (NCEP) is also commonly used as a reference dataset in many conterminous U.S. (CONUS) precipitation product comparisons (Lin and Mitchell, 2005).

Satellite precipitation products are derived from passive and active microwave (MW) sensors onboard low Earth orbiting satellites, and visible/infrared (VIS/IR) sensors onboard geostationary satellites (Hou et al., 2014). So far, the raw satellite precipitation data has been mainly retrieved from three spaceborne precipitation radars: the Ku-band precipitation radar onboard the Tropical Rainfall Measuring Mission (TRMM) satellite that was in orbit from 1997 to 2015, the W-band Cloud Profiling Radar (CPR) onboard the CloudSat operating from 2006 to the present, and the Dual-frequency Precipitation Radar (DPR) onboard the Global Precipitation Measurement (GPM) Core Observatory operating from 2014 to the present (Tang et al., 2018b). Unlike ground-based products, satellite products provide spatially homogeneous coverage with low latency. Some of the currently available satellite products, such as the Integrated Multi-satellite Retrievals for GPM (IMERG) (Huffman et al., 2015) and TRMM Multi-satellite Precipitation Analysis (Huffman et al., 2007), not only assimilate information from multiple MW/IR sensors, but also are corrected by ground observations. Currently, the most common resolution of satellite precipitation products is 0.25° per 3 h (Sun Q. et al., 2018).

Reanalysis products are generated by assimilating irregular observations into earth system models to generate a synthesized estimate of the state of the system (e.g., precipitation) across a uniform model grid, with spatial homogeneity and temporal continuity (Sun Q. et al., 2018). The commonly used reanalysis products include the NCEP/NCAR Reanalysis system (1.875°, 1979–2010) (Kistler et al., 2001), European Center for Medium-Range Weather Forecasts (ECMWF) reanalysis systems (0.25/0.75°, 1979–present) (Dee et al., 2011), and the NCEP Climate Forest System Reanalysis system (CFSR, 38 km, 1979–2010) (Saha et al., 2010).

Recent trends in precipitation product development are geared toward merging multi-source and multi-sensor data to leverage information existing at multiple scales. Examples include the Multi-Source Weighted-Ensemble Precipitation (MSWEP, 0.1/0.5°, 1979–present) (Beck et al., 2017) and Modern-Era Retrospective Analysis for Research and Application system (MERRA-2) (Rienecker et al., 2011), both combining gauge, satellite, and reanalysis data. These products typically adopt an optimal weighting scheme to merge information. In MSWEP, for example, weights assigned to the gauge-based data are determined from the gauge network density, while weights assigned to the satellite and reanalysis-based estimates are calculated from their comparative performance at the surrounding gauges (Beck et al., 2017).

Notwithstanding the tremendous effort dedicated to developing various products, precipitation forcing remains a major source of uncertainty in global hydrological and land surface models (Wood et al., 2011; Scanlon et al., 2018) because of its inherent high variability in space and time, especially in topographically complex, convection-dominated, and snow-dominated regions (Tang et al., 2018a; Beck et al., 2019). The accuracy of rain gauge data may be affected by a number of environmental factors, such as wind, wetting and evaporation loss, and undercatch (Sun Q. et al., 2018; Tang et al., 2018a). The uncertainty in satellite precipitation data may stem from different sources, including algorithms used for retrieving, downscaling, and merging multi-sensor data, as well as from the acquisition instrument itself (Sorooshian et al., 2011).

Recently, Sun Q. et al. (2018) reviewed 30 currently available global precipitation datasets, including gauge-based, satellite-related, and reanalysis datasets. They found that the magnitude of annual precipitation estimates over global land deviated by as much as 300 mm/yr among the products. They also noted that the degree of variability in precipitation estimates varied by region, with large differences found over tropical oceans, complex mountain areas, northern Africa, and some high-latitude regions. Beck et al. (2019) evaluated the performance of 26 gridded daily precipitation products over the CONUS for the period 2008–2017. Among the 15 uncorrected datasets considered, they found that ERA5-HRES (the 5th global reanalysis product released by ECMWF, 0.28°, 2008–present) gives better performance than others across most of CONUS, especially in the west; among the 11 gauge-corrected products, MSWEP V2.2 gives the best performance, which was attributed to applying daily gauge corrections and accounting for gauge reporting times during product development. Both product reviews suggest that the reliability of precipitation datasets depends on the number and spatial coverage of surface stations, the accuracy of satellite data retrieval algorithms, as well as the data assimilation models used. Most data assimilation and bias correction methods, in turn, rely on the understanding and characterization of precipitation error distributions, which are typically non-stationary and product dependent. AghaKouchak et al. (2012) investigated the systematic and random errors in several major satellite precipitation products against the NCEP Stage IV data. A major finding of their study is that the spatial distribution of the systematic error had similar patterns for all precipitation products they considered, for which the error is remarkably higher during the winter than in summer; the error was also found to be proportional to rain rates, with larger errors tending to be associated with higher rain rates. Parameterization of the precipitation error model is thus critically important for improving precipitation products, but remains a challenging task, partly because of the strong spatial and temporal variability in rainfall patterns (Sorooshian et al., 2011; AghaKouchak et al., 2012).

The advent of deep learning (DL) algorithms in recent years has revolutionized the field of statistical pattern recognition, enabling machines to achieve human-like classification accuracy (Goodfellow et al., 2016). Precipitation product development represents a research domain that can readily benefit from the DL because of the explosive growth of multiscale, multi-source Earth observation data (Ma et al., 2015; Sun and Scanlon, 2019). Pan et al. (2019) recently presented a convolutional neural network (CNN) method for precipitation estimation using numerical weather model outputs. The CNN model architecture follows an end-to-end design, in which a fully connected dense layer is used at the output layer to recover the dimensions of the input images. The input predictors they used include 3-h geopotential height and precipitable water at 500, 850, and 1,000 hPa, which were taken from the NCEP regional reanalysis at 32 km (~0.29°) resolution; and the predictand is the total precipitation. Their results show CNN obtained better skills in the northwest and east parts of CONUS, but performed poorer than the reference Climate Prediction Center (CPC) gauge-based dataset in the mid-U.S. Tang et al. (2018b) applied a four-layer, deep multilayer perceptron network to predict precipitation rates (at single locations), by mapping passive microwave data from GPM and MODerate resolution Imaging Spectroradiometer (MODIS) to spaceborne radar data. Kim et al. (2017) used ConvLSTM, a combination of convolutional neural nets and long short-term memory (LSTM) neural net (Shi et al., 2015), for precipitation nowcasting using weather radar data. Their results showed ConvLSTM was able to obtain better results than the simple linear regression method. Similarly, ConvLSTM was recently used for precipitation estimation based on atmospheric dynamical fields simulated by ERA-Interim (a predecessor of ERA5) (Miao et al., 2019).

Tremendous interests exist in using machine learning techniques for statistical precipitation downscaling, which has long been studied even before the DL era to refine coarse-resolution precipitation products and global climate model projections for local water management and hydrological modeling needs (Maraun et al., 2010; Jia et al., 2011; Duan and Bastiaanssen, 2013; Chen et al., 2018). To correct biases arising during downscaling, two types of traditional methods may be identified, quantile mapping (Li et al., 2010; Shen et al., 2014; Yang et al., 2016) and multiplicative/additive correction (or linear scaling) (Vila et al., 2009; Jakob Themeßl et al., 2011). In the realm of DL, Vandal et al. (2018) introduced DeepSD, which is a stacked superresolution CNN for statistical downscaling of climate and Earth system model simulations. In their experiments, Vandal et al. (2018) upsampled the 4-km PRISM data progressively to 1°, and then tried to restore the original high-resolution data by training a CNN model. He et al. (2016) used the random forest algorithm to downscale the precipitation forcing field used in North-American Land Data Assimilation System Project Phase 2 (NLDAS-2). Their main research question was whether the upsampled NLDAS-2 precipitation forcing (in spatial resolutions of 0.25, 0.5, and 1°) could be restored to its native resolution (0.125°) by using additional dynamic and static information (e.g., air temperature, wind speed, elevation, slope) as auxiliary inputs.

So far, however, few studies have attempted to directly map coarse-resolution precipitation products (e.g., satellite or reanalysis products) to fine-resolution, gauge-based precipitation products using DL. As mentioned previously, gauge products tend to have higher resolutions but are often created using proprietary data processing algorithms that may not be readily accessible to local users. Inconsistencies in data release times may also prevent end users from accessing the information when they need it the most. The main motivation of this research was thus to investigate a data-driven, DL-based statistical downscaling procedure by learning covariational patterns between the coarse- and fine-resolution precipitation products. A novel, attention-based, deep convolutional neural network model was adopted to help capture multiscale spatial and temporal patterns. Once trained, end users may apply the DL-based model to generate downscaled high-resolution precipitation maps using only coarse-resolution products, which are available operationally. Ultimately, such a DL-based downscaling procedure may be applied to regions without high resolution products through transfer learning, in which models trained for data-rich domains are "transferred" to inform models for data-sparse domains (Pan and Yang, 2009; Goodfellow et al., 2016; Jean et al., 2016; Sun and Scanlon, 2019). Like in all regression studies, a main hypothesis underneath this research is that certain spatial and temporal covariational patterns exist between the predictand and its predictors, which has been confirmed to a certain degree by previous validation studies (Beck et al., 2017), but is also shown to vary significantly across space and time (AghaKouchak et al., 2012), creating a major challenge for the pattern-based learning algorithms.

For demonstration, we focus on Central Texas, which is a region of low hydrometeorological predictability (AghaKouchak et al., 2012; Sun et al., 2014; Beck et al., 2019; Pan et al., 2019), and yet, is frequented by flooding and drought events (Lowrey and Yang, 2008; Long et al., 2013; Sun A. Y. et al., 2018). PRISM data is used as the high-resolution training target. The performance of three coarse-resolution satellite and reanalysis products, along with other auxiliary variables, are evaluated. This paper is organized as follows. Section 2 describes the study area and datasets used. Section 3 presents the design of the deep CNN model. Results are provided in section 4, followed by discussion and conclusions. For reference, a table of major abbreviations and acronyms used this paper is provided in the **Appendix**.

2. STUDY AREA AND DATA USED

2.1. Study Area

Central Texas represents the fastest-growing region in the U.S. among metros with at least 1 million people (Austin Statesman, 2019). The region is also known for its severe precipitation events, resulting from a juxtaposition of meteorological factors, including moisture influx from the Gulf of Mexico, easterly wave moving across the area, and orographic uplift from the Balcones Escarpment (a physiographic feature of steep elevation gradient at the boundary between the Edwards Plateau and the Gulf Coast Plain) (Hirschboeck, 1987; Nielsen-Gammon et al., 2005; Lowrey and Yang, 2008; Sun A. Y. et al., 2018).

The area of study is a 3 × 3° region bounded between latitudes 29–32°N and longitudes 100–97°W (**Figure 1**). It encompasses two major Central Texas cities, Austin and San Antonio, as well as their surrounding regions. Central Texas is part of the Texas Hill Country, which is within the Edwards Plateau, a geographic region known by its rugged karstic terrains and thin top soils (Mace et al., 2000). Major land cover types include forest lands, rangeland, agricultural lands, urban, barren land, and wetlands (Omranian and Sharif, 2018). Elevation is highest (736 m) near the west boundary of the study area and gradually decreases toward the east boundary to 42 m (**Figure 1A**). Climate in the region is humid subtropical, and precipitation exhibits a distinctive bimodal pattern: spring is the wettest season, with April and May the wettest months; a secondary peak of rainfall occurs in September and October (Slade and Patton, 2003). Spatially, the annual rainfall in the 3 × 3° region ranges from 575 to 1,005 mm, which is the highest in the east and decreases toward the west (**Figure 1B**). Tropical cyclones (hurricanes and tropic storms) typically occur in late summer or early fall, bringing the largest amount of rainfall. Moreover, Balcones Escarpment acts as a major mechanism of localization and intensification of rainfall (Nielsen-Gammon et al., 2005).

Hydrology wise, the study area is part of two major river basins, the Lower Colorado River Basin that drains to Lower Colorado River and its major tributaries (San Saba River, Llano River, and Pedernales River), and the Brazos River Basin. The former includes a cascade of surface reservoirs (e.g., Lake Buchanan, Lake Travis) that provide surface water supply to the City of Austin. In addition to flooding, severe drought is a major concern, often causing significant loss to the regional economy (Long et al., 2013). The accuracy and reliability of precipitation estimate is thus of paramount importance to local water agencies, for continuously evaluating flood/drought potential, as well as for quantifying groundwater recharge, reservoir storage, and water availability. For those reasons, the Lower Colorado River Authority (LCRA), the primary water management agency of the area, has established a dense gauge network in recent years to provide continuous rainfall data at relatively high spatial and temporal resolutions (open circles in **Figure 1A**). The *in situ* data offers important additional information for precipitation downscaling in this study, as discussed below in section 4.

2.2. Datasets

The study period is from Jan 2001 to Dec 2018, which was chosen based on the common period of coverage of all products considered. The monthly scale was chosen because of our interest in downscaling precipitation for supporting subseasonal water management activities. In the following, the gridded and gauge data used are described in details, and a summary of all data used is also provided in **Table 1**, including the data URLs.

2.2.1. Gauge Data

Monthly gauge precipitation data was obtained from Texas Mesonet. **Figure 1A** shows the locations of rain gauges as of Jan 2017 (the first month of our test period), which are distributed more densely within the LCRA boundary than in the surrounding areas. The number of valid gauges increased from 361 in Jan 2017 to 532 in Dec 2018. As mentioned before, the number of rain gauges only increased in recent years in light of the severe 2012–2013 Texas drought. Before that event, the number of *in situ* data was generally much smaller. For example, the number of gauge data available in Jan 2003 was 24 and in Jan 2013 it was 53. Thus, the quality of the PRISM data evolved with time. In other words, patterns used for the training period may be less constrained than the patterns used during validation.

In this study, the gridded, gauge-based precipitation product PRISM is the training target or predictand. The Stage IV data from NCEP was used for cross-examining the PRISM patterns. Stage IV data includes merged operational radar data and rain gauge measurements in hourly accumulations. Both datasets have a spatial resolution of 4 km and were temporally aggregated to the monthly scale.

2.2.2. Satellite and Reanalysis Data

Three coarse-resolution satellite and reanalysis precipitation products, TRMM, ERA5-Land, and IMERG, were tested for generating PRISM like data. The TRMM data used in this study is 3B43 v7 (0.25°, 1998–2019), which is a post-real-time, gauge-corrected monthly product that merges precipitation estimates from multi-sensors, as well as monthly precipitation gauge analysis from the Global Precipitation Climatology Center (https://www.dwd.de) (see also **Table 1**). The main motivation behind developing the 3B43 algorithm was to produce the best estimate of precipitation rate from sensors onboard TRMM, as well as from other satellites including Advanced Microwave Scanning Radiometer for Earth Observing Systems (AMSR-E), Special Sensor Microwave Imager (SSMI), Special Sensor Microwave Imager/Sounder (SSMIS), Advanced Microwave Sounding Unit (AMSU), Microwave Humidity Sounder (MHS), and microwave-adjusted merged geo-infrared (IR) (Huffman et al., 2010). After TRMM was decommissioned in 2015, TRMM data continued to be produced using the climatological calibrations/adjustments until 2019 (Bolvin and Huffman, 2015).

ERA5 is the latest generation of reanalysis data from ECMWF. It is produced using the 4D-variational data assimilation system in ECMWF's Integrated Forecast System, and features several improvements over its predecessor (i.e., ERA-Interim), including an updated model and data assimilation system, higher spatial

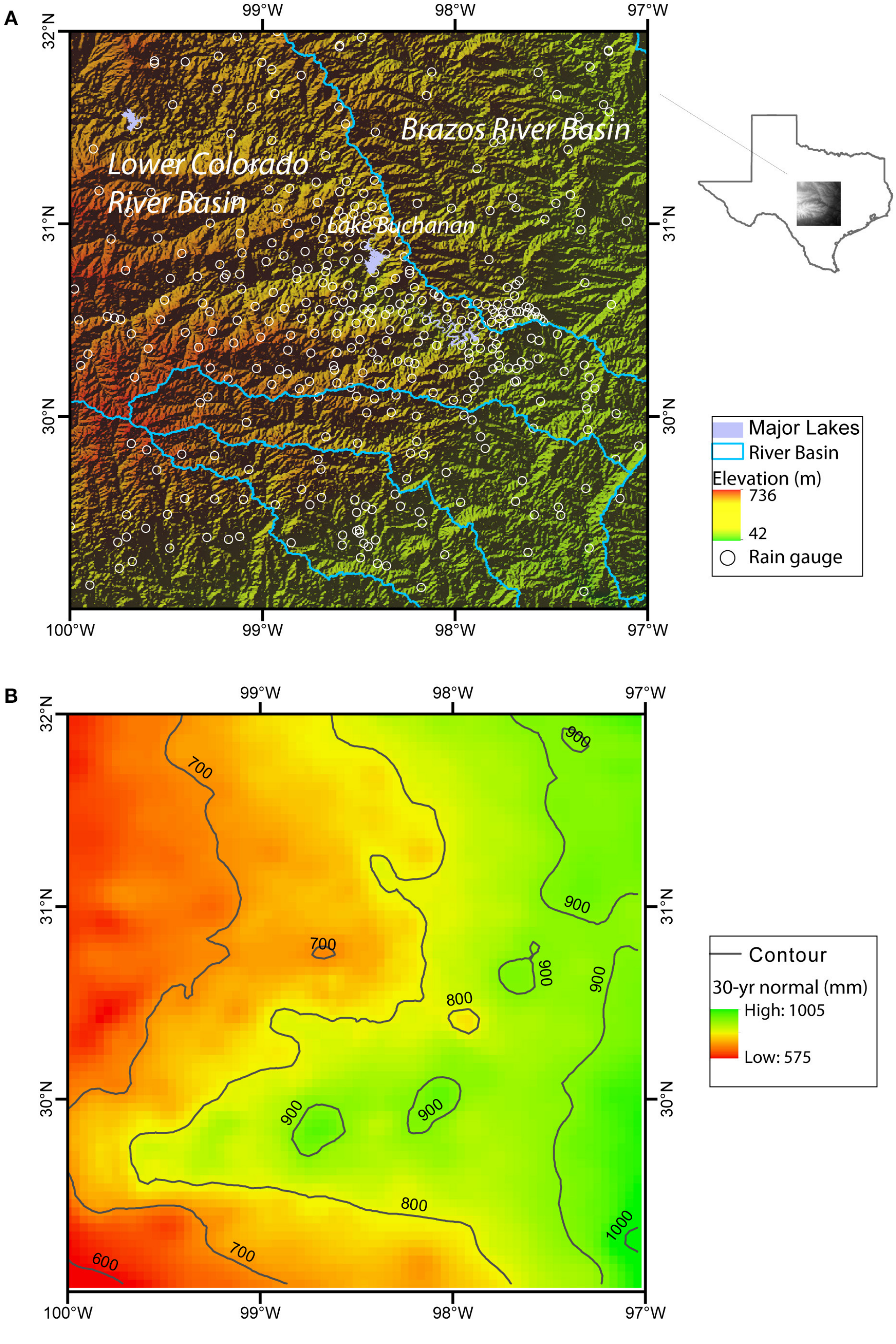


FIGURE 1 | (A) Study area boundary (lat: 29–32°N, lon: 100–97°W) and the shaded relief map (open circles correspond to the existing rain gauges as of Jan 2017); **(B)** 30-years precipitation normal extracted from PRISM, where color and contour lines represent total rain amount in mm.

resolution (0.28 vs. 0.75°) and temporal resolution (1 vs. 6-h), more vertical levels (137 vs. 60), and assimilation of more observations (Hennermann and Berrisford, 2017). Currently, the ERA5 dataset includes a high-resolution realization (HRES, ~0.28°) and a reduced-resolution, 10-member ensemble, and are available at both sub-daily and monthly intervals (Hennermann and Berrisford, 2017). For this study, the ERA5-Land data (0.1°) was used and was downloaded from the Climate Data Store (see **Table 1**). The grid resolution of ERA5-Land is higher than the native ERA5 resolution of 0.28°. Thus, during processing, the input air temperature, air humidity, and pressure used to run ERA5-Land were corrected to account for the altitude difference between the grid of the forcing and the higher resolution grid of ERA5-Land (Hennermann and Berrisford, 2017).

IMERG supersedes the TRMM 3B42 product as the next-generation precipitation product developed using the GPM data. The original purpose of IMERG was to calibrate, merge, and interpolate all satellite microwave precipitation estimates, together with microwave-calibrated infrared (IR) satellite estimates, precipitation gauge analyses, and potentially other precipitation estimators at fine temporal (30 min) and spatial resolution (0.1°) over the entire globe (Huffman et al., 2015). Previously, Tang et al. (2016) compared Day-1 IMERG with TRMM 3B42V7 over a well-gauged, mid-latitude basin in China, and concluded that the Day-1 IMERG product could be an adequate replacement of TRMM products, both statistically and hydrologically. Probably more relevant to this study, Omranian and Sharif (2018) compared the quality and accuracy of Day-1 IMERG product over the entire Lower Colorado River Basin. They showed the Day-1 IMERG product can be potentially used at small basin scales with errors comparable to those of weather radar products, provided that gauge-based real-time adjustment algorithms are available for correction. For this study, monthly IMERG V06 data (0.1°, 2000–present) was downloaded from NASA's data repository (see **Table 1**).

In addition to exploring temporal and spatial correlation in precipitation itself, other auxiliary predictors commonly considered during precipitation downscaling include elevation, vegetation index, and air temperature (Duan and Bastiaanssen, 2013; He et al., 2016; Vandal et al., 2018). For this study, elevation, slope, and aspect were tested as possible static auxiliary variables. The elevation data (DEM) was extracted from the Global Multi-resolution Terrain Elevation Data (GMTED2010) developed by U.S. Geological Survey and National Geospatial-Intelligence Agency (15 arc s or ~450 m). Slope and aspect were derived from the DEM using the Python package, `RichDEM` (Barnes, 2018). Monthly enhanced vegetation index (EVI) was also evaluated as a dynamic auxiliary variable. The response of vegetation to precipitation can have a lag time of 2–3 months in semi-arid areas (Quiroz et al., 2011). Previous studies utilizing the vegetation index were done at both monthly (López López et al., 2018) and annual (Duan and Bastiaanssen, 2013) scales. For this study, EVI was extracted from the Level-3 vegetation index product derived from MODIS, MOD13C2 (0.05°). Finally, the 2-m air temperature data was extracted from the ERA5-Land forcing data available from the Climate Data Store.

3. METHODOLOGY

3.1. Problem Formulation

Here a regression model is sought to relate a pair of low- and high-resolution precipitation maps that are created from different types/sources of data. Formally, the problem may be stated as the following statistical learning problem (Goodfellow et al., 2016)

$$\boldsymbol{\kappa} : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

where domain $\mathcal{X}$ represents the input space, including the low-resolution precipitation data and any auxiliary information, as explained later in section 3.3; and domain $\mathcal{Y}$ represents the high-resolution target space. In reality, the true mapping operator $\boldsymbol{\kappa}$ is not accessible. Thus, we seek an approximation to $\boldsymbol{\kappa}$, namely, finding $\mathbf{y} = f(\mathbf{X}, \Theta)$, where $\mathbf{y} \in \mathcal{Y}$, $\mathbf{X} \in \mathcal{X}$, f is a statistical mapping that is trained using the labeled training dataset $\{\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$ consisting of input samples $\mathbf{X}^{(i)} \in \mathbb{R}^{H\times W\times C}$ and output samples $\mathbf{y}^{(i)} \in \mathbb{R}^{H\times W}$ (H, W, and C denote the height, width, and channel dimensions of inputs) and Θ is a set of trainable parameters of f. In this work, we adopt an attention-based, U-Net model (AU-Net) for f.

3.2. Attention-Based Deep Convolutional Neural Net

Deep CNN models consist of a cascade of convolution blocks, each including one or more convolutional layers that perform convolution operations on inputs from the previous layer (Goodfellow et al., 2016)

$$\begin{aligned}
\mathbf{x}_c^l &= \sigma\left(\sum_{c'} \mathbf{W}_{c',c}^l \otimes \mathbf{x}_{c'}^{l-1} + b_c^l\right),\\
x_{i,j,c}^l &= \sigma\left(\sum_m \sum_n \sum_{c'} w_{m,n,c',c}^l x_{i+m,j+n,c'}^{l-1} + b_c^l\right),\\
&\quad i = 1, \ldots, H,\ j = 1, \ldots, W,\ c = 1, \ldots, C_f
\end{aligned} \tag{2}$$

where $\mathbf{x}^{l-1}$ and $\mathbf{x}^l$ are the input and output tensors of the l-th layer; subscripts m, n denote indices along the width and height dimensions of a kernel, c' is the index along channel dimension; c represents the index of output channel dimension C_f, which is equal to the number of kernels used for convolving the l-th layer; $\otimes$ is a convolution operator as defined in the second line of the above equation; $\mathbf{W}_{c',c}^l = \{w_{m,n,c',c}^l\}$ represents the weight matrix of the c-th kernel for the input channel c', and $\mathbf{b}^l = \{b_c^l\}$ represents a bias vector, both are trainable parameters; and σ represents the activation function. In practice, a number of other types of layers, such as batch normalization and pooling, are used in the convolution block to increase the learning efficiency while keeping the number of trainable parameters manageable (Goodfellow et al., 2016).

U-Net is a type of deep CNN and more specifically, an image-to-image autoencoder that was originally introduced in biomedical image segmentation (Ronneberger et al., 2015). Unlike some early deep CNN model designs that use dense (or fully connected) layers at the output end, U-Net is fully convolutional (i.e., consisting of only convolutional layers). The downsampling step (encoder) is designed to capture fine-scale image contexts by using repeated convolutional blocks to progressively extract downsampled feature maps, while the upsampling step (decoder) is designed to progressively enlarge the feature maps until the original image dimension is restored. In the final step, a 1 × 1 (kernel size) convolutional layer is used to condense the stack of feature maps along the channel dimension and generate a single output image, completing the image-to-image regression process (**Figure 2**). For this work, the rectified

TABLE 1 | Summary of datasets used in this study, where P–precipitation, T–temperature, DEM–elevation, EVI –enhanced vegetation index.

Data (Variable)	Format (resolution)	Source
Texas Mesonet (P)	Gauge	https://www.texmesonet.org
PRISM (P)	Gridded (4 km)	http://www.prism.oregonstate.edu
Stage IV (P)	Gridded (4 km)	https://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4
TRMM3B43 V7 (P)	Gridded (0.25°)	https://earthdata.nasa.gov
ERA5 (P,T)	Gridded (0.1°)	https://cds.climate.copernicus.eu
IMERG V06 (P)	gridded (0.1°)	ftp://arthurhou.pps.eosdis.nasa.gov
GMTED2010 (DEM)	Gridded (450 m)	https://www.usgs.gov/land-resources
MOD13C2 (EVI)	Gridded (0.05°)	https://modis.gsfc.nasa.gov/data/dataprod/mod13.php

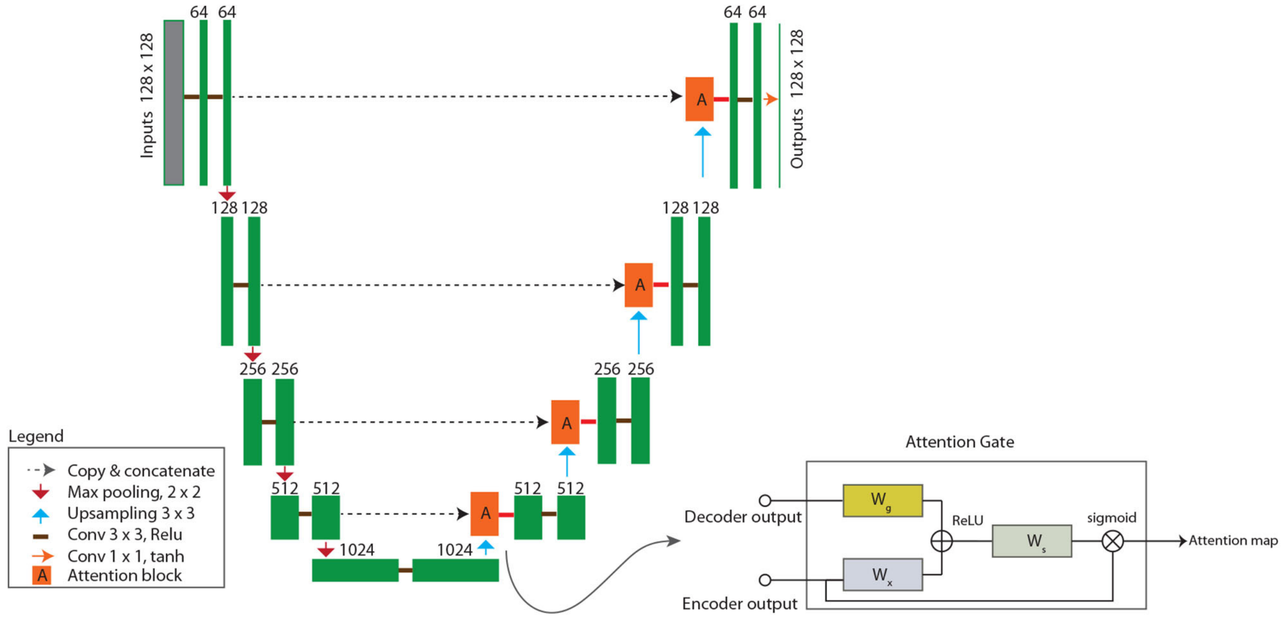


FIGURE 2 | Model architecture of attention-based U-Net (AU-Net), which consists of a pair of encoder and decoder for end-to-end learning. Green blocks are convolutional blocks, for which the number of kernels used is labeled on top of each block. Red blocks are attention gate blocks, the design of which is shown in the callout on bottom right. Dashed arrow lines are skip connections. Most hidden convolutional layers use 3 × 3 kernels and the ReLU activation function, except in the attention block and in the output layer, where 1 × 1 kernels and tanh are used. Meanings of other symbols are explained in the legend shown at bottom left.

linear unit (ReLU) is used as the activation function for all hidden layers except in the output layer, where the hyperbolic tangent function (tanh) is used. The pooling size is 2 so that the input layer dimension is halved after each pooling operation.

A key feature of the U-Net design is the skip connection, which is a combination of copy and concatenation operations to merge the fine-scale features from the downsampling step with the upsampled coarse-scale feature maps to better learn representations (dashed arrow lines in **Figure 2**) (Ronneberger et al., 2015). Mao et al. (2016) showed that the use of skip connections helps the training process to converge much faster and attain a higher-quality local optimum. So far, U-Net and its variants have been used in a large number of DL applications in geosciences (Sun, 2018; Arge et al., 2019; Karimpouli and Tahmasebi, 2019; Mo et al., 2019; Sun et al., 2019; Zhong et al., 2019; Zhu et al., 2019).

In CNN design, the size of the receptive field (e.g., kernel dimensions) directly affects the learning performance. If a single, fixed-size kernel is used to scan the inputs, global information may be missed, especially when the resolution of the input image is high. In the literature, several methods have been proposed to circumvent the issue. The skip connection used in U-Net is one example. As another example, Ren et al. (2016) proposed a multiscale CNN model consisting of a pair of coarse-scale and fine-scale autoencoders, the former uses a 11 × 11 kernel, while the latter uses a 7 × 7 kernel; the output of the coarse network is fed to the fine network as additional information to refine the coarse prediction with details. In recent years, self-attention has emerged as yet another alternative for capturing multiscale contexts in an image.

Simply speaking, attention refers to the biological capability of certain animals, including humans, to direct their gaze rapidly toward objects of interest in a visual environment, transforming the understanding of a visual scene into a series of computationally less demanding, localized visual analysis problems (Itti and Koch, 2001). Significant interests exist in computational neuroscience to replicate such capability in pattern recognition algorithms. In machine translation (natural language processing), for example, self-attention has been proposed as a mechanism for relating different positions of a single sequence in order to compute a representation of the sequence (Vaswani et al., 2017). In image processing, attention has been used to model the image as a sequence of regions, allowing for better capturing of large-scale features while producing sharper local details, thus leading to improved performance in object tracking, object detection, and image caption generation applications (Xu et al., 2015; Oktay et al., 2018; Bello et al., 2019; Hou et al., 2019). For this study, we hypothesize that the same attention mechanism that helps to capture multiscale spatial and temporal interactions may also be useful in learning the covariational patterns between high- and low-resolution precipitation products.

In general, attention-based algorithms work by suppressing the irrelevant background and enabling salient features to dynamically come to the forefront (Xu et al., 2015). We adopt the attention-gate module proposed by Oktay et al. (2018), which has the advantage of being compatible with the standard CNN models (e.g., U-Net) and can be added as an additional block without incurring significant computational overhead. As illustrated in **Figure 2**, the attention block is attached to the upsampling step of the U-Net (i.e., red blocks in **Figure 2**). For the l-th layer, the inputs to the attention block are outputs from the coarse-scale decoder ($\mathbf{g}^l$) and that from the encoder (via the skip connection) ($\mathbf{x}^l$). Inside the attention block, the inputs are passed through separate 1×1 convolutional layers, concatenated, and then passed through another 1×1 convolutional layer (with activation) to arrive at an attention map. In essence, the attention block may be regarded as a sub-network and its role is to suppress irrelevant features from the skip connections using information from the decoder. Mathematically, the series of attention gate operations may be described as (Oktay et al., 2018)

$$\mathbf{q}^l = \mathbf{W}_s \otimes \left(\sigma_1\left(\mathbf{W}_x \otimes \mathbf{x}^l + \mathbf{b}_x + \mathbf{W}_g \otimes \mathbf{g}^l + \mathbf{b}_g\right)\right) + b_s, \quad (3)$$

$$\boldsymbol{\alpha}^l = \sigma_2\left(\mathbf{q}^l\left(\mathbf{x}^l, \mathbf{g}^l; \Theta\right)\right), \quad (4)$$

where $\Theta = \{\mathbf{W}_x, \mathbf{W}_g, \mathbf{b}_g, \mathbf{b}_x, b_s\}$ represents a set of trainable weight matrices and bias terms, σ_1 is ReLU activation function, σ_2 is sigmoid activation function, $\boldsymbol{\alpha}^l$ is the resulting attention map for weighting different regions in the input.

3.3. Network Training and Performance Metrics

Monthly data from 2001/01 to 2018/12 were divided into three parts, training ($N_r = 168$), validation ($N_v = 24$), and testing ($N_t = 24$). After preliminary analyses, four predictor groups were considered, including coarse-resolution satellite/reanalysis precipitation products (P), enhanced vegetation index (EVI), air temperature (T), elevation (DEM), slope, and aspect,

$$M_1 : P_{t-2:t},\ EVI_{t-1:t},\ DEM,\ Slope,\ Aspect, \quad (5)$$

$$M_2 : P_{t-2:t},\ DEM,\ Slope,\ Aspect, \quad (6)$$

$$M_3 : P_{t-2:t},\ DEM, \quad (7)$$

$$M_4 : P_{t-2:t},\ T_{t-2:t}, \quad (8)$$

where the subscript t denotes the month index, and the target is PRISM data at month t. Models M_1 to M_3 include both static and dynamic variables, while M_4 only includes coarse-resolution dynamic variables. All AU-Net models are developed at the 128×128 grid resolution, which is about 2.6 km/pixel for the current problem. The lags for the dynamic variables are chosen based on preliminary analyses. Higher-resolution grids are tested as part of the sensitivity study. Before training, all inputs are resampled to the same grid resolution through bilinear interpolation, and then normalized before passing to the DL model for training.

All models were developed in the `PyTorch` (v1.1) machine learning framework. The loss function used in training the AU-Net models is the mean square error (MSE) defined as

$$\text{MSE} = \frac{1}{N_r}\sum_{i=1}^{N_r} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2, \quad (9)$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ represent the true precipitation data used for training and the predicted data, respectively. The ADAM solver (Kingma and Ba, 2014) was used to train the neural nets, with a learning rate $\alpha = 5 \times 10^{-4}$, first moment decay rate $\beta_1 = 0.5$, and second moment decay rate $\beta_2 = 0.999$. During training and validation, the data samples were randomly shuffled to improve generalization. Training of the AU-Net was carried out on a dual-processor computing node equipped with 128Gb RAM and Nvidia 1080-TI GPU. A total of 100 epochs were used for each model and the batch size (i.e., number of samples used in each solver iteration) was set to 10. Early stopping was implemented by monitoring the validation loss to mitigate overfitting. Training time depends on the model size and grid resolution and generally takes about 15 min for each model at the 128×128 grid resolution.

Model performance evaluation includes comparison with both *in situ* gauge data and PRISM data for the testing period. For comparison with *in situ* data, three metrics are used, namely, the root mean square error (RMSE), bias, and correlation coefficient (CC),

$$RMSE = \sqrt{\frac{1}{N_G}\sum_{i=1}^{N_G}(y_{g,i} - \hat{y}_i)^2}, \quad (10)$$

$$BIAS = \frac{1}{N_G}\frac{\sum_{N_G}\left(\hat{y}_i - y_{g,i}\right)}{\sum_{N_G}\left(y_{g,i}\right)} \times 100, \quad (11)$$

$$CC = \frac{1}{N_G\sigma_y\sigma_g}\sum_{i}^{N_G}(y_{g,i} - \mu_g)(\hat{y}_i - \mu_y), \quad (12)$$

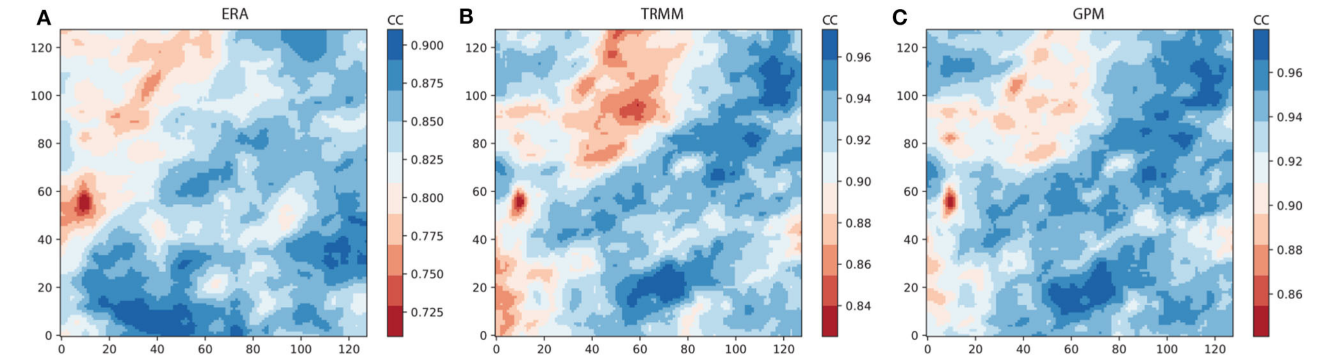


FIGURE 3 | Correlation maps between PRISM and **(A)** ERA, **(B)** TRMM, and **(C)** GPM over the entire study area. Correlation at each grid cell is calculated as the Pearson correlation coefficient between pairs of time series for that cell.

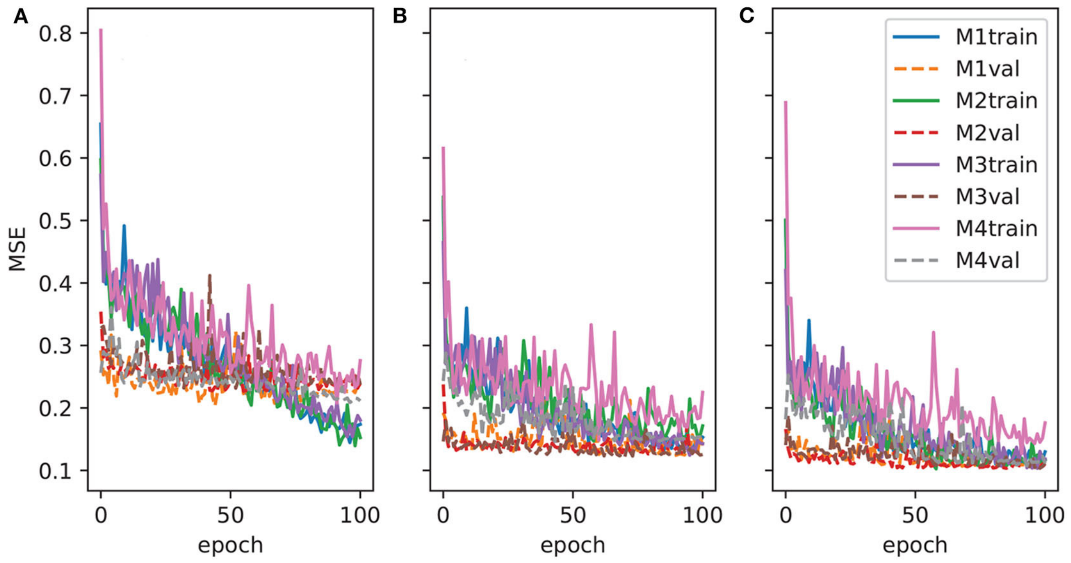


FIGURE 4 | Convergence history of AU-Net training and validation: **(A)** ERA5, **(B)** TRMM, and **(C)** GPM.

where y_g and $\hat{y}$ are measured and predicted data at a gauge location, N_G is the number of usable gauge data for a month, and μ and σ denote mean and standard deviation. If multiple gauges exist in a grid cell, we used the average of gauge values for that cell. Mean metric values were then obtained by averaging over all months in the testing period.

For image-to-image comparison, RMSE is calculated over all grid cells. In addition, the structural similarity index metric (SSIM) is calculated, which is a metric widely used in computer vision to measure similarity between two images (Wang et al., 2004). Specifically, for two sliding windows u and v operating separately on the testing and reference images (grayscale), SSIM is defined as

$$\mathrm{SSIM}(u,v)=\frac{(2\mu_u\mu_v+c_1)(2\sigma_{uv}+c_2)}{(\mu_u^2+\mu_v^2+c_1)(\sigma_u^2+\sigma_v^2+c_2)}, \qquad (13)$$

where μ and σ represent the mean and standard deviation of image patches falling in the sliding windows, and c_1 and c_2 are small constants introduced to avoid numerical instability (Wang et al., 2004). The global SSIM is obtained by averaging the patch SSIM values and is in the range $[-1,1]$, with higher values indicating better pattern matches. The sizes of sliding windows used are 11 × 11.

3.4. Residual Correction

Residual correction is commonly used in the final step of downscaling to fuse gauge observations (Haylock et al., 2006; Duan and Bastiaanssen, 2013; Chen et al., 2018). We experimented with both kriging and inverse distance weighting (IDW), and found that the latter gave better results. Thus, the IDW scheme was adopted to interpolate the residual errors between AU-Net results and gauge observations, e_i, to the entire grid, which were then added to the AU-Net estimates $\hat{y}$.

Specifically, the final estimate $\tilde{y}$ is obtained by

$$\tilde{y}(\mathbf{x}) = \begin{cases} \hat{y}(\mathbf{x}) + \left(\sum_{i=1}^{N_G} w_i(\mathbf{x}) e_i\right) \Big/ \left(\sum_{i=1}^{N_G} w_i(\mathbf{x})\right), & \text{if } d(\mathbf{x},\mathbf{x}_i) \neq 0 \\ y_{g,i}, & \text{if } d(\mathbf{x},\mathbf{x}_i) = 0 \end{cases}$$

where $e_i = y_{g,i} - \hat{y}_i$ is the residual calculated at a gauge location, $d = \| \mathbf{x} - \mathbf{x}_i \|$ is the distance between a grid cell location $\mathbf{x}$ and a gauge location $\mathbf{x}_i$, and the weight factor is $w_i = d^{-\beta}$. For this study, we set $\beta = 2$ based on error statistics calculated against PRISM data.

4. RESULTS

For each coarse-resolution precipitation product (i.e., ERA5-Land, TRMM, and IMERG), the performance of four groups of predictors ($\mathrm{M}_1 - \mathrm{M}_4$) that are defined under section 3.3 were evaluated, leading to a total of 12 different AU-Net models. For brevity, IMERG will be simply referred to as GPM, and ERA5-Land as ERA5 in the following discussions.

As part of the exploratory analyses, the temporal correlations between PRISM and the coarse-resolution products at all 128 × 128 grid cell locations (i.e., after resampling via bilinear interpolation) were calculated and are shown in **Figures 3A–C**. The correlation maps of both TRMM and GPM exhibit similar spatial patterns, which tend to be higher in the eastern part and lower in the northwest high-elevation areas; all correlation values are above 0.8 (Note: despite the similarity, TRMM and IMERG processing are different in a number of ways, e.g., the Level-2 and Level-3 algorithms, the infrared data used for gap filling and replacement, as well as the spatiotemporal resolutions). In comparison, the correlation between ERA5 and PRISM is generally lower, except near the southwestern corner of the study area. Nevertheless, the large-scale spatial patterns of all three coarse-resolution products are similar, all exhibiting this diagonally oriented (southwest to northeast) stripe pattern.

4.1. Performance of AU-Net Models

Figure 4 plots the training and validation errors vs. training epochs for all models. The ERA5 model group (**Figure 4A**) tends to have larger training and validation errors than TRMM (**Figure 4B**) and GPM (**Figure 4C**) groups. In each group, the M_4 AU-Net model tends to have slower training convergence rate and stronger oscillations than the rest of the models.

Table 2 summarizes the mean performance metrics of all products and (uncorrected) AU-Net models against the gauge data for the test period 2017/01–2018/12. The target data, PRISM, has the smallest RMSE (3.59 cm) and highest CC (0.637) values. Among the three satellite/reanalysis products, TRMM and GPM have similar mean RMSE values (4.19 and 4.15 cm, respectively), which are all lower than that of the ERA5 (4.89 cm), but are more than 16% higher than that of the PRISM. The bias of TRMM (0.41) is lowest among all. GPM shows a slightly higher mean CC value (0.408), but is still about 35% lower than that of PRISM. Note that the CC values shown in **Table 2** are lower than those seen previously in **Figure 3**. This is because the former quantifies the spatial correlation between gridded products and point measurements at gauge locations, while the latter measures CC between harmonized time series at each grid cell. Also rain gauges are subject to various errors as mentioned in the Introduction, and point-scale gauge measurements may deviate significantly from areal precipitation (Tang et al., 2018a).

TABLE 2 | Summary of gauge comparison metrics on testing data (2017/01–2018/12) for three precipitation products, ERA5, TRMM, and GPM, and AU-Net models (best performing member in each category is highlighted).

Product	RMSE (cm)	Bias	CC
PRISM	3.59	10.58	0.637
ERA5	4.89	25.18	**0.283**
M_1	4.88	**1.14**	0.247
M_2	4.86	6.13	0.251
M_3	**4.80**	9.33	0.243
M_4	4.96	1.54	0.247
TRMM	**4.19**	**0.41**	0.396
M_1	4.23	−3.37	**0.410**
M_2	4.21	−4.95	0.407
M_3	4.22	3.78	0.407
M_4	4.26	3.91	0.399
GPM	**4.15**	4.27	0.408
M_1	4.19	−12.11	**0.416**
M_2	4.16	−1.76	0.414
M_3	4.19	**0.49**	0.412
M_4	4.20	−4.97	0.407

At the gauge level, **Table 2** suggests that the AU-Net models in the ERA5 group show similar mean RMSE and CC, but the bias is reduced significantly compared to the original ERA5. The AU-Net models for TRMM show slightly worse RMSE than the original TRMM, and slightly better CC, but the bias is also larger. The same is true for GPM AU-Net models. All the metrics are summarized in **Figures 5A–C** in separate Taylor diagrams (Taylor, 2001), which help to visualize model performance in terms of CC, standard deviation (SD), and RMSE relative to the reference gauge observations. The Taylor diagrams suggest that all gridded precipitation products underestimate the data spread as seen in the gauge data, which is due to the harmonization process behind the gridded products. ERA5 has the greatest SD, while TRMM and GPM have similar SD values. The AU-Net models for different data groups are mostly clustered together, although the M_1 models seem to do better than others.

In **Supplementary Figures 1–3**, boxplots of the monthly values of MSE, bias, and CC for each AU-Net model are provided. In general, the boxplots support the aforementioned observations. Moreover, they also suggest that the range of model performance metrics depends on the quality of the coarse-resolution inputs. For example, models trained using TRMM and GPM, both are already gauge corrected, generally exhibit smaller variations in metric values than the models trained using ERA5 do.

Overall, on the basis of rain gauge data comparison, the best performers are scattered among predictor groups M_1–M_3 for the three products considered. The M_4 group seems to underperform

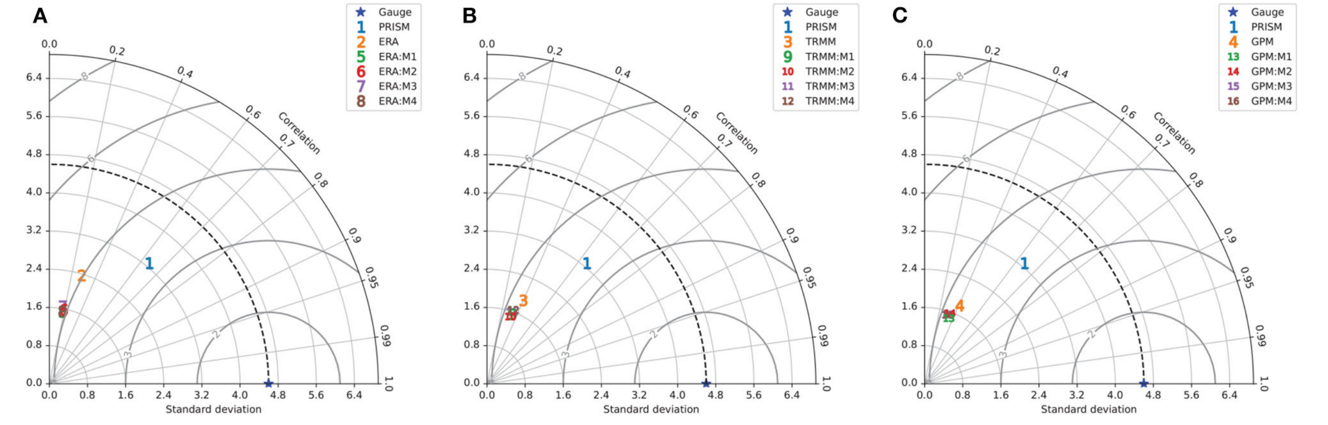


FIGURE 5 | Taylor diagrams summarizing statistics of all gridded rainfall products and uncorrected AU-Net models, against the gauge data and PRISM data for test period (2017/01–2018/12): **(A)** ERA, **(B)** TRMM, and **(C)** GPM. The blue star on the horizontal axes corresponds to the gauge dataset result.

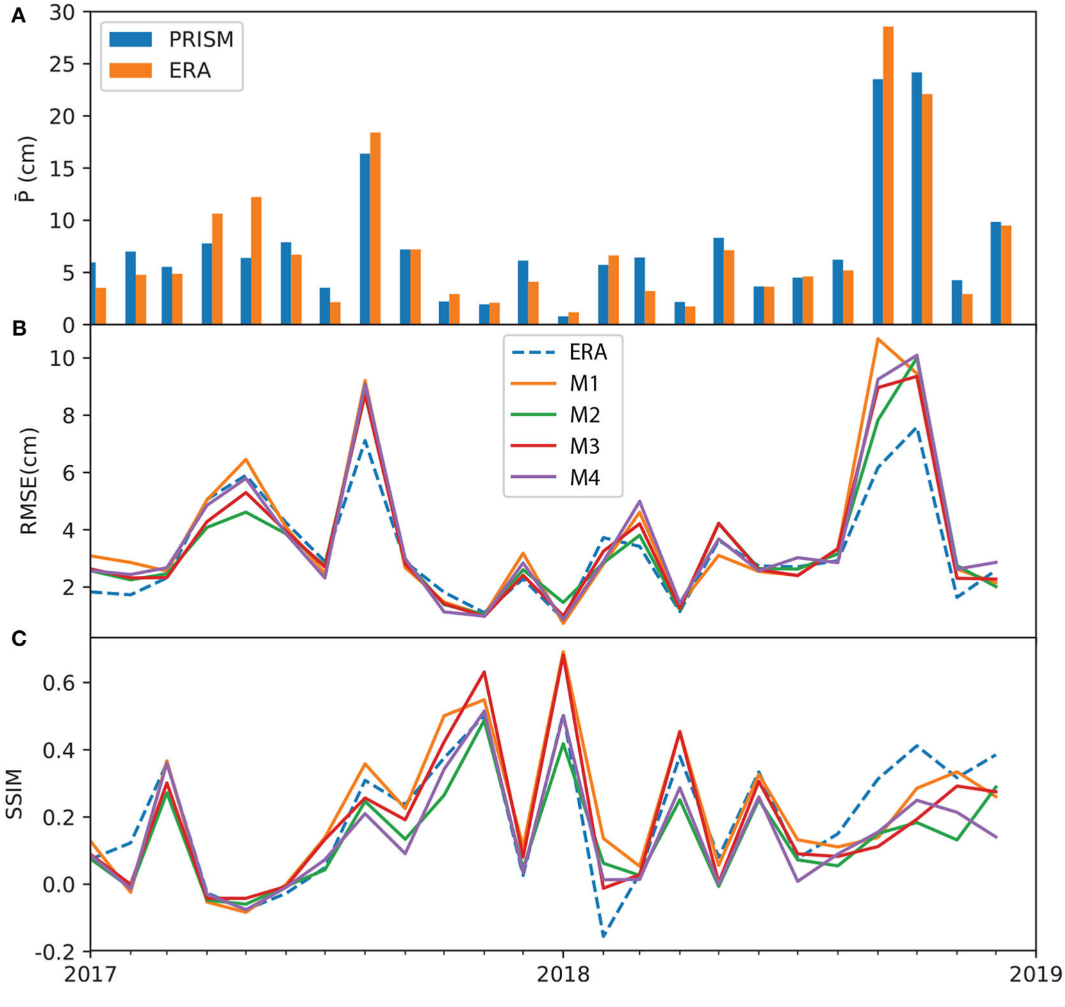


FIGURE 6 | Time series of ERA metrics over the test period (2017/01–2018/12): **(A)** monthly averaged PRISM and ERA data; **(B)** mean grid averaged RMSE; and **(C)** SSIM. All subplots share the same x axis.

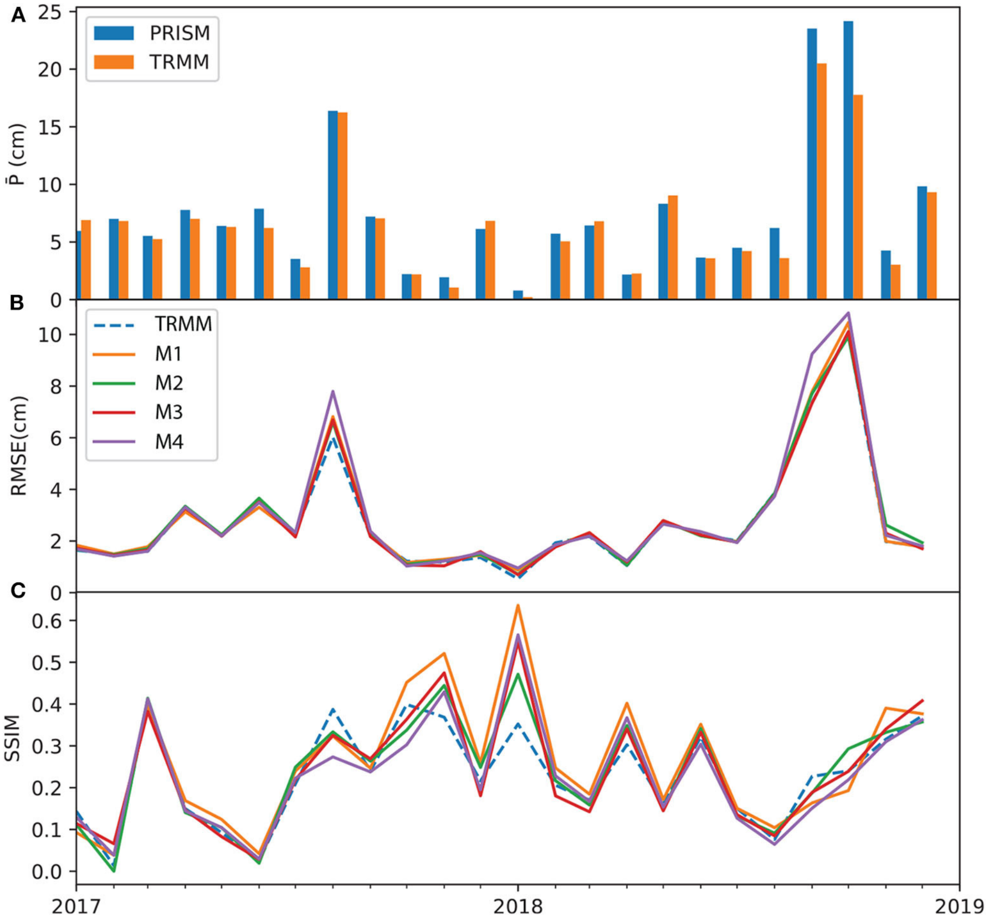


FIGURE 7 | Time series of TRMM metrics over the test period: **(A)** monthly averaged PRISM and TRMM data; **(B)** mean grid averaged RMSE; and **(C)** SSIM. All subplots share the same x axis.

compared to the other three predictor groups, due to its use of only coarse-scale information.

At the grid-level, **Figures 6–8** show the monthly time series of performance metrics for each product. The top panel of each figure compares the spatially averaged precipitation calculated on PRISM and the respective data product. In general, results suggest that the model performance is sensitive to the amount of precipitation, as well as to antecedent conditions of dynamic variables (i.e., P, T, and EVI). The models tend to outperform the baseline coarse-resolution product in dry months than in wet months. For ERA5, models M_1 and M_3 improve over ERA5 (as measured in SSIM) in the mid part of the test period, ranging from 2017/03 to 2018/06. Most ERA5 models, however, underperform the original ERA5 data during two extreme wet events, one in 2017/08 when Hurricane Harvey made landfall and the other during the record-breaking wet period 2018/08–2018/09. This may be attributed to the extremity of the events and the lack of predictability at the monthly scale. In the case of TRMM and GPM models, the metrics time series are less oscillatory—all models tend to have very similar RMSE values, and the model SSIM values are better than the TRMM and GPM during the dry period in the middle of the test period. Compared to ERA, the metrics of TRMM and GPM model stay close to the original data products even during the two extreme events.

To give examples of learned patterns, in **Figure 9** we plot the AU-Net results (left three columns), together with the original coarse-resolution data (top row) and the fine-resolution PRISM and Stage-IV datasets (the rightmost column) for 2017/01, which was a relatively wet month. The SSIM between each image and PRISM is shown on top of each subplot. PRISM and Stage-IV have very similar spatial patterns. Compared to the PRISM and Stage-IV data, ERA5 (upper-left) did not capture the higher rainfall zone near the eastern side, while both TRMM and GPM were able to capture the same zone in a large-scale sense. The AU-Net models that include static information (i.e., M_1–M_3) introduce more fine-scale features in the results, such as near the southwestern corner of the domain and inside the wetter zone; however, the improvements over the original products are rather limited in terms of SSIM. Only the M_2 model under the GPM group predicts the location of the high-precipitation relatively accurately. The M_4 models, which only

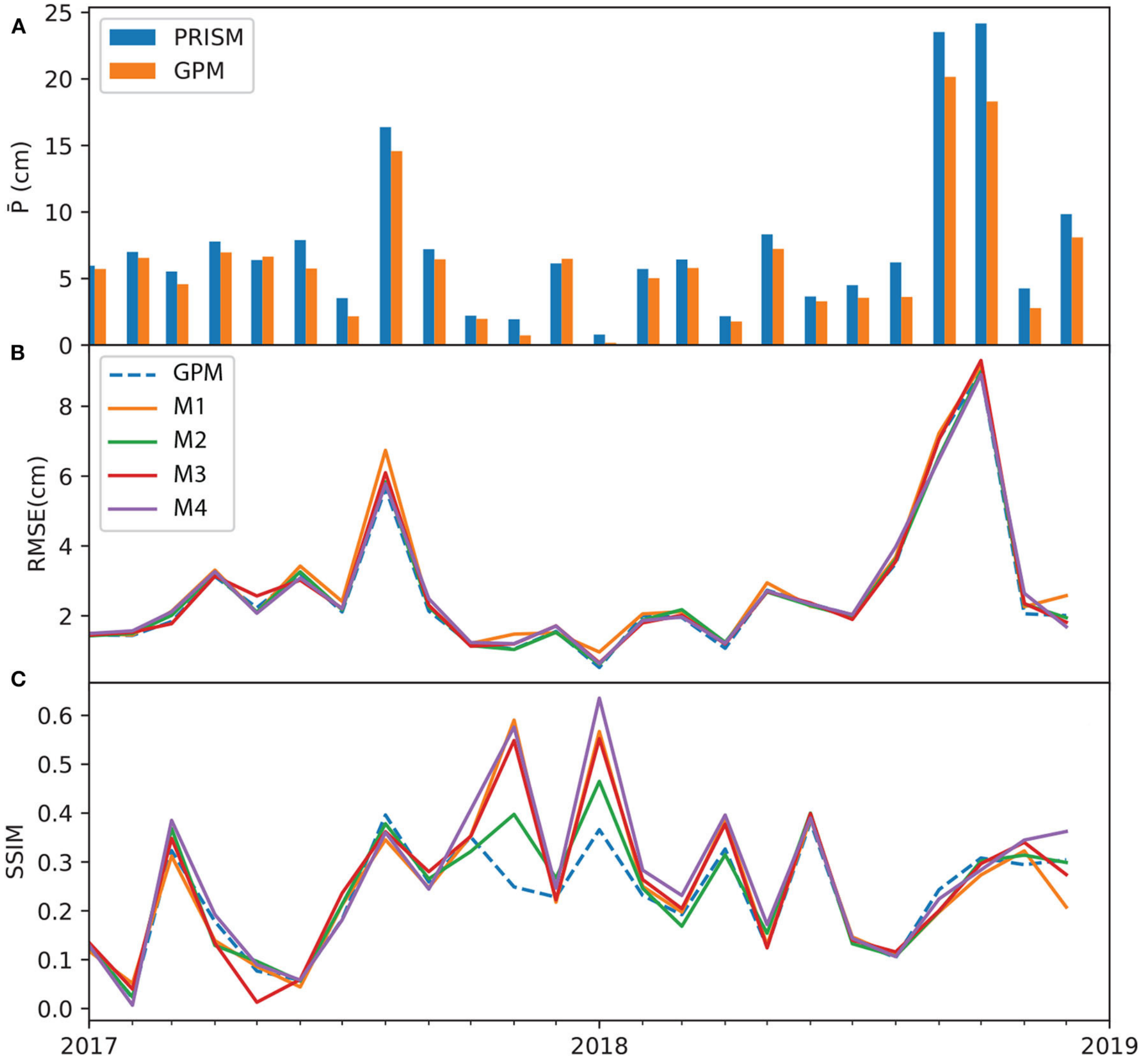


FIGURE 8 | Time series of GPM metrics over the test period: **(A)** monthly averaged PRISM and GPM data; **(B)** mean grid averaged RMSE; and **(C)** SSIM. All subplots share the same x axis.

use coarse-resolution information, yield more smooth features than the other models do.

As another example, we compare the AU-Net results for the month 2018/01, which was a relatively dry month. **Figure 10** shows that all three coarse-resolution products are able to delineate the large low-precipitation zone near the northwestern corner. Under the ERA group, the M_1 model gives the best pattern match (SSIM = 0.69), while in the cases of TRMM and GPM, M_1 (SSIM = 0.63) and M_4 (SSIM = 0.67) give the best pattern match, respectively. In this case, even the M_4 model, which only uses dynamic variables, is able to infill some fine-scale features.

Results in **Figures 9**, **10** highlight the promises, as well as challenges, in extracting and learning the fine-scale features in precipitation data for this low-predictability study area. The use of antecedent conditions and auxiliary static information helped to improve the baseline coarse-resolution products in some cases, but deteriorated the baseline in other cases. In particular, we note that static information tends to be more useful under dry conditions, while autocorrelation in precipitation itself seems to play a major role in predictability. No predictor group consistently performed better than the others. The inherent high variability of precipitation in space and time, especially in topographically complex regions, makes the pattern-based downscaling challenging, without further correction using *in situ* data.

4.2. Corrected AU-Net Models

The AU-Net models were corrected by first calculating the error residual between model and gauge data, and then interpolating to the grid using the IDW scheme described under this section. **Figure 11** shows the gauge-corrected AU-Net results on the 2017/01 data used in **Figure 9**. Similarly, **Figure 12** shows the gauge corrected results for 2018/01 data. Results suggest that

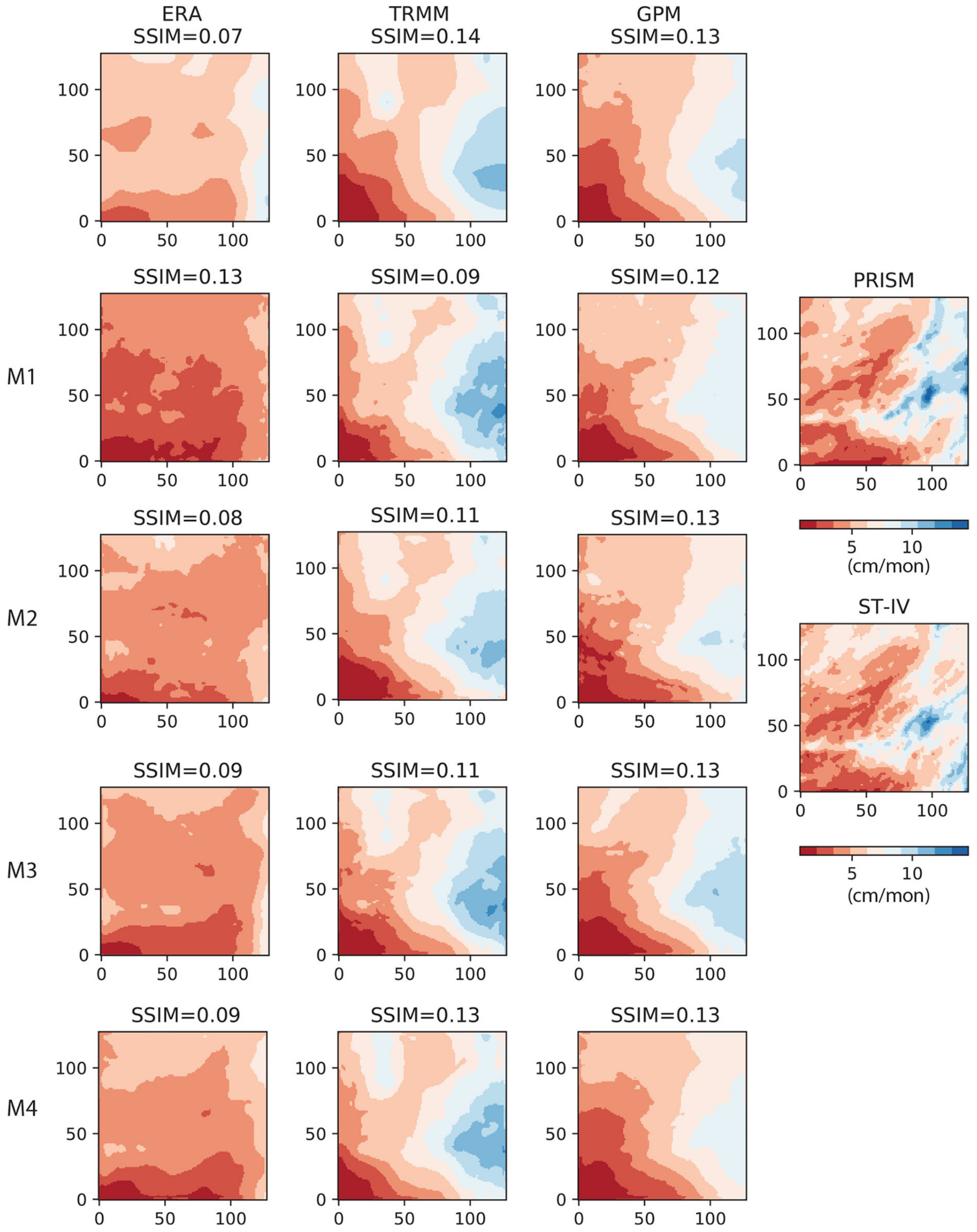


FIGURE 9 | AU-Net results on 2017/01 data (128 × 128 grid). Left column (from top to bottom) ERA5 and the corresponding $M_1 - M_4$ AU-Net models; 2nd column: TRMM and the corresponding AU-Net models; 3rd column: GPM and the corresponding AU-Net models; rightmost column: PRISM and the reference Stage-IV data for the same month. All subplots are scaled to the same color range.

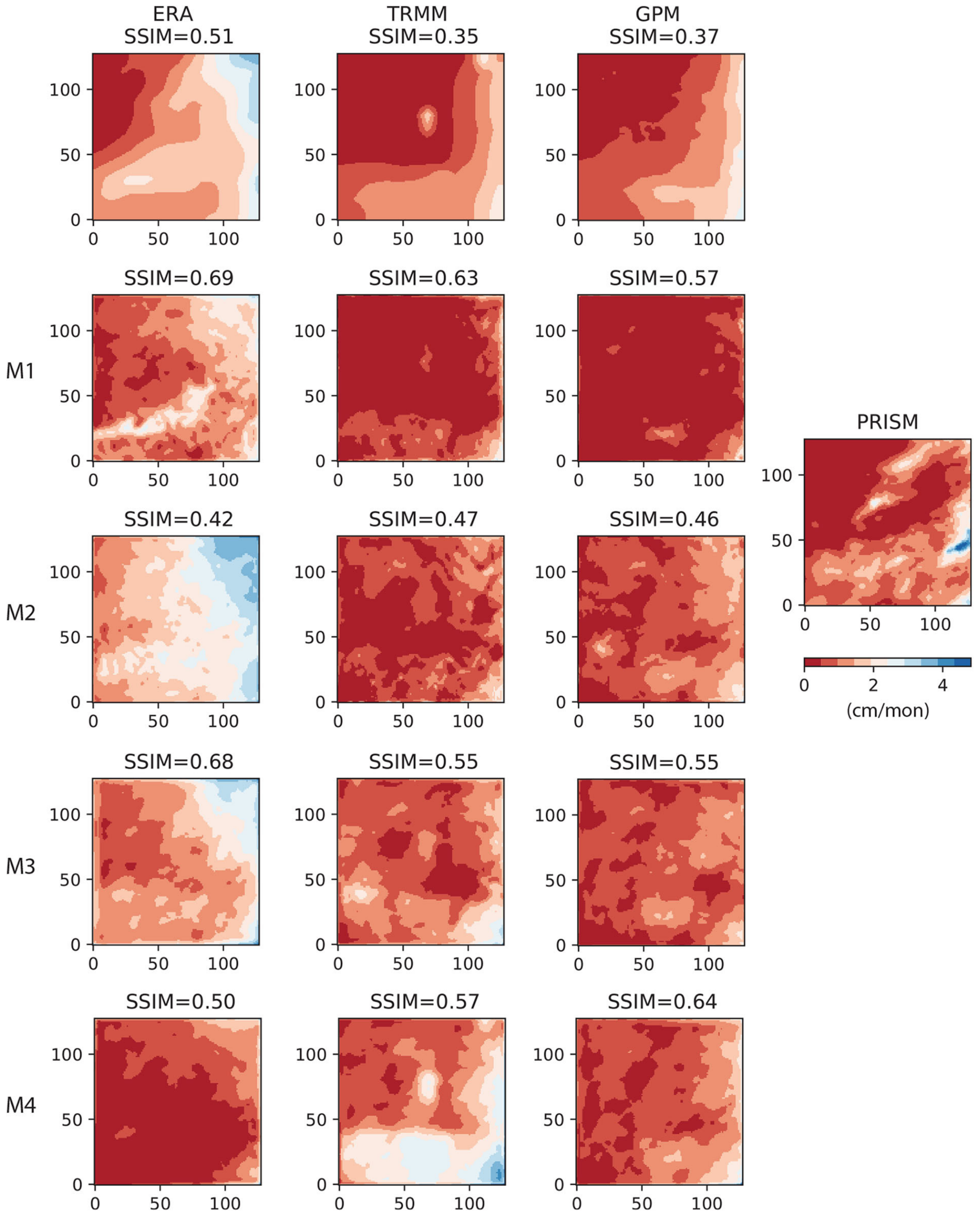


FIGURE 10 | AU-Net results on 2018/01 data (128 × 128 grid). Left column (from top to bottom): ERA5 and the corresponding $M_1 - M_4$ AU-Net models; 2nd column: TRMM and the corresponding AU-Net models; 3rd column: GPM and the corresponding AU-Net models; rightmost column: PRISM data for the same month.

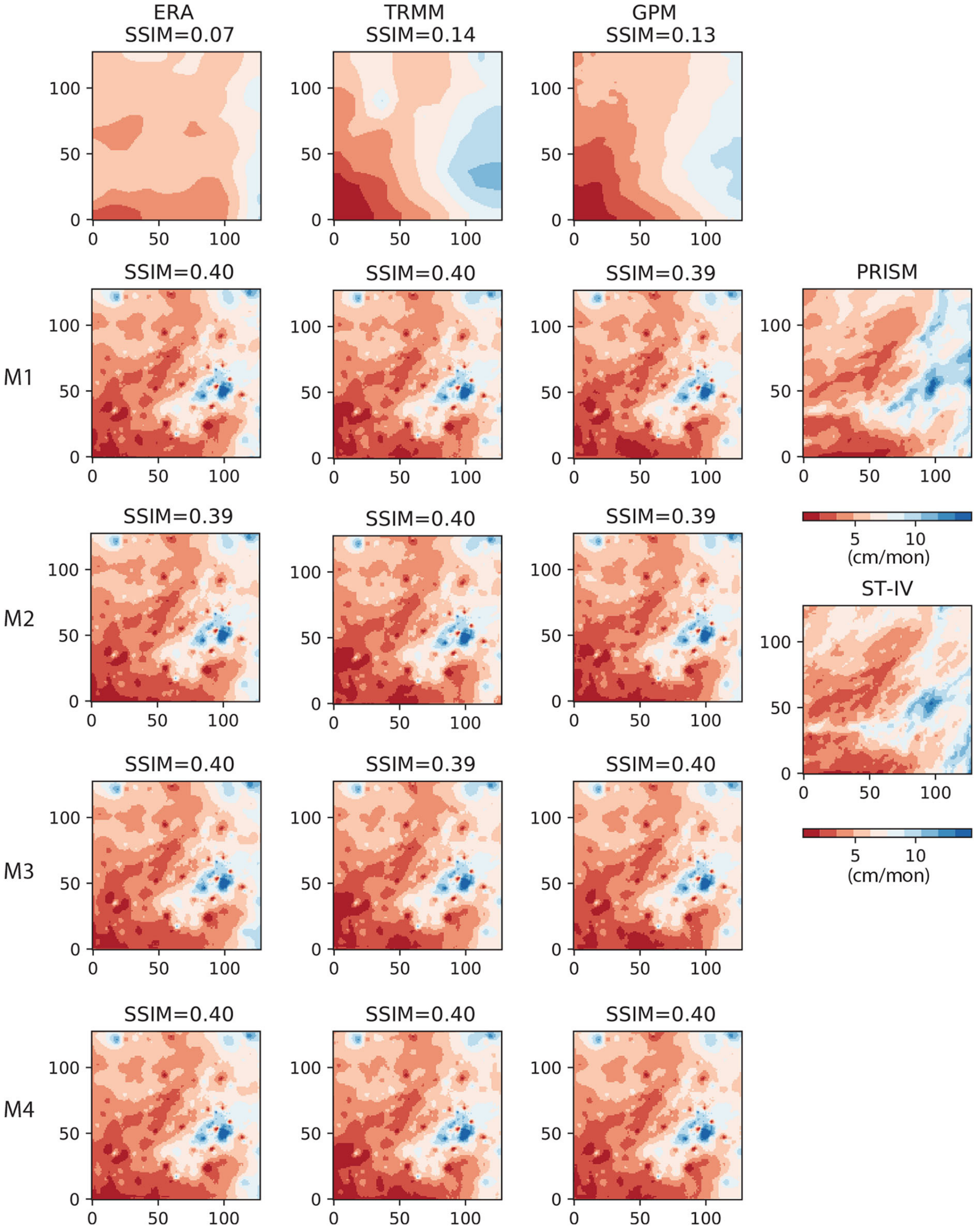


FIGURE 11 | Results of gauge correction using inverse distance weighting on 2017/01 data. Left column (from top to bottom) ERA5 and the corresponding $M_1 - M_4$ AU-Net models; 2nd column: TRMM and the corresponding AU-Net models; 3rd column: GPM and the corresponding AU-Net models; rightmost column: reference PRISM and Stage-IV data for the same month.

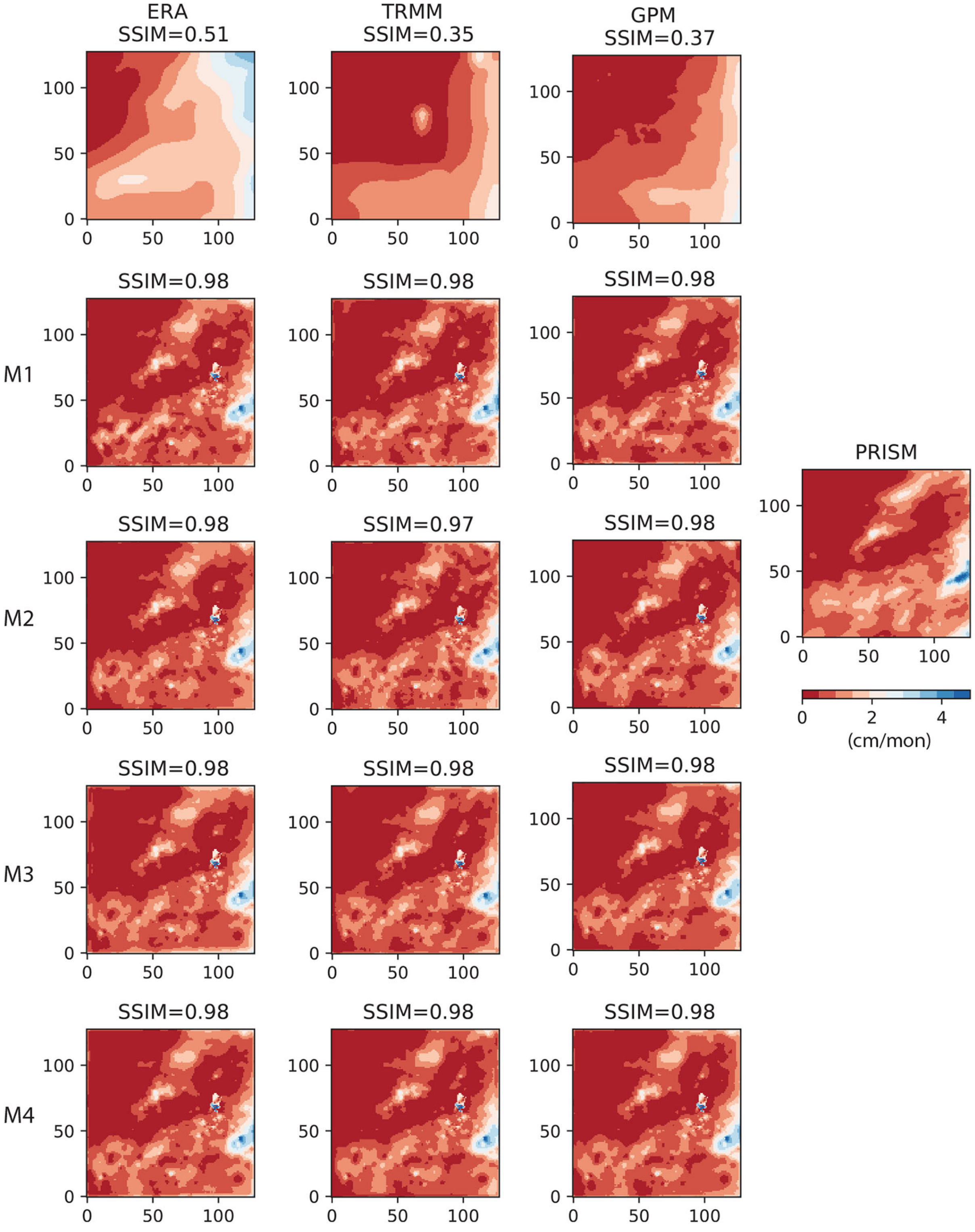


FIGURE 12 | Results of gauge correction on 2018/01 data. Left column (from top to bottom) ERA5 and the corresponding $M_1 - M_4$ AU-Net models; 2nd column: TRMM and the corresponding AU-Net models; 3rd column: GPM and the corresponding AU-Net models; rightmost column: reference PRISM data for the same month.

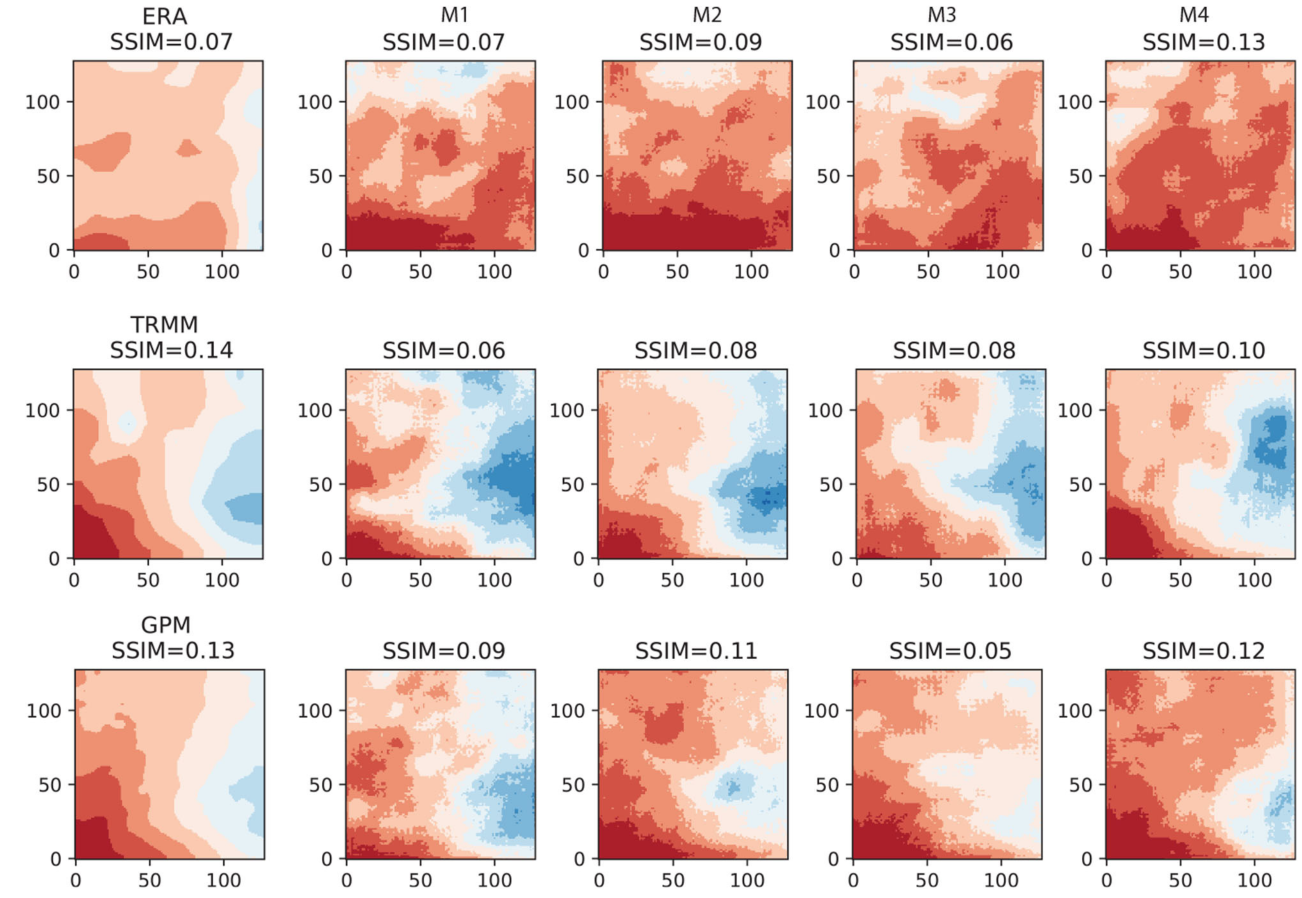


FIGURE 13 | Results of U-Net models for 2017/01. Top row: ERA5 results; mid-row: TRMM results; and bottom row: GPM results.

gauge correction significantly improved the pattern match for all AU-Net models, leading to convergence in model patterns among all models. In the case of 2017/01, gauge correction actually introduced more fine-scale features than that are present in the PRISM image. This may be caused by the difference in point measurement set used, and also in the gauge data interpolation algorithm. In the case of 2018/01, gauge correction almost resulted in identical patterns to the PRISM image. The dominating effect of gauge correction observed here is not surprising, given the large number of gauges available for the study area (361 in 2017/01 and 459 in 2018/01). RMSE of the gauge-corrected AU-Net models (not reported here) is also significantly reduced, compared to the uncorrected results.

4.3. Effect of Attention Mechanism

A main motivation of this work to explore the use of attention mechanism for multiscale pattern extraction. To demonstrate the effect of attention mechanism, we train the classic U-Net models using the same model structures, but with the attention block removed (see **Figure 2**). The kernel size used in the U-Net is 4 × 4 and stride size is 2. As mentioned before, attention mechanism helps to capture the large-scale patterns while producing sharper local details. **Figure 13** shows the result for the same 2017/01 data, as shown earlier in **Figure 9**. An immediate observation from **Figure 9** is that all images produced by the U-Net are more blurry than those generated by the AU-Net models. The SSIM values are also smaller than their counterparts in AU-Nets.

5. DISCUSSION

In this work, the feasibility of using AU-Net to downscale precipitation data was investigated over Central Texas, U.S., on three coarse-resolution satellite/reanalysis products. Climate in the region ranges from semi-arid in the west to subtropical in the east. The climate and hilly terrain of the region lead to strong spatial and temporal variations in precipitation patterns, making downscaling for the study area at the monthly level especially challenging. The AU-Net models, which can extract features at multiple scales, are used to learn the mappings between coarse- and fine-resolution products.

At the regionally aggregated scale, all three coarse-resolution products (baseline) are shown to have relatively strong temporal correlation with the fine-resolution PRISM product (>0.7) at the monthly level. The question is whether this correlation can be propagated down to the grid level. A main finding of this study is that the efficacy of downscaling and thus, model improvement, depends on the precipitation amount and information content embedded in antecedent conditions and auxiliary variables, as well as the quality of the original product. Under drier conditions, the precipitation patterns are more contiguous and are easier for the AU-Net models to learn. In addition, the static information tends to be more useful under drier conditions. On the other hand, in wet months the precipitation patterns become spatially heterogeneous and are more difficult to downscale without additional constraints. This observation is largely in agreement with the previous studies that show systematic errors in precipitation products are proportional to precipitation rates, which is higher for higher rates (e.g., AghaKouchak et al., 2012).

The fine-resolution auxiliary variables considered in this study only include EVI and DEM. EVI may be less reliable as a predictor at the monthly level (Duan and Bastiaanssen, 2013), even with added lags. Thus, future effort should focus on experimenting with alternative fine-resolution remotely sensed information, which is especially valuable when high-density rain gauge networks are not available.

Performance of DL models may depend on grid resolution. Higher grid resolution models, however, also increase training time significantly. In this work, we mainly experiment with 128 × 128 grids. As a sensitivity analysis, we also trained the same AU-Net models on 256 × 256 grids. The average training time is about 30 min on the same computing node. The results, which are compared in **Supplementary Material S2**, indicate that finer resolution tends to improve error metrics across all groups. The relative performance between models and the original products remains about the same.

The monthly scale considered in this work limits the number of data samples available for training which, in turn, may also affect the network performance. Future work will examine daily scales. At last, in this work we assume that the ground truth is PRISM, which itself may be subject to uncertainties present in the rain gauge data.

6. SUMMARY

High-resolution precipitation data is needed in a large number of hydrological planning and emergency management activities. Currently, a number of coarse-resolution remotely sensed products are produced on operational basis. To maximize the societal benefits of these products, some type of downscaling is necessary, which is a highly ill-posed inverse problem. This work investigates the feasibility of deep-learning-based downscaling approaches by considering different combinations of static and dynamic variables as predictors. The state-of-the-art, end-to-end deep learning (DL) framework adopted in this study allows for stacking of multi-source and multi-resolution inputs. In addition, we explore a new attention mechanism for learning multiscale features (i.e., AU-Net). The efficacy of the AU-Net is demonstrated over Central Texas, U.S., for downscaling three coarse-resolution precipitation products, namely, ERA, TRMM, and IMERG data. Results suggest that the trained AU-Net models achieve different degrees of success in downscaling the coarse-resolution products. In general, the model performance depends on the precipitation rate, and the performance is better under nominal and dry conditions than in extremely wet conditions.

Although we mainly demonstrate an attention-based, DL framework for a low-predictability study area in the U.S., the problem setup is general and the approach can be applied to other regions and at different spatial and temporal resolutions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

AS and GT conceived the original idea, discussed the results, and contributed to the final manuscript. AS developed the machine learning code and performed the computations. All authors contributed to the article and approved the submitted version.

FUNDING

AS was partly supported by Planet Texas 2050 Initiative funded by VP for Research Office at the University of Texas at Austin.

ACKNOWLEDGMENTS

The authors were grateful to the handling editor and the reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2020.536743/full#supplementary-material

REFERENCES

AghaKouchak, A., Mehran, A., Norouzi, H., and Behrangi, A. (2012). Systematic and random error components in satellite precipitation data sets. *Geophys. Res. Lett.* 39, 1–4. doi: 10.1029/2012GL051592

Arge, L., Grønlund, A., Svendsen, S. C., and Tranberg, J. (2019). Learning to find hydrological corrections. *arXiv[Preprint].arXiv:1909.07685.* doi: 10.1145/3347146.3359095

Austin Statesman (2019). Available online at: https://www.statesman.com/news/20190418/austin-region-fastest-growing-large-metro-in-nation-8-years-running-data-shows (accessed October 15, 2019).

Barnes, R. (2018). *RichDEM: High-Performance Terrain Analysis*. Technical Report. PeerJ Preprints.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., et al. (2019). Daily evaluation of 26 precipitation datasets using stage-iv gauge-radar data for the conus. *Hydrol. Earth Syst. Sci.* 23, 207–224. doi: 10.5194/hess-23-207-2019

Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Gonzalez Miralles, D., Martens, B., et al. (2017). Mswep: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.* 21, 589–615. doi: 10.5194/hess-21-589-2017

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., et al. (2013). A description of the global land-surface precipitation data products of the global precipitation climatology centre with sample applications including centennial (trend) analysis from 1901–present. *Earth Syst. Sci. Data* 5, 71–99. doi: 10.5194/essd-5-71-2013

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). Attention augmented convolutional networks. *arXiv[Preprint].arXiv:1904.09925*. doi: 10.1109/ICCV.2019.00338

Bolvin, D., Huffman, G. (2015). *Transition of 3B42/3B43 Research Product From Monthly to Climatological Calibration/Adjustment*. NASA Precipitation Measurement Missions Document. Washington, DC: NASA.

Chen, Y., Huang, J., Sheng, S., Mansaray, L. R., Liu, Z., Wu, H., et al. (2018). A new downscaling-integration framework for high-resolution monthly precipitation estimates: combining rain gauge observations, satellite-derived precipitation data and geographical ancillary data. *Rem. Sens. Environ.* 214, 154–172. doi: 10.1016/j.rse.2018.05.021

Daly, C., Taylor, G., and Gibson, W. (1997). "The PRISM approach to mapping precipitation and temperature," in *Proceedings of 10th AMS Conference on Applied Climatology*, Oct 20–23, Reno, NV. p. 1–4.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The era-interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* 137, 553–597. doi: 10.1002/qj.828

Duan, Z., and Bastiaanssen, W. (2013). First results from version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure. *Rem. Sens. Environ.* 131, 1–13. doi: 10.1016/j.rse.2012.12.002

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., and Goodess, C. M. (2006). Downscaling heavy precipitation over the united kingdom: a comparison of dynamical and statistical methods and their future scenarios. *Int. J. Climatol.* 26, 1397–1415. doi: 10.1002/joc.1318

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.* 52, 8217–8237. doi: 10.1002/2016WR019034

Hennermann, K., and Berrisford, P. (2017). *Era5 Data Documentation*. Copernicus Knowledge Base. Available online at: https://confluence.ecmwf.int/display/CKB/ERA5 (accessed November 09, 2020).

Hirschboeck, K. (1987). "Catastrophic flooding and atmospheric circulation anomalies (USA)," in *Catastrophic Flooding*. eds. L. Mayer and D. B. Nash (Boston: Allen & Unwin), p. 23–56.

Hong, Y., Adler, R. F., Negri, A., and Huffman, G. J. (2007). Flood and landslide applications of near real-time satellite rainfall products. *Nat. Hazards* 43, 285–294. doi: 10.1007/s11069-006-9106-x

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., et al. (2014). The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* 95, 701–722. doi: 10.1175/BAMS-D-13-00164.1

Hou, Y., Ma, Z., Liu, C., and Loy, C. C. (2019). Learning lightweight lane detection CNNs by self attention distillation. *arXiv[Preprint].arXiv:1908.00821*. doi: 10.1109/ICCV.2019.00110

Huffman, G. J., Adler, R. F., Bolvin, D. T., and Nelkin, E. J. (2010). "The TRMM multi-satellite precipitation analysis (TMPA)," in *Satellite Rainfall Applications for Surface Hydrology*. eds. M. Gebremichael and F. Hossain (Dordrecht: Springer). doi: 10.1007/978-90-481-2915-7_1

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., et al. (2015). NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm Theor. Basis Doc.* 4:30. Available online at: https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_ATBD_V06.pdf (accessed November 09, 2020).

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM multisatellite precipitation analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.* 8, 38–55. doi: 10.1175/JHM560.1

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2:194. doi: 10.1038/35058500

Jakob Themeßl, M., Gobiet, A., and Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.* 31, 1530–1544. doi: 10.1002/joc.2168

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353, 790–794. doi: 10.1126/science.aaf7894

Jia, S., Zhu, W., Lű, A., and Yan, T. (2011). A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam basin of China. *Rem. Sens. Environ.* 115, 3069–3079. doi: 10.1016/j.rse.2011.06.009

Karimpouli, S., and Tahmasebi, P. (2019). Segmentation of digital rock images using deep convolutional autoencoder networks. *Comput. Geosci.* 126, 142–150. doi: 10.1016/j.cageo.2019.02.003

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., et al. (2017). So, how much of the earth's surface is covered by rain gaugee? *Bull. Am. Meteorol. Soc.* 98, 69–78. doi: 10.1175/BAMS-D-14-00283.1

Kim, S., Hong, S., Joh, M., and Song, S. K. (2017). Deeprain: convLSTM network for precipitation prediction using multichannel radar data. *arXiv* 1711.02316.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv[Preprint].arXiv:1412.6980*.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., et al. (2001). The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bull. Am. Meteorol. Soc.* 82, 247–268. doi: 10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2

Li, H., Sheffield, J., and Wood, E. F. (2010). Bias correction of monthly precipitation and temperature fields from intergovernmental panel on climate change AR4 models using equidistant quantile matching. *J. Geophys. Res. Atmos.* 115, 1–20. doi: 10.1029/2009JD012882

Lin, Y., and Mitchell, K. E. (2005). "1.2 the NCEP stage II/IV hourly precipitation analyses: development and applications," in *19th Conference on Hydrology* (San Diego, CA: American Meteorological Society; Citeseer).

Long, D., Scanlon, B. R., Longuevergne, L., Sun, A. Y., Fernando, D. N., and Save, H. (2013). Grace satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas. *Geophys. Res. Lett.* 40, 3395–3401. doi: 10.1002/grl.50655

López López, P., Immerzeel, W. W., Rodríguez Sandoval, E. A., Sterk, G., and Schellekens, J. (2018). Spatial downscaling of satellite-based precipitation and its impact on discharge simulations in the Magdalena river basin in Colombia. *Front. Earth Sci.* 6:68. doi: 10.3389/feart.2018.00068

Lowrey, M. R. K., and Yang, Z. L. (2008). Assessing the capability of a regional-scale weather model to simulate extreme precipitation patterns and flooding in central Texas. *Weather Forecast.* 23, 1102–1126. doi: 10.1175/2008WAF2006082.1

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., et al. (2015). Remote sensing big data computing: challenges and opportunities. *Fut. Gen. Comput. Syst.* 51, 47–60. doi: 10.1016/j.future.2014.10.029

Mace, R. E., Chowdhury, A. H., Anaya, R., and Way, S. C. (2000). Groundwater availability of the Trinity Aquifer, Hill Country Area, Texas: numerical simulations through 2050. *Texas Water Dev. Board Rep.* 353:117. Available online at: https://www.twdb.texas.gov/publications/reports/numbered_reports/doc/R353/Report353.pdf (accessed November 09, 2020).

Mao, X., Shen, C., and Yang, Y. B. (2016). "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona. p. 2810–2818.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., et al. (2010). Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* 48, 1–34. doi: 10.1029/2009RG000314

Miao, Q., Pan, B., Wang, H., Hsu, K., and Sorooshian, S. (2019). Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. *Water* 11:977. doi: 10.3390/w11050977

Mo, S., Zhu, Y., Zabaras, N., Shi, X., and Wu, J. (2019). Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. *Water Resour. Res.* 55, 703–728. doi: 10.1029/2018WR023528

Nielsen-Gammon, J. W., Zhang, F., Odins, A. M., and Myoung, B. (2005). Extreme rainfall in Texas: patterns and predictability. *Phys. Geogr.* 26, 340–364. doi: 10.2747/0272-3646.26.5.340

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-net: learning where to look for the pancreas. *arXiv[Preprint].arXiv:1804.03999*.

Omranian, E., and Sharif, H. O. (2018). Evaluation of the global precipitation measurement (GPM) satellite rainfall products over the lower Colorado river basin, Texas. *J. Am. Water Resour. Assoc.* 54, 882–898. doi: 10.1111/1752-1688.12610

Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resour. Res.* 55, 2301–2321. doi: 10.1029/2018WR024090

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Quiroz, R., Yarlequé, C., Posadas, A., Mares, V., and Immerzeel, W. W. (2011). Improving daily rainfall estimation from NDVI using a wavelet transform. *Environ. Model. Softw.* 26, 201–209. doi: 10.1016/j.envsoft.2010.07.006

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., and Yang, M. H. (2016). "Single image dehazing via multi-scale convolutional neural networks," in *European Conference on Computer Vision Oct 8–16* (Amsterdam: Springer), 154–169.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., et al. (2011). MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Clim.* 24, 3624–3648. doi: 10.1175/JCLI-D-11-00015.1

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bull. Am. Meteorol. Soc.* 91, 1015–1058. doi: 10.1175/2010BAMS3001.1

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., van Beek, L. P., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to grace satellite data. *Proc. Natl. Acad. Sci. U.S.A.* 115, E1080–E1089. doi: 10.1073/pnas.1704665115

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3245–3250. doi: 10.1073/pnas.1222460110

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture–climate interactions in a changing climate: a review. *Earth Sci. Rev.* 99, 125–161. doi: 10.1016/j.earscirev.2010.02.004

Shen, Y., Zhao, P., Pan, Y., and Yu, J. (2014). A high spatiotemporal gauge-satellite merged precipitation analysis over china. *J. Geophys. Res. Atmos.* 119, 3063–3075. doi: 10.1002/2013JD020686

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., and Woo, W. C. (2015). "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*. eds. C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and M. Garnett (Montreal) 802–810.

Slade, R. M. Jr., and Patton, J. M. (2003). *Major and Catastrophic Storms and Floods in Texas: 215 Major and 41 Catastrophic Events From 1953 to September 1, 2002*. Technical Report. US Geological Survey.

Sorooshian, S., AghaKouchak, A., Arkin, P., Eylander, J., Foufoula-Georgiou, E., Harmon, R., et al. (2011). Advanced concepts on remote sensing of precipitation at multiple scales. *Bull. Am. Meteorol. Soc.* 92, 1353–1357. doi: 10.1175/2011BAMS3158.1

Sun, A. Y., Scanlon, B. R., Zhang, Z., Walling, D., Bhanja, S. N., Mukherjee, A., et al. (2019). Combining physically based modeling and deep learning for fusing grace satellite data: can we learn from mismatch? *Water Resour. Res.* 55, 1179–1195. doi: 10.1029/2018WR023333

Sun, A. Y., and Scanlon, B. R. (2019). How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environ. Res. Lett.* 14:073001. doi: 10.1088/1748-9326/ab1b7d

Sun, A. Y., Wang, D., and Xu, X. (2014). Monthly streamflow forecasting using Gaussian process regression. *J. Hydrol.* 511, 72–81. doi: 10.1016/j.jhydrol.2014.01.023

Sun, A. Y., Xia, Y., Caldwell, T. G., and Hao, Z. (2018). Patterns of precipitation and soil moisture extremes in Texas, US: a complex network analysis. *Adv. Water Resour.* 112, 203–213. doi: 10.1016/j.advwatres.2017.12.019

Sun, A. Y. (2018). Discovering state-parameter mappings in subsurface models using generative adversarial networks. *Geophys. Res. Lett.* 45, 11–137. doi: 10.1029/2018GL080404

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K. L. (2018). A review of global precipitation data sets: data sources, estimation, and intercomparisons. *Rev. Geophys.* 56, 79–107. doi: 10.1002/2017RG000574

Tang, G., Behrangi, A., Long, D., Li, C., and Hong, Y., (2018a). Accounting for spatiotemporal errors of gauges: a critical step to evaluate gridded precipitation products. *J. Hydrol.* 559, 294–306. doi: 10.1016/j.jhydrol.2018.02.057

Tang, G., Long, D., Behrangi, A., Wang, C., and Hong, Y. (2018b). Exploring deep neural networks to retrieve rain and snow in high latitudes using multisensor and reanalysis data. *Water Resour. Res.* 54, 8253–8278. doi: 10.1029/2018WR023830

Tang, G., Zeng, Z., Long, D., Guo, X., Yong, B., Zhang, W., et al. (2016). Statistical and hydrological comparisons between TRMM and GPM level-3 products over a midlatitude basin: is day-1 IMERG a good successor for TMPA 3B42V7? *J. Hydrometeorol.* 17, 121–137. doi: 10.1175/JHM-D-15-0059.1

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* 106, 7183–7192. doi: 10.1029/2000JD900719

Trenberth, K. E., Smith, L., Qian, T., Dai, A., and Fasullo, J. (2007). Estimates of the global water budget and its annual cycle using observational and model data. *J. Hydrometeorol.* 8, 758–769. doi: 10.1175/JHM600.1

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A. R., Nemani, R. R., and Ganguly, A. R. (2018). "Generating high resolution climate change projections through single image super-resolution: an abridged version," in *IJCAI Jul 13–19* (Stockholm), 5389–5393.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*. eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Long Beach, CA), 5998–6008.

Vila, D. A., De Goncalves, L. G. G., Toll, D. L., and Rozante, J. R. (2009). Statistical evaluation of combined daily gauge observations and rainfall satellite estimates over continental south america. *J. Hydrometeorol.* 10, 533–543. doi: 10.1175/2008JHM1048.1

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: meeting a grand challenge for monitoring earth's terrestrial water. *Water Resour. Res.* 47, 1–10. doi: 10.1029/2010WR010090

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention," in: *International Conference on Machine Learning Jul 6–15*. Lile. p. 2048–2057.

Yang, Z., Hsu, K., Sorooshian, S., Xu, X., Braithwaite, D., and Verbist, K. M. (2016). Bias adjustment of satellite-based precipitation estimation using gauge observations: a case study in chile. *J. Geophys. Res. Atmos.* 121, 3790–3806. doi: 10.1002/2015JD024540

Zhong, Z., Sun, A. Y., and Jeong, H. (2019). Predicting CO_2 plume migration in heterogeneous formations using conditional deep convolutional generative adversarial network. *Water Resour. Res.* 55, 5830–5851. doi: 10.1029/2018WR024592

Zhu, Y., Zabaras, N., Koutsourelakis, P. S., and Perdikaris, P. (2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* 394, 56–81. doi: 10.1016/j.jcp.2019.05.024

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

Definition of Abbreviations

AU-Net	Attention-based U-Net
CFSR	NCEP Climate Forest System Reanalysis
CNN	Convolutional Neural Network
DL	Deep learning
ECMWF	European Center for Medium-Range Weather Forecasts
ERA	ECMWF reanalysis
GPM	Global Precipitation Measurement
IDW	Inverse distance weighting
IMERG	Integrated Multi-satellite Retrievals for GPM
LSTM	Long short-term memory
MERRA	Modern-Era Retrospective Analysis for Research and Application
MODIS	MODerate resolution Imaging Spectroradiometer
MSWEP	Multi-Source Weighted-Ensemble Precipitation
NCEP	National Centers for Environmental Prediction
PRISM	Parameter-elevation Regressions on Independent Slopes Model
RMSE	Root mean square error
SSIM	Structural similarity index metric
TRMM	Tropical Rainfall Measuring Mission

ORIGINAL RESEARCH
published: 27 November 2020
doi: 10.3389/frwa.2020.583000

Revealing Causal Controls of Storage-Streamflow Relationships With a Data-Centric Bayesian Framework Combining Machine Learning and Process-Based Modeling

Wen-Ping Tsai[1], *Kuai Fang*[1,2], *Xinye Ji*[1,3], *Kathryn Lawson*[1] *and Chaopeng Shen*[1]*

[1] Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, United States, [2] Earth System Science, Stanford University, Stanford, CA, United States, [3] Shenzhen State High-Tech Industrial Innovation Center, Shenzhen, China

OPEN ACCESS

Edited by:
Bhavna Arora,
Lawrence Berkeley National
Laboratory, United States

Reviewed by:
Jeffrey Michael Sadler,
United States Geological Survey
(USGS), United States
Jie Niu,
Jinan University, China

***Correspondence:**
Chaopeng Shen
cshen@engr.psu.edu

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: 13 July 2020
Accepted: 30 September 2020
Published: 27 November 2020

Citation:
Tsai W-P, Fang K, Ji X, Lawson K and Shen C (2020) Revealing Causal Controls of Storage-Streamflow Relationships With a Data-Centric Bayesian Framework Combining Machine Learning and Process-Based Modeling. Front. Water 2:583000. doi: 10.3389/frwa.2020.583000

Some machine learning (ML) methods such as classification trees are useful tools to generate hypotheses about how hydrologic systems function. However, data limitations dictate that ML alone often cannot differentiate between causal and associative relationships. For example, previous ML analysis suggested that soil thickness is the key physiographic factor determining the storage-streamflow correlations in the eastern US. This conclusion is not robust, especially if data are perturbed, and there were alternative, competing explanations including soil texture and terrain slope. However, typical causal analysis based on process-based models (PBMs) is inefficient and susceptible to human bias. Here we demonstrate a more efficient and objective analysis procedure where ML is first applied to generate data-consistent hypotheses, and then a PBM is invoked to verify these hypotheses. We employed a surface-subsurface processes model and conducted perturbation experiments to implement these competing hypotheses and assess the impacts of the changes. The experimental results strongly support the soil thickness hypothesis as opposed to the terrain slope and soil texture ones, which are co-varying and coincidental factors. Thicker soil permits larger saturation excess and longer system memory that carries wet season water storage to influence dry season baseflows. We further suggest this analysis could be formulated into a data-centric Bayesian framework. This study demonstrates that PBM present indispensable value for problems that ML cannot solve alone, and is meant to encourage more synergies between ML and PBM in the future.

Keywords: Machine Learning (ML), process-based model (PBM), streamflow-storage relationships, data-centric, Bayes law, classification tree, soil texture

BACKGROUND

Basin water storage has deep connections with streamflow (Reager et al., 2014; Fang and Shen, 2017). Hence terrestrial water storage anomalies (TWSA) data could, under certain circumstances, be used to increase flood forecast lead time (Reager et al., 2015). From a physical hydrologic point of view, more water stored in a basin could mean a higher groundwater table or wetter soils which lead to more runoff source areas (Dingman, 2015). The storage-streamflow relationship is also important for predicting baseflow (Thomas et al., 2013) and related ecosystem (Poff and Allan, 1995) and water supply issues. The issue is that these relationships vary widely in space. Fang and Shen (2017) (hereafter named FS17, more description in section The Background Story) conducted an analysis of the correlation between TWSA annual extrema and different streamflow percentiles in a year, and found very interesting patterns of these correlations over the conterminous United States (CONUS). The correlations between TWSA annual extrema and high-percentile flows are strong in certain parts of the CONUS, e.g., the southeastern Coastal Plain and northern Great Plains, but are weak in other areas such as the Appalachian Plateau, northern Indiana, and Florida. *Why are there wildly different storage-streamflow relationships,* i.e., *what physical factors caused them?* Our limited understanding of this question hampered our use of water storage and groundwater data in flood forecasting.

In general, to answer "*why*" questions such as the one raised above, one could resort to two avenues: process-based models (PBMs) or data-driven analyses. They are often regarded as two separate roads that do not cross. PBMs embody our *beliefs* about how the system functions. We can use PBMs to conduct numerical experiments to assess causal relationships, as we can alter measurable physical factors to directly examine their impacts on the outputs. We typically employ a "model-centric" framework, where we (i) deploy some prior distributions or beliefs of model structures; (ii) create an ensemble of model simulations (with different parameter sets, inputs, or model structures); (iii) confront these models with observations by evaluating likelihood functions either formally or informally by visually examining the outcomes; and (iv) identify the model(s) that best describe(s) the data. It is easy to see that paradigms like model calibration (Vrugt et al., 2003) or Monte Carlo Markov Chain (Vrugt et al., 2009) fit into this framework. Moreover, numerical experiments where the modelers perturb model physics on an *ad-hoc* basis (e.g., Maxwell and Condon, 2016; Shen et al., 2016; Ji et al., 2019) could also be placed in this framework. Potential issues with this framework are that it can be both subjective and inefficient, as many competing hypotheses remain un-tested. The priors are often based on one's own beliefs, and one needs to throw a huge amount of simulations to capture the plausible model structure. It has been argued that hydrologic models are necessarily degenerate (Nearing et al., 2016) and even sampling exhaustively from its parameter distribution does not capture the whole possible model space.

In contrast to PBMs, various interpretable machine learning approaches could be used to generate possible explanations, or "hypotheses" in machine learning language (Russell and Norvig, 2009), of an observed behavior. For example, the weights from linear regression could inform us of the relative importance of factors. Classification and regression tree (CART) analysis (Breiman et al., 1984; Mitchell, 1997), which iteratively separates data points based on predictors and their thresholds, is another explanatory tool that has often been employed. For example, Verhougstraete et al. (2015) used the first level split in a CART model to draw the conclusion that septic systems are the primary driver of fecal bacteria levels in 64 US rivers. An advantage of machine learning approaches is that they are highly efficient to execute compared to PBMs, and the models they generate are already consistent with data. They also carry the appeal of relying less on subjective assumptions and model choices.

However, the "Achilles heel" for machine learning as an explanatory tool is arguably their inability to distinguish between causal and associated relationships. If we had a large enough training dataset that covered all possible combinations of physical factors, machine learning should theoretically be able to extract the causal factor. However, we are limited by the combinations that exist in the real world and for which we have data, posing limits on the power of data. Naturally, one might wonder if PBMs' strength in causality analysis could be exploited to complement machine learning algorithms.

Recently, there have emerged increasing interest in combining physics with data-driven models. One could adopt a variety of methods loosely termed "physics-guided machine learning" (PGML) or "theory-guided machine learning" (Ganguly et al., 2014; Karpatne et al., 2017; Jia et al., 2019; Read et al., 2019; Yang et al., 2019), such as modifying the loss function to accommodate physical constraints (Jia et al., 2019) or pre-training a ML model using PBM outputs (Jia et al., 2018). These constructive ideas have made ML more robust and have enriched our means of investigations. Nevertheless, PGML frameworks have not taken advantage of PBM's ability to conduct experiments and assess causes and effects. Here we propose that the evaluation of competing hypotheses could be accomplished by running numerical experiments with a PBM to utilize the physics encoded in the PBM (**Figure 2**), as an example of the alternative research avenue proposed earlier (Shen et al., 2018). We then compare the probability of each hypothesis and reject those with low probability. Bayes' law allows information from different sources to be merged in a sequential manner given some evidence. In the context of hydrology (Beven and Binley, 1992; Kavetski et al., 2006; Raje and Krishnan, 2012; Viglione et al., 2013), the gist is that a likelihood function based on (oftentimes subjective) assumptions of error or data distribution replaces the conditional probability of observing a data point given model parameters. While such kinds of likelihood functions have been well-established, Bayes' law itself is quite generic and not restricted to this use. An opportunity exists to explore using Bayes' law to use process-based models to provide a quantification of the likelihood. Because this framework first starts with data, we call it a data-centric framework, in contrast to a conventional model-centric Bayesian framework where a model's inputs and parameters are perturbed and the posterior probability of each realization is calculated. We will use the storage-streamflow question to showcase the effectiveness of this framework and help us understand the main controlling factors of streamflow in the Susquehanna River basin to inspire best modeling practices. This work is a first exploration of this particular method of coupling data-driven hypotheses with process-based modeling capabilities, and by no means do we indicate this method is optimal or the most efficient.

In the following, we first provide some background for the case study of streamflow-storage correlations and the competing hypotheses that explain them (section The Background Story). Then we describe the process-based model and the experimental setup (sections Process-Based Hydrologic Model and Competing Hypotheses and the Implementation of Perturbation Experiments). We make sure the model produces

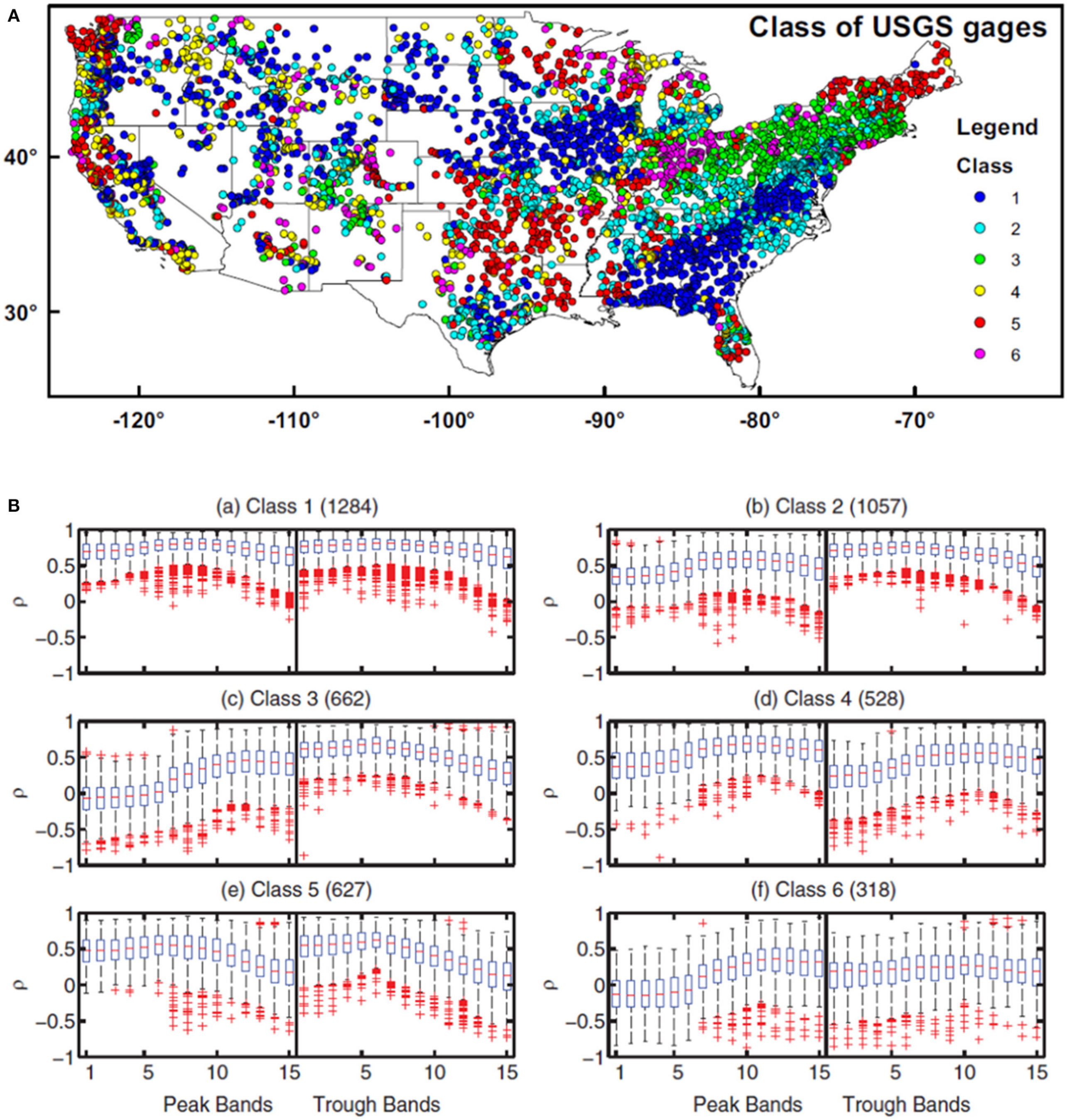


FIGURE 1 | Class map **(A)** and boxplots of the SSCS for Class #1 to Class #6 **(B)**. The boxes contain 25–75% percentiles, and the crosses are those considered outliers (Reprinted from FS17 with permission).

reasonable hydrologic dynamics (section Performance of the Physically-Based Model), and then finally we use the perturbation experiments to test the competing hypotheses from ML (section Testing Competing Hypotheses).

THE BACKGROUND STORY

The Storage-Streamflow-Correlation Spectrum

In FS17 we introduced a hydrologic signature termed the Storage-Streamflow-Correlation Spectrum (SSCS), which quantifies how water storage is correlated with streamflow at different flow regimes. SSCS is the collection of Pearson's correlation coefficients (R) between annual extrema (peaks or troughs) of the terrestrial water storage anomalies (TWSA) and different streamflow percentiles (15 percentiles extracted are: {0.5%, 1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%}) in a window around the extrema for the same basin. The correlations are calculated on an annual scale, using the water year (the 12-month period from October 1 through September 30 of the following year). The study period of FS17 is from 1 October 2002 to 31 September 2012. Treating each flow percentile as a "band," we obtained a correlation "spectrum." The SSCS gives a snapshot of the correlations across all bands, as compared to previous studies that focused only on high flow regimes.

If streamflow is disconnected from storage, e.g., when most rainfall runs off or evaporates directly without entering the subsurface, the system would exhibit low correlation between flows and storage during peak flows. Generally, the high-flow bands have lower R than low-flow bands because peak streamflows result from large storms whose magnitudes are poorly correlated to water storage. In contrast, if groundwater exerts a significant influence over streamflow, we expect the correlation to be higher. A high correlation between TWSA peaks and low flows indicates a long system memory: when such basins receive plenty of precipitation in the wet season, the excess storage is carried over the seasons and is reflected in low flows. Therefore, SSCS gives us a window of observation into how varied surface and subsurface hydrologic systems function. Please see FS17 for more details.

When applying the SSCS over the conterminous United States (CONUS), a large variety of SSCS behaviors emerged (FS17). To facilitate our interpretation, we clustered these responses into 6 different classes using K-means and a distance measure (the Euclidean distance in the SSCS space). The correlation values for different classes and the spatial distribution of classes are shown in **Figure 1**. We can clearly observe regional clusters and spatial gradients in the SSCS patterns. Class #1 was described as "full-spectrum responsive" since it had the highest correlations and the smallest variability across all SSCS bands. Class #1 concentrated on the southeast Coastal Plain and northern Great Plains. Class #2 and #3 catchments had weaker SSCS values and were concentrated along the northern Appalachian Plateau. For Class #3, in peak-TWSA bands, streamflow-storage correlation was low for flow percentiles below 20%, but higher for percentiles above 60%; in trough-TWSA bands, there were high streamflow-storage correlations at percentiles below 60%, but correlations were a little lower for high streamflow percentiles (80% above). Class #2 can be considered a transition type between class #1 and class #3.

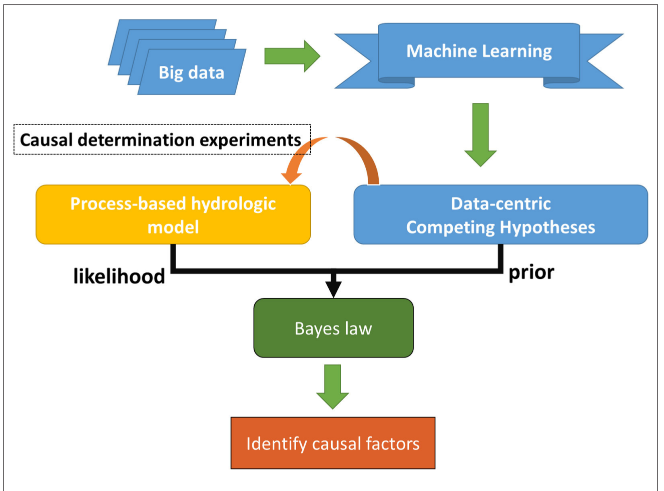

Big data
Machine Learning
Causal determination experiments
Process-based hydrologic model
Data-centric Competing Hypotheses
likelihood
prior
Bayes law
Identify causal factors

FIGURE 2 | A framework for integrating PBM and machine learning.

Explaining the Controls of SSCS

When observing the large spatial gradients of SSCS classes over CONUS in **Figure 1**, one cannot help asking, "***what causes the SSCS behavior to differ between Appalachia and the Coastal Plain?***", which was the central question of this study. FS17 employed a CART analysis to learn simple and interpretable decision rules (the split criteria and thresholds) from the data. Focusing on the differences between basins in Appalachia (Appalachian Plateau, Piedmont, and Valley and Ridge physiographic provinces) and basins in the southeastern Coastal Plain, FS17 trained a specific CART model to predict the distances of basins to class centers in SSCS space. They used a number of predictors including the aridity index, depth to bedrock, rainfall seasonality, and the fraction of precipitation as snow (supporting information **Table S1** in FS17). In other words, they asked what factors made the two clusters of basins different in terms of their SSCS patterns. From this *ad hoc* tree, the CART-based model automatically identified soil thickness (RockDep), obtained by merging soils-survey-based depth to bedrock with bedrock depth simulated by a geomorphological model (Pelletier et al., 2016) as the main difference between the two types of streamflow-storage correlation patterns.

The problem of learning an optimal CART is that a CART is not robust. This can be mitigated by training multiple trees in an ensemble as in the random forest (RF) algorithm (Ho, 1995), where the features and samples are randomly sampled with replacement. The RF generalizes from the CART and provides an estimation of probability. While RF models are more robust and can be used to infer probabilities, they are more difficult for humans to interpret.

While the RockDep explanation does make physical sense, it could be dangerous to take this hypothesis as the truth. First, even though soil thickness appeared to be the stronger explanatory model, there could be other slightly weaker but nonetheless valid models. We have yet to explore what would happen if we slightly alter the training dataset. Because we rely on available data, the results may be dependent on a few data points that critically cover certain parts of the input space. However, such critical data points may happen to be missing in our training data; the robustness of the model has not been established.

METHODS AND DATASETS

To answer our central question, here we propose a novel framework that combines the strengths of machine learning and process-based modeling. In this framework, machine learning first presents competing hypotheses and assigns them prior probabilities. Then, we construct numerical perturbation experiments with a process-based model to implement and test the hypotheses (**Figure 2**). The testing of the hypotheses could be achieved by visual examination of the outcome of the experiments, or via a more quantitative Bayesian approach.

Study Area—Susquehanna River Basin (SRB)

The Susquehanna River (watershed area: 71,225 km^2) is a major river located in the northeastern and mid-Atlantic United States (**Figure 3A**), which has historically been the source of many instances of flooding damage along the main river floodplains (Yarnal et al., 1997; May, 2011). The basin spans the physiographic provinces of the Appalachian Plateau, Piedmont, Valley and Ridge, and Coastal Plain regions. In general, most of the northern subbasins of the SRB consist of mountains mantled by thin soils which are mostly thinner than 2 m (**Figure 3B**). We show the SSCS behaviors of 13 randomly selected subbasins in the Susquehanna River Basin. We found that the all 13 stations in the Susquehanna River Basin belong to either Class #2 or Class #3 (**Figure 3C**, the original pattern of SSCS from 13 USGS gauge stations is similar to class #3).

We further chose 4 subbasins (**Figure 3A**), namely, the Otselic River basin (OR), the Pine Creek basin (PIN), the Raystown Branch Juniata River basin (RAY), and the Octoraro Creek basin (OCT) in the south to create process-based hydrologic models. Both soils survey data and global modeled soil thickness data were used to parameterize soil thickness: in most of the basin where the bedrock is within the limit of the soils survey depth (1.52 m), the RockDep attribute in SSURGO (NRCS, 2010) was used; outside of these areas, we used the average soil and sedimentary layer thickness from Pelletier et al. (2016), which has global coverage with 1 km resolution. Among the subbasins modeled, OR and PIN are headwater subbasins in the Appalachian Plateau, RAY is a headwater subbasin in the Valley and Ridge physiographic division, and OCT is near the Coastal Plain region. OCT has a visibly larger soil thickness.

Process-Based Hydrologic Model

To be able to conduct causal experiments, we employed the Process-based Adaptive Watershed Simulator coupled with the Community Land Model (PAWS+CLM) (Shen and Phanikumar, 2010; Shen et al., 2013, 2014, 2016; Ji et al., 2015, 2019; Niu et al., 2017; Ji and Shen, 2018; Fang et al., 2019). First introduced in Shen and Phanikumar (2010), the PAWS model was coupled to the Community Land Model (CLM) (Collins et al., 2006; Dickinson et al., 2006; Oleson et al., 2010; Lawrence et al., 2011) which describes the land surface and vegetation dynamics (Shen et al., 2013). The PAWS model has been used to explain the relative importance of different controlling processes on hydrologic and ecosystem dynamics. CLM incorporates comprehensive physical and biogeochemical processes including vapor and momentum transfer, surface radiative transfer, soil heat transfer, freeze-thaw phase changes, and biochemical photosynthesis, as well as plant carbon and nitrogen cycles (Shen et al., 2014). PAWS+CLM inherits the land surface processes from CLM, including surface energy fluxes, ET, vegetation growth, and carbon cycling, while solving physically-based conservative laws for flow processes including 2D overland flow, quasi-3D subsurface (soil and groundwater) flow, vectorized channel networks, and the exchanges among these domains. The flow module starts with throughfall, stemflow, and snowmelt as the precipitation inputs, and converts the CLM-computed evapotranspiration term into a sink. The surface water layer is divided into the flow domain, which can flow laterally, and the ponding domain, which exchanges with the main soil column and does not circulate laterally. The flow domain water is routed downstream as overland flow, described by a diffusive wave equation (DWE). Infiltrated water is governed by the Richards equation. Water reaching the phreatic water table may move laterally, as described by Dupuit-Forchheimer flow in an unconfined aquifer. 1D columns of vertical soil flow are coupled to the saturated lateral flow at the bottom. The confined aquifers below are described by a 3D saturated groundwater flow equation. The channel flow is governed by DWE in a 1D cascade network. More information about PAWS can be found in Shen et al. (2016).

Configuration of the Hydrologic Model

In this study, a 1,040 × 1,040 m horizontal grid was used to discretize the domain. Precipitation and climate forcing data used in PAWS+CLM were obtained from the North American Land Data Assimilation System (NLDAS) (Mitchell, 2004). Information from the Soil Survey Geographic Database (SSURGO) was used to provide initial values for the soil properties. In PAWS+CLM, we extracted topographic information from the National Elevation Dataset (30 m) to parameterize the river bed elevations, and used the mean elevation to parameterize the gridcell elevation (Shen et al., 2016). The climatic forcing datasets that come from NLDAS are on an hourly basis.

The channel network is represented by an explicit, vectorized channel network for larger rivers and the implicit, gridded overland flow for smaller headwater streams. As an advance of PAWS+CLM, the channel network topology is now established based on the National Hydrography Dataset Plus Version 2 (NHDPlus V2) shapefiles. In NHDPlus V2, each segment is encoded with a unique ID number and the downstream ID. Combing through this connectivity information, our pre-processing package traces the rivers from downstream to upstream and records the river distances of each segment. The

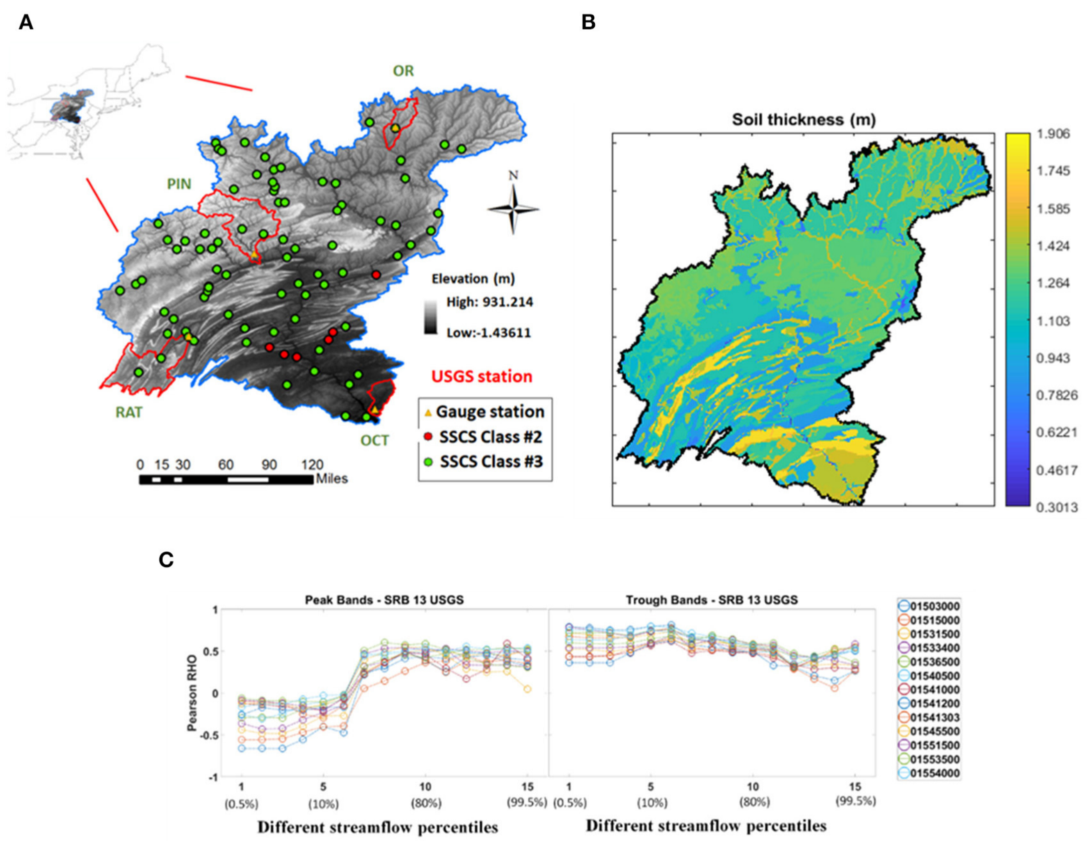


FIGURE 3 | Study area, Susquehanna River Basin (SRB). Main class of SRB observation data is class #3 and #2 in FS17. **(A)** Study area, Susquehanna River Basin. Among the subbasins modeled, OR and PIN are headwater subbasins in the Appalachian Plateau, RAY is a headwater subbasin in the Valley and Ridge physiographic division, and OCT is near the Coastal Plain region. **(B)** Soil thickness. **(C)** SSCS in Susquehanna River Basin for 13 USGS gauge stations (station numbers in legend).

available channels from NHD are vastly greater than what can be explicitly represented in the vectorized channel network in the model. In previous work, the selection of the explicitly modeled streams was manual. We have now implemented an automatic selection procedure: our pre-processing utility iteratively selects the longest rivers from the candidate pool built from NHDPlus V2, so that the total selected river length satisfies a prescribed river density (river length : basin area). Based on these explicitly represented rivers, we then establish a network structure, recording names of the streams, network topology, upstream/downstream nodes in the hierarchy, boundary condition types (headwater, inflow, connecting streams, or outflow), tributaries, and locations of confluences. For each explicitly modeled river, the discretization procedure evenly distributes the river polyline into river cells. We then overlay the river cell with high resolution DEM and groundwater data, extracting information, e.g., bank and bed elevation (inferred through regional regression equation), during discretization (Shen et al., 2016).

In PAWS the soil water retention and unsaturated hydraulic conductivity are parameterized using the van Genuchten formulation. To obtain spatially distributed van Genuchten parameters, we incorporated a range of well-established pedotransfer functions (PTFs) (Guber et al., 2009) and the Rosetta (Schaap et al., 2001) program which employs a hierarchy of PTFs, ranging in complexity from a soil textural lookup table to algorithms based on Artificial Neural Networks (ANN). We also exported soil textural information (sand, clay, and silt percentages), bulk density, and water contents from soil horizon data from the SSURGO database (NRCS, 2010) into Rosetta, wherever they were available. Rosetta was then used to predict van Genuchten parameters, and the results were subsequently read into PAWS. Normally, we chose the "best possible model" option in Rosetta. The SSURGO database contains fine resolution (1:24,000 map scale) soil type maps, which are encoded as "map unit" keys (mukey). A mukey value serves as an index key to the SSURGO relational databases that detail the characteristics of that soil type. A mukey may contain several "soil components," each taking up a certain fraction of the map unit. Every component then describes the vertical soil horizons and their depths.

The runtime of PAWS-CLM for 18 years of simulation in an SRB subbasin (OR) is about 4 h on a machine with CPU. Even with the help of the pedotransfer functions, process-based hydrologic model parameters need to be further adjusted or calibrated. The SRB is large, and it is difficult to perform calibration for the whole basin. We thus defined our objective function as the mean of the Nash-Sutcliffe model efficiency coefficients (Nash and Sutcliffe, 1970) for the four subbasins. This way, the resulting parameter set may not produce the best achievable performance for each subbasin, but presents a balance between them for the whole basin. Model performance was evaluated against USGS streamflow records.

Competing Hypotheses and the Implementation of Perturbation Experiments

To identify potential competing hypotheses, we first ran a CART analysis, which combines classification tree and linear regression algorithms, for both southeastern and Appalachian basins with multiple random seeds and randomized removal of training data points (basins). A classification tree is used to split data points with a binary decision rule, and a linear regression is used to predict the distance to clusters' centroids. The runtime of the CART-based model was 3∼5 s for both the southeastern and Appalachian areas. Then we further ran RF analysis with an expanded list of attributes. With the CART-based model, we considered all basin physiographic parameters that were deemed as important for SSCS in FS17, including: RockDep, sand, slope, soil bulk density, watershed percent agriculture, watershed percent developed, and standard deviation of elevation. In FS17, we employed sand and clay as representatives for soil texture and removed silt, since they add up to one. In the present analysis we also followed this practice. We then implemented changes in these factors via perturbing corresponding parameters in the process-based model. Essentially, we first replaced the values of these factors in the SRB by their counterparts from the Southeastern CONUS, and ran experiments to determine their individual impacts on the SSCS classes. We also considered the combinatory impacts of these factors by altering them at the same time.

Some climatic variables such as relative humidity, annual precipitation, and fraction of precipitation as snow could overtake as the top-level split, but are ignored in the manual CART analysis because we are interested in the relative impacts of physical basin parameters. We nonetheless included them in the RF model and PBM perturbation experiments by replacing forcing data on the SRB with those from some locations in the Coastal Plain region, to compare their impacts with the physical basin parameters.

One of the important physical basin parameters is soil thickness. The difference in average soil thickness between the thinly-mantled Appalachian basins and their southeastern neighbors is about 30 m. Hence, for the perturbation experiments, we added 30 m of soil thickness to each subbasin of SRB.

The second factor of importance is soil texture (sand or clay percentages). We replaced the soil van Genuchten parameters in the SRB with those from soil classes that were randomly selected from two survey areas in the Southeast. One survey area had many map units, each of which had many soil component and horizons. We randomly selected one soil horizon from each survey area (GA603 and GA632). The soil van Genuchten parameters were obtained using the Rosetta program. We also selected two SSURGO horizons where one had the maximum sand content (FL131) and the other one had the minimum sand content (TN081). Hence, in these experiments, the SRB basins effectively are given the same soil texture as the Coastal Plain region. The characteristics of soil texture of these four SSURGO entries are shown in **Table 1** (sand, silt, and clay percentages). One could note that basins in the Coastal Plain region have much more sandy soils, and thus have high infiltration capacity.

The third factor to be analyzed was the terrain slope. We examined the difference between the slopes of the southeastern CONUS (Class #1) and SRB, which are <10% and ∼30%, respectively. Thus, we implemented an experiment where the terrain slope was reduced by 80%, by changing the digital elevation data that were inputs to the data pre-processor (Shen et al., 2014) of PAWS+CLM. 80% was chosen because after this treatment, the average slopes of the SRB basins were similar to those in the Coastal Plain region.

Besides single factor experiments, we also evaluated how multiple factors interacted to impact hydrologic fluxes. After implementing the numerical experiments, we recalculated the SSCS from each perturbed simulation. The total simulated water stored in the soil column and groundwater in the model was used as the water storage, while streamflow was extracted from the simulated daily outflow from each subbasin.

The Data-Centric Bayesian Learning Framework

The effects of the ML hypotheses can be demonstrated solely by visualizing the results from the experiments. However, as an exploratory step, here we also propose a quantitative, data-centric Bayesian framework to integrate data and the results from the modeling experiments. Essentially, the machine learning provides the prior, and the numerical experiments compute a likelihood for a factor being the causal factor. The two probabilities can be integrated using Bayes' law.

Here, we define y as the observed patterns and F as the list of perturbations of the "process parameters", i.e. physical factors whose effects can be represented by perturbing our PBM. In the present example, F can take one of three values in {"soil thickness," "soil texture," "slope"}. When F is equal to "soil thickness," the setup of the PBM experiment is to increase soil

TABLE 1 | The characteristics of alteration of soil texture.

Soil category	Sand percentage (%)	Silt percentage (%)	Clay percentage (%)
GA603	86	4	10
GA632	43	40	17
FL131	85	10	5
TN081	21	55	25
SRB Average	32.8	51.7	15.5

thickness, while leaving soil texture and slope untouched. We can then identify the factors *causing* the differences in observed patterns between instances using Bayes' law:

$$P\left(F|y\right) = \frac{L(y|F)P(F)}{P(y)} \tag{1}$$

where $P(F)$ is the prior probability of the process parameters being the cause of the observed differences between instances, to be obtained from the pure data-driven analysis (more below), $L(y|F)$ is the likelihood that, after making the process perturbations in F, the differences in patterns in y are observed, $P\left(y\right) = \int L(y|F)P(F)dF$ is the marginalized probability, and $P\left(F|y\right)$ is then the probability that, given the evidence with the model experiments, F is the causal factor for the observed differences. In the Bayesian analysis here, we only consider the top three individual factors as potential values for F, and do not consider parameter interactions.

More specifically for this case, we start from basins that are by default of SSCS class #2 and #3 in the SRB, and ask whether a change in one of the physical factors could turn them into class #1. Therefore, $P(F)$ is the prior probability of each process perturbation, and was calculated as the frequency that F appears as the first level split in the RF model trained to predict the distance to the class center #1; $L(y|F)$ is the likelihood function for the perturbed model to produce class #1 basins. This likelihood was assessed using a Gaussian Mixture Model (GMM), which is a generalization from K-means clustering. Instead of predicting one class membership, the GMM generates a fuzzy membership for all classes. Our GMM used the clustering results of FS17, including the clusters' centroids, clusters' covariances, and the fraction of data points belonging to each class (more details of the GMM are in **Appendix A**). The marginalized probability, $P(y)$, was computed by integration.

The definitions of $P(F)$, which uses model visit frequency, may seem unestablished. However, in the world-shocking event where AlphaGo defeated the Go world champion, the algorithm selected the most visited move during its Monte Carlo tree search as its actual action (Silver et al., 2016). Their choice, also reliant on model visit frequency, also seemed informal, but it performed marvelously well. Our choices were based on the current best tool we have given the overall objective of this paper.

RESULTS AND DISCUSSIONS

In this section, we first show the limitations of CART analysis and ML in general, and present multiple competing hypotheses from ML. After demonstrating the performance of the PAWS+CLM model for the Susquehanna River basin, we show results from the perturbation experiments. Finally, we put those results in the exploratory Bayesian framework and examine its usefulness.

The Robustness of CART and the Competing Hypotheses

While soil thickness was the most frequent factor that can predict the SSCS difference between class #1 and class #3 basins (**Figures 4A,B**), we found that soil texture (**Figures 4C,D** display the result for sand percentage), and terrain slope (**Figures 4E,F**) are competing hypotheses. The CART experiments with 20 different random seeds showed that there is a 75% chance that RockDep was selected as the top-level split, followed by Sand and then Slope. From the RF modeling, RockDep, Sand, and Slope have 21%, 17%, and 2% chances to be selected as the top-level split, respectively, with the other remaining chances mostly taken by climatic variables. The performance of these alternative models are weaker than soil thickness, but the difference, especially between soil thickness and soil texture, was not big enough to warrant confident rejection. These competing hypotheses exist because terrain slope, soil texture, and depth to bedrock covary in space. As we go from Appalachia (Appalachian Plateau, Piedmont, Valley and Ridge) to the Coastal Plain, simultaneously the terrain flattens, the soil texture becomes more sandy, and the soil thickness increases substantially.

Besides random seeds, we also ran experiments with reduced training data points to examine the robustness of CART. We found that the frequency of the first-level criterion of the classification tree changed significantly when we randomly removed ∼22% of the data. Moreover, in the extreme case, if we purposefully removed as few as 7 data points with the lowest sand percentages out of 693 total data points, the most important variable would change from "RockDep" to "Sand."

These results all suggest that the CART analysis is not robust. CART is indeed problematic; however, this is not just an issue with CART, but more generically an issue with the statistical power of the data. It can be argued that there is not enough statistical power in the data to differentiate between the causal and the coincidental factors. Geoscientists are opportunistic in the sense that we can only examine basins with the combinations of land use, geology, soil texture, and slope that naturally exist in the world and have been, or are, under study. It is not be hard to imagine missing some critical combinations which would lead to erroneous conclusions.

More importantly, from these results, we extracted three factors that are treated as competing hypotheses that explains the main difference in SSCS between the Appalachian basins and their Southeast neighbors: soil thickness (RockDep), soil texture (Sand, Silt, or Clay), and terrain (Slope). Other basin parameters such as soil bulk density and land use have very low importance and can be ignored in later analysis. We then implemented changes in these factors in the process-based model to examine their impacts on the SSCS.

Performance of the Physically-Based Model

The daily observed USGS streamflow and simulated flow for a period of 18 years (2000–2017) were compared in **Figure 5**. The model had decent performance for streamflow simulation, especially within the baseflow and low flow periods (**Figure 5**), and captures the long-term streamflow pattern as well as some extreme high flows. The Nash-Sutcliffe model efficiency coefficient is not as high as in some of our previous applications (e.g., Shen and Phanikumar, 2010; Shen et al.,

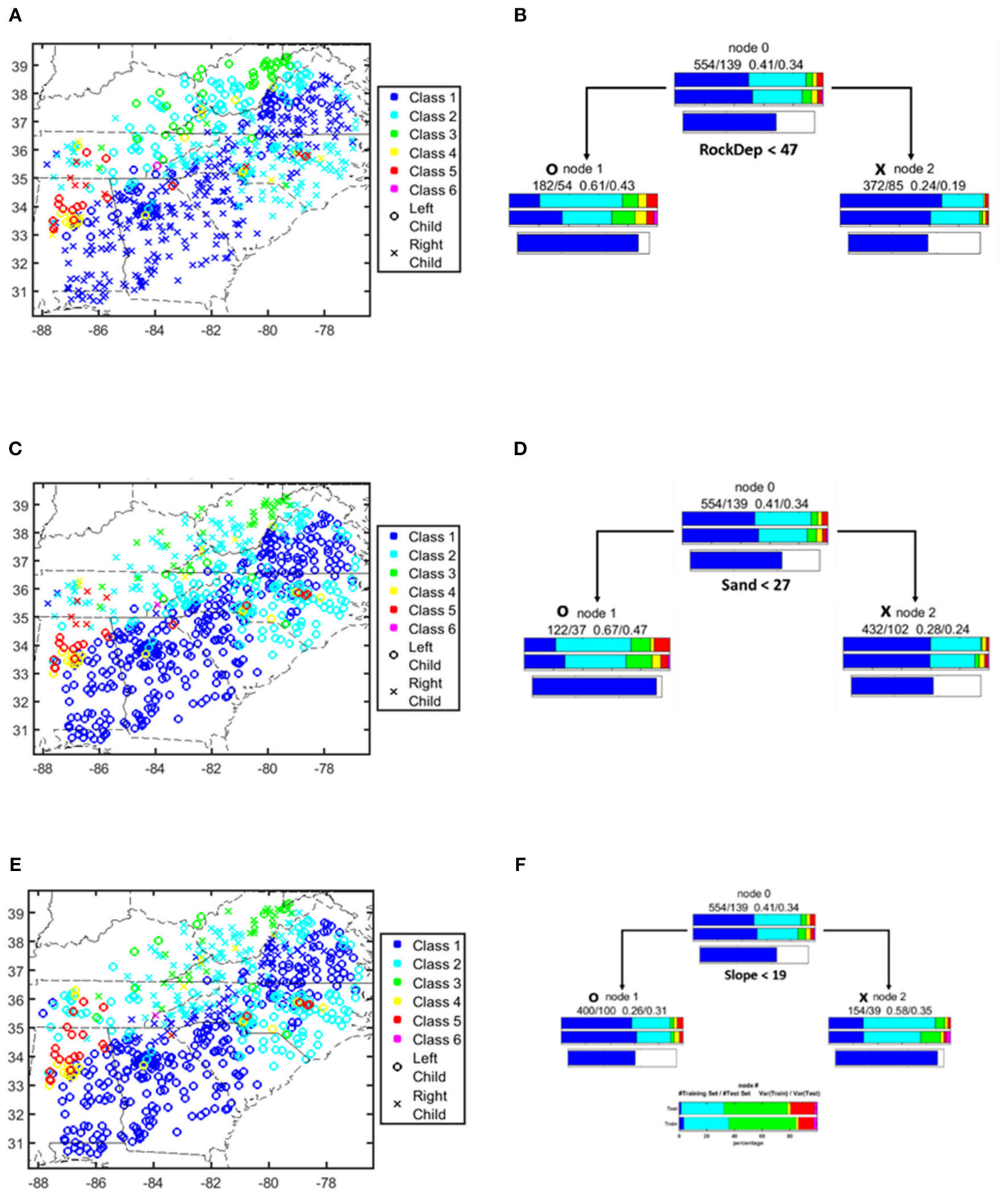


FIGURE 4 | A one-level classification tree model picks up soil thickness (RockDep) as the main difference between two types of storage-stream flow correlation patterns compared to other physical factors like soil texture (result for sand percentage shown here, Sand) and terrain (Slope) (Reproduced from FS17 with permission). **(A,C,E)** Southeastern region of CONUS. Color indicates SSCS class, symbols indicate tree nodes for physical factor (RockDep, Sand, Slope). **(B,D,F)** One-level *ad hoc* trees to predict class #1 in **(A,C,E)** via physical factor (RockDep, Sand, Slope).

2014; Niu et al., 2017), due to the compromise in the 4 subbasins' parameter calibration. While the largest dam on the Susquehanna River, the Conowingo Dam, is downstream from our gage, there are other smaller dams in the basin that could have contributed to the mismatch. In addition, our experiences have indicated that NLDAS precipitation often underestimates the peak storms, leading to an underestimation of peaks. As the main focus of the paper is

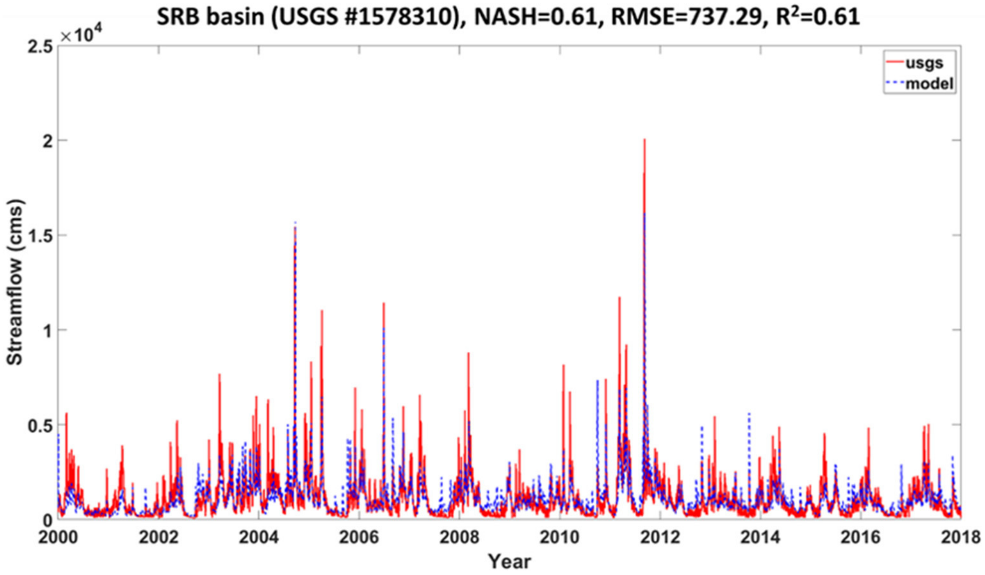


FIGURE 5 | Model streamflow simulation of whole SRB streamflow simulation. The red solid line indicates the USGS measured streamflow, and the blue dashed line indicates the model's simulated flow.

not streamflow prediction, our calibration of the model is not extensive.

Testing Competing Hypotheses

It is easy to observe the impacts of soil thickness on the SSCS curves extracted from the default and perturbed simulations (**Figure 6**). On this figure, we colored experiments by whether they do have thicker soil implemented (adding 30 m to the soil thickness, shown in blue) or do not (shown in red). All four basins have similar patterns. The default SSCS (red x) curves are similar to SSCS classes #2 and #3 of FS17 (except the trough band of PIN, which is similar to Class #4), in that they have low correlations in peak-storage-low-flow bands, medium correlations in peak-storage-high-flow bands, and low correlations in trough-storage bands. These patterns all indicate a limited system memory; the water storage in the wet season has no impact on baseflow later in the water year. When we increased the soil thickness, the correlations in peak-storage-low-flow increased substantially, indicating that the annual-scale system memory had been enhanced. Except for the OCT subbasin, there is a clear separation between the red and blue points.

On the other hand, when soil texture was modified from the default (red x) into those from the Southeast (red plus, asterisk, square, and diamond), SSCS barely fluctuated, and results based on these southeastern soil textures were clustered closely with the default simulation. We could see that soil texture has a small impact: FL131 (red square) appears to encourage higher correlations across the spectrum as compared to the others. The notable soil texture characteristics were that GA603 had a high sand percentage (most were higher than 70%); GA632 had high sand and high silt percentages (summation of both were higher than 70%); FL131 was high in sand percentage (most were higher than 80%); and TN081 was high in silt percentage (most were higher than 50%). However, the magnitude of the impact of soil texture was not comparable to that of the soil thickness. According to the likelihood value calculated by the GMM, with all default parameters, OR belongs to Class #2 (highest probability, almost 1) and PIN belongs to Class #2 with a likelihood of 0.75 (**Figures 7A,B**). In contrast, all experiments with "thick soil" had SSCS class #1. Some parameter interaction can be observed, but its effects were minor compared to the impact of soil thickness.

From the experiments where we replaced forcing data in the SRB with those from the Coastal Plain region, we found the impacts of climate on SSCS classes (or GMM likelihoods) to be small (data not shown here). In fact, going from Appalachia in the North to the Coastal Plain in the South, we saw a lower fraction of precipitation as snow, which should have reduced storage-streamflow relationships, but this effect ran counter to the observation of higher correlations between storage and streamflow in the south. Apparently, the effects of climatic variables were not as strong as the physical basin parameters, and were also coincidental factors. Hence, they were not further examined.

The Data-Centric Bayesian Inference Results

According to the Bayesian inference framework in Equation 1, the soil thickness factor had the highest posterior probability (**Table 2**). Although soil texture also had a prior that was comparable to that of soil thickness, experiments that only perturbed soil had very low likelihood functions, lowering its posterior to almost zero. Terrain slope had a lower prior (although it was higher than other physical factors which were

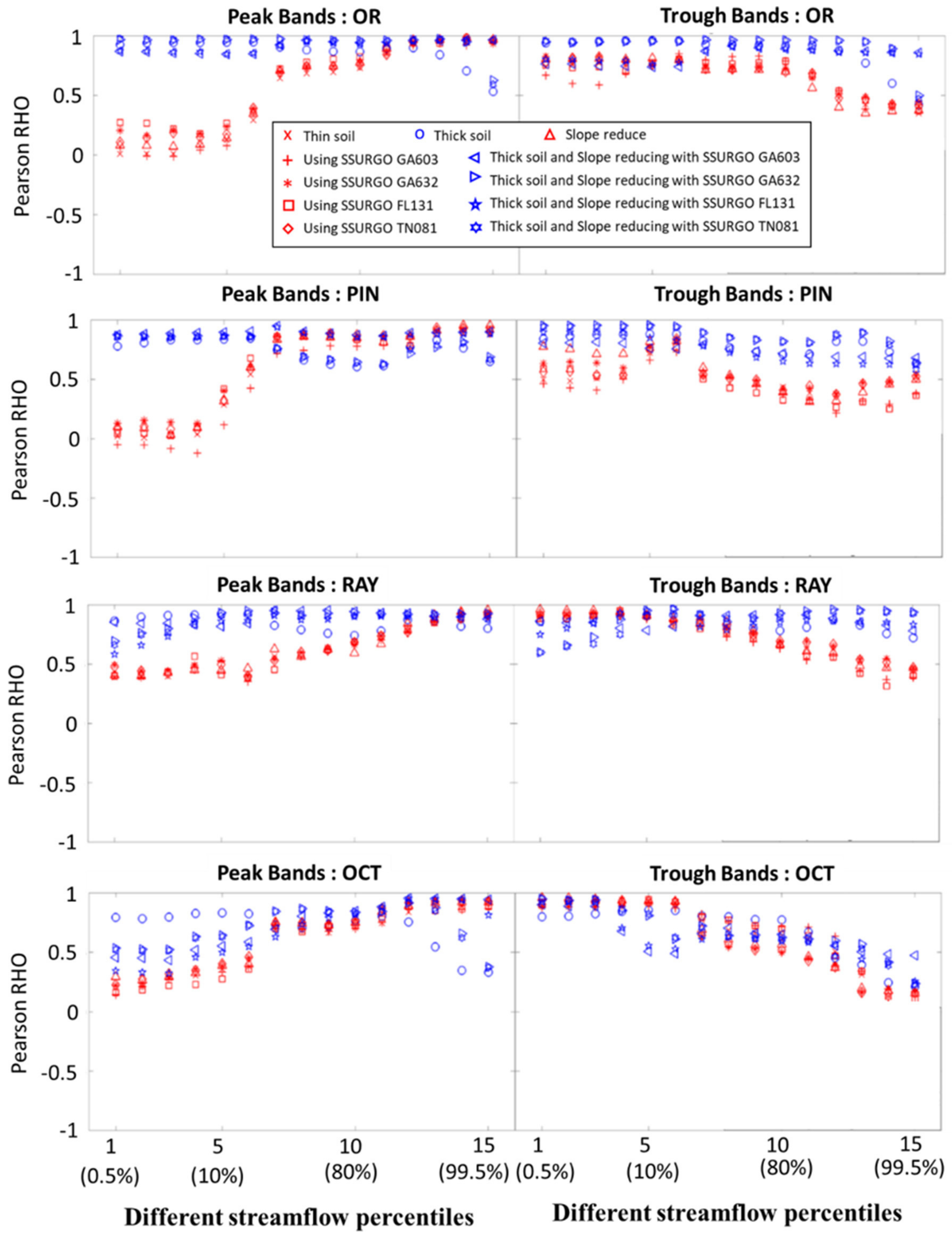
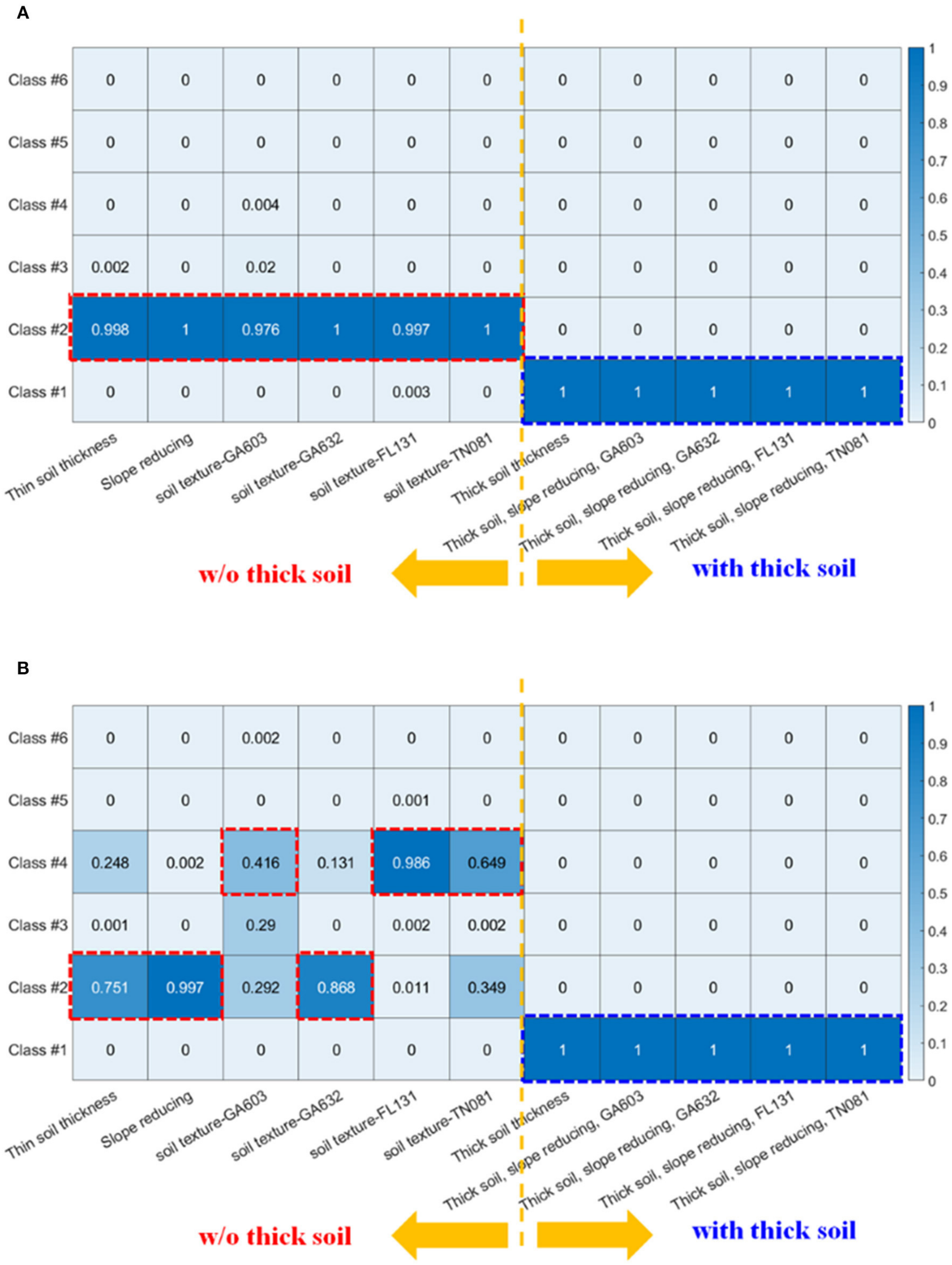


FIGURE 6 | SSCS extracted from the numerical experiments. “Thin soil” is the default simulation with SRB-default parameters.

examined but not mentioned here), and its likelihood was also low, indicating that it was only a coincidental factor, not causal.

These results unequivocally support soil thickness as the causal factor of SSCS differences between Appalachian basins and those on the southeastern Coastal Plain, whereas soil texture and slope were merely coincidental factors. It is notable that the PBM was needed to break the practical tie between the priors of soil texture and soil thickness. From these results, we can



	Thin soil thickness	Slope reducing	soil texture-GA603	soil texture-GA632	soil texture-FL131	soil texture-TN081	Thick soil thickness	Thick soil, slope reducing, GA603	Thick soil, slope reducing, GA632	Thick soil, slope reducing, FL131	Thick soil, slope reducing, TN081
Class #6	0	0	0	0	0	0	0	0	0	0	0
Class #5	0	0	0	0	0	0	0	0	0	0	0
Class #4	0	0	0.004	0	0	0	0	0	0	0	0
Class #3	0.002	0	0.02	0	0	0	0	0	0	0	0
Class #2	0.998	1	0.976	1	0.997	1	0	0	0	0	0
Class #1	0	0	0	0	0.003	0	1	1	1	1	1

	Thin soil thickness	Slope reducing	soil texture-GA603	soil texture-GA632	soil texture-FL131	soil texture-TN081	Thick soil thickness	Thick soil, slope reducing, GA603	Thick soil, slope reducing, GA632	Thick soil, slope reducing, FL131	Thick soil, slope reducing, TN081
Class #6	0	0	0.002	0	0	0	0	0	0	0	0
Class #5	0	0	0	0	0.001	0	0	0	0	0	0
Class #4	0.248	0.002	0.416	0.131	0.986	0.649	0	0	0	0	0
Class #3	0.001	0	0.29	0	0.002	0.002	0	0	0	0	0
Class #2	0.751	0.997	0.292	0.868	0.011	0.349	0	0	0	0	0
Class #1	0	0	0	0	0	0	1	1	1	1	1

FIGURE 7 | The likelihood function L(y|F) as calculated by GMM in different PAWS+CLM experiments. Deeper blue color highlights higher probability. Here, we only show the **(A)** OR and **(B)** PIN subbasins, but the other 2 subbasins have similar results (**Appendix B**).

TABLE 2 | Calculations of the data-centric Bayesian inference framework for three factors.

OR basin		***P*(*F*)**	***L*(*y*|*F*)** ***(Class 1)***	***P*(*y*)** **(P1*L1+P2*L2+P3*L3)**	***P* (*F*|*y*)** ***(Class 1)***
Thickness	30 m addition	0.21 (P1)	0.99999 (L1)	0.21012	***1.00***
Slope	80% reduction	0.02 (P2)	0.00001 (L2)		0.00
Soil texture	Different SSURGO	0.17 (P3)	0.00070 (L3)		0.00
PIN basin		***P*(*F*)**	***L*(*y*|*F*)** ***(Class 1)***	***P*(*y*)**	***P* (*F*|*y*)** ***(Class 1)***
Thickness	30 m addition	0.21	0.99997	0.21001	***1.00***
Slope	80% reduction	0.02	0.00020		0.00
Soil texture	Different SSURGO	0.17	0.00004		0.00

The remaining P(F) was mostly taken by climatic variables. Bold font indicates the factor with the highest posterior probability.

conclude that in general, systems with large soil thickness have longer memory, allowing water from the recharge season to accumulate, which thus impacts the baseflow in the hot summers. Although more sandy soil could allow for more infiltration and hence mildly boost storage-streamflow correlations, its impact was apparently not comparable to that of soil thickness. This contrast was automatically highlighted by the Bayesian framework proposed here.

Further Discussion

In this case study, ML allowed us to focus on only three factors prior to running any numerical experiments. If we were to run the hydrologic model to assess all of the 11 factors analyzed in CART, assuming 3 levels for each factor, 3^{11} model runs would be needed, but in this analysis we only ran 11 jobs. Not only does this provide savings of computational power and time, but also means that we need to objectively confront our PBMs with the identified ML hypotheses. If the PBM at hand is not able to represent the effects of these factors, one needs to take note and either refine the PBM or select a different one. Because of the target, inputs, training data, and other aspects of ML still needing to be defined by humans, it is not unbiased, and fairness in artificial intelligence is a big topic (Zou and Schiebinger, 2018). However, as long as the initial ML problem is posed inclusively, ML can be relatively impartial compared to only using one PBM and starting only from expert-conceived hypotheses. The PBM was also critically important here, allowing us to study causal relationships and nuances of parameter interactions, where data may not be sufficient for complete analysis via ML.

The proposed framework is very different from that of physics-guided machine learning (PGML) (Ganguly et al., 2014; Jia et al., 2019) in that it utilizes established PBMs, which are valuable assets which the geoscience community has accumulated over the past decades, as the backbone of the analysis, whereas PGML relies on ML algorithms as the backbone. While one can easily encode simple principles such as mass and energy conservation in the loss function for PGML, it will be quite difficult to similarly express the complex physical processes and cross-domain interactions encoded in complex PBMs. Another PGML method is to pre-train a ML network with outputs from the PBM; in the future it will certainly be interesting to compare these methods in terms of their capability and clarity of finding explanations.

The proposed data-centric Bayesian framework is raised here for the first time, and is thus only exploratory. It requires the definition of a prior (from ML), a proper PBM, a likelihood function (calculated by the GMM), and a marginalization strategy. Upon proper definition of the prior and likelihood functions, this framework can be autonomously executed. The prior is obtained purely from data analysis of GRACE and streamflow data while the posterior mostly depends on the assumed model dynamics which were built from physical laws such as the Richards equation, diffusive flow equations, and ecosystem equations. Each one of these choices can have alternatives, and may involve arbitrary decisions that lead to debates. We fully recognize that the choices we made could be improved in the future. However, our goal here was to highlight the value of both PBM and ML, and to inspire exploration into the diverse ways that both approaches can be coupled together for the advancement of knowledge.

Here we used an interpretable machine learning method (classification tree) for illustrative purposes, essentially to obtain a parameter importance ranking and an estimate of a prior. Other methods such as linear regression, support vector machines, or deep learning neural networks could also be used to provide the prior. Time series deep learning-based models (Fang et al., 2017, 2018; Fang and Shen, 2020; Feng et al., 2020), have also emerged and are transforming hydrology (Shen, 2018), but they are less interpretable. The main purpose of the ML algorithm is to obtain a parameter importance ranking and an estimate of prior. Besides algorithm-specific methods such as layer-wise relevance propagation (Bach et al., 2015), many model-agnostic methods, e.g. permutation feature importance (Fisher et al., 2019) or forward/backward feature selection, exist to obtain parameter importance rankings and priors. On the other hand, interpretability is not necessarily required if the purpose is to autonomously discover knowledge, e.g., if the purpose is for an AI agent to reduce uncertainty in the framework of active learning (Settles, 2012). The only true requirements are that the hypothesis generated by the ML algorithm can be translated into a PBM configuration and used to make perturbations, and that the likelihood of those configurations can be evaluated.

CONCLUSIONS

Here we have proposed a Bayesian framework that combines machine learning and process-based modeling to overcome limitations of both approaches. In this framework, machine learning is first used to generate competing hypotheses that are consistent with existing data. These hypotheses are subsequently implemented as perturbed process-based model simulations, which help to distinguish between causal and coincidental factors. This framework can be executed by a program and could be regarded as giving PBMs to machine learning as diagnosis tools. ML has its limitations regarding robustness, the statistical power of limited data, and causal reasoning, but it allows us to rapidly focus on several competing hypotheses and limit our subjective bias when choosing a model.

We tested the framework using the example of inferring the physical factor that controls storage-streamflow correlation behaviors across the gradients from Appalachia to the Coastal Plain. Although machine learning suggested that soil thickness and soil texture have similar prior probabilities of being the causal factor, the PBM experiments unequivocally supported soil thickness. This example highlights the value of the PBM in the era of big data, and promotes an alternative ML-PBM integration methodology to physics-guided machine learning, as it works with complicated, established PBMs.

DATA AVAILABILITY STATEMENT

The details of the Gaussian Mixture model and a map of physiographic provinces are available in **Supplemental Materials**. Streamflow data can be downloaded from the U.S. Geological Survey Water Data for the Nation website (http://dx.doi.org/10.5066/F7P55KJN). GRACE TWSA data can be downloaded from GRACE monthly mass grids (https://grace.jpl.nasa.gov/data/get-data/). CART, RF, and GMM codes can be downloaded from Scikit-learn (https://scikit-learn.org/stable/). Our CART code and the PAWS-CLM code are available at doi: 10.5281/zenodo.4019836. The input to one of the subbasins is available at http://water.engr.psu.edu/shen/Data/PIN.zip.

AUTHOR CONTRIBUTIONS

CS conceived the study. WP-T ran the experiments, prepared the visualization, and wrote the initial draft. KF and XJ assisted with the experiments. CS and KL edited the manuscript.

FUNDING

This work was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under contract DE-SC0016605.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2020.583000/full#supplementary-material

REFERENCES

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Beven, K., and Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Process.* 6, 279–298. doi: 10.1002/hyp.3360060305

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, G. B., Bretherton, C. S., Carton, J. A., et al. (2006). The community climate system model version 3 (CCSM3). *J. Clim.* 19, 2122–2143. doi: 10.1175/JCLI3761.1

Dickinson, R. E., Oleson, K. W., Bonan, G., Hoffman, F., Thornton, P., Vertenstein, M., et al. (2006). The community land model and its climate statistics as a component of the community climate system model. *J. Clim.* 19, 2302–2324. doi: 10.1175/JCLI3742.1

Dingman, S. L. (2015). *Physical Hydrology (Third)*. Long Grove, IL: Waveland Press.

Fang, K., Ji, X., Shen, C., Ludwig, N., Godfrey, P., Mahjabin, T., et al. (2019). Combining a land surface model with groundwater model calibration to assess the impacts of groundwater pumping in a mountainous desert basin. *Adv. Water Resourc.* 130, 12–28. doi: 10.1016/j.advwatres.2019.05.008

Fang, K., Pan, M., and Shen, C. (2018). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Trans. Geosci. Remote Sens.* 57, 2221–2233. doi: 10.1109/TGRS.2018.2872131

Fang, K., and Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resourc. Res.* 53, 8064–8083. doi: 10.1002/2016WR020283

Fang, K., and Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *J. Hydrometeorol.* 21, 399–413. doi: 10.1175/JHM-D-19-0169.1

Fang, K., Shen, C., Kifer, D., and Yang, X. (2017). Prolongation of SMAP to spatio-temporally seamless coverage of continental US using a deep learning neural network. *Geophys. Res. Lett.* 44, 11030–11039. doi: 10.1002/2017GL075619

Feng, D., Fang, K., and Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resourc. Res.*56:e2019WR026793. doi: 10.1029/2019WR026793

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81. Available online at: https://jmlr.org/papers/v20/18-760.html

Ganguly, A. R., Kodra, E. A., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S. N., et al. (2014). Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. *Nonlinear Process. Geophys.* 21, 777–795. doi: 10.5194/npg-21-777-2014

Guber, A. K., Pachepsky, Y. A., Th van Genuchten, M., Simunek, J., Jacques, D., Nemes, A., et al. (2009). Multimodel simulation of water flow in a field soil using pedotransfer functions. *Vadose Zone J.* 8, 1–10. doi: 10.2136/vzj2007.0144

Ho, T. K. (1995). "Random decision forests," in *Proceeding ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition* (Montreal, QC: IEEE).

Ji, X., Lesack, L., Melack, J. M., Wang, S., Riley, W. J., and Shen, C. (2019). Seasonal and inter-annual patterns and controls of hydrological fluxes in an Amazon floodplain lake with a surface-subsurface processes model. *Water Resourc. Res.* 55, 3056–3075. doi: 10.1029/2018WR023897

Ji, X., and Shen, C. (2018). The introspective may achieve more: enhancing existing geoscientific models with native-language structural reflection. *Comput. Geosci.* 110, 32–40. doi: 10.1016/j.cageo.2017.09.014

Ji, X., Shen, C., and Riley, W. J. (2015). Temporal evolution of soil moisture statistical fractal and controls by soil texture and regional groundwater flow. *Adv. Water Resourc.* 86, 155–169. doi: 10.1016/j.advwatres.2015.09.027

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., et al. (2018). "Physics guided recurrent neural networks for modeling dynamical systems: application to monitoring water temperature and quality in Lakes," in *8th International Workshop on Climate Informatics*.

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., et al. (2019). "Physics guided RNNs for modeling dynamical systems: a case study in simulating lake temperature profiles," in *Proceedings of the 2019 SIAM International Conference on Data Mining* (Calgary: Society for Industrial and Applied Mathematics Publications), 558–566. doi: 10.1137/1.9781611975673.63

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* 29, 2318–2331. doi: 10.1109/TKDE.2017.2720168

Kavetski, D., Kuczera, G., and Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resourc. Res.* 42. doi: 10.1029/2005WR004376

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., et al. (2011). Parameterization improvements and functional and structural advances in version 4 of the community land model. *J. Adv. Model. Earth Syst.* 3. doi: 10.1029/2011MS00045

Maxwell, R. M., and Condon, L. E. (2016). Connections between groundwater flow and transpiration partitioning. *Science* 353, 377–380. doi: 10.1126/science.aaf7891

May, J. (2011). *On Ancient Susquehanna River, Flooding's a Frequent Fact*. Associated Press. Available online at: http://cumberlink.com/news/local/on-ancient-susquehanna-river-flooding-s-a-frequent-fact/article_ee769266-db2f-11e0-945d-001cc4c002e0.html (accessed November 01, 2020).

Mitchell, K. E. (2004). The multi-institution North American land data assimilation system (NLDAS): utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.* 109:D07S90. doi: 10.1029/2003JD003823

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Nash, J. E., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — a discussion of principles. *J. Hydrol.* 10, 282–290. doi: 10.1016/0022-1694(70)90255-6

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., and Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *J. Hydrometeorol.* 17, 745–759. doi: 10.1175/JHM-D-15-0063.1

Niu, J., Shen, C., Chambers, J., Melack, J. M., and Riley, W. J. (2017). Interannual variation in hydrologic budgets in an Amazonian watershed with a coupled subsurface—land surface process model. *J. Hydrometeorol.* 18, 2597–2617. doi: 10.1175/JHM-D-17-0108.1

NRCS (2010). *SSURGO Soil Survey Geographic Database*. Natural Resources Conservation Service; Natural Resources Conservation Service, United States Department of Agriculture. Available online at: https://data.nal.usda.gov/dataset/soil-survey-geographic-database-ssurgo (accessed November 01, 2020).

Oleson, K., Lawrence, D. M., Bonan, G. B., Flanner, M., Kluzek, E., Lawrence, P., et al. (2010). *Technical Description of version 4.0 of the Community Land Model (CLM)*. Boulder, CO: NCAR Technical Note, NCAR/TN-478+STR.

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., et al. (2016). A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *J. Adv. Model. Earth Syst.* 8, 41–65. doi: 10.1002/2015MS000526

Poff, N. L., and Allan, J. D. (1995). Functional organization of stream fish assemblages in relation to hydrological variability. *Ecology* 76, 606–627. doi: 10.2307/1941217

Raje, D., and Krishnan, R. (2012). Bayesian parameter uncertainty modeling in a macroscale hydrologic model and its impact on Indian river basin hydrology under climate change. *Water Resourc. Res.* 48. doi: 10.1029/2011WR011123

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 Resourc. Res.* 55, 9173–9190. doi: 10.1029/2019WR024922

Reager, J. T., Thomas, A., Sproles, E., Rodell, M., Beaudoing, H., Li, B., et al. (2015). Assimilation of GRACE terrestrial water storage observations into a land surface model for the assessment of regional flood potential. *Remote Sens.* 7, 14663–14679. doi: 10.3390/rs71114663

Reager, J. T., Thomas, B. F., and Famiglietti, J. S. (2014). River basin flood potential inferred using GRACE gravity observations at several months lead time. *Nat. Geosci.* 7, 588–592. doi: 10.1038/ngeo2203

Russell, S., and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach 3rd Edn*. Upper Saddle River, NJ: Prentice Hall.

Schaap, M. G., Leij, F. J., and Th van Genuchten, M. (2001). Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J. Hydrol.* 251, 163–176. doi: 10.1016/S0022-1694(01)00466-8

Settles, B. (2012). "Active learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning,* eds R. J. Brachman, W. W. Cohen, and T. G. Dietterich (Williston, VT: Norgan and Claypool), 1–114. doi: 10.2200/S00429ED1V01Y201207AIM018

Shen, C. (2018). A trans-disciplinary review of deep learning research and its relevance for water resources scientists. *Water Resourc. Res.* 54, 8558–8593. doi: 10.1029/2018WR022643

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS opinions: incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci.* 22, 5639–5656. doi: 10.5194/hess-22-5639-2018

Shen, C., Niu, J., and Fang, K. (2014). Quantifying the effects of data integration algorithms on the outcomes of a subsurface–land surface processes model. *Environ. Model. Softw.* 59, 146–161. doi: 10.1016/j.envsoft.2014.05.006

Shen, C., Niu, J., and Phanikumar, M. S. (2013). Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface—land surface processes model. *Water Resourc. Res.* 49, 2552–2572. doi: 10.1002/wrcr.20189

Shen, C., and Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resourc.* 33, 1524–1541. doi: 10.1016/j.advwatres.2010.09.002

Shen, C., Riley, W. J., Smithgall, K. M., Melack, J. M., and Fang, K. (2016). The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics. *Water Resourc. Res.* 52, 880–902. doi: 10.1002/2015WR018086

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Thomas, B. F., Vogel, R. M., Kroll, C. N., and Famiglietti, J. S. (2013). Estimation of the base flow recession constant under human interference. *Water Resourc. Res.* 49, 7366–7379. doi: 10.1002/wrcr.20532

Verhougstraete, M. P., Martin, S. L., Kendall, A. D., Hyndman, D. W., and Rose, J. B. (2015). Linking fecal bacteria in rivers to landscape, geochemical, and hydrologic factors and sources at the basin scale. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10419–10424. doi: 10.1073/pnas.1415836112

Viglione, A., Merz, R., Salinas, J. L., and Blöschl, G. (2013). Flood frequency hydrology: 3. *A bayesian analysis. Water Resourc. Res.* 49, 675–692. doi: 10.1029/2011WR010782

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resourc. Res.* 39. doi: 10.1029/2002WR001746

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlinear Sci. Numerical Simul.* 10, 273–290. doi: 10.1515/IJNSNS.2009.10.3.273

Yang, X., Barajas-Solano, D., Tartakovsky, G., and Tartakovsky, A. M. (2019). Physics-informed CoKriging: a gaussian-process-regression-based multifidelity method for data-model convergence. *J. Comput. Phys.* 395, 410–431. doi: 10.1016/j.jcp.2019.06.041

Yarnal, B., Johnson, D. L., Frakes, B. J., Bowles, G. I., and Pascale, P. (1997). The flood of '96 and its socioeconomic impacts in the susquehanna river basin. *J. Am. Water Resourc. Assoc.* 33, 1299–1312. doi: 10.1111/j.1752-1688.1997.tb03554.x

Zou, J., and Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature* 559, 324–326. doi: 10.1038/d41586-018-05707-8

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 in Water

REVIEW
published: 29 January 2021
doi: 10.3389/frwa.2020.562304

Toward Urban Water Security: Broadening the Use of Machine Learning Methods for Mitigating Urban Water Hazards

Melissa R. Allen-Dumas, Haowen Xu, Kuldeep R. Kurte and Deeksha Rastogi*

Computational Urban Sciences Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Climate Change Science Institute, Oak Ridge, TN, United States

Due to the complex interactions of human activity and the hydrological cycle, achieving urban water security requires comprehensive planning processes that address urban water hazards using a holistic approach. However, the effective implementation of such an approach requires the collection and curation of large amounts of disparate data, and reliable methods for modeling processes that may be co-evolutionary yet traditionally represented in non-integrable ways. In recent decades, many hydrological studies have utilized advanced machine learning and information technologies to approximate and predict physical processes, yet none have synthesized these methods into a comprehensive urban water security plan. In this paper, we review ways in which advanced machine learning techniques have been applied to specific aspects of the hydrological cycle and discuss their potential applications for addressing challenges in mitigating multiple water hazards over urban areas. We also describe a vision that integrates these machine learning applications into a comprehensive watershed-to-community planning workflow for smart-cities management of urban water resources.

Keywords: urban water security, hazard mitigation, machine learning, watershed modeling, integrated water resource management

OPEN ACCESS

Edited by:
Chaopeng Shen,
Pennsylvania State University (PSU),
United States

Reviewed by:
Andrea Cominola,
Technical University of Berlin,
Germany
Xi Chen,
University of Cincinnati, United States

****Correspondence:***
Haowen Xu
xuh4@ornl.gov

Specialty section:
This article was submitted to
Water and Hydrocomplexity,
a section of the journal
Frontiers in Water

Received: *14 December 2020*
Accepted: *15 May 2020*
Published: *29 January 2021*

Citation:
Allen-Dumas MR, Xu H, Kurte KR and Rastogi D (2021) Toward Urban Water Security: Broadening the Use of Machine Learning Methods for Mitigating Urban Water Hazards. Front. Water 2:562304. doi: 10.3389/frwa.2020.562304

1. INTRODUCTION

A recent United Nations report projects that 60% of the world's total population will live in cities by the year 2030 (U.N., 2018). This highly-urbanized population will face vulnerability to water-related hazards in many ways. For example, the combined effect of natural changes and human intervention on the landscape can lead to flooding, drought, and morphologic instabilities (e.g., stream erosion and instability, erosion, and sedimentation at structures) in and around urban areas, as well as deterioration of water quality, riverine ecology, and natural habitats (Crossman et al., 2013; Krajewski et al., 2016). Because of the accelerated pace of anthropogenic activity, hazard frequency, and intensity is exacerbated requiring immediate delivery of science-based solutions for mitigation, resilience, and adaptation that can be quickly deployed in any hazard-prone area. Mitigating these urban water hazards is challenging for watershed management and the urban planning community (Eriksson et al., 2015) due to the following hydro-complexities. First, these hazards exist in a variety of forms (e.g., floods, droughts, increased soil erosion, and water pollution) and are associated with multiple urban risks (e.g., property inundation and infrastructure failure, water shortage, landslide, and eco-habitat deterioration) (Carson et al., 2018). Second, these urban water hazards may occur separately or in a multi-hazard chain (Kappes et al., 2012; Komendantova et al., 2014), in which the occurrence of one hazard (e.g., urban flooding) may trigger another hazard (e.g., bank erosion and landslide). Third, the occurrences of different urban water hazards are connected through the flow of the water and watershed processes over a range of spatial scales (Souchère et al., 2010; Santelmann et al., 2019), pressing the need for multiscale mitigation strategies that target hazard drivers at both watershed and urban neighborhood scale (Bertolotto et al., 2007; Xu et al., 2019b).

Given these challenges, a holistic approach to water security is articulated by Ait Kadi and Arriens (2012), as one that produces a world in which each community has access to enough water for social and economic development, and for ecosystems in and beyond those communities; and where those communities are protected from floods, droughts, landslides, erosion, and waterborne diseases (Carson et al., 2018; Aboelnga et al., 2019). Additionally, ensuring urban water security is a complex endeavor, as it involves dynamic processes and requires the interaction and participation of multiple planning actors (stakeholders, resource managers, and policy makers) to safeguard the integrity and security of urban water systems and assets in a continuous, physical, and legal manner. Subsequently, these actors must formulate policies and make investments using robust, adaptive, and accessible strategies that balance the socioeconomic and ecological benefits and urban sustainability with the cost of mitigation measures and management practices, and increase the resilience and preparedness of urban communities against extreme weather and natural disasters (Medema et al., 2014; Carson et al., 2018).

Fundamentally, these methods must have the capability of identifying and assessing the risk of multiple interconnected urban water hazards simultaneously (Kappes et al., 2012; Komendantova et al., 2014). Further, these methods must include system-based techniques for providing generalized predictions and acquiring unseen data in order to obtain reliable and accurate depictions of both current and future states of water resources in both urban areas and their associated watersheds. The projections and updates provided through these techniques must be easy to interpret and to understand, so that researchers, decision makers, and communities can readily obtain useful insights that support the planning of urban water resources, including the mitigation of existing hazards and the prevention of future hazards (Carson et al., 2018; Zaidi et al., 2018).

To fulfill these management needs, comprehensive disaster management frameworks are proposed to promote the collaborative planning and management of water, land, and related resources (Selin and Chevez, 1995; Emerson et al., 2012). These frameworks are developed to reduce the risk of multiple water hazards equitably without compromising the sustainability of vital ecosystems. Examples of these frameworks include Integrated Water Resources Management (IWRM), Adaptive Management (AM), and the Ecosystem Approach (EA) (Cardwell et al., 2009; Dörendahl, 2013; Palmer et al., 2013; Carson et al., 2018). In general, these frameworks entail a series of planning processes that can be categorized into four major stages (Yu et al., 2018; Sun and Scanlon, 2019):

1. Long-term planning and mitigation
2. Early warning and prediction of hazards
3. Rapid response and rescue
4. Recovery and restoration.

Within the long-term planning and mitigation stage, we summarize here a list of common planning processes from several planning frameworks (Yoe and Orth, 1996; NRCS, 2003; USEPA, 2012), and we address machine learning (ML) methods for application to these processes throughout the paper. These steps are as follows:

1. Identification and assessment of multi-hazard risk in urban water systems.
2. Determination of the objectives of urban water planning and hazard mitigation.
3. Inventory of useful data resources that can define urban water hazards and risks, indicate the performances of existing urban water systems, and reflect the current state of the urban water system and the watershed to which it pertains.
4. Identification, evaluation, and selection of Best Management Practices (BMPs) from a variety of planning alternatives for water quality improvement, stormwater management, and erosion controls (NRCS, 2011; USEPA, 2018).
5. Evaluation of the performance and effectiveness of the implemented plan by examining information and monitoring data collected from pilot studies.
6. Identification, evaluation, and selection of proposed modifications for ongoing or existing plans and implementation schedules based on the future scenarios of urban water.

Despite the usefulness of these planning directives, the implementation of these processes is sophisticated and faces both methodological and technical challenges. Methodological challenges are associated with the long-term planning and mitigation processes and include: (a) assessing the multi-hazard risk and vulnerability of a municipal water system (Kappes et al., 2012; Jetten et al., 2014; Lambert, 2014), and (b) optimizing the selection of the BMPs from a variety of mitigation alternatives based on multiple criteria and objectives (FHWA, 2000). Technical challenges are associated with the implementation of multiple planning processes. One of the major technical challenges is related to the discovery and integration of a large volume of interdisciplinary data and simulation models (Adamala, 2017), which is essential for supporting the multi-hazard risk assessment in the long-term planning and mitigation process, as well as for informing rapid response and rescue during a hazardous event. These information resources can provide data-driven and model-driven insights for informing the current and future state of urban water systems and watersheds. Another major technical challenge is related to the accurate and timely prediction of hazardous events, which help facilitate early warning and prevention of hazard.

Conventionally, these challenges are approached using domain models and human justification of decision-makers, and therefore require computation- and labor-intensive efforts for coupling multiple models and investigating the underlying physical processes of different hazards. In recent decades, developments in advanced ML techniques has offered a more time efficient method for overcoming these challenges in an intelligent manner. Many review papers have enumerated ML and big data applications for enhancing various water resources management related applications and hydrological analysis (Adamala, 2017; Holzbecher et al., 2019) and for mitigating a specific water hazard, such as flooding (Mosavi et al., 2018), water pollution (Haghiabi et al., 2018), and erosion (Abdulkadir et al., 2019). In this paper, we explore and discuss benefits and potential opportunities of the ML applications for enhancing the mitigation of multiple urban water hazards. Herein, we review a selection of successful studies that apply various ML techniques and hybrid modeling techniques (i.e., the fusion of ML methods with process-based domain models) to overcome challenges encountered by different planning processes for integrated urban water management. Hybrid models are a mixture of inductive (data-driven) and deductive (process-based) approaches (Goldstein and Coco, 2015; Hajigholizadeh et al., 2018; Frame, 2019) and are referred to by Goldstein and Coco (2015) as the use of empiricisms built from ML in process-based models. Other researchers (e.g., Karpatne et al., 2016) approach hybrid modeling from the opposite direction—as "theory-guided data science," in which data analysis, given sufficient grounding in physical principles, can represent causative relationships among parameters.

Additionally, we provide a vision for ways in which ML techniques can be used to facilitate different processes in the planning framework for the future. Different from previous review articles that focus on the machine learning application in the water management sector (Sun and Scanlon, 2019; Chen et al., 2020), we review innovative and application-ready machine learning solutions to facilitate urban water hazard mitigation from the practical aspect of addressing technical and methodological challenges in water resources and disaster management frameworks. The target audience of this paper includes watershed management authorities (WMAs), urban and regional planners, and research professionals in the water resources management sectors. To retrieve the relevant literature in this field that applies various ML techniques for urban water management, we conducted searches using tools such as Google scholar (https://scholar.google.com) and Scopus (https:www.scopus.com). **Figure 1** shows the result of the query: *("Random Forest" OR "Artificial Intelligence" OR "ANN" OR "Support Vector Machine" OR "ANN" OR "Artificial Neural Network" OR "Neural Network" OR "SVM" OR "Machine Learning") AND ("water management" OR "water resources management" OR "watershed management" OR "watershed planning" OR "urban water systems" OR "multi-hazard" OR "water hazard" OR "flood disaster" OR "water pollution") AND [EXCLUDE (PUBYEAR, 2020)]*. We executed the query for years 1999–2019, and excluded year 2020. The above query retrieved a total of 46,145 documents from Scopus such that either article title, list of keywords or abstract satisfies the query. It is clear from **Figure 1A** that there is a significant growth in ML based approaches for water related areas such as water management and urban water hazards. **Figure 1B** shows the top four scientific journals which receive research on ML application to water related areas. The graph in **Figure 1B** also confirms the increasing trends in the applications of ML techniques in water management and hazards.

Among the thousands of literature identifies from the Scopus, we select a handful of studies that are either published in recent years or are most relevant to and practical for improving specific processes and steps in the generic hazard mitigation stages and long-term water planning frameworks that are discussed early in the introduction section. We also consider the diversity and novelty of the machine learning techniques during the selection of studies for more detailed reviews and discussions. Based on the challenge and planning process targeted by these studies, we divide our review here into the following sections. Section 2 reviews the predictive data analytics powered by various ML techniques that help planners predict water-related hazards (e.g., flood, drought, water quality, and soil erosion and sediment transport). Multiple applications of hybrid modeling are also discussed in this section. Additionally, a subsection reviewing innovative combinations of ML and remote sensing technologies for disaster management is included, as remote sensing technologies are increasingly applied for improving the discovery and extraction of useful information and features (e.g., land use and land cover, flood inundation extent, and reservoir storage from satellite imagery) that are critical for early warning of hazards and rapid response and rescue during hazardous events (Hodgson et al., 2010). Section 3 presents the ML applications for the identification and assessment of water-related multi-hazard risks and vulnerability (e.g., building inundation, infrastructure failure, and economic loss) in urban water systems. In section 4, we review a few case studies that utilize ML algorithms to optimize the selection of urban BMPs, which can improve long-term planning and mitigation and recovery and restoration processes. Finally, in section 5, we present our vision for the application of next-generation ML techniques to efficient generation of mitigation strategies in response to urban water hazards. ML methods and their performance as applied to each issue are summarized in **Table 1**.

2. EARLY WARNING AND PREDICTION OF URBAN WATER HAZARDS

The capability to predict timely and accurate occurrence, intensity, and frequency of natural hazards is essential to every planning process that develops disaster preparedness and response to ensure public safety and mitigate unfavorable consequences associated with hazardous events (de Goyet et al., 2006). Traditionally, hydrological processes that contribute to water-related hazards have been analyzed using probabilistic modeling and physics based modeling approaches. The probabilistic approaches are devised to estimate the available stock over relatively short future time horizons (Philbrick and

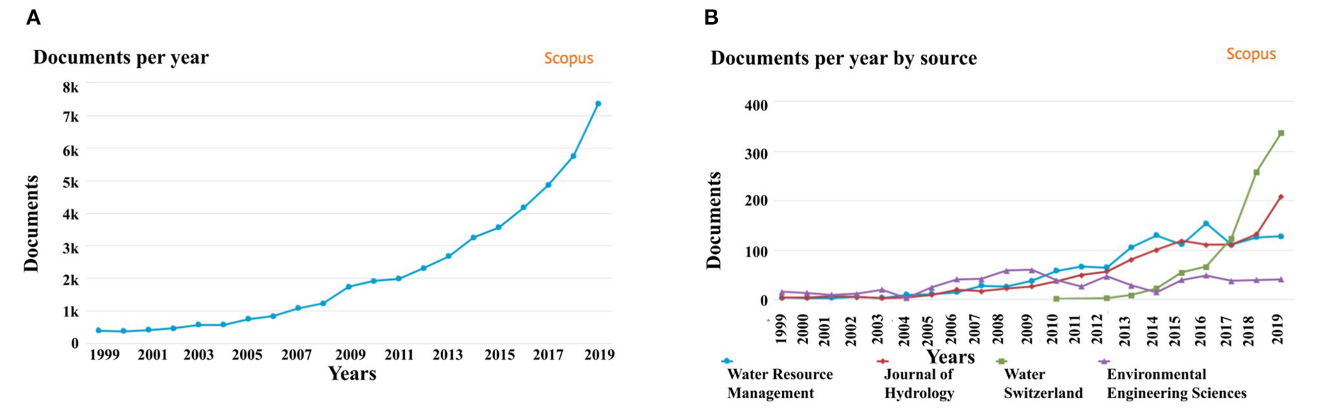


FIGURE 1 | Research trend showing increased application of machine learning techniques in water management and hazard (Copyright 2020 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V). **(A)** Documents per year during 1999–2019. **(B)** Documents per year source during 1999–2019.

Kitanidis, 1999). However, since the overall global climate is changing, rainfall data in any given area are non-stationary; thus the past does not necessarily predict the future, and the information given in recent data points may be more predictive than that of the data points from the more distant past (Tay and Cao, 2002). Limitations of probabilistic methods to produce realistic and specific results for water security planning have required the employment of physics based models for these predictions. Modeling hazardous events using physics based approaches requires the theoretical understanding of the atmospheric, land, and human processes and their interconnections; along with dynamics behind multiple hazards. However, many physics based models are designed to simulate pristine watersheds where hydrology is assumed to behave in a "pure" way, untainted by human interference (Joslin, 2016); therefore these physics based models are not suitable alone for predicting water-related hazards in urban watersheds. In addition, physics based models require large parallel machines and long periods of time for computation, neither of which may be available to water managers. Compared with the traditional modeling approaches, predictive data analytics powered by ML models can directly extract knowledge of natural disaster processes based on previous disaster occurrences and geo-environmental factors without prior knowledge (Pham et al., 2016; Rahmati et al., 2019). Unlike physics based modeling approaches, ML techniques can provide a bridge between physics based and probabilistic models because they can highlight patterns, trends, and regularities in data without requiring detailed understanding of the physical processes (Dibike and Solomatine, 2000; Rahmati et al., 2019), even when data are sparse, and with less complexity of construction and at relatively low computational cost (Mekanik et al., 2013).

Based on the scientific reasoning behind them, ML applications for predicting water-related parameters can be categorized either as inductive, whereby classifications are made based on statistical similarity in the hydrologic data directly; or deductive, whereby environmental variables (e.g., watershed characteristics) are analyzed as key drivers of hydrology to create classification (Wagener et al., 2007, 2010; Olden et al., 2012; Auerbach et al., 2015). Because the inductive approach requires abundant hydrologic data (although all watersheds are ungauged at some point with unavailable or insufficient measurements; Joslin, 2016) many studies have favored the deductive approach, which classifies rivers and watersheds based on readily available environmental data that reflect the main drivers of hydrologic processes (Auerbach et al., 2015). Many researchers have utilized the deductive approach to relate stream condition (e.g., flow regimes, biodiversity, streamflow) with upstream watershed characteristics for different water resource management purposes (Poff and Allan, 1995; Snelder and Biggs, 2007; Carlisle et al., 2008; Reidy Liermann et al., 2012; Rice et al., 2015). The rationale for deductive classification methods, such as hydrologic regionalization, environmental regionalization, and environmental classification is to group river hydrological characteristics by spatial representation (e.g., river basin, region, catchment) based on environmental, hydrological, physical, and climatic similarity (Olden et al., 2012) to develop reliable class and empirical relationships between predictor and watershed characterizations.

2.1. Floods

Long term processes of change, including changes in climate, shifts in population, and increases in urbanization, will likely increase future urban flood risk changing the assumptions upon which flood risk analysis and management has long been based (Gangrade et al., 2019), and requiring new tools for risk assessment (Milly et al., 2008). In order to understand how to predict floods and to mitigate their effects on urban areas using new tools, it is important to understand the events that lead to flooding. The locations and processes that contribute to floods include atmospheric processes, catchment-level floods, river flooding, and accumulation of water in flood-prone urban areas (Merz et al., 2010). We discuss next the ML methods applied to each of these processes.

TABLE 1 | Machine learning methods discussed in each section.

Topic	Machine learning	Summary	References
Atmosphere (section 2.1.1)	ANN	Use of ANN has proved to be efficient in analyzing and representing complex, non-linear relationships between multiple atmospheric and hydrological parameters. Compared with the traditional modeling approach, ANNs have varying performance and are less time consuming.	Sahoo et al., 2017; Zaidi et al., 2018
	SVM	The performance of SVMs for forecasting regional rainfall varies across the geographic area. For non-stationary time series forecasting, DSVMs generalize better than the standard SVMs. The evaluation of these methods is conducted using both real data and simulated data.	Cao and Gu, 2002; Mohanty and Mohapatra, 2018
	Anomaly detection	Various anomaly detection algorithms have been proposed for detecting point anomalies to improve hydrological and climate data quality, as well as to mine potentially meaningful pattern anomalies within a given time series or spatio-temporal data.	Chandola et al., 2009a; Das and Parthasarathy, 2009; Sun et al., 2017
Catchment (section 2.1.2)	Evolutionary algorithms	Genetic programming approach performs better than the traditional hydrological models during scenarios where surface water movement and water losses are poorly understood.	Whigham and Crapper, 2001
	Cellular automata (CA)	The CA technique provides a versatile approach for modeling complex physical systems using a simplified 5-feature cell-based system. Compared with physically-based models, CA can dramatically reduce computational load, while providing a minimum required accuracy for rapid flood analysis in large-scale applications.	Guidolin et al., 2016
Rivers (section 2.1.3)	ANN	Compared with statistical models, ANNs tend to perform better for simulations containing non-linear patterns. Among popular ANNs, FFBPNN has been proven to have the best performances by many studies, and GRNN performs better than RBFNN in most cases. Thus, ANNs can serves as helpful tools for predicting river floods, as well as for mitigating missing flow data records.	Shamseldin, 2010; Badrzadeh et al., 2013; Tayyab et al., 2016
Urban Flood (section 2.1.4)	ANN, Bayesian linear regression, boosted decision tree regression, decision forest regression, linear regression	Noymanee et al. (2017) compared multiple ML techniques for predicting urban flood peak using a list of error metrics. The performance of different ML techniques varies for predicting urban flood stage at different timestamps. The study demonstrated that for predicting flood peaks, ANNs and Boosted trees performed best.	Noymanee et al., 2017
Indirect effects (section 2.1.5)	Reinforcement learning with agent-based models	Yang S. et al. (2019) used two studies to demonstrate the effect of Reinforcement learning with agent-based models in supporting the decisions of recovery actions after a flood disaster. The case that adopts the ML technique outperformed the other case and achieved a shorter recovery time.	Yang S. et al., 2019
Drought prediction (section 2.2.1)	ARF, BRT, Cubist	Model performance varies by drought type and across different regions. Park et al. (2016) demonstrated a case study showing that boosted random forests generally produced better results than the other two models for both arid and humid regions.	Park et al., 2016
	CART, random forests	Kuswanto and Naufal (2019) used a case study (based on both the TRMM and MERRA-2 datasets) to demonstrate that random forests perform well for prediction of droughts in East Nusa Tenggara, Indonesia.	Kuswanto and Naufal, 2019
	CART, BRT, random forests, MARS, FDA, SVM	Rahmati et al. (2019) demonstrated that the performance of different models varies when predicting the risk for different types of hazards. For example, the SVM model showed the highest accuracy for avalanches, while BRT demonstrated the best performance for flood hazards.	Rahmati et al., 2019
	ANN, SVR	For predicting the Standardized Precipitation Index (SPI) (in this case SPI 3, SPI 12, and SPI 24), a meteorological drought index, the wavelet boosting ANN (WBS-ANN) and wavelet boosting SVR (WBS-SVR) models produced better prediction results compared to the SVM.	Belayneh et al., 2016
	XGBoost	Zhang R. et al. (2019) demonstrated that the incorporation of non-linear and lag effects of predictors into the XGBoost method can significantly improve prediction accuracy of Standardized Precipitation Evapotranspiration Index (SPEI) and drought, providing a new modeling strategy for drought predictions based on multistation data.	Zhang R. et al., 2019

(Continued)

TABLE 1 | Continued

Topic	Machine learning	Summary	References
	W-QEISS	Zaniolo et al. (2018) applied a variable subset selection algorithm to improve the FRIDA's (FRamework for Index-based Drought Analysis) capability for automating the design of basin-customized drought indexes across different types of basins. The algorithm is based on a Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS) and is capable of maximizing the wrapper accuracy, minimizing the number of selected variables, and optimizing relevance and redundancy of the subset.	Zaniolo et al., 2018
Water quality prediction (section 2.3)	ANN	Several studies have shown the ability of ANNs to simulate water quality variables and to produce simulated values for un-gauged locations.	Palani et al., 2008; Singh et al., 2009; García-Alba et al., 2019
	XGBoost, RF	Lu and Ma (2020) evaluated the prediction performances of two novel hybrid decision tree-based ML models (based on XGBoost and RF) using the absolute percentage errors. The RF-based model has the best performance for predicting temperature, dissolved oxygen, and specific conductance, and the XGBoost-based model is best for predicting the pH value, turbidity, and fluorescent dissolved organic matter.	Lu and Ma, 2020
	Random forests, M5P, RT, REPT	Bui et al. (2020) demonstrated the capability of hybrid algorithms to improve the predictive power of several standalone ML models. Among these models, the Hybrid BA-RT showed the best performance.	Bui et al., 2020
Soil erosion (section 2.4)	Tree-based ML methods	Rahmati et al. (2017) found that many tree-based models (e.g., RF, RBF-SVM, BRT, and P-SVM) performed excellently both in the degree of fit and in performance for predicting gully headcuts. Hosseinalizadeh et al. (2019) proved that random forests were the most effective of these models for predicting and mapping gully headcuts in the future.	Rahmati et al., 2017; Hosseinalizadeh et al., 2019
	SVM	Mustafa et al. (2018) demonstrated that SVMs with different kernel functions have different performance levels for predicting soil erosion. They found that the polynomial kernel function had the highest performance, followed by linear and radial basis functions. Pourghasemi et al. (2017) explored multiple individual and ensemble ML methods (e.g., ABB, SVM, maximum entropy) for soil erosion prediction and concluded that the ANN-SVM ensemble performed best.	Pourghasemi et al., 2017; Mustafa et al., 2018
	ANN	Rahmati et al. (2017) demonstrated that the ANN could be applied to produce accurate and robust gully erosion susceptibility maps for decision-making and soil and water management practices, even though the random forests outperform ANN in many cases. Abdollahzadeh et al. (2011) demonstrated that ANN outperforms Multi Linear Regression (MLR) for predicting soil erosion.	Abdollahzadeh et al., 2011; Pourghasemi et al., 2017; Rahmati et al., 2017
Sediment transport (section 2.4.1)	ANN	The performance of ANN varies based on the training dataset (e.g., the time span and data quality) and the type of sediment for prediction. These ANN predictions are often tested against domain models and theories.	Tayfur, 2002; Lin and Montazeri Namin, 2005; Bhattacharya et al., 2007; Yang et al., 2009
	Adaptive-network-based fuzzy inference system (ANFIS)	Wieprecht et al. (2013) demonstrated that the ANFIS approach could be a useful alternative technique for predicting both bedload and total bed-material load. Lin and Montazeri Namin (2005) found that the method can be used to model both uniform and non-uniform suspended sediment. Bakhtyar et al. (2008) revealed that the ANFIS model provides higher accuracy and reliability for longshore sediment transport techniques than other methods, such as Fuzzy Inference System and CERC.	Lin and Montazeri Namin, 2005; Bakhtyar et al., 2008; Wieprecht et al., 2013
	M5 model trees	Goyal (2014) presented a comparative evaluation of the performance of M5 Model Tree and wavelet regression vs. ANN clearly demonstrating that M5 Model Tree and wavelet regression outperform ANN models in estimation of sediment yield. Onderka (2012) compared the M5 model tree with the conventional power-law rating curves, and concluded that the M5 model has better performance for modeling suspended sediments in a headwater catchment.	Onderka, 2012; Goyal, 2014

(Continued)

TABLE 1 | Continued

Topic	Machine learning	Summary	References
Sediment load (section 2.4.1)	Random forests	Francke et al. (2008) demonstrated that Random forests and quantile regression forests, compared with generalized linear models, are more accurate and favorable for reproducing sediment dynamics.	Francke et al., 2008; López-Tarazón et al., 2012
	Genetic algorithms (GA)	Yadav et al. (2019b) suggested that GA models outperform other models, such as ANN and SVM, for estimating suspended sediment yield. Altunkaynak (2009) found that GA models outperform the regression method for predicting sediment loads.	Altunkaynak, 2009; Yadav et al., 2019b
	Unsupervised techniques	These methods are self-organizing, and their results are often validated using domain models and knowledge. For example, Xu et al. (2019a) used the concept of geological landform regions to verify the clustering results of sedimentation potential from a self-organizing map.	Ahmed et al., 2018; Xu et al., 2019a
Urban infrastructure (section 2.4.1)	Random forests	Xu et al. (2019a) demonstrated decent performance of random forests for forecasting sedimentation risks at culverts by validating the results with field inspection.	Xu et al., 2019a
	ANFIS	Azamathulla et al. (2011, 2012) demonstrated that the ANFIS approach can give more satisfactory results for predicting both the scour depth at culvert outlets and sediment transport in clean sewers compared with other methods (regression equations and ANN).	Azamathulla et al., 2011, 2012
Flood management with RS (section 2.5.1)	ANN	Tsintikidis et al. (1997) used a shallow neural network with one hidden layer to estimate rainfall from a passive microwave radiometer SSM/I data. The network considered brightness temperature and associated polarization information as inputs and it output the rainfall rates.	Tsintikidis et al., 1997
	Random forests	Kühnlein et al. (2014) performed a precipitation estimate using random forests with satellite-derived information on cloud-op height, cloud-top temperature, cloud phase, and cloud water path retrieved from Meteosat Second Generation (MSG) Spinning Enhanced Visible and Infrared Imager (SEVIRI). Feng et al. (2015) developed a random forest based approach to map accurately a flooded area using high-resolution (0.2 m) imagery obtained from Unmanned Areal Vehicle (UAV) imagery.	Kühnlein et al., 2014; Feng et al., 2015
	K-NN	Shahabi et al. (2020) employed a ML ensemble method with four different k-nearest neighbor (kNN) algorithms for flood detection and susceptibility mapping using Sentinel-1 images to generate the flood inventory and SRTM DEM to obtain various flood-related conditioning factors.	Shahabi et al., 2020
	LSTM	A spatio-temporal sequence forecasting using Convolutional Long-Short Term Memory (ConvLSTM) for precipitation nowcasting. RADAR echo data in 2D from a ground-based RADAR was used in this study. ConvLSTM forecasted the echo data.	Shi et al., 2015
	CNN	Pan et al. (2019) used CNN to improve the precipitation estimates from NWP models. Based on the work by Hong et al. (2004), Hayatbini et al. (2019) proposed a CNN model to estimate precipitation using geostationary satellite data GOES-16. Jain et al. (2020) used water indices with CNN to detect flood water. Potnis et al. (2019) used a CNN based architecture called ERFNet to detect flooded urban regions from high resolution Worldview-2 satellite imagery. Jiang et al. (2020) proposed an approach to obtain waterlogging depth from video images using CNN.	Hong et al., 2004; Hayatbini et al., 2019; Pan et al., 2019; Potnis et al., 2019; Jain et al., 2020; Jiang et al., 2020
	Knowledge-based approaches	Kurte et al. (2017) used a semantics-driven framework to enable spatial relationships based semantic queries to detect flooded regions from satellite imagery and further extended the framework (Kurte et al., 2019) to accommodate temporal dimension that enabled spatio-temporal queries over flooded regions. In a similar approach, Potnis et al. (2018) developed a flood scene ontology (FSO) which formally defines complex classes such as *Accessible Residential Buildings*, to classify flooded regions in urban area from satellite imagery.	Kurte et al., 2017; Potnis et al., 2018; Kurte et al., 2019

(Continued)

TABLE 1 | Continued

Topic	Machine learning	Summary	References
Water quality monitoring with RS (section 2.5.2)	ANN	Dogan et al. (2009) used ANN to improve the accuracy of biological oxygen demand (BOD) estimation from RS imagery. Wu et al. (2014) used ANN for TSS turbidity estimations to analyze data measured with a hyperspectral spectroradiometer. Hafeez et al. (2019) compared various ML techniques for estimating water quality indicators form RS imagery and found that ANN worked well. Govedarica and Jakovljević (2019) found that the SVM algorithm worked better than ANN with Landsat 8 data and ANN worked better than SVM when Sentinel-2 data was used for water quality monitoring.	Dogan et al., 2009; Wu et al., 2014; Govedarica and Jakovljević, 2019; Hafeez et al., 2019
	SVM	Wang et al. (2011) used the support vector regression (SVR) method to retrieve various water quality estimators from SPOT-5 satellite data. Huo et al. (2014) used genetic algorithms combined with support vector machines (GA-SVM) to build an inversion model for eutrophic indicators such as Chl-a from Landsat ETM imagery.	Wang et al., 2011; Huo et al., 2014
Impervious surface detection with RS (section 2.5.3)	Random forests	Bian et al. (2019) used a random forest algorithm and time-series data from multiple satellites HJ-1A/B and GF-1/2 to estimate the changes in the impervious surface percentage over the years 2009–2017.	Bian et al., 2019
	PBL, PUL, SVM	Yao et al. (2017) adopted a one-class classification approach to detect impervious surfaces using high-resolution GF-1 satellite images, and found that Presence and Background Learning (PBL) and Positive Unlabeled Learning (PUL) outperformed SVM models.	Yao et al., 2017
	CNN	Zhang H. et al. (2019) used a deep CNN approach with data fusion from optical and SAR satellites WV-3, Sentinel-2, and Radarsat-2. Similar other works, Sun et al. (2019) (used 3D CNN with WV-3 and LiDAR), McGlinchy et al. (2019) (used UNet with WV-2), show increasing trends of using deep learning based approaches with multi-satellite data fusion.	McGlinchy et al., 2019; Sun et al., 2019; Zhang Z. et al., 2019
Multi-hazard assessment (section 3)	BRT, GAM, SVM	Rahmati et al. (2019) investigated and mapped multi-hazard exposure using a combination of ML models. They found that the different ML models differed in their accuracy in predicting the different hazards, but that the applied ML models were nevertheless useful and generalizable for multi-risk mapping.	Rahmati et al., 2019
	Random forests, RBF neural network	Chen et al. (2019) evaluated the risk of regional flood disaster in the Yangtze River Delta (YRD) region. They discovered that the level of urban flood disaster is closely related to rainfall, topography, economic development, land use, soil erosion, urban flood control investment, and disaster emergency response capability.	Chen et al., 2019
	Random forests, SOM	Xu et al. (2019a) showed that ML application can be used not only for multi-risk assessment and hazard prediction but also for exploring the complex and interconnected processes behind multiple hazards.	Xu et al., 2019a
	Random forests	Pourghasemi et al. (2020) developed the Sendai framework, which used random forests to produce a reasonable understanding of the factors controlling flood, forest fire, and landslide occurrence, and to produce a multi-hazard probability map for facilitating integrated and comprehensive watershed management and land use planning.	Pourghasemi et al., 2020
	LSTM	Yang T. et al. (2019) used long short-term memory units (LSTM) to improve the timing component of the amplitude of peak discharge for flood simulations generated by global hydrological models over different climate zones.	Yang T. et al., 2019
Best management practices (BMP)	GA/adaptive search	Hadka and Reed (2013) developed a high-performance adaptive search "Borg" algorithm, which was shown to be the most scalable and the best performing of five best performing multi-objective optimization algorithms applied to rainfall-runoff calibration, long-term groundwater monitoring, and risk-based water supply portfolio planning. Others applied GA-based optimization models to find solutions to water quality problems for several watersheds in the United States by connecting non-point pollution reduction models with economic components.	Srivastava et al., 2002; Hadka and Reed, 2013; Limbrunner et al., 2013; Reed and Kollat, 2013; Chen et al., 2015

2.1.1. Atmospheric Process Methods

One ML method that is used to capture the underlying relationship between independent and dependent variables in atmospheric processes is Artificial Neural Networks (ANNs). ANNs are interconnected networks comprising an input layer, some number of hidden layers, and an output layer. Each layer contains several processors, or nodes, referred to as artificial neurons. The neurons in each layer are connected to the neurons in the previous and next layers, and they transfer information from one layer to the next. Synaptic weights and biases, along with activation functions applied to the input layer, modulate the input signals sent from one layer to the next. The processed information is then sent as output to the connected neurons in the output layer (Zounemat-Kermani et al., 2020). The power of ANNs is their ability to learn functional relationships, with minimal empirical error, between these variables. Additionally, the use of activation functions with ANNs allows them to handle non-linear data effectively (Zaidi et al., 2018). In fact, many water related studies (e.g., Sahoo et al., 2017) using ANNs have shown that complex, reproducible, non-linear relationships exist among, for example, precipitation, temperature, streamflow, climate indices, irrigation demand, and groundwater levels.

Another ML method that has been used for predicting average rainfall is a classification algorithm known as Support Vector Machines (SVM) (e.g., Mohanty and Mohapatra, 2018). This method, developed by Vapnik (1995), is based on *Structural Risk Minimization*, which, rather than minimizing empirical error, as ANNs do, minimizes an upper bound of the generalization error ε. Dynamic Support Vector Machines (DSVMs), a modified version of the SVM, can be used to accommodate the structural changes in non-stationary rainfall data because it uses, instead of a static ε and static regularization constants, an exponentially decreasing ε, and exponentially increasing regularization constants (Cao and Gu, 2002) to allow room for analysis of changing patterns in the data.

The probabilities of hydrological extreme events such as floods and drought are modeled using different distributions from those that predict future average values. Traditionally, these events and their return periods are estimated with distributions associated with Extreme Value Theory (e.g., Kao and Ganguly, 2011). However, ML techniques for anomaly detection have begun to be applied to hydrological extremes problems. Anomaly detection is the identification of outliers in the data, or items that differ significantly from the overall trend of the data. Typically, anomalous data is related to issues such as measurement equipment failure or an extreme hydrological event. For example, Das and Parthasarathy (2009) used unsupervised spatio-temporal distance-based and neighborhood-based anomaly detection method with global climate data to identify extreme drought and heavy rainfall at specific locations. Characterization of short-term and long-term future extreme events have also been made with anomaly detection using trends found in historical time series. For these analyses, techniques such as kernel-based (rule-based classification), window-based (examination of the data in smaller "windows" in space or time), predictive, and segmentation (partitioning data into even smaller, possibly unequal, segments) algorithms are employed along with anomaly detection for locating extremely low and extremely high temperature and precipitation events (Chandola et al., 2009b). In the case of the research by Sun et al. (2017), a density-based method was applied to anomaly detection in a hydrological time series. That is, the data were transformed to a piecewise linear representation through the important feature points of the data before mapping their slope, length, and mean to three-dimensional space for examination.

2.1.2. Catchment-Level Methods

Flood models at the catchment level analyze mainly issues of runoff generation and concentration leading to flood discharge. Because flood flow predictions are complex, non-linear, and not well-understood, ML may be required to *evolve* algorithms to derive characteristics of a particular flow. One way of evolving these algorithms is with the use of genetic programming, or genetic algorithms (GA), which produce, using routines imitating Darwin's "natural selection," algorithms directed to perform tasks defined by a set of training examples. Whigham and Crapper (2001) applied a type of genetic programming system to discover rainfall-runoff relationships for two meteorologically and topographically different catchments, one in Wales and one in Australia, and compared the results to those obtained with a traditional deterministic lumped parameter model. While both models did well when rainfall and runoff were correlated, the genetically programmed model performed better on the more poorly correlated data because it was allowed not to assume any underlying relationships, only to demonstrate its "fitness" to solve the problem.

Guidolin et al. (2016) used a two-dimensional cellular-automata-based model employing simple transition rules and a weight-based system to model catchment-level runoff. This diffusive-like method is designed to work with various general grids (rectangular, hexagonal, triangular) and with different neighborhood types (e.g., Moore or von Neumann). It also allows for model parallelization to increase its efficiency in large compute environments. To propagate a flood using this method, ratios of water to be transferred from a central cell to downstream neighbor cells are calculated using a weight-based system, with water volume transferred limited by Manning's formula (Manning et al., 1890), and the critical flow equation. Water velocity and an adaptive time step are evaluated within a larger updated timestep. The results of the emergent behavior of this process shows good agreement with much more computationally intensive physical methods.

2.1.3. Machine Learning for Analyzing River Floods

Flood hazard in rivers can be characterized by the probability and intensity of large river flows and their consequent inundations, and it depends on the atmospheric and catchment processes preceding river flood generation (Merz et al., 2010). In fact, river floods are generally defined in hydrological terms by their water level or amount of discharge. Thus, Shamseldin (2010) explore the use of ANN for forecasting discharge from the Blue Nile river in Sudan. The type of neural network they chose was that of a multi-layer perceptron (MLP) feedforward network, a non-linear input–output model consisting of a network of interconnected neurons, or computational units, linked together by connection pathways. The input layer is essentially a set vectors of independent variable values, whereas the output layer is a set of possible dependent variable vectors of values. Between these two layers is a hidden layer containing an unknown number of neurons which are usually estimated by a trial-and-error procedure based on a mathematical non-linear transfer function (Shamseldin, 2010). Input variables in this case were weighted historical rainfall estimates, weighted seasonal rainfall estimates, and seasonal expectation of discharge; and the output variables were the river discharge values. Results showed strong correlation with observations for the river.

In addition to the multilayer perceptron ANN approach, other types of ANNs have been used to analyze river floods. For example, Tayyab et al. (2016) applied and compared three different types of ANNs to predict stream discharge for the Jinsha River Basin in China. The methods included feedforward back propagation neural networks (FFBPNN), generalized regression neural networks (GRNN), and radial basis function neural networks (RBFNN). The differences among these approaches lies in the hidden layer functions and activation functions that are applied to the problem. Badrzadeh et al. (2013) expanded on these ANN approaches by coupling wavelet (transforms that identify trends in the data normally not revealed by signal analysis approaches and also help to de-noise a dataset) multi-resolution analysis and adaptive neuro-fuzzy interface system (ANFIS) techniques (integration of neural networks and fuzzy logic) as preprocessing techniques to the ANN and show improved daily river flow forecasting over the use of ANNs alone, especially for long lead times. Mosavi et al. (2018) demonstrated the application of ANNs, neuro-fuzzy, SVM, and support vector regression (SVR) (SVM with regression only), in forecasting river floods and predicting the runoff hydrograph. The robustness of these techniques was evaluated and was found to be in good agreement with the observations.

2.1.4. Methods for Addressing Flood-Prone Urban Areas

Building resilience to natural disasters is one of the most pressing challenges for achieving sustainable urban development in flood-prone regions (Chang et al., 2019). River flooding in urban areas can cause high levels of damage, and while a relationship between hydrological characteristics and damaging floods may exist, knowing about an area's hydrological characteristics does not always indicate understanding of its vulnerability to damaging floods (Pielke, 2000). This understanding is imperative for hazard-mitigation planning for urban areas because these areas' responses to rainfall extremes tend to be faster than those for natural surfaces (Rodriguez et al., 2003). Thus, strategies for flood mitigation in these areas such as detention ponds, soakaways, permeable concrete, and green spaces, or upstream solutions such as river training and construction of dams and levees (Shamseldin, 2010) should be evaluated and implemented based on a thorough understanding of flood risks and responses of the area. For example, for predicting urban floods for the city of Pattani south of Thailand, Noymanee et al. (2017) examined the entire Pattani basin, which includes two dams for water management: a diversion-type, Pattani Dam, and a hydropower plant, Bang Lang Dam. It is known that the most frequent floods are a result of overflow from flash flooding of the Pattani Dam rushing toward the city. The researchers acknowledge that a comprehensive approach to controlling floods in the area must include both structural and non-structural measures such as the development of improved technology for data management of the drainage network, and an increase in the sensors' frequency and extent of coverage. Thus, Noymanee et al. (2017) tested five different ML methods using open data pertaining to the area hydrology, the dam structures, the drainage network, and the technological components of the dams to explain the occurrence of extreme floods estimating dam water levels and cumulative precipitation amounts to forecast flood peaks in the urban area. The five methods tested included an ANN, Bayesian linear regression (statistical inference using Bayes' theorem), boosted decision tree regression and decision forest regression (both similar to random forest analysis discussed in section 2.2.1) and linear regression. Results showed the lowest error and highest correlation with the observations in the urban area from the Bayesian linear regression. This favorable result for that method may have occurred because it was informed by probability distributions drawn from prior data.

Often, in order to understand and manage risks of urban flooding beyond purely hydrological considerations, integration of decision support tools with predictive models is instructive. For example, one study (Rozos, 2019), combined a hydrological model, a demand management model called a network flow programming model (NFP), and an Feed Forward Neural Network (FFNN) to simulate a water supply system in Athens, Greece. The NFP optimizes and simulates the operation of a water supply system given hydrological inputs. FFNNs are the simplest type of ANN, whereby information moves in a forward direction from input nodes to the hidden layer to the output nodes (Mosavi et al., 2018) and they lend themselves to multi-model coupling. In this case, the NFP used synthetic data of a length capable of capturing the risk of each policy. Then the penalty functions of the NFP were selected to reflect the operating policies with different levels of risk acceptance. This process provided a large set of training data over a long period of time that was then used as input to the FFNN. This process allowed optimal decisions to be identified and made for the Athens system.

2.1.5. Predicting Indirect Flood Effects in Urban Areas

Indirect flood effects are those that cause damage to assets outside the flooded area. These assets can be physical, economic, social, or ecological in nature with impacts lasting for days, months, or even years after a large flooding event (Costello et al., 2019). In order to evaluate the extent of these effects, multi-agent-based simulations have been applied. Agent-based models simulate actions and interactions of autonomous agents, which can be individual actors or groups of actors, to assess the effects of these individual actions on the system as a whole. In one study (Yang S. et al., 2019), reinforcement learning, which rewards software agents for actions taken to maximize their cumulative reward, was used with the agent-based simulation for the optimization of post-disaster recovery for both individual companies and supply chains for Tokyo, Japan. That study showed improved indirect damage estimation accuracy and mitigation potential over statistical methods and rough empirical models.

2.2. Drought

Drought is a prolonged period of precipitation deficit that may occur at varying spatiotemporal scales ranging from local to regional, lasting for weeks, months, multiple years, or even decades (Pendergrass et al., 2020; Hao et al., 2018). Drought may be exacerbated by extreme heat, soil moisture deficit, land atmosphere feedbacks, sea surface temperature anomalies, atmospheric circulation, and human activities such as land use and land cover changes and increased water demand (Cook et al., 2007; Dai, 2011; Kam et al., 2014). Droughts are high-impact weather hazards that affect agriculture, economy, ecosystem, water supply, and human lives (Hao et al., 2018). Over the past two decades, the total cost associated with drought is estimated to be billions of dollars (Huntingford et al., 2019). In a warming climate, the duration and intensity of drought is further projected to increase (Pagán et al., 2016; Pendergrass et al., 2020). Therefore, an advancement in the capability of timely prediction and development of early warning systems is crucial for drought risk management and strategic planning.

2.2.1. Advancement in the Use of Machine Learning Techniques for Drought Prediction

Drought is a complex weather hazard (Van Loon, 2015); therefore, a comprehensive understanding of the physical mechanisms that drive drought is essential to improving drought prediction (Huang et al., 2016). Numerous studies have been conducted to understand the intricate physical processes that lead to the extreme low moisture conditions of drought. Scientists have employed dynamical methods that involve climate and hydrological model simulations, statistical models using a suite of predictors and drought indices, as well as hybrid models for drought prediction (Fernández et al., 2009; Dutra et al., 2014; AghaKouchak, 2015; Mo and Lyon, 2015; Wood et al., 2015; Hao et al., 2017, 2018).

During the last decade, there has been an increase in the use of ML techniques to improve drought predictability (Hao et al., 2018). For instance, random forest ML algorithms have been increasingly used in drought prediction studies (Park et al., 2016; Kuswanto and Naufal, 2019; Rahmati et al., 2020). Random forests are extensions of decision tree analysis that start with classification trees–types of decision trees that can be grown together as a "forest" in a computational system. They provide highly accurate classification and characterization of complex predictor variable interactions while maintaining flexible analytical technique selection (Allen et al., 2018). Random forests also provide the capability to deal with the issue of overfitting and multicollinearity as compared to the traditional linear regression models (Konapala and Mishra, 2020). Park et al. (2016) employed random forests, boosted regression tree, and Cubist ML algorithms (rule-based model trees on which the terminal leaves contain linear regression models) for meteorological and agricultural drought monitoring using 16 remote sensing based drought factors over arid and humid regions in the United States. Their findings suggest that among the three approaches, random forests provide the best performance for Standardized Precipitation Index (SPI) prediction. Similarly, Kuswanto and Naufal (2019) found the performance of random forests to be optimal when using SPI derived from Modern-Era Retrospective analysis for Research and Applications (MERRA-2) for drought prediction over the East Nusa Tenggara Province in Indonesia. A more recent study, Rahmati et al. (2020) compared the performance of six different ML techniques [classification and regression trees (CART), boosted regression trees (BRT), random forests, multivariate adaptive regression splines (MARS), flexible discriminant analysis (FDA), and SVM] for mapping agricultural drought hazard in the southeast region of Queensland, Australia. Similar to Park et al. (2016) and Kuswanto and Naufal (2019), they found that random forests had the best goodness-of-fit and predictive performance among the six models. Zaniolo et al. (2018) contributed to the FRIDA (FRamework for Index-based Drought Analysis) for the automatic design of basin-customized drought indexes across different types of basins by applying a ML-powered variable selection algorithm. The algorithm is based on a Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS), which applies a multi-objective evolutionary algorithm to identify Pareto-efficient subsets of variables. This technique is able to maximize the wrapper accuracy, minimize the number of selected variables, and optimize relevance and redundancy of the subset. As a result, the framework is able to build an index that represents a surrogate of the drought conditions in a basin through the computation and combination of all the relevant available information regarding the water cycle in the system identified using the feature selection algorithm.

ANN ML techniques (see section 2.1.1) have also been used for drought forecasting (Mishra et al., 2007; Morid et al., 2007; Belayneh and Adamowski, 2012; Belayneh et al., 2014). Belayneh et al. (2016) coupled a wavelet transform data processing technique (see section 2.1.3), bootstrapping and boosting ensemble approaches with ANN and Support Vector Regression (SVR) (see section 2.1.1) for drought prediction in the Awash river basin of Ethiopia. Bootstrapping is a resampling technique with replacement that was used to create bootstrap ANN and SVR ensemble models to reduce model prediction uncertainty. Boosting techniques improve the performance of an algorithm by producing a series of models focusing on training cases that were not well predicted previously. The researchers found that the coupled models showed an improved performance and provided more robust SPI predictions as compared to either of ANN or SVR alone.

ANN models can be limited by model interpretability, local minima traps, and computational efficiency issues. Thus, alternatively, XGBoost has been gaining popularity due to its high execution speed and improved model performance as compared to other ML techniques such as SVM, ANN, and random forests (Fan et al., 2018; Shimoda et al., 2018; Zhang R. et al., 2019). XGBoost is an ensemble technique that implements a gradient boost decision tree algorithm to produce an ensemble of weak prediction models. Models are subsequently added to improve errors until an optimum performance is achieved. Zhang R.

et al. (2019) compared the performance of XGBoost with a traditional statistical model and an ANN model for Standardized Precipitation Evapotranspiration Index (SPEI) prediction with a lead time of 1–6 months for 32 weather stations in the Shaanxi Province of China. In their study, the XGBoost model showed the best performance for SPEI prediction, achieved highest user's and producer's accuracies and was much faster than the ANN model.

2.3. Water Quality

The deterioration of water quality in both groundwater and surface water has become a major concern causing negative impacts on human well-being, eco-systems, water supply, and infrastructure around the world (UN, 2012; Khan and See, 2016). According to United Nations (UN), more than 880 million people are living in water scarcity without adequate safe drinking water, and 2.6 billion people lack access to basic sanitation due to water shortage (UN, 2010, 2012). Effective management of water supply systems and watersheds often requires reliable and timely approaches for predicting water quality and forecasting future water quality trends (Wang et al., 2017; Bui et al., 2020). Based on established water quality standards (Nowell and Resek, 1994; EPA, 2012), water quality is often estimated using a combination of water quality parameters that reflect the physical, biological, or chemical characteristics of the air, watershed hydrology, soils, and sediment transported in the aquatic system (Hou et al., 2013; EPA, 2019). Developing accurate and timely prediction of water quality is a challenging effort. The traditional approaches utilize water quality models for analyzing and predicting water quality parameters. Most of these models consist of mathematical representations of physical mechanisms that determine (a) the fate, transport, and degradation of pollutants within a water body, and (b) the movement of pollutants from land-based sources to a water body (Refsgaard and Henriksen, 2004). Despite their usefulness for modeling specific scenarios, water quality models can only provide one line of evidence that serves as an imperfect approximation of reality (Kebede, 2009). This is because of process complexity of the water quality problems in that (1) there is a large number of interconnected multi-domain processes (e.g., physical transport, hydrological, chemical, and biological); and that (2) many underlying mechanisms that may affect water quality are still unknown. Complex water quality models often involve time-consuming and labor-intensive processes (Ahmed et al., 2019), rendering them costly and ineffective for supporting many time-critical water resources management tasks that have limited budgets. Compared with process-based (mechanistic) models, the newly emerging data-driven approaches for water quality predictions often rely on a large volume of water quality and hydrological data from various sources (Khan and See, 2016). Examples of these data sources include the United States Geological Survey (USGS) online resource—National Water Information System (NWIS) and the United States Environmental Protection Agency's (USEPA) STORET Data Warehouse (Beran and Piasecki, 2008). These analyses normally consider the combined effect of multiple water quality parameters, such as ammoniacal nitrogen (NH3-N), suspended solid (SS), dissolved oxygen (DO), pH, and salinity. As many of these parameters are dynamic and affected by natural watershed hydrology, their influences on water quality may vary across watersheds (EPA, 2019). In different watersheds, some parameters may have greater and more noticeable influences on water quality than others (Khan and See, 2016). In response to this challenge, the water quality index (WQI) has been proposed as a representation of several water quality variables simultaneously considered. However, calculating WQI using traditional approaches consumes time and is often filled with errors during derivations of sub-indices (Bui et al., 2020). To address these limitations and improve water quality analysis and prediction, researchers have applied many ML techniques (Khan and See, 2016; Ahmed et al., 2019; Bui et al., 2020), as well as developed a few hybrid approaches that combine various traditional methods with ML techniques (Taskaya-Temizel and Casey, 2005; Wang et al., 2017). We discuss the application of some of these approaches next.

Palani et al. (2008) and Singh et al. (2009) applied ANN models to predict river and coastal water quality in India and Singapore respectively. Each found that the ANN-computed values of water quality indicators were in close agreement with their respective measured values in the river water. García-Alba et al. (2019) developed an ANN model to estimate bathing water quality in estuaries and found that ANN models are able to estimate *Escherichia coli* concentrations comparable to those extimated by process-based models, and at much lower computational cost. In more recent studies, combinations of multiple ML and data analytic techniques applied to a problem are preferred to analysis with a single ML technique. For example, Lu and Ma (2020) proposed coupling two ML models to improve water quality prediction: XGBoost (section 2.2.1), and a random forest algorithm (section 2.2.1). They found that while the hybrid XGBoost model performed better for PH values, turbidity, and fluorescent dissolved organic matter predictions, and the random forest model performed better for temperature, dissolved oxygen, and specific conductance prediction; the combined performance of the two models was the best for optimizing the calculation of a water quality index. Barzegar et al. (2020) applied two standalone deep learning (DL) models, a convolutional neural network (CNN), an ANN with a convolutional activation function, and the long short-term memory (LSTM) model, which includes feedback in addition to feedforward networks, and a combined CNN–LSTM model to predict two water quality variables, dissolved oxygen (DO; mg/L), and chlorophyll-a (Chl-a; μ/L), in the Small Prespa Lake in Greece. Assessment of the model performance using statistical metrics, showed that LSTM outperformed the CNN model for DO prediction, but the standalone DL models yielded similar performances for Chl-a prediction. The combined CNN–LSTM model, however, outperformed the standalone models for predicting both DO and Chl-a. By coupling the LSTM and CNN models, both the low and high levels of water quality parameters were successfully captured, particularly for the DO concentrations (Barzegar et al., 2020). Similar successful approaches involving the coupling of multiple ML algorithms for the short-term prediction of water quality parameters include Li et al. (2018) and Lu and Ma (2020). Bui et al. (2020) applied four standalone algorithms [random forests and three variants: M5P (similar to Cubist, section 2.2.1), random tree (RT), reduced error pruning tree (REPT)], and developed 12 algorithm combinations among these methods to predict water quality in northern Iran. They found fecal coliform concentrations to have the most effect and total solids to have the least effect on the predictions. Finally, Read et al. (2019) integrated theory with state-of-the-art ML techniques to improve predictions of water quality related parameters guided by physical laws. The study presented a use case for a Process-Guided Deep Learning (PGDL) hybrid modeling framework for predicting depth-specific lake water temperature, which serves as an important water quality parameter. The PGDL consisted of three primary components: a deep learning (many-layered neural network) model with temporal awareness (long short-term memory recurrence), theory-based feedback (model penalties for violating conversation of energy), and model pre-training to initialize the network with synthetic data (water temperature predictions from a process-based model) (Read et al., 2019). Through the use case the researchers demonstrated that the integration of scientific knowledge into deep learning tools shows promise for improving predictions of many important environmental variables.

2.4. Soil Erosion and Sediment Transport

Erosion and sedimentation are naturally occurring processes that include the detachment, transportation, and deposition of soil particles through the action of wind, water, and ice (NRCS, 2008). However, excessive soil erosion and sedimentation rates are results of anthropogenic activities (e.g., urbanization and agriculture) where soil surfaces are exposed and initially not revegetated (e.g., construction sites). Without proper mitigation, erosion and sedimentation in urban areas can cause a series of adverse impacts to the environment and urban areas (Guy, 1970; Hewett et al., 2018), which include water pollution, degradation of aquatic habitat, infrastructure damage (e.g., sediment blockage in urban waterways, storm sewer, and stream crossings, as well as silting of roadways, utility supply networks, and fences), increase in water-treatment costs, and stream bank instabilities (e.g., gullying and land-slides) (NRCS, 2008).

2.4.1. Machine Learning Techniques for Sediment Research

To tackle sediment-related problems, the predictions of sediment production and transport are required to inform urban planning and watershed management communities of the major source of sediment and erosion-prone areas. Conventionally, these predictions are addressed through a wide variety of erosion and sediment transport models (Merritt et al., 2003; Nearing et al., 2005). Despite the usefulness and maturity of these traditional approaches, the prediction of sediment-related parameters (e.g., soil losses, in-stream sediment load, and sediment delivery ratio) is still challenging because of the following model limitations: (a) running many physically-based erosion and sediment transport models are time- and resource-intensive, and requires the consideration of more physical processes in addition to the hydrological process making models are less applicable to sediment-related predictions in large watersheds and areas (Abaci and Papanicolaou, 2009); (b) most models are designed to simulate a specific type of erosion (e.g., rill, gully, and stream bank erosion) and sediment transport (e.g., suspended load and bed load) (Wischmeier and Smith, 1978; Ganasri and Gowda, 2015), while sediment-related problems in urban areas and urban waterways often entail multiple types of erosions and sediment transport therefore requiring the integration of a variety of models; and (c) most erosion and sediment transport models do not cover sediment transport and deposition at man-made structures (Rowley, 2014) in urban areas. A comparative study conducted by Liang et al. (2019) showed that data-driven models can effectively inform and complement the simulations conducted with physics based models. Currently, there are many studies that utilize various ML methods to address various issues in sediment research. We summarize a list of example studies by their application areas and their applied ML methods:

1. Modeling sediment transport
   - (a) Artificial neural networks (Tayfur, 2002; Lin and Montazeri Namin, 2005; Bhattacharya et al., 2007; Yang et al., 2009),
   - (b) Adaptive-network-based fuzzy inference system: (Lin and Montazeri Namin, 2005; Bakhtyar et al., 2008; Wieprecht et al., 2013),
   - (c) M5 Model trees (Onderka, 2012; Goyal, 2014).
2. Predicting sediment load
   - (a) Random forests (Francke et al., 2008; López-Tarazón et al., 2012),
   - (b) Genetic algorithms (Altunkaynak, 2009; Yadav et al., 2019b),
   - (c) Unsupervised techniques (Ahmed et al., 2018; Xu et al., 2019a).
3. Predicting soil erosion
   - (a) Tree-based ML methods (e.g., random forest, gradient boosted regression tree, naïve Bayes tree, and tree ensemble models) (Rahmati et al., 2017; Hosseinalizadeh et al., 2019),
   - (b) Support vector machine (SVM) (Pourghasemi et al., 2017; Mustafa et al., 2018),
   - (c) Artificial neural networks (Abdollahzadeh et al., 2011; Pourghasemi et al., 2017; Rahmati et al., 2017).
4. Sediment-related impacts on urban infrastructure
   - (a) Random forest (Xu et al., 2019a),
   - (b) Adaptive-Network-based Fuzzy Inference System (ANFIS) (Azamathulla et al., 2011, 2012).

In general, erosion and sediment research is a broad subject that provides numerous opportunities for ML applications. By reviewing the above-mentioned example studies, we have summarized that (a) compared with traditional erosion and sedimentation transport models, ML methods are easier and cheaper (Cigizoglu, 2002; Tayfur and Guldal, 2006; Yadav et al., 2019a), and can be readily applied to solve complex sediment problems that entail human factors and multiple erosion and sediment transport processes (Xu et al., 2019a), (b) ML models that rely on field data generally produce better and more reliable results than those obtained from experimental models (Kitsikoudis et al., 2014).

2.4.2. Hybrid Modeling Techniques for Sediment Research

In addition to its application to previously described hydrological studies, hybrid modeling has also been applied to sediment research (Merritt et al., 2003; Hajigholizadeh et al., 2018). Through the fusion of inductive data-driven models and deductive process-based models (Goldstein and Coco, 2015), hybrid models inherit the strengths of both the ML methods and physics-based models in a single model that has an increased performance in terms of speed (Babovic et al., 2001; Hall, 2004), accuracy (Krasnopolsky and Fox-Rabinovitz, 2005; Goldstein and Coco, 2015), and the capability of addressing soil-water problems with complex and multi-scale physical processes (Hajigholizadeh et al., 2018). An additional benefit of hybrid modeling is that ML models and data can be directly coupled to improve the calibration of process-based models (Knaapen and Hulscher, 2003; Ruessink, 2005; Mekonnen et al., 2012). Hajigholizadeh et al. (2018) summarized a table of hybrid modeling applications that integrate statistical models with process-based models in sediment research including:

- Modified Morgan, Morgan and Finney (MMMF) (Morgan et al., 1984),
- Sediment river network model (SEDNET) (Prosser et al., 2001),
- Erosion Assessment Tool of MIKE BASIN & MILW (SEAGIS) (DHI, 2003),
- Automated Geospatial Watershed Assessment (AGWA) (Scott et al., 2002).

2.5. Application of Machine Learning to Remotely-Sensed Data for Water Hazard Prediction and Mitigation

Remotely-sensed (RS) data, due to its wide spatial coverage, provides a synoptic view of disaster affected areas. It is also frequently available during the disaster response phase providing a temporal overview of the disaster situation. Due to the recent advancements in satellite sensor technology, RS data is now available at various spatial resolutions (i.e., low, medium, and high) affording local, regional, and global coverage, and various spectral resolutions, from a few spectral bands in optical sensors to several hundreds of spectral bands in hyperspectral sensors. Additionally, advancements in the RS field have resulted in a continuous growth in Earth Observation (EO) data archives. Due to these characteristics, RS data is a potential data source for each stage during hydrological pre-event planning and post-event countermeasures (Ge et al., 2020). Nevertheless, it is not always possible and is often dangerous to conduct ground surveys of disaster affected areas. Often the disaster destroys the transportation and communication facilities making ground-based survey impossible. In such time-critical situations, the proper selection of the sensor type, spatial resolution, and satellite revisit period is crucial, as pre-disaster and ancillary data can provide a wide coverage of the disaster affected area (Ge et al., 2020). Despite these occasional limitations, various powerful approaches have been developed recently in the context of advanced ML and computer vision to exploit the wealth of information that can be found in RS data to address various urban water hazards related events (Kurte et al., 2017).

2.5.1. Flood Management

Over the last two decades, RS data have successfully contributed to various stages of flood management (Rahman and Di, 2017) such as flood risk assessment and flood emergency planning and management. Flood risk assessment requires the performance of flood hazard assessment, exposure risk assessment, and vulnerability assessment. As a part of the flood hazard assessment, RS data have been analyzed for flood forecasting and evaluation of flood inundation. As a part of flood emergency planning and management, RS data have been widely used in flood early warning systems, rescue and relief operations, post-flood damage assessment and policy making. Various recent approaches have used advanced ML techniques and RS during various stages of flood management.

Flood forecasting requires accurate estimation of rainfall. Although satellite RS has limited direct applicability to flood forecasting, it has been widely used for precipitation estimation, which is an important input for flood forecasting models. In the late 90s, Tsintikidis et al. (1997) used a shallow neural network with one hidden layer to estimate rainfall from a passive microwave radiometer SSM/I data. The network considered brightness temperature and associated polarization information as inputs and it output the rainfall rates. A random forest based ML algorithm was used to estimate the precipitation which used satellite-derived information on cloud-top height, cloud-top temperature, cloud phase, and cloud water path retrieved from the Meteosat Second Generation (MSG) Spinning Enhanced Visible and Infrared Imager (SEVIRI) (Kühnlein et al., 2014). Recently, Shi et al. (2015) proposed a spatio-temporal sequence forecasting approach using Convolutional Long-Short Term Memory (ConvLSTM) with RADAR echo data in 2D from a ground-based RADAR for precipitation nowcasting by forecasting the RADAR echo data. Pan et al. (2019) proposed a Convolutional Neural Network (CNN) based approach to improve the precipitation estimates from numerical weather prediction (NWP) models. The authors stated that the method outperformed reanalysis precipitation products as well as statistical downscaling (SD) products using linear regression, nearest neighbors, random forests, or fully connected deep neural networks. In an another recent work, Hayatbini et al. (2019) proposed a precipitation estimation framework using a fully convolutional neural network and the advanced baseline imager data from GOES-16, a multispectral geostationary satellite. Specifically, they proposed that the U-net CNN architecture could perform rain/no-rain classification using satellite imagery. The study was based on the earlier work of Hong et al. (2004) on precipitation estimation using remote sensing data and an ANN.

Flash flood susceptibility mapping is another important process in flood risk assessment. Recently, Costache et al. (2019) used a Digital Elevation Model (DEM) with 30 m spatial resolution obtained from Shuttle Radar Topography Mission (SRTM), and which was developed using the technique called SAR interferometry, to derive seven flash-related conditioning factors such as slope angle, aspect, profile curvature, and other factors. In addition, the authors used aerial imagery from Google Earth to delineate the torrential areas along with the land use/cover data, CORINE, which was derived from Sentinel-2 and Landsat-8 RS images. K-nearest neighbors (kNN), K-Start (KS), and Anlytical Hierarchy Process (AHP) algorithms were then applied to obtain the flash-flood susceptibility mapping. Thus, RS techniques played a crucial role in obtaining eight out of 10 flash-flood conditioning factors. In a similar work, Shahabi et al. (2020) used a ML ensemble method with four different k-nearest neighbor (kNN) algorithms for flood detection and susceptibility mapping. Authors used Sentinel-1 images to generate the flood inventory and SRTM DEM to obtain various flood-related conditioning factors. These two works show that ML ensemble methods are gaining traction in flood susceptibility mapping.

Mapping of flooded areas is important to performing damage assessment, deploying rescue and relief operations and developing policies. An example of applying RS and ML to this undertaking is Feng et al. (2015), who developed a random forest based approach to map accurately a flooded area using high-resolution (0.2 m) imagery obtained from Unmanned Areal Vehicle (UAV) imagery. The data were obtained for Yuyao City of Zhejiang Province in Eastern China during the flooding that occurred due to the extreme rainfall event on October 7, 2013. Additionally, Jain et al. (2020) developed a hybrid approach to combine the strength of the traditional water indices from RS imagery and generalization capability of Convolutional Neural Networks (CNN). The authors proposed a new water index which minimized cloud interference in the RS image and used it with a pre-trained VGG-16 model (Simonyan and Zisserman, 2014) and a transfer learning based approach to re-train the model for a new task of flood water detection. In a similar work, Potnis et al. (2019) used an Encoder-Decoder Neural Network based on the Efficient Residual Factorized Convnet (ERFNet) architecture for multi-class segmentation of urban floods satellite imagery from WorldView-2 of floods in Srinagar, India during September 2014. Recently, Jiang et al. (2020) proposed an approach to obtain waterlogging depth from video images using CNN. The approach generated synthetic images from the set of images of reference objects and flood surface, which was further used to train the CNN model to obtain the waterlogging depth. This method can also be employed to obtain waterlogging depth from the images taken of the flooded area using recent drone-based video surveillance. Cervone et al. (2017) added to these techniques a methodology to fuse social media data with the RS data during a flood situation to improve the flood mapping capability.

Recently, a few approaches to model the semantics in RS images were proposed for flood detection and mapping. Kurte et al. (2017) proposed a semantics enabled framework to model the spatial relationships among various regions in the RS images to enable spatial-relationships-based queries such as *Retrieve all images in the ALI repository having Built Up region externally connected to the Stagnated Flood Water*. Later this work was extended to accommodate the temporal aspect to enable the spatio-temporal semantic queries such as *Show road segments which were completely submerged during 9th September 2014 to 22nd September 2014* (Kurte et al., 2019). In a similar semantics based approach, Potnis et al. (2018) developed a flood scene ontology (FSO) which formally defines complex classes such as *Flooded_Residential_Buildings*, *Accessible_Residential_Buildings*, *Operational_Roads*. After detecting various objects in the RS imagery using any supervised classification approach, the ontology can be used to infer complex classes which are very important for flood mapping.

2.5.2. Water Quality Monitoring

RS data has been used over the past 50 years to monitor water quality. For instance, RS data can be used to measure water turbidity, or lack of transparency, which is a good measure of the water quality. Clear water shows high absorptivity in the infrared and near-infrared wavelength regions. It also shows some reflectivity in the visible regions. Reflectivity in this application can reveal variations in water quality due to salinity, temperature, and turbidity. In the past decade, much research has been published in which remote sensing and ML approaches are used to estimate additional water quality parameters. For example, Dogan et al. (2009) explored the non-linear capability of ANN to improve the accuracy of biological oxygen demand (BOD) estimation. Wu et al. (2014) compared multiple regression (MR) with ANN for total suspended solid (TSS) turbidity estimations using data measured with a hyperspectral spectroradiometer and found that the non-linear transformation function of ANN performed better than MR. Wang et al. (2011) used the support vector regression (SVR) method to retrieve various water quality estimators from SPOT-5 satellite data. SVRs showed potential in solving problems with small sample size, non-linearity, or high dimension (Vapnik, 1995). Huo et al. (2014) stated that the lakes near urban areas or inside urban areas are becoming eutrophied or even hypereutrophied due to excessive urbanization and a fast growing economy. The authors used genetic algorithms combined with support vector machines (GA-SVM) to build an inversion model for eutrophic indicators such as Chl-a from Landsat ETM imagery. They showed that the GA-SVM based method had better prediction accuracy than the traditional statistical regression methods and ANN based approaches. According to Sharaf El Din et al. (2017), modeling water quality using satellite data is a complex problem, and conventional regression-based approaches can not perform well while modeling such complex relationships between water quality and RS data. The authors claimed that the proposed Landsat8-based-BPNN—back propagation neural network—to estimate water quality (both optical and non-optical) worked better than SVM-based methods. Moreover, the authors mentioned that, compared to the BPNN-based methods, the SVM-based methods could produce very different results due to differences in parameter selections, kernel-selection, high algorithmic complexity, and extensive memory requirement. The developed model showed $R^2 > 0.9$ for the water quality indicators such turbidity, total suspended solids (TSS), chemical oxygen demand (COD), biological oxygen demand (BOD), and dissolved oxygen (DO). Recently, Hafeez et al. (2019) compared several ML techniques including artificial neural networks, random forests, cubist regression, and support vector regression for estimating the concentrations of suspended solids (SS), Chl-a, and turbidity using Landsat data. The results showed that the ANN-based model achieved the highest accuracy in estimating the above mentioned water quality indicators. In an another recent study, Govedarica and Jakovljević (2019) used 4-years of time-series data of *in-situ* monitoring of surface water bodies for the calibration and validation of a water quality estimation based on SVM and ANN algorithms using Landsat 8 data. The work also compared the estimations based on Landsat 8 with the Sentinel-2 data and found that, due to higher spatial and spectral resolution, Sentinel-2 data is a better alternative for water quality monitoring. Interestingly, the results showed that SVM produced more accurate results than ANN when used with Landsat data, whereas ANN provided better estimation accuracy for turbidity and TSS than SVM, and lower accuracy for TN and TP than SVM when used with Sentinel-2 data. Finally, Wang et al. (2017) conducted a study that combined a ML algorithm and remote sensing spectral indices [difference index (DI), ratio index (RI), and normalized difference index (NDI)] through fractional derivatives methods and in turn establishes a model for estimating and assessing the water quality index (WQI) (2.3). For this study, the WQI was calculated using sensitive wave bands and a spectral index of hyperspectral data, and particle swarm optimization (Kennedy and Eberhart, 1995; Shi and Eberhart, 1998)—support vector regression models (PSO-SVR), which deploy a population of candidate solutions over the SVR search space. Through comparisons of the predictive effects of the 22 water quality index estimations determined by the PSO-SVR, Wang et al. (2017) demonstrated that the model based on RI, DI, and NDI values of the 1.6 order was better performing than the others for predicting the water quality index of the semi-arid area of central Asia [R2 (0.92), RMSE = 58.4, RPD (2.81) and a slope of curve fitting of 0.97].

2.5.3. Impervious Surface Detection

Urban impervious surfaces such as roads, driveways, sidewalks, and parking lots prevent water from infiltrating into soil, which has impacts on urban hydrology, groundwater, and water quality. Impervious surfaces facilitate pollutant's movements to nearby water bodies during heavy rain and urban flooding (Hall and Hossain, 2020). In the context of ML, identifying impervious surfaces from RS data is fundamentally a classification approach. However, many index-based approaches for sighting impervious surfaces using RS (e.g., Weng, 2012) focus on the developments in this area that use ML algorithms. Recently, Yao et al. (2017) adopted a one-class classification approach to detect impervious surfaces using high-resolution GF-1 satellite images, and found that Presence and Background Learning (PBL) and Positive Unlabeled Learning (PUL) outperformed SVM models in detecting impervious surfaces. Miao et al. (2019) also used a one class classification technique and Landsat-8 imagery for impervious surface classification. In a similar study, Bian et al. (2019) used a random forest algorithm and time-series data from multiple satellites HJ-1A/B and GF-1/2 to estimate the changes in the impervious surface percentage over the years 2009–2017. Lin et al. (2019) addressed the challenges in detecting impervious surfaces due to the diversity of land use and shadow effects in high-resolution satellite imagery using a dictionary sparse representation classification and data fusion approach with WV-2, GeoEye-1, TerraSAR-X, and LiDAR. Zhang H. et al. (2019) addressed similar issues by using a deep CNN approach with data fusion from optical and SAR satellites WV-3, Sentinel-2, and Radarsat-2. Similar other works, Sun et al. (2019) (used 3D CNN with WV-3 and LiDAR), McGlinchy et al. (2019) (used UNet with WV-2), show increasing trends of using deep learning based approaches with multi-satellite data fusion.

3. IDENTIFICATION AND ASSESSMENT OF MULTI-HAZARD RISK

Multi-hazard identification and compound risk assessment inform effective planning activities and strategies (FEMA, 2015), and help water managers prioritize attention, investment, and recourse (Dickson-Anderson et al., 2016) to target the most urgent and the highest impact risks. Risk is defined as a combination of hazard, exposure, and vulnerability (Garrick and Hall, 2014). Because exposure in urban areas is relatively high due to the high density of population and man-made structures (Hoekstra et al., 2018), cities without proper preparedness and adaptation strategies are vulnerable to a wide variety of urban water hazards (Shaw et al., 2016; Eldho et al., 2018; Hoekstra et al., 2018; Gangrade et al., 2019; Rahmasary et al., 2019) that are often causally linked to further hazards. Additionally, coincidental hazards may occur, resulting in a compounding effect overwhelming the ability of local or national governments to respond (Liu and Huang, 2014). For example, a specific urban water hazard such as flooding can lead to multiple risks (Dai et al., 2017; Cook et al., 2019) that include inundation of building structures, damage to infrastructure, and/or the spread of water-borne diseases (Gangrade et al., 2018; Pereira, 2018). Consequently, multi-hazard risk assessment techniques must be conducted in the urban water management sector in a manner that considers the combined effects and interactive reactions of multiple urban water hazards in urban areas (Garcia-Aristizabal and Marzocchi, 2013; Gruber and Mergili, 2013; FEMA, 2015; Karlsson et al., 2017).

Despite its usefulness for hazard mitigation planning, multi-hazard risk assessment has been under-emphasized in natural disaster management and planning (Rahmati et al., 2019) due to the difficulty of analyzing the risk for more than one hazard in the same area, and of analyzing their interaction. In the past, studies have focused primarily on forecasting and controlling hazards, and their physical processes (Kalantari et al., 2019) in natural areas, without considering the social and economic impacts of these hazards in urban areas (e.g., hazard effects on buildings, infrastructures, and agriculture). Previous studies, which intended to analyze hazard risk and social vulnerabilities, only analyzed the risks of single hazards separately (Bühler et al., 2013; Statham et al., 2017) using physical or statistical models [e.g., flood impact using the HEC-FIA model (Lehman and Light, 2016) or economic damage to fisheries caused by surface water pollution using AQUATOX model (Park et al., 2008)]. In general, most past studies do not consider the multi-hazard chain (hazard interaction) and the combined risk of coupled hazard events (Garcia-Aristizabal and Marzocchi, 2013; Rahmati et al., 2019). Although a few studies (Freeman and Warner, 2001; Newman et al., 2017) analyze the components of different types of vulnerability and risk by evaluating physical, social, and economic consequences of a chain of urban hazards, developing a systematic approach for multi-hazard risk assessment using conventional modeling methods faces multiple challenges. These challenges are primarily associated with (a) integrating multiple physical or statistical models and domain-data that only target single hazards to simulate a multi-hazard chain and predict the combined effect of multiple urban water hazards, and (b) in-depth understanding of hazards, including interconnections between different hazards, and dynamics behind multiple hazards. In the presence of hydro-complexities, many underlying mechanisms of urban water hazards remain unknown. Therefore, conventional methods based on physical modeling alone may not be the best way to assess multi-hazard risk in urban water systems.

In recent years, advanced ML methods have been used to develop innovative multi-hazard risk assessment frameworks and workflows, which are able to address the challenges associated with conventional risk assessment techniques. The feasibility of applying ML to multi-hazard risk assessment is shown by the following: (a) ML is a subfield of artificial intelligence and data-driven analysis where ML models can easily identify trends, patterns, and empirical relationships in a large volume of data without considering detailed physical processes behind a phenomenon, such as the interactive reactions between multiple water hazards (Dibike and Solomatine, 2000; Rahmati et al., 2019), and (b) ML models are capable of handling data that are multi-dimensional and multi-domain (Anzai, 2012). In this section, we review several ML workflows and applications that are designed to support the analysis of multi-hazard risk for mitigating water-related hazards.

For example, Rahmati et al. (2019) investigated and mapped multi-hazard exposure using several ML models including BRT (Boosted Regression Trees), GAM (Generalized Additive Model, a regression which can include linear or non-linear predictor variables and predicted values potentially following any of a variety of probability distribution functions), and SVM (Support Vector Machines), and they evaluated the performance of these ML models using threshold-dependent and threshold-independent methods. The study consists of several steps: (1) selection of predictive factors for modeling multiple hazards (e.g., flood, landslide, soil erosion, and debris flow), (2) creation of Multi-Hazard Inventory using records from road organization and the regional water company (RWC) to document the occurrence of various hazards, (3) application of ML models to predict and map the exposure of multiple hazards, and (4) evaluation of the accuracy of these models. The results of this study indicate that (a) different ML models differed in their accuracy of predicting the different hazards (Rahmati et al., 2019), and (b) the applied ML models are useful and generalizable for multi-risk mapping around the world.

Another example of a multi-hazard multi-model approach is Chen et al. (2019), in which the researchers evaluate the risk of regional flood disaster in the Yangtze River Delta (YRD) region. Based on the driving force, pressure, state, impact, and response (DPSIR) conceptual framework, the study first applies a random forest algorithm to screen important indices of flood risk. They then construct a radial basis function (RBF) neural network to evaluate the flood risk level. In this study, the radial basis function is the activation function for the ANN. The study approaches the urban flood risk assessment as a multi-classification problem using ML methods and indicates that only a few of the previous studies use ML theory to assess the urban flood disaster risks that are complex and associated with multiple sources and contributing factors. The study concludes that the level of urban flood disaster is closely related to rainfall, topography, economic development, land use, soil erosion, urban flood control investment, and disaster emergency response capability, shedding light on effective regulation measures for improving flood prevention in urban environments.

3.1. Exploration of Complex and Interconnected Hazards and Risks

To explore complex and interconnected hazards and risks, Xu et al. (2019a) present a visual analytics framework that combines various types of ML applications (e.g., feature selection, classification, and multivariate clustering analysis) with different geo-visualization techniques to analyze multi-hazard risk at culverts due to flooding and sedimentation. ML models applied in this study include the classification schemes, random forests and Self Organizing Maps (SOM), and are used for exploratory data analysis, aiming to improve the understanding of the factors and interconnected hazards (e.g., flooding, excessive erosion, and sediment transport in rivers) that contribute to the sedimentation and flood over-topping of culverts (transportation infrastructure). The results of the study show that ML application can be used not only for multi-risk assessment and hazard prediction but also for exploring the complex and interconnected processes behind multiple hazards. Additionally, the same framework can be readily extended to analyze multiple hazards at other hydraulic structures, such as bridges and weirs. Pourghasemi et al. (2020) presented a ML workflow, debuted as the Sendai framework, for assessing and mapping multi-hazard risk susceptibility, with an overall objective of reducing hazard risk and increasing sustainable development in urban areas. The workflow entails three main steps: (1) data preparation for obtaining the location of various hazards (floods, forest fires, and landslides), (2) recognition of the most important factors contributing to the occurrence of different hazards using the Boruta algorithm (a wrapper around random forest classification that iteratively removes irrelevant features from the data), and (3) construction of multi-hazard susceptibility maps along with validation processes using the random forest model and the preparation of a Multi-hazard Probability Index (MHPI) for the study area. The significance of the Sendai framework is that it (a) creates a reasonable understanding of the factors controlling flood and forest fire through ML-powered variable ranking and landslide occurrence, and (b) produces a multi-hazard probability map for facilitating integrated and comprehensive watershed management and land use planning.

3.2. Hybrid Modeling for Multi-Hazard Risk Assessment

A few researchers have applied hybrid models to water-related multi-hazard risk assessment. For example, Yang T. et al. (2019) used long short-term memory units (LSTM) to improve the timing component of the amplitude of peak discharge for flood simulations produced with global hydrological models over different climate zones. Hajigholizadeh et al. (2018) used hybrid models for predicting and assessing water erosion vulnerability and risks, as well as for the optimization of management strategies for agricultural or soil and water conservation practices. Application of hybrid models to these multi-hazard hydrological risks is still emerging within the domain, but the utility of this approach continues to be demonstrated across a variety of hydrological applications.

4. SELECTION OF BEST MANAGEMENT PRACTICES

The proper selection and placement of Best Management Practices (BMPs) is a critical planning process that helps many watershed and urban planning communities effectively mitigate water-related hazards and manage urban water resources (e.g., stormwater management, water pollution reduction, and erosion controls) (Cheng et al., 2006; NRCS, 2011; USEPA, 2018). These BMPs are carefully selected from a pool of planning and mitigation alternatives that exists in various forms. Based on their spatial scales, these alternatives can be categorized as either localized alternatives, which are city-scale practices for protecting the municipal water supply and infrastructure through structural actions and non-structural actions, and watershed alternatives, which represent the management of land cover and land-use at the watershed scale (Carson et al., 2018). The selection of BMPs is a complex multi-objective optimization problem that requires the consideration of multiple planning objectives and criteria, which aim to maximize the environmental and social benefits for multiple urban communities, while minimizing the economic cost for the implementation of these management practices (Maringanti et al., 2008; Rodriguez et al., 2011). The development and advancement in GA (section 2.4.1) have provided watershed management communities with a method for solving complicated optimization problems that are associated with the selection of BMPs. GA are capable of handling complex and irregular solution spaces when searching for a global optimum (Chambers, 2000; Rodriguez et al., 2011) in a multiobjective optimization. Multiobjective optimization has been defined as "vector optimization" (Cohon and Marks, 1975) for which the objective function is a vector containing scalar objectives subject to a set of constraints, and for which Pareto optimal solutions show the best performance. Reed et al. (2013) evaluated a variety of multiobjective optimization GA as applied to rainfall-runoff calibration, long-term groundwater monitoring, and risk-based water supply portfolio planning. They found five best performing algorithms, of which their high-performance adaptive search Borg algorithm (Hadka and Reed, 2013) was the most scalable and the best performing, and has shown particular stakeholder usefulness in its incorporation into a visual and interactive decision support framework (Reed and Kollat, 2013).

In the water quality management sector, several studies applied GA-based optimization models to find optimal solutions to water quality problems for several watersheds in the United States by connecting non-point pollution reduction models with economic components (Srivastava et al., 2002; Chen et al., 2015). In the stormwater management sector, Limbrunner et al. (2013) applied classic optimization techniques to stormwater and non-point source pollution management at the watershed scale, and compared their effectiveness for finding optimal solutions to that of genetic algorithms, and linear and dynamic programming. Dynamic programming proved to find the most efficient solution to the sediment-management-optimization problem.

In addition to the optimization of planning alternatives, ML methods can enable selection for optimal management practices (Savic, 2019). AI-driven applications are envisioned to learn from the human decision-making process, during which best management practices are selected by planners and watershed managers based on their past experiences.

5. VISION: NEW APPLICATIONS OF MACHINE LEARNING TO URBAN WATER SECURITY

In order to ensure high-quality and timely water availability in the right quantities for urban areas, water resources must be managed well. In order for water resources to be managed well, a planning system leading to actions that promote sustainability and urban water security must be in place at the municipal level. We have shown that ML can help with this system as it applies to every stage of disaster management and planning, as outlined sequentially on the left hand side of **Figure 2** and shown as an interconnected and cyclical process on the right side. That is, we have outlined a variety of ML applications for facilitating the individual disaster management stages and planning processes. For long-term planning and mitigation, we have presented studies that use ML methods to identify and assesses multi-hazard risks and vulnerability in urban water systems, taking into account socio-economic factors and the multi-hazard chain. We have also discussed how ML can help optimize the selection of urban best management practices for reducing water pollution and supporting storm water management. For early warning and hazards prediction, we have examined a range of ML applications for supporting the prediction of various water-hazard related parameters. We included studies that combine ML methods with process-based models (e.g., conceptual and physics-based hydrological and sediment transport models) into hybrid models to increase the accuracy and speed of the predictions for water hazard-related parameters. We have also discussed how

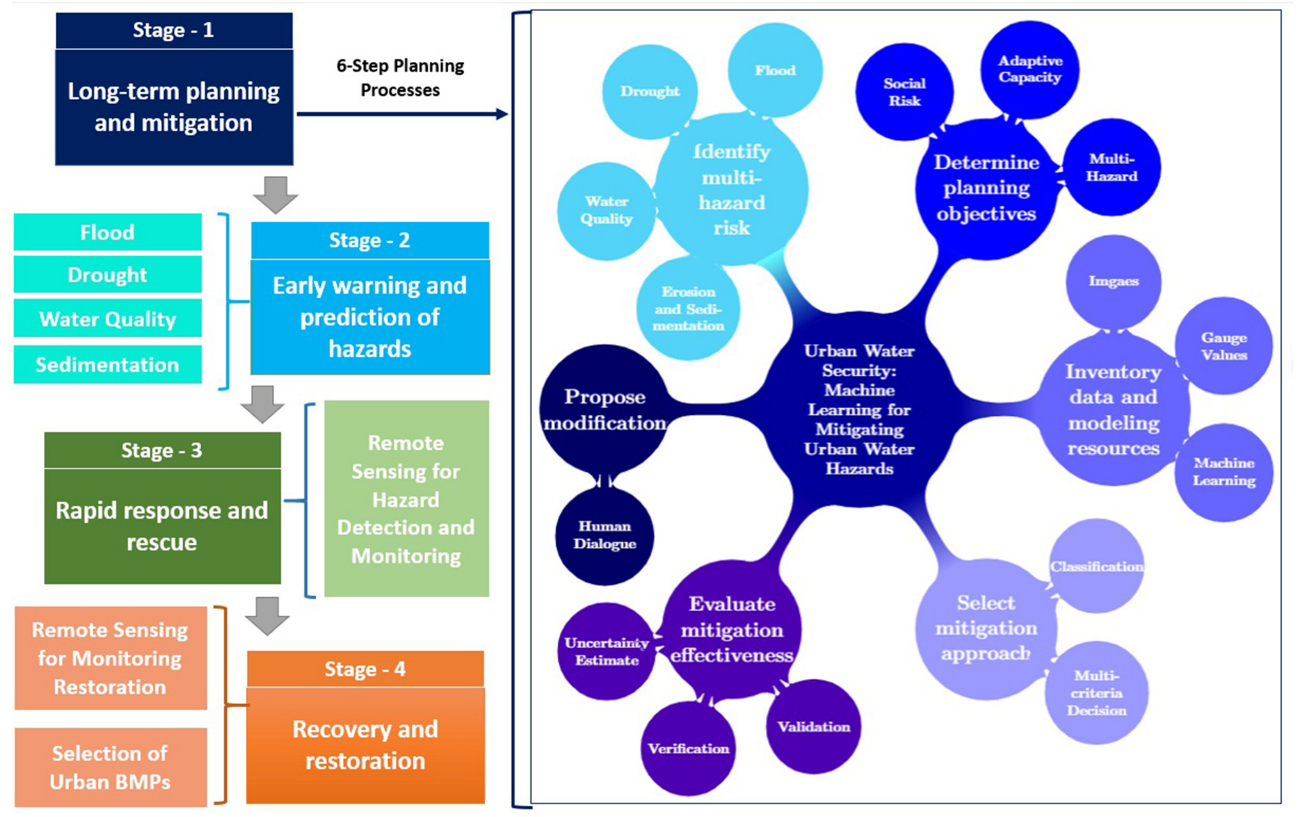

Stage - 1
Long-term planning and mitigation
6-Step Planning Processes
Flood
Drought
Water Quality
Sedimentation
Stage - 2
Early warning and prediction of hazards
Stage - 3
Rapid response and rescue
Remote Sensing for Hazard Detection and Monitoring
Remote Sensing for Monitoring Restoration
Selection of Urban BMPs
Stage - 4
Recovery and restoration
Urban Water Security: Machine Learning for Mitigating Urban Water Hazards
Identify multi-hazard risk
Drought
Flood
Water Quality
Erosion and Sedi-mentation
Determine planning objectives
Social Risk
Adaptive Capacity
Multi-Hazard
Inventory data and modeling resources
Imgaes
Gauge Values
Machine Learning
Select mitigation approach
Classification
Multi-criteria Decision
Evaluate mitigation effectiveness
Uncertainty Estimate
Verification
Validation
Propose modification
Human Dialogue

FIGURE 2 | Potential ML opportunities for improving both the generic hazard mitigation stages **(left)** and detailed long-term planning steps **(right)**.

innovative combinations of ML and remote sensing technologies can improve the discovery and extraction of useful hazard information and features that are critical to early-warning, rapid response and rescue, and recovery and restoration.

Our vision is that these methods can be combined into ML water management workflows that build on those already in use for characterizing and predicting multi-hazard hydrological events. By weaving together the ML methods we have described, long-term management processes including the six steps shown on the right hand side of **Figure 2** and outlined in the introduction can be captured. For example, risks associated with flood, drought and water quality can be identified using genetic algorithms, artificial neural networks, support vector machines, random forests, and other types of regression and hybrid models. Then planning objectives can be determined by weighing social risk and adaptive capacity using agent-based models, boosted regression trees, generalized additive models, and support vector machines. To inventory data, ground-based and satellite-based data can be reckoned, cataloged, and formatted for use in spatial-relationships-based queries, k-nearest neighbors, analytical hierarchy processes, and convolutional neural networks. To select mitigation approaches, classification schemes can be used along with multi-criteria decision methods. Uncertainty estimates can be used to evaluate the mitigation approaches selected. Finally, the insight gained from the ML results may be discussed by the planners to modify and implement the approaches determined.

ML is often not the first choice of analytical tools for planners for a variety of reasons. The first is that reasonably robust methods with known uncertainty for analyzing water risks are well established and accepted in the water management community. ML methods are less proven even if they often can perform better on data than the traditional methods. To address the uncertainty in ML methods, some researchers (e.g., Morrison et al., 2003; Duncan, 2014) use metrics such as Receiver Operating Characteristic Curves for scoring the diagnostic ability of a binary (or higher dimensional) classifier system, or alternative goodness-of-fit measures for evaluating the reliability of ML output. Others (e.g., Munafò and Smith, 2018) suggest a method of investigation called *triangulation*, in which multiple approaches (at least 3) are used to address one question. The uncertainty associated with a complete model chain is large, especially at the required level of decision-making under climate change, urbanization (Dessai et al., 2009), and the accumulation of uncertainty at each level of the assessment (Merz et al., 2010). However, while each ML method may have its own strengths, weaknesses, and unrelated assumptions, uncertainty quantification can help assign some degree of confidence to results obtained.

We observe that many aspects of urban water security and hazard modeling are still underrepresented as ML problems, in particular, those pertaining to the prediction of indirect effects of water-related hazards and their associated risks. Additionally, the use of ML techniques often requires additional mathematical and computational training (and often large high performance compute resources) beyond traditional statistical methods, and time constraints of working water managers may not allow for this additional training. Nevertheless, understanding the development of sustainable urban water management planning, we can draw lessons from history and devise sensible approaches for the future that include ML. If we view hydrological systems as "structurally co-constituted of natural, engineered, and social elements," (Brelsford et al., 2020), we may more readily employ ML to integrate disparate data and discover new perspectives on management practices based on the new patterns these methods reveal. In the near future, We also envision an increase in the applications of the hybrid modeling approaches (i.e., theory-guided ML) (Mekonnen et al., 2012; Karpatne et al., 2017; Frame, 2019) in the urban water management sector through the integration of data-driven ML methods and conventional process-based domain models.

AUTHOR CONTRIBUTIONS

MA-D suggested the focus on urban hydrology, co-wrote the introduction and vision sections, and wrote the text on machine learning for flooding. HX developed the outline and organization of the paper based on urban water management practices, co-wrote the introduction and vision, and wrote the text for multi-hazard risk, soil erosion and sediment transport, and selection of best management practices. KK co-wrote the introduction, provided the research trend analysis on the historical application of machine learning to water management and hazard, and wrote the text on applications of machine learning to disaster management using remote sensing. DR wrote the sections on machine learning application to drought and water quality characterization and prediction. All authors edited and revised the document throughout.

FUNDING

Funding was provided by the Oak Ridge National Laboratory Computational Science and Engineering Division.

REFERENCES

Abaci, O., and Papanicolaou, T. (2009). Long-term effects of management practices on water-driven soil erosion in an intense agricultural sub-watershed: monitoring and modelling. *Hydrol. Process*. 23, 2818–2837. doi: 10.1002/hyp.7380

Abdollahzadeh, A., Mukhlisin, M., and Shafie, A. (2011). Predict soil erosion with artificial neural network in Tanakami (japan). *WSEAS Trans. Comput*. 10, 51–60.

Abdulkadir, T. S., Muhammad, R. U. M., Yusof, K. W., Ahmad, M. H., Aremu, S. A., Gohari, A., et al. (2019). Quantitative analysis of soil erosion causative factors for susceptibility assessment in a complex watershed. *Cogent Eng*. 6:1594506. doi: 10.1080/23311916.2019.1594506

Aboelnga, H. T., Ribbe, L., Frechen, F.-B., and Saghir, J. (2019). Urban water security: Definition and assessment framework. *Resources* 8:178. doi: 10.3390/resources8040178

Adamala, S. (2017). An overview of big data applications in water resources engineering. *Mach. Learn. Res*. 2, 10–18. doi: 10.11648/j.mlr.20170201.12

AghaKouchak, A. (2015). A multivariate approach for persistence-based drought prediction: application to the 2010-2011 east Africa drought. *J. Hydrol*. 526, 127–135. doi: 10.1016/j.jhydrol.2014.09.063

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., et al. (2019). Machine learning methods for better water quality prediction. *J. Hydrol*. 578:124084. doi: 10.1016/j.jhydrol.2019.124084

Ahmed, N., Mahmud, S., Elahi, M., Ahmed, S., and Sujauddin, M. (2018). Forecasting river sediment deposition through satellite image driven unsupervised machine learning techniques. *Remote Sens. Appl. Soc. Environ*. 133, 435–444. doi: 10.1016/j.rsase.2018.12.011

Ait Kadi, M., and Arriens, W. L. (2012). "Increasing water security: a development imperative," in *Perspectives Paper of the Global Water Partnership Technical Committee*. Stockholm: Global Water Partnership.

Allen, M. R., Zaidi, S. M. A., Chandola, V., Morton, A. M., Brelsford, C. M., McManamay, R. A., et al. (2018). A survey of analytical methods for inclusion in a new energy-water nexus knowledge discovery framework. *Big Earth Data* 2, 197–227. doi: 10.1080/20964471.2018.1524344

Altunkaynak, A. (2009). Sediment load prediction by genetic algorithms. *Adv. Eng. Softw*. 40, 928–934. doi: 10.1016/j.advengsoft.2008.12.009

Anzai, Y. (2012). *Pattern Recognition and Machine Learning*. Elsevier: Amsterdam.

Auerbach, D., Buchanan, B., Alexiades, A., Anderson, E., Encalada, A., Larson, E., et al. (2015). Towards catchment classification in data-scarce regions. *Ecohydrology*. 9, 1235–1247 doi: 10.1002/eco.1721

Azamathulla, H., Ab Ghani, A., and Fei, S. (2012). Anfis-based approach for predicting sediment transport in clean sewer. *Appl. Soft Comput*. 12, 1227–1230. doi: 10.1016/j.asoc.2011.12.003

Azamathulla, H., Asce, M., and Ab Ghani, A. (2011). Anfis-based approach for predicting the scour depth at culvert outlets. *J. Pipeline Syst. Eng. Practice* 3, 1227–1230. doi: 10.1061/(ASCE)PS.1949-1204.0000066

Babovic, V., Canizares, R., Jensen, H., and Klinting, A. (2001). Neural networks as routine for error updating of numerical models. *J. Hydraul. Eng. ASCE* 127:3. doi: 10.1061/(ASCE)0733-9429(2001)127:3(181)

Badrzadeh, H., Sarukkalige, R., and Jayawardena, A. (2013). Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting. *J. Hydrol*. 507, 75–85. doi: 10.1016/j.jhydrol.2013.10.017

Bakhtyar, R., Ghaheri, A., Yeganeh-Bakhtiary, A., and Baldock, T. (2008). Longshore sediment transport estimation using fuzzy inference system. *Appl. Ocean Res*. 30, 273–286. doi: 10.1016/j.apor.2008.12.001

Barzegar, R., Aalami, M. T., and Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model. *Stochast. Environ. Res. Risk Assess*. 34, 1–19. doi: 10.1007/s00477-020-01776-2

Belayneh, A., and Adamowski, J. (2012). Standard precipitation index drought forecasting using neural networks, wavelet neural networks, and support vector regression. *Appl. Comput. Intell. Soft Comput*. 2012:794061. doi: 10.1155/2012/794061

Belayneh, A., Adamowski, J., Khalil, B., and Ozga-Zielinski, B. (2014). Long-term SPI drought forecasting in the awash river basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *J. Hydrol*. 508, 418–429. doi: 10.1016/j.jhydrol.2013.10.052

Belayneh, A., Adamowski, J., Khalil, B., and Quilty, J. (2016). Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmos. Res*. 172, 37–47. doi: 10.1016/j.atmosres.2015.12.017

Beran, B., and Piasecki, M. (2008). Availability and coverage of hydrologic data in the US geological survey national water information system (NWIS) and us environmental protection agency storage and retrieval system (storet). *Earth Sci. Inform*. 1, 119–129. doi: 10.1007/s12145-008-0015-2

Bertolotto, M., Martino, S. D., Ferrucci, F., and Kechadi, T. (2007). Towards a framework for mining and analysing spatio-temporal datasets. *Int. J. Geogr. Inform. Sci*. 21, 895–906. doi: 10.1080/13658810701349052

Bhattacharya, B., Price, R., and Solomatine, D. (2007). A machine learning approach to modeling sediment transport. *J. Hydraul. Eng. ASCE* 133, 440–450. doi: 10.1061/(ASCE)0733-9429(2007)133:4(440)

Bian, J., Li, A., Zuo, J., Lei, G., Zhang, Z., and Nan, X. (2019). Estimating 2009-2017 impervious surface change in Gwadar, Pakistan using the HJ-1a/b constellation, GF-1/2 data, and the random forest algorithm. *ISPRS Int. J. Geo-Inform*. 8:443. doi: 10.3390/ijgi8100443

Brelsford, C., Dumas, M., Schlager, E., Dermody, B. J., Aiuvalasit, M., Allen-Dumas, M. R., et al. (2020). Developing a sustainability science approach for water systems. *Ecol. Soc*. doi: 10.5751/ES-11515-250223

Bühler, Y., Kumar, S., Veitinger, J., Christen, M., Stoffel, A., and Snehmani (2013). Automated identification of potential snow avalanche release areas based on digital elevation models. *Nat. Hazards Earth Syst. Sci*. 13, 1321–1335. doi: 10.5194/nhess-13-1321-2013

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., and Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Tot. Environ*. 2020:137612. doi: 10.1016/j.scitotenv.2020.137612

Cao, L., and Gu, Q. (2002). Dynamic support vector machines for non-stationary time series forecasting. *Intell. Data Anal*. 6, 67–83. doi: 10.3233/IDA-2002-6105

Cardwell, H. E., Langsdale, S., and Stephenson, K. (2009). "Developing best practices for computer aided dispute resolution," in *World Environmental and Water Resources Congress May 17* (Kansas City, MI: American Society of Civil Engineers (ASCE) Library). doi: 10.1061/41036(342)484

Carlisle, D., Falcone, J., and Meador, M. (2008). Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds. *Environ. Monitor. Assess*. 151, 143–160. doi: 10.1007/s10661-008-0256-z

Carson, A., Windsor, M., Hill, H., Haigh, T., Wall, N., Smith, J., et al. (2018). Serious gaming for participatory planning of multi-hazard mitigation. *Int. J. River Basin Manage*. 16, 1–47. doi: 10.1080/15715124.2018.1481079

Cervone, G., Schnebele, E., Waters, N., Moccaldi, M., and Sicignano, R. (2017). "Using social media and satellite data for damage assessment in urban areas during emergencies," in *Seeing Cities Through Big Data*, eds P. Thakuriah, N. Tilahun, and M. Zellner (Cham: Springer), 443–457. doi: 10.1007/978-3-319-40902-3_24

Chambers, L. (2000). *The Practical Handbook of Genetic Algorithms: Applications, 2nd Edn* Boca Raton, FL: CRC Press. doi: 10.1201/9781420035568

Chandola, V., Cheboli, D., and Kumar, V. (2009a). Anomaly detection: a survey. *ACM Comput. Surveys* 41, 1–58. doi: 10.1145/1541880.1541882

Chandola, V., Cheboli, D., and Kumar, V. (2009b). *Detecting Anomalies in a Time Series Database*. Computer Science Department, University of Minnesota, Tech. Rep.

Chang, F.-J., Hsu, K., and Chang, L.-C. (2019). *Flood Forecasting Using Machine Learning Methods*. MDPI.

Chen, C.-Y., Li, Wang, and Deng (2019). A machine learning ensemble approach based on random forest and radial basis function neural network for risk evaluation of regional flood disaster: a case study of the Yangtze river delta, China. *Int. J. Environ. Res. Public Health* 17:49. doi: 10.3390/ijerph17010049

Chen, L., Wei, G., and Shen, Z. (2015). An auto-adaptive optimization approach for targeting nonpoint source pollution control practices. *Sci. Rep*. 5:15393. doi: 10.1038/srep15393

Chen, S., Dong, S., Cao, Z., and Guo, J. (2020). A compound approach for monthly runoff forecasting based on multiscale analysis and deep network with sequential structure. *Water* 12:2274. doi: 10.3390/w12082274

Cheng, M.-S., Zhen, J., and Shoemaker, L. (2006). BMP decision support system for evaluating stormwater management alternatives. *Front. Environ. Sci. Eng. China* 3, 453–463. doi: 10.1007/s11783-009-0153-x

Cigizoglu, H. (2002). Suspended sediment estimation for rivers using artificial neural networks and sediment rating curves. *Turkish J. Eng. Environ. Sci*. 26, 27–36.

Cohon, J. L., and Marks, D. H. (1975). A review and evaluation of multiobjective programing techniques. *Water Resour. Res*. 11, 208–220. doi: 10.1029/WR011i002p00208

Cook, E. R., Seager, R., Cane, M. A., and Stahle, D. W. (2007). North American drought: reconstructions, causes, and consequences. *Earth Sci. Rev*. 81, 93–134. doi: 10.1016/j.earscirev.2006.12.002

Cook, S., [van Roon], M., Ehrenfried, L., LaGro, J., and Yu, Q. (2019). "Chapter 27 - WSUD 'best in class' Case Studies from Australia, New Zealand, United States, Europe, and Asia," in *Approaches to Water Sensitive Urban Design*, eds A. K. Sharma, T. Gardner, and D. Begbie (Amsterdam: Elsevier; Woodhead Publishing), 561–585. doi: 10.1016/B978-0-12-812843-5.00027-7

Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S., Vojtek, M., et al. (2019). Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and GIS techniques. *Remote Sens*. 12, 1–26. doi: 10.3390/rs12010106

Costello, C., Ensor, C., and Chadwick, S. (2019). *Indirect Losses From Flood Disasters*. Available online at: https://firststreet.org/flood-lab/research/indirect-losses-from-flood-disasters (accessed May 02, 2020).

Crossman, J., Futter, M., Oni, S., Whitehead, P., Jin, L., Butterfield, D., et al. (2013). Impacts of climate change on hydrology and water quality: future proofing management strategies in the Lake Simcoe Watershed, Canada. *J. Great Lakes Res*. 39, 19–32. doi: 10.1016/j.jglr.2012.11.003

Dai, A. (2011). Drought under global warming: a review. *Wiley Interdisc. Rev*. 2, 45–65. doi: 10.1002/wcc.81

Dai, L., Rijswick, H., Driessen, P., and Keessen, A. (2017). Governance of the sponge city programme in china with Wuhan as a case study. *Int. J. Water Resour. Dev*. 34, 1–19. doi: 10.1080/07900627.2017.1373637

Das, M., and Parthasarathy, S. (2009). "Anomaly detection and spatio-temporal analysis of global climate system," in *Proceedings of the Third International Workshop on Knowledge Discovery From Sensor Data* (Paris), 142–150. doi: 10.1145/1601966.1601989

de Goyet, C. D. V., Marti, R. Z., and Osorio, C. (2006). *Natural Disaster Mitigation and Relief. In Disease Control Priorities in Developing Countries*. 2nd edition. The International Bank for Reconstruction and Development/The World Bank.

Dessai, S., Hulme, M., Lempert, R., and Pielke R. Jr. (2009). Do we need better predictions to adapt to a changing climate? *EOS Trans. Am. Geophys. Union* 90, 111–112. doi: 10.1029/2009EO130003

DHI (2003). *A Versatile Decision Support Tool for Integrated Water Resources Management Planning*. Available online at: https://www.slideshare.net/indiawrm/14-dhi-integrated-decision-support-tools-for-water-resources-managementsep17 (accessed April 14, 2020).

Dibike, Y., and Solomatine, D. (2000). River flow forecasting using artificial neural networks. *Phys. Chem. Earth Part B* 26, 1–7. doi: 10.1016/S1464-1909(01)85005-X

Dickson-Anderson, S., Wallace, C., and Newton, J. (2016). Water security assessment indicators: the rural context. *Water Resour. Manage*. 30, 1567–1604. doi: 10.1007/s11269-016-1254-5

Dogan, E., Sengorur, B., and Koklu, R. (2009). Modeling biological oxygen demand of the Melen river in turkey using an artificial neural network technique. *J. Environ. Manage*. 90, 1229–1235. doi: 10.1016/j.jenvman.2008.06.004

Dörendahl, E. I. (2013). "Boundary work and water resources: towards improved management and research practice?," in *ZEF Working Paper Series*, No. 122; Zentrum f?r Entwicklungsforschung (Bonn).

Duncan, A. P. (2014). *The analysis and application of artificial neural networks for early warning systems in hydrology and the environment* (Ph.D. thesis). University of Exeter, Exeter, United Kingdom. doi: 10.13140/RG.2.1.1602.4806

Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., et al. (2014). Global meteorological drought-part 2: seasonal forecasts. *Hydrol. Earth Syst. Sci*. 18, 2669–2678. doi: 10.5194/hess-18-2669-2014

Eldho, T., Zope, P., and Kulkarni, A. (2018). "Urban flood management in coastal regions using numerical simulation and geographic information system," in *Integrating Disaster Science and Management* (Elsevier), 205–219. doi: 10.1016/B978-0-12-812056-9.00012-9

Emerson, K., Nabatchi, T., and Balogh, S. (2012). An integrated framework for collaborative governance. *J. Publ. Administr. Res. Theory* 22:1. doi: 10.1093/jopart/mur011

EPA (2012). *Water Quality Standards 101*. Technical report, U.S. Environmental Protection Agency.

EPA (2019). *Examples of Water Quality Assessments for Watershed Health*. Available online at: https://www.epa.gov/hwp/examples-water-quality-assessments-watershed-health (accessed October 12, 2020).

Eriksson, M., Nutter, J., Day, S., Guttman, H., James, R., and Quibell, G. (2015). Challenges and commonalities in basin-wide water management. *Aquat. Proc.* 5, 44–57. doi: 10.1016/j.aqpro.2015.10.007

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., et al. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in china. *Energy Convers. Manage.* 164, 102–111. doi: 10.1016/j.enconman.2018.02.087

FEMA (2015). *Integrating Disaster Data into Hazard Mitigation Planning. Federal Emergency Management Agency*.

Feng, Q., Liu, J., and Gong, J. (2015). Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier-a case of Yuyao, China. *Water* 7, 1437–1455. doi: 10.3390/w7041437

Fernández, C., Vega, J. A., Fonturbel, T., and Jiménez, E. (2009). Streamflow drought time series forecasting: a case study in a small watershed in North West Spain. *Stochast. Environ. Res. Risk Assess.* 23:1063. doi: 10.1007/s00477-008-0277-8

FHWA (2000). *Stormwater Best Management Practices in an Ultra-Urban Setting: Selection and Monitoring*. Technical report, Federal Highway Administration.

Frame, J. M. (2019). "Toward global terrestrial hydrology with theory guided machine learning," in *Proceedings of the AGU Annual Meeting* (San Francisco, CA: NASA Goddard Space Flight Center).

Francke, T., López-Tarazón, J., and Schröder, B. (2008). Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrol. Process.* 22, 4892–4904. doi: 10.1002/hyp.7110

Freeman, P., and Warner, K. (2001). *Vulnerability of Infrastructure to Climate Variability: How Does This Affect Infrastructure Lending Policies*? The World Bank, ProVention Consortium.

Ganasri, B., and Gowda, R. (2015). Assessment of soil erosion by Rusle model using remote sensing and gis - a case study of Nethravathi basin. *Geosci. Front.* 7, 953–961. doi: 10.1016/j.gsf.2015.10.007

Gangrade, S., Kao, S.-C., Dullo, T. T., Kalyanapu, A. J., and Preston, B. L. (2019). Ensemble-based flood vulnerability assessment for probable maximum flood in a changing environment. *J. Hydrol.* 576, 342–355. doi: 10.1016/j.jhydrol.2019.06.027

Gangrade, S., Kao, S.-C., Naz, B. S., Rastogi, D., Ashfaq, M., Singh, N., et al. (2018). Sensitivity of probable maximum flood in a changing environment. *Water Resour. Res.* 54, 3913–3936. doi: 10.1029/2017WR021987

García-Alba, J., Bárcena, J. F., Ugarteburu, C., and García, A. (2019). Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. *Water Res.* 150, 283–295. doi: 10.1016/j.watres.2018.11.063

Garcia-Aristizabal, A., and Marzocchi, W. (2013). New methodologies for multi-hazard and multi-risk assessment methods for Europe. EU Matrix project.

Garrick, D., and Hall, J. (2014). Water security and society: Risks, metrics, and pathways. *Annu. Rev. Environ. Resour.* 39, 611–639. doi: 10.1146/annurev-environ-013012-093817

Ge, P., Gokon, H., and Meguro, K. (2020). A review on synthetic aperture radar-based building damage assessment in disasters. *Remote Sens. Environ.* 240:111693. doi: 10.1016/j.rse.2020.111693

Goldstein, E., and Coco, G. (2015). Machine learning components in deterministic models: hybrid synergy in the age of data. *Front. Environ. Sci.* 3:33. doi: 10.3389/fenvs.2015.00033

Govedarica, M., and Jakovljević, G. (2019). "Monitoring spatial and temporal variation of water quality parameters using time series of open multispectral data," in *Seventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2019), Vol. 11174* (Paphos: International Society for Optics and Photonics). doi: 10.1117/12.2533708

Goyal, M. (2014). Modeling of sediment yield prediction using m5 model tree algorithm and wavelet regression. *Water Resour. Manage.* 28, 1991–2003. doi: 10.1007/s11269-014-0590-6

Gruber, F., and Mergili, M. (2013). Regional-scale analysis of high-mountain multi-hazard and risk indicators in the Pamir (Tajikistan) with grass GIS. *Nat. Hazards Earth Syst. Sci.* 13, 2779–2796. doi: 10.5194/nhess-13-2779-2013

Guidolin, M., Chen, A. S., Ghimire, B., Keedwell, E. C., Djordjević, S., and Savić, D. A. (2016). A weighted cellular automata 2d inundation model for rapid flood analysis. *Environ. Model. Softw.* 84, 378–394. doi: 10.1016/j.envsoft.2016.07.008

Guy, H. P. (1970). *Sediment Problems in Urban Areas*. United States Geological Survey. doi: 10.3133/cir601E

Hadka, D., and Reed, P. (2013). Borg: an auto-adaptive many-objective evolutionary computing framework. *Evol. Comput.* 21, 231–259. doi: 10.1162/EVCO_a_00075

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., et al. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-ii waters: a case study of Hong Kong. *Remote Sens.* 11:617. doi: 10.3390/rs11060617

Haghiabi, A., Nasrolahi, A., and Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Qual. Res. J.* 53:wqrjc2018025. doi: 10.2166/wqrj.2018.025

Hajigholizadeh, M., Melesse, A., and Fuentes, H. (2018). Erosion and sediment transport modelling in shallow waters: a review on approaches, models and applications. *Int. J. Environ. Res. Publ. Health* 15:518. doi: 10.3390/ijerph15030518

Hall, J., and Hossain, A. (2020). Mapping urbanization and evaluating its possible impacts on stream water quality in Chattanooga, Tennessee, using GIS and remote sensing. *Sustainability* 12:1980. doi: 10.3390/su12051980

Hall, J. W. (2004). Comment on 'of data and models'. *J. Hydroinform.* 6, 75–77. doi: 10.2166/hydro.2004.0006

Hao, Z., Hao, F., Singh, V. P., Ouyang, W., and Cheng, H. (2017). An integrated package for drought monitoring, prediction and analysis to aid drought modeling and assessment. *Environ. Model. Softw.* 91, 199–209. doi: 10.1016/j.envsoft.2017.02.008

Hao, Z., Singh, V. P., and Xia, Y. (2018). Seasonal drought prediction: advances, challenges, and future prospects. *Rev. Geophys.* 56, 108–141. doi: 10.1002/2016RG000549

Hayatbini, N., Kong, B., Hsu, K.-l., Nguyen, P., Sorooshian, S., Stephens, G., et al. (2019). Conditional generative adversarial networks (cGANs) for near real-time precipitation estimation from multispectral GOES-16 satellite imageries-PERSIANN-cGAN. *Remote Sens.* 11:2193. doi: 10.3390/rs11192193

Hewett, C., Simpson, C., Wainwright, J., and Hudson, S. (2018). Communicating risks to infrastructure due to soil erosion: a bottom-up approach. *Land Degrad. Dev.* 29, 1282–1294. doi: 10.1002/ldr.2900

Hodgson, M., Davis, B., and Kotelenska, J. (2010). "Remote sensing and GIS data/information in the emergency response/recovery phase," in *Geospatial Techniques in Urban Hazard and Disaster Analysis*, eds P. Showalter and Y. Lu (Dordrecht: Springer), 327–354. doi: 10.1007/978-90-481-2238-7_16

Hoekstra, A. Y., Buurman, J., and van Ginkel, K. C. H. (2018). Urban water security: a review. *Environ. Res. Lett.* 13:053002. doi: 10.1088/1748-9326/aaba52

Holzbecher, E., Hadidi, A., Barghash, H., and Balushi, K. (2019). "Application of big data and technologies for integrated water resources management - a survey," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (Grenada), 309–315. doi: 10.1109/SNAMS.2019.8931722

Hong, Y., Hsu, K.-L., Sorooshian, S., and Gao, X. (2004). Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *J. Appl. Meteorol.* 43, 1834–1853. doi: 10.1175/JAM2173.1

Hosseinalizadeh, M., Kariminejad, N., Chen, W., Pourghasemi, H. R., Alinejad, M., Mohammadian Behbahani, A., et al. (2019). Gully headcut susceptibility modeling using functional trees, naive Bayes tree, and random forest models. *Geoderma* 342, 1–11. doi: 10.1016/j.geoderma.2019.01.050

Hou, D., Song, X., Zhang, G., Zhang, H., and Loaiciga, H. (2013). An early warning and control system for urban, drinking water quality protection: China's experience. *Environ. Sci. Pollut. Res.* 20, 4496–4508. doi: 10.1007/s11356-012-1406-y

Huang, J., Svoboda, M., Wood, A., Schubert, S. D., Peters-Lidard, C. D., Wood, E. F., et al. (2016). NOAA drought task force 2016: research to advance national drought monitoring and prediction capabilities.

doi: 10.1201/9781315265551-10 Available online at: https://repository.library.noaa.gov/view/noaa/11194

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environ. Res. Lett*. 14:124007. doi: 10.1088/1748-9326/ab4e55

Huo, A., Zhang, J., Qiao, C., Li, C., Xie, J., Wang, J., et al. (2014). Multispectral remote sensing inversion for city landscape water eutrophication based on genetic algorithm-support vector machine. *Water Qual. Res. J. Can*. 49, 285–293. doi: 10.2166/wqrjc.2014.040

Jain, P., Schoen-Phelan, B., and Ross, R. (2020). "Automatic flood detection in sentinei-2 images using deep convolutional neural networks," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (Brno), 617–623. doi: 10.1145/3341105.3374023

Jetten, V. G., Alkema, D., van Westen, C. J., and Brussel, M. J. G. (2014). "Development of the Caribbean handbook on disaster risk information management," in *International Conference on Analysis and Management of Changing Risks for Natural Hazards 2014* (Padua).

Jiang, J., Qin, C.-Z., Yu, J., Cheng, C., Liu, J., and Huang, J. (2020). Obtaining urban waterlogging depths from video images using synthetic image data. *Remote Sens*. 12:1014. doi: 10.3390/rs12061014

Joslin, P. (2016). *Data-driven analyses of watersheds as coupled human-nature systems* (Ph.D. thesis). University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Kalantari, Z., Ferreira, C., Koutsouris, A., Ahlmer, A.-K., Cerdà, A., and Destouni, G. (2019). Assessing flood probability for transportation infrastructure based on catchment characteristics, sediment connectivity and remotely sensed soil moisture. *Sci. Tot. Environ*. 661, 393–406. doi: 10.1016/j.scitotenv.2019.01.009

Kam, J., Sheffield, J., and Wood, E. F. (2014). A multiscale analysis of drought and pluvial mechanisms for the southeastern united states. *J. Geophys. Res*. 119, 7348–7367. doi: 10.1002/2014JD021453

Kao, S.-C., and Ganguly, A. R. (2011). Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios. *J. Geophys. Res*. 116, 1–14. doi: 10.1029/2010JD015529

Kappes, M., Keiler, M., v. Elverfeldt, K., and Glade, T. (2012). Challenges of dealing with multi-hazard risk: a review. *Nat. Hazards* 64, 1925–1958. doi: 10.1007/s11069-012-0294-2

Karlsson, C., Kalantari, Z., Mörtberg, U., Olofsson, B., and Lyon, S. (2017). Natural hazard susceptibility assessment for road planning using spatial multi-criteria analysis. *Environ. Manage*. 60, 823–851. doi: 10.1007/s00267-017-0912-6

Karpatne, A., Atluri, G., Faghmous, J., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2016). Theory-guided data science: a new paradigm for scientific discovery. *arXiv[Preprint].arXiv:1612.08544*. doi: 10.1109/TKDE.2017.2720168

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng*. 29, 2318–2331.

Kebede, A. (2009). *Water Quality Modeling An Overview*. Available online at: https://files.nc.gov/ncdeq/Water Quality/Planning/TMDL/Modeling/Modeling 101 for FON stakeholder May09.pdf (accessed October 10, 2020).

Kennedy, J., and Eberhart, R. (1995). "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4* (Perth, WA), 1942–1948. doi: 10.1109/ICNN.1995.488968

Khan, Y., and See, C. S. (2016). "Predicting and analyzing water quality using machine learning: a comprehensive model," in *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (Farmingdale, NY), 1–6. doi: 10.1109/LISAT.2016.7494106

Kitsikoudis, V., Sidiropoulos, E., and Hrissanthou, V. (2014). Assessment of sediment transport approaches for sand-bed rivers by means of machine learning. *Hydrol. Sci. J*. 60, 1566–1586. doi: 10.1080/02626667.2014.909599

Knaapen, M., and Hulscher, S. (2003). Use of a genetic algorithm to improve predictions of alternate bar dynamics. *Water Resour. Res*. 39, 1–11. doi: 10.1029/2002WR001793

Komendantova, N., Mrzyglocki, R., Mignan, A., Khazai, B., Wenzel, F., Patt, A., et al. (2014). Multi-hazard and multi-risk decision-support tools as a part of participatory risk governance: feedback from civil protection stakeholders. *Int. J. Disast. Risk Reduct*. 8, 50–67. doi: 10.1016/j.ijdrr.2013.12.006

Konapala, G., and Mishra, A. (2020). Quantifying climate and catchment control on hydrological drought in continental United States. *Water Resour. Res*. 56:e2018WR024620. doi: 10.1029/2018WR024620

Krajewski, W., Ceynar, D., Demir, I., Goska, R., Kruger, A., Langel, C., et al. (2016). Real-time flood forecasting and information system for the state of Iowa. *Bull. Am. Meteorol. Soc*. 98, 539–554. doi: 10.1175/BAMS-D-15-00243.1

Krasnopolsky, V., and Fox-Rabinovitz, M. (2005). "Complex hybrid models combining deterministic and machine learning components as a new synergetic paradigm in numerical climate modeling and weather prediction, Vol. 3," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005* (Montreal, QC), 1615–1620.

Kühnlein, M., Appelhans, T., Thies, B., and Nauß, T. (2014). Precipitation estimates from msg seviri daytime, nighttime, and twilight data with random forests. *J. Appl. Meteorol. Climatol*. 53, 2457–2480. doi: 10.1175/JAMC-D-14-0082.1

Kurte, K., Potnis, A., and Durbha, S. (2019). "Semantics-enabled spatio-temporal modeling of earth observation data: an application to flood monitoring," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities* (Chicago, IL), 41–50. doi: 10.1145/3356395.3365545

Kurte, K. R., Durbha, S. S., King, R. L., Younan, N. H., and Vatsavai, R. (2017). Semantics-enabled framework for spatial image information mining of linked earth observation data. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 10, 29–44. doi: 10.1109/JSTARS.2016.2547992

Kuswanto, H., and Naufal, A. (2019). Evaluation of performance of drought prediction in Indonesia based on TRMM and merra-2 using machine learning methods. *MethodsX* 6, 1238–1251. doi: 10.1016/j.mex.2019.05.029

Lambert, P. (2014). *Caribbean Handbook on Risk Management*. Technical report, ACP-EU NDRR Program.

Lehman, W., and Light, M. (2016). Using hec-fia to identify indirect economic losses. *E3S Web Confer*. 7:05008. doi: 10.1051/e3sconf/20160705008

Li, C., Li, Z., Wu, J., Zhu, L., and Yue, J. (2018). A hybrid model for dissolved oxygen prediction in aquaculture based on multi-scale features. *Inform. Process. Agric*. 5, 11–20. doi: 10.1016/j.inpa.2017.11.002

Liang, J., Li, W., Bradford, S., and Simunek, Jiri, J. (2019). Physics-informed data-driven models to predict surface runoff water quantity and quality in agricultural fields. *Water* 11:200. doi: 10.3390/w11020200

Limbrunner, J., Chapra, S., and Kirshen, P. (2013). Classic optimization techniques applied to stormwater and nonpoint source pollution management at the watershed scale. Journal of Water Resources *Plann. Manage*. 139, 486–491. doi: 10.1061/(ASCE)WR.1943-5452.0000361

Lin, B., and Montazeri Namin, M. (2005). Modelling suspended sediment transport using an integrated numerical and ANNs model. *J. Hydraul. Res*. 43, 302–310. doi: 10.1080/00221680509500124

Lin, Y., Zhang, H., Li, G., Wang, T., Wan, L., and Lin, H. (2019). Improving impervious surface extraction with shadow-based sparse representation from optical, SAR, and LIDAR data. *IEEE Journal of Selected Top. Appl. Earth Observ. Remote Sens*. 12, 2417–2428. doi: 10.1109/JSTARS.2019.2907744

Liu, M., and Huang, M. C. (2014). *COMPOUND DISASTERS AND COMPOUNDING PROCESSES: Implications for Disaster Risk Management.* Prepared for the Global Assessment Report on Disaster Risk Reduction 2015, United Nations Office for Disaster Risk Reduction. Asian Development Bank Institute / National Graduate Institute for Policy Studies. Available online at: https://www.preventionweb.net/english/hyogo/gar/2015/en/bgdocs/inputs/Liu%20and%20Huang,%202014.%20Compound%20disasters%20and%20compounding%20processes%20-%20Implications%20for%20Disaster%20Risk%20Management.pdf

López-Tarazón, J., Batalla, R. J., Vericat, D., and Francke, T. (2012). The sediment budget of a highly dynamic mesoscale catchment: the river isbena. *Geomorphology* 138, 15–28. doi: 10.1016/j.geomorph.2011.08.020

Lu, H., and Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* 249:126169. doi: 10.1016/j.chemosphere.2020.126169

Manning, R., Griffith, J. P., Pigot, T., and Vernon-Harcourt, L. F. (1890). On the flow of water in open channels and pipes. *Inst. Civil Eng. Trans*. 20, 161–207.

Maringanti, C., Chaubey, I., Arabi, M., and B, E. (2008). A multi-objective optimization tool for the selection and placement of bmps for pesticide control. *Hydrol. Earth Syst. Sci. Discuss*. 5, 1821–1862. doi: 10.5194/hessd-5-1821-2008

McGlinchy, J., Johnson, B., Muller, B., Joseph, M., and Diaz, J. (2019). "Application of UNET fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* (Yokohama), 3915–3918. doi: 10.1109/IGARSS.2019.8900453

Medema, W., Wals, A., and Adamowski, J. (2014). Multi-loop social learning for sustainable land and water governance: towards a research agenda on the potential of virtual learning platforms. *Wageningen J. Life Sci*. 69, 23–38. doi: 10.1016/j.njas.2014.03.003

Mekanik, F., Imteaz, M., Gato-Trinidad, S., and Elmahdi, A. (2013). Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *J. Hydrol*. 503, 11–21. doi: 10.1016/j.jhydrol.2013.08.035

Mekonnen, B., Nazemi, A., Elshorbagy, A., Mazurek, K., and Putz, G. (2012). Hybrid modelling approach to prairie hydrology: fusing data-driven and process-based hydrological models. *Hydrol. Sci. J*. 60:6562. doi: 10.1080/02626667.2014.935778

Merritt, W., Letcher, R., and Jakeman, A. (2003). A review of erosion and sediment transport models. *Environ. Model. Softw*. 18, 761–799. doi: 10.1016/S1364-8152(03)00078-1

Merz, B., Hall, J., Disse, M., and Schumann, A. (2010). Fluvial flood risk management in a changing world. *Nat. Hazards Earth Syst. Sci*. 3, 509–527. doi: 10.5194/nhess-10-509-2010

Miao, Z., Xiao, Y., Shi, W., He, Y., Gamba, P., Li, Z., et al. (2019). Integration of satellite images and open data for impervious surface classification. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 12, 1120–1133. doi: 10.1109/JSTARS.2019.2903585

Milly, P. C., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., et al. (2008). Stationarity is dead: whither water management? *Science* 319, 573–574. doi: 10.1126/science.1151915

Mishra, A., Desai, V., and Singh, V. (2007). Drought forecasting using a hybrid stochastic and neural network model. *J. Hydrol. Eng*. 12, 626–638. doi: 10.1061/(ASCE)1084-0699(2007)12:6(626)

Mo, K. C., and Lyon, B. (2015). Global meteorological drought prediction using the North American multi-model ensemble. *J. Hydrometeorol*. 16, 1409–1424. doi: 10.1175/JHM-D-14-0192.1

Mohanty, J. R., and Mohapatra, M. R. (2018). Rainfall prediction using support vector machine (SVM). *IOSR J. Comput. Eng*. 20, 6–13. doi: 10.9790/0661-2003020613

Morgan, R., Morgan, D., and Finney, H. (1984). A predictive model for assessment of erosion risk. *J. Agric. Eng. Res*. 30, 245–253. doi: 10.1016/S0021-8634(84)80025-6

Morid, S., Smakhtin, V., and Bagherzadeh, K. (2007). Drought forecasting using artificial neural networks and time series of drought indices. *Int. J. Climatol*. 27, 2103–2111. doi: 10.1002/joc.1498

Morrison, A. M., Coughlin, K., Shine, J. P., Coull, B. A., and Rex, A. C. (2003). Receiver operating characteristic curve analysis of beach water quality indicator variables. *Appl. Environ. Microbiol*. 69, 6405–6411. doi: 10.1128/AEM.69.11.6405-6411.2003

Mosavi, A., Ozturk, P., and Chau, K.-W. (2018). Flood prediction using machine learning models: literature review. *Water* 10:1536. doi: 10.3390/w10111536

Munafó, M. R., and Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399–401. doi: 10.1038/d41586-018-0 1023-3

Mustafa, M., Abdulkadir, T. S., Wan Yusof, K., Hashim, A., Waris, M., and Shahbaz, M. (2018). SVM-based geospatial prediction of soil erosion under static and dynamic conditioning factors. *MATEC Web Conf*. 203:04004. doi: 10.1051/matecconf/201820304004

Nearing, M., Jetten, V., Baffaut, C., Cerdan, O., Couturier, A., Hernandez, M., et al. (2005). Modeling response of soil erosion and runoff to changes in precipitation and cover. *Catena* 61, 131–154. doi: 10.1016/j.catena.2005.03.007

Newman, J., Maier, H., Riddell, G., Zecchin, A., Daniell, J., Schfer, A., et al. (2017). Review of literature on decision support systems for natural hazard risk reduction: current status and future research directions. *Environ. Model. Softw*. 96, 378–409. doi: 10.1016/j.envsoft.2017.06.042

Nowell, L. H., and Resek, E. A. (1994). "National standards and guidelines for pesticides in water, sediment, and aquatic organisms: application to water-quality assessments," in *Reviews of Environmental Contamination and Toxicology*, ed G. W. Ware (New York, NY: Springer), 1–154. doi: 10.3133/ofr9444

Noymanee, J., Nikitin, N. O., and Kalyuzhnaya, A. V. (2017). Urban pluvial flood forecasting using open data with machine learning techniques in Pattani basin. *Proc. Comput. Sci*. 119, 288–297. doi: 10.1016/j.procs.2017.11.187

NRCS (2003). *National Planning Procedures Handbook (NPPH), Amendment 4*. Available online at: https://nutrientmanagement.tamu.edu/content/resources/nrcs_handbook.pdf (accessed February 5, 2020).

NRCS (2008). *Urban Soil Erosion and Sediment Control*. USDA.

NRCS (2011). *Conservation Planning and Regulatory Compliance Handbook*. Technical report, Pennsylvania NRCS.

Olden, J., Kennard, M., and Pusey, B. (2012). A framework for hydrologic classification with a review of methodologies and applications in ecoyhydrology. *Ecohydrology* 5, 503–518. doi: 10.1002/eco.251

Onderka, M. (2012). Dynamics of storm-driven suspended sediments in a headwater catchment described by multivariable modeling. *J. Soils Sedim*. 12, 620–635. doi: 10.1007/s11368-012-0480-6

Pagán, B. R., Ashfaq, M., Rastogi, D., Kendall, D. R., Kao, S.-C., Naz, B. S., et al. (2016). Extreme hydrological changes in the southwestern us drive reductions in water supply to southern California by mid century. *Environ. Res. Lett*. 11:094026. doi: 10.1088/1748-9326/11/9/094026

Palani, S., Liong, S.-Y., and Tkalich, P. (2008). An ANN application for water quality forecasting. *Mar. Pollut. Bull*. 56, 1586–1597. doi: 10.1016/j.marpolbul.2008.05.021

Palmer, R., Cardwell, H., Lorie, M., and Werick, W. (2013). Disciplined planning, structured participation, and collaborative modeling-applying shared vision planning to water resources. *JAWRA J. Am. Water Resour. Assoc*. 49, 614–628. doi: 10.1111/jawr.12067

Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resour. Res*. 55, 2301–2321. doi: 10.1029/2018WR024090

Park, R., Clough, J., and Wellman, M. (2008). Aquatox: Modeling environmental fate and ecological effects in aquatic ecosystems. *Ecol. Model*. 213, 1–15. doi: 10.1016/j.ecolmodel.2008.01.015

Park, S., Im, J., Jang, E., and Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agric. For. Meteorol*. 216, 157–169. doi: 10.1016/j.agrformet.2015.10.011

Pendergrass, A. G., Meehl, G. A., Pulwarty, R., Hobbins, M., Hoell, A., AghaKouchak, A., et al. (2020). Flash droughts present a new challenge for subseasonal-to-seasonal prediction. *Nat. Clim. Change* 10, 191–199. doi: 10.1038/s41558-020-0709-0

Pereira, J. (2018). Science and technology to enhance disaster resilience in a changing climate. *Sci. Technol. Diaster Risk Reduc. Asia* 2018, 31–38. doi: 10.1016/B978-0-12-812711-7.00003-1

Pham, B., Tien Bui, D., Prakash, I., and Dholakia, M. (2016). Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena* 149, 52–63. doi: 10.1016/j.catena.2016.09.007

Philbrick, C. R. Jr., and Kitanidis, P. K. (1999). Limitations of deterministic optimization applied to reservoir operations. *J. Water Resour. Plann. Manage*. 125, 135–142. doi: 10.1061/(ASCE)0733-9496(1999)125:3(135)

Pielke, R. Jr. (2000). Flood impacts on society: damaging floods as a framework for assessment. *Floods* 1, 133–156.

Poff, N., and Allan, J. D. (1995). Functional organization of stream fish assemblages in relation to hydrologic variability. *Ecology* 76, 606–627. doi: 10.2307/1941217

Potnis, A. V., Durbha, S. S., and Kurte, K. R. (2018). "A geospatial ontological model for remote sensing scene semantic knowledge mining for the flood disaster," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (Valencia), 5274–5277. doi: 10.1109/IGARSS.2018.8517680

Potnis, A. V., Shinde, R. C., Durbha, S. S., and Kurte, K. R. (2019). "Multi-class segmentation of urban floods from multispectral imagery using deep learning," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (Yokohama), 9741–9744. doi: 10.1109/IGARSS.2019.8900250

Pourghasemi, H. R., Kariminejad, N., Amiri, M., Zarafshar, M., thomasBlaschke, and Edalat, M. (2020). Assessing and mapping multi- hazard risk susceptibility using a machine learning technique. *Sci. Rep*. 10:3203. doi: 10.1038/s41598-020-60191-3

Pourghasemi, H. R., Yousefi, S., Kornejady, A., and Cerdà, A. (2017). Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. *Sci. Tot. Environ*. 609, 764–775. doi: 10.1016/j.scitotenv.2017.07.198

Prosser, I., Young, W., Rustomji, P., Hughes, A., and Moran, C. (2001). "A model of river sediment budgets as an element of river health assessment," in *Proceedings of the International Congress on Modelling and Simulation MODSIM'2001*, 861–866.

Rahman, M. S., and Di, L. (2017). The state of the art of spaceborne remote sensing in flood management. *Nat. Hazards* 85, 1223–1248. doi: 10.1007/s11069-016-2601-9

Rahmasary, A. N., Robert, S., Chang, I.-S., Jing, W., Park, J., Bluemling, B., Koop, S., and Van Leeuwen, K. V. L. (2019). Overcoming the challenges of water, waste and climate change in asian cities. *Environ. Manage*. 63, 520–535. doi: 10.1007/s00267-019-01137-y

Rahmati, O., Falah, F., Dayal, K. S., Deo, R. C., Mohammadi, F., Biggs, T., et al. (2020). Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. *Sci. Tot. Environ*. 699:134230. doi: 10.1016/j.scitotenv.2019.134230

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R., and Feizizadeh, B. (2017). Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology* 298, 118–137. doi: 10.1016/j.geomorph.2017.09.006

Rahmati, O., Yousefi, S., Kalantari, Z., Uuemaa, E., Teimurian, T., Keesstra, S., et al. (2019). Multi-hazard exposure mapping using machine learning techniques: a case study from Iran. *Remote Sens*. 11:1943. doi: 10.3390/rs11161943

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

Reed, P. M., Hadka, D., Herman, J. D., Kasprzyk, J. R., and Kollat, J. B. (2013). Evolutionary multiobjective optimization in water resources: the past, present, and future. *Adv. Water Resour*. 51, 438–456. doi: 10.1016/j.advwatres.2012.01.005

Reed, P. M., and Kollat, J. B. (2013). Visual analytics clarify the scalability and effectiveness of massively parallel many-objective optimization: a groundwater monitoring design example. *Adv. Water Resour*. 56, 1–13. doi: 10.1016/j.advwatres.2013.01.011

Refsgaard, J. C., and Henriksen, H. J. (2004). Modelling guidelines–terminology and guiding principles. *Adv. Water Resour*. 27, 71–82. doi: 10.1016/j.advwatres.2003.08.006

Reidy Liermann, C., Nilsson, C., Robertson, J., and Ng, R. (2012). Implications of dam obstruction for global freshwater fish diversity. *BioScience* 62, 539–548. doi: 10.1525/bio.2012.62.6.5

Rice, J., Emanuel, R., Vose, J., and Nelson, S. (2015). Continental U.S. streamflow trends from 1940 to 2009 and their relationships with watershed spatial characteristics. *Water Resour. Res*. 51, 6262–6275. doi: 10.1002/2014WR016367

Rodriguez, F., Andrieu, H., and Creutin, J.-D. (2003). Surface runoff in urban catchments: morphological identification of unit hydrographs from urban databanks. *J. Hydrol*. 283, 146–168. doi: 10.1016/S0022-1694(03)00246-4

Rodriguez, H., Popp, J., Maringanti, C., and Chaubey, I. (2011). Selection and placement of best management practices used to reduce water quality degradation in Lincoln lake watershed. *Water Resour. Res*. 47, 1–13. doi: 10.1029/2009WR008549

Rowley, K. J. (2014). *Sediment transport conditions near culverts* (Master's thesis). Brigham Young University, Provo, UT. doi: 10.1061/97807844135 48.141

Rozos, E. (2019). Machine learning, urban water resources management and operating policy. *Resources* 8:173. doi: 10.3390/resources8040173

Ruessink, G. (2005). Calibration of nearshore process models - application of a hybrid genetic algorithm. *J. Hydroinform*. 7, 135–149. doi: 10.2166/hydro.2005.0012

Sahoo, S., Russo, T. A., Elliott, J., and Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resour. Res*. 53, 3878–3895. doi: 10.1002/2016WR0 19933

Santelmann, M., Hulse, D., Wright, M., Enright, C., Branscomb, A., Tchintcharauli-Harrison, M., et al. (2019). Designing and modeling innovation across scales for urban water systems. *Urban Ecosyst*. 22, 1149–1164. doi: 10.1007/s11252-019-00882-6

Savic, D. (2019). "What is artificial intelligence and how can water planning and management benefit from it?," in *International Association for Hydro-Environment Engineering and research (IAHR)*, ed S. Wieprecht (Universit?t Stuttgart). Available online at: https://www.researchgate.net/profile/Dragan_Savic3/publication/335126234_What_is_Artificial_Intelligence_and_how_can_water_planning_and_management_benefit_from_it/links/5d51795a299bf1995b78d9e4/What-is-Artificial-Intelligence-and-how-can-water-planning-and-management-benefit-from-it.pdf

Scott, S., Burns, I., Levick, L., Hernandez, M., and Goodrich, D. (2002). Automated geospatial watershed assessment (AGWA) - a GIS-based hydrologic modeling tool: documentation and user manual. Available online at: https://www.researchgate.net/publication/253514869_Automated_Geospatial_Watershed_Assessment_AGWA_-_A_GIS-Based_Hydrologic_Modeling_Tool_Documentation_and_User_Manual

Selin, S., and Chevez, D. (1995). Developing a collaborative model for environmental planning and management. *Environ. Manage*. 19, 189–195. doi: 10.1007/BF02471990

Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J. J., et al. (2020). Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. *Remote Sens*. 12:266. doi: 10.3390/rs12020266

Shamseldin, A. Y. (2010). Artificial neural network model for river flow forecasting in a developing country. *J. Hydroinform*. 12, 22–35. doi: 10.2166/hydro.2010.027

Sharaf El Din, E., Zhang, Y., and Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *Int. J. Remote Sens*. 38, 1023–1042. doi: 10.1080/01431161.2016.1275056

Shaw, R., Rahman, A.-U., Surjan, A., and Parvin, G. (2016). Urban Disasters and Resilience in Asia. Elsevier: Butterworth-Heinemann. 368.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and WOO, W.-C. (2015). "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems* (Montreal, QC), eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc.), 802–810.

Shi, Y., and Eberhart, R. (1998). "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings* (Anchorage, AK), 69–73. doi: 10.1109/ICEC.1998.699146

Shimoda, A., Ichikawa, D., and Oyama, H. (2018). Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Comput. Methods Prog. Biomed*. 163, 39–46. doi: 10.1016/j.cmpb.2018.05.032

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv[Preprint].arXiv:1409.1556*.

Singh, K. P., Basant, A., Malik, A., and Jain, G. (2009). Artificial neural network modeling of the river water quality-a case study. *Ecol. Model*. 220, 888–895. doi: 10.1016/j.ecolmodel.2009.01.004

Snelder, T., and Biggs, B. (2007). Multi-scale river environment classification for water resources management. *JAWRA J. Am. Water Resour. Assoc*. 38, 1225–1239. doi: 10.1111/j.1752-1688.2002.tb04344.x

Souchère, V., Millair, L., Echeverria, J., Bousquet, F., Le Page, C., and Etienne, M. (2010). Co-constructing with stakeholders a role-playing game to initiate collective management of erosive runoff risks at the watershed scale. *Environ. Model. Softw*. 25, 1359–1370. doi: 10.1016/j.envsoft.2009.03.002

Srivastava, P., Hamlett, J., Robillard, P., and Day, R. (2002). Watershed optimization of best management practices using annaGNPS and a genetic algorithm. *Water Resour. Res*. 38, 3-1–3-14. doi: 10.1029/2001WR000365

Statham, G., Haegeli, P., Greene, E., Birkeland, K., Israelson, C., Tremper, B., et al. (2017). A conceptual model of avalanche hazard. *Nat. Hazards*. 90, 663–691. doi: 10.1007/s11069-017-3070-5

Sun, A., and Scanlon, B. (2019). How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environ. Res. Lett*. 14:073001. doi: 10.1088/1748-9326/ab1b7d

Sun, J., Lou, Y., and Ye, F. (2017). "Research on anomaly pattern detection in hydrological time series," in *2017 14th Web Information Systems and Applications Conference (WISA)* (Liuzhou), 38–43. doi: 10.1109/WISA.2017.73

Sun, Z., Zhao, X., Wu, M., and Wang, C. (2019). Extracting urban impervious surface from worldview-2 and airborne LIDAR data using 3d convolutional neural networks. *J. Indian Soc. Remote Sens*. 47, 401–412. doi: 10.1007/s12524-018-0917-5

Taskaya-Temizel, T. and Casey, M. C. (2005). A comparative study of autoregressive neural network hybrids. *Neural Netw*. 18, 781–789. doi: 10.1016/j.neunet.2005.06.003

Tay, F. E., and Cao, L. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing* 48, 847–861. doi: 10.1016/S0925-2312(01)00676-2

Tayfur, G. (2002). Artificial neural networks for sheet sediment transport. *Hydrol. Sci. J*. 47, 879–892. doi: 10.1080/02626660209492997

Tayfur, G., and Guldal, V. (2006). Artificial neural networks for estimating daily total suspended sediment in natural streams. *Nordic Hydrol*. 37, 69–79. doi: 10.2166/nh.2006.0006

Tayyab, M., Zhou, J., Zeng, X., and Adnan, R. (2016). Discharge forecasting by applying artificial neural networks at the Jinsha river basin, china. *Eur. Sci. J*. 12, 108–127. doi: 10.19044/esj.2016.v12n9p108

Tsintikidis, D., Haferman, J. L., Anagnostou, E. N., Krajewski, W. F., and Smith, T. F. (1997). A neural network approach to estimating rainfall from spaceborne microwave data. *IEEE Trans. Geosci. Remote Sens*. 35, 1079–1093. doi: 10.1109/36.628775

U.N. (2018). *World Urbanization Prospects: The 2018 Edition Highlights*. Department of Economic and Social Affairs.

UN (2010). *Clean Water for a Healthy World*. Technical report, United Nations.

UN (2012). *Un Water Decade Program on Advocacy and Communication*. Technical report, United Nations.

USEPA (2012). *Integrated Planning for Municipal Stormwater and Wastewater*. Available online at: https://www.epa.gov/npdes/integrated-planning-municipal-stormwater-and-wastewater (accessed February 5, 2020).

USEPA (2018). *Best Management Practices (BMPs) Siting Tool*. Available online at: https://www.epa.gov/water-research/best-management-practices-bmps-siting-tool (accessed February 5, 2020).

Van Loon, A. F. (2015). Hydrological drought explained. *Wiley Interdisc. Rev*. 2, 359–392. doi: 10.1002/wat2.1085

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer Verlag. doi: 10.1007/978-1-4757-2440-0

Wagener, T., Sivapalan, M., Troch, P., McGlynn, B., Harman, C., Gupta, H., et al. (2010). The future of hydrology: an evolving science for a changing world. *Water Resour. Res*. 46, 1–10. doi: 10.1029/2009WR008906

Wagener, T., Sivapalan, M., Troch, P., and Woods, R. (2007). Catchment classification and hydrologic similarity. *Geogr. Compass* 1, 901–931. doi: 10.1111/j.1749-8198.2007.00039.x

Wang, X., Fu, L., and He, C. (2011). Applying support vector regression to water quality modelling by remote sensing data. *Int. J. Remote Sens*. 32, 8615–8627. doi: 10.1080/01431161.2010.543183

Wang, X., Zhang, F., and Ding, J. (2017). Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Sci. Rep*. 7, 1–18. doi: 10.1038/s41598-017-12853-y

Weng, Q. (2012). Remote sensing of impervious surfaces in the urban areas: requirements, methods, and trends. *Remote Sens. Environ*. 117, 34–49. doi: 10.1016/j.rse.2011.02.030

Whigham, P., and Crapper, P. (2001). Modelling rainfall-runoff using genetic programming. *Math. Comput. Model*. 33, 707–721. doi: 10.1016/S0895-7177(00)00274-0

Wieprecht, S., Habtamu, G., and Yang, C. (2013). A neuro-fuzzy-based modelling approach for sediment transport computation. *Hydrol. Sci. J*. 58, 587–599. doi: 10.1080/02626667.2012.755264

Wischmeier, W. H., and Smith, D. D. (1978). *Predicting-Rainfall Erosion Losses: A Guide to Conservation Planning*. USDA.

Wood, E. F., Schubert, S. D., Wood, A. W., Peters-Lidard, C. D., Mo, K. C., Mariotti, A., et al. (2015). Prospects for advancing drought understanding, monitoring, and prediction. *J. Hydrometeorol*. 16, 1636–1657. doi: 10.1175/JHM-D-14-0164.1

Wu, J.-L., Ho, C.-R., Huang, C.-C., Srivastav, A. L., Tzeng, J.-H., and Lin, Y.-T. (2014). Hyperspectral sensing for turbid water quality monitoring in freshwater rivers: empirical relationship between reflectance and turbidity and total solids. *Sensors* 14, 22670–22688. doi: 10.3390/s141222670

Xu, H., Demir, I., Koylu, C., and Muste, M. (2019a). A web-based geovisual analytics platform for identifying potential contributors to culvert sedimentation. *Sci. Tot. Environ*. 692, 806–817. doi: 10.1016/j.scitotenv.2019.07.157

Xu, H., Windsor, M., Muste, M., and Demir, I. (2019b). A web-based decision support system for collaborative mitigation of multiple water-related hazards using serious gaming. *J. Environ. Manage*. 255:109887. doi: 10.1016/j.jenvman.2019.109887

Yadav, A., Chatterjee, S., and Equeenuddin, S. (2019a). Prediction of suspended sediment yield by artificial neural network and traditional mathematical model in Mahanadi River Basin, India. *Sustain. Water Resour. Manage*. 4, 745–759. doi: 10.1007/s40899-017-0160-1

Yadav, A., Chatterjee, S., and Equeenuddin, S. (2019b). Suspended sediment yield estimation using genetic algorithm-based artificial intelligence models: case study of Mahanadi River, India. *Hydrol. Sci. J*. 63, 1162–1182. doi: 10.1080/02626667.2018.1483581

Yang, C., Marsooli, R., and AALAMI, M. (2009). Evaluation of total load sediment transport formulas using ann. *Int. J. Sedim. Res*. 24, 274–286. doi: 10.1016/S1001-6279(10)60003-0

Yang, S., Ogawa, Y., Ikeuchi, K., Akiyama, Y., and Shibasaki, R. (2019). "Firm-level behavior control after large-scale urban flooding using multi-agent deep reinforcement learning," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoSpatial Simulation* (Chicago, IL), 24–27. doi: 10.1145/3356470.3365529

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., et al. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environ. Res. Lett*. 14:114027. doi: 10.1088/1748-9326/ab4d5e

Yao, Y., He, J., Zhang, J., and Zhang, Y. (2017). Extracting urban impervious surface from GF-1 imagery using one-class classifiers. *arXiv[Preprint].arXiv:1705.04824*.

Yoe, C., and Orth, K. (1996). *Planning Manual. U.S. Army Corps of Engineering*. Water Resources Support Center.

Yu, M., Yang, C., and Li, Y. (2018). Big data in natural disaster management: a review. *Geosciences* 8:165. doi: 10.3390/geosciences8050165

Zaidi, S. M. A., Chandola, V., Allen, M. R., Sanyal, J., Stewart, R. N., Bhaduri, B. L., et al. (2018). Machine learning for energy-water nexus: challenges and opportunities. *Big Earth Data* 2, 228–267. doi: 10.1080/20964471.2018.1526057

Zaniolo, M., Giuliani, M., Castelletti, A. F., and Pulido-Velazquez, M. (2018). Automatic design of basin-specific drought indexes for highly regulated water systems. *Hydrol. Earth Syst. Sci*. 22, 2409–2424. doi: 10.5194/hess-22-2409-2018

Zhang, H., Wan, L., Wang, T., Lin, Y., Lin, H., and Zheng, Z. (2019). Impervious surface estimation from optical and polarimetric sar data using small-patched deep convolutional networks: a comparative study. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 12, 2374–2387. doi: 10.1109/JSTARS.2019.2915277

Zhang, R., Chen, Z.-Y., Xu, L.-J., and Ou, C.-Q. (2019). Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, china. *Sci. Tot. Environ*. 665, 338–346. doi: 10.1016/j.scitotenv.2019.01.431

Zhang, Z., Xu, L., and Liu, Q. (2019). "The identification of impervious area from sentinel-2 imagery using a novel spectral spatial residual convolution neural network," in *Proceedings of the 2019 3rd International Conference on Advances in Image Processing* (New York, NY, USA: Association for Computing Machinery).

Zounemat-Kermani, M., Matta, E., Cominola, A., Xia, X., Zhang, Q., Liang, Q., et al. (2020). Neurocomputing in surface water hydrology and hydraulics: a review of two decades retrospective, current status and future prospects. *J. Hydrol*. 2020:125085. doi: 10.1016/j.jhydrol.2020.125085

Disclaimer: This manuscript has been authored by UT-Battelle LLC under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

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
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact