frontiers | Research Topics

Democratizing data: Environmental data access and its future

Edited by
Michael C. Kruk, Lauren A. Jackson, Kevin A. Butler, Tiffany C. Vance and Nazila Merati

Published in
Frontiers in Climate

ISSN 1664-8714
ISBN 978-2-83251-523-5
DOI 10.3389/978-2-83251-523-5

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

Democratizing data: Environmental data access and its future

Topic editors

Michael C. Kruk — National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), United States

Lauren A. Jackson — National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), United States

Kevin A. Butler — Environmental Systems Research Institute, United States

Tiffany C. Vance — U.S. Integrated Ocean Observing System, United States

Nazila Merati — National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), United States

Citation

Kruk, M. C., Jackson, L. A., Butler, K. A., Vance, T. C., Merati, N., eds. (2023). *Democratizing data: Environmental data access and its future*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-523-5

Table of contents

frontiers | Frontiers in Climate

TYPE Editorial
PUBLISHED 17 January 2023
DOI 10.3389/fclim.2022.1081021

OPEN ACCESS

EDITED AND REVIEWED BY
Chris C. Funk,
College of Letters & Science (UC),
United States

*CORRESPONDENCE
Lauren A. Jackson
✉ lauren.jackson@noaa.gov

SPECIALTY SECTION
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

RECEIVED 26 October 2022
ACCEPTED 15 December 2022
PUBLISHED 17 January 2023

CITATION
Butler KA, Jackson LA, Kruk MC, Merati N and Vance TC (2023) Editorial: Democratizing data: Environmental data access and its future. *Front. Clim.* 4:1081021. doi: 10.3389/fclim.2022.1081021

Editorial: Democratizing data: Environmental data access and its future

Kevin A. Butler[1], Lauren A. Jackson[2]*, Michael C. Kruk[2], Nazila Merati[3] and Tiffany C. Vance[4]

[1]Environmental Systems Research Institute (ESRI), Redlands, CA, United States, [2]National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), Stennis Space Center, MS, United States, [3]National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), Boulder, CO, United States, [4]United States (U.S.) Integrated Ocean Observing System, National Oceanic and Atmospheric Administration (NOAA), Silver Spring, MD, United States

KEYWORDS

data discoverability, data access, data and service equity, data usability, reproducibility, environmental data

Editorial on the Research Topic
Democratizing data: Environmental data access and its future

Brief introduction to the Research Topic

Data democratization is the equal and interdependent responsibility of data producers, consumers, and curators to make data discoverable, accessible, equitable, and usable (Figure 1). The genesis of democratization stems from critical changes in the supply and demand of data over the last decades. First, the supply of data, particularly environmental data, has exploded, partly due to improvements in the number and resolution of observational platforms and governmental entities and private organizations embracing the philosophy of open public data declaring data as a public good. Likewise, the increased demand for environmental data stems from the realization that planet-scale problems require planet-scale analyses and an increased emphasis on data-driven environmental decision-making (Tonn et al., 2000).

The global phenomenon of *datafication* (Mayer-Schönberger and Cukier, 2013) has resulted in ever-increasing availability, demand, and use of environmental data at unprecedented physical and social scales. Taylor (1997, p. 327) prophetically states, "in the twenty-first century, the community—not the federal government—will be the principal unit of solution to social and economic difficulties." This shift has expanded the collection and reach of environmental data to previously unforeseen data consumers such as citizen scientists. However, even traditional consumers of environmental data, science practitioners, face data challenges in this era of big data. Meeting the goals of data as a public good and supporting science at local scales requires a new perspective on the production, management, and curation of data. We refer to this new perspective as data democratization—introducing democratic principles into all aspects of data processes

(see Tilly, 2001 for a complete discussion of democratization). The submissions in this Research Topic raise important questions and provide practical examples for making data more discoverable, accessible, equitable, and usable.

Demographics of participation

The nine articles on the Research Topic were submitted by 43 authors representing 27 institutions or organizations. The majority (58%) were submitted by authors affiliated with educational institutions, with the remainder from commercial (7%), governmental (21%), or non-profit (14%) entities. The heterogeneous institutional affiliations suggest that democratizing data presents challenges and opportunities for all data producers.

Topics of the collection

Three broad themes emerged from the collection:

- Maintaining a **user focus** in all aspects of the data lifecycle: Virapongse et al. challenge Earth scientists to be introspective about the methods and processes they use to produce more effective information products to help place-based communities build resilience. Cantor et al. suggest that the decision-maker needs to be incorporated directly into data systems design. Finally, Gärtner-Roer et al. highlight the recent increased availability of glacier data and the role a centralized user-focused repository and standards organization can play.
- Making **data usability** a priority: Stern et al. describe an open-source platform for extracting archival data and creating analysis-ready, cloud-optimized data stores that empower a broader community of scientists. Contributors also highlighted the challenges of repurposing and making existing data repositories more usable (Rossi et al.) and unique usability challenges in 3D data (Paxton et al.).
- Ensuring **data veracity** and **equity**: In addition to the traditional data veracity challenges in big environmental data, ensuring veracity in time series (Sweeney) and simulation data (Mullendore et al.) presented unique challenges. Finally, Dosemagen and Williams state that prioritizing environmental data as a public good is a key to data usability, and usability is key to addressing environmental justice issues.

Increased pressure from funding agencies that promote or even mandate open data sharing has resulted in a new perspective on data—an explicit focus on the user. Historically, the data lifecycle supported the process of *knowledge production*, collecting, and analyzing data for peer-reviewed journal publications (Baker and Mayernik, 2020). A complementary

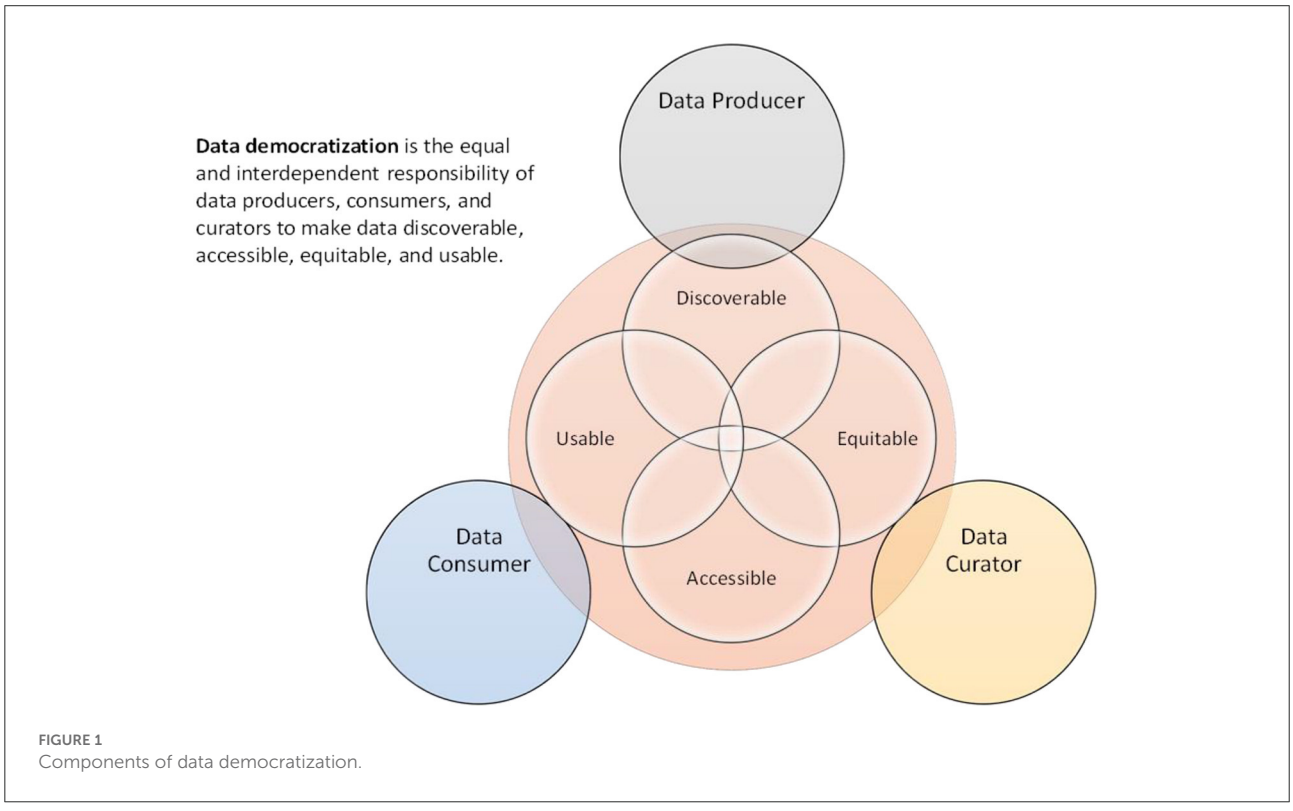


FIGURE 1
Components of data democratization.

data lifecycle, *data production*, "creates data intended for release to a data repository that makes data accessible for reuse by others" (p. 4). This user-focused perspective dramatically impacts how data are documented, managed, and shared. Just as a democratic system of government requires an informed citizenry, democratized systems of data require informed data users.

While big environmental data presents tremendous opportunities, they also come with tremendous challenges. For example, data's sheer volume, variety, and velocity can impede its usability. Contributors touched on many aspects of the findability, accessibility, interoperability, and reuse (FAIR) principles (Wilkinson et al., 2016), providing the following insights that help reach the goals of democratization:

- View data not as stand-alone datasets but as a system in and of itself (Stern et al.).
- Design data systems to meet the data needs of decision-makers (decision-driven data systems) rather than requiring decision-makers to adapt to existing systems (Cantor et al.).
- Include data users and producers in the design of data access systems (Dosemagen and Williams).
- Strive for *effective* use of data (Virapongse et al.).

Data veracity is, arguably, the riskiest aspect of data democratization. Issues of data quality and fitness for purpose become more critical as data sharing and data use networks grow beyond the data producer. Producers and curators are responsible for summarizing and communicating data quality issues and fitness for purpose in a form and tone approachable by data users outside the subject domain and the technical expertise of the data producers. This approach is consistent with the trend in several academic conferences and journals requesting a plain-language summary of scientific research.

Beyond FAIR

How is data democratization different from the FAIR principles? While the FAIR principles are an essential first step to promoting the democratization of data, they are, in our view, focused on the data provider. Boeckhout et al. (2018, p. 931) argue that "even though the principles create a powerful platform for furthering data sharing and improving data stewardship, they do not address the normative issues and challenges associated with data sharing." Strict check-listed adherence to the FAIR principles is a necessary but insufficient first step. Data democratization is a more holistic, comprehensive view of a process to make data discoverable, accessible, equitable, and usable.

Although not mentioned in our Research Topic, the CARE principles are a seminal example of shifting the focus from data providers to data consumers and moving beyond FAIR (Carroll et al., 2021). Developed by and for indigenous communities, these principles promote data ecosystems that provide **collective benefit**, where the **authority to control** the data resides with the data subjects and where there is a recognized **responsibility** to engage respectfully with data subjects. In addition, the **ethics** of the data subjects should inform data use.

Author contributions

KB and LJ drafted the content of this document. MK, TV, and NM provided edits. KB contributed figures. All authors contributed to the article and approved the submitted version.

Acknowledgments

The topic editors thank the authors that contributed papers to this topic. We also thank the reviewers for their dedication and time invested in providing invaluable feedback to the authors.

Conflict of interest

KB was employed by Environmental Systems Research Institute (ESRI).

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

Publisher's note

References

Baker, K. S., and Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere* 11, e03191. doi: 10.1002/ecs2.3191

Boeckhout, M., Zielhuis, G. A., and Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough? *Euro. J. Hum. Genet*. 26, 931–936. doi: 10.1038/s41431-018-0160-0

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., and Stall, S. (2021). Operationalizing the CARE and FAIR principles for indigenous data futures. *Sci Data* 8, 108. doi: 10.1038/s41597-021-00892-0

Mayer-Schönberger, V., and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Taylor, H. L. Jr. (1997). No more ivory towers: connecting the research university to the community. *J. Plan. Liter*. 11, 327–332. doi: 10.1177/088541229701100304

Tilly, C. (2001). Mechanisms in political processes. *Ann. Rev. Polit. Sci*. 4, 21–41. doi: 10.1146/annurev.polisci.4.1.21

Tonn, B., English, M., and Travis, C. (2000). A framework for understanding and improving environmental decision making. *J. Environ. Plan. Manage*. 43, 163–183. doi: 10.1080/09640560010658

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi: 10.1038/sdata.2016.18

frontiers in Climate

ORIGINAL RESEARCH
published: 17 November 2021
doi: 10.3389/fclim.2021.761444

Making a Water Data System Responsive to Information Needs of Decision Makers

Alida Cantor[1,2,3], Michael Kiparsky[2,3], Susan S. Hubbard[4], Rónán Kennedy[5], Lidia Cano Pecharroman[3,6], Kamyar Guivetchi[7], Gary Darling[7], Christina McCready[7] and Roger Bales[2,8]*

[1] Department of Geography, Portland State University, Portland, OR, United States, [2] Water Security and Sustainability Research Initiative, University of California, Merced, Merced, CA, United States, [3] Center for Law, Energy & the Environment, Berkeley School of Law, University of California, Berkeley, Berkeley, CA, United States, [4] Lawrence Berkeley National Laboratory, University of California, Berkeley, Berkeley, CA, United States, [5] School of Law, National University of Ireland Galway, Galway, Ireland, [6] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States, [7] California Department of Water Resources, Sacramento, CA, United States, [8] School of Engineering, University of California, Merced, Merced, CA, United States

OPEN ACCESS

Edited by:
Tiffany C. Vance, U.S. Integrated Ocean Observing System, United States

Reviewed by:
Nancy Wilkinson, San Francisco State University, United States
Austin Becker, University of Rhode Island, United States

***Correspondence:**
Alida Cantor
acantor@pdx.edu

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *19 August 2021*
Accepted: *07 October 2021*
Published: *17 November 2021*

Citation:
Cantor A, Kiparsky M, Hubbard SS, Kennedy R, Pecharroman LC, Guivetchi K, Darling G, McCready C and Bales R (2021) Making a Water Data System Responsive to Information Needs of Decision Makers. Front. Clim. 3:761444. doi: 10.3389/fclim.2021.761444

Evidence-based environmental management requires data that are sufficient, accessible, useful and used. A mismatch between data, data systems, and data needs for decision making can result in inefficient and inequitable capital investments, resource allocations, environmental protection, hazard mitigation, and quality of life. In this paper, we examine the relationship between data and decision making in environmental management, with a focus on water management. We focus on the concept of *decision-driven data systems*—data systems that incorporate an assessment of decision-makers' data needs into their design. The aim of the research was to examine the process of translating data into effective decision making by engaging stakeholders in the development of a water data system. Using California's legislative mandate for state agencies to integrate existing water and other environmental data as a case study, we developed and applied a participatory approach to inform data-system design and identify unmet data needs. Using workshops and focused stakeholder meetings, we developed 20 diverse use cases to assess data sources, availability, characteristics, gaps, and other attributes of data used for representative decisions. Federal and state agencies made up about 90% of the data sources, and could readily adapt to a federated data system, our recommended model for the state. The remaining 10% of more-specialized data, central to important decisions across multiple use cases, would require additional investment or incentives to achieve data consistency, interoperability, and compatibility with a federated system. Based on this assessment, we propose a typology of different types of data limitations and gaps described by stakeholders. We also propose technical, governance, and stakeholder engagement evaluation criteria to guide planning and building environmental data systems. Data-system governance involving both producers and users of data was seen as essential to achieving workable standards, stable

funding, convenient data availability, resilience to institutional change, and long-term buy-in by stakeholders. Our work provides a replicable lesson for using decision-maker and stakeholder engagement to shape the design of an environmental data system, and inform a technical design that addresses both user and producer needs.

Keywords: water management, data systems, stakeholder engagement, environmental decision making, California

INTRODUCTION

Evidence-based environmental management requires data that are sufficient, accessible, useful and used (California Department of Water Resources, 2020). If data systems are to effectively inform environmental decision making, then development of such systems can be improved through assessment and incorporation of decision-makers' data needs. The concept of *data-driven decision making* describes the practice of making decisions based on analysis of data (Provost and Fawcett, 2013). In this paper, we develop a related and equally important concept of *decision-driven data systems*: data systems that are designed based on an understanding of decision-makers' data needs. Development of such systems can be improved through first assessing these needs and then incorporating this assessment into system design and content prioritization.

We define "data systems" broadly as the assemblage of hardware, software, people, and institutions that collect, organize, archive, distribute, integrate, process, analyze, and synthesize data and information. There are a growing number of efforts that seek to advance earth and environmental data systems through integration and collaboration in order to maximize applicability to both research and decision making. For example, National Science Foundation (NSF) has supported Hydroshare, a collaborative environment for sharing hydrologic and critical-zone data and models geared toward research users. In the European Union, the INSPIRE Directive seeks to create a spatial-data infrastructure to inform E.U. environmental policies, and the Copernicus project focuses on meeting earth-science data-user needs. Copernicus developers have created a use case library demonstrating how data are applied to real-world problem solving.

Water management presents an important case for strengthening the relationship between environmental data and decision making. Provisioning and use of adequate information are central to effectively making investments in water infrastructure, confirming environmental regulatory compliance, managing risks and uncertainties, guiding operations, evaluating and encouraging innovation, and making rapid and effective decisions during droughts, floods, or crisis events (Kiparsky et al., 2013; Escriva-Bou et al., 2016; Larsen et al., 2016; Green Nylen et al., 2018a,b). Researchers have worked to strengthen connections between data and decision making related to water. For example, researchers have assessed decision-makers' demand for and use of forecasting data for water resources management (Viel et al., 2016; Neumann et al., 2018). Researchers and computational/data scientists are advancing new approaches to quantify watershed behavior to inform management decisions. Recent examples highlight the promise of machine learning for advancing tractable watershed-data processing, parameter estimation, sensor optimization, early warning, groundwater-level prediction, and process understanding (e.g., Ahmad et al., 2010; Oroza et al., 2016; Pau et al., 2016; Mosavi et al., 2018; Schmidt et al., 2018; Müller et al., 2019). Researchers are also developing watershed-centric data tools that seek to improve integration of data management, analysis, modeling and interpretation of diverse watershed datasets (Varadharajan et al., 2019; Hubbard et al., 2020). These examples indicate significant potential for new tools to aid in the tractable translation of water data into information for decision making.

The complexity of water systems means that managers must integrate and analyze multiple types of data and information (Kallis et al., 2006; Bakker, 2012; Vogel et al., 2015). Modern information technology promises, in concept, to make such multi-faceted integration possible, but providing data does not in and of itself ensure that data can or will be used for more effective and sustainable water management. Here, *water data* refers to a broad suite of data and information used to inform water-related research and decision making. Water data includes both measured data and model-output data, and can be used both to characterize systems and to monitor conditions over time. Our definition of water data goes beyond hydrologic data such as streamflow, precipitation, and groundwater-level measurements to include many related and relevant areas, such as land use, ecological, and agricultural data. We primarily address public data sources in this paper.

As a case study, we focus on California water, which is one of the most complex and politically contentious environmental management challenges in the world. California's water challenges require a wide range of data to solve problems including managing drought and climate change, balancing environmental and agricultural water demands, and meeting water needs of endangered species and cities alike (Hanak, 2011). Yet despite California's prominence in the technology sphere, the state's water data have not proven up for these challenges (California Council on Science and Technology, 2014; Escriva-Bou et al., 2016). California water data are diverse and fragmented, and are produced, housed, and maintained by multiple entities from disparate sectors. Recent legislation has attempted to address this issue. California's Open and Transparent Water Data Act (Assembly Bill, or AB 1755), passed in 2016 (Cal. Water Code §12,400 et seq.), requires California state agencies to integrate existing water and other

environmental data from local, state, and federal agencies for the purpose of creating and maintaining a statewide integrated water data platform. In this research, we developed a process to systematically explore data needs for decision making to inform the design of data systems, focusing on California.

The aim of this paper is to contribute a better understanding of the practice of translating data into effective decision making by engaging stakeholders in data system development. The research has three main contributions. First, we develop the concept of a decision-driven data system, and assess how it might support improvements in informing management across a wide range of environmental sectors. Second, we examine and illustrate the concept's application in the California case study by defining attributes of a user-centered data and information system through stakeholder engagement. Third, we identify and characterize types of data limitations, and evaluate how a decision-driven, user-defined data system can address the data limitations experienced by users.

We first describe our methods, which involved working with stakeholders in California water management to develop and analyze a set of "use cases," short descriptions of decision making and the data needed to inform those decisions. We then develop a typology of different types of data limitations and gaps described by stakeholders, including gaps in data availability, accessibility, interoperability, and resolution. We propose technical, governance, and stakeholder engagement evaluation criteria to guide planning and building environmental data systems that account for these needs. By developing and describing a method for engaging stakeholders in the development of data systems, this article contributes to a better understanding of a crucial but understudied aspect of the practice of translating data into effective decision making, and offers recommendations applicable to a broad range of environmental and climate data and information systems.

METHODS

Leaders from the California Department of Water Resources (DWR), the California Council on Science and Technology (CCST) and researchers from University of California collaborated on a process of engaging stakeholders and evaluating data needs with the goal of ensuring that California's Open and Transparent Water Data Act results in an effective data system that improves water management in practice[1]. Our stakeholder engagement was centered around identification and analysis of "use cases"—brief descriptions of decision making associated with a specific outcome (such as balancing a basin water budget or responding to a harmful algal bloom) and the data needed to inform those decisions (fully described in Cantor et al., 2018). The idea of use cases was initially articulated in the field of computer sciences, based on the concept of developing data systems by starting with the end users' goals in mind in order to increase efficiency and efficacy (Alexander and Maiden, 2005; Kulak and Guiney, 2012). We adapted the use case approach from computer sciences to first systematically assess the data needs of California's water decision makers and other data users, then evaluate whether existing data and data systems met these needs, and finally to communicate these needs with technical developers of data systems and applications.

Use Case Development

We developed our application of the use case concept in collaboration with technical data system developers as well as data users. To begin, we asked the interrelated questions of *who* needs *what data* in *what form* to make *what decisions* (Kiparsky and Bales, 2017). We created a template (**Table 1**) to guide stakeholders in answering these questions in a systematic way, centered around a particular decision or goal.

Using the template in **Table 1**, we identified and developed 20 use cases (see Cantor et al., 2018). The use cases were compiled during three full-day-long facilitated workshops as well as additional meetings with stakeholders. We defined "stakeholder" broadly as including data producers and consumers with an interest in the outcomes of California's progress on

TABLE 1 | Use case template: Elements and definitions of a use case (adapted from Cantor et al., 2018).

Use case element	Definition
Objective	The decision, goal or desired action. The objective describes what the user is trying to accomplish. The objective is the goal or desired action on the part of the system user. Decisions could be investment and policy decisions (longer-term); programmatic implementation (medium-term); regulatory compliance; or operational decisions (short term).
Description	The description provides important context and background information that might help a reader understand the objective.
Participants	The participants include the main actor(s) or decision maker(s). Participants may also include other parties involved or affected by the decision or objective (in this case, note the main decision-maker).
Regulatory context	Regulatory context deriving from specific statutes or regulations and activities; legal operational constraints; specific government-agency programs or those under development; reporting requirements; and other regulated activities. It also includes physical and fiscal boundaries, frequency of reporting requirements and constraints.
Workflow	The workflow describes a progression of steps and specific actions taken by the participants in order to accomplish the objective.
Data sources	Data sources include existing data sources as well as gaps. This section describes the data already in use, along with additional sources that data users would like to see developed.
Data characteristics	Data characteristics includes notes about the type, form, and format of data that would be most useful for making decisions, and anything peculiar about the data.

[1]In this article, we build on and extend a 2018 report published by the Center for Law, Energy & the Environment at Berkeley Law, available at: https://doi.org/10.15779/J28H01. The initial report was published as a white paper intended largely for a California-based water policy and decision-maker audience. In this article, we strive to speak to a broader scholarly audience by expanding the theoretical framing, putting key ideas from the 2018 report into a more in-depth conversation with scholarly literature, extending the generalizable observations, and more fully developing and discussing the typology of data limitations.

water data, including academics, state and local agency representatives, non-governmental-organization representatives, community members, the private sector, and other water management practitioners. Workshop participants were selected through purposive sampling (Aarons et al., 2012; Ritchie et al., 2013) based on their relevant experience with data use or production related to the selected use cases.

The first two workshops, which produced eight use cases in total, each included 60–80 attendees. The majority of attendees worked with one of the state agencies named in California's Open and Transparent Water Data Act (AB 1755), so they attended in the capacity of their agencies, which had a direct stake in the process. Other attendees included academics, non-profit organization representatives, and others who saw themselves as having an interest in participating in water data system design and development. Lunch and opportunities for networking were provided as part of the workshops. Workshops began with an overview of the concept of data for decision making and the specific task of informing development of a data system. Participants then formed smaller breakout groups of 10–20

TABLE 2 | Example of completed use case: Groundwater recharge project planning.

Use case element	Use case: Planning a groundwater recharge project
Source	Data for Water Decision Making Workshop 1, February 9, 2017
Objective	To determine when, where, and how to recharge groundwater, with what water, in order to avoid declining groundwater levels through the recharge of groundwater.
Description	Under California's Sustainable Groundwater Management Act (SGMA), Groundwater Sustainability Agencies (GSAs) must avoid undesirable results including chronic lowering of groundwater levels. Managed Aquifer Recharge (MAR) is the use of, e.g., infiltration basins, green infrastructure, aquifer storage, and recovery wells to actively increase the amount of water that enters an aquifer. MAR can offset reductions in groundwater levels by increasing storage of water.
Participants	• GSA • Consultants • Local land use planners • State Water Resources Control Board and CA Department of Water Resources (interested in results of groundwater sustainability plan) • GSA constituents
Regulatory context	• Sustainable Groundwater Management Act • Other regulatory contexts: for example, CEQA, NEPA, water rights issues, water quality issues • Possible permits from SWRCB
Workflow	Identify potential source(s), quantity, timing, and cost of water available for recharge. Examine options for recharge areas based on geology, basin capacity, available land and land values, and water quality implications. Take into account basin characteristics such as subsurface characteristics, soil types, topography, current and planned land use, and basin capacity.
Data sources	• Water availability data: Water rights information, precipitation data, projected flows, projections/forecasts of water availability. ○ DWR California Data Exchange Center datasets: "California Statewide Water Conditions" (includes precipitation, snowpack, runoff forecasts, river runoff, and reservoir storage) ○ Executive Update on Hydrologic Conditions in CA (03/31/2017; updated monthly) ○ Annual Water Year Precipitation Summary ○ Reservoir Water Storage, by hydrologic region ○ USGS Current Water Data for California: Daily Streamflow Conditions ○ NOAA Precipitation Frequency Data Server (PFDS) ○ CA Water Board Electronic Water Rights Information Management System • Basin characteristics data: Soil types, basin capacity, subsurface characteristics, assimilative capacity, models of basin characteristics, evidence for natural recharge. ○ DWR Groundwater Basin Maps and Descriptions (Bulletin 118) ○ USGS Groundwater Modeling: California Groundwater Model Archive ○ UC Davis California Soil Research Lab Soil Agricultural Groundwater Banking Index (SAGBI) suitability index for groundwater recharge • Land use data: Available land, water quality concerns from past land use history, historical data on land use (requires both temporal and spatial dimensions). ○ DWR Land Use Survey data (available at county scale; available years vary) ○ USDA National Agricultural Statistics Service "Cropscape" Cropland Data Layer ○ USGS Global Land Cover Characteristics Data Base, Version 2.0 ○ CA Department of Conservation Farmland Mapping and Monitoring Program • Data gaps: ○ Water rights data may be incomplete or unavailable. ○ Groundwater pumping data may not be readily available. ○ Data on water demands for managed habitat, including state, federal and private wildlife refuges, hunting clubs, and incidental habitat areas
Data characteristics & further notes	To capture potential impacts of previous land uses (including contamination), land use data must include both historical and spatial dimensions. Spatial analysis can help find areas of overlap between various characteristics. Groundwater models may be required to make decisions in some cases, but not all. Existing groundwater models may be useful in some cases, but in other cases existing models may be insufficient. Not all required data is digitized, which presents problems for those seeking to access and use data. Uncertainties in this case include land use impacts on groundwater, as well as climate change and other uncertainties.

participants to develop use cases on pre-identified topics. Each group was given the use case template (**Table 1**) and had an assigned facilitator and note taker from the project team. We next identified and developed four additional use cases through a series of more-targeted, facilitated meetings with smaller groups of water data users and data producers with specific subject area expertise (for example, employees at the California State Water Resources Control Board involved in water rights), and worked directly with a range of non-governmental organizations and state agencies to identify and develop the remaining eight use cases using the template. Finally, a third, larger workshop was held toward the end of the use case process to present the initial use cases and findings to ∼100 attendees, and to solicit their feedback. The process thus evolved over time—from medium-sized workshops with a variety of water data users, to targeted meetings and one-on-one work to generate specific use cases, to a more general forum to present initial results.

The use cases encompassed a diversity of topics relevant to California water management, including groundwater management, environmental restoration, wetland monitoring, fishery management, urban and agricultural water management, water rights and water availability, capital investment, and drought contingency planning[2]. For example, some of the specific use case topics included "Management of environmental flows to protect salmon habitat," "Groundwater basin water budgets," "Water shortage contingency planning vulnerability assessment," and "Decision support system for harmful algal bloom response, communication, and mitigation." To provide a more detailed example, **Table 2** shows a completed use case on the topic of groundwater recharge project planning, and **Table 3** summarizes the specific data sources listed by stakeholders for this example use case.

While the sample of use cases does not comprehensively represent the entire landscape of California water management (for example, the cases covered many themes related to water quality, habitat, and water allocation, but water treatment utilities were largely unaddressed in the overall use case portfolio), the cases represent the complexity and breadth of water-management topics, and the selection of use cases was deliberately aligned with broader goals for California water (California Natural Resources Agency, 2016).

TABLE 3 | Specific data sources for groundwater recharge use case.

Topic	Description	Data source description
Water	Precipitation	DWR CDEC 2017 WY Precipitation Summary
Water	Hydrologic conditions	DWR CDEC Executive Update on Hydrologic Conditions in CA (03/31/2017; updated monthly)
Water	Reservoir storage	DWR CDEC reservoir storage by hydrologic region
Water	Statewide water conditions	DWR CDEC information on precipitation; snowpack; runoff forecasts; river runoff; and reservoir storage
Water	Precipitation	NOAA Precipitation Frequency Data Server (PFDS)
Agriculture, mapping	Farmland maps	California Department of Conservation Farmland Mapping and Monitoring Program (county-level data)
Water, mapping	Groundwater basin maps	DWR Bulletin 118 basin boundaries
Land use	Land use surveys	DWR Land Use Survey data (available at county scale; years vary)
Water	Water rights	SWRCB Electronic Water Rights Information Management System (eWRIMs)
Water	Groundwater models	USGS Groundwater Modeling: California Groundwater Model Archive
Water	Groundwater recharge suitability	SAGBI (Soil Ag Groundwater Banking Index) suitability index
Land use, mapping	Land cover maps	USGS Global Land Cover Characteristics Data Base Version 2.0
Agriculture	Agricultural land use	USDA National Agricultural Statistics Service Cropscape Cropland Data Layer
Water	Streamflow	USGS California streamflow data
Data gaps		
Water	Water rights	Incomplete or inaccessible; not digitized
Water	Groundwater pumping	Incomplete or unavailable records
Water	Water demands for habitat	Data not readily available

Analysis of Use Cases

We analyzed the collected use cases to identify patterns. We compiled the data sources listed for each use case and coded them according to thematic categories, including data topic and data provider. At least two members of the research team coded each data source and cross-checked their categorizations to enhance reliability. An emergent coding scheme (Holton, 2007) was used in order to capture the wide range of stakeholder-generated themes that were included in the use cases. Use case information was then cross checked and verified to remove errors and redundancy. We then identified data gaps, which we defined as data that were unavailable, inconsistently available, available only in formats that did not allow for interoperability, or that contained gaps in measurement or analysis. Data gaps were also coded and checked by multiple researchers for reliability. Finally, qualitative comments and feedback were coded using an emergent coding scheme, and were grouped according to themes to better understand stakeholder perspectives (see Cantor et al., 2018 for more detail). These classifications allowed us to systematically examine the availability of data sources, origin of data sources, the thematic topics covered, and gaps in data.

RESULTS

Data Types and Sources

Stakeholders used (or saw potential to use) water-related data for a wide variety of decisions. Some use cases were oriented toward directly answering a question, while other use cases involved collecting and integrating data into models or

[2]A full, detailed compilation of all 20 use cases and the specific data sources associated with each is available online at: https://doi.org/10.15779/J28H01.

decision support tools that in turn could be used to inform a number of different decisions. Some use cases focused on high-level investment and policy decisions, some on mid-level programmatic implementation, and others on day-to-day operational decisions, and regulatory compliance. Some cases represented concrete, already-existing decision processes, while others were more aspirational in describing desired goals.

Analysis of the use cases confirmed that water decision makers require a wide diversity of data types. While this may be no surprise to those versed in environmental management, it is important to consider the implications for data-system design. Water decision making requires a variety of data related to various natural, built, and socioeconomic systems in addition to data more traditionally associated with the hydrologic cycle (including precipitation and streamflow, water demand, groundwater, water quality, and water storage data) (**Table 4**). As illustrated in **Table 4**, the heterogeneity of data included in the use cases underscores the point that water data systems need to incorporate not only data obviously related to water (e.g., precipitation, streamflow), but also a wide range of related data—from agricultural land use to population data to climate-change projections—to fully support water-related decisions. The diversity of data and their associated spatial and temporal resolutions presents a challenge to data-system designers seeking to prioritize accessibility and interoperability for water decision making.

A relatively small number of state and federal public agencies provided the bulk of the data: just six federal and state agencies (including, at the federal level, the U.S. Geological Survey, the U.S. Department of Agriculture, and the National Oceanic and Atmospheric Administration, and at the California state level, the Department of Water Resources, the State Water Resource Control Board, and the Department of Fish and Wildlife) provided ∼two-thirds of the data sources mentioned by decision makers. Federal and state agencies made up about 90% of the data sources, while a variety of university, private, and non-governmental sources together made up the remaining 10%. Data systems seeking to integrate public data from the full range of federal and state data providers contributing to water management will need to rely upon common data standards between public agencies to ensure interoperability—a large task currently underway in California. At the same time, there was a long list of more specialized data that were cited for specific use in a single case. Water data users drew not only from public data from state and federal agencies, but also from a wide range of less-frequently-used other sources that were still highly important in certain decisions.

TABLE 4 | Broad range of data needs and topics represented within data needed for water decision making (adapted from Cantor et al., 2018).

Topic	Examples of data needed
Water-related data needs & topics	
Water demand & use	Water demand for different uses, water rights, water transfers, water usage, conservation, conjunctive use, urban water use, water deliveries, imports and diversions, pump locations, per capita water use, consumptive use, environmental use, domestic well data
Water supply	Precipitation, hydrologic conditions, streamflow, hydrographs, full natural flow, flow projections, snowpack, return flows, river stages, annual or seasonal volume, water year type
Water storage	Reservoir capacity, reservoir levels, reservoir surveys, snowpack storage, flood storage capacity, groundwater storage capacity
Water quality	Water quality, temperature, Total Maximum Daily Loads (TMDLs), water chemistry, sediments, contaminants, bacteria, algal blooms, biological indicators
Groundwater	Groundwater basin maps, elevation, models, pumping, quality, recharge suitability, storage, groundwater-dependent ecosystems, groundwater-surface water connectivity, Groundwater Sustainability Agency boundaries, well locations, well logs, aquifer storage capacity
Further data needs & topics beyond water-specific data	
Agriculture	Land use, crop types, evapotranspiration, pesticide use
Ecology	Species counts, habitat attributes, biodiversity, invasive species, wildlife population estimates, forest type, vegetation classification, aquatic resources, wetland boundaries
Geology & soils	Soil types, subsidence, geologic and hydrogeologic attributes
Infrastructure	Service area boundaries, water utility boundaries, pumping records, roads, water and energy use
Land use	Aerial imagery, city and county land use, land cover, land-use surveys, remote sensing data
Mapping & modeling	Watershed boundaries, surface waterways, terrain models, topographic surveys, elevation, county boundaries
Socioeconomic	Population, demographics, cost-benefit analyses, water pricing data, economic impact assessments, policy analyses
Weather and climate	Temperature, seasonal forecasts, climate projections, drought scenarios

Data Limitations

Stakeholder input and use cases revealed significant limitations in data and information availability (**Figure 1**). Some critical data were not available at all (limitation type 1). For example, data about groundwater extraction by individual water users was not systematically collected. As another example, data related to water demand by different interests such as recreation, or socioeconomic data such as valuation by different interests, pricing, or willingness to pay, was not readily available.

Other data were inaccessible or hard to use (limitation type 2). For example, some datasets were only published as PDF files or were not machine readable, and other data were password protected, required a fee to access, or were otherwise inaccessible. Other data had been transformed into maps or visualization tools, but the underlying data were not readily available. In one notable example, most information on California water rights only existed in paper form in a vault in the state capitol, rather than in an accessible digital database (although there have since been efforts to digitize this information).

Other data had low interoperability (limitation type 3). For example, stakeholders described datasets that were collected for specific purposes and were therefore not intended for interoperability. Multiple data producers had their own processes

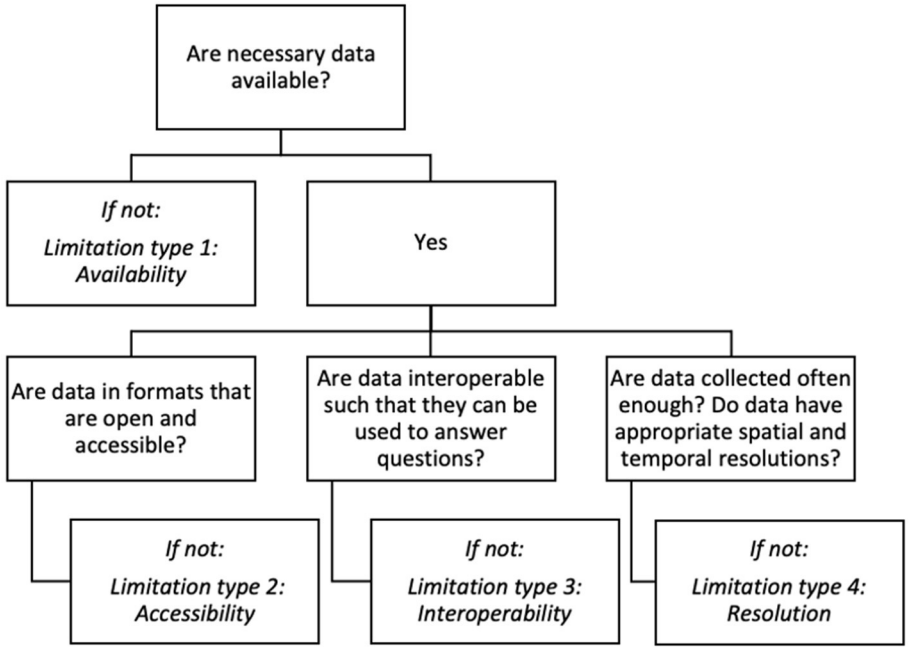


FIGURE 1 | Types of data limitations.

for data collection, storage, and documentation. The result was that data and IT systems could not exchange information with each other in standard ways allowing for comparison, aggregation, and analysis.

Finally, some data were not gathered using standardized approaches, or were not collected at useful time intervals or consistent spatial resolutions (limitation type 4). For example, data can be collected seasonally, monthly, or daily but this may not line up with decision-making needs. As another specific example, the California Department of Water Resources divides California into different hydrologic regions, but these boundaries did not exactly match USGS hydrologic boundaries, making it difficult to integrate multiple data sets.

Limitations in accessibility, interoperability, and resolution (types 2, 3, and 4) mean that some data sources can effectively constitute data gaps even if data technically exist.

DISCUSSION

Scholarship from environmental science and management has outlined guiding principles for how data can ideally guide decision making (Cortner, 2000; Cash et al., 2003; Holmes and Clark, 2008; Lemos and Rood, 2010). Data and information, beyond providing a snapshot of the state of the environment, should be *useful*, which refers to functionality and desirability for decision makers, as well as *usable*, which refers to how well data inform decision making processes in practice (Lemos and Rood, 2010). Data and information must also be *salient* (relevant to decision makers), *credible* (accurate from a scientific perspective), and *legitimate* (produced in a way that is perceived as respectful, unbiased, and fair) (Cash et al., 2003).

In this paper, we apply these principles to the mechanisms through which data are stored, published, accessed, and used. Drawing from our stakeholder engagement and analysis, we identified three categories of considerations for developing useful and usable water data systems that are salient, credible, and legitimate: (1) technical elements, including data interoperability, spatiotemporal resolution, documentation and quality; (2) governance, including funding and operating of systems across institutions; and (3) stakeholder engagement. Here we discuss each of these categories, then use them to inform criteria to evaluate a water data system.

Technical Considerations

Most of the use cases in our analysis integrated multiple data sources spanning a variety of thematic categories and sourced from a range of different data providers. The extraordinary heterogeneity of water data (**Table 4**) reflects how water decisions must often consider hydrologic, ecological, climate and other natural-system phenomena (e.g., streamflow, groundwater levels, species abundance, temperature, etc.) as well as characteristics associated with human and built systems (e.g., land use, crop types, built infrastructure, etc.). It also reflects institutional realities: water data are produced, housed, and maintained by multiple entities from disparate sectors.

Our analysis showed that there are significant limitations in data availability (**Figure 1**), including non-existent data and available but difficult-to-access data. Interoperability (limitation type 3) presented a particularly significant problem, and based on our analysis, it became evident that interoperability of

multiple data sources from different providers is key to the success of an environmental data system (**Figure 1**). The current lack of uniform, accessible, interoperable, and ultimately usable data hampers evidence-based water management in California (Escriva-Bou et al., 2016). Datasets are produced for a variety of primary purposes, and thus do not always share metadata or data-quality standards. Given our finding that a relatively small number of state and federal agencies provided a large fraction of needed data, there is significant potential for interoperability to improve by focusing on those agencies. Stakeholders also noted challenges related to spatial and temporal resolution of data collection (limitation type 4), which are related to interoperability (Gibson et al., 2000).

To address the interoperability challenge, participants in our project discussed the relative benefits of centralized vs. federated data systems. A centralized system such as those used by multiple federal agencies can readily implement uniform data standards and respond to diverse user needs. Yet federated data systems were preferred by many participants. Federated data systems connect multiple independent data systems through common standards, conventions, and protocols, while keeping those independent systems autonomous (Busse et al., 1999; Blodgett et al., 2016). Our research showed that data users relied upon a wide range of data produced and distributed by a variety of state and federal agencies and other data producers. Given the reliance on a range of distributed data sources from independent organizations, a federated data system may have advantages. A successful interoperable federated system requires *clear* standards for data quality, metadata, and technical requirements. Standards do not have to be created from scratch: for example, projects such as Hydroshare and the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), a cyberinfrastructure system to integrate diverse environmental datasets, have laid significant groundwork for methods to define and store metadata (Peckham and Goodall, 2013; Agarwal et al., 2017; Varadharajan et al., 2019). Here, it is worth highlighting the importance of clear standards, as data managers across different agencies and organizations may believe their standards are aligned but in practice, they may not be aligned sufficiently to support an effective federated system.

Workshop participants emphasized the importance of traceability, clear identification of sources, and documentation of uncertainties, all of which contribute to an assessment of data limitations (**Figure 1**). A data system drawing from multiple sources requires clear protocols for data quality assurance and documentation throughout all stages of the data life cycle. Structuring data according to set standards can facilitate integration between multiple data providers (Blodgett et al., 2016). Georeferencing of data is also critical for many water-related analyses. Archiving practices also require thought, as they are important to prevent data losses. One solution is the use of unique digital object identifiers (DOIs) for data sets (Paskin, 2010; Wilkinson et al., 2016), which can address traceability concerns by ensuring that data sets persist even if websites are reorganized and can assist with versioning, quality assistance/quality control, and referencing. For continually updated datasets, making versioned DOI sets of data would be a helpful best practice across agencies.

The range of use cases identified in this research also showed that different data users need data in different formats. In some cases, stakeholders and researchers preferred raw data which they could analyze and translate themselves into information. In other cases, stakeholders required quality-controlled data with transformed formats that could be readily input into decision-support systems, hydrologic models, workflows, visualization software, water-budget calculation, or other analytical tools.

Governance Considerations

Open data are important for sustainable and inclusive environmental management and water governance in particular (De Stefano et al., 2012; Chini and Stillwell, 2020), and can help make environmental governance more transparent, accountable, and efficient (Blodgett et al., 2016; Mayton and Story, 2018). Stakeholders in our research emphasized that developing and maintaining an open and transparent water data system requires not just making existing data more readily available, but also requires thoughtful governance and sustainable funding. Strategies for generating a sustainable funding source and governance model for a water data system have been proposed and adopted by the state of California. These involve a consortium of state, NGO, and private-sector actors working collaboratively (Huttner et al., 2018).

Participants in our stakeholder engagement noted that resources are needed throughout the information pipeline: this includes data system design, quality control, decision support and analysis tools, archiving, user support and continued system innovation. Building and maintaining a sustainable data system will therefore require investment in addressing limitations in data availability, accessibility, interoperability and resolution (**Figure 1**). To maximize usability over time, long-term funding models must be carefully thought out, with special consideration given to openness of data systems. Again, a federated system has benefits in this area: while a federated system with multiple funding streams may be vulnerable to losing one or more data streams, it also provides resilience by being distributed. It can also incorporate incremental additions from legislative actions that introduce new data sources or systems that meet new or emerging needs.

In addition to funding, an effective data system relies upon robust institutions to coordinate decision making and actions around how the data system is structured and used (Huttner et al., 2018). A framework that does not address institutional concerns increases the risk of data system failure from lack of coordination, underinvestment, or lack of trust and buy-in. Stakeholders noted the importance of trust, confidence, and credibility within and between institutions, which are widely recognized as important in water resources management generally, but can be forgotten when the focus is on the technical aspects of data systems (Jackson, 2006).

Data systems benefit from participation of data providers because their adherence to standards is important for interoperability and their involvement in those standards is

a way to facilitate that adherence. Governance mechanisms such as mandates for incorporating standard metadata and data-quality procedures could help ensure that agencies participate in a federated system. The bulk of the data used by stakeholders in our analysis came from public agencies. Legislative and regulatory mandates could be a way to encourage participation of these agencies. Still, a large handful of data sources identified as useful or necessary came from a wide variety of non-governmental stakeholders. Such smaller data providers may require incentives to fully participate in a system if adhering to protocols involves costs. For example, "intervener funding" (financial support that helps stakeholders to effectively participate in agency proceedings) could help support engagement of non-governmental data producers (Kiparsky et al., 2016). Another mechanism to encourage participation could involve requiring that state-funded projects make data interoperable and publicly available (similar to current National Science Foundation requirements for data management plans and data publication).

This raises a particular conundrum for environmental data systems design: the distinction between public and non-public data. While it may be possible (although far from straightforward) to require openness and transparency of data from federal, state, and local agencies, there remains a large category of non-public data. Other sources of data include nonprofit data sources, but also private data sources that present additional complications with regards to openness and transparency. It also may be more difficult to enact requirements or incentives for interoperability with these non-public data sources, meaning that they are likely to be more difficult to integrate, even though they may provide valuable information.

Stakeholder Engagement

Ensuring that an environmental data system is sufficient, accessible, useful and used (California Department of Water Resources, 2020) hinges on meaningful, ongoing relationships with data users. Successful stakeholder engagement requires many things: recognition of common goals, time to develop functional relationships, common vocabulary, careful facilitation and ongoing maintenance of relationships, and resources. Developing environmental data systems that are sufficient, accessible, useful, and used requires both usable technical cyberinfrastructure, good governance, and funding sufficient to support both technical infrastructure and governance.

We found that engaging knowledgeable stakeholders with detailed understanding of data needs and workflows involved in different aspects of water-related decision making is essential to identifying key aspects of data system usability. We also note the importance of engaging those who hold a stake in water decisions but do not have in-depth technical knowledge. To support communication, we used professional facilitation in larger meetings to ensure that project goals were articulated clearly and concisely. We also found it useful to engage stakeholders through different formats to serve different project goals. Larger workshops were helpful in communicating overall aims to a broader audience, including those with influence over policy decisions. Smaller meetings enabled focused conversations with specific groups of people with targeted technical knowledge. Working directly with organizations to identify use cases was an effective way to engage additional stakeholders.

User-focused data-system development can thus be framed as an adaptive management cycle (Pahl-Wostl, 2007) that includes multiple iterations of planning, implementation, and evaluation. Stakeholder engagement should be formally integrated into this cycle from an early stage to increase usability of the data system (Welp et al., 2006; Reed, 2008). Because decision-maker needs and technological capacities change over time, a data system must be adaptable (McNie, 2007; Hanseth and Lyytinen, 2016), and as new decision-maker needs and new technologies arise, a data system must evolve to remain useful. The process of identifying stakeholder objectives, translating these objectives into functional and technical requirements, and using these objectives to inform the development of data systems, can be built into the life cycle of data system design.

Evaluating Decision-Driven Data Systems

To integrate the technical, governance, and stakeholder-engagement considerations identified during our research and outlined here, we propose a set of questions to guide evaluating the success of an environmental data system (**Table 5**). This set of evaluation criteria incorporates the multiple types of data limitations identified in this paper (see **Figure 1**) and includes technical considerations, governance considerations, and stakeholder engagement considerations.

TABLE 5 | Proposed criteria for evaluating success of an environmental data system (adapted from Cantor et al., 2018).

	Evaluation criteria
Addressing data limitations (see **Figure 1**)	Are appropriate data readily available? Are data accessible in open, transparent, and usable formats? Are data from multiple sources interoperable? Are data available at appropriate spatial and temporal resolution?
Technical considerations	Is documentation adequate? Are standards for metadata, data quality, and technical requirements clear to data managers? Does the data system effectively support synthesis and analysis? Are systems regularly updated?
Governance considerations	Is there institutional commitment by key organizations to use and maintain the system? Do incentives exist to ensure participation by data providers and users? Are data providers participating, in practice? Are sufficient resources allocated to long-term maintenance? Is there a plan to ensure financial stability over time?
Stakeholder engagement considerations	Are data users engaged meaningfully at key points in data system development? Is involvement of stakeholders an ongoing process? Is the system based on an understanding of decision-making contexts and user needs? Do users believe the system is useful and usable? Is the system used in practice to inform decision making?

These evaluation questions are in line with those developed by others, such as the "FAIR" (Findable, Accessible, Interoperable, Reusable) Guiding Principles (Wilkinson et al., 2016), but also add to these guiding principles through inclusion of governance and stakeholder engagement criteria, which we argue are crucial to data system success and should therefore be included alongside the more technical considerations. These questions are targeted at data providers, although many of the evaluation questions require the input of data users. The questions do not provide quantitative measurements or metrics, which would need to be specific to an individual data system; instead, these questions provide a guide for data providers to consider how well their system is serving users. Our evaluation criteria include the very important question of whether the data system is ultimately used in practice to inform decision making—perhaps the key indicator of success.

A crucial indicator of the success of our process can be found in the formal uptake of the concepts of decision-driven water data systems into state processes required by statute (California Department of Water Resources, 2020). Based on the results of our workshops and analysis, our recommendation of a federated, use case-driven water data platform that connects independent databases while prioritizing and managing data based on how data will be used has been adopted by California's AB 1755 Partner Agency Team. Another indicator of success is in the influence of other subsequent processes. For example, organizers of a recent workshop on water data in Texas used a use case approach based on our template and model (Rosen and Roberts, 2018). Drawing from our approach, the Texas workshop organizers also started from the basic principle that water data systems must be responsive to stakeholder needs in order to support decision making in practice (Rosen and Roberts, 2018).

Challenges and Limitations

In the course of our study, we experienced inevitable obstacles related to the challenges of working with stakeholders. We found that (as might be expected) engaging with stakeholders meaningfully is time consuming and takes resources, and it is important not to underestimate the capacity needed to conduct effective stakeholder engagement. We also learned that developing a sufficiently clear articulation of an objective or decision around which to anchor a use case was not a simple task. In practice, it proved difficult for larger groups with greater diversity in their topical expertise to agree upon objectives. At the same time, engaging participants in groups helped ensure that different stakeholders with various types of expertise could provide different types of knowledge.

The work presented in this paper has several limitations. First, many problems in the water sector are highly complex. They may involve multiple levels or stages of decisions: in this project we mainly tested the use case approach on single-stage decisions and the concept would need to be adapted or used iteratively to account for multi-stage decisions. Second, the use case framework is helpful for identifying data gaps, but does not necessarily provide a mechanism for evaluating the relevance or significance of such gaps. That is, some limitations represent a critical bottleneck to decision processes, while other limitations do not actively constrain decisions from going forward but still impact the quality of those decisions. Future efforts to implement use cases and identify data limitations could ask participants about the relative impact of a particular data limitation. Third, we developed this methodology with the creation of a new data system in mind; we did not test the applicability of the methodology to existing data systems that already have established formats and tools. Future work could test our proposed evaluation criteria by applying it to an existing system. Finally, given growing interest in water data from global organizations (for example, the World Water Data Initiative, led by the World Meteorological Organization) there may be opportunity for future research to examine how these concepts apply to different scales.

We also acknowledge that conflicts in water management go beyond data. Water issues and proposed solutions frequently evoke controversy and can be hotly contested. In this project we did not directly address the complex politics and disagreements between different stakeholder groups that frequently emerge in environmental governance and problem-solving. While data can, ideally, help inform and evaluate solutions to difficult and controversial issues, we recognize that lack of data is not the only issue preventing good water governance, and that conflict will not be resolved solely through data availability.

CONCLUSIONS

Applying the concept of decision-driven data systems to environmental management is an important contribution to the overarching goal of enhancing data-informed environmental decision making. Our case study of water data in California identified specific ways in which less-than-adequate data sources and systems are currently constraining decision making, resulting in data gaps, ineffective delivery of overlapping data needs across sectors, and limiting secondary uses of data. Based on this research, we argue that to effectively inform water management, data systems must begin with a strong understanding of decision makers' data needs, and should engage decision makers to identify and address different types of data gaps and limitations. Otherwise, data systems risk being of limited utility, an inefficient use of resources, and a source of frustration for users.

Our work shows that useful and usable environmental data systems must consider not only technical elements, but also data system governance and stakeholder engagement. In the case we examined, given the distributed nature of data required by stakeholders, the independence of disparate agencies, and the need for interoperability, federated data systems have the potential to address technical and governance issues. In terms of stakeholder engagement, a responsive data system requires ongoing analysis of stakeholder objectives and translation of those objectives into functional and technical

requirements. Resources for engagement should be considered part of infrastructure investment, because they ultimately can help inform usability of a data system and prevent wasting future resources.

Supporting environmental decision making through decision-driven data systems is a long-term project involving ongoing attention to meaningful engagement with decision makers and other data stakeholders. As is true of other forms of infrastructure, the full value of investments in environmental data may only become apparent when it is sorely needed: for example, the value of water data becomes apparent during droughts, floods, or other crisis events. In such events, access to information may be a crucial factor in determining whether or not rapid and effective decisions can be reached. This prospect alone justifies the forward-looking efforts described in this article, and, more generally, greater attention to the role of data in environmental management and sustainability.

DATA AVAILABILITY STATEMENT

A full, detailed compilation of all 20 use cases developed for this project and the specific data sources associated with each is available online at: https://doi.org/10.15779/J28H01. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

AC: conceptualization, methodology, investigation, data curation, analysis, and writing—original draft. MK: conceptualization, methodology, investigation, analysis, writing—original draft, supervision, project administration, and funding acquisition. SH and RK: conceptualization and writing—review and editing. LP: analysis, data curation, and writing—review and editing. KG: project administration, investigation, and writing—review and editing. GD and CM: resources, investigation, and writing—review and editing. RB: conceptualization, supervision, project administration, funding acquisition, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the University of California Office of the President (UCOP), through the UC Water Security and Sustainability Research Initiative (UCOP Grant No. 13941), and by the Water Foundation. Support for SH was provided by U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-AC02-05CH1123.

ACKNOWLEDGMENTS

An earlier report on this study, including a complete description of the California use case development process can be found in a 2018 report written by the authors of this article and published as a report by UC Berkeley School of Law, Center for Law, Energy & the Environment. The 2018 report is available at https://doi.org/10.15779/J28H01. Thanks to the workshop participants, facilitators, and use case contributors for sharing their time and expertise. We thank workshop sponsors, including the California Council on Science and Technology (CCST), UC Water, Lawrence Berkeley National Laboratory, the California Department of Water Resources, and the Water Foundation, Leigh Bernacchi, Luke Sherman, and Amber Mace for assistance in organizing the workshops, John Helly, Richard Roos-Collins, Holly Doremus, and Nell Green Nylen for discussions on the concepts presented in this paper, and reviewers for helpful comments on versions of this paper.

REFERENCES

Aarons, G. A., Fettes, D. L., Sommerfeld, D. H., and Palinkas, L. A. (2012). Mixed methods for implementation research: application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child Maltreat.* 17, 67–79. doi: 10.1177/1077559511426908

Agarwal, D., Varadharajan, C., Cholia, S., Snavely, C., Hendrix, V., Gunter, D., et al. (2017). "Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)-a new US DOE data archive," in *American Geophysical Union Fall Meeting* (New Orleans, LA).

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

Alexander, I. F., and Maiden, N. (2005). *Scenarios, Stories, Use Cases: Through the Systems Development Life-Cycle*. Hoboken, NJ: John Wiley & Sons.

Bakker, K. (2012). Water security: research challenges and opportunities. *Science (80-.)* 337, 914–915. doi: 10.1126/science.1226337

Blodgett, D., Read, E., Lucido, J., Slawecki, T., and Young, D. (2016). An analysis of water data systems to inform the open water data initiative. *JAWRA* 52, 845–858. doi: 10.1111/1752-1688.12417

Busse, S., Kutsche, R. D., Leser, U., and Weber, H. (1999). Federated information systems: Concepts, terminology and architectures. *Forschungsberichte des Fachbereichs Informatik* 99, 1–38.

California Council on Science and Technology (2014). *Achieving a Sustainable California Water Future Through Innovations in Science and Technology*. Sacramento, CA.

California Department of Water Resources (2020). *Open and Transparent Water Data Act- Implementation Journal*. Sacramento, CA.

California Natural Resources Agency (2016). *California Water Action Plan 2016*. Sacramento, CA.

Cantor, A., Kiparsky, M., Kennedy, R., Hubbard, S., Bales, R., Pecharroman, L. C., et al. (2018). *Data for Water Decision Making: Informing the Implementation of California's Open and Transparent Water Data Act through Research and Engagement*. Berkeley, CA: UC Berkeley Law, Center for Law, Energy & the Environment.

Cash, D. W., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., Guston, D. H., et al. (2003). Knowledge systems for sustainable development. *Proc. Natl. Acad. Sci.* 100, 8086–8091. doi: 10.1073/pnas.1231332100

Chini, C. M., and Stillwell, A. S. (2020). Envisioning blue cities: urban water governance and water footprinting. *J. Water Resour. Plan. Manag.* 146:4020001. doi: 10.1061/(ASCE)WR.1943-5452.0001171

Cortner, H. J. (2000). Making science relevant to environmental policy. *Environ. Sci. Policy* 3, 21–30. doi: 10.1016/S1462-9011(99)00042-8

De Stefano, L., Hernández-Mora, N., López Gunn, E., Willaarts, B., and Zorrilla-Miras, P. (2012). "Public participation and transparency in water management," in *Water, agriculture and the environment in Spain: Can we square the circle?*, eds L. De Stefano and M. Ramon Llamas (Boca Raton, FL: CRC Press/Balkema; Taylor & Francis Group), 217–225. doi: 10.1201/b13078-22

Escriva-Bou, A., McCann, H., Hanak, E., Lund, J., and Gray, B. (2016). *Accounting for California's Water*. San Francisco, CA: Public Policy Institute of California. doi: 10.5070/P2CJPP8331936

Gibson, C. C., Ostrom, E., and Ahn, T.-K. (2000). The concept of scale and the human dimensions of global change: a survey. *Ecol. Econ.* 32, 217–239. doi: 10.1016/S0921-8009(99)00092-0

Green Nylen, N., Kiparsky, M., Owen, D., Doremus, H., and Hanemann, M. (2018a). *Addressing Institutional Vulnerabilities in California's Drought Water Allocation, Part 1: Water Rights Administration and Oversight During Major Statewide Droughts, 1976–2016*. Berkeley, CA: California's Fourth Climate Change Assessment, California Natural Resources Agency.

Green Nylen, N., Kiparsky, M., Owen, D., Doremus, H., and Hanemann, M. (2018b). *Addressing Institutional Vulnerabilities in California's Drought Water Allocation, Part 2: Improving Water Rights Administration and Oversight for Future Droughts*. Berkeley, CA: UC Berkeley Law, Center for Law, Energy & the Environment.

Hanak, E. (2011). *Managing California's Water: From Conflict to Reconciliation*. San Francisco, CA: Public Policy Institute of CA.

Hanseth, O., and Lyytinen, K. (2016). "Design theory for dynamic complexity in information infrastructures: the case of building internet," in *Enacting Research Methods in Information Systems* (Berlin: Springer), 104–142. doi: 10.1007/978-3-319-29272-4_4

Holmes, J., and Clark, R. (2008). Enhancing the use of science in environmental policy-making and regulation. *Environ. Sci. Policy* 11, 702–711. doi: 10.1016/j.envsci.2008.08.004

Holton, J. A. (2007). "The coding process and its challenges," in *The SAGE Handbook of Grounded Theory,* ed K.Charmaz (Thousand Oaks, CA: SAGE Publications), 265–290. doi: 10.4135/9781848607941.n13

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

Huttner, N., King, K., and Whitney, J. (2018). *Governance and Funding for Open and Transparent Water Data*. Redwood City, CA: Redstone Strategy Group.

Jackson, S. (2006). "Water models and water politics: design, deliberation, and virtual accountability," in *Proceedings of the 2006 International Conference on Digital Government Research* (San Diego, CA: Digital Government Society of North America), 95–104. doi: 10.1145/1146598.1146632

Kallis, G., Kiparsky, M., Milman, A., and Ray, I. (2006). Glossing over the complexity of water. *Science (80-.)* 314, 1387–1388. doi: 10.1126/science.314.5804.1387c

Kiparsky, M., and Bales, R. (2017). Advanced data would improve how California manages water. *Sacramento Bee*.

Kiparsky, M., Owen, D., Green Nylen, N., Doremus, H., Christian-Smith, J., Cosens, B., et al. (2016). *Designing Effective Groundwater Sustainability Agencies: Criteria for Evaluation of Local Governance Options*. Berkeley, CA: UC Berkeley Law, Center for Law, Energy & the Environment.

Kiparsky, M., Sedlak, D. L., Thompson, B. H. Jr, and Truffer, B. (2013). The innovation deficit in urban water: the need for an integrated perspective on institutions, organizations, and technology. *Environ. Eng. Sci.* 30, 395–408. doi: 10.1089/ees.2012.0427

Kulak, D., and Guiney, E. (2012). *Use Cases: Requirements in Context*. Boston, MA: Addison-Wesley.

Larsen, S., Hamilton, S., Lucido, J., Garner, B., and Young, D. (2016). Supporting diverse data providers in the open water data initiative: communicating water data quality and fitness of use. *JAWRA* 52, 859–872. doi: 10.1111/1752-1688.12406

Lemos, M. C., and Rood, R. B. (2010). Climate projections and their impact on policy and practice. *Wiley Interdiscip. Rev. Clim. Chang.* 1, 670–682. doi: 10.1002/wcc.71

Mayton, H., and Story, S. D. (2018). Identifying common ground for sustainable water data management: the case of California. *Water Policy* 20, 1191–1207. doi: 10.2166/wp.2018.047

McNie, E. C. (2007). Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature. *Environ. Sci. Policy* 10, 17–38. doi: 10.1016/j.envsci.2006.10.004

Mosavi, A., Ozturk, P., and Chau, K. (2018). Flood prediction using machine learning models: literature review. *Water* 10:1536. doi: 10.3390/w10111536

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B., et al. (2019). Surrogate Optimization of Deep Neural Networks for Groundwater Predictions. arxiv [preprint].arxiv:1908.10947.

Neumann, J., Arnal, L., Emerton, R., Griffith, H., Hyslop, S., Theofanidi, S., et al. (2018). Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity. *Geosci. Commun.* 1, 35–57. doi: 10.5194/gc-1-35-2018

Oroza, C. A., Zheng, Z., Glaser, S. D., Tuia, D., and Bales, R. C. (2016). Optimizing embedded sensor network design for catchment-scale snow-depth estimation using LiDAR and machine learning. *Water Resour. Res.* 52, 8174–8189. doi: 10.1002/2016WR018896

Pahl-Wostl, C. (2007). Transitions towards adaptive management of water facing climate and global change. *Water Resour. Manag.* 21, 49–62. doi: 10.1007/s11269-006-9040-4

Paskin, N. (2010). Digital object identifier (DOI®) system. *Encycl. Libr. Inf. Sci.* 3, 1586–1592. doi: 10.1081/E-ELIS3-120044418

Pau, G. S. H., Shen, C., Riley, W. J., and Liu, Y. (2016). Accurate and efficient prediction of fine-resolution hydrologic and carbon dynamic simulations from coarse-resolution models. *Water Resour. Res.* 52, 791–812. doi: 10.1002/2015WR017782

Peckham, S. D., and Goodall, J. L. (2013). Driving plug-and-play models with data from web services: a demonstration of interoperability between CSDMS and CUAHSI-HIS. *Comput. Geosci.* 53, 154–161. doi: 10.1016/j.cageo.2012.04.019

Provost, F., and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data* 1, 51–59. doi: 10.1089/big.2013.1508

Reed, M. S. (2008). Stakeholder participation for environmental management: a literature review. *Biol. Conserv.* 141, 2417–2431. doi: 10.1016/j.biocon.2008.07.014

Ritchie, J., Lewis, J., Nicholls, C. M., and Ormston, R. (2013). *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. Thousand Oaks, CA: Sage Publications.

Rosen, R. A., and Roberts, S. V. (2018). *Connecting Texas Water Data Workshop: Building an Internet for Water*. San Antonio, TX: Water Resources Science and Technology Book and E-Book Publications and Reports.

Schmidt, F., Wainwright, H. M., Faybishenko, B., Denham, M., and Eddy-Dilek, C. (2018). In situ monitoring of groundwater contamination using the Kalman filter. *Environ. Sci. Technol.* 52, 7418–7425. doi: 10.1021/acs.est.8b00017

Varadharajan, C., Agarwal, D. A., Brown, W., Burrus, M., Carroll, R. W. H., Christianson, D. 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

Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P. (2016). How seasonal forecast could help a decision maker: an example of climate service for water resource management. *Adv. Sci. Res.* 13, 51–55. doi: 10.5194/asr-13-51-2016

Vogel, R. M., Lall, U., Cai, X., Rajagopalan, B., Weiskel, P. K., Hooper, R. P., et al. (2015). Hydrology: the interdisciplinary science of water. *Water Resour. Res.* 51, 4409–4430. doi: 10.1002/2015WR017049

Welp, M., de la Vega-Leinert, A., Stoll-Kleemann, S., and Jaeger, C. C. (2006). Science-based stakeholder dialogues: theories and tools. *Glob. Environ. Chang.* 16, 170–181. doi: 10.1016/j.gloenvcha.2005.12.002

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., and Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18

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

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

frontiers in Climate

PERSPECTIVE
published: 24 November 2021
doi: 10.3389/fclim.2021.763420

Open Science Expectations for Simulation-Based Research

Gretchen L. Mullendore[*†], *Matthew S. Mayernik*[†] *and Douglas C. Schuster*[†]

National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), Boulder, CO, United States

There is strong agreement across the sciences that replicable workflows are needed for computational modeling. Open and replicable workflows not only strengthen public confidence in the sciences, but also result in more efficient community science. However, the massive size and complexity of geoscience simulation outputs, as well as the large cost to produce and preserve these outputs, present problems related to data storage, preservation, duplication, and replication. The simulation workflows themselves present additional challenges related to usability, understandability, documentation, and citation. These challenges make it difficult for researchers to meet the bewildering variety of data management requirements and recommendations across research funders and scientific journals. This paper introduces initial outcomes and emerging themes from the EarthCube Research Coordination Network project titled "What About Model Data? - Best Practices for Preservation and Replicability," which is working to develop tools to assist researchers in determining what elements of geoscience modeling research should be preserved and shared to meet evolving community open science expectations.

Specifically, the paper offers approaches to address the following key questions:

- How should preservation of model software and outputs differ for projects that are oriented toward knowledge production vs. projects oriented toward data production?
- What components of dynamical geoscience modeling research should be preserved and shared?
- What curation support is needed to enable sharing and preservation for geoscience simulation models and their output?
- What cultural barriers impede geoscience modelers from making progress on these topics?

Keywords: data, preservation, replicability, model, simulation

OPEN ACCESS

Edited by:
Lauren A. Jackson, National Oceanic and Atmospheric Administration (NOAA), United States

Reviewed by:
Derrick P. Snowden, U.S. Integrated Ocean Observing System, United States
Scott Cross, National Oceanic and Atmospheric Administration (NOAA), United States
Kemal Cambazoglu, University of Southern Mississippi, United States

***Correspondence:**
Gretchen L. Mullendore
gretchen@ucar.edu

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *23 August 2021*
Accepted: *04 November 2021*
Published: *24 November 2021*

Citation:
Mullendore GL, Mayernik MS and Schuster DC (2021) Open Science Expectations for Simulation-Based Research. Front. Clim. 3:763420. doi: 10.3389/fclim.2021.763420

INTRODUCTION

Dynamical models are central to the study of Earth and environmental systems as they are used to simulate specific localized phenomena, such as tornadoes and floods, as well as large-scale changes to climate and the environment. High-profile projects such as the Coupled Model Intercomparison Project (CMIP) have demonstrated the potential value of sharing simulation output data broadly within scientific communities (Eyring et al., 2016). However, more focus is needed on open science

challenges related to simulation output. Researchers face a bewildering variety of data management requirements and recommendations across research funders and scientific journals, few of which have specific and useful guidance for how to deal with simulation output data.

Simulation-based research presents a number of significant data-related problems. First, simulations can generate massive volumes of output. Increased computing power enables researchers to simulate weather, climate, oceans, watersheds, and many other phenomena at ever-increasing spatial and temporal resolutions. It is common for simulations to generate tens or hundreds of terabytes of output, and larger projects like the CMIPs generate petabytes of output.

Second, interdependencies between hardware and software can limit the portability of models, and make the long-term accessibility of their output problematic. Many current data management guidance documents provided by scientific journal publishers conflate scientific computational models with software, thereby not addressing whether/how to archive model outputs. Equating computational models with software does not add much clarity to these recommendations, as ensuring "openness" of software is itself a significant challenge (Easterbrook, 2014; Irving, 2016). Models in many cases involve interconnections between community models, open source software components, and custom code written to investigate particular scientific questions. Large-scale models often also borrow and extend specific components from other models (Masson and Knutti, 2011; Alexander and Easterbrook, 2015).

Third, the lack of standardization and documentation for models and their output makes it difficult to achieve the goals of open and FAIR data initiatives (Stall et al., 2018). While this problem is not unique to simulation-based research, it has stimulated a number of initiatives to develop more consistency in how variables are named within simulation models, how models themselves are documented, and in how model output data are structured and described (Guilyardi et al., 2013; Heydebreck et al., 2020; Eaton et al., 2021).

The result is that the long-term value of simulation outputs is harder to assess than of observational data, and requires focused effort if the value is to be achieved. Key questions that challenge researchers who use such models are "what data to save" and "for how long?" Guidance on these questions is particularly vague and inconsistent across funders and publishers. "Reproducibility" is likewise difficult to define and achieve for computational simulations. Many different approaches have been proposed for what is required to successfully reproduce prior research (Gundersen, 2021). Within climate science, for example, bitwise reproducibility of model runs has not been a primary focus due to the non-linear nature of the phenomena being simulated, as well as the differences in bitwise output that occur when transferring models to different computing hardware (Bush et al., 2020).

Following the terminology of the recent US National Academies of Sciences, Engineering, and Medicine (2019) report on "Reproducibility and Replicability in Science," the primary goal in Earth and environmental science research is replicability of findings related to the physical system being simulated, not bitwise computational reproducibility. In other words, the goal is to have enough information about research workflows and selected derived data outputs to communicate the important configurational characteristics to allow a future researcher to build from the original study.

This paper builds on the initial findings of the EarthCube Research Coordination Network (RCN) project titled "What About Model Data? - Best Practices for Preservation and Replicability" (https://modeldatarcn.github.io/) to address the following key questions related to open science and simulation-based research:

- How should preservation of model software and outputs differ for projects that are oriented toward knowledge production vs. projects oriented toward data production?
- What elements of dynamical geoscience modeling research should be preserved and shared?
- What curation support is needed to enable sharing and preservation for geoscience simulation models and their output?
- What cultural barriers impede geoscience modelers from making progress on these topics?

The goal of this discussion is to highlight initial findings and selected themes that have emerged from the RCN project. The discussion is not intended to provide prescriptive guidelines for what and how long data should be preserved and shared from simulation based research to fulfill community open science expectations. Instead, we share here initial progress toward guidelines, and, importantly, the broader themes that we have identified as crucial to understand and address in order to reach community open science goals. We plan to share detailed guidance for specific datasets in a future article.

RESEARCH COORDINATION NETWORK PROJECT OVERVIEW

The ultimate goal of the RCN project is to provide guidance on what data and software elements of simulation based research, specifically from dynamical models, need to be preserved and shared to meet community open science expectations, including those of funders and publishers. To achieve this goal, two virtual workshops were held in 2020, and ongoing engagement with selected stakeholders has taken place through professional society based town halls and webinars. Workshop participants included representatives from a variety of communities, including atmospheric, hydrologic, and oceanic sciences, data managers, funders, and publishers.

Project deliverables developed through the workshops and follow-on discussions include: (1) a preliminary rubric that can be used to inform a researcher on what simulation output needs to be preserved and shared in a FAIR aligned community data repository to support replicability of research results, and allow others to easily build upon research findings, (2) draft rubric usage instructions, and (3) an initial set of reference use cases, which are intended to provide researchers with examples of what has been preserved and shared by other projects that attained similar rubric scores. The current version of all project

deliverables can be accessed at https://modeldatarcn.github.io/. After further workshops are held and additional community input is gathered in 2022, project stakeholders plan to refine the project outputs further for later publication.

INITIAL RCN PROJECT FINDINGS

Knowledge Production vs. Data Production

A primary determinant of the data archiving for modeling projects is whether they are oriented toward knowledge or data production. Most scientific research projects are undertaken with the main goal of knowledge production (e.g., running an experiment with the goal of publishing research findings). Other projects are designed and undertaken with the specific goal of data production, that is, they produce data with the intention that those data will be used by others to support knowledge production research. For example, regional and global oceanic and atmospheric reanalysis products produced by numerical weather prediction centers would fall into the category of data production. The importance of this distinction is that different kinds of work are involved in knowledge vs. data production (Baker and Mayernik, 2020, **Figure 1**).

In particular, data production cannot occur without well-planned and funded data curation support, whereas knowledge production-oriented projects can be quite successful at generating new findings with minimal data curation. In some cases, such as the CMIPs, projects are designed for both knowledge and data production.

It is difficult to achieve either knowledge or data production if a project does not have that orientation from the beginning. Projects with a knowledge production orientation may generate significant amounts of data, and may want other scientists to use their outputs. But if data production is not the explicit goal and orientation from the beginning of a project, it is difficult for data to be used by others without direct participation by the initial investigator(s). If preservation and broad sharing of most project-generated data is intended to take place, a data production-orientation is necessary. This must encompass data preparation and curation tasks, such as ensuring that data and metadata conform to standards, that files are structured in consistent formats, that data access and preservation are possible, that data biases and errors are documented, and that data can be accessed and cited via persistent identifiers (McGinnis and Mearns, 2021; Petrie et al., 2021).

Determining What to Preserve and Share

While each project is unique, certain data and software elements should be preserved and shared for all projects to support research replicability and allow researchers to more easily build upon the work of others. Accordingly, workshop participants found that it would be best to preserve and share all elements of the simulation workflow, not just model source code (**Figure 2**). Simply sharing model code doesn't provide the level of understanding needed to easily build upon existing research. Also, if initialization and forcing data are provided by an outside provider, such as a national meteorological center, it should be the responsibility of that center to provide access to those data.

As discussed above, most scientific research projects are focused on knowledge production and as such should be saving little to no raw simulation data in repositories, instead focusing on smaller derived fields that help communicate to future researchers the environmental state or other information important for building similar studies in the future. Particularly for highly non-linear case studies, the goal is not exact reproducibility, but rather enough output to understand the environmental state that forced, and the impacts of, the features being investigated. There may be unique projects in which bitwise reproducibility is deemed necessary; in those cases, containerization can be useful (Hacker et al., 2016). However, to build upon prior research, most knowledge production research does not require bitwise reproducibility. Conversely, as described above, data production projects should have well-structured plans to preserve and share all model outputs needed for downstream users to successfully develop knowledge production research from those outputs.

Need for Curation Support

Development of research data and software that adheres to community best practice expectations for reuse requires specialized knowledge, and can be resource intensive. For example, data management includes a broad spectrum of activities in the data lifecycle, including proposal planning, data collection and organization, metadata development, repository selection, and governance (Wilkinson et al., 2016; Lee and Stvilia, 2017). Model code, output data, and any platforms being used to deliver code and/or software need to be documented clearly to provide guidance for potential users. Research software should be made available through collaborative development platforms such as GitHub (github.com) or Bitbucket (bitbucket.org), versioned, and licensed to describe terms of reuse and access (Lamprecht et al., 2020; American Meteorological Society, 2021). Both the data and snapshots of software versions that were used to support research outcomes should be archived in trusted data (e.g., https://repositoryfinder.datacite.org) and software (e.g., https://zenodo.org, https://figshare.com) repositories for long-term preservation and sharing, and assigned digital object identifiers to facilitate discovery and credit (Data Citation Synthesis Group, 2014; Katz et al., 2021).

The RCN project is working to develop strategies for deciding what needs to be preserved and shared, and communicate those practices clearly to researchers, repositories, and publishers. This should decrease the volume of simulation-related output that needs to be preserved, but conversely there is an expectation for researchers to share simulation configuration, model and processing codes that can reasonably be understood and reused by others with discipline specific knowledge. Researchers are currently spending a significant portion of their own time dealing with data curation; in some cases, over 50% of their funded time. Developing and stewarding software that adheres to community best practice expectations adds an additional burden on the researcher that may take up more of their funded time. Additionally, the availability of community data repositories in selected disciplines, such as the atmospheric sciences, is sparse, making it challenging for researchers to find an appropriate

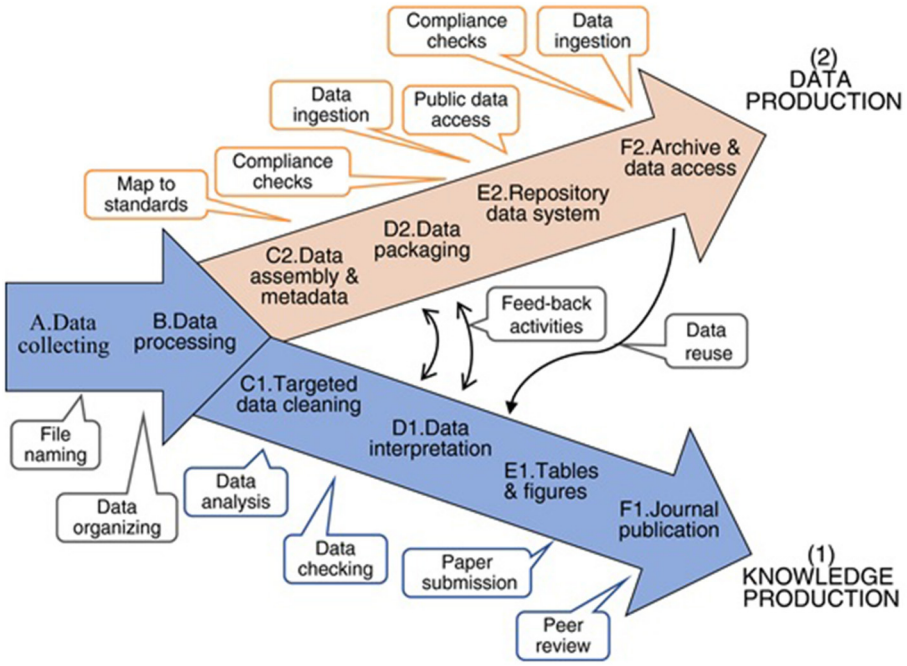


FIGURE 1 | Baker and Mayernik (2020). The two-stream model shows two branches: (1) knowledge production using data optimized for local use with the final form optimized for publication of papers; and (2) data production creates data intended for release to a data repository that makes data accessible for reuse by others. This figure was published *via* the Creative Commons Attribution 4.0 International copyright license (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/.

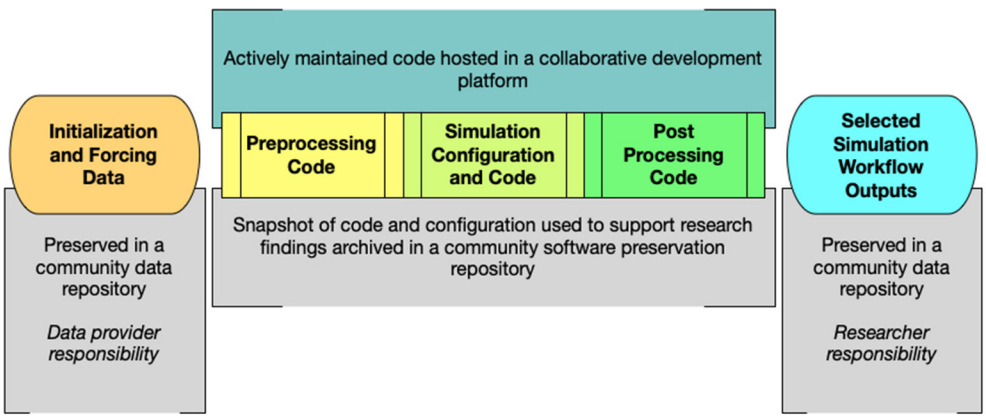


FIGURE 2 | Data and software elements to be preserved and shared by all projects.

repository to deposit their data. A coordinated effort is needed to fund personnel to assist researchers in data and software curation, as well as investment in the needed repository preservation and stewardship services, to complement existing capabilities (Gibeaut, 2016; Mayernik et al., 2018). It is unreasonable to expect already overloaded researchers to become expert data managers and software developers, and find time to complete their research activities.

Cultural Barriers to Progress

As was discussed already, resources (time, money, personnel) remain a significant barrier to implementation of data management best practices that promote increased scientific replicability, reduced time-to-science, and broadened participation. But there is also resistance to change as these practices are often in opposition to the way much of the community has built a successful career. Career advancement for scientists in typical scientific career pathways at research centers and universities is based on long-used metrics of "scientific success." The primary traditional metrics are number of publications, citations, and amount of proposals awarded. Often observational instrument researchers have built careers by leveraging use of their instrument in field campaigns to secure proposal dollars and subsequent publications. Some

model researchers and theorists argue that limiting access to their software is thereby an equivalent path that they take for building their career. However, instrument researchers are only a small subset of observationalists, and many scientists have built a strong career without limiting access to their software or data.

That said, we do recognize existing challenges in sharing of data and software. One challenge is the lack of adoption in formally citing datasets and software in peer reviewed journals, and consideration of such impact measurements in evaluations for promotion. Initiatives to add data and software contributions to the evaluation process for promotion and awards exist at various institutions, but adoption of new practices tends to be arduous. To protect early career scientists and researchers at smaller institutions, we propose using the practices of data and software sharing embargoes and curation waivers. Embargoes on data sharing are often used in field campaigns to give graduate students a certain amount of time to work with the data before sharing more broadly, in recognition of their need to publish on this data as part of their career development (and possibly needing more time to do so than more experienced researchers). This practice should be continued for new data and software to protect early career researchers. Additionally, waivers are often used to reduce requirements (e.g., publication fees) for researchers without sufficient resources. Here, we can extend waivers to curation requirements for researchers at institutions lacking in data/software curation expertise. Embargoes and waivers should not be used as excuses, however, to fall back on "data available upon request" statements that are proven to be problematic (Tedersoo et al., 2021). Overall, for these kinds of considerations, we emphasize not disproportionately punishing researchers with fewer resources.

Other common concerns from scientists about more open access, and particularly to software, are misuse of the software and fear of sharing suboptimal code. Misuse of software is a real outcome, as any open source software may ultimately be misused by some. However, the benefit of a more inclusive user base far outweighs the dangers of misuse (American Meteorological Society, 2021). A significant challenge is often determining who is responsible, if anyone, for user support, as this is rarely documented or formalized within research teams. As for sharing suboptimal code, most researchers in the Earth sciences are not formally trained programmers and many feel that their code is clunky, sometimes embarrassingly so. However, while some documentation is needed, elegant code is not a requirement for success in the earth sciences. In general, the community is accepting of code as long as it gets the correct physical answer. An added benefit of sharing code is that later users may streamline and optimize it, benefiting everyone.

DISCUSSION

We must work as a community to overcome the barriers to open data and software because our current practices impede broadened participation in the Earth sciences. Scientific equity cannot be fully achieved when individual scientists act as gatekeepers for new models, data, and software. However, these new initiatives need to be supported financially, with expertise and infrastructure, and incentivized through modernized merit review criteria (Moher et al., 2018). As discussed above, researchers are already struggling with data curation and code documentation and sharing; the community needs help from researchers trained in these areas (possibly as staff support at shared repositories). Without financial support and infrastructure provided for the scientific community, researchers at smaller institutions will be the hardest hit by these changes, negating the very advances we are trying to achieve in broadening participation. We can mitigate to some degree with embargos and waivers, but in the long run, we need federal commitment to data and software curation services.

Funding agencies are already paying for data work, if indirectly, by adding open data requirements to research grants but not increasing the investment in data infrastructures and data curation expertise. The result has been that scientists and graduate students re-allocate grant funding intended for scientific research to complete data tasks. If open science expectations for simulation-based research are to be achieved, the investment in data work should be more direct and intentional. Investigators spending research grant dollars on minimal curation by untrained graduate students is inefficient and will not lead to the intended outcomes of high-quality data sets being deposited in well-curated data repositories.

Finally, we emphasize that it is important to consider more than just the extremes for many of the questions and topics discussed in this paper. From our project's discussions, it is clear that we must get past the poles of either all or no model output being preserved. The best outcome in most use cases discussed within our project has been somewhere in the middle, namely that some output be preserved, but not all. Likewise, software need not be all open or closed. Some software may be released openly even if other software components are withheld from public view due to security or proprietary concerns. Similarly, questions about curation work should not be limited to a scientist vs. curator debate. Ideally, curation tasks should involve partnerships between scientific and data experts to take advantage of their respective knowledge and skills.

The next steps for our project and for the community broadly will be to address other important questions that have come up in our project activities, but have not been discussed in detail. For example, how long should simulation output be preserved and shared? Needs for data longevity are almost impossible to assess up front, due to the unknown future value and user bases of archived data sets (Baker et al., 2016). Such assessments have to be done downstream. But what are the best measures of a data set's value over time? Ideally this would be based on robust metrics, but there is not yet community agreement on what metrics are most appropriate. The overall goal is to make sure that we are preserving materials that can enable follow-on research, whether that be data, software, or both. More discussion and use cases will be necessary going forward to address these difficult challenges.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

GM was the Principal Investigator of the NSF-funded RCN project. MM and DS are Co-Principal Investigators. All authors contributed to the article and approved the submitted version.

FUNDING

This project was funded by the NSF EarthCube program, NSF Awards 1929757 and 1929773. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

REFERENCES

Alexander, K., and Easterbrook, S. M. (2015). The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geosci. Model Dev.* 8, 1221–1232. doi: 10.5194/gmd-8-1221-2015

American Meteorological Society (2021). *Software Preservation, Stewardship, and Reuse: A Professional Guidance Statement of the American Meteorological Society*. Available online at: https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/software-preservation-stewardship-and-reuse/

Baker, K. S., Duerr, R. E., and Parsons, M. A. (2016). Scientific knowledge mobilization: co-evolution of data products and designated communities. *Int. J. Digital Curat.* 10, 110–135. doi: 10.2218/ijdc.v10i2.346

Baker, K. S., and Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere* 11:3191. doi: 10.1002/ecs2.3191

Bush, R., Dutton, A., Evans, M., Loft, R., and Schmidt, G. A. (2020). Perspectives on data reproducibility and replicability in paleoclimate and climate science. *Harvard Data Sci. Rev.* 2:4. doi: 10.1162/99608f92.00cd8f85

Data Citation Synthesis Group (2014). *Joint Declaration of Data Citation Principles*. San Diego CA: FORCE. doi: 10.25490/a97f-egyk

Easterbrook, S. M. (2014). Open code for open science? *Nat. Geosci.* 7, 779–781. doi: 10.1038/ngeo2283

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2021). *NetCDF Climate and Forecast (CF) Metadata Conventions*. Available online at: https://cfconventions.org/cf-conventions/cf-conventions.html (accessed October 11, 2021).

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

Gibeaut, J. (2016). Enabling data sharing through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). *Oceanography* 29, 33–37. doi: 10.5670/oceanog.2016.59

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., et al. (2013). Documenting climate models and their simulations. *Bull. Am. Meteorol. Soc.* 94, 623–627. doi: 10.1175/BAMS-D-11-00035.1

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philo. Transact. R. Soc. A.* 379:2197. doi: 10.1098/rsta.2020.0210

Hacker, J., Exby, J., Gill, D., Jimenez, I., Maltzahn, C., See, T., et al. (2016). A containerized mesoscale model and analysis toolkit to accelerate classroom learning, collaborative research, and uncertainty quantification. *Bull. Am. Meteorol. Soc.* 98, 1129–1138. doi: 10.1175/BAMS-D-15-00255.1

Heydebreck, D., Kaiser, A., Ganske, A., Kraft, A., Schluenzen, H., and Voss, V. (2020). *The ATMODAT Standard enhances FAIRness of Atmospheric Model Data*. Washington, DC: American Geophysical Union. doi: 10.1002/essoar.10504946.1

Irving, D. (2016). A minimum standard for publishing computational results in the weather and climate sciences. *Bull. Am. Meteorol. Soc.* 97, 1149–1158. doi: 10.1175/bams-d-15-00010.1

Katz, D. S., Chue Hong, N., Clark, T., Muench, A., Stall, S., Bouquin, D., et al. (2021). Recognizing the value of software: a software citation guide [version 2; peer review: 2 approved]. *F1000Research* 9:1257. doi: 10.12688/f1000research.26932.2

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., et al. (2020). Towards FAIR principles for research software. *Data Sci.* 3, 37–59. doi: 10.3233/DS-190026

Lee, D. J., and Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE* 12:e0173987. doi: 10.1371/journal.pone.0173987

Masson, D., and Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.* 38:46864. doi: 10.1029/2011gl046864

Mayernik, M., Schuster, D., Hou, S., and Stossmeister, G. J. (2018). *Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report*. Boulder, CO: National Center for Atmospheric Research. doi: 10.5065/D6NC601B

McGinnis, S., and Mearns, L. (2021). Building a climate service for North America based on the NA-CORDEX data archive. *Climate Serv.* 22:100233. doi: 10.1016/j.cliser.2021.100233

Moher, D., Naudat, F., Cristea, I., Miedema, F., Ioannidis, J., and Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLoS Biol.* 16:e2004089. doi: 10.1371/journal.pbio.2004089

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. doi: 10.17226/25303

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., et al. (2021). Coordinating an operational data distribution network for CMIP6 data. *Geosci. Model Dev.* 14, 629–644. doi: 10.5194/gmd-14-629-2021

Stall, S., Yarmey, L. R., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J., et al. (2018). Advancing FAIR data in Earth, space, and environmental science. *Eos* 99:9301. doi: 10.1029/2018EO109301

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., et al. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* 8:192. doi: 10.1038/s41597-021-00981-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

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

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

frontiers in Climate

PERSPECTIVE
published: 20 December 2021
doi: 10.3389/fclim.2021.732174

Timeline Visualization Uncovers Gaps in Archived Tsunami Water Level Data

Aaron D. Sweeney[1,2]*

[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, United States, [2] National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information, Boulder, CO, United States

We demonstrate that data abstraction via a timeline visualization is highly effective at allowing one to discover patterns in the underlying data. We describe the rapid identification of data gaps in the archival time-series records of deep-ocean pressure and coastal water level observations collected to support the NOAA Tsunami Program and successful measures taken to rescue these data. These data gaps had persisted for years prior to the development of timeline visualizations to represent when data were collected. This approach can be easily extended to all types of time-series data and the author recommends this type of temporal visualization become a routine part of data management, whether one collects data or archives data.

Keywords: data management, data rescue, timeline, tsunamis, visualization, water level

OPEN ACCESS

Edited by:
Kevin Butler,
Environmental Systems Research Institute, United States

Reviewed by:
James O'Donnell,
Environmental Systems Research Institute, United States
Jared Rennie,
National Oceanic and Atmospheric Administration (NOAA), United States

***Correspondence:**
Aaron D. Sweeney
aaron.sweeney@colorado.edu

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *28 June 2021*
Accepted: *30 November 2021*
Published: *20 December 2021*

Citation:
Sweeney AD (2021) Timeline Visualization Uncovers Gaps in Archived Tsunami Water Level Data. Front. Clim. 3:732174. doi: 10.3389/fclim.2021.732174

INTRODUCTION

Timelines are an effective way of visualizing data inventory, leading to the improvement of both data curation and exploration (Kräutli, 2016; Shneiderman et al., 2017). They offer a two-dimensional, graphical representation of history, often with important events annotated. The NOAA National Centers for Environmental Information (NCEI) is the long-term archive (hereafter, referred to as the Archive) for ocean-bottom pressure data (National Oceanic and Atmospheric Administration, 2005) and coastal tide gauge data (Center for Operational Oceanographic Products Services, 2007) collected in support of the NOAA Tsunami Program. The Archive also provides quality control and tidal analysis of these data and maintains authoritative information regarding past tsunamis (including sources) in the NCEI and the collocated World Data Service for Geophysics (WDS) Global Historical Tsunami Database (National Geophysical Data Center/World Data Service, 2018). Roughly speaking, the pressure goes up one atmosphere for every 10 meters of water (Fofonoff and Millard, 1983). When we first started working at what was then the NOAA National Geophysical Data Center in Boulder, CO, we wanted to understand the extent of water level data we were managing. One had access to archive tape listings, maps, and spreadsheets, but one could not make sense of the when and where of the data and any underlying patterns. Inspired by the data visualization work of Robert Aspinall at the NOAA Center for Operational Oceanographic Products and Services (CO-OPS, https://tidesandcurrents.noaa.gov/inventory.html?id=9410230) and our own experience with web development, we began to play with a number of open-source, Javascript libraries to display information about the data. Vis.js (http://visjs.org/) had the best functionality out-of-the-box, had good documentation, and is free and dual-licensed under Apache-2.0 and MIT. The construction of the timeline was much easier once we converted all the data, archived as a mix of tabular data and hierarchical XML, to a common standard, array-based format (netCDF) meeting the interoperable goal of the FAIR

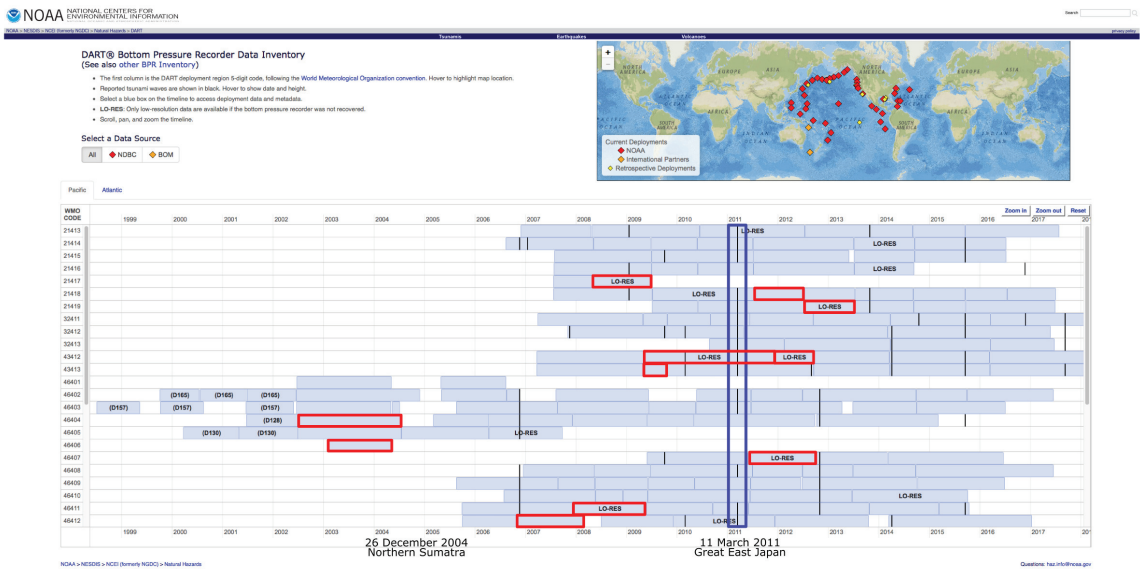


FIGURE 1 | A timeline and map view of DART® ocean bottom pressure data archived at NCEI. As a result of the 26 December 2004 Northern Sumatra earthquake and tsunami, the U.S. DART® network was expanded from 6 to 39 stations by 2008. Superposed on the timeline are the dates of observed tsunami waves. As an example, observations of the 11 March 2011 Great East Japan earthquake and tsunami are circled in dark blue. More than 20 gaps in the Archive were identified by visual inspection of this timeline and filled by NDBC (circled in red, not all visible in this screenshot), including one deployment that observed two tsunamis. High resolution, 15-s data is stored on the seafloor instrument until recovery. If the instrument was not recoverable, lower resolution data (i.e., "LO-RES") reported in real-time is archived instead. See https://www.ngdc.noaa.gov/hazard/dart.

Data Principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016). The construction of time-series netCDF files followed the guidance provided by NCEI at https://www.ncei.noaa.gov/data/oceans/ncei/formats/netcdf/v2.0/index.html. We could then more easily extract and add the temporal bounds of the data—start and end times as well as data gaps—to the timeline. In an era of big data, visualizing and interacting with information about the data, before diving into the details, can leverage the power of human perception and insight (Shneiderman et al., 2017). Visualization reduces the need to formulate, in advance, specific questions about the data. Unforeseen patterns emerge through interaction with and exploration of the visualization. To our surprise, the visualization of the ocean-bottom pressure data and coastal tide gauge data revealed data gaps and patterns.

FINDING AND FILLING GAPS

It is extremely important to have continuous, uninterrupted time-series of water level measurements, even in the absence of tsunami observations. These data are the ground truth for tsunami propagation models. In other words, if the model predicts a tsunami to arrive at a given tide gauge location, but that tide gauge observation showed no tsunami present, then that model needs to be refined and corrected. If we don't have a measurement, we cannot validate the model. We cannot go back and remeasure the past, so every observation counts.

Data are considered at-risk of being lost or, at least "un-FAIR," until they are archived and managed at NCEI. Success in finding and filling gaps depends on data providers being interested and willing collaborators. Our primary data provider for ocean bottom pressure data is the National Data Buoy Center (NDBC), and when the Deep-ocean Assessment and Reporting of Tsunamis (DART®) data inventory timeline (see https://www.ngdc.noaa.gov/hazard/dart and the associated Javascript code at https://www.ngdc.noaa.gov/hazard/js/dart_inventory_timeline.js) went live in 2016, NDBC examined it closely and compared it with their record of seafloor pressure sensor deployments (spreadsheets). Each sensor deployment on the seafloor produces 1–3 years of data and supporting metadata, generally stored in separate files. These data and metadata files from a single deployment are referred to as a "data package." NDBC identified over 20 deployment data packages, going back as far as 12 years, that had not been submitted to NCEI for archive—including one that detected tsunami waves (**Figure 1**). Without the timeline, these data may have been lost forever.

With the coastal tide gauge inventory published (https://www.ngdc.noaa.gov/hazard/tide), a clear pattern of data transfer hiccups to NCEI really stood out: we observed a coincident absence of archived data from the entire network on specific

days. Fortunately, CO-OPS holds onto all of their data for characterizing long-term changes and for defining tidal datums, so filling in the gaps at NCEI was possible. Specifically, we identified gaps in the Archive of 1-min water-level data going back to 2008 and rescued an average of 3 months of data per station across 243 coastal tide gauge stations operated by CO-OPS. We rescued more than 6 months of data from 13 stations—the longest record was 3 years' worth of data from Pago Pago, American Samoa. These rescued data include observations of two tsunamis at Pago Pago and Kwajalein, Marshall Islands.

These backfilled data have been incorporated into our quality-controlled and de-tided products (Mungov et al., 2013). One unexpected benefit of this Archive reconciliation effort was the discovery and correction of a reporting error: the CO-OPS Data API reported two heights for the same time for 20 stations on a select number of dates. We provided details to CO-OPS so they could troubleshoot and resolve this issue. This work directly supports NOAA's data stewardship role by ensuring the ocean environmental record is complete, preserved, and accessible.

IMPROVING QUALITY AND COMPLETENESS

Constructing a timeline from the NCEI/WDS Global Historical Tsunami Database of reported times and heights of maximum

FIGURE 2 | Clicking on one of the blue segments on the timeline (**Figure 1**) takes you to a page like this one. The deployment or station page shows a plot preview of quality-controlled data, clearly showing tidal oscillations that correlate with the dates of the new and full moon (gray and white circles, respectively), and de-tided data. The observation of the 11 March 2011 Great East Japan tsunami is circled in black.

tsunami waves and the inventory of water level data can help us fill in our record of reported tsunami waves. We can immediately see where adjacent stations observed tsunami waves and begin to take a look at stations where no report was made, but where a closer examination of the quality-controlled and de-tided water level data may reveal additional tsunami detections.

The NCEI/WDS Global Historical Tsunami Database consists of dates and heights as reported in real-time by the National Tsunami Warning Center (NTWC) or the Pacific Tsunami Warning Center (PTWC). The centers are selective about which stations they typically examine, and they don't always report all stations, especially if an event was small. With our timeline, we can zero in on times and locations when and where a tsunami ought to have been detected. Because we have access to the digital time-series data, we then have an opportunity to fill in the database with additional reports.

TOWARD EVENT-DRIVEN DISCOVERY AND ACCESS

Timeline visualizations promote event-driven data exploration and discovery. The most difficult aspect for a data curator is deciding which events and other sources of information to include and which sources are authoritative. During a tsunami event, NTWC and PTWC provide products listing the times and maximum heights of tsunami waves at select stations in their respective areas of operation. These are superposed on the DART® and tide gauge data inventory timelines to draw attention to events that may be of interest. Station pages (**Figure 2**) are linked from the timelines and provide access to raw data, quality-controlled and de-tided data products, modeled tidal constituents, summary time-series plots, and supporting metadata. In the future, we plan to enhance the station pages with interactive plots. We also hope to archive and redistribute data from international partners, with appropriate agreements in place.

These products support the operational forecasting efforts of the Tsunami Warning Centers and the research efforts of the NOAA Center for Tsunami Research (NCTR) under the auspices of the NOAA Tsunami Program, as well as the broader tsunami modeling community. If you have data you would like us to archive and re-distribute, please contact us at haz.info@noaa.gov.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://doi.org/10.7289/V59884XF, https://doi.org/10.7289/V5F18WNS, https://doi.org/10.7289/V5PN93H7.

AUTHOR CONTRIBUTIONS

AS was responsible for the research and development of the timelines, including data abstraction, and writing the manuscript.

FUNDING

This work was supported by the NOAA Cooperative Agreement with CIRES, NA17OAR4320101. The Archive was supported by the NOAA National Weather Service and the NOAA National Environmental Satellite, Data, and Information Service. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

ACKNOWLEDGMENTS

A special thank to Dr. George Mungov for providing water level quality-control and tidal analysis and to Nicolas Arcos and Paula Dunbar for providing quality assurance of historical tsunami information. Main data providers include the National Data Buoy Center, the Center for Operational Oceanographic Products and Services, the National Tsunami Warning Center, the Pacific Tsunami Warning Center, and the Pacific Marine Environmental Laboratory. We are grateful to Richard Bouchard at NDBC for identifying and submitting DART® packages not previously archived, to Heather McCullough at NCEI for web page design review and suggestions, to John Cartwright at NCEI for mapping library considerations, and to Dr. Danielle Szafir at CU Boulder for discussion of user interfaces and design studies.

REFERENCES

Center for Operational Oceanographic Products and Services (2007). *CO-OPS 1-Minute Tsunami Water Level Data*. NOAA National Centers for Environmental Information.

Fofonoff, N. P., and Millard, R. C. (1983). *Algorithms for Computation of Fundamental Properties of Seawater. No. 44*. Paris: UNESCO Technical Papers in Marine Sciences.

Kräutli, F. (2016). *Visualising Cultural Data: Exploring Digital Collections Through Timeline Visualisations.* (Dissertation), Royal College of Art. ProQuest Dissertations Publishing, London (United Kingdom).

Mungov, G., Eblé, M., and Bouchard, R. (2013). DART® tsunameter retrospective and real-time data: a reflection on 10 years of processing in support of tsunami research and operations. *Pure Appl. Geophys.* 170, 1369–1384. doi: 10.1007/s00024-012-0477-5

National Geophysical Data Center/World Data Service (2018). *NCEI/WDS Global Historical Tsunami Database*. NOAA National Centers for Environmental Information. doi: 10.7289/V5PN93H7

National Oceanic and Atmospheric Administration (2005). *Deep-Ocean Assessment and Reporting of Tsunamis (DART®)*. NOAA National Centers for Environmental Information. doi: 10.7289/V5F18WNS

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2017). *Defining the User Interface: Strategies for Effective Human-Computer Interaction, 6th Edn*. Boston, MA: Pearson.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

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

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

frontiers in Climate

TECHNOLOGY AND CODE
published: 10 February 2022
doi: 10.3389/fclim.2021.782909

Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production

Charles Stern[1], *Ryan Abernathey*[1]*, *Joseph Hamman*[2,3], *Rachel Wegener*[4], *Chiara Lepore*[1], *Sean Harkins*[5] *and Alexander Merose*[6]

[1] Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, United States, [2] CarbonPlan, San Francisco, CA, United States, [3] National Center for Atmospheric Research, Boulder, CO, United States, [4] Department of Atmospheric and Ocean Science, University of Maryland, College Park, MD, United States, [5] Development Seed, Washington, DC, United States, [6] Google Research, Mountain View, CA, United States

Pangeo Forge is a new community-driven platform that accelerates science by providing high-level recipe frameworks alongside cloud compute infrastructure for extracting data from provider archives, transforming it into analysis-ready, cloud-optimized (ARCO) data stores, and providing a human- and machine-readable catalog for browsing and loading. In abstracting the scientific domain logic of data recipes from cloud infrastructure concerns, Pangeo Forge aims to open a door for a broader community of scientists to participate in ARCO data production. A wholly open-source platform composed of multiple modular components, Pangeo Forge presents a foundation for the practice of reproducible, cloud-native, big-data ocean, weather, and climate science without relying on proprietary or cloud-vendor-specific tooling.

Keywords: data, community, cloud, ARCO, NetCDF, Zarr, Python

OPEN ACCESS

Edited by:
Michael C. Kruk, National Oceanic and Atmospheric Administration (NOAA), United States

Reviewed by:
Mark Capece, General Dynamics Information Technology, Inc., United States
Micah James Wengren, National Oceanic and Atmospheric Administration (NOAA), United States

***Correspondence:**
Ryan Abernathey
rpa@ldeo.columbia.edu

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *25 September 2021*
Accepted: *30 November 2021*
Published: *10 February 2022*

Citation:
Stern C, Abernathey R, Hamman J, Wegener R, Lepore C, Harkins S and Merose A (2022) Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production. Front. Clim. 3:782909. doi: 10.3389/fclim.2021.782909

1. INTRODUCTION

In the past 10 years, we have witnessed a rapid transformation in environmental data access and analysis. The old paradigm, which we refer to as the *download model*, was to search for files from a range of different data providers, download them to a local laptop or workstation, and analyze the data in a traditional desktop-based analysis environment (e.g., IDL, MATLAB, and ArcGIS). The new paradigm, which we call *data-proximate computing*, instead brings compute resources adjacent to the data, with users performing their data analysis in a web browser and retrieving data on demand via APIs or HTTP calls (Ramamurthy, 2017). Data-proximate environmental data analysis tools and platforms are often deployed in the commercial cloud, which provides scalable, on-demand computing and high-throughput data access, but are not necessarily limited to cloud environments. Data-proximate computing removes the burden on the data user to provide local computing; this has the potential to massively expand access to environmental data, empowering communities that have been historically marginalized and lack such local computing resources (Gentemann et al., 2021). However, this democratization is not guaranteed. FAIR data, open standards, and equitable access to resources must be actively pursued by the community (Wilkinson et al., 2016; Stall et al., 2019).

Many different platforms exist to analyze environmental data in the cloud; e.g., Google Earth Engine (GEE) and Microsoft's Planetary Computer (Gorelick et al., 2017; Microsoft, 2021). A common need for all such platforms is access to analysis-ready, cloud optimized (ARCO) data. While a range of powerful ARCO data formats exist (e.g., Cloud Optimized GeoTIFF, Zarr, TileDB Embedded, and Parquet), ARCO data production has remained a bespoke, labor-intensive process.

Recent sessions devoted to cloud computing at meetings of the American Geophysical Union (AGU) and Earth System Information Partners (ESIP) enumerated the considerable toil involved in creating ARCO data in the cloud (Hua et al., 2020; Quinn et al., 2020). For example, when GEE partnered with the European Center for Medium-Range Weather Forecasting (ECMWF) to bring a portion of the ERA5 reanalysis data to GEE, the data ingestion process was incredibly time and resource intensive, spanning 9 months and involving a suite of specialized tools (Wagemann, 2020).

In addition to demanding computing resources and specialized software, ARCO data production also requires knowledge in a range of areas, including: legacy and ARCO data formats, metadata standards, cloud computing APIs, distributed computing frameworks, and domain-specific knowledge sufficient to perform quality control on a particular dataset. In our experience, the number of individuals with this combination of experience is very small, limiting the rate of ARCO data production overall.

This paper describes Pangeo Forge, a new platform for the production of ARCO data (Pangeo Forge Community, 2021). A central goal of Pangeo Forge is to reduce the toil associated with downloading, cleaning, and preparing data for analysis, particularly for the large, complex datasets associated with high-bandwidth observing systems, Earth-system simulations, and weather reanalyses. Recognizing that individuals with domain-specific data knowledge are not necessarily experts in cloud computing or distributed data processing, Pangeo Forge aims to lower the barrier for these scientists to contribute to ARCO data curation. Finally, we hope to build a platform that encourages open and inclusive participation, crowdsourcing ARCO data production from the diverse community of environmental data specialists across the world, for the mutual benefit of all.

At the time of writing, Pangeo Forge is still a work on progress. This paper describes the motivation and inspiration for building the platform (section 2) and reviews its technical design and implementation (section 3). We then describe some example datasets that have been produced with Pangeo Forge (section 4) and conclude with the future outlook for the platform (section 5).

2. MOTIVATION AND INSPIRATION

2.1. Analysis-Ready, Cloud-Optimized Data

In the context of geospatial imagery, remote sensing instruments collect raw data which typically requires preprocessing, including color correction and orthorectification, before being used for analysis. The term analysis-ready data (ARD) emerged originally in this domain, to refer to a temporal stack of satellite images depicting a specific spatial extent and delivered to the end-user or customer with these preprocessing steps applied (Dwyer et al., 2018; Holmes, 2018). In the context of this paper, however, we use the term "analysis-ready" more generally to refer to any dataset that has been preprocessed such that it fulfills the quality standards required by the analysis which will be performed on it. This may include merging and alignment of many individual source files or file-like objects into a single cohesive entity. For remotely sensed measurements, it may involve signal processing to correct for known atmospheric or other distortions. For synthetic (i.e., simulation) data, quality control may include ensuring that output values fall within test parameters defined by the model developers, as well as homogenization of metadata across simulation ensembles.

Analysis-ready data is not necessarily or always cloud-optimized. One way of understanding this is to observe that just because an algorithm *can* be applied to a given dataset, that fact alone does not guarantee the algorithm will execute expediently or efficiently. In a context where even efficient algorithms can take hours or days to run, optimization matters. Computational performance is affected by many factors including algorithm design and hardware specifications, but in the case of big data analytics, the rate-limiting aspect of the system is often I/O throughput, i.e., the rate at which bytes can be read into the algorithm from the data storage location (Abernathey et al., 2021). This rate is itself influenced by variables such as network bandwidth, hardware characteristics, and data format. When we refer to "cloud-optimized" data it is this third variable, format, which we are most concerned with. Cloud-optimized data formats are unique insofar as they support direct access to data subsets without the computational overhead of opening and navigating through a massive data object simply to retrieve a small subset of bytes within it. Implementations of this functionality vary according to the specific cloud-optimized format: some formats include a metadata header which maps byte-ranges within a single large data object, while others opt to split a large object up into many small blocks stored in an organized hierarchical structure. Regardless of the specific implementation, the end result is an interface whereby algorithms can efficiently access data subsets. Efficient access to data subsets is especially impactful in the context of cloud object storage, where simultaneous read/write of arbitrary numbers of data subsets does not decrease the throughput to any individual subset. As such, parallel I/O dramatically increases cumulative throughput.

Analysis-ready, cloud-optimized datasets are, therefore, datasets which have undergone the preprocessing required to fulfill the quality standards of a particular analytic task and which are also stored in formats that allow efficient, direct access to data subsets.

2.2. Open Science, Open Source

The Pangeo Forge codebase, which is written in Python, is entirely open source, as are its Python dependencies including packages such as NumPy, Xarray, Dask, Filesystem Spec, and Zarr (Dask Development Team, 2016; Hoyer and Hamman, 2017; Harris et al., 2020; Durant, 2021; Miles et al., 2021). We see open source software as a scientific imperative. Production of ARCO datasets involves considerable preprocessing and reformatting. Data corruptions can easily be introduced at any step of these multi-stage transformations, either due to user error or, less commonly but more consequentially, due to bugs in the software packages used to perform the ARCO transformation. In an open source context, the scientific user community can readily introspect every step of the process, building trust in its effectiveness as well as contributing to

its robustness by identifying bugs when they arise. The core scientific tenet of reproducibility is also served by open source: the exact provenance of each byte of data that passes through Pangeo Forge is entirely transparent, traceable, and recreatable.

Where Pangeo Forge must unavoidably rely on commercial technology providers, we strive always to uphold the user's Right to Replicate (2i2c.org, 2021). In practice, this means that even if an underlying cloud provider technology is closed source, the application code defining our particular implementation of that technology is always open source, allowing anyone the option to replicate our system exactly as we've deployed it. Version control hosting, continuous integration, compute infrastructure, storage resources, and workflow automation are arenas in which commercial solutions are implemented. The former two services are provided through GitHub repositories and GitHub Actions, respectively, and the latter three through the "big three" cloud service providers (Google Cloud, Amazon Web Services, Microsoft Azure) and Prefect, a dataflow automation provider.

2.3. Crowdsourcing Complexity: the Conda Forge Model

The incredible diversity of environmental science datasets and use cases means that a fully generalized and automatic approach for transforming archival data into ARCO stores is likely neither achievable nor desirable. Depending on the analysis being performed, for example, two users may want the same archival source data in ARCO form, but with different chunking strategies (Chunking, i.e., the internal arrangement of a dataset's bytes, is often adjusted to optimize for different analytical tasks). Transforming just a single dataset from its archival source into an ARCO data store is an incredibly complex task which unavoidably requires human expertise to ensure the result is fit for the intended scientific purpose. Fantasies of cookie-cutter algorithms automatically performing these transformations without human calibration are quickly dispelled by the realities of just how unruly archival data often are, and how purpose-built the ARCO data stores created from them must be. As with all of science, ARCO transformations require human interpretation and judgement.

The necessity of human participation, combined with the exponentially increasing volumes of data being archived, means that ARCO data production is more work than any individual lab, institution, or even federation of institutions could ever aspire to manage in a top-down manner. Any effort to truly address the present scarcity of high-quality ARCO data must by necessity be a grassroots undertaking by the international community of scientists, analysts, and engineers who struggle with these problems on a daily basis.

The software packaging utility Conda Forge, from which Pangeo Forge draws both inspiration and its name, provides a successful example of solving a similar problem via crowdsourcing (Conda-Forge Community, 2015). Conda Forge emerged in 2015 in response to frustrations scientific software users consistently faced when attempting to install system package dependencies in the course of their research. Just like ARCO data production, installing open source software packages with binary dependencies is frequently a multi-step process involving an intricate sequence of software compilation.

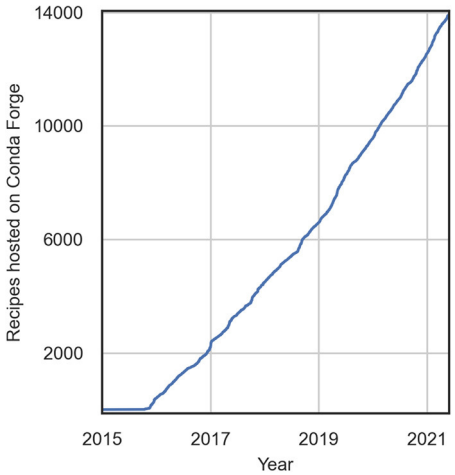


FIGURE 1 | Number of software installation recipes hosted on Conda Forge by year.

If any one step is completed out of order, or perhaps if one of the sub-packages installed is of the wrong version, the end result will be non-functional. This struggle devoured countless years worth of human effort on the part of researchers who required a specific software configuration to pursue their investigations.

Conda Forge introduced the simple yet revolutionary notion that two people, let alone hundreds or thousands, should not be duplicating effort to accomplish the same tedious tasks. As an alternative to that toil, Conda Forge established a publicly-licensed and freely-accessible storehouse, hosted on the open internet, to hold blueprints for performing these arcane yet essential engineering feats. It also defined a process for contributing blueprints to that storehouse and established a build system compatible with the Conda package manager, a component of the open-source Anaconda Software Distribution, itself a popular collection of data science tooling (Anaconda Inc., 2021). This interconnection with the Conda package manager, in addition to serving as the inspiration for Conda Forge's name, means that a given Conda Forge package can be built from the public storehouse onto a community member's system with just a one-line command: `conda install`.

It is not an understatement to say that this simple invocation, `conda install`, and the system built by Anaconda undergirding it, fundamentally transformed for the better the practice of computational science with open source software. The crowdsourcing model defined by Conda Forge then leveraged this technology to maximal advantage for the open source scientific community. For evidence of this fact, we need look no further than the incredible growth rate of community contributed "recipes" (as these installation blueprints are known) in the Conda Forge storehouse (**Figure 1**). The summed impact of this solution totals untold numbers of reclaimed hours which are now dedicated to scientific research itself, rather than tinkering with finicky engineering issues.

In the case of Conda Forge, community members contribute recipes to a public storehouse which define steps for building software dependencies. Then they, along with anyone else, can

avoid ever needing to revisit the toil and time of manually building that specific piece of software again. Contributions to Conda Forge, while they often include executable software components, consist minimally of a single metadata file, named `meta.yaml`, which conforms to a specification established in accordance with the build system. This design is explicitly copied in Pangeo Forge.

3. TECHNICAL DESIGN AND IMPLEMENTATION OF PANGEO FORGE

Pangeo Forge follows an agile development model, characterized by rapid iteration, frequent releases, and continuous feedback from users. As such, implementation details will likely change over time. The following describes the system at the time of publication.

At the highest level, Pangeo Forge consists of three primary components:

- `pangeo-forge-recipes`: A standalone Python package which provides a data model ("recipes") and scalable algorithms for ARCO data production. This package can be used by itself, without the platform's cloud automation tools.
- An automation system which executes recipes using distributed processing in the cloud.
- A catalog which exposes the ARCO data to end users.

3.1. Recipes: Object-Oriented Extraction, Optimization, and Storage (EOS) Algorithms

Inspired directly by Conda Forge, Pangeo Forge defines the concept of a recipe, which specifies the logic for transforming a specific data archive into an ARCO data store. All contributions to Pangeo Forge must include an executable Python module, named `recipe.py` or similar, in which the data transformation logic is embedded (**Figure 2**). The recipe contributor is expected to use one of a predefined set of template algorithms defined by Pangeo Forge. Each of these templated algorithms is designed to transform data of a particular source type into a corresponding ARCO format, and requires only that the contributor populate the template with information unique to their specific data transformation, including the location of the source files and the way in which they should be aligned in the resulting ARCO data store.

Pangeo Forge implements template algorithms with object-oriented programming (OOP), the predominant style of software design employed in Python software packages. In this style, generic concepts are represented as abstract *classes* which gain meaning once *instantiated* with details relevant to a particular use case. Once instantiated, class instances (as they are known) can perform operations on or with the attributes (i.e., details) they ve been given. In Pangeo Forge, the operations embedded in the template algorithms are, broadly speaking, those of data extraction, optimization, storage (EOS). First, data is extracted from a traditional source file server, most commonly via HTTP or FTP request; next, the source data is transformed into an ARCO format; and finally, the data is deposited into cloud object storage.

Within a given class of these EOS algorithms, it's possible to largely generalize the esoteric transformation logic itself, while leaving the specific attributes, such as source file location and alignment criteria, up to the recipe contributor to fill in. The completed `recipe.py` module containing a specific instance of the generic EOS algorithm can then be executed in one of a number of ways. While recipe developers are certainly free to run these open source algorithms on private compute clusters, they are strongly encouraged to submit their recipes to be run on Pangeo Forge's shared infrastructure, which has the dual benefit of being a freely accessible resource and, perhaps even more importantly, results in the ARCO data being written to a publicly-accessible cloud storage bucket and added to the Pangeo Forge catalog for discovery and shared use by the global community. It is through scaling contributions to our public ARCO data catalog that Pangeo Forge aspires to do for ARCO data production what Conda Forge has already accomplished for software dependency management.

3.2. Base Abstractions: Insulating Scientific Domain Expertise From Cloud Automation Concerns

Pangeo Forge consists of multiple interrelated, modular components. Each of these components, such as the recipes described above, consists of some abstracted notions about how a given aspect of the system typically functions. These abstractions are for the most part implemented as Python classes. They include classes related to source file location, organization, and access requirements; the recipe classes themselves; classes which define storage targets (both for depositing the eventual ARCO data store, as well as for intermediate caching); and multiple different models according to which the algorithms themselves can be executed.

The boundaries between these abstraction categories have been carefully considered with the aim of insulating scientific domain expertise (i.e., of the recipe contributor) from the equally rigorous yet wholly distinct arena of distributed computing and cloud automation. Among ocean, weather, and climate scientists today, Python is a common skill, but the ability to script advanced data analyses by no means guarantees an equivalent fluency in cloud infrastructure deployments, storage interfaces, and workflow engines. Moreover, Pangeo Forge aims to transform entire global datasets, the size of which is often measured in terabytes or petabytes. This scale introduces additional technical challenges and tools which are more specialized than the skills required to convert a small subset of data.

By abstracting data sourcing and quality control (i.e., the recipe domain) from cloud deployment and workflow concerns, Pangeo Forge allows recipe contributors to focus exclusively on defining source file information along with setting parameters for one of the predefined recipe classes. Recipe contributors are, importantly, *not* expected to understand or manipulate the storage and execution aspects of the system, which are maintained by community members with expertise in those areas. In what follows, well examine four aspects of the system in closer detail.

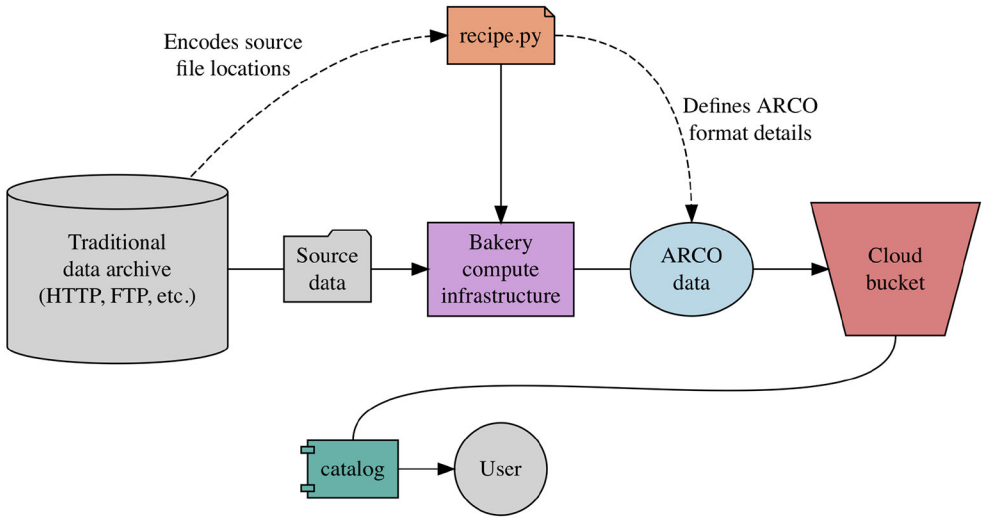


FIGURE 2 | A recipe in relation to Pangeo Forge architecture.

3.2.1. Source File Patterns

In Pangeo Forge, all data transformations begin with a `FilePattern`. This Python class encodes information about archival source files including their location, access requirements, and alignment criteria. Data providers such as NASA and NOAA commonly distribute source files over HTTP. File Transfer Protocol (FTP) is also a common means for distribution of source data in the earth and atmospheric sciences. In either case, contributors specify the access URLs for their source files as part of a `FilePattern`. If the archival data URLs correspond to a dynamic API such as OPeNDAP (Cornillon et al., 2009; Hankin et al., 2010), rather than a static file server, that information is specified at this stage. In cases where authorization credentials such as a password or API token are required to access the source data, the names of environment variables which will point to these values at runtime are included here as well.

```
from pangeo_forge_recipes.patterns import (
    ConcatDim,
    FilePattern,
    MergeDim,
)

def make_full_path(variable, time):
    url_base = "http://data-provider.org/data"
    return f"{url_base}/{variable}_{time}.nc"

merge_dim = MergeDim(
    "variable", ["temperature", "humidity"],
)
concat_dim = ConcatDim("time", list(range(1, 11)))
pattern = FilePattern(
    make_full_path, merge_dim, concat_dim,
)
```

Listing 1 | Defining a source file pattern with alignment criteria.

Almost all ARCO datasets are assembled from many source files which are typically divided by data providers according to temporal, spatial, and/or variable extents. In addition to defining the location(s) of the source files, the `FilePattern` is where contributors define how the specified set of source files should be aligned to create a single cohesive ARCO dataset. Alignment operations include concatenation, for arranging files end-to-end; and merging, for layering files which cover the same spatial or temporal extent, but for different variables. **Listing 1** demonstrates how a recipe contributor would define a `FilePattern` for archival data accessed via the imaginary file server `http://data-provider.org/`. The `pattern` defined in the final line of this snippet encodes not just the location of the source files, but also the fact that any resulting ARCO data store should concatenate these files in the time dimension, and merge them in the variable dimension. This encoding relies on the near-universal practice among data providers of defining URL naming schemes which are descriptive of a given file server's contents; i.e., the access endpoint for a file covering specific extents will name those extents as part of its URL. The objects `merge_dim` and `concat_dim`, in the example provided in **Listing 2**, map our imaginary file server's URL character string representation of dataset dimensions onto Pangeo Forge internal datatypes for consumption by downstream recipe classes.

3.2.2. Recipe Classes

Ocean, climate, and weather data is archived in a wide range of formats. The core abstractions of Pangeo Forge, including `FilePattern`, are designed to be agnostic to data formats, and can be leveraged to transform any archival source file format into any corresponding ARCO format. The transformation from a specific archival format (or category of formats) into a corresponding ARCO format does require a dedicated algorithm, however. In Pangeo Forge, recipe classes are the modular template algorithms which perform a specific category of ARCO transformation. As modular components, an arbitrary number

of these classes can be added to the platform over time, with each new class adding support for a new type of ARCO data production.

As of the writing of this paper, Pangeo Forge defines two such recipe classes, `XarrayZarrRecipe` and `HDFReferenceRecipe`, each of which is most commonly used to transform one or many NetCDF files into a single consolidated Zarr dataset. The difference between these algorithms lies in the nature of their outputs. Whereas, `XarrayZarrRecipe` creates an actual Zarr store by mirroring the source file bytes into a new format, `HDFReferenceRecipe` leverages the Python library `kerchunk` to write lightweight metadata files which map the location of bytes within the archival source files, allowing users to read the original data in a cloud-optimized manner with the Zarr library, but without duplicating bytes (Durant et al., 2021).

```
from pangeo_forge_recipes.recipes import (
  XarrayZarrRecipe
)

recipe = XarrayZarrRecipe(pattern)
```

Listing 2 | Instantiating a recipe algorithm with a source file pattern.

As an algorithm case study, we ll take a closer look at the internals of the `XarrayZarrRecipe`. To begin, let's consider how we would create an instance of this algorithm. In the simplest case each algorithm instance requires only a `FilePattern` instance as input. Using the instance we defined in **Listing 1**, we define a recipe as shown in **Listing 2**. In just these few simple lines, we have created an algorithm containing all of the information needed to extract data from our specified provider archive and transform it into the cloud-optimized Zarr format.

Real-world use cases will likely necessitate additional options be specified for the `XarrayZarrRecipe` instance. Pangeo Forge supports many such options. One worth highlighting is the `target_chunks` option, which is used to indicate the desired chunking scheme of the resulting ARCO data store. As mentioned in section 2.3, chunking, the internal subsetting of a large dataset, is often optimized for a particular analytical aim, with a classic example being the divergent chunking required for optimizing timeseries vs. spatial analyses. Contributors pass a mapping of a dimension name to an integer value to specify their desired chunking; e.g., `target_chunks={"time": 10}` tells the algorithm to divide the ARCO dataset into chunks of length 10 in the time dimension. Should downstream data users require a variation on this or another contributor-defined option, they can make or request changes to the recipe and release those changes as a new dataset version (see section 3.3.2 for further discussion of dataset versioning).

A full treatment of the Zarr specification is beyond the scope of this paper, but a brief overview will provide a better context for understanding. In a Zarr store, compressed chunks of data are stored as individual objects within a hierarchy that includes a single, consolidated JSON metadata file. In actuality, cloud object stores do not implement files and folders, but in a colloquial sense we can imagine a Zarr store as a directory containing a

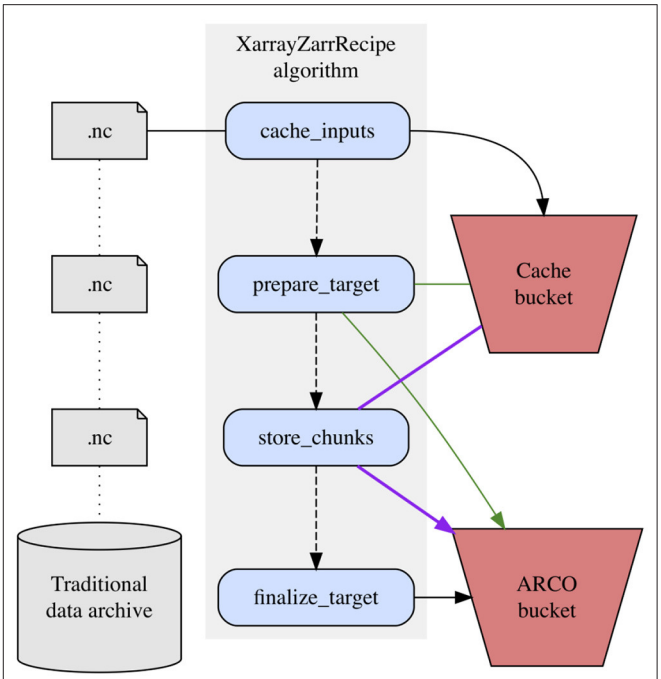


FIGURE 3 | XarrayZarrRecipe algorithm.

single metadata file alongside arbitrary numbers of data files, each of which contains a chunk of the overall dataset (Miles et al., 2021). The `XarrayZarrRecipe` algorithm which transforms archival data into this format consists of four sequential steps, each of which performs a series of sub-operations. Depending on the specific use case, one or more of these steps may be omitted, but we will consider them here for the scenario in which they are all performed (**Figure 3**).

Caching input files is the first step of the `XarrayZarrRecipe` algorithm. This step copies all archival files required for the dataset into temporary storage in a cloud storage bucket. This affords downstream steps of the algorithm fast, parallelizable access to the source data. Typically, the cached source files will be in NetCDF format (Rew et al., 2006). As the name of the algorithm suggests, however, the actual requirement is not for NetCDF inputs specifically, but rather for input files compatible with Xarray, a widely-used Python interface for labeled multidimensional arrays that supports multiple backend file formats, including GRIB, COG, and some flavors of HDF5 (Hoyer and Hamman, 2017).

Before any actual bytes are written to the Zarr store, the target storage location must first be initialized with the skeletal structure of the ARCO dataset. We refer to this step, which immediately follows caching, as `prepare_target`. Preparing the target entails reading metadata from a representative subset of the source files to establish an empty Zarr store of the proper dimensions at the target location.

Once this framework has been established, the algorithm moves on to actually copying bytes from the source data into the Zarr store, via the `store_chunks` task. Internally, this step performs a lot of heavy lifting, insofar as it determines

which specific byte ranges within which source files are required to build each output chunk. Because both the cached source bytes and target dataset reside on cloud object storage, which supports scalable parallel reads and writes, this computationally intensive step is designed to be executed in parallel; specifically, each `store_chunks` task can be executed in any order, without communication or synchronization needed between processes. Parallelization of this step is essential to Pangeo Forge's performance, given that ARCO datasets are often hundreds of gigabytes in size on the low end, and can easily reach multi-petabyte scale.

Following the mirroring of all source bytes into their corresponding Zarr chunks, the `XarrayZarrRecipe` algorithm concludes with a finalization step which consolidates the dataset's metadata into a single lightweight JSON object.

Duplicating bytes is a costly undertaking, both computationally, and because cloud storage on the order of terabytes is not inexpensive. This is a primary reason why sharing these ARCO datasets via publicly accessible cloud buckets is so imperative: a single copy per cloud region or multi-region zone can serve hundreds or thousands of scientists. A clear advantage of the `HDFReferenceRecipe` algorithm is that it does not require byte duplication, however it has certain limitations. This approach requires that the data provider's server support random access to source file subsets, a common but non-universal feature of HTTP and FTP servers. Because the bytes on the data provider's server are not duplicated, use of `HDFReferenceRecipe` precludes forms of data preprocessing which modify the data itself; only metadata preprocessing is supported. Finally, opening data stores created by this algorithm requires the Python package `kerchunk`, effectively preventing access from languages other than Python, as of writing. The interface specification for virtual Zarr stores is clearly defined in the `kerchunk` documentation, therefore we anticipate it may be implemented in other languages in the future (Durant et al., 2021). `HDFReferenceRecipe` presents a remarkably efficient pathway for certain use cases, however as with most efficiencies, it comes with inevitable tradeoffs.

Pangeo Forge's initial algorithms produce Zarr outputs because this format is well-suited to ARCO representation of the gridded multidimensional array data that our early scientific adopters use in their research. Disadvantages of Zarr include the fact that popular data science programming languages such as R do not yet have an interface for the format (Durbin et al., 2020). As our community grows, we anticipate future recipe implementations to include support for most common ARCO formats. These include TileDB Embedded, for multidimensional arrays; Cloud Optimized GeoTIFF (COG), which is widely used in the geospatial imagery community; Parquet, for optimized tabular data stores; and the recently announced Cloud Optimized Point Cloud (COPC) format for, among other uses, light detection and ranging (LiDAR) measurements (Holmes, 2021; Le Dem and Blue, 2021; Hobu, Inc., 2021; TileDB, Inc., 2021). As with our Zarr algorithms, which depend on Xarray as an I/O interface, our path to implementing these algorithms will build on the standard Python interfaces for each data structure; e.g., Rasterio for raster data and Pandas for tabular data (Gillies et al., 2013; Pandas Development Team, 2021).

3.2.3. Storage Abstractions

In the discussion of source file patterns, above, we referred to the fact that input data may be arbitrarily sourced from a variety of different server protocols. The backend file transfer interface which enables this flexibility is the Python package Filesystem Spec, which provides a uniform API for interfacing with a wide range of storage backends (Durant, 2021). This same package provides the engine behind our storage abstractions, a set of modular components which handle various permutations of file caching, reading, and writing. These classes need not be enumerated here; the interested reader can find details about them in the Pangeo Forge documentation. One aspect of these components worth highlighting, however, is that even though cloud object storage is the typical destination of datasets processed by Pangeo Forge, the platform is just as easily able to read from and write to a local POSIX file system or, for that matter, any Filesystem Spec-compatible storage location. Among other things, this capability allows recipe contributors to experiment with recipe algorithms by writing ARCO dataset subsets to local disk during the development process. For our typical cloud storage interfaces, the Filesystem Spec implementations we employ most frequently are `s3fs` (for Amazon Web Services S3), `gcsfs` (for Google Cloud Storage), and `adlfs` (for Azure Datalake and Azure Blob Storage).

3.2.4. Execution Modes

Instantiating a recipe class does not by itself result in any data transformation actually occurring; it merely specifies the steps required to produce an ARCO dataset. In order to actually perform this workflow, the recipe must be executed. A central goal of the software design of `pangeo-forge-recipes` is to be as flexible as possible regarding the execution framework. A wide range of frameworks for parallel and/or distributed computing exist, and `pangeo-forge-recipes` seeks to be compatible with as many of these as possible. For example, high-performance computing (HPC) users may prefer to use traditional job-queue based execution, while cloud users may want to use Kubernetes (Brewer, 2015).

`pangeo-forge-recipes` does not directly implement any parallel computing. Rather, the library has the ability to compile recipes into several different formats used by common distributed computing frameworks. As of writing, we currently support four different flavors of compilation:

- **Compilation to a single Python function:** This is a reference implementation for serial execution.
- **Compilation to Dask Delayed graph**: Dask is a general purpose parallel computing framework widely used in the scientific Python world (Dask Development Team, 2016). By compiling recipes to Dask graphs, `pangeo-forge-recipes` users are able to leverage the variety of different schedulers Dask has implemented for a wide range of different computing platforms. These include `dask-jobqueue` for HPC systems using PBS, SLURM,

SGE, etc. (Henderson, 1995; Gentzsch, 2001; Yoo et al., 2003); Dask Kubernetes for cloud; and Dask-Yarn for Hadoop (Shvachko et al., 2010). Dask's single machine schedulers enable recipes to be executed in parallel using threads or processes on a single large server.

- **Compilation to Prefect Flow**: Prefect is a suite of workflow automation tools encompassing both open source and software-as-a-service (SaaS) components: Prefect Core is an open source workflow engine for Python; a Prefect Flow is a set of interrelated individual tasks, structured in a graph; and Prefect Cloud is a SaaS platform which helps manage and monitor Flow execution (Prefect Technologies, Inc., 2021). Prefect provides a robust and observable way of running recipes and is our current default model for the Pangeo Forge cloud automation.
- **Compilation to Apache Beam Pipeline**: Apache Beam is an open source framework for defining parallel processing pipelines for batch and streaming data (Apache Software Foundation, 2016). Beam Pipelines are high-level dataflow graphs, composed of distributed datasets and globally optimized, lazily evaluated processing steps. By compiling to Beam Pipelines, `pangeo-forge-recipes` can be executed on major distributed computation systems including Apache Spark (Zaharia et al., 2016), Apache Fink (Apache Software Foundation, 2015), and Google Cloud Dataflow (Akidau et al., 2015), as well as through intermediaries such as Hadoop, Yarn, Mesos, and Kubernetes (Shvachko et al., 2010; Hindman et al., 2011; Brewer, 2015). Beam is a multi-language framework capable of executing multiple languages in a single Pipeline. This makes it possible to incorporate recipes into execution workflows outside of the Python ecosystem.

In addition to these execution frameworks, recipe steps can be manually run in sequential fashion in a Jupyter Notebook or other interactive environment (Ragan-Kelley et al., 2014). This facilitates user introspection and debugging.

3.3. Cloud Automation Platform

The nuclei of Pangeo Forge cloud automation are Bakeries, cloud compute clusters dedicated specifically to executing recipes. Bakeries provide a setting for contributors to run their recipes on large-scale, distributed infrastructure and deposit ARCO datasets into performant publicly-accessible cloud storage, all entirely free of cost for the user. By running their recipes in a Bakery, contributors are not only gaining access to free compute and storage for themselves, but are also making a considerable contribution back to the global Pangeo Forge community in the form of ARCO datasets which will be easily discoverable and reusable by anyone with access to a web browser.

Pangeo Forge follows the example of Conda Forge in managing its contribution process through the cloud-hosted version control platform GitHub. Recipe contributors who wish to run their recipes in a Bakery first submit their draft recipes via a Pull Request (PR) to the Pangeo Forge `staged-recipes` repository which, as the name implies, is a holding area for incoming recipes. Following an iterative review process, described in detail below, recipe PRs are approved by Pangeo Forge maintainers, at which point their contents are automatically transferred out of the `staged-recipes` repository and incorporated into a new, standalone repository known as a Feedstock. It is from this Feedstock repository that recipe execution is dispatched to the Bakery compute cluster. The details of and rationale behind this workflow are provided in the following subsections.

3.3.1. Contribution Workflow

Continuous integration (CI) is a software development practice whereby code contributions are reviewed automatically by a suite of specialized test software prior to being incorporated into a production codebase. CI improves code quality by catching errors or incompatibilities that may escape a human reviewer's attention. It also allows code contributions to a large project to scale non-linearly to maintainer effort. Equipped with a robust CI infrastructure, a single software package maintainer can review and incorporate large numbers of contributions with high confidence of their compatibility with the underlying codebase.

Pangeo Forge currently relies on GitHub's built-in CI infrastructure, GitHub Actions, for automated review of incoming recipe PRs (**Figure 4**). The first stage of this review process consists of checks that the submitted files, including the `meta.yaml` metadata and the `recipe.py` algorithm module, conform to the technical and stylistic specifications defined in the Pangeo Forge documentation. If errors are identified at this stage, the contributor is notified automatically and given a list of recommended changes, which must be incorporated prior to advancing to the next stage of evaluation.

Once the PR passes this first gate, a human project maintainer dispatches a command to run an automated execution test of the recipe. This test of a reduced subset of the recipe runs the same Prefect workflows on the same Bakery infrastructure which will be used in the full-scale data transformation. Creation of the reduced recipe is performed by a Pangeo Forge function which prunes the dataset to a specified subset of increments in the concatenation dimension. Any changes required to the recipe's functionality are identified here. For datasets expected to conform to Climate and Forecast (CF) Metadata Conventions, we plan to implement compliance checks at this stage using established validation tooling such as the Centre for Environmental Data Analysis (CEDA) CF Checker and the Integrated Ocean Observing System (IOOS) Compliance Checker (Adams et al., 2021; Eaton et al., 2021; Hatcher, 2021). Following an iterative process of corrections based on the results of the automated execution test (or a series of such tests, as necessary), the recipe PR is accepted by a human maintainer. At this point, a Feedstock repository is programmatically generated by incorporating the recipe PR files into a predefined repository template.

Creation of a Feedstock repository from the recipe PR triggers the full build of the ARCO dataset, after which the only remaining step in the contribution workflow is the generation of a catalog listing for the dataset, an automated process dispatched by GitHub Actions.

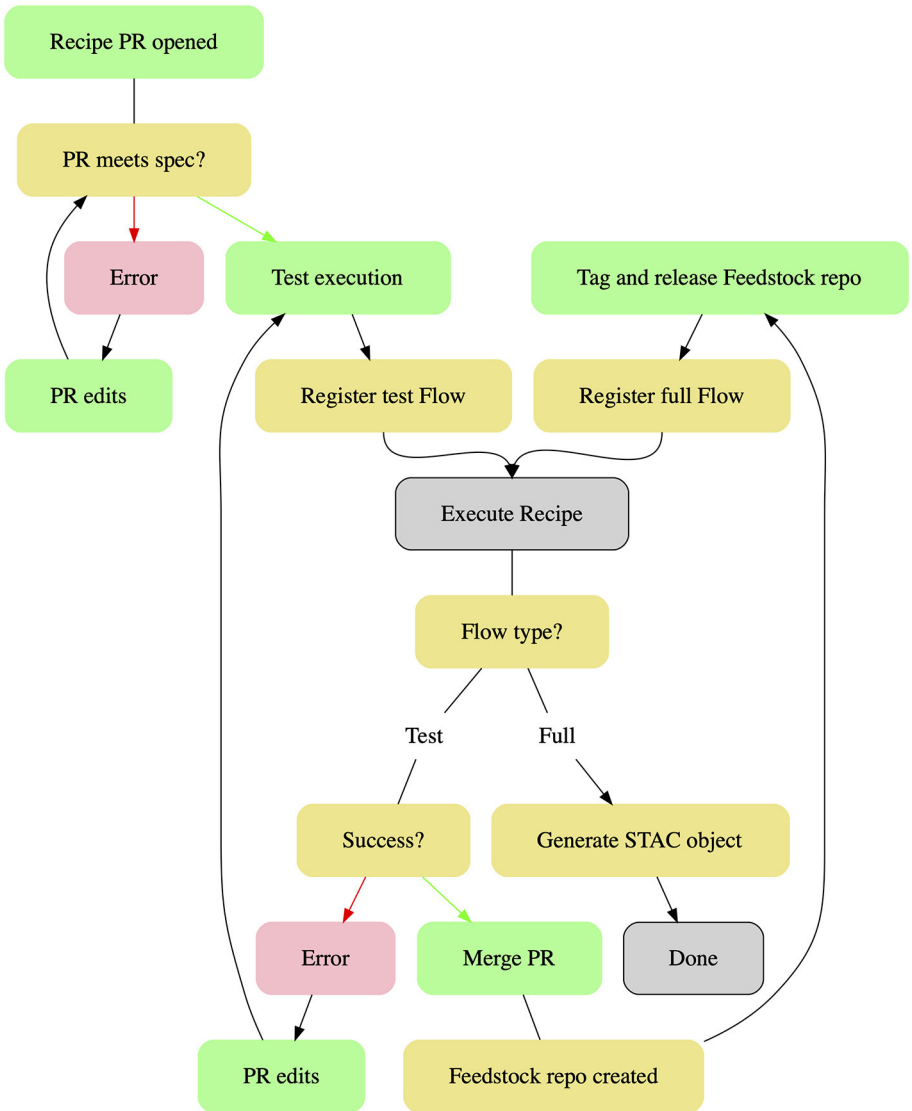


FIGURE 4 | Pangeo Forge contribution workflow.

3.3.2. Feedstocks

Feedstocks are GitHub repositories which place user-contributed recipes adjacent to Pangeo Forge's cloud automation tools and grant access to Pangeo Forge credentials for authentication in a Bakery compute cluster. The Feedstock repository approach mirrors the model successfully established in Conda Forge. Those familiar with software version control processes will know that, most often, *merging* a PR results in proposed code changes being incorporated into an existing repository's codebase. As in Conda Forge, merging a PR to `staged-recipes` takes on a slightly different meaning in Pangeo Forge. Rather than incorporating a recipe's code into `staged-recipes`, merging a recipe PR results in the creation of a new, dedicated GitHub repository for the recipe called a Feedstock.

We can think of this new Feedstock repository as the deployed or productionalized version of the recipe. The template from which GitHub Actions automatically generates this repository includes automation hooks which register the recipe's ARCO dataset build with the specified Bakery infrastructure. All of these steps are orchestrated automatically by GitHub Actions and abstracted from the recipe code itself. As emphasized throughout this paper, this separation of concerns is intended to provide a pathway for scientific domain experts to participate in ARCO data curation without the requirement that they understand the highly-specialized domain of cloud infrastructure automation.

As public GitHub repositories, Feedstocks serve as invaluable touchstones for ARCO dataset provenance tracking. Most users will discover datasets through a catalog (more on cataloging in section 3.4). Alongside other metadata, the catalog entry for

each dataset will contain a link to the Feedstock repository used to create it. This link connects the user to the precise recipe code used to produce each ARCO dataset. Among other benefits, transparent provenance allows data users to investigate whether apparent dataset errors or inconsistencies originate in the archival source data, or are artifacts of the ARCO production process. If the latter, the GitHub repository provides a natural place for collaboration on a solution to the problem. Each time a Feedstock repository is tagged with a new version number, the recipe it contains is re-built to reflect any changes made since the prior version.

Pangeo Forge implements a two-element semantic versioning scheme for Feedstocks (and, by extension, the ARCO datasets they produce). Each Feedstock is assigned a version conforming to the format `MAJOR.MINOR` and beginning at `1.0`. Increments to the minor version are made for changes which are likely to be backwards-compatible with user code that relies on an earlier version of the data. Such updates include metadata corrections, adding new variables, or extending the temporal range of existing variables. Major version increments are triggered for non-backwards-compatible edits such as changing existing variable names or revising preprocessing functions such that they alter existing variable arrays. Pangeo Forge will maintain prebuilt copies of all major versions of a given dataset, but in the interest of storage efficiency will only retain the latest minor version for each of these major versioned datasets. (For example, if prebuilt copies of `1.0` and `2.0` exist in storage when `2.1` is released, `1.0` will be retained but `2.0` will be overwritten by `2.1`.) In cases where a user may have a need for a specific minor version of a dataset that has already been superseded in storage by a new minor version release, the corresponding Feedstock can be used to rebuild any version of the ARCO data store on an as-needed basis.

3.3.3. Bakeries: On-Demand Cloud Clusters Paired With Cloud Storage Targets

As in Conda Forge, the majority of Pangeo Forge users will not execute recipes themselves, but rather interact with recipe outputs which are pre-built by shared cloud infrastructure. As such, execution typically only occurs once per recipe (or, in the case of updated recipe versions, once per recipe version). This one-time execution builds the ARCO dataset to a publicly-accessible cloud storage bucket. Arbitrary numbers of data users can then access the pre-built dataset directly from this single shared copy. This approach has many advantages for our use case, including:

- Shared compute is provisioned and optimized by cloud infrastructure experts within our community to excel at the specific workloads associated with ARCO dataset production.
- As a shared resource, Pangeo Forge cloud compute can be scaled to be larger and more powerful than most community users are likely to be able to provide themselves.
- Storage and compute costs (financial, and in terms of environmental footprint) are not duplicated unnecessarily.

Costs for these shared resources are currently covered through a combination of free credits provided by technology service providers and grants awarded to Pangeo Forge.

Bakeries, instances of Pangeo Forge's shared cloud infrastructure, can be created on Amazon Web Services, Microsoft Azure, and Google Cloud Platform cloud infrastructure. In keeping with the aforementioned Right to Replicate, an open source template repository, tracing a clear pathway for reproducing our entire technology stack, is published on GitHub for each supported deployment type (2i2c.org, 2021). In practice, the cost and complexity of these deployments likely means they will be undertaken by organizations rather than individuals. Over time, we anticipate the benefits of participating in Pangeo Forge will motivate a wide range of both non-profit and commercial partners to establish Bakeries for community use. The greater the number and scale of Bakeries in operation, the greater the capacity of Pangeo Forge to democratize the means of ARCO data production.

When a community member submits a Pangeo Forge recipe, they use the `meta.yaml` file included as part of each recipe submission to specify the Bakery on which to execute it, and the target storage location within that Bakery in which to deposit the resulting dataset. Each Bakery will manage their own compliment of storage buckets. Available Bakeries and their specifications, including storage bucket protocols and locations, are recorded in a public database for reference. Selection of one Bakery over another may be based on factors including the geographic location of the associated storage bucket(s), given that physical proximity of compute resources to data impacts performance for big data analytics.

3.4. Cataloging and Loading

The SpatioTemporal Asset Catalog (STAC) is a human and machine readable cataloging standard gaining rapid and broad traction in the geospatial and earth observation (EO) communities (Alemohammad, 2019; Emanuele, 2020; Holmes et al., 2021). The value of STAC is enhanced by its tooling ecosystem, which includes interfaces for many programming languages and a community-supported web frontend (Emanuele et al., 2021; Fitzsimmons et al., 2021). STAC was not originally conceived as a cataloging solution for the Earth-system model (ESM) data which will constitute a majority of Pangeo Forge's ARCO data holdings, however extensions such as the Datacube Extension bring descriptive cataloging of ESM data with STAC within reach (Mohr et al., 2021). Despite the imperfect fit of ESM data into STAC, the momentum behind this specification and its associated ecosystem recommends it as the best option for implementation of our user-facing catalog.

Following the completion of each ARCO production build, GitHub Actions automatically generates a STAC listing for the resulting dataset and adds it to the Pangeo Forge root catalog. Information which can be retrieved from the dataset itself (including dimensions, shape, coordinates, and variable names) is used to populate the catalog listing whenever possible. Fields likely not present within the dataset, such as a long description and license type, are populated with values from the `meta.yaml` file which contributors include as part of each recipe.

STAC provides not only a browsing interface, but also defines a streamlined pathway for loading datasets. Catalog-mediated loading simplifies the user experience as compared to the added complexity of loading directly from a cloud storage Uniform Resource Identifier (URI). Pangeo Forge currently provides documentation for loading datasets with Python into Jupyter Notebooks, given that our early adopters are likely to be Python users (Perkel, 2018). One distinct advantage of STAC's JSON-based specification over other language-specific cataloging options is its current (or in some cases, planned) interoperability with a wide variety of programming languages. We look forward to documenting catalog access from JavaScript, R, Julia, and many other contemporary languages as our user community grows.

Discoverability is the ease with which someone without prior knowledge of a particular dataset can find out about its existence, locate the data, and make use of it. As the project grows, we aspire to enhance data discoverability by offering a range of search modalities for the Pangeo Forge ARCO dataset catalog, enabling users to explore available datasets by spatial, temporal, and variable extents.

4. EXAMPLES

In the course of development and validation, we employed Pangeo Forge to transform a selection of archival NetCDF datasets, collectively totalling more than 2.5 terabytes in size, into the cloud-optimized Zarr format. The resulting ARCO datasets were stored on the Open Storage Network (OSN), an NSF-funded instance of Amazon Web Services S3 storage infrastructure, and have already been featured in multiple presentations and/or played a central role in ongoing research initiatives. We offer a brief summary of these example results below followed by some general reflections, drawn from these experiences, on the performance of the platform to date.

4.1. SWOT Ocean Model Intercomparison

The upcoming Surface Water and Ocean Topography (SWOT) satellite mission will measure sea-surface height at high resolution with synthetic aperture radar (Morrow et al., 2019). In coordination with this mission, an international consortium of oceanographers are currently undertaking modeling and *in-situ* field campaigns for purposes of comparison to the forthcoming SWOT satellite measurements (Li, 2019). As part of these efforts, we have transformed portions of the outputs from the FESOM, GIGATL, HYCOM, eNATL60, and ORCA36 ocean models into ARCO datasets with Pangeo Forge (Chassignet et al., 2007; Danilov et al., 2017; Brodeau et al., 2020; Castrillo, 2020; Gula, 2021). From a technical perspective, these transformations involved caching approximately a terabyte of ocean model data from FTP servers in France and Germany onto Google Cloud Storage in Iowa, USA via Pangeo Forge's internal file transfer utilities. This experience highlighted the persisting influence of geographic distance on network communication speeds and led to many improvements in how we manage file transfer internally within the platform. From the standpoint of data structure, the multigigabyte-scale array sizes contained within some of these model outputs encouraged the development of a specialized subsetting pathway within `pangeo-forge-recipes` for handling larger-than-memory input arrays.

4.2. NOAA Optimal Interpolation Sea Surface Temperature

NOAA's Optimal Interpolation Sea Surface Temperature (OISST) is a daily resolution data product combining *in-situ* field measurements with satellite temperature observations from the Advanced Very High Resolution Radiometer (AVHRR) (Huang et al., 2021). With Pangeo Forge, we created a single consolidated Zarr store from 14,372 NOAA OISST source files spanning a time range from 1981 to 2021. This Zarr store was subsequently used as part of investigations into the morphology of ocean temperature extremes (Scannell et al., 2021). In many ways, this flavor of recipe (concatenation of NetCDF timeseries archives into a consolidated ARCO store) is what the earliest versions of Pangeo Forge were designed to excel at. We therefore relied heavily on this recipe during early development as a useful test case for our cloud automation infrastructure.

```
import gcsfs
import xarray as xr
# open data
url = (
    'gs://pangeo-forge-us-central1/pangeo-forge/'
    'cmems/sea-level-anomalies.zarr'
)
gcs = gcsfs.GCSFileSystem(requester_pays=True)
ds = xr.open_zarr(
    gcs.get_mapper(url), consolidated=True,
)
# calculate mean
sla_zm = ds.sla.mean('longitude', keep_attrs=True)
# compute using Dask cluster
with cluster.get_client():
    sla_zm.load()
sla_zm.plot(robust=True, x='time')
```

Listing 3 | Code used to generate **Figure 5** from the Pangeo Forge ARCO sea-level data.

4.3. CMEMS Sea Surface Altimetry

A 70 gigabyte ARCO dataset of gridded sea surface altimetry measurements was assembled by Pangeo Forge from nearly 9,000 files sourced from the Copernicus Marine Service (Copernicus Marine Environment Monitoring Service, 2021). For researchers wishing to study trends in sea level, downloading so many files is a laborious barrier to science. With the Pangeo Forge ARCO dataset, a reduction over the entire dataset to visualize the global patterns of sea-level rise can be accomplished in less than a minute and with just a few lines of code (shown in **Listing 3**). This calculation was performed as part of live demonstrations of Pangeo Forge presented at recent ESIP and Research Running on Cloud Compute and Emerging Technologies (RRoCCET) conferences (Barciauskas et al., 2021; Stern, 2021).

4.4. CESM POP 1-Degree

Processing this low-resolution output of the Community Earth System Model (CESM) became an unexpected but

welcome opportunity to examine how Pangeo Forge handles user credentials for accessing source files and resulted directly in the addition of query string authentication features to `pangeo-forge-recipes`. Regarding the data transformation itself, the source files for this recipe represented yet another example of containing larger-than-memory variable arrays (National Center for Atmospheric Research, 2021). The development team's swift and successful adaptation of Pangeo Forge to accommodate this use case is a testament to the extensibility of the platform's base abstractions.

4.5. SODA 3.4.2 ICE

The Simple Ocean Data Assimilation (SODA) model aims to reconstruct twentieth century ocean physics (Carton et al., 2018). We transformed a subset of this model's output consisting of roughly 2,100 source files into a consolidated ARCO data store to aid a colleague's ongoing research.

4.6. Challenges, Performance, and Costs

We have had no difficulty converting any of the NetCDF files from our use cases into Zarr, thanks to Xarray's sophisticated metadata and encoding management. Xarray faithfully replicates all variables, metadata, and datatypes present in the archival NetCDF files into their Zarr analogs such that the resulting Zarr stores, when opened with Xarray, are identical to the dataset present in the original (aggregated) NetCDF files. The main known limitation of Xarray's Zarr interface is that it does not support hierarchically nested NetCDF groups, only flat groups; this particular limitation has not affected our above-listed use cases.

Pangeo Forge is generally I/O bound. The greatest challenge we have experienced is slow downloading from source data archives during the caching phase of recipe execution. If too many simultaneous requests are made to an HTTP or FTP source server, this will typically result in the per-file transfer throughput decreasing considerably. Therefore, caching source files for a given recipe is not highly parallelizable. As noted in section 4.1, transcontinental data transfer can be slow process, even for sequential requests. Once the source files are cached into the cloud, however, platform performance scales out well, since all I/O is happening against cloud object storage.

We have not yet made a systematic assessment of cost and performance. Regarding minimum hardware requirements, Bakery workloads are typically distributed across large numbers of lightweight compute nodes. In a typical implementation, each node may be provisioned with roughly 4 gigabytes of RAM and one CPU core. The larger-than-memory archival arrays referenced in sections 4.1, 4.4 challenged this computational model and prompted the addition of subsetting routines to the platform that facilitate division of arbitrarily-sized input arrays along one or multiple dimensional axes. This allows our lightweight compute nodes to handle inputs in excess of their RAM allocation. As we move from the initial software development phase into productionalization of increasing numbers of Bakeries, we look forward to sharing more fine-grained assessments of the platform's performance and resource requirements.

5. FUTURE OUTLOOK

As of the time of writing this paper, all of the major components of Pangeo Forge (with the exception of the data catalog) have been released openly on GitHub, tested thoroughly, and integrated through end-to-end workflows in the cloud. Dozens of actual and potential users have interacted with the project via GitHub issues and bi-weekly meetings. However, the platform has not been officially "launched," as in, advertised broadly to the public as open for business. We anticipate taking this step in early 2022. After that point, development will continue indefinitely into the future as we continue to refine and improve the service in response to user feedback.

The current development of Pangeo Forge is supported by a 3 year grant from the National Science Foundation (NSF) EarthCube program. Storage expenses are covered through our partnership with the Open Storage Network (OSN), which provides Pangeo Forge with 100 terabytes of cloud storage space, accessible over the S3 protocol for free (Public cloud storage buckets often implement a "requester-pays" model in which users are responsible for the cost of moving data; our OSN storage does not). All three major cloud providers offer programs for free hosting of public scientific datasets. We anticipate engaging in these programs as our storage needs grow. We have also begun to evaluate distributed, peer-to-peer storage systems such as the InterPlanetary FileSystem (IPFS) and Filecoin as an alternative storage option.

Pangeo Forge is not itself an archival repository but rather a platform for transforming data, sourced from archival repositories, into optimized formats. We therefore do not commit to preserving every recipe's materialized data in perpetuity. In nearly all cases, recipes source data from archival repositories with a long-term stewardship plan. It should therefore almost always be possible to regenerate a Pangeo Forge ARCO dataset by re-running the recipe contained in the versioned Feedstock repository from which it was originally built.

We hope that the platform we create during the course of our NSF award will gain traction that merits long-term financial support for the project. The level of support required will depend on the volume of community interest and participation. In any scenario, however, it is not feasible for Pangeo Forge core development team to personally maintain every Feedstock. Instead, via the crowdsourcing model, we aspire to leverage the expertise of a large community of contributors, each of whom will be responsible for keeping their recipes up to date. The core team will support these maintainers to the greatest degree possible, via direct mentorship as well as more scalable modes of support such as documentation and automated integration tests. Recipe contribution is not the only thing we envision crowdsourcing. Indeed, the platform itself is designed to be "franchisable": any organization can run a Bakery. We envision Pangeo Forge not as a single system with one owner but rather as a federation. Participating organizations will bear the compute and storage costs of the datasets they care about supporting and recipes will be routed to an appropriate Bakery as part of the GitHub contribution workflow.

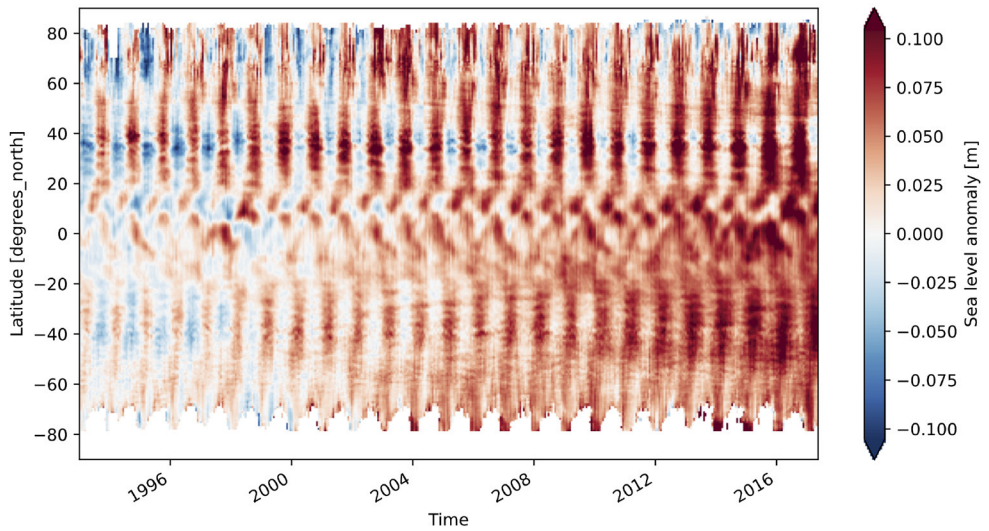


FIGURE 5 | Daily zonal mean sea-level anomaly, calculated from Pangeo Forge ARCO dataset.

In the remainder of this final section, we conclude by imagining a future state, several years from now, in which Pangeo Forge has cultivated a broad community of recipe contributors from across disciplines, who help populate and maintain a multi-petabyte database of ARCO datasets in the cloud. How will this transform research and applications using environmental data? What follows is inherently speculative, and we look forward to revisiting these speculations in several years time to see how things turn out.

5.1. An Ecosystem for Open Science

Pangeo Forge and the ARCO data repositories it generates are most valuable as part of a broader ecosystem for open science in the cloud (Gentemann et al., 2021). In particular, Pangeo Forge ARCO data is designed to be used together with scalable, data-proximate computing. For interactive data analysis, Jupyter (including Jupyter Lab and Jupyter Hub) is emerging as a consensus open-source platform for the scientific community (Kluyver et al., 2016). Jupyter supports interactive computations in all major scientific computing languages, including Python, R, and Julia (We note especially that, although Pangeo Forge itself is written in Python, the data formats and catalogs it generates are all based on open standards, accessible from any major programming language). Jupyter in the cloud, combined with cloud-native parallel computing tools such as Dask (Rocklin, 2015) and Spark (Zaharia et al., 2016), creates a complete end-to-end solution for data-intensive research based purely on open-source software. By accelerating the production and sharing of ARCO data, we hope to stimulate further development and broad adoption of this new model for scientific research.

Beyond expert analysis, we also hope that the datasets produced by Pangeo Forge will enable a rich downstream ecosystem of tools to allow non-experts to interact with large, complex datasets *without writing code*. ARCO formats like Zarr are ideal for powering APIs, dashboards, and interactive websites, since they are based on open standards and can be read quickly from any programming language, including JavaScript, the language of the web. As an example, the sea-level data shown in **Figure 5** could be used to create an interactive data visualization website for high-school students to study sea level change. Students wishing to go beyond the visual exploration could transition to an interactive Jupyter notebook and write their first lines of code, all pointing at the same underling data. Similarly, industry experts and policy makers could use such tools to examine climate impacts on their sector of interest. The direct provenance chain from the interactive tool, to the ARCO data copy, to the original upstream data provider would provide a fully transparent and trustworthy foundation for decision making.

5.2. Collaboration and Recognition Around Data Production

While nearly all scientists recognize the importance of data for research, scientific incentive systems do not value data production nearly as much as other types of scientific work, such as model development (Pierce et al., 2019). This was emphasized in a recent paper from Google Research, warning of the impact of data quality issues in the context of artificial intelligence research (Sambasivan et al., 2021). The undervaluing of "data work" is pervasive in the sciences, as evidenced by the existence of pejorative terms such as "data janitor." Data work often occurs in the shadows of science, not talked about much in papers or recognized via honors and awards. One of our central hopes with Pangeo Forge is that the preparation of well curated, quality controlled datasets immediately accessible to high-performance computing will become an area of increased collaboration and visibility in environmental science research. By leveraging the interactivity inherent in GitHub discussions, we hope to see researchers from different institutions and countries coming together around building shared datasets of use to many different

groups. By establishing a community storehouse of datasets themselves, as well as Feedstock repositories containing dataset provenances, we hope to offer citable artifacts of data production which, if reused and credited by the community, may serve to elevate the profile of this essential scientific work. Perhaps 1 day we will give an award for "most valuable recipe"!

Pangeo Forge does not currently implement a system for assigning unique persistent identifiers, such as Digital Object Identifiers (DOIs), to either Feedstocks or the datasets they produce. We certainly appreciate the tremendous benefit such identifiers provide, particularly for purposes of academic citation. As noted above, at this stage of our development we are not making a commitment to keeping datasets online in perpetuity, as would be required for a DOI. This reflection leads us to conclude that the Pangeo Forge Feedstock, a lightweight repository which will be permanently stored on GitHub, may in fact be the most appropriate object for DOI-assignment and citation. Feedstocks (and within them, recipes) are also the products which are most plainly expressive of contributors' technical and domain expertise. We welcome community feedback on how to best support contributors to receive the credit and recognition they deserve.

5.3. Asking More Ambitious Questions From Data

A recurring theme of the examples in section 4 is the relative simplicity of aligning thousands of source files into a single consolidated dataset with Pangeo Forge. The ARCO datasets which result from this process are not simply faster to work with than archival data, in many cases they enable an entirely new worldview. When working within the confines of traditional filesystems, it can be difficult for the scientific imagination to fly nimbly across the grand spatial and temporal scales permitted by ARCO workflows. By making entire worlds (observed or synthetic, past or future) accessible in an instant through shared ARCO data stores, we wholly expect that Pangeo Forge to not only *accelerate* existing science, but to also play a pivotal role in the *reimagination* of what's possible in ocean, weather, and climate science at scale.

5.4. Reproducibility in Action

Each Pangeo Forge recipe encodes data provenance starting from an archival source, all the way to the precise derived version used for a given research project. Tracking an unbroken provenance chain is particularly important in the context of ARCO data, which undergoes significant transformation prior to being used for analysis. The algorithms used to create ARCO datasets encode assumptions about what types of homogenization and/or simplification may serve the investigation for which the dataset is being produced. These judgement calls can easily be as impactful to the scientific outcome as the analysis itself. By tracking the ARCO production methodology through a recipe's Feedstock repository, Pangeo Forge affords visibility into the choices made at the data curation stage of research.

The oft-quoted eighty-twenty rule describes a typical ratio of time required for cleaning and preparing data vs. actually performing analysis. Depending on the type of preprocessing applied to a dataset, the time and technical knowledge required to reproduce previous derived datasets, let alone results, represents a major barrier to reproducibility in computational science. Duplication of data preparation is unnecessary and can be avoided if the dataset used for a given study, along with the recipe used to create it, are made publicly accessible.

5.5. Broadening Participation

Traditionally, working with big environmental datasets has required considerable infrastructure: big computers, hard drives, and IT staff to maintain them. This severely limits who can participate in research. One of the great transformative potentials of cloud-native science is the ability to put powerful infrastructure into the hands of anyone with an internet connection (Gentemann et al., 2021). In our recent experience, we have observed that it is easy enough to get started with cloud computing; the hard part is getting the right data into the cloud in the right format.

Pangeo Forge not only shifts the infrastructure burden of data production from local infrastructure to the cloud; it also lightens the *cognitive burden* for potential contributors by encouraging them to focus on the domain-specific details of the data, rather than the data engineering. As a recipe contributor to Pangeo Forge, anyone with a laptop can run their ARCO transformation algorithm at a scale previously only available to a small organizationally-affiliated group.

The true success of Pangeo Forge depends on creation of a space where a diverse community of recipe contributors can come together to curate the ARCO datasets which will define the next decade of cloud-native, big-data ocean, weather, and climate science. How we best nurture this community, and ensure they have the education, tools, and support they need to succeed, remains an open question, and an area where we seek feedback from the reader.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CS drafted the manuscript with contributions from all other authors. All authors contributed to the design and implementation of Pangeo Forge. All authors read and approved the submitted version of the manuscript.

FUNDING

Pangeo Forge development is funded by the National Science Foundation (NSF) Award 2026932.

ACKNOWLEDGMENTS

Pangeo Forge benefits from the thoughtful contributions and feedback of a broad community of scientists and software engineers. We extend heartfelt thanks to all those who have contributed in any form including code, documentation, or via participation in community meetings and discussions.

REFERENCES

2i2c.org (2021). *The Customer Right to Replicate*. Available online at: https://2i2c.org/right-to-replicate/ (accessed September 22, 2021).

Abernathey, R. P., Augspurger, T., Banihirwe, A., Blackmon-Luca, C. C., Crone, T. J., Gentemann, C. L., et al. (2021). Cloud-native repositories for big scientific data. *Comput. Sci. Eng.* 23, 26–35. doi: 10.1109/MCSE.2021.3059437

Adams, B., Campbell, L., Kell, D., Fernandes, F., Fratantonio, B., Foster, D., et al. (2021). *IOOS Compliance Checker*. Available online at: https://github.com/ioos/compliance-checker

Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., et al. (2015). The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proc. VLDB Endowment* 8, 1792–1803. Available online at: https://research.google/pubs/pub43864/

Alemohammad, H. (2019). "Radiant mlhub: A repository for machine learning ready geospatial training data," in *AGU Fall Meeting Abstracts*, (IN11A–05) (Washington, DC).

Anaconda Inc. (2021). *Anaconda Software Distribution*. Available online at: https://docs.anaconda.com/

Apache Software Foundation (2015). *Apache Fink*. Available online at: https://flink.apache.org/

Apache Software Foundation (2016). *Apache Beam*. Available online at: https://beam.apache.org/

Barciauskas, A., Shrestha, S., Casey, R., Signell, R., Friesz, A., Olson, S., et al. (2021). "The saga continues: cloud-optimized data formats," in *Earth Science Information Partners (ESIP) Summer Meeting 2021* (Severna Park, MD: ESIP).

Brewer, E. A. (2015). Kubernetes and the path to cloud native. in *Proceedings of the Sixth ACM Symposium on Cloud Computing, SoCC '15, (Association for Computing Machinery)* (New York, NY), 167.

Brodeau, L., Sommer, J. L., and Albert, A. (2020). *Ocean-next/eNATL60: Material Describing the Set-up and the Assessment of NEMO-eNATL60 Simulations (Version v1)*. Zenodo. doi: 10.5281/zenodo.4032732

Carton, J. A., Chepurin, G. A., and Chen, L. (2018). Soda3: a new ocean climate reanalysis. *J. Climate* 31, 6967–6983. doi: 10.1175/JCLI-D-18-0149.1

Castrillo, M. (2020). "The nemo orca36 configuration and approaches to increase nemo4 efficiency," in *The 6th European Network for Earth System Modelling (ENES) Workshop on High Performance Computing for Climate and Weather* (ENES). Available online at: https://www.esiwace.eu/events/6th-hpc-workshop/presentations/the-nemo-orca36-configuration-and-approaches-to-increase-nemo4-efficiency

Chassignet, E. P., Hurlburt, H. E., Smedstad, O. M., Halliwell, G. R., Hogan, P. J., Wallcraft, A. J., et al. (2007). The hycom (hybrid coordinate ocean model) data assimilative system. *J. Marine Syst.* 65, 60–83. doi: 10.1016/j.jmarsys.2005.09.016

Conda-Forge Community (2015). *The Conda-Forge Project: Community-Based Software Distribution Built on the Conda Package Format and Ecosystem*. Zenodo. doi: 10.5281/zenodo.4774217

Copernicus Marine Environment Monitoring Service (2021). *Global Ocean Gridded l4 Sea Surface Heights and Derived Variables Reprocessed (1993-ongoing)*. Available online at: https://resources.marine.copernicus.eu/product-detail/SEALEVEL_GLO_PHY_L4_REP_OBSERVATIONS_008_047/INFORMATION

Cornillon, P., Adams, J., Blumenthal, M. B., Chassignet, E., Davis, E., Hankin, S., Kinter, J., et al. (2009). Nvods and the development of opendap. *Oceanography* 22, 116–127. doi: 10.5670/oceanog.2009.43

Danilov, S., Sidorenko, D., Wang, Q., and Jung, T. (2017). The finite-volume sea ice–ocean model (fesom2). *Geosci. Model Develop.* 10, 765–789. doi: 10.5194/gmd-10-765-2017

Dask Development Team (2016). *Dask: Library for Dynamic Task Scheduling*. Avaialable online at: https://dask.org

Durant, M. (2021). *fsspec: Filesystem Interfaces for Python*. Avaialable online at: https://filesystem-spec.readthedocs.io/

Durant, M., Sterzinger, L., Signell, R., Jelenak, A., Maddox, L., Bell, R., et al. (2021). *kerchunk*. Avaialable online at: https://github.com/fsspec/kerchunk

Durbin, C., Quinn, P., and Shum, D. (2020). *Task 51-cloud-optimized format study*. NTRS—NASA Technical Reports Server.

Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., and Lymburner, L. (2018). Analysis ready data: Enabling analysis of the landsat archive. *Remote Sens.* 10, 1363. doi: 10.3390/rs10091363

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2021). *NetCDF Climate and Forecast (CF) Metadata Conventions*. Available online at: https://cfconventions.org/

Emanuele, R. (2020). "Using spatiotemporal asset catalogs (stac) to modularize end-to-end machine learning workflows for remote sensing data," in *AGU Fall Meeting Abstracts*, (IN007–01) (Washington, DC).

Emanuele, R., Duckworth, J., Engmark, V., Kassel, S., Schwehr, K., Olaya, V., et al. (2021). *PySTAC: A library for working with SpatioTemporal Asset Catalog in Python 3*. Available online at: https://github.com/stac-utils/pystac

Fitzsimmons, S., Mohr, M., Emanuele, R., Blackmon-Luca, C., et al. (2021). *STAC Browser: A Vue-Based STAC Browser for Static Catalogs and APIs*. Available online at: https://github.com/radiantearth/stac-browser

Gentemann, C. L., Holdgraf, C., Abernathey, R., Crichton, D., Colliander, J., Kearns, E. J., et al. (2021). Science storms the cloud. *AGU Adv.*, 2:e2020AV000354. doi: 10.1029/2020AV000354

Gentzsch, W. (2001). Sun grid engine: towards creating a compute power grid. in *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 35–36. doi: 10.1109/CCGRID.2001.923173

Gillies, S. et al. (2013). *Rasterio: Geospatial Raster I/O for Python Programmers*. Mapbox. Available online at: https://github.com/rasterio/rasterio

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. doi: 10.1016/j.rse.2017.06.031

Gula, J. (2021). *Mesharou/GIGATL: Description of the GIGATL Simulations (v1.1)*. Zenodo. doi: 10.5281/zenodo.4948523

Hankin, S. C., Blower, J. D., Carval, T., Casey, K. S., Donlon, C., Lauret, O., et al. (2010). Netcdf-cf-opendap: Standards for ocean data interoperability and object lessons for community data standards processes. in *Oceanobs 2009, Venice Convention Centre, 21-25 septembre 2009, Venise*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hatcher, R. (2021). *cf-checker*. Available online at: https://github.com/cedadev/cf-checker

Henderson, R. L. (1995). "Job scheduling under the portable batch system," in *Job Scheduling Strategies for Parallel Processing*, eds D. G. Feitelson and L. Rudolph (Berlin; Heidelberg: Springer), 279–294.

Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A. D., Katz, R. H., et al. (2011). Mesos: A platform for fine-grained resource sharing in the data center. in *NSDI, vol. 11*. 22–22. Available online at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=+Mesos%3A+A+platform+for+fine-grained+resource+sharing+in+the+data+center.&btnG=

Hobu, Inc. (2021). *Cloud Optimized Point CLoud (COPC)*. Available online at: https://copc.io/

Holmes, C. (2018). *Analysis Ready Data Defined*. Available online at: https://medium.com/planet-stories/analysis-ready-data-defined-5694f6f48815. (accessed September 09, 2021).

Holmes, C. (2021). *Cloud Optimized GeoTIFF Specification*. Available online at: https://github.com/cogeotiff/cog-spec

Holmes, C., Mohr, M., Hanson, M., Banting, J., Smith, M., Mathot, E., et al. (2021). *SpatioTemporal Asset Catalog Specification-Making Geospatial Assets Openly Searchable and Crawlable*. Available online at: https://github.com/radiantearth/stac-spec

Hoyer, S. and Hamman, J. (2017). xarray: N-D labeled arrays and datasets in Python. *J. Open Res. Softw.* 5, 10. doi: 10.5334/jors.148

Hua, H., Barciauskas, A., Chang, G., and Lynnes, C. (2020). "In042-lessons learned on supporting analysis ready data (ard) with analytics optimized data stores/services (aods) in collaborative analysis platforms posters," in *American Geophysical Union (AGU) Fall Meeting 2020* (Washington, DC: AGU).

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., et al. (2021). Improvements of the daily optimum interpolation sea surface temperature (doisst) version 2.1. *J. Climate* 34, 2923–2939. doi: 10.1175/JCLI-D-20-0166.1

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., et al. (2016). "Jupyter notebooks a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, eds F. Loizides and B. Scmidt (Amsterdam: IOS Press), 87–90. Available online at: https://www.iospress.com/catalog/books/positioning-and-power-in-academic-publishing-players-agents-and-agendas

Le Dem, J., and Blue, R. (2021). *Apache Parquet*. Available online at: https://github.com/apache/parquet-format

Li, J. (2019). *SWOT Adopt-A-Crossover Consortium has been endorsed by CLIVAR*. Available online at: https://www.clivar.org/news/swot-%E2%80%98adopt-crossover%E2%80%99-consortium-has-been-endorsed-clivar (accessed September 09, 2021).

Microsoft (2021). *The Planetary Computer*. Available online at: https://planetarycomputer.microsoft.com/

Miles, A., Bussonnier, M., Moore, J., Fulton, A., Bourbeau, J., Onalan, T., et al. (2021). *zarr-developers/zarr-python: v2.10.3*. Zenodo. doi: 10.5281/zenodo.5712786

Mohr, M., Hanson, M., Augspurger, T., Emanuele, R., Holmes, C., Scott, R., et al. (2021). *Datacube Extension Specification*. Available online at: https://github.com/stac-extensions/datacube

Morrow, R., Fu, L.-L., Ardhuin, F., Benkiran, M., Chapron, B., Cosme, E., et al. (2019). Global observations of fine-scale ocean surface topography with the surface water and ocean topography (swot) mission. *Front. Marine Sci.*, 6:232. doi: 10.3389/fmars.2019.00232

National Center for Atmospheric Research. (2021). *One degree, standard resolution CESM simulation from the Accelerated Scientific Discovery Phase of Yellowstone*. NCAR Climate Data Gateway. Available Online at: https://www.earthsystemgrid.org/dataset/ucar.cgd.asd.cs.v5_rel04_BC5_ne30_g16.ocn.proc.daily_ave.html

Pangeo Forge Community (2021). *Pangeo Forge*. Available online at: https://pangeo-forge.readthedocs.io/

Perkel, J. M. (2018). Why jupyter is data scientists' computational notebook of choice. *Nature* 563, 145–147. doi: 10.1038/d41586-018-07196-1

Pierce, H. H., Dev, A., Statham, E., and Bierer, B. E. (2019). Credit data generators for data reuse. *Nature* 570, 30–32. doi: 10.1038/d41586-019-01715-4

Prefect Technologies, Inc. (2021). *Prefect*. Available online at: https://docs.prefect.io/

Quinn, P., Abernathey, R., Signell, R., Neufeld, D., Privette, A., Killick, P., et al. (2020). "Cloud-optimized data," in *Earth Science Information Partners (ESIP) Summer Meeting 2020*. ESIP.

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J., et al. (2014). "The jupyter/ipython architecture: a unified view of computational research, from interactive exploration to communication and publication," in *AGU Fall Meeting Abstracts*, vol. 2014, H44D–07.

Ramamurthy, M. (2017). "Geoscience cyberinfrastructure in the cloud: data-proximate computing to address big data and open science challenges," in *2017 IEEE 13th International Conference on e-Science (e-Science)*, 444–445.

Rew, R., Hartnett, E., Caron, J., et al. (2006). "Netcdf-4: software implementing an enhanced data model for the geosciences," in *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, vol. 6.

Rocklin, M. (2015). "Dask: Parallel computation with blocked algorithms and task scheduling," in *Proceedings of the 14th python in science conference*, vol. 130, 136. Citeseer.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., and Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes ai.

Scannell, H., Abernathey, R., Busecke, J., Gagne, D. J., Thompson, L., and Whitt, D. (2021). Ocetrac: morphological image processing for monitoring ocean temperature extremes. in *Scientific Computing with Python (SciPy) 2021*. SciPy.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10.

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., et al. (2019). Make scientific data fair. *Nature* 570, 27–29. doi: 10.1038/d41586-019-01720-7

Stern, C. (2021). "Analysis ready data in the cloud," in *Research Running on Cloud Compute and Emerging Technologies (RRoCCET) 2021*. RRoCCET. Available at Available online at: https://na.eventscloud.com/file_uploads/25629138ed86f9d6e6b4d8b8189e3b87_ConferenceProceedings.v2.pdf (accessed September 09, 2021).

The Pandas Development Team. (2021). *pandas-dev/pandas: Pandas*. Zenodo. doi: 10.5281/zenodo.3509134

TileDB, Inc. (2021). *TileDB*. Available online at: https://docs.tiledb.com/

Wagemann, J. (2020). *ERA5 Reanalysis Data Available in Earth Engine*. Available online at: https://www.ecmwf.int/en/newsletter/162/news/era5-reanalysis-data-available-earth-engine (accessed September 09, 2021).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Yoo, A. B., Jette, M. A., and Grondona, M. (2003). "Slurm: Simple linux utility for resource management," in *Job Scheduling Strategies for Parallel Processing*, eds D. Feitelson, L. Rudolph, and U. Schwiegelshohn (Berlin; Heidelberg: Springer), 44–60.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., et al. (2016). Apache spark: a unified engine for big data processing. *Commun. ACM* 59, 56–65. doi: 10.1145/2934664

Conflict of Interest: AM was employed by company Google. SH was employed by company Development Seed.

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

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

frontiers | Frontiers in Climate

POLICY AND PRACTICE REVIEWS
published: 09 May 2022
doi: 10.3389/fclim.2022.761499

How Can Earth Scientists Contribute to Community Resilience? Challenges and Recommendations

Arika Virapongse[1,2], Rupanwita Gupta[3], Zachary J. Robbins[4], Jonathan Blythe[5], Ruth E. Duerr[2] and Christine Gregg[6]*

[1] Middle Path EcoSolutions, Boulder, CO, United States, [2] The Ronin Institute, Montclair, NJ, United States, [3] Independent Scholar, New York, NY, United States, [4] Department of Forestry, North Carolina State University, Raleigh, NC, United States, [5] Bureau of Ocean Energy Management, Washington, DC, United States, [6] School of Information, University of Michigan, Ann Arbor, MI, United States

Community resilience increases a place-based community's capacity to respond and adapt to life-changing environmental dynamics like climate change and natural disasters. In this paper, we aim to support Earth science's understanding of the challenges communities face when applying Earth science data to their resilience efforts. First, we highlight the relevance of Earth science in community resilience. Then, we summarize these challenges of applying Earth science data to community resilience:

- Inequity in the scientific process,
- Gaps in data ethics and governance,
- A mismatch of scale and focus, and
- Lack of actionable information for communities.

Lastly, we offer the following recommendations to Earth science as starting points to address the challenges presented:

- Integrate community into the scientific data pathway,
- Build capacity to bridge science and place-based community needs,
- Reconcile openness with self-governance, and
- Improve access to data tools to support community resilience.

Keywords: community resilience, data governance, actionable data, data ethics, data pathways, Earth science data

OPEN ACCESS

Edited by:
Nazila Merati, National Oceanic and Atmospheric Administration (NOAA), United States

Reviewed by:
Tasnuva Rouf, University of California, Berkeley, United States
Kripa Jagannathan, Lawrence Berkeley National Laboratory, United States
Mathew Biddle, U.S. Integrated Ocean Observing System, United States

****Correspondence:***
Arika Virapongse
av@middlepatheco.com

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *19 August 2021*
Accepted: *31 March 2022*
Published: *09 May 2022*

Citation:
Virapongse A, Gupta R, Robbins ZJ, Blythe J, Duerr RE and Gregg C (2022) How Can Earth Scientists Contribute to Community Resilience? Challenges and Recommendations. Front. Clim. 4:761499. doi: 10.3389/fclim.2022.761499

INTRODUCTION

Climate change, natural disasters, and public health threats test the durability of our society. A place-based community's capacity to respond to and recover from life-changing events is called community resilience. Science plays a key role in providing evidence that people and groups can use to make informed decisions about their community's resilience. Earth science data, which are qualitative and quantitative products of observation representing properties of objects, events, and their environments (Rowley, 2007), have been historically produced, curated, and managed for use by scientists. Due to the nature of Earth science data, data support has been mostly provided by and restricted to large institutions with access to high computing power (Dutton et al., 1995; Ramapriyan and Behnke, 2019). However, the societal, political, technical, and cultural landscape

around the meaning and use of data is changing. There are growing expectations in society for more open data (ESIP interview: Mayernik and Virapongse, 2019). Decision-makers, such as practitioners, planners, industry leaders, and the general public, seek to leverage scientific data and information to develop new tools, make decisions, and act. There is still much work to be done by Earth scientists and their data science partners to address issues of data access and use by decision-makers (Wee and Piña, 2019), and particularly for community resilience.

The Earth Science Information Partners (ESIP) Community Resilience cluster addresses how Earth science data can be better utilized to support community resilience. Over the past 5 years, members of the Community Resilience cluster (co-authors of this paper) have been leading discussions within ESIP on this topic during ESIP's biannual meetings, through monthly teleconferencing calls, and across other ESIP clusters. With over 150 member organizations, ESIP includes individuals from federal, state, and local government, non-profit organizations, and the private sector. The discussions that we had through these activities greatly informed this paper.

In this paper, we aim to facilitate the contribution of Earth science data to community resilience efforts by:

- summarizing challenges related to how data and information are accessed, trusted, and made actionable for the purposes of community resilience,
- framing how data use and community resilience can work together, and
- presenting recommendations to address the challenges.

We pose the question: **How can Earth science data and information be used more effectively to enhance community resilience?** We aim for this paper to be useful for people that generate, analyze, and manage Earth science data, support the translation of scientific data to information, and apply science-based information for societal benefit.

COMMUNITY RESILIENCE

Community resilience functions within a complex social-ecological system, where people, environment, climate, and other entities interact. In this context, "community" is geographically localized (place-based), while referring to the social interactions that occur among people in a place (Theodori, 2005). Communities can be defined by geopolitical units (e.g., town or neighborhood), social groupings (e.g., demographic profile, social class structure, culture, language), and entities organized around special interests. We draw on a definition of "community resilience" that describes it as the capacity of a system to prevent, adapt, and recover from shocks and stressors so that it grows stronger and is more prepared in the future [United States Agency for International Development (USAID), 2012]. This framing enables us to conceptualize "community resilience" expansively to consider a community's capacity to effectively respond to natural hazards such as flooding, wildfires, and earthquakes (e.g., Cutter et al., 2013), human-caused disasters such as mass shootings (e.g., Aldrich and Meyer, 2015), as well as systemic trauma experienced by indigenous people through repression and colonization (e.g., Kirmayer et al., 2009). While climate resilience in response to adverse climate events has been a recent focus in the literature, our definition encompasses it, acknowledging the multiple scenarios that communities must cope with in the face of external challenges. Our thesis in this paper considers this broader focus of community resilience, with recommendations that can be applied across different contexts and circumstances.

Building the capacity and flexibility of a community to adapt to an ever-changing socio-environment is central to improving community resilience (Magis, 2010). In application, community resilience goals help to frame and guide local- to global-scale decision-making to improve human livelihoods, address environmental change, and prepare communities to cope with hazards, risks, and disasters (PCAST-Executive Office of the President, 2011; Cutter et al., 2013; Bone et al., 2016). Environmental justice goals can also be addressed through community resilience efforts aiming to reduce inequitable exposure from toxic waste in industrial waterfront areas and improving residents' health and quality of life (Bautista et al., 2015); a community's resilience is dependent on the strength of the entire social-ecological system as a whole. Improving community resilience involves accounting for future change and uncertainty in decision-making, including assessing when resisting change is beneficial or detrimental, and developing plans that allow flexibility as needed. To effectively impact a community's resilience, decision-making must be informed and empowered at multiple socio-political scales (including individual, city, and national levels) and across sectors (including community members, private businesses, and government stakeholders) (**Table 1**).

Navigating change to community resilience is often framed within an adaptive cycle, which posits that growth is not constant and reorganization is required to maintain the community as a functioning entity (Fath et al., 2015). This emphasizes the need for communities to adapt to and build capacity for a changing suite of problems, while being aware of common pitfalls that they may experience when responding to disturbances. Efforts to create a more resilient community often require addressing the tension between the inherent qualities of a system (e.g., physical and ecological structure, function, or states) and the values that people (and often specific groups of people) associate with different components of a system (Higuera et al., 2019). Community resilience has been used to address industry changes (King, 2008), diminishing natural resources (Smith et al., 2012), climate change (Adger et al., 2005; Funfgeld and McEvoy, 2012), health crises (Chandra et al., 2011), health and wellbeing of indigenous people (Kirmayer and Valaskakis, 2009), forest fires (McWethy et al., 2019), and environmental management (Virapongse et al., 2016). Fundamental concepts of resilience can be operationalized to enhance a community's ability to use available Earth science data and information for their benefit.

TABLE 1 | How geopolitical scale matters for community resilience data/information needs.

Geopolitical unit	Example decision-maker	Example resilience need	Example of data and information needs, and their infrastructure and policies
Individual	A resident of a town or village	Planning for evacuation or sheltering-in-place at home for a hurricane	Predicted heights of a tidal surge on resident's home and status of evacuation routes
Community	Leaders and members of subsistence-based indigenous communities	Adapting subsistence lifestyles to climate and ecosystem changes that are occurring	Localized ecosystem data with observed trends
City	City Planners	Long to medium term uncertainties regarding climate/natural disasters/ COVID-19	Improved data capability for smaller cities; improved integration of diverse dataset and information for decision-making
County	Leaders, organizers, and members of local farming/forestry cooperatives or granges	Understanding short term climate forecasts for crop/financial planning	Climate data scaled appropriately, greater short-term certainty
Region	Leaders and participants in recovery restoration efforts, like in the US Gulf Coast	Long term restoration planning, implementation, monitoring, and disaster recovery	Challenges with data quality, documentation, storage, product integration, discovery, accessibility, and archiving
Country	Members of legislature	Equitably setting land use policies that support the needs and interests of the populace	Information collecting processes that adequately represent diverse perspectives equitably; Improved data and information on the true effect, cost, and benefit of land use practices to people and their environment
Global	Members of the UN and other multinational entities	Meeting climate action sustainable development goals	Global statistics indicating the effect of climate change on people and their environment

DATA CHALLENGES IN COMMUNITY RESILIENCE

Closing the gap between community resilience needs and Earth science data is an ongoing, multifaceted challenge. While we note some well-known barriers that already have relatively extensive literature describing these issues (for example, downscaling), we concentrate primarily on aspects that are less well-studied.

Inequity in the Scientific Process

Without equitable representation in the scientific process, it is unlikely that the resilience needs of communities—particularly among those segments of the population that are historically underrepresented across society—will be sufficiently addressed. Scientific bodies, composed of academics and professionals, largely determine who asks questions and makes decisions in science. They decide the objectives of research, the purpose and methods for data collection, the types of data products created, and what research should be funded. Inequities and lack of representation (e.g., race, ethnicity, class, political perspective; Funk, 2012) among scientific decision-makers increases the likelihood that scientific narratives and science agendas are biased and perceived as untrustworthy. Indeed, the stakes are high: scientific decision-making can determine the priority placed on issues and the types of information and knowledge generated to address them, potentially depriving more socially vulnerable groups the right to scientific benefits (Klinsky et al., 2017).

The harmful societal impact of such power discrepancies in decision-making has been recognized in sectors like environmental conservation in the US, where it is increasingly apparent that conservation agendas are being drawn from a primarily White perspective (Green2.0, 2021). Similarly, STEM disciplines (e.g., geoscience) have historically created systemic barriers and thwarted success (e.g., hostile work environments, limited access to resources and opportunities) for researchers minoritized due to their race, ethnicity, gender identity, sexual orientation, and other aspects of their identities (e.g., Berhe et al., 2022). In terms of topical areas, climate research is one where such inequities exist in production as well as implementation for impact. For example in the global context, the majority of climate change research is conducted within the developed world (Tai and Robinson, 2018), even though it is well established that climate impacts less developed countries disproportionately more (IPCC, Allen et al., 2014). The latest Intergovernmental Panel on Climate Change report further documents the miniscule amount of funding available for climate-related research in Africa, despite worsening impacts of a warmer climate in the continent such as biodiversity loss, droughts, reduced crop productivity, and economic growth (Trisos et al., 2022, Chapter 9, pp. 9–18).

Poor representation in science perpetuates environmental injustice ("a situation in which a specific social group is disproportionately affected by environmental hazards") (Brulle and Pellow, 2006). A well-known example of environmental injustice is the drinking water crisis in Flint, Michigan, which the ESIP Community Data cluster (Diggs et al., 2021) examined as a case study for the role of Earth Science data in environmental justice. As a predominantly Black and socioeconomically depressed community, Flint represents one of the most

disadvantaged and marginalized groups in our society and in the scientific process. The failure of the government to provide adequate drinking water protections resulted in disastrous health impacts and further erosion of trust between communities of color and the local, state, and federal U.S. government. With underrepresentation in the regulatory decision making process, including any advocates for them, it was extremely challenging for the Flint community to convince governmental agencies to address water contamination issues, despite the alarm being raised within the community (Butler et al., 2016). After attention was drawn to the issue by the U.S. Environmental Protection Agency (EPA), further deception of the EPA by state regulators eventually led to criminal charges (Butler et al., 2016). It also became evident that data collection protocols to test and monitor drinking water were willfully ignored or misinterpreted–not only in Flint but in other underserved communities in the US as well (Balazs and Ray, 2014; Katner et al., 2016). The Flint example highlights how scientists supporting government decision-making must recognize and mitigate the challenges that exist in applying the scientific process in communities striving for environmental justice.

Fundamental issues regarding ownership of and access to scientific data and information exacerbate inequity in the scientific process by limiting who can use data and information. Despite the fact that many scientific datasets are at least partially funded by place-based communities (e.g., *via* taxes), much of the scientific literature that reports Earth science research findings is published in journals that are inaccessible to the public or behind paywalls that are insurmountable by marginalized communities and nations. Open licensing presents its own unique set of challenges regarding the ownership of federally funded research data (Khayyat and Bannister, 2015). Moreover, even when the data and literature are available, it isn't always clear that results are easily and broadly usable by communities. While progress toward public access to data continues to be incremental, the equitable, technical, and legal challenges associated with data inhibit the societal benefit of science.

Social structures inherent in science create rigidity in the scientific process that inhibits the due consideration of stake and rights holder input, and the equitable distribution of benefits from science. These are especially pronounced given the social inequities that exist in society, and the fact that science has significant barriers to overcome with its implicit biases, including learning to see how inequities show up within the scientific establishment (Tanner, 2009). The question is, who do scientists take an oath to when performing their duties? Are they serving science and the continuation of its norms and practices? Community resilience applications offer an option to help scientists understand who science is being performed for, so the due representation of disadvantaged and historically underrepresented communities in the scientific process can be improved. Only by adapting the scientific process to the community context, can science address these inequities.

Gaps in Data Ethics and Governance

Improved community resilience relies on data that authentically represents the specific context and needs of a place-based community. Collecting such data, however, can be fraught and contested if people are not able to control their data and trust how it will be used. In this section, we focus on the ethical lapses that occur when managing, using, and reusing data products. Data governance entails formalizing ethical processes to ensure data are correctly managed after collection–this can include data preparation, maintenance, and security (Thompson et al., 2015).

In the community resilience context, data governance can include community members designing restrictions on data collection and use, and assigning responsibility for collecting, maintaining, and protecting data in alignment with their cultural identity. This is particularly crucial for indigenous peoples and other racial and ethnic minority groups and the institutions within which they are embedded (Smith, 2016). The ability of groups to control their own data has been defined as part of the right to self-determination as outlined in the United Nations Declaration on the Rights of Indigenous Peoples (UN General Assembly, 2007). The authority to control data collection and use are also components of more recent CARE Principles for Indigenous Data Governance (Carroll et al., 2020).

Data reuse can be ethically problematic. Interpretation and reuse of data by third parties can be harmful if such use is uninformed by the scientific and cultural context or selective in its analysis of available data (Reimsbach-Kounatze, 2021). Many datasets lack the information needed to inform data users about how data should or should not be used. For example, the U.S. Bureau of Land Management (BLM) collected oral histories from indigenous elders in Alaska Arctic boroughs in order to establish indigenous rights to federal lands. The BLM controlled access to the data for many years despite it being a valuable source of cultural information for the associated communities (Pratt, 2004).

Systematic data collection from disenfranchized communities can be intertwined with discrimination, reinforcing narratives that may be detrimental to the communities themselves. In addition, while data on their own may not disclose sensitive information, they often can be combined with other data to unintentionally or intentionally reveal sensitive information about vulnerable populations. A broadly cited example is the re-identification of individuals' names and addresses by piecing together multiple de-identified environmental quality datasets from a public health study (Sweeney et al., 2017). The availability of spatial data can similarly present a problem, such as when personal information is unintentionally re-identified through a geographic information system (GIS; Scassa, 2010).

Existing global inequality is perpetuated by data injustices through surveillance, economic exploitation, algorithmic profiling, and loss of the right to privacy (Heeks and Renken, 2018). Globally, approximately 19% of countries have no data protection laws, leaving their residents acutely vulnerable to personal security breaches and algorithmic prejudice (UNCTAD DPR., 2016; UNCTAD, 2020). Refugees are even more vulnerable. They lack access, control, and protection of their data, since they may not have rights within an asylum country to their data or be defended by the country they leave (Rolan et al., 2020). As the collection, storage, sharing, and use of Earth science data are increasingly facilitated by technological

advances, ethical advances must also keep pace to ensure that the interests of vulnerable communities are adequately protected.

Despite recent interest in co-production of scientific knowledge between producers and users of science (e.g., Lemos et al., 2018; Jagannathan et al., 2020), the literature remains largely theoretical without addressing the specific challenges we have described in this section (e.g., the step-by-step process of co-development and relationship). While there are promising trends to be more inclusive of communities in co-developed climate services (i.e., customized products such as forecasts and predictions), climate adaptation projects, environmental decision making, and environmental sustainability projects (Kirchhoff et al., 2013; Laursen et al., 2018; Bremer et al., 2019; Mach et al., 2020), many projects still fail to authentically address the inequitable power dynamics that are inherent in the processes of collaboration. For initiatives aiming to enhance community resilience in response to external events, including but not limited to climate-related damage, the gap in skills, capacities, and awareness needed for data producers to build trust with users to enable co-creation continues to be a limitation.

A Mismatch of Scale and Focus

Earth science research and place-based communities often work on different time and spatial scales, as well as have differing needs and goals for data collection. Community resilience decision-makers need data to make urgent and consequential decisions, while addressing multiple spatial and temporal scales (**Table 1**) and the needs of diverse stakeholders. An example of such tradeoffs are near- and long-term economic stability, or a city vs. an individual's exposure to climate risks (Chelleri et al., 2015).

Earth science data products that are needed for community resilience planning (e.g., climate projections, Earth systems modeling) are often produced at vast spatial and temporal scales, which are relevant to the understanding of long-term natural phenomena (Kirchmeier-Young et al., 2019). Yet, communities and industries often require data products that incorporate long-term trends with more actionable near-term, higher accuracy data (Vera et al., 2010; Dunn et al., 2015). Therefore, additional research and processing of Earth science data are needed to make these larger-scaled data more appropriate for community-level resilience needs (Bhuvandas et al., 2014), but many communities lack such data processing capacity (O'Neill, 2011). Without correctly scaling data and information, decision-makers may be taking (or not taking) actions that ultimately reduce their community's resilience. For example, Earth science phenomena described at a coarse scale (i.e., large areas represented as uniform depicted with lower resolution) may have directionally opposite effects at a local scale (Keskitalo et al., 2016). The lack of appropriately scaled Earth science data can limit its use to community resilience efforts for larger geopolitical units (such as nations), while local communities are left out of such efforts.

While the previously described issue of downscaling is well-addressed in literature, there are other kinds of scale mismatches. One is illustrated by the 2012 U.S. National Science Foundation (NSF) sponsored 2nd Semantic Sea Ice Interoperability Initiative (SIII)—a workshop that brought together Alaskan indigenous sea ice experts, National Oceanic and Atmospheric Administration (NOAA) specialists, and researchers around the topic of sea ice. Some members of indigenous communities of the Arctic are sea ice experts. However, their ability to understand and predict sea-ice behavior has been becoming increasingly strained by climate change. For years, NOAA had been providing sea ice charts, which provide a low resolution but comprehensive picture of sea ice in the Arctic. These charts are based on a variety of much higher resolution data, including Synthetic Aperture RADAR (SAR) full-resolution imagery; imagery that is too large to be transmitted to indigenous coastal villages given their available bandwidth. During the workshop, members of the indigenous community suggested producing a cropped SAR data product covering only the sea ice within a few miles of their community. Such a product would be useful for making decisions about travel and food harvesting. Existing ice charts and SAR products extend for 100 miles or more, far outside the bounds that a typical hunter would travel, thus providing many times more data than are useful at the community level. A cropped data product, however, would fit within the bandwidth limitations of the communities and could be directly compared to what community members could see from shore and on shore fast ice; an insight that NOAA data managers appreciated and acted upon.

Scientific research is often organized as individual research projects (i.e., principal investigator-led projects) that range from global efforts to local groups focusing on a single domain area. While there is an assumption that the results of different research projects will naturally inform each other, the reality is that these connections often fail to form, resulting in a splintered scientific approach that lacks place-based synthesis and fails to address emerging disasters (Cutter et al., 2013). For example, scientists studying rare earthquakes and floods do not automatically integrate their research with those studying modern disaster response or engineering cities (Ismail-Zadeh et al., 2016). In addition, the relationships between the projects may be too weak to support upward scaling of their goals and extrapolation from their results (Taylor, 1984; Parsons et al., 2011). Data products may neglect cross-scale relationships and address communities as if they are stand-alone and isolated entities (Sharifi, 2016), or provide sweeping results that are not fine-grained enough for tangible and useful decision-making. In contrast, information useful for community resilience must be well-coordinated across various scales to accurately assess risk, illuminate knowledge gaps and solutions, and communicate hazards. The information available to decision-makers depends on the strength of the interactions between data creators operating at multiple scales (Pulsifer et al., 2020). Better interaction across scales of this "information ecosystem" can improve data for interdisciplinary, systemic research (Parsons et al., 2011), such as for community resilience.

Lack of Actionable Information for Communities

Earth science data are often created for a particular science community with the expectation that others will find a way to use them (Baker et al., 2015). The lack of attention given to the nuanced needs and worldviews of specific groups that could

use Earth science data contributes to this challenge (Bhargava and Manoli, 2015). For example, different communities use different terminology, which impacts how information can reach and influence people, and in turn impacts their knowledge development (Eitzel et al., 2017). Further, existing power dynamics in knowledge generation often undermine data and knowledge that originate from outside of conventional scientific frameworks, such as among traditional and tacit knowledge systems (Brun and Schumacher, 1987; Roux et al., 2006; Dunlop, 2009). Often this issue is framed as a lack of data literacy, which implies that the solution involves improving a community's ability to read, work with, analyze, and argue with data (D'Ignazio and Bhargava, 2016). The problem with this framing is that all of the responsibility for "learning" about data is placed on community members, rather than Earth science researchers making efforts to provide data and information that are useful for those communities.

The application of Earth science data toward societal benefit is often conceptualized as a data, information, knowledge, and wisdom (DIKW) pathway (Ackoff, 1989; Sharma, 2008). This DIKW pathway emphasizes the one-directionality of data to application. An oversimplified and idealized version of the climate change discourse illustrates an example of the DIKW pathway from data to societal benefit: (1) Earth scientists notice trends and describe climate change to their peers as part of their typical scientific discourse, (2) scientists communicate technical information about climate change outside their disciplines, and their implications become evident in closely allied fields, (3) climate change begins to be recognized broadly across a number of disciplines and by a subset of the general public, and (4) climate change is now a central part of public policy discourse with analysis occurring cross-sectorally. There are many assumptions within this process, such as the trickling effect of scientific knowledge into the general public. However, it serves as a useful example of how scientists might envision how their data efforts contribute to societal benefit through the DIKW pathway.

Figure 1 depicts the data equivalent of such a DIKW pathway, where source data is processed through a number of steps into intermediate data products that become publicly accessible final data products to be interpreted for community consumption. Even with this simple, linear model it is clear that unless all of the products along the path are continuously funded and available, neither the final data products nor their interpretations will be available for any community to consume.

An unexplored assumption behind long-term investments in data management, and in today's policy and decision making context, is that the DIKW pathway justifies the initial investment and creation of structured data, for example, climate data records (Meier et al., 2021), even when existing knowledge frameworks are too rigid to address today's pressing needs and priorities. Indeed, the example of the climate change information pathway demonstrates how much easier it is to address an Earth science problem from one worldview, rather than taking on the challenge of considering the complex and nuanced perspectives of multiple user communities. The challenge here is to present more accurate conceptual frameworks that demonstrate the role of users (community) in the data pathway.

In reality, the Earth science data "pathway" is often more like a "network system" (Li and Whalley, 2002) with multiple intended products stemming from data sets or combined data sets, and people who play multiple roles on the path and in the network. **Figure 2** depicts a fragment of an existing data product network currently available at the National Snow and Ice Data Center (NSIDC). Data products that are tailored to support a particular user group's needs have potential for greater use and applicability by that user group, in contrast to data products that are developed without a specific end user in mind. However, creation of such downstream data and interpretive products for end users can be challenging. Baker et al. (2015) describe a situation where data products created by a specific Earth science community (e.g., the sea ice scientific community) were confusing and unhelpful for various groups outside of that community, leading to a process of data development that continued for more than a dozen years.

RECOMMENDATIONS TO ENHANCE COMMUNITY RESILIENCE

To address the challenges identified in the previous section, this section describes four categories of recommendations to help improve how Earth Science enhances community resilience:

- Integrate community into the scientific data pathway,
- Build capacity to bridge science and place-based community needs,
- Balance openness with self-governance, and
- Improve access to data tools to support community resilience.

These recommendations present elements of decentralized, transdisciplinary, and systems thinking, which are needed to address the complex social-ecological systems that underlie community resilience. **Figure 3** provides a summary of how the recommendations (section Recommendations to Enhance Community Resilience) map to the challenges presented previously in this paper (section Data Challenges in Community Resilience).

Integrate Community Into the Scientific Data Pathway

Develop new conceptual frameworks that incorporate all relevant participants within data usage pathways for societal benefit (e.g., communities, scientists, data managers, analysts, translators, consultants, science communicators, non-profit groups, and other intermediaries).

Earth science applications affect people, the world that we live in, and the resilience of our communities. For this reason, Earth science project design and development, including data creation and curation, should be inclusive and representative of all relevant perspectives, values, and needs. It is challenging to consider the needs of different stakeholders and participants present in a community context, but by doing so, Earth science projects can be designed to ensure that the groups most vulnerable to adverse effects of climate and other environmental events are fairly represented in scientific processes. Progress to that end involves co-production within the conceptual

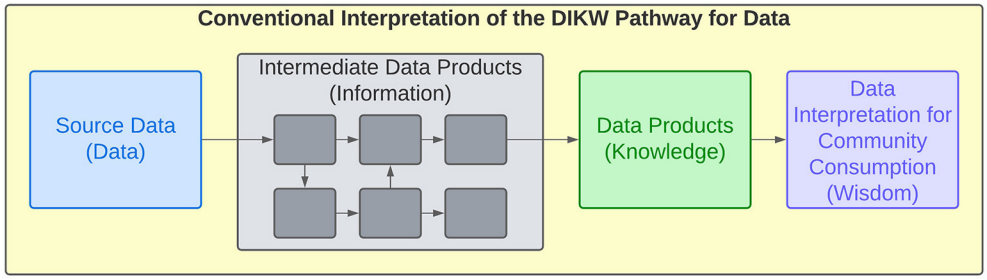


FIGURE 1 | The linear equivalent of the DIKW pathway for data.

framing of the Earth science data pathway for community resilience. Such a process encourages more interaction between scientific information and the applied context, helping to overcome the entrenched knowledge systems that support one-way linear processes where science disseminates information and knowledge as a commodity.

While programs such as ELOKA have begun to facilitate these processes (in ELOKA's case for Arctic communities), more long-term programs that facilitate co-production in individual research activities for a wide variety of community contexts are needed (Pulsifer et al., 2012). Similarly, large agency missions should explicitly consider the entire range of communities that could potentially find utility in the agency data produced and should design their initial data products accordingly. In general, it can be expected that different products will be needed for each kind of community, so such products should be developed in conjunction with those communities. Moreover, agencies should expect that as conditions change over time, new needs and uses for the data will be found. Consequently, ongoing resources should be allocated to support co-development of useful, new products as they are identified. NASA's long-standing Advancing Collaborative Connections for Earth System Science (ACCESS) program is a prototype of such a program though focused primarily on science community needs (Ramapriyan and Murphy, 2017). Their Earth Science Applications programs that call for projects in specific topic areas, such as health and air quality, use Earth observations to improve decision-making and service to the public.

Several ESIP clusters contribute in this area. For example, the long-standing Agriculture and Climate cluster is working with county and agency fire chiefs and managers on improved data for dealing with wildfires and their aftermaths. A nascent Space Weather group, whose members have worked with national grid providers, are discussing how best to represent space weather impacts on the electrical grid. The Community Data cluster also grappled with how to better include community perspectives in the analysis of data for environmental justice purposes.

Community resilience projects are diverse, and their needs for different co-production processes can vary. Such contextual variation is considered in typologies of community participation (Cornwall, 2008) that range from a more passive information-seeking approach to a more highly interactive transdisciplinary approach, where users are more deeply involved in developing the goals, methods, and analyses of a project. Similarly Meadow et al. (2015) describe strategic co-development of science knowledge through four increasingly cooperative modes of engagement ranging from contractual, consultative, collaborative, to collegial that reflect one-directional flow of information to more shared exchange of different forms of science knowledge. Their examples of conducting action research incorporating social science approaches suggests an openness to transdisciplinary learning that is critical to advance community resilience.

An often used type of co-production approach in science is citizen science, or "community science," which centers around involving members of the public in the scientific process to both benefit the scientific process and the involved community members (Craglia and Shanley, 2015). Areas of convergence where science and communities can build beneficial co-productive relationships include motivations to benefit society, social location (i.e., the user groups that must be involved to succeed at change), and ethics to support inclusive knowledge generation (Jull et al., 2017).

Co-production in Earth science occurs as scientific data and information interfaces with existing knowledge and wisdom from communities to support decision-making for community resilience. For such a model to work, mutual respect between scientists and community members, as well as a more pluralistic perspective of research expertise is needed to provide an important starting point for successful co-production. This can look like evidence-building activities that incorporate scientific research findings and data (e.g., as expressed in the Foundations of Evidence Based Policy-making Act of 2018) with those based on traditional knowledge systems and tacit place-based knowledge (Kendall et al., 2017; Rainie et al., 2017). Strategic reframing is one approach for conflict resolution in environmental management decisions that allows government agencies to adapt to different stake and rights holder perspectives and consider different scenarios (Auad et al., 2018). Such integration of world views helps to build trust around science, while also ensuring that scientific information is relevant and useful within real-world contexts. The goal of these processes is for Earth scientists to become more sensitive to the historical colonial context of science that influences and colors their assumptions, approaches, implicit biases, and applications within their work (D'Ignazio and Klein, 2020).

FIGURE 2 | Each bubble represents a data product, most of which are publicly available from NSIDC. Orange products are raw data; while dark blue indicates data external to NSIDC used in creating NSIDC data products. Light blue indicates science products available from NSIDC; while Green indicates products that are NSIDC terms "Easy-to-use". In NSIDC's view these data products are in "formats that require little or no processing or programming. These may be of particular interest to K-12 teachers and students, the press, the general public, or non-cryospheric researchers." (NSIDC, 2022). Purple represents science products (blogs) also available from the NSIDC homepage that provide data interpretation for broad audiences. Figure adapted and updated from Baker et al. (2015).

Build Capacity to Bridge Science and Place-Based Community Needs

Provide more skills development for scientists, community members, and other participants within the data usage pathway to help bridge gaps and develop intermediaries.

Training and professional development for scientists can help them engage more meaningfully with communities and better understand the real-world context that their work applies to. While many scientists are unlikely to interact directly with communities, some do interact with community members that use and interface with data based on their unique skills and to fulfill different roles (as intermediaries). These include: *Communicators* (who make sense of and tell stories about data for others to digest), *Readers* (who need skills to interpret data),

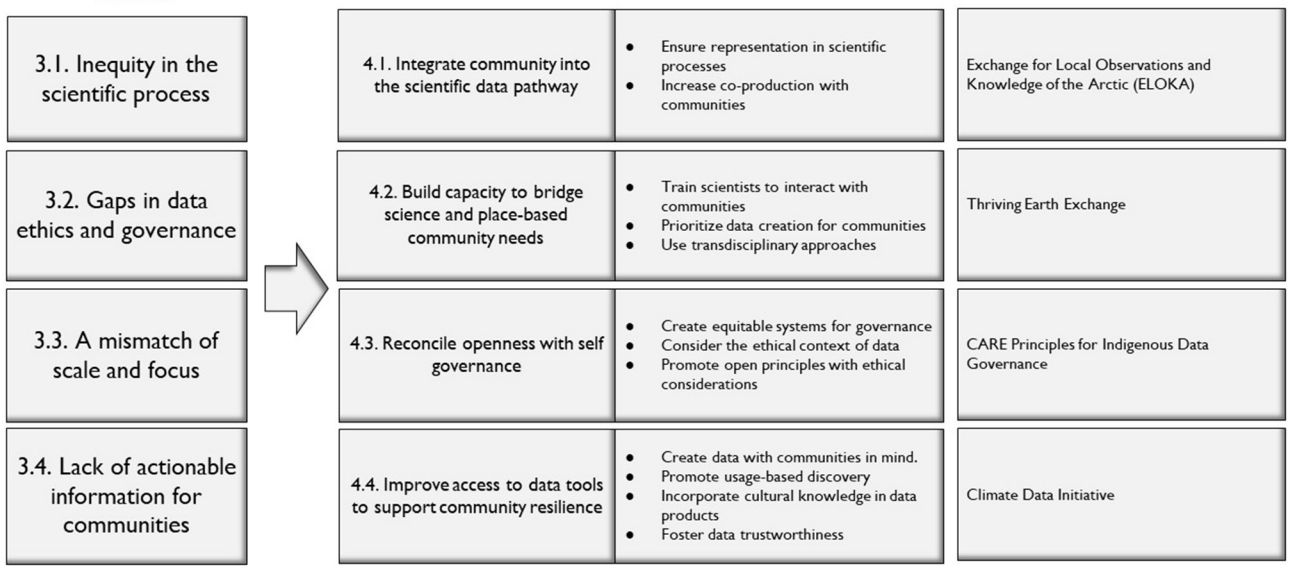


FIGURE 3 | Recommendations to the Earth sciences as to how changes in Earth science can enhance community resilience, stemming from the challenges described in this paper.

Makers (who need the skills to use data for problem solving), and *Scientists* (who are knowledgeable about the data domain and need to leverage strong technical data use and communication skills) (Wolff et al., 2016). Community members with technical skills (e.g., *Scientists* as described above) can be valuable brokers, representing community needs in collaboration with the Earth science data community. The Pacific Islands Climate Adaptation Science Center's University of Hawai'i's Manager Climate Corps program (Laursen et al., 2018) demonstrates the value of a careful needs assessment as researchers embarked on a collaboration with natural and cultural resource managers to co-create climate adaptation strategies–their processes of harnessing the knowledge and experiential capacities of local experts through close interpersonal engagement helped establish relationships that can jointly work toward social-ecological change.

Recent trends in data science education are promising for creating a new generation of application-focused data scientists (e.g., Irizarry, 2020). Scholars have advocated for multiple strategies and new curricular foci to train students to effectively use data to understand and tackle societal problems–these suggestions include creating opportunities for students to actively engage with real-world projects, collaborate across disciplines, and across sectors like industry and government (Song and Zhu, 2016); learn to interact with multivariate phenomena that depict the complexity and interconnectedness of variables underlying societal issues (Engel, 2017); and inculcate "habits of mind" (p. 5) in practice and theory that encourage critical thinking, inquiry-based reasoning, and problem solving (Finzer, 2013). Greater attention to skills-based education is a start to help data scientists develop data products for community members in a manner that prioritizes community members' use (rather than Earth scientists' use). However, the burden should not fall on communities to defend themselves with data, and the education of data skills and capability to work with data should not be isolated to the privileged. This would help communities overcome barriers in access and use of available information to advocate on their own behalf.

To address the problems of scale and community data needs that we identify in our challenges, new research pathways can be created by pursuing transdisciplinary approaches based on communities of practice (Wenger-Trayner and Wenger-Trayner, 2015). Such communities of practice would include Earth science and Community Resilience practitioners, who are brought together by common needs and goals. These communities of practice would act as transdisciplinary teams to "work jointly to grasp the complexity of problems from diverse scientific and societal perspectives, integrate natural and social science disciplines, alter discipline-specific approaches, and focus on problem-solving for what is perceived to be the common good" (Yates et al., 2015). These teams would require long-term and stable co-operation between Earth Scientists and the community. Working together would help develop translators or intermediaries between different stakeholders in the information pathway, such as translating between communities to understand information needs and what data is appropriate for resilience planning. As an example, the Bureau of Ocean Energy Management, a US federal agency, could integrate a multi-scale resilience framework into its

scientific research and management enterprise in order for science to more clearly articulate and navigate the multi-scale dynamics of social-ecological systems (Auad et al., 2018). Building connections between Earth science and communities may benefit from intermediary organizations and initiatives, like the Thriving Earth Galkiewicz and Pandya (2014) which can help identify communities with Earth science challenges and facilitate interactions with Earth science experts.

Reconcile Openness With Self-Governance

Open science reduces barriers to information while increasing governance & agency over data to protect the interests of people and communities that are subjects of data.

We articulated two problems concerning access to data. The first is the limited access to relevant data by those outside the Earth science community. The second problem is marginalized communities' limited agency over data use. While these may seem to have different solutions, we argue that both problems require the same solution: creating equitable systems for data products, infrastructure, and governance. Such management of Earth science data is not only scientific, but also dependent on societal norms and cultural best practices. Yet transformative shifts in norms of practice or systems change are inherently challenging as highlighted by Jagannathan et al. (2020) in their analysis of outcomes of knowledge co-production processes observed in practice compared to those theorized. However, initiatives are already in place to change data systems that enable more transparency and access to data and will hopefully spur additional efforts. For example, the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016) is an example of one solution that has been offered to help increase community members' access to data.

Additionally, there have been some good examples of how the consideration of the ethical context around data collection, governance, and use has occured at different scales that affect the Earth science data community. The 2019 decadal U.S. Federal Data Strategy emphasizes ethical governance, conscious design, and learning culture to "continually challenge and guide agencies, practitioners, and policymakers to improve the government's approach to data stewardship and the leveraging of data to create value" (United States Government et al., 2019). Recent findings from an open forum co-hosted by the Data Coalition and the Data Foundation concluded that future Federal Action Plans should emphasize "equity and inclusion, the importance of sharing data in a way that protects and respects privacy, and prioritizes transparency and openness" (Turbes, 2020). The American Geophysical Union (AGU) has also published ethics questions for practicing scientists to consider during the data life cycle (Gundersen, 2017) and the AGU Position Statement on Data states "all players in the science ecosystem should ensure … that relevant scientific evidence is processed, shared, and used ethically…" (American Geophysical Union, 2019). Privacy concerns can involve the release of information by different entities, so it may benefit from control by an overarching governance body (Commission on Evidence-Based Policymaking, 2017).

Openness, without governance, might infringe upon the rights and privacy of people who are subjects of data, but the CARE principles present an example of an ethical bookend to FAIR. The CARE principles were formulated to further the self-determination of indigenous people against the ongoing process of colonial oppression and exploitation of indigenous knowledge and data (Carroll et al., 2020). Similar principles could also be more broadly applied beyond the indigenous context and among other groups of people who are historically marginalized in society.

We note that principles alone will not resolve these issues. Principles must be carried over into practice. While several groups are working to translate FAIR principles into action as well as measure the FAIRness of research data (Bahim et al., 2020); practices and measures for the CARE principles are not as far along. However, it should be noted that the ESIP Sustainable Data Management cluster has been working with the Global Indigenous Data Alliance (GIDA) to define the responsibilities and actions data repositories should take to become more CARE-compliant (Global Indigenous Data Alliance, 2019). Moreover, work on translating the CARE principles into theory and action for researchers is forthcoming from IEEE (a project on "Recommended Practice for Provenance of Indigenous Peoples' Data" https://standards.ieee.org/ieee/2890/10318/) and Research Data Alliance (International Indigenous Data Sovereignty Interest Group, https://www.rd-alliance.org/groups/international-indigenous-data-sovereignty-ig).

Reconciling openness with governance offers the opportunity to be deliberative with incremental progress that carefully considers the ethics behind data use. Science and society has no other option but to reconcile these objectives, as further promotion of open principles along with due consideration for the ethical dimension are two critical components for maintaining the social contract (Lubchenco and Rapley, 2020).

Improve Access to Data Tools to Support Community Resilience

Reduce the burden for community resilience practitioners to discover and access Earth science data.

People in communities use a diverse set of sources to make decisions impacting their lives, including those that address problems that arise within the complex systems that they live in. Decisions are rapid and responsive to evolving information contexts, and may not be based in the sciences, including Earth sciences. In other words, in the absence of data that are specific to a problem, people use whatever information they have at hand. Decision-makers may base their decisions on data that have been reinterpreted within a personal or specific context. They might also seek out experts to help interpret the landscape of the Earth sciences and present contextualized information relevant to the

community's predicament. Secondary data sources that provide a level of interpretation, which people in communities consume as scientific products, can form the basis of people's world views. For example, the Climate Data Initiative brought together earth science datasets and reframed them in a global resilience context to enable broader consumption by practitioners (Sisco et al., 2019). In this derived data context, discovery of scientific data may occur when pre-digested and cited secondary scientific information is cataloged to allow people to back track to originating discoveries and datasets more easily, and challenge misconceptions that may inadvertently arise about the secondary sources. Cvitanovic et al., 2014 also recommends the creation of management-oriented summaries of research articles to describe the policy implications of research outcomes in publicly meaningful ways; many journals have instituted or are in the process of instituting the concept of plain language summaries.

ESIP's Discovery cluster emphasizes that making assumptions about data's future utility should be avoided. Instead, data discovery informed by the end-user's actual usage (Usage-based data discovery; Lynnes et al., 2020) can better align existing knowledge frameworks to re-prioritize research investments. As a result of this process, intended audiences and beneficiaries of data management labors can be clearly identified, so the relevance and value of highly technical endeavors can be better targeted (e.g., decision-makers who seek to support resilience in their community). Usage-based data discovery can further increase the utility of data by providing examples of how data have been used by other communities, while communicating the larger application context for the data. With the right processes in place, a feedback mechanism to data providers would also allow for iterative improvements.

To further this process, Earth science repositories can develop controlled vocabularies to tag datasets according to specific information needs (e.g., Semnacher and Chong, 2019). This allows federal agencies to enhance discovery and demonstrate the applicability of their research products to society and sustainable development goals (e.g., the Arctic report card by Starkweather et al., 2020). In these scenarios, data products (e.g., summaries, subsets, trends, or infographic representations) that are context- and culturally-specific are more easily utilized by decision-makers within their community (Baker et al., 2015). Correctly scaled data in a format usable by communities can interact with their cultural knowledge to produce data subsets and syntheses that directly improve community resilience. For example, city governments develop resilience plans that cover different sectors of the government, including energy, food, and urban infrastructure. These plans aim to bring together government, industry, community groups, and residents using a collaborative transdisciplinary effort. Earth scientists would be more closely involved in these efforts as the developers of data being used to guide decision-making. Data are often used by communities *post-hoc*, so the purpose of data and their relevant metadata need to be carefully and clearly articulated so they can be used appropriately.

Information about the reliability and trustworthiness of data should also be provided. A basic overview of the database, including how it was developed, who was involved, and the audiences for whom it was intended are some basic metadata that are fundamental to improving the usability of data. For example, websites displaying analyses of Earth science can provide a level of transparency to understand data sources and versions of data used in the analytical products (Lynnes et al., 2020). The Earth sciences could provide data that have verifiable data trust measures as a means to pre-assess the quality of data for applications (see example below). While there will always be inherent uncertainty in Earth science data products, certain datasets are categorically wrong for certain applications, and an established and transparent evaluation process would help mitigate untrained users from missapplying data (Ekstrom et al., 2015; Nissan et al., 2019) and clearly communicate the uncertainty within the dataset. This would enable increased trust and reliability in external data sources by users of this data in community contexts. Recent guidelines around the trustworthiness of data can help the scientific community (Jamieson et al., 2019) and repositories (Lin et al., 2020) with their assessment of datasets. Trusted data are highly reusable, broadly applicable, of verified quality, and clarify the source of the data and its intended applicability. Data trust assessments should be provided to datasets that meet transparent guidelines or rules on data maintenance, governance, and storage that emphasize access, accountability, and long-term management. This is another area where ESIP clusters have been actively working. For example, the Information Quality Cluster recently published a paper calling for the development of practical guidelines for representing and sharing data quality information (Peng et al., 2021).

In another example, Operational Readiness Levels (ORL) developed within the ESIP Disaster Lifecyle cluster provide a ranking by which to provide decision-makers with an understanding of the operational reliability of datasets using predefined criteria (case study provided by Hicks in the ESIP webinar Moe et al., 2018). This is an example of how credentials can allow for diverse datasets to be accessed and used in decision-making, while still accounting for the level of uncertainty inherent to a given data set. This can drive community decision-making by forwarding value-neutral information as to how complete or ready a data set is. Similar ORLs could be developed for community planning data, such as climate/weather/natural disaster forecasts.

CONCLUSION

In this paper, we described the challenges of applying Earth science data to community resilience decision-making. *Inequity in the scientific process* presents challenges in identifying solutions that are both innovative and reflective of community needs. *Gaps in data ethics and governance* highlight the need to be aware of how misuse of data can negatively impact people who are subjects of that data. *A mismatch of scale and focus* emphasizes the importance of how data must address the different sociopolitical, temporal, and geographical scales that are relevant to place-based community resilience. *Lack of actionable information for communities*

speaks to the gaps between the types of data products that are produced and the expectations of communities to use them.

We encourage the Earth science data community to develop practices for improving how Earth science data and analyses are disseminated to and used for community resilience purposes. To that end, we have suggested the following starting points:

- *Integrate community into the scientific data pathway* to emphasize the importance of re-framing how we think about data literacy, and re-conceptualize the DIKW pathway to be more inclusive of community world views.
- *Build capacity to bridge science and place-based needs* to help identify opportunities for skills development among both scientists and community decision-makers to reduce the disconnect between the groups.
- *Reconcile openness with self-governance* to create more ethical and equitable systems of data products, infrastructure, and governance.
- *Improve access to data tools to support community resilience* by prioritizing usage-based discovery to ensure data can better meet the needs of communities.

The framing that we present here can help organizations like ESIP mobilize Earth science data scientists and practitioners to develop innovative solutions to Earth science data challenges, and affect durable and effective change with the most impactful benefit across society. Such organizations have the capacity and connections to integrate the suggestions that we offer within the Earth science ecosystem. Through our work in ESIP's Community Resilience cluster and in collaboration with other ESIP clusters, we see the potential to create spaces within Earth science that better support place-based community needs.

Contributing to these Earth science data opportunities is our professional responsibility to the communities that we are a part of, since as scientists we have access to resources, skills, and tools that many in our society do not. Without more attention given to bridging the gaps between Earth science and community resilience, we risk continuing to exclude and marginalize the most vulnerable place-based communities in our world, and continue to waste scarce resources on producing data and investing in scientific initiatives that do not meet the urgent needs of our society. By working in concert with communities, Earth Scientists can contribute to making the systemic changes needed to help overcome some of the biggest challenges of our generation.

AUTHOR CONTRIBUTIONS

AV coordinated meetings and drafts, developed initial concept of the paper, contributed original writing, contributed revisions, acted as corresponding author, and submitted the article. RG, JB, and RD developed initial concept of the paper, contributed original writing, and contributed revisions. ZR contributed original writing, contributed revisions, formatted and completed citations, and drafted and completed figure. CG contributed revisions, formatted citations, formatted paper, and drafted and completed figure. All authors contributed to the article and approved the submitted version.

FUNDING

CG and ZR were supported through the Earth Science Information Partners (ESIP) Community Fellowship in 2021 and 2019–2020, respectively. Funding for open access publication was provided by Earth Science Information Partners (ESIP).

ACKNOWLEDGMENTS

We would like to thank the participants of previous community resilience-themed ESIP meeting sessions, particularly the Summer Meeting 2015 in Monterey, CA; Winter Meeting January 2017 in Bethesda, MD; Winter Meeting January 2018 in Bethesda, MD; Winter Meeting 2019 in Bethesda, MD; Winter Meeting 2021 (virtual); and Summer Meeting 2021 (virtual). We acknowledge Steve Young and the expertise he provided on Environmental Protection Agency (EPA) regulations and the environmental justice issue in Flint, Michigan. We thank Drs. Lindsay Barbieri, Brian Wee, and Christine White who were co-authors of a synthesis report entitled Community Resilience: Demonstrating the socioeconomic value of Earth Science data from ESIP Winter Meeting 2018 in Bethesda, MD that provided inspiration for this paper. Two reviewers also provided helpful comments and suggestions that improved the paper. The views expressed in this article do not necessarily represent the views of the Bureau of Ocean Energy Management, the Department of the Interior, or the United States.

REFERENCES

Ackoff, R. L. (1989). From data to wisdom. *J. Appl. Syst. Anal.* 16, 3–9.

Adger, W. N., Hughes, T. P., Folke, C., Carpenter, S. R., and Rockström, J. (2005). Social-ecological resilience to coastal disasters. *Science* 309, 1036–1039. doi: 10.1126/science.1112122

Aldrich, D. P., and Meyer, M. A. (2015). Social capital and community resilience. *Am. Behav. Scientist* 59, 254–269. doi: 10.1177/0002764214550299

Allen, M. R., Pachauri, R. K., Barros, V. R., Broome, J., Cramer, W., Christ, R., et al. (2014). "Climate change 2014: synthesis report," in *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds R. K. Pachauri and L. Meyer (EPIC3Geneva; IPCC), 151. Available online at: https://epic.awi.de/id/eprint/37530/

American Geophysical Union (2019). *AGU Position Statement on Earth and Space Science Data. Poistion Statment on Data.* Available online at: https://www.agu.org/Share-and-Advocate/Share/Policymakers/Position-Statements/Position_Data

Auad, G., Blythe, J., Coffman, K., and Fath, B. D. (2018). A dynamic management framework for socio-ecological system stewardship: a case study for the United States Bureau of Ocean Energy Management. *J. Environ. Manage.* 225, 32–45. doi: 10.1016/j.jenvman.2018.07.078

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., et al. (2020). The FAIR data maturity model: an approach to harmonise FAIR assessments. *Data Sci. J.* 19, 41. doi: 10.5334/dsj-2020-041

Baker, K. S., Duerr, R. E., and Parsons, M. A. (2015). Scientific knowledge mobilization: co-evolution of data products and designated communities. *Int. J. Digital Curation* 10, 110–135. doi: 10.2218/ijdc.v10i2.346

Balazs, C. L., and Ray, I. (2014). The drinking water disparities framework: on the origins and persistence of inequities in exposure. *Am. J. Public Health* 104, 603–611. doi: 10.2105/AJPH.2013.301664

Bautista, E., Hanhardt, E., Osorio, J. C., and Dwyer, N. (2015). New York City environmental justice alliance waterfront justice project. *Local Environ.* 20, 664–682. doi: 10.1080/13549839.2014.949644

Berhe, A. A., Barnes, R. T., Hastings, M. G., Mattheis, A., Schneider, B., Williams, B. M., et al. (2022). Scientists from historically excluded groups face a hostile obstacle course. *Nat. Geosci.* 15, 2–4. doi: 10.1038/s41561-021-00868-0

Bhargava, S., and Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: evidence from an IRS field experiment. *Am. Econ. Rev.* 105, 3489–3529. doi: 10.1257/aer.20121493

Bhuvandas, N., Timbadiya, P. V., Patel, P. L., and Porey, P. D. (2014). Review of downscaling methods in climate change and their role in hydrological studies. *Int. J. Geol. Environ. Eng.* 8, 713–718.

Bone, C., Moseley, C., Vinyeta, K., and Bixler, R. P. (2016). Employing resilience in the United States Forest Service. *Land Use Policy* 52, 430–438. doi: 10.1016/j.landusepol.2016.01.003

Bremer, S., Wardekker, A., Dessai, S., Sobolowski, S., Slaattelid, R., and van der Sluijs, J. (2019). Toward a multi-faceted conception of co-production of climate services. *Climate Serv.* 13, 42–50. doi: 10.1016/j.cliser.2019.01.003

Brulle, R. J., and Pellow, D. N. (2006). Environmental justice: human health and environmental inequalities. *Annu. Rev. Public Health* 27, 103–124. doi: 10.1146/annurev.publhealth.27.021405.102124

Brun, V., and Schumacher, T. (1987). *Traditional Herbal Medicine in Northern Thailand*. Berkeley, CA: University of California Press.

Butler, L. J., Scammell, M. K., and Benson, E. B. (2016). The Flint, Michigan, water crisis: a case study in regulatory failure and environmental injustice. *Environ. Justice* 9, 93–97. doi: 10.1089/env.2016.0014

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., et al. (2020). The CARE principles for indigenous data governance. *Data Sci. J.* 19, 43. doi: 10.5334/dsj-2020-043

Chandra, A., Acosta, J., Howard, S., Uscher-Pines, L., Williams, M., Yeung, D., et al. (2011). Building community resilience to disasters. *Rand Health Q.* 1.

Chelleri, L., Waters, J. J., Olazabal, M., and Minucci, G. (2015). Resilience trade-offs: addressing multiple scales and temporal aspects of urban resilience. *Environ. Urban.* 27, 181–198. doi: 10.1177/0956247814550780

Commission on Evidence-Based Policymaking (2017). *The Promise of Evidence-Based Policymaking*. Bipartisan Policy Center. Available online at: https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2019/03/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Comission-on-Evidence-based-Policymaking.pdf

Cornwall, A. (2008). Unpacking 'participation': models, meanings and practices. *Commun. Dev. J.* 43, 269–283. doi: 10.1093/cdj/bsn010

Craglia, M., and Shanley, L. (2015). Data democracy - increased supply of geospatial information and expanded participatory processes in the production of data. *Int. J. Digital Earth* 8, 679–693. doi: 10.1080/17538947.2015.1008214

Cutter, S. L., Ahearn, J. A., Amadei, B., Crawford, P., Eide, E. A., Galloway, G. E., et al. (2013). Disaster resilience: a national imperative. *Environ. Sci. Policy Sustain. Dev.* 55, 25–29. doi: 10.1080/00139157.2013.768076

Cvitanovic, C., Fulton, C. J., Wilson, S. K., van Kerkhoff, L., Cripps, I. L., and Muthiga, N. (2014). Utility of primary scientific literature to environmental managers: an international case study on coral-dominated marine protected areas. *Ocean Coast. Manag.* 102, 72–78. doi: 10.1016/j.ocecoaman.2014.09.003

Diggs, S., Thomer, A., and McKenzie, M. (2021). *Community Data Cluster Ideation and Gap Analysis, Facilitated by the Community Data Cluster*. 2021 ESIP Winter Meeting. Available online at: https://2021esipwintermeeting.sched.com/event/g49o/community-data-cluster-ideation-and-gap-analysis

D'Ignazio, C., and Bhargava, R. (2016). DataBasic: design principles, tools and activities for data literacy learners. *J. Commun. Informatics* 12, 83–107. doi: 10.15353/joci.v12i3.3280

D'Ignazio, C., and Klein, L. F. (2020). *Data Feminism*. Cambridge, MA: MIT Press.

Dunlop, C. A. (2009). Policy transfer as learning - capturing variation in what decision-makers learn from epistemic communities. *Policy Stud.* 30, 289–311. doi: 10.1080/01442870902863869

Dunn, M. R., Lindesay, J. A., and Howden, M. (2015). Spatial and temporal scales of future climate information for climate change adaptation in viticulture: a case study of user needs in the Australian winegrape sector. *Aust. J. Grape Wine Res.* 21, 226–239. doi: 10.1111/ajgw.12138

Dutton, J., Bretherton, F. P., Jenne, R. L., Karin, S., Volansky, S., Webster, F., et al. (1995). *"The Earth Sciences Information System" (Appendix F) in A Review of the U.S. Global Change Research Program and NASA's Mission to Planet Earth/Earth Observing System*. National Research Council. Available online at: https://ntrs.nasa.gov/api/citations/19960016634/downloads/19960016634.pdf

Eitzel, M. V., Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., et al. (2017). Citizen science terminology matters: exploring key terms. *Citizen Sci. Theory Prac.* 2, 1–20. doi: 10.5334/cstp.96

Ekstrom, J. A., Suatoni, L., Cooley, S. R., Pendleton, L. H., Waldbusser, G. G., Cinner, J. E., et al. (2015). Vulnerability and adaptation of US shellfisheries to ocean acidification. *Nat. Clim. Chang.* 5, 207–214.

Engel, J. (2017). Statistical literacy for active citizenship: a call for data science education. *Stat. Educ. Res. J.* 16, 44–49. doi: 10.52041/serj.v16i1.213

Fath, B. D., Dean, C. A., and Katzmair, H. (2015). Navigating the adaptive cycle: an approach to managing the resilience of social systems. *Ecol. Soc.* 20. doi: 10.5751/ES-07467-200224

Finzer, W. (2013). The data science education dilemma. *Technol. Innov. Stat. Educ.* 7.doi: 10.5070/T572013891

Funfgeld, H., and McEvoy, D. (2012). Resilience as a useful concept for climate change adaptation? *Plann. Theory Prac.* 13, 324–328. doi: 10.1080/14649357.2012.677124

Funk, C. (2012). *Key Findings About Americans' Confidence in Science and Their Views on Scientists' Role in Society*. Pew Research Center. Available online at: https://www.pewresearch. org/fact-tank/2020/02/12/key-findings-about-americans-confidence-in-science-and-their-views-on-scientists-role-in-society

Galkiewicz, J., and Pandya, R. (2014). Meeting people where they are: Thriving earth exchange. *Eos Trans. Am. Geophys Union*. 95, 44–44. doi: 10.1002/2014EO050006

Global Indigenous Data Alliance (2019). *CARE Principles for Indigenous Data Governance*. GIDA. Available online at: https://www.Gida-Global.Org/Care

Green2.0 (2021). *2021 NGO & Foundation Transparency Report Card, Green2.0*. Available online at: https://diversegreen.org/transparency-cards/2021-green-2-0-ngo-foundation-transparency-report-card/

Gundersen, L. C. (ed.). (2017). *Scientific Integrity and Ethics in the Geosciences*. Hoboken, NJ: John Wiley & Sons, Inc.

Heeks, R., and Renken, J. (2018). Data justice for development: what would it mean? *Information Dev.* 34, 90–102. doi: 10.1177/0266666916678282

Higuera, P. E., Metcalf, A. L., Miller, C., Buma, B., McWethy, D. B., Metcalf, E. C., et al. (2019). Integrating subjective and objective dimensions of resilience in fire-prone landscapes. *Bioscience* 69, 379–388. doi: 10.1093/biosci/biz030

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Sci. Rev.* 2. doi: 10.1162/99608f92.dd363929

Ismail-Zadeh, A. T., Cutter, S. L., Takeuchi, K., and Paton, D. (2016). Forging a paradigm shift in disaster science. *Natural Hazards* 2, 969–988. doi: 10.1007/s11069-016-2726-x

Jagannathan, K., Arnott, J. C., Wyborn, C., Klenk, N., Mach, K. J., Moss, R. H., et al. (2020). Great expectations? Reconciling the aspiration, outcome, and possibility of co-production. *Curr. Opin. Environ. Sustain.* 42, 22–29. doi: 10.1016/j.cosust.2019.11.010

Jamieson, K. H., McNutt, M., Kiermer, V., and Sever, R. (2019). Signaling the trustworthiness of science. *Proc. Nat. Acad. Sci. U.S.A.* 116, 19231–19236. doi: 10.1073/pnas.1913039116

Jull, J., Giles, A., and Graham, I. D. (2017). Community-based participatory research and integrated knowledge translation: advancing the co-creation of knowledge. *Implement. Sci.* 12, 150. doi: 10.1186/s13012-017-0696-3

Katner, A., Pieper, K. J., Lambrinidou, Y., Brown, K., Hu, C. Y., Mielke, H. W., et al. (2016). Weaknesses in federal drinking water regulations and public health policies that impede lead poisoning prevention and environmental justice. *Environ. Justice* 9, 109–117. doi: 10.1089/env.2016.0012

Kendall, J. J. K. Jr., Brooks, J. J., Campbell, C., Wedemeyer, K. L., Coon, C. C., Warren, S. E., et al. (2017). Use of traditional knowledge by the United States Bureau of Ocean Energy Management to support resource management. *Czech Polar Rep.* 7, 151–163. doi: 10.5817/CPR2017-2-15

Keskitalo, E. C. H., Horstkotte, T., Kivinen, S., Forbes, B., and Käyhkö, J. (2016). "Generality of mis-fit"? The real-life difficulty of matching scales in an interconnected world. *Ambio* 45, 742–752. doi: 10.1007/s13280-015-0757-2

Khayyat, M., and Bannister, F. (2015). Open data licensing: more than meets the eye. *Information Polity* 20, 231–252. doi: 10.3233/IP-150357

King, C. A. (2008). Community resilience and contemporary agri-ecological systems: Reconnecting people and food, and people with people. *Syst. Res. Behav. Sci.* 25, 111–124. doi: 10.1002/sres.854

Kirchhoff, C. J., Carmen Lemos, M., and Dessai, S. (2013). Actionable knowledge for environmental decision making: broadening the usability of climate science. *Annu. Rev. Environ. Resour.* 38, 393–414. doi: 10.1146/annurev-environ-022112-112828

Kirchmeier-Young, M. C., Wan, H., Zhang, X., and Seneviratne, S. I. (2019). Importance of framing for extreme event attribution: the role of spatial and temporal scales. *Earth's Future* 7, 1192–1204. doi: 10.1029/2019EF001253

Kirmayer, L. J., Sehdev, M., Whitley, R., Dandeneau, S. F., and Isaac, C. (2009). Community resilience: models, metaphors and measures. *Int. J. Indigenous Health* 5, 62–117. doi: 10.3138/ijih.v5i1.28978

Kirmayer, L. J., and Valaskakis, G. G. (2009). *Healing Traditions: The Mental Health of Aboriginal Peoples in Canada*. Vancouver, BC: UBC Press.

Klinsky, S., Roberts, T., Huq, S., Okereke, C., Newell, P., Dauvergne, P., et al. (2017). Why equity is fundamental in climate change policy research. *Global Environ. Change* 44, 170–173. doi: 10.1016/j.gloenvcha.2016.08.002

Laursen, S., Puniwai, N., Genz, A. S., Nash, S. A., Canale, L. K., and Ziegler-Chong, S. (2018). Collaboration across worldviews: managers and scientists on Hawaiî Island utilize knowledge coproduction to facilitate climate change adaptation. *Environ. Manage.* 62, 619–630. doi: 10.1007/s00267-018-1069-7

Lemos, M. C., Arnott, J. C., Ardoin, N. M., Baja, K., Bednarek, A. T., Dewulf, A., et al. (2018). To co-produce or not to co-produce. *Nat. Sustain.* 1, 722–724. doi: 10.1038/s41893-018-0191-0

Li, F., and Whalley, J. (2002). Deconstruction of the telecommunications industry: from value chains to value networks. *Telecomm. Policy* 26, 451–472. doi: 10.1016/S0308-5961(02)00056-3

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., et al. (2020). The TRUST principles for digital repositories. *Scientific Data* 7, 144. doi: 10.1038/s41597-020-0486-7

Lubchenco, J., and Rapley, C. (2020). Our moment of truth: the social contract realized? *Environ. Res. Lett.* 15, 110201. doi: 10.1088/1748-9326/abba9c

Lynnes, C., Zhu, M. Q., Blythe, J., Williamson, T. N., Burnett, J., Huffer, E., et al. (2020). *Usage-Based Discovery of Earth Observations*. 2020, IN012-02. Available online at: https://ui.adsabs.harvard.edu/abs/2020AGUFMIN012..02L

Mach, K. J., Lemos, M. C., Meadow, A. M., Wyborn, C., Klenk, N., Arnott, J. C., et al. (2020). Actionable knowledge and the art of engagement. *Curr. Opin. Environ. Sustain.* 42, 30–37. doi: 10.1016/j.cosust.2020.01.002

Magis, K. (2010). Community resilience: an indicator of social sustainability. *Soc. Nat. Resources* 23, 401–416. doi: 10.1080/08941920903305674

Mayernik, M., and Virapongse, A. (2019). *Making Data Matter with Matt Mayernik*. Making Data Matter. doi: 10.6084/m9.figshare.7914197.v1

McWethy, D. B., Schoennagel, T., Higuera, P. E., Krawchuk, M., Harvey, B. J., Metcalf, E. C., et al. (2019). Rethinking resilience to wildfire. *Nat. Sustain.* 2, 797–804. doi: 10.1038/s41893-019-0353-8

Meadow, A. M., Ferguson, D. B., Guido, Z., Horangic, A., Owen, G., and Wall, T. (2015). Moving toward the deliberate coproduction of climate science knowledge. *Weather Climate Soc.* 7, 179–191. doi: 10.1175/WCAS-D-14-00050.1

Meier, W., Fetterer, F., Windnagel, A., and Stewart, S. (2021). *NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, Version 4*. doi: 10.7265/EFMZ-2T65

Moe, K., Moran, T., Jones, D., Hicks, K., Glasscoe, M., Virapongse, A., et al. (2018). *ESIP Webinar #5: Managing Disasters Through Improved Data-Driven Decision-Making*. doi: 10.6084/m9.figshare.7361327

Nissan, H., Goddard, L., de Perez, E. C., Furlow, J., Baethgen, W., Thomson, M. C., et al. (2019). On the use and misuse of climate change projections in international development. *Wiley Interdiscip. Rev. Clim. Chang.* 10, e579. doi: 10.1002/wcc.579

NSIDC (2022). Easy-to-use Data Products. NSIDC. Available online at: https://nsidc.org/data/resources/easy-to-use.

O'Neill, E. T. (2011). Frbr: functional requirements for bibliographic records. *Libr. Resour. Tech. Serv.* 46, 150–159. doi: 10.5860/lrts.46n4.150

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., et al. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *J. Information Sci.* 37, 555–569. doi: 10.1177/0165551511412705

PCAST-Executive Office of the President (2011). *Sustaining Environmental Capital: Protecting Society and the Environment*. President's Council of Advisors on Science and Technology. Available online at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast_sustaining_environmental_capital_report.pdf

Peng, G., Downs, R.R., Lacagnina, C., Ramapriyan, H., Ivánová, I., Moroni, D., et al. (2021). Call to action for global access to and harmonization of quality information of individual earth science datasets. *Data Sci. J.* 20, 19. doi: 10.5334/dsj-2021-019

Pratt, K. L. (2004). Observations on researching and managing alaska native oral history: a case study. *Alaska J. Anthropol.* 2, 138–153.

Pulsifer, P., Gearheard, S., Huntington, H. P., Parsons, M. A., McNeave, C., and McCann, H. S. (2012). The role of data management in engaging communities in Arctic research: Overview of the Exchange for Local Observations and Knowledge of the Arctic (ELOKA). *Polar Geogr.* 35, 271–90. doi: 10.1080/1088937X.2012.708364

Pulsifer, P. L., Kontar, Y., Berkman, P. A., and Taylor, D. R. F. (2020). "Information ecology to map the arctic information ecosystem," in *Governing Arctic Seas: Regional Lessons from the Bering Strait and Barents Sea*, Vol. 1, eds O. R. Young, P. A. Berkman, and A. N. Vylegzhanin (Cham: Springer International Publishing), 269–291.

Rainie, S. C., Rodriguez-Lonebear, D., and Martinez, A. (2017). *Policy Brief(Version2): Data Governance for Native Nation Rebuilding*. Tucson: Native Nations Institute.

Ramapriyan, H., and Behnke, J. (2019). Importance and incorporation of user feedback in earth science data stewardship. *Data Sci. J.* 18, 24. doi: 10.5334/dsj-2019-024

Ramapriyan, H. K., and Murphy, K. (2017). Collaborations and partnerships in nasa's earth science data systems. *Data Sci. J.* 16, 51–55. doi: 10.5334/dsj-2017-051

Reimsbach-Kounatze, C. (2021). "Enhancing access to and sharing of data: striking the balance between openness and control over data," in *Data Access, Consumer Interests and Public Welfare* (Nomos Verlagsgesellschaft mbH & Co. KG), 25–68.

Rolan, G., McKemmish, S., Oliver, G., Evans, J., and Faulkhead, S. (2020). "Digital equity through data sovereignty: a vision for sustaining humanity," in *IConference 2020 Proceedings*.

Roux, D., Rogers, K., Biggs, H., Ashton, P., and Sergeant, A. (2006). Bridging the science-management divide: moving from unidirectional knowledge transfer to knowledge interfacing and sharing. *Ecol. Soc.* 11. doi: 10.5751/ES-01643-110104

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *J. Information Sci.* 33, 163–180. doi: 10.1177/0165551506070706

Scassa, T. (2010). Geographical information as 'personal information.' *Oxford Univ. Commonwealth Law J.* 10, 185–214. doi: 10.5235/147293410794895322

Semnacher, C., and Chong, S. (2019). *Improving Information Retrieval: The Arctic Data Center Unveils New Semantic Search Product*. Witness the Arctic. Available online at: https://arcus.org/witness-the-arctic/2019/2/article/30127

Sharifi, A. (2016). A critical review of selected tools for assessing community resilience. *Ecol. Indic.* 69, 629–647. doi: 10.1016/j.ecolind.2016.05.023

Sharma, N. (2008). *The Origin of the "Data Information Knowledge Wisdom"(DIKW) Hierarchy*. Available online at: http://www.researchgate.net/publication/292335202_The_Origin_of_Data_Information_Knowledge_Wisdom_DIKW_Hierarchy

Sisco, A., Cook, K., Bugbee, K., Davidson, J., Duffy, L., Dabolt, T., et al. (2019). *Changing Climate, Changing Data: Exposing Climate Data to New Users Through GeoPlatform.gov's Resilience Community*. American Geophysical Union Fall Meeting, Poster Contribution: IN33B-0822, San Francisco, CA. Available online at:: agu.confex.com/agu/fm19/meetingapp.cgi/Paper/615144

Smith, D. E. (2016). "Governing data and data for governance: the everyday practice of indigenous sovereignty, " in *Indigenous Data Sovereignty: Toward an Agenda*, 117–135.

Smith, J. W., Moore, R. L., Anderson, D. H., and Siderelis, C. (2012). Community resilience in Southern Appalachia: a theoretical framework and three case studies. *Hum. Ecol.* 40, 341–353. doi: 10.1007/s10745-012-9470-y

Song, I. Y., and Zhu, Y. (2016). Big data and data science: what should we teach?. *Expert Syst.* 33, 364–373. doi: 10.1111/exsy.12130

Starkweather, S., Shapiro, H., Vakhutinsky, S., and Druckenmiller, M. (2020). *The Observational Foundation of the Arctic Report Card-A 15-Year Retrospective Analysis on the Arctic Observing Network (AON) and Insights for the Future System*. National Oceanic and Atmospheric Administration.

Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., and Brody, J. G. (2017). *Re-Identification Risks in HIPAA Safe Harbor Data: A Study of Data From One Environmental Health Study*. Technology Science, 2017. Available online at: *https*://www.ncbi.nlm.nih.gov/pmc/articles/PMC6344041/

Tai, T. C., and Robinson, J. P. (2018). Enhancing climate change research with open science. *Front. Environ. Sci.* 6, 115. doi: 10.3389/fenvs.2018.00115

Tanner, K. D. (2009). Learning to see inequity in science. *CBE Life Sci. Educ.* 8, 265–270. doi: 10.1187/cbe.09-09-0070

Taylor, H. A. (1984). Information ecology and the archives of the 1980s. *Archivaria* 25–37. Available online at: https://archivaria.ca/index.php/archivaria/article/view/11075

Theodori, G. L. (2005). Community and community development in resource-based areas: operational definitions rooted in an interactional perspective. *Soc. Nat. Resour.* 18, 661–669. doi: 10.1080/08941920590959640

Thompson, N., Ravindran, R., and Nicosia, S. (2015). Government data does not mean data governance: lessons learned from a public sector application audit. *Govern. Infm. Q.* 32, 316–322. doi: 10.1016/j.giq.2015.05.001

Trisos, C. H., Adelekan, I. O., Totin, E., Ayanlade, A., Efitre, J., Gemeda, A., et al. (2022). "Africa," in *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, eds H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, et al. (Cambridge: Cambridge University Press), 9–225.

Turbes, C. (2020). *Continuing the Federal Data Strategy is a Must-Do for the Next Administration*. Data Coalition. Available online at: http://www.datacoalition.org/continuing-the-federal-data-strategy-is-a-must-do-for-the-next-administration,/

UN General Assembly (2007). United Nations declaration on the rights of indigenous peoples. *UN Wash* 12, 1–18.

UNCTAD (2020). *Data Protection and Privacy Legislation Worldwide*. UNCTAD. Available online at: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

UNCTAD and DPR. (2016). *International Data Flows: Implications for Trade and Development*. United Nations. Available online at: https://unctad.org/system/files/official-document/dtlstict2016d1_en.pdf

United States Agency for International Development (USAID) (2012). *Building Resilience to Recurrent Crisis: USAID Policy and Program Guidance*. Available online at: http://www.usaid.gov/sites/default/files/documents/1870/USAIDResiliencePolicyGuidanceDocument.pdf

United States Government, Droegemeier, K., Kelley, K., Kent, S., Potok, N., and Roat, M. (2019). *Federal Data Strategy 2020 Action Plan*. Available online at: https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf

Vera, C., Barange, M., Dube, O. P., Goddard, L., Griggs, D., Kobysheva, N., et al. (2010). Needs assessment for climate information on decadal timescales and longer. *Procedia Environ. Sci.* 1, 275–286. doi: 10.1016/j.proenv.2010.09.017

Virapongse, A., Brooks, S., Metcalf, E. C., Zedalis, M., Gosz, J., Kliskey, A., et al. (2016). A social-ecological systems approach for environmental management. *J. Environ. Manage.* 178, 83–91. doi: 10.1016/j.jenvman.2016.02.028

Wee, B., and Piña, A. (2019). *A Vision for Adapting at the Pace of Socioenvironmental Change*. EOS. Available online at: https://eos.org/opinions/a-vision-for-adapting-at-the-pace-of-socioenvironmental-change

Wenger-Trayner, E., and Wenger-Trayner, B. (2015). *Introduction to Communities of Practice*. Available online at: https://wenger-trayner.com/introduction-to-communities-of-practice/

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

Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., and Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *J. Commun. Informatics* 12, 3. doi: 10.15353/joci.v12i3.3275

Yates, K. K., Turley, C., Hopkinson, B. M., Todgham, A. E., Cross, J. N., Greening, H., et al. (2015). Transdisciplinary science: a path to understanding the interactions among ocean acidification, ecosystems, and society. *Oceanography* 28, 212–225. doi: 10.5670/oceanog.2015.43

Conflict of Interest: During this work, AV was the sole employee of the Middle Path EcoSolutions consultancy which aims to empower communities to become more healthy and sustainable. RG was a Researcher at Knology for the majority of the writing process.

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

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

frontiers | Frontiers in Climate

PERSPECTIVE
published: 22 June 2022
doi: 10.3389/fclim.2022.785269

Data Usability: The Forgotten Segment of Environmental Data Workflows

Shannon Dosemagen and Emelia Williams*

Open Environmental Data Project, New Orleans, LA, United States

While there has been a rapid increase in the use of participatory science methods over the last decade, the usability of resulting data in addressing situations of environmental injustice is often overlooked, neglected, or used as political fuel for ignoring inconvenient truths. The inability of data to be used for policy, regulation, and enforcement impedes its usefulness in various situations depending on user requirements and governance scales. On the other hand, there are vast open datasets that could be useful for communities and researchers, but these data are often difficult to find, use, or repurpose, beyond their original intent. This article unpacks the data usability problem at the frontier of environmental governance and decision-making, suggesting that by prioritizing environmental data as a public good, there are clear mechanisms for ensuring data usability toward participatory environmental governance. The authors are interested in uncovering the policies and behavioral and bureaucratic patterns that have remained static as participatory science methods and tools have advanced. It is necessary to understand where and when associated tools, methods, and platforms have failed to ensure that data is usable and useful for communities attempting deeper engagement and representation in environmental governance.

Keywords: community data, environmental governance, environmental justice, data usability, public good, open source, environmental data

OPEN ACCESS

Edited by:
Michael C. Kruk, National Oceanic and Atmospheric Administration (NOAA), United States

Reviewed by:
Stephen Formel, National Oceanic and Atmospheric Administration (NOAA), United States
Mathew Biddle, U.S. Integrated Ocean Observing System, United States

****Correspondence:***
Shannon Dosemagen
shannon@openenvironmentaldata.org

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *28 September 2021*
Accepted: *25 May 2022*
Published: *22 June 2022*

Citation:
Dosemagen S and Williams E (2022) Data Usability: The Forgotten Segment of Environmental Data Workflows. Front. Clim. 4:785269. doi: 10.3389/fclim.2022.785269

INTRODUCTION

Data about the environment and its impact on health come from many places, including scientists and researchers, government, and communities who are activated to collect their own data. There are an equal number of issues with environmental data: scarcity in some places and overabundance in others; difficulties collecting data based on timing, accessibility of tools, and technical complexities of data requirements; figuring out where and how data can be disseminated for use in different scenarios. While citizen science (Shirk et al., 2012) has dominated the language and landscape of participatory science, this article is interested in forms of participatory science such as community science (Dosemagen and Parker, 2019) and community-owned and managed research (Heaney et al., 2007). These center scientific practice around the questions of communities, seek to build co-equal partnerships between communities and scientists, and aim to leverage multiple forms of data (e.g., quantitative data from sensors, traditional, and local knowledge)

to an actionable end[1], often in support of addressing environmental injustices. This article is also interested in the role of already available open government datasets[2] and the benefit for communities and other researchers beyond the original intent of use.

In efforts to collect and share data and information as part of environmental governance processes, there are limited cases or examples that show how community data follows a streamlined process from collection by communities to its use in decision-making within and between communities and tribal, local, state, national, and global governance processes. While there is demonstrable progress—notably the inclusion of water quality monitoring data in local, state, federal, and multi-lateral processes such as the Sustainable Development Goals[3] and through legislation such as the Crowdsourcing and Citizen Science Act of 2016[4]—the problem of *data usability* is often overlooked or neglected, in part allowing inconvenient environmental truths to perpetuate. Likewise, data streams coming from scientific institutions (e.g., research institutes, government agencies, and universities) often struggle to solve the data "last-mile" usability problem (Celliers et al., 2021). In other words, frameworks have been created for ensuring these data streams meet certain standards for enabling access and usability, but we have yet to figure out how to format data for a variety of different user needs and governance scales beyond original intent. This perspective article unpacks the data usability problem at the frontier of environmental governance and decision-making.

The sophistication of participatory science continues to grow, and progress has been made toward increasing actionable data, yet there are limited instances of data and information from communities being used in ways that demonstrate clear integration with policymaking and ongoing, collaborative interaction between government and communities around environmental governance and management. Often, data collected to demonstrate a potential environmental issue (see for instance Allen, 2003) is paired with long-term and ongoing community activism. The popular route of public notice-and-comment leaves much to be desired as it does not account for power differentials, creates further inequity through lack of access to political know-how, and allows for an information request without having a feedback loop through which response is guaranteed to the comment provider (Rahman, 2011). Data from communities can provide rich contextual and time-sensitive information in environmental governance decisions, like permitting affordances from industrial plants or land stewardship practices for endangered species. The overused *data pipeline* analogy suggests a clear route from community data collection to enforcement of rights (see **Figure 1**), but our workflows underperform when it comes to ensuring data are usable and useful for communities attempting deeper engagement and representation in environmental governance (see **Figure 2**).

The onus of working within existing data systems has long been placed on communities. This is exemplified through priority placed on training communities to interact with existing workflows, rather than internal agency self-reflection on where data workflows complicating community involvement could better function. It is the responsibility of government, with insight and advice from civil society, to correct data workflow issues and to modernize and update the infrastructure that supports them. This can happen by prioritizing environmental data as a public good (i.e., data that works for all) which can emphasize the necessity and value of diverse data and information in environmental governance (Williams et al., 2021). While methods and tools for monitoring have proliferated (for instance through next-generation sensors or the value that local and traditional knowledge can bring toward adding context to environmental datasets), the policies, behavioral patterns, and bureaucratic systems around data have remained relatively stagnant.

WHY NOW: THE OPPORTUNITY FOR A WHOLE OF GOVERNMENT APPROACH

With the Biden-Harris Administration declaring a "whole-of-government"[5] approach to environmental justice (Justice40[6] and the need for climate action, it is an opportune time to think differently about where data can be useful in governance processes and also the ways in which data moves between actors—from community to government, government to community, researchers to communities. Historically, community data is used by communities and, in certain cases, researchers and government, to call attention to potential environmental and health issues, often resulting in establishing a baseline for further research, indicating the need for additional monitoring, or assisting media campaigns in support of community goals (National

[1]See for instance the American Geophysical Union's Thriving Earth Exchange, Public Lab and the Association of Science and Technology Centers.

[2]In 2019, the OPEN Government Data Act, a component of the Foundations for Evidence Based Policymaking Act (https://www.congress.gov/bill/115th-congress/house-bill/4174/text#toc-H8E449FBAEFA34E45A6F1F20EFB13ED95), was turned into law. This, "requires federal agencies to publish their information online as open data, using standardized, machine-readable data formats, with their metadata included in the Data.gov catalog." (https://www.data.gov/meta/data-gov-at-ten-and-the-open-government-data-act/).

[3]See for example: Monitoring drinking water quality for the Sustainable Development Goals (https://www.nature.com/collections/gdiahjefdh/) and UNEP Monitoring Water Quality (https://www.unep.org/explore-topics/water/what-we-do/monitoring-water-quality).

[4]15 U.S.C. §3,724. Crowdsourcing and citizen science (https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title15-section3724&num=0&edition=prelim).

[5]Fact Sheet: President Biden Takes Executive Actions to Tackle the Climate Crisis at Home and Abroad, Create Jobs, and Restore Scientific Integrity Across Federal Government (https://www.whitehouse.gov/briefing-room/statements-releases/2021/01/27/fact-sheet-president-biden-takes-executive-actions-to-tackle-the-climate-crisis-at-home-and-abroad-create-jobs-and-restore-scientific-integrity-across-federal-government/).

[6]Justice40 does not explicitly mention data accessibility or usability, though the recommendations from the White House Environmental Justice Advisory Council reference community data input extensively and could signal a future push to incorporate (https://www.whitehouse.gov/omb/briefing-room/2021/07/20/the-path-to-achieving-justice40/).

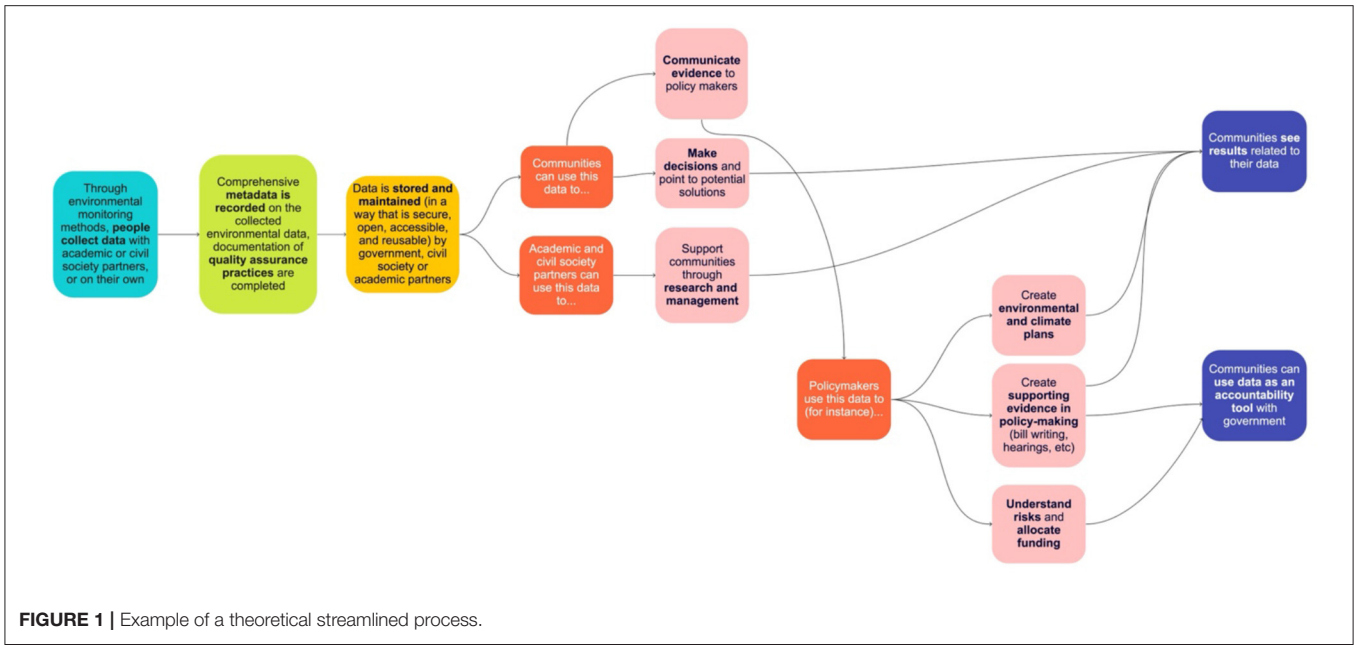


FIGURE 1 | Example of a theoretical streamlined process.

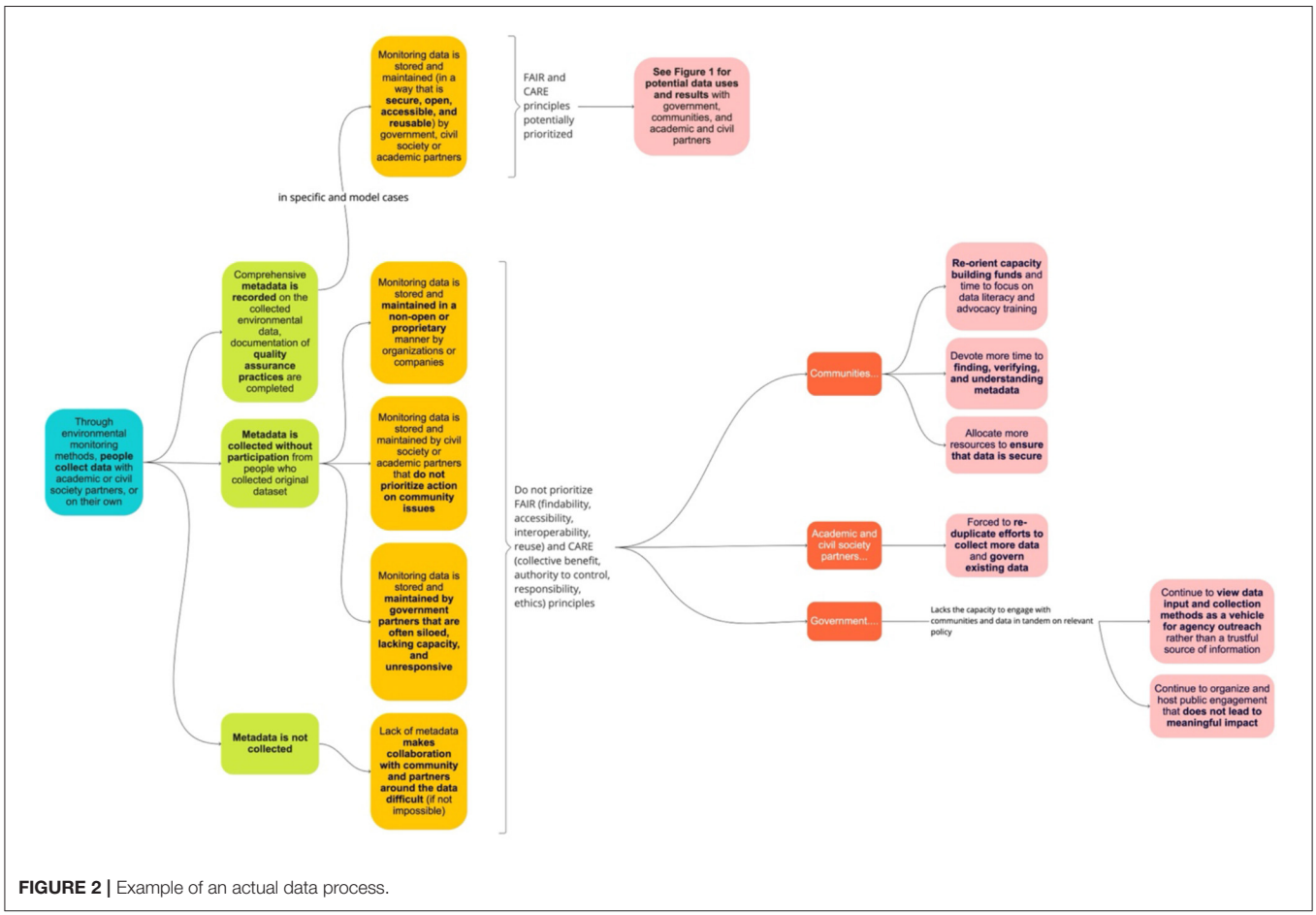


FIGURE 2 | Example of an actual data process.

Advisory Council on Environmental Policy Technology, 2016). Rarely though, do we see environmental data used to provide an ongoing system of collective accountability between community, government, academia, and industry. The use of data as a tool should be prioritized by government and the broader public as a public good; explicit nomenclature that designates data as such can highlight its use as a tool for collective accountability. It is here that we need to focus our efforts.

To achieve a whole-of-government approach to environmental justice and climate change mitigation efforts, agencies should embrace addressing and solving questions of data use and accessibility that people have pressed for years[7]. In addition to the Administration's openness to incorporating environmental justice into national agendas, there is also demonstrated intent toward action from Congress such as the introduction of the *Environmental Justice Mapping and Data Collection Act of 2021*[8], *Environmental Justice Act of 2021*[9] and the *Environmental Justice for All Act*[10]. The federal government's movement on this stands on the foundation of more localized and regional action, largely catalyzed over the past 5 years, in state, city, and mayoral offices.

Though tools such as *EJScreen*[11] and the *Climate and Economic Justice Screening Tool*[12] offer useful demographic data, these tools primarily are a means for understanding and identifying environmental justice communities (Barnes et al., 2021), rather than increasing access to national systems of environmental governance. These tools, as well as new modes of enhancing public data literacy and education, are valuable and needed, but must be paired with programs such as environmental justice training for federal employees and the increased distribution of funds to environmental justice communities. Such programs and support should seek to identify and leverage places where data, information, and input from communities could be used to create multi-stakeholder, collaborative models of governance that value and encourage the use of environmental data and information. This can include a range of inputs from traditional ecological knowledge to "good enough data" (Gabrys et al., 2016) that demonstrate where trends might be emerging.

The Environmental Protection Agency (EPA)[13] and other federal agencies have a complicated road ahead in which they'll be required to address large-scale systems change across a gamut of activities from transportation and land-use to infrastructure upgrades. Environmental data has a role to play in these scaled changes, to provide a clear understanding of what resources are needed in which geographies. To create truly just systems of environmental governance in which data that already exists and data that is created by communities is valued as part of the process, it is necessary for government to (1) consider administrative justice[14] alongside environmental justice, and (2) to understand the entrenched behavioral and cultural challenges that government faces before becoming open to this form of data collection. It is also necessary to reconsider administrative justice as, "a set of principles for shaping humane relationships between citizen and state" in the "small places," in the interactions between civil servants, between government and community (Doyle and O'Brien, 2020). A truly whole-of-government approach must include these "small places" and data questions (Doyle and O'Brien, 2020). Federal agencies that collect and share environmental data, such as NOAA, USGS, and NASA, can and should support these efforts, but as a regulatory agency whose mission is to ensure human health (in addition to environmental protection), the EPA is best positioned to lead such change.

WHY THIS: DATA AS A TOOL TOWARD ENVIRONMENTAL JUSTICE

Early signals from the Biden-Harris Administration point to environmental justice as a route for building conversations about the role of environmental data and information from impacted communities. The focus on a whole-of-government approach to environmental justice will increase the propensity of government to identify the needs of environmental justice communities. However, the authors contend there is a more significant role for community data in decision-making. Amplifying this role can be accomplished by not only creating more data and maps to show the distribution of environmental injustices, but also by creating

[7]For instance, see the case studies in the 2016 NACEPT report (National Advisory Council on Environmental Policy Technology, 2016), the 2018 NACEPT report (National Advisory Council on Environmental Policy Technology, 2018), and reports by the Environmental Law Institute in 2020 (Moodley and Wyeth, 2020).

[8]The *Environmental Justice Mapping and Data Collection Act of 2021* notes that it "aims to create and authorize funding for a system to comprehensively identify the demographic factors, environmental burdens, socioeconomic conditions, and public health concerns that are related to environmental justice and collect high-quality data through community engagement and a government-wide interagency process. These data would be used to build layered maps depicting which communities experience environmental injustices" (https://www.congress.gov/bill/117th-congress/senate-bill/101).

[9]The *Environmental Justice Act of 2021* aims "to improve research and data collection relating to the health and environment of populations of color, communities of color, indigenous communities, and low-income communities, including through the increased use of community-based science" (https://www.congress.gov/bill/117th-congress/senate-bill/2630?q=%7B%22search%22%3A%5B%22S.+2630%22%2C%22S.%22%2C%222630%22%5D%7D&s=2&r=1).

[10]The *Environmental Justice for All Act* seeks "to improve Federal research and data collection efforts related to— (I) the health and environment of communities of color, low-income communities, and Tribal and Indigenous communities..." (https://www.congress.gov/bill/116th-congress/house-bill/5986/text).

[11]EJScreen is the Environmental Justice Screen and Mapping Tool, developed and used by the EPA "to screen for areas that may be candidates for additional consideration, analysis or outreach as EPA develops programs, policies and activities that may affect communities" (https://www.epa.gov/ejscreen/how-does-epa-use-ejscreen).

[12]The Climate and Economic Justice Screening Tool is in its public beta form, developed by the Council of Environmental Quality "to help Federal agencies identify disadvantaged communities that are marginalized, underserved, and overburdened by pollution" (https://screeningtool.geoplatform.gov/en/#3/33.47/-97.5).

[13]While we acknowledge framing this paper to focus on EPA limits the whole-of-government approach to one agency, the authors do so to help narrow the discussion. Further exploration of how other agencies are addressing these topics is a possible future route of work.

[14]We note that this term is specifically used as a concept in law and judicial systems, but we use it here in parallel with the concept of environmental justice to underline the complexity of administrative systems.

data accessibility, literacy and transparency for a plethora of researchers from community to academic. Government should also look to the less acknowledged places where communities can provide direct guidance on program rollouts. For instance, in funding programs, identifying where points of input in grantmaking processes about how funds are spent can lead to a stronger balance in the distribution of these funds. The work of identifying places of input has begun with the White House Environmental Justice Advisory Committee and their Interim Final Recommendations for the Climate and Economic Justice Tool (WHEJAC, 2021), but there is additional work needed to streamline this process and ensure that less acknowledged communities are involved.

Additionally, the whole-of-government approach to infrastructure and environmental justice seems to have its limits within the Biden-Harris administration, namely when it comes to the oil and gas industry, as they recently announced an increase of exports of liquefied natural gas (Natter and Dlouhy, 2022) and are outpacing the Trump administration in issuing drilling permits on public lands (Phillips, 2022). These types of environmentally harmful activities point to places where community environmental data could bolster calls for government accountability by providing, for instance, information on the lived experiences of and impacts on communities in proximity to this harm. Only when our infrastructure allows for access to both data and decision-making across the places where influence sits, will we move from a *whole-of-government* to a *multi-sector collaborative* governance model.

While environmental justice is the focus, a method of collecting data to address these injustices[15]—citizen science (and to some extent community science)—has previously made large strides in becoming part of agency agendas. There is an interagency working group on crowdsourcing and citizen science[16], the National Advisory Council on Environmental Policy and Technology (NACEPT)[17], wrote two substantial reports on citizen science, and there is a law encouraging the increased use of citizen science in Federal Government[18] However, one of the key agencies required to interface with community data, EPA, has historically viewed data input and the methods for collecting it as a vehicle for agency outreach and engagement. While EPA has created resources such as the quality assurance toolkit (Environmental Protection Agency, 2022b), houses an environmental monitoring tool loan program (Environmental Protection Agency, 2022a), and provides regional funding for citizen and community science projects, less capacity has been directed at systemic Agency-wide integration of community data and the infrastructure needed to ensure this data is used. There is also a problematic history of how communities seeking to be part of the environmental data infrastructure have been categorically dismissed or viewed as data contributors (rather than co-equal partners); they have filled in gaps for government agencies that lack the political, social, or economic capacity to achieve their mission of environmental and health protection[19] and management. Community data is not a replacement for government inaction, or an avenue leading to community-industry partnerships, but should be seen as a way for communities to build agency in political decision-making (Ottinger, 2013; Shapiro et al., 2017).

The reason we place value on community data is that this data and information can serve to socially situate issues, provide different perspectives, and communicate how people are experiencing environmental injustices and the burden of pollution[20]. Notably, ensuring the role of community data and information in environmental governance can show us the value of pairing scientific data alongside contextual information, for instance indicating there are multiple truths to how people experience living in polluted environments[21]. Community data can also help agencies forecast areas where future interventions are required with trend data collected by communities. Being able to proactively point to out-of-pattern events is invaluable—especially as we see the increasing effects of the climate crisis.

Community data can additionally provide new partnership and outreach opportunities for agencies to work with scientists, community organizers and advocates, educators, designers, and technologists. These partnerships are integral to ensuring that, as our innovation landscape around the next generation of environmental sensors increases, technology, and its resulting data are usable by agencies. The incorporation of environmental data from communities requires an openness and willingness on the part of agencies to examine and explore both these new environmental data technology frontiers and their own complex and difficult-to-navigate administrative systems. Working with communities, and their data and information, can demonstrate a willingness for agencies to collaboratively achieve EPA's mandate[22] environmental and human health protection. This participatory collaboration will require a switch from the mindset of being a gatekeeper of this responsibility to being a conduit for working in partnership with the public. Building in processual transparency and points of clear input for communities, can work against the legacy of distrust in government by environmental justice communities.

In the Biden-Harris Administration, there is also a notable financial commitment, for instance, to increased air quality monitoring (Environmental Protection Agency, 2021), and billions allocated to cleaning up legacy pollution and investments in the nation's water infrastructure (Mock and

[15]Citizen science and environmental justice and their influences can be described as a feedback loop, which can be positive or negative. See Figure 12.1 in *Citizen Science, Health, and Environmental Justice* (Ceccaroni et al., 2021).

[16]Crowdsourcing and Citizen Science (https://digital.gov/communities/crowdsourcing-citizen-science/).

[17]National Advisory Council on Environmental Policy and Technology (https://archive.epa.gov/epa/faca/nacept.html).

[18]15 U.S.C. §3724. Crowdsourcing and citizen science (https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title15-section3724&num=0&edition=prelim).

[19]The EPA mission is to protect human health and the environment. Read more at: https://www.epa.gov/aboutepa/our-mission-and-what-we-do.

[20]For more on the situation of knowledge, see Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective (Haraway, 1988).

[21]Communities can use environmental data to create information that demonstrates experience of lived situations and other forms of knowing (see for instance Celliers et al., 2021).

[22]The EPA mandate is derived from the Reorganization Plan No. 3 of 1970 (https://archive.epa.gov/epa/aboutepa/reorganization-plan-no-3-1970.html).

Lowenkron, 2021). Paired with a focus on squarely placing routes of community input (not simply data collection) and accountability into the infrastructure of these financial commitments would signal the potential for a transformative approach to environmental injustices.

HOW WE GET THERE: SYSTEMS FOR COLLECTIVE ACCOUNTABILITY

To deepen the whole-of-government approach and create an ongoing commitment to both the work of environmental justice and environmental justice communities, policy frameworks should incorporate the willingness to explore and expand an ongoing system of collective accountability for environmental protection, management, and governance. This is a complex problem that requires a multi-faceted approach through which we collaboratively, (1) build models for new ways to think about incorporating diverse datasets and their metadata, while also considering how to strengthen the current governance landscape data lives in, (2) standardize across both new and old data systems to support collective accountability, and (3) build legitimacy in new data systems through this accountability.

To be successful in creating new systems of collective accountability, changes in bureaucratic culture that lead to administratively just systems should be created. While many people in government agencies are proponents of incorporating community data and information and recognizing the value of collaborative governance, these are not agency-wide mandates. For instance, only recently has the White House Office of Science and Technology Policy explicitly advised that "where appropriate, ITEK[23] can and should inform Federal decision making along with scientific inquiry" (Lander and Mallory, 2021). Instead, many times community input is seen as an administrative burden in a system already weighed down by bureaucracy (Harrison, 2017). Those that see community input as a burden often actively resist—in both conscious and unconscious ways—the work necessary for environmental justice (Harrison, 2019). When the bureaucratic system of data input and analysis by the agency causes additional delays, there are failures in systems leaving limited choices for remedy. These delays and blockades increase the failures of systems in addressing environmental justice concerns (Goldman, 2000).

There are also places for considering hybrid social, legal, and technical approaches to the way data becomes available for government use. These places must consider and design for the representation of a diversity of perspectives, respect the boundaries of communities in sharing (as well as the necessity of sovereignty) and ensure that there is a place of input beyond public comment processes and the mechanics of town halls and public hearings. For instance, the Open Environmental Data Project (OEDP) has been working on conceptualizing a community data hub model that is (a) decentralized for collaborative ownership within each community and (b) reflective of collective governance models[24]; at the same time, it recognizes the importance and necessity of federated systems[25] so that communities (and their data) can speak to each other and government infrastructure (see **Figure 3**). OEDP pairs these concepts and prototypes with models that tell the story of the pain points these types of systems would encounter through network amplifying conversations (i.e., OEDP's Brain Trusts or *Data Dialogues* series). These dialogues help us to identify the complexities of usable data in ways that look at them as opportunities for creating new systems or thinking in different ways that will help us to alleviate environmental data burdens.

Through a dual approach that ensures *environmental data is a public good*—it is non-rival and non-excludable—we mechanize its ability to be an accountability measure in both directions—from government to communities and from communities to government. The space of community environmental data and governance is ripe for this change.

DISCUSSION: ENVIRONMENTAL DATA AS A PUBLIC GOOD

The roots of environmental injustice in the United States span further back than the start of the environmental justice movement or the EPA (Altman, 2021) into the early industrial period of U.S. history (or, one could argue, early colonization). Yet a century later, we are just starting to acknowledge that a whole-of-government approach is needed to address these issues. To make this whole-of-government (and community) approach it is necessary to ensure the place of *environmental data as a public good* (see **Figure 4**).

A public good serves the well-being of a populace. It is past time to ensure both our *available* environmental data and environmental data that are *collected* both on hyper-local scales and by, for instance, sensor networks, are allowed this position in society. The state of pollution combined with the climate crisis means we need an all-hands-on-deck approach to solve these problems. Our solutions are in the data, the technology, and the ability for people to share what they know based on local and lived experience. But it is necessary to put structures in place so that in building environmental data as a public good, we ensure a stronger and clearer emphasis on data "reusability"[26]. There are three main ways in which this can happen:

Ensure that administrative justice[27] is part of the environmental justice whole-of-government approach. The whole-of-government approach to addressing environmental injustices will not work without putting specific attention toward

[23] Indigenous Traditional Ecological Knowledge.

[24] Collective, or collaborative governance models refers to a "multi-actor collaboration, usually led by a public sector organization aimed at building consensus among stakeholders on a formal set of policies designed and implemented to generate public value" (Bianchi et al., 2021).

[25] Federated systems enable "queries to be sent between disparate data repositories, or nodes in a federation." For more on the benefits, risks, and elements of a federated system (see Herrman, 2019).

[26] FAIR Principles address reusability, the closest acknowledgement of data usability beyond original intent, though known problems exist with maintenance of data under these principles (Wilkinson et al., 2016).

[27] In "New Directions in Environmental Justice Research at the U.S. Environmental Protection Agency: Incorporating Recognitional and Capabilities Justice Through Health Impact Assessments" the authors also note the importance of capabilities and recognitional justice in relation to environmental justice work with communities (Eisenhauer et al., 2021).

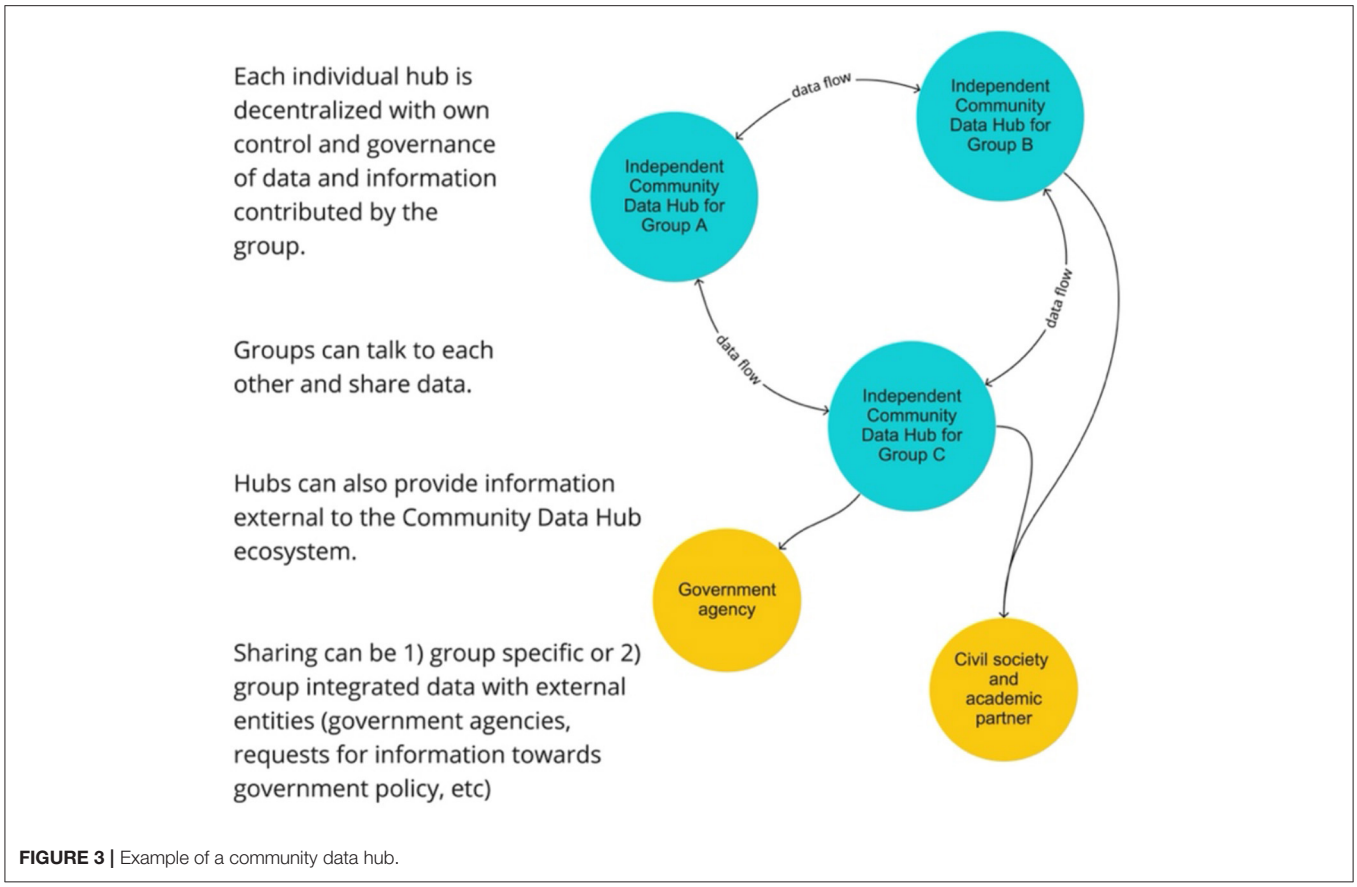


FIGURE 3 | Example of a community data hub.

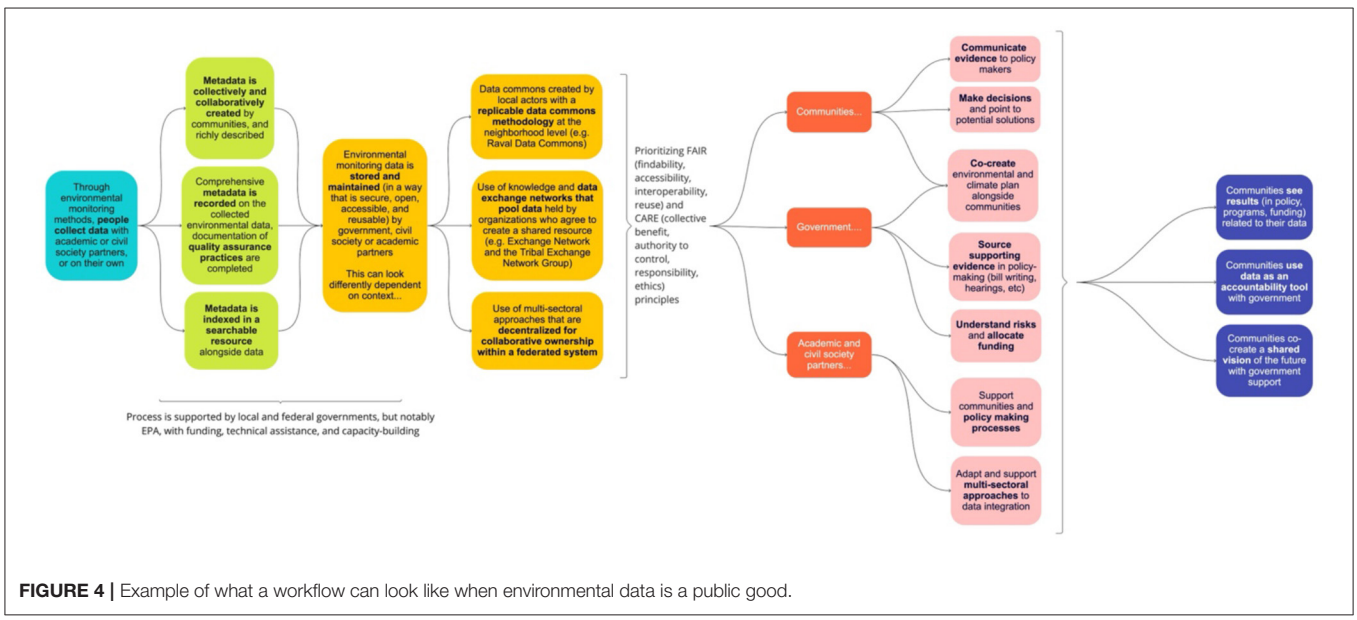


FIGURE 4 | Example of what a workflow can look like when environmental data is a public good.

addressing behaviors and bureaucratic systems that are unjust in themselves and then also exemplified as acceptable in agency workflows (including those related to data). To ensure existing data is usable, findable, accessible, and that there are routes of input for communities, the administrative behaviors of agencies will need to be examined, alongside technical workflows.

While working in current governance systems, create new ones that are responsive to communities and their data needs.

Change within government processes is slow and introducing and adapting innovative and responsive governance models for scaled use will be incremental. Though it is necessary to work within current environmental governance systems (i.e., public commentary frameworks), the rise of new technology and methods for data collection should encourage us to think about how to use more representative forms of data and information that allow for robust models of community governance. The civic technology movement[28] of over a decade ago provided a plethora of valuable lessons (for instance, Costanza-Chalk, 2020; Harrell, 2020), and models from within government that have been reinterpreted by non-profits[29] and vice versa. The models and frameworks are there to build representative data systems for collaborative environmental governance. While doing so, consideration for what the connective tissues between old and new systems are—and specifically what usability structures need to be put into place—should be central.

Consider socio-technical touchpoints. Many times, new technology for data management, storage, and collection does not need to be built from scratch; instead, there is a needed investment in critical digital infrastructure and features that will make environmental data usable and useful. To create better representation, the focus should be on the appropriateness of models of collaborative governance[30], community ownership, direct routes of input and checks and balances that data provides, and how the data fits into current data systems and yet is proactively designed for future systems[31]. As previously discussed, current problems with bureaucratic workflows point out that these developments must work toward unburdensome governance structures or they simply will not be used. Also known, through the reflections of researchers and practitioners on the past decade of civic technology (see for instance Costanza-Chalk, 2020), is that any design or technology development that leaves people out, or is created *for*, not *with,* will further problematize the push toward addressing environmental injustices.

Across sectors and working in collaboration, this is an opportune moment to grasp the momentum we're seeing at the top levels of the administration, Congress, and federal agencies to do differently and do better for and with environmental justice communities. To grasp this opportunity, we must recognize the deep histories of misaligned bureaucratic practices that have complicated, or even intentionally or unintentionally prevented, how environmental justice can happen in practice. Building workable routes in our current data systems should be prioritized, while simultaneously encouraging spaces of innovation in which we can consider legacy systems, setting the tone for new ones that allow for proactive and collaborative environmental governance. The rise of environmental data from multiple sources should be considered a public good and we have a collective responsibility to ensure it becomes a *workable* public good for both communities and the elected officials that represent them.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SD researched, analyzed, and wrote the article. EW researched, analyzed, and added written contributions to the article. All authors contributed to the article and approved the submitted version.

[28] Civic technology is a "loosely integrated movement that brings the strengths of the private-sector tech world (its people, methods, or actual methodology) to public entities with the aim of making government more responsive, efficient, modern, and more just" (Harrell, 2020).

[29] See for instance, *In the Realm of the Barely Feasible (Prabhakar, 2020)*.

[30] Part of creating these touchpoints is to understand and test models that have risen around collaborative governance of resources in other sectors. Communities who are collectively contributing data should have control mechanisms and ownership boundaries in place. Extensive work has been done to this point around the sharing of health data where models such as trusts and collectives have been tested [e.g., Aapti Institute (https://www.aapti.in/) and GovLab's Data Collaboratives initiative (https://datacollaboratives.org/)]. Querying if a focus on governance could bridge the conversation between data sovereignty and representation in data-based governance decisions for communities is also important.

[31] Future responsive systems could look many ways, but Open Environmental Data Project has suggested a system for a hardware and software platform that integrates dynamic environmental trend data, of known quality, about ecological integrity (https://www.openenvironmentaldata.org/work/new-models-environmental-context-part-1).

REFERENCES

Allen, B. (2003). *Uneasy Alchemy: Citizens and Experts in Louisiana's Chemical Corridor Disputes*. Cambridge: MIT Press.

Altman, R. (2021). *Upriver*. Orion. Available online at: https://orionmagazine.org/article/upriver/

Barnes, A., Luh, A., and Gobin, M. (2021). *Mapping Environmental Justice in the Biden-Harris Administration*. Center for American Progress. Available online at: https://www.americanprogress.org/article/mapping-environmental-justice-biden-harris-administration/

Bianchi, C., Nasi, G., and Rivenbark, W. (2021). Implementing collaborative governance: models, experiences, and challenges. *Publ. Manage. Rev.* 23, 1581–1589. doi: 10.1080/14719037.2021.1878777

Ceccaroni, L., Woods, S., Springs, J., Wilson, S., Faustman, E., Bonn, A., et al. (2021). "Citizen science, health, and environmental justice," in *The Science of Citizen Science*, eds K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, and K. Wagenknecht (Cham: Springer), 219–239. doi: 10.1007/978-3-030-58278-4_12

Celliers, L., Máñez Costa, M., Williams, D. S., and Rosendo, S. (2021). The 'last mile' for climate data supporting local adaptation. *Glob. Sustain.* 4, 1–8. doi: 10.1017/sus.2021.12

Costanza-Chalk, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need.* Cambridge: MIT Press. doi: 10.7551/mitpress/12255.001.0001

Dosemagen, S., and Parker, A. (2019). Citizen science across a spectrum: building partnerships to broaden the impact of citizen science. *Sci. Technol. Stud.* 32, 24–33. doi: 10.23987/sts.60419

Doyle, M., and O'Brien, N. (2020). *Reimagining Administrative Justice: Human Rights in Small Places*. London: Palgrave. doi: 10.1007/978-3-030-21388-6

Eisenhauer, E., Williams, K. C., Warren, C., Thomas-Burton, T., Julius, S., and Geller, A. M. (2021). New directions in environmental justice research at the U.S. Environmental Protection Agency: incorporating recognitional and capabilities justice through health impact assessments. *Environ. Just.* 14, 322–331. doi: 10.1089/env.2021.0019

Environmental Protection Agency (2021). *EPA Announces an Additional $50 Million Under the American Rescue Plan to Enhance Air Pollution Monitoring*. Environmental Protection Agency. Available online at: https://www.epa.gov/newsreleases/epa-announces-additional-50-million-under-american-rescue-plan-enhance-air-pollution

Environmental Protection Agency (2022a). *EPA's Equipment Loans Program*. Environmental Protection Agency. Available online at: https://www.epa.gov/citizen-science/epas-equipment-loan-programs

Environmental Protection Agency (2022b). *Quality Assurance Handbook and Guidance Documents for Citizen Science Projects*. Environmental Protection Agency. Available online at: https://www.epa.gov/citizen-science/quality-assurance-handbook-and-guidance-documents-citizen-science-projects

Gabrys, J., Pritchard, H., and Barratt, B. (2016). Just good enough data: figuring data citizenships through air pollution sensing and data stories. *Big Data Soc.* 3, 1–14. doi: 10.1177/2053951716679677

Goldman, B. (2000). An environmental justice paradigm for risk assessment. *Hum. Ecol. Risk Assess.* 6, 541–548. doi: 10.1080/10807030008951327

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Femin. Stud.* 14, 575–599. doi: 10.2307/3178066

Harrell, C. (2020). *A Civic Technologist's Practice Guide*. San Francisco, CA: Five Seven Five Books.

Harrison, J. L. (2017). We do ecology, not sociology': interactions among bureaucrats and the undermining of regulatory agencies' environmental justice efforts. *Environ. Sociol.* 3, 197–212. doi: 10.1080/23251042.2017.1344918

Harrison, J. L. (2019). *From the Inside Out: The Fight for Environmental Justice Within Environmental Agencies*. Cambridge: MIT Press. doi: 10.7551/mitpress/12063.001.0001

Heaney, C., Wilson, S., and Wilson, O. (2007). The West End Revitalization Association's community-owned and -managed research model: development, implementation, and action. *Prog. Commun. Health Partnersh.* 1, 339–349. doi: 10.1353/cpr.2007.0037

Herrman, A. (2019). *Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data*. World Economic Forum. Available online at: https://www3.weforum.org/docs/WEF_Federated_Data_Systems_2019.pdf

Lander, E., and Mallory, B. (2021). *Memorandum for the Heads Of Departments and Agencies: Indigenous Traditional Ecological Knowledge and Federal Decision Making*. Available online at: https://www.whitehouse.gov/wp-content/uploads/2021/11/111521-OSTP-CEQ-ITEK-Memo.pdf (accessed December 1, 2021).

Mock, B., and Lowenkron, H. (2021). *The Infrastructure Bill Is a Trillion-Dollar Test for Environmental Justice*. Bloomberg CityLab. Available online at: https://www.bloomberg.com/news/articles/2021-08-11/an-infrastructure-bill-built-on-environmental-justice

Moodley, K., and Wyeth, G. (2020). *Citizen Science Programs at Environmental Agencies: Case Studies*. Environmental Law Institute. Available online at: https://www.eli.org/research-report/citizen-science-programs-environmental-agencies-case-studies

National Advisory Council on Environmental Policy and Technology (2016). *Environmental Protection Belongs to the Public: A Vision for CITIZEN Science at EPA*. Eds S. Dosemagen and A. Parker. Report to the Environmental Protection Agency.

National Advisory Council on Environmental Policy and Technology (2018). *Information to Action: Strengthening EPA Citizen Science Partnerships for Environmental Protection*. Eds S. Dosemagen, A. Parker and D. Bator. Report to the Environmental Protection Agency.

Natter, A., and Dlouhy, J. (2022). *Biden Risks Undercutting Climate Goal With Wartime Gas Pivot*. Bloomberg Green. Available online at: https://www.bloomberg.com/news/articles/2022-03-24/biden-risks-undercutting-climate-goals-with-wartime-pivot-to-gas

Ottinger, G. (2013). *Refining Expertise: How Responsible Engineers Subvert Environmental Justice Challenges*. New York, NY: NYU Press. doi: 10.18574/nyu/9780814762370.001.0001

Phillips, A. (2022). *Biden Outpaces Trump in Issuing Drilling Permits on Public Lands*. The Washington Post. Available online at: https://www.washingtonpost.com/climate-environment/2022/01/27/oil-gas-leasing-biden-climate/

Prabhakar, A. (2020). In the realm of the barely feasible. *Issues Sci. Technol.* 31:1. doi: 10.1111/phis.12167

Rahman, K. S. (2011). Envisioning the regulatory state: technocracy, democracy, and institutional experimentation in the 2010 financial reform and oil spill statues. *Harvard J. Legislat*. 48:555.

Shapiro, N., Zakariya, N., and Roberts, J. (2017). A wary alliance: from enumerating the environment to inviting apprehension. *Engag. Sci. Technol. Soc.* 3, 575–602. doi: 10.17351/ests2017.133

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., et al. (2012). Public participation in scientific research: a framework for deliberate design. *Ecol. Soc.* 17:29. doi: 10.5751/ES-04705-170229

WHEJAC (2021). *Interim Final Recommendations on Justice40 Climate and Economic Justice Screening Tool & Executive Order 12898 Revisions*. Report to the Council on Environmental Quality.

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

Williams, E., Dosemagen, S., and Hoeberling, K. (2021). *Opportunity Brief: Environmental Data as a Public Good*. Open Environmental Data Project. Available online at: https://www.openenvironmentaldata.org/research-series/environmental-data-as-a-public-good (accessed December 17, 2021).

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

frontiers | Frontiers in Climate

ORIGINAL RESEARCH
published: 27 June 2022
doi: 10.3389/fclim.2022.841103

Democratizing Glacier Data – Maturity of Worldwide Datasets and Future Ambitions

Isabelle Gärtner-Roer[1], Samuel U. Nussbaumer[1], Bruce Raup[2], Frank Paul[1], Ethan Welty[1], Ann K. Windnagel[2], Florence Fetterer[2] and Michael Zemp[1]*

[1] World Glacier Monitoring Service (WGMS), Department of Geography, University of Zurich, Zurich, Switzerland, [2] National Snow and Ice Data Center (NSIDC), Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, United States

The creation and curation of environmental data present numerous challenges and rewards. In this study, we reflect on the increasing amount of freely available glacier data (inventories and changes), as well as on related demands by data providers, data users, and data repositories in-between. The amount of glacier data has increased significantly over the last two decades as remote sensing techniques have improved and free data access is much more common. The portfolio of observed parameters has increased as well, which presents new challenges for international data centers, and fosters new expectations from users. We focus here on the service of the Global Terrestrial Network for Glaciers (GTN-G) as the central organization for standardized data on glacier distribution and change. Within GTN-G, different glacier datasets are consolidated under one umbrella, and the glaciological community supports this service by actively contributing their datasets and by providing strategic guidance via an Advisory Board. To assess each GTN-G dataset, we present a maturity matrix and summarize achievements, challenges, and ambitions. The challenges and ambitions in the democratization of glacier data are discussed in more detail, as they are key to providing an even better service for glacier data in the future. Most challenges can only be overcome in a financially secure setting for data services and with the help of international standardization as, for example, provided by the CoreTrustSeal. Therefore, dedicated financial support for and organizational long-term commitment to certified data repositories build the basis for the successful democratization of data. In the field of glacier data, this balancing act has so far been successfully achieved through joint collaboration between data repository institutions, data providers, and data users. However, we also note an unequal allotment of funds for data creation and projects using the data, and data curation. Considering the importance of glacier data to answering numerous key societal questions (from local and regional water availability to global sea-level rise), this imbalance needs to be adjusted. In order to guarantee the continuation and success of GTN-G in the future, regular evaluations are required and adaptation measures have to be implemented.

Keywords: glacier data, maturity matrix assessment, data repositories, Essential Climate Variable (ECV), Global Terrestrial Network for Glaciers (GTN-G)

OPEN ACCESS

Edited by:
Tiffany C. Vance, U.S. Integrated Ocean Observing System, United States

Reviewed by:
Wesley Van Wychen, University of Waterloo, Canada
Laura Thomson, Queen's University, Canada

****Correspondence:***
Isabelle Gärtner-Roer
isabelle.roer@geo.uzh.ch

Specialty section:
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

Received: *21 December 2021*
Accepted: *17 May 2022*
Published: *27 June 2022*

Citation:
Gärtner-Roer I, Nussbaumer SU, Raup B, Paul F, Welty E, Windnagel AK, Fetterer F and Zemp M (2022) Democratizing Glacier Data – Maturity of Worldwide Datasets and Future Ambitions. Front. Clim. 4:841103. doi: 10.3389/fclim.2022.841103

BACKGROUND

The amount of glacier data has increased significantly over the last two decades. In the year 2000, data from around a hundred glaciers with direct mass-balance measurements and from around 1,000 glaciers with annual observations of terminus fluctuations were available. Today, satellite data in combination with the Randolph Glacier Inventory (RGI) that became available in 2012 (Pfeffer et al., 2014), enables the observation of all 215,000 glaciers worldwide. As a result, further observational parameters have been included in glacier repositories, such as glacier area and volume changes, flow velocities, ice-thickness distribution and snow-covered areas. This presents new challenges for the international data centers that provide access to glacier data, and fosters new expectations from users. In parallel, the storage (archiving), documentation (metadata), and access to the data and related products have become much more complex.

The Global Terrestrial Network for Glaciers (GTN-G) is the framework for the internationally coordinated monitoring of glaciers in support of the United Nations Framework Convention on Climate Change (UNFCCC). The network, authorized under the Global Climate Observing Systems (GCOS), is jointly run by the World Glacier Monitoring Service (WGMS), the U.S. National Snow and Ice Data Center (NSIDC), and the Global Land Ice Measurements from Space initiative (GLIMS). GTN-G represents the glacier monitoring community and provides an umbrella for existing and operational data services and related working groups such as of the International Association of Cryospheric Sciences (IACS). This setting is largely responsible for its success.

In addition to qualitative data, such as photographs and maps, the GTN-G provides standardized observations on changes in mass, volume, area and length of glaciers with time (glacier fluctuations), as well as statistical information on the spatial distribution of perennial surface ice (glacier inventories) (**Figure 1**). Such glacier fluctuation and inventory data are high-priority key variables in climate system monitoring; they form a basis for hydrological modeling with respect to possible effects of global warming, and provide fundamental information in glaciology, glacial geomorphology, and quaternary geology. The increased amount of glacier data from the last decade has enhanced the understanding of geophysical processes, improved glacier-related modeling, and resulted in higher-confidence statements in the last report of the International Panel on Climate Change (IPCC, 2021). Beyond this, the data are needed for the development of sustainable adaptation strategies and related decision-making processes in glacierized mountain regions (Nussbaumer et al., 2017; Gärtner-Roer et al., 2019). These urgent demands are accompanied by equally urgent challenges, such as the rapidly increasing number of glacier observations from space that need to be managed in a functioning database infrastructure.

GTN-G facilitates free access to data through different channels, depending on the level of interest and detail required, and addresses issues such as the standardization of measurement methods. Most important, it gathers high-level information about and access to all available glacier datasets on one platform (https://www.gtn-g.ch/data_catalog/). This ensures that all data are equally available for any user: findable, accessible, interoperable, and reusable, following the FAIR principles (Wilkinson et al., 2016). While the public or mountain tourists might use the "wgms Glacier App" for a quick overview of available glacier data, scientists typically access data offered by the glacier services using the GTN-G data browser or directly from the catalog listing of data collections held by member repositories. Decision makers make use of edited products such as reports or specific country profiles. Finally, journalists often approach the GTN-G or constituent services directly and ask for support in filtering the main messages out of the full database and in showing different perspectives. Thus, the different repositories serve different user communities and purposes. However, whereas the FAIR principles emphasize the needs of data users, the right of the data providers to be acknowledged should not be neglected. Acknowledgment is accomplished through versioning of the datasets, e.g., *via* digital object identifiers (DOIs). When users cite datasets and include a DOI, the DOI provides traceability between data creation and use. During the whole process, the proper citation of data origin must be followed and ideally should be controlled by repositories, journals, and funding agencies. However, such control mechanisms have yet to be established by the international community.

Each GTN-G dataset nicely reflects the history of glacier monitoring, which began in 1894 with the internationally coordinated systematic observations on glacier variation (**Figure 1**). The history mostly followed the overall paradigms in science: after empirical and theoretical investigations, focus was given on simulations and, more recently, on "big data." For the long-term monitoring of environmental variables, continuous and standardized measurements are of highest priority. The *in situ* measurements, where methodology has changed little over the last 125 years, are fundamental to this long-term monitoring. On the other hand, in order to capture uniform information on a large scale (glacier distribution, changes in ice thickness), remote sensing data are indispensable. The rise of "big data" in glaciology is a direct result of the rapid increase in remote sensing techniques and corresponding data, as well as free data-access policies (e.g., Landsat; Wulder et al., 2012) and the availability of "analysis ready data," for example pre-orthorectified satellite scenes in GeoTIF format that can be easily processed and analyzed.

With the increase in volume, timeliness, and variety of data, as well as variety of data users, access becomes ever more challenging and requires improved interfaces (Pospiech and Felden, 2012). Citizens increasingly use data from different sources (maps, tides, etc.) and glaciers all around the world can now be explored and measured without much effort. This has implications for the management and handling of monitoring datasets that affect data providers as well as data users. Hence it is time to critically reflect on the democratization of glacier data. In the context of this study we understand the term "democratizing" as the free access to glacier data for everyone. As this is an active verb, it implies the transition of a former "closed" system to a more "open" system, even if access to most glacier data has been open already for many years (WGMS, 1998). For the future, it is the process of proper

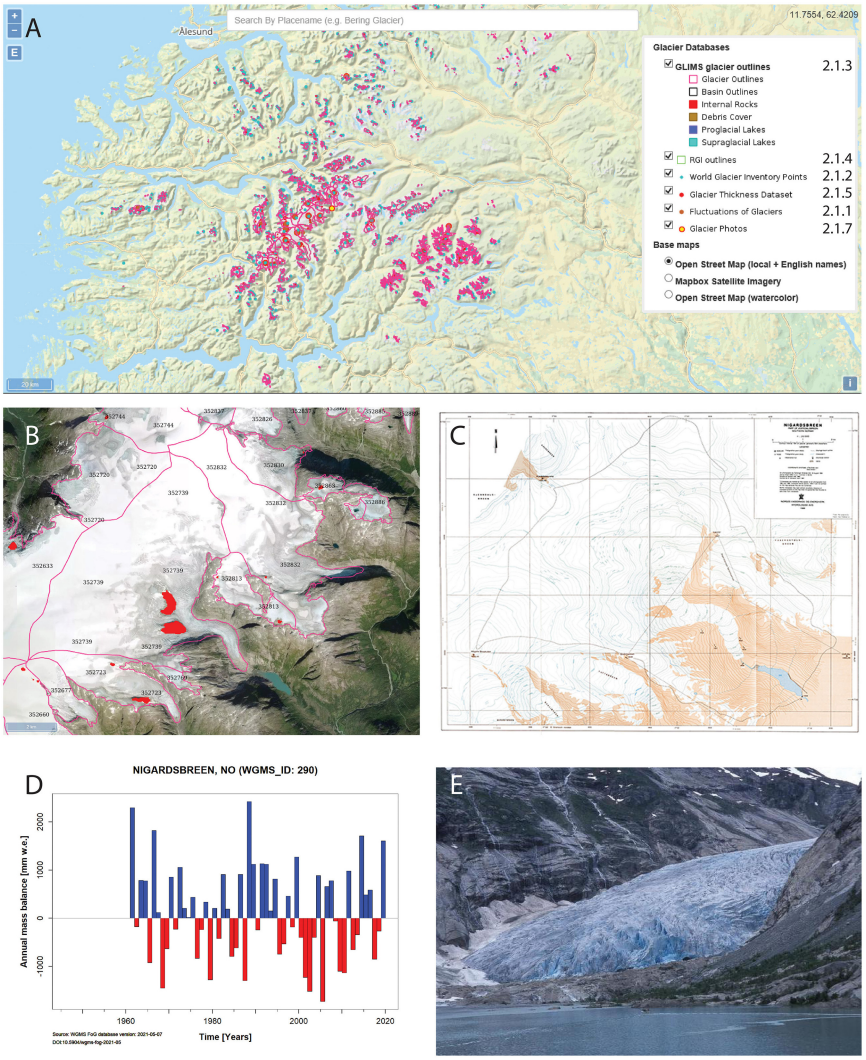


FIGURE 1 | (A) The GTN-G data browser (zoomed) showing available glacier data for a region in southern Norway. The legend shows the different datasets and the related sections in this paper where the datasets are described. Examples are given for Nigardsbreen, an outlet glacier of the Jostedalsbreen ice cap (Norway), as represented in different glacier datasets accessible *via* the GTN-G data browser: **(B)** GLIMS outlines as of 2006 (ID: 352739); **(C)** topographic map of Nigardsbreen as of 1998 (Norwegian Water Resources and Energy Directorate, NVE); **(D)** annual mass balance since 1960 (B. Kjøllmoen and colleagues, NVE; WGMS, 2021a); **(E)** photo of the glacier tongue as of August 3rd, 2000 (E. Roland; Glacier Photograph Collection).

data stewardship and international standardization that ensures the democratization of data. In this context, certification is provided by the CoreTrustSeal (https://www.coretrustseal.org/), an international, community based, non-governmental, and non-profit organization promoting sustainable and trustworthy data infrastructures.

We here systematically assess all seven GTN-G datasets with a focus on data preservability, accessibility, usability, production sustainability, quality assurance, quality control, quality assessment, transparency/traceability, and integrity, as described by Peng et al. (2015). The individual performance is analyzed with regard to the historical development and the current funding situation of individual datasets, but also with regard to each dataset's significance and function for environmental monitoring and related decision-making procedures. Particular challenges are stressed and suggestions for solutions are provided by good-practice recommendations. During this process, the requirements of both data providers and data users are considered. Expectations from and ambitions of GTN-G are also formulated, as they direct the way toward the further democratization of glacier data.

DATA AND METHODS

Description of Datasets Available Within the GTN-G

Internationally coordinated collection and distribution of standardized information about glaciers was initiated in 1894 and is, since 1998, coordinated within GTN-G. Since 2008, an international steering committee coordinates, supports, and advises the operational bodies responsible for the international glacier monitoring, which are the WGMS, the NSIDC, and

GLIMS (Zemp, 2011). GTN-G ensures (1) the integration of the various operational databases and (2) the development of a one-stop web interface to these databases. The datasets all have different purposes, formats and histories, reflecting the history of glaciological science (**Figure 2**). By joint effort, consistency and interoperability of the different glacier databases has had to be developed; the different historical developments and methodological contexts of the datasets are challenges for linking individual glaciers throughout the databases.

For the analysis of the data, the interoperability with web-based services (e.g., cloud services) need to be improved. So far, most of the glacier datasets can be downloaded directly in their entirety and can be integrated into a programmatic local or cloud-based workflow. However, the linking between different GTN-G datasets is not very mature and urgently needs to be developed further. Developed in 2010 and updated since, a map-based web interface spatially links the available data and provides data users a fast overview of all available data (https://www.gtn-g.ch/data_browser/; see **Figure 1**). The interface was adapted for GTN-G from one developed for the constituent NASA-sponsored GLIMS initiative. It provides fast access to information on glacier outlines from about 215,000 glaciers mainly based on satellite images, length-change time series from 2,581 glaciers, glaciological mass-balance time series from 482 glaciers, geodetic mass-balance series from 37,446 glaciers, special events (e.g., hazards, surges, calving instabilities) from 2,747 glaciers, as well as more than 25,000 photographs (RGI Consortium, 2017; National Snow and Ice Data Center, 2021; WGMS, 2021b). By choosing the browser layer for a particular dataset one can quickly see the spatial distribution of that dataset. Whereas some datasets are fed continuously by an active community (such as the FoG (Fluctuations of Glaciers) and GLIMS datasets), others are created on an *ad-hoc* basis (GlaThiDa glacier thickness database and RGI dataset), have a random community (Glacier Photograph Collection, Glacier Map Collection) or have been discontinued (World Glacier Inventory, WGI).

The spatio-temporal coverage of the different datasets varies largely, because of their individual histories. For the *in situ* data there is a significant spatial bias toward the Northern Hemisphere, in particular to Europe and North America, whereas the Andes and Antarctica are underrepresented. In GlaThiDa, the largest spatial gaps persist in Asia, the Russian Arctic, and the Andes. With the recent developments in satellite remote sensing of the cryosphere, the extended sharing of data, and the free availability of a globally complete baseline glacier inventory (the RGI), near global coverage has been achieved for many datasets during the last decades (e.g., Farinotti et al., 2019; Hugonnet et al., 2021). Other temporal gaps in the datasets are related to the limited lifetime of individual projects or institutions. In addition, political crises can have a direct influence on the long-term continuation of data series. An assessment of national glacier distribution and changes, delineating also spatio-temporal gaps, is provided in Gärtner-Roer et al. (2019).

Fluctuations of Glaciers

Internationally collected and standardized data on changes in glacier length, area, volume, and mass, based on *in situ* and remotely sensed observations, as well as on model-based reconstructions, are compiled in the Fluctuations of Glaciers (FoG) database. The standardized compilation and free dissemination of glacier data from all over the world, as undertaken by the WGMS and its predecessor organizations, are a major contribution to international initiatives and bodies such as the United Nations Framework Convention on Climate Change (UNFCCC) and the Intergovernmental Panel on Climate Change (IPCC) (**Figure 2**). Since the beginning of coordinated glacier monitoring, the collected data have been published in written reports. The first reports were written in French, but from 1967 on, all reports are published in English (see https://wgms.ch/literature_published_by_wgms). The comprehensive FoG reports represented the backbone of the scientific data compilation, which comes with full documentation on principal investigators, national correspondents, their sponsoring agencies, and publications related to the reported data series. These reports, issued every 5 years, were complemented by the bi-annual Glacier Mass Balance Bulletin, which presented the data in summary form for non-specialists through the use of graphic presentations rather than as purely numerical data. In 2015, these two publication series were merged into the "Global Glacier Change Bulletin" series with the aim of providing an integrative assessment of worldwide and regional glacier changes at two-year intervals. Beyond these synthesis reports, the FoG data are accessed by downloadable files of past and current versions since 2008 (https://wgms.ch/data_databaseversions/), direct visualizations *via* the FoG Browser (https://wgms.ch/fogbrowser/), and the "wgms Glacier App" for mobile devices (https://wgms.ch/glacierapp/).

With the inclusion of near real-time measurements at high temporal resolution (e.g., hourly data) for selected study sites and the increasing amount of satellite-derived observations (number of records evolved from a few hundred to more than 200,000 glaciers), the database experienced growing pains. In order to address these challenges, plans for migration to advanced database structures are currently under development.

World Glacier Inventory

The WGI was planned as a snapshot of glacier occurrence on Earth during the second half of the 20th century. In 1976, the United Nations Environment Programme (UNEP), through its Global Environment Monitoring System (GEMS) started supporting activities of a Temporary Technical Secretariat for the World Glacier Inventory (TTS/WGI) established at the Geography Department of ETH (Eidgenössische Technische Hochschule) Zurich. Detailed and preliminary regional inventories were compiled all over the world. From these inventories, statistical measures of the geography of glaciers could be extracted. The WGI completed and updated earlier compilations (e.g., by Mercer, 1967 and Field, 1975). Instructions and guidelines for the compilation of standardized glacier inventory data were developed by UNESCO/IASH (1970), Müller et al. (1977), Müller (1978), and Scherler (1983). The publication of the WGI report (WGMS, 1989) presented the status at the end of 1988, and is the first such compilation to give a systematic global overview. It contains information for

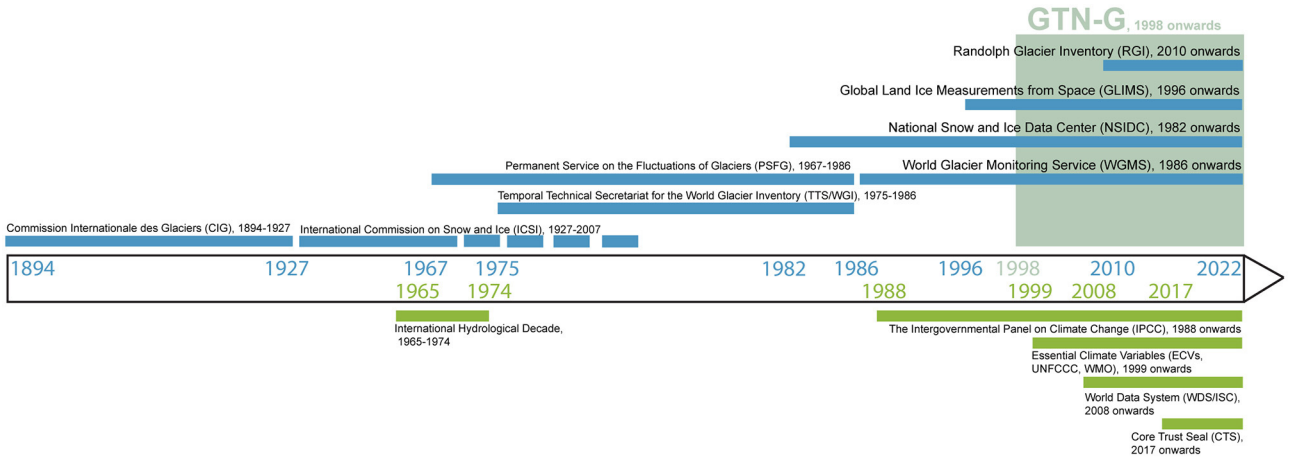


FIGURE 2 | Timeline related to the operational bodies and partners within GTN-G and their predecessor bodies (blue bars), as well as influencing international efforts in relation to use and valorization of glacier data (green bars). Figure modified after Allison et al. (2019).

approximately 130,000 glaciers. Inventory parameters include geographic location, area, length, orientation, elevation, and classification. The WGI is based primarily on aerial photographs and topographic maps with most glaciers having one data entry only. Hence, the dataset can be viewed as a snapshot of the glacier distribution in the second half of the 20th century. An update of the WGI was performed in 2012 (WGMS, and National Snow and Ice Data Center, 2012).

The data collection presents a fairly complete, albeit preliminary, picture of the world's glacierized regions at the given time. The WGI database is stored both at WGMS in Zurich and in the National Snow and Ice Data Center's NOAA collection, part of the World Data Center for Glaciology in Boulder, Colorado. It is most easily accessed through the website of GTN-G (www.gtn-g.org). The WGI database is searchable by glacier ID, glacier name, or latitude/longitude (as well as other parameters) using the main "Search Inventory" interface. In addition, the "Extract Selected Regions" interface can be used.

It was the sincere wish of organizations and people who have been involved in WGI activities over the years that the information in the publication, together with the data available in the database, be of service to scientists and decision makers concerned with various applications of glacier data both then and in the future (WGMS, 1989). For instance, it was suggested that the information available within the WGI together with other data provided by the WGMS could be usefully applied in studies of the impact of a global warming on the availability of water resources in frozen form, particularly in semi-arid and arid regions bordering glacierized areas. Inventory data had already proven useful for estimating precipitation amounts in some mountainous regions where stations for direct measurements are difficult to establish, and it was expected they would be used for the same purpose in many more regions (WGMS, 1989).

Independent of the high scientific value of the glacier information stored in the WGI, it has some disadvantages when considering today's applications. The foremost problem is its storage as point information. The shapes and the extents of the glaciers to which the data belong are unknown. It cannot be used for change assessment or any application that requires glacier outlines. The technological revolution in the 1990's providing Geographic Information Systems (GIS), digital elevation models (DEMs) and satellite data covering nearly each region in the world with glaciers, has made it possible to generate, store and manipulate related vector data. As a consequence, the GLIMS database (see Global Land Ice Measurements From Space) was initiated at the turn of the century, superseding the WGI. The compilation of a near-globally complete dataset of glacier outlines as available from the RGI (see Randolph Glacier Inventory) was, however, only possible once free access to orthorectified satellite data, DEMs, and GIS environments was in place.

Global Land Ice Measurements From Space

The GLIMS glacier database (GLIMS Consortium, 2005) contains multi-temporal outlines of glaciers in a vector format with additional data about each glacier (e.g., name, area, length or mean elevation). All data are stored in a PostGIS relational database, providing support for geographic objects allowing location queries. It emerged from the increasing need for improved calculation of glacier changes and glacier-specific assessments, which were impossible using the point data provided by the WGI (see above). As the WGI and its extended format WGI-XF (Cogley, 2009) was still spatially incomplete, there was also an urgent need to obtain complete global coverage, i.e., to have outlines from all glaciers in the world rather than just 2/3. At the inception of GLIMS in 2010 it was still not known how many glaciers we had on Earth, where they were located and how large they were. Accordingly, all calculations concerned with regional scale hydrology in mountain regions or global scale sea-level rise were highly error prone. With the free availability of multispectral images at 15 m spatial resolution from the ASTER sensor onboard the Terra satellite (after its launch in 1999), the dream of a global glacier database suddenly became realistic. Data

acquisition requests for ASTER were prepared (Raup et al., 2000) and a geospatial database was created (Raup et al., 2007). The relational database included everything that possibly could be derived from a satellite image and slowly filled over the years.

Whereas algorithms for automated mapping of clean glacier ice were already established at that time (e.g., Bayr et al., 1994; Paul et al., 2002), two major bottlenecks hindered rapid and efficient data processing: (a) debris-covered glacier parts were not included and had to be delineated manually and (b) image analysts had to manually orthorectify all ASTER scenes. With no money available for such time-consuming activities both could only be performed as a part of funded research projects that mostly analyzed small regions (e.g., Paul and Kääb, 2005). Fortunately, the opening of the Landsat archive in 2008 (Wulder et al., 2012) suddenly provided free access to all Landsat scenes in an already orthorectified format and obviated the need to use manually orthorectified ASTER imagery. This encouraged glacier mapping over larger regions (Bolch et al., 2010; Frey et al., 2012; Rastner et al., 2012; Guo et al., 2015) which filled the GLIMS database with more, better quality and also multi-temporal data. At the time of writing, the GLIMS database hosts approximately 300,000 glacier outlines (including perennial snow patches), i.e., 40% of the 215,000 glaciers have multi-temporal outlines. The data have been widely used for a range of hydrological and glaciological applications. The datasets stored in GLIMS also formed the base for the compilation of a first globally complete single snapshot inventory (RGI; see next section).

Randolph Glacier Inventory

The RGI was born from two ideas: to have (i) an easily accessible temporal snapshot of glacier extents available that is (ii) globally complete, i.e., there is one outline for each glacier in the world with the relevant attribute information. This idea was motivated primarily by the preparation of IPCC AR5, where a clear need for such a dataset was communicated to the glaciological community to improve the assessment of glacier-related questions (e.g., their contribution to sea-level rise) compared to IPCC AR4. With glacier outlines from the GLIMS database and a special community effort in glacier mapping (for details see Pfeffer et al., 2014), first versions of this dataset were created and provided for the global-scale calculations presented in IPCC AR5 (Vaughan et al., 2013). Given the limited time available for finalizing the product, shortcomings in quality were accepted, noting that the outlines were produced for global to continental scale assessments rather than regional or local ones. Over time, the RGI was continuously improved (version 6.0 appeared in 2017) and the regionally most complete datasets were collected and combined for the best possible product.

Whereas, the initial effort to get all data together in a consistent format was enabled by a couple of engaged individuals, the current effort for compilation of a further improved RGI (version 7) is coordinated by a dedicated IACS working group (https://cryosphericsciences.org/activities/working-groups/rgi-working-group/) that is organizing and structuring the related work. A detailed technical specification about RGI contents, its development over time, and all its contributors is available in the form of a Technical Note from the RGI web page (https://www.glims.org/RGI/00_rgi60_TechnicalNote.pdf). The RGI is split into 19 first order regions, each having its own glacier outlines shapefile and hypsometric data file. When summed up, it contains about 215,000 glacier entities covering an area of more than 723,000 km^2 (excluding glaciers on the Antarctic Peninsula).

The RGI has likely become one of the single most important datasets for glaciological and hydrological research. It is widely accepted as the best possible dataset for large scale applications and the number of studies using it might exceed 1000. The related study by Pfeffer et al. (2014) describing version 3.2 in detail is now the most cited publication in the Journal of Glaciology. The intense use of the dataset is also a main reason for ongoing efforts to further improve it, being careful not to lessen its usefulness. For the new version 7 of the RGI it was decided to bring the individual datasets closer to the year 2000 (e.g., to facilitate mass-balance calculations starting with the SRTM or ASTER-derived DEMs from 2000) and swap out datasets with known problems (e.g., too much seasonal snow mapped as glaciers in the Andes) for "better" ones.

A dataset such as the RGI is never perfect nor complete. Whereas, obvious errors such as too much seasonal snow being mapped as glaciers, wrongly mapped debris-covered glaciers or (frozen) lakes or missing glacier parts due to clouds can be detected and corrected, variability in the interpretation (is this a rock glacier or a debris-covered glacier?, where is the drainage divide?) or topological issues (is this one ice cap or many glaciers?) are much harder to address. They will persist in future versions of the RGI as there is no unique right or wrong answer to these questions. In the end, a user of the dataset can always consult the larger GLIMS database when searching for an alternative interpretation of glacier extent or the timing of the outline does not fit to the intended application. Apart from the glacier mapping itself that should become more precise over time as increasingly high resolution satellite images (e.g., Sentinel-2) are available (Paul et al., 2016), the extraction of a "new" RGI version from the GLIMS database is not a button-press application but requires considerable effort. It is yet unclear if funding will be available for this in the future. The creation of RGI version 7 is largely automated now so that future RGI versions can be extracted from the GLIMS glacier database according to a set of pre-scribed criteria with limited effort. However, due to topological inconsistencies and the different internal handling of glacier datasets the creation of this automation has been time-consuming.

Glacier Thickness Database

GlaThiDa is the only worldwide database of glacier ice thickness observations, and thus plays an important role in studies of glacier ice volumes and their potential sea-level rise contributions (e.g., Farinotti et al., 2017; MacGregor et al., 2021). The measurements are compiled from literature reviews (e.g., Gärtner-Roer et al., 2014), imported from published datasets, or submitted by researchers in response to calls for data. While major versions of GlaThiDa are archived at the WGMS (e.g., https://doi.org/10.5904/wgms-glathida-2020-10), the dataset is developed online as a version-controlled "git"

repository (https://gitlab.com/wgms/glathida). The development environment (described in Welty et al., 2020) automatically records changes to the dataset, continuously checks the integrity of the dataset, and provides an interface for bug reports, feature requests, and other community dialogue. Although a few suspicious ice thickness measurements have been flagged manually, source data are not automatically checked for plausibility, and in some cases they may be very wrong (e.g., https://gitlab.com/wgms/glathida/-/issues/25). Additional checks could be developed to automatically flag data that are inconsistent with neighboring measurements, modeled ice thicknesses (e.g., Farinotti et al., 2019), or glacier outlines (e.g., GLIMS, RGI).

Glacier Map Collection

Many glacier maps were published in the FoG reports between 1967 and 2012. They often show individual glaciers and their spatio-temporal changes in very detailed mappings, some of them with outstanding quality. Several glaciers, e.g., Lewis glacier (Kenya), were mapped repeatedly over many decades. This additional dataset complements the FoG database with more qualitative and comprehensive environmental information. To enable a direct access and use, the maps were digitized and made available online in 2018 (wgms.ch/products_fog_maps). Sporadically, additional maps (newly created and digitized old maps) are added to the collection.

Glacier Photograph Collection

The Glacier Photograph Collection (National Snow and Ice Data Center, 2021) is an online (https://nsidc.org/data/glacier_photo/search/), searchable database of digital photographs of glaciers from around the world, some dating back to the mid-19th century, which provide a historical reference for glacier extent. The photos are either scanned from physical objects such as photographic prints or slides or they originated in digital form from a digital camera. As of May 2022, the database contains over 25,500 photographs. Most of the photographs are of glaciers in the Rocky Mountains of North America, the Pacific Northwest, Alaska, and Greenland. However, the collection does include a smaller number of photos of glaciers in Europe, South America, the Himalayas, and Antarctica. The collection includes a number of sub-collections or Special Collections that are distinguished in some way. For example, there is a special collection of Repeat Photography of glaciers that provides a unique look at changes in glaciers over time. These photographs constitute an important historical record, as well as a data collection of interest to those studying the response of glaciers to climate change. Educators use the photographs frequently and artists have found inspiration in the photographs.

The collection is maintained by NSIDC. New photographs are submitted from a wide community and are added to the database sporadically. The collection is accessible on the NSIDC website, using a detailed search interface that allows request for regional or national data and individual glaciers, as well for specific years and single photographers.

Data Stewardship Assessment: Parameters

The assessment of the different GTN-G datasets is performed by the compilation of individual maturity matrices. Each matrix compiles all information on preservability, accessibility, sustainability, quality, reproducibility, and integrity of the data and metadata in each dataset following the approach by Peng et al. (2015, 2019), which is explained in more detail below. The assessment is based on the conceptual model and the related scoreboard presented by Peng et al. (2015). The individual evaluation criteria are slightly adapted to the "language" of glaciologists (see **Table 1**). The individual assessments of the seven datasets are compiled in a separate score table (**Table 2**).

This maturity scale contains nine key components. For each dataset, a maturity score from 1 to 5 is assigned, representing five levels of maturity. The levels range from Level 1, corresponding to a dataset that was developed *ad-hoc* and that is not managed, to Level 5, representing a dataset that is optimally managed and developed on the long-term and that is externally audited (**Table 1**). The assessment was compiled by the authors of the present paper who are managers of the different GTN-G datasets. This expert evaluation is characterized by a multi-step approach. First, each manager completed a full assessment of their respective dataset (self-assessment) based on their interpretation of the criteria as compiled in **Table 1** and their reading of the original work by Peng et al. (2015). Second, as each dataset is run by several experts, the individual assessments were discussed with the other people responsible for the dataset in an iterative process to achieve a consensus. In a last iteration, this consensus assessment was presented to and discussed with the managers of the other datasets, representing also the GTN-G Executive Board. Related key words for the assignment of the score are given in **Table 1** for each assessment criteria (key component). In the following, the assessment criteria are described from a more glaciological perspective:

Preservability

Are there any archiving standards (e.g., CoreTrustSeal) for the dataset? Is there redundancy? Do the archiving processes follow certain standards? Is there any predictive planning for future changes?

Accessibility

Are the data publicly available? Do the data services follow the sense of community standards? Is there additional dissemination of data products to enhance data accessibility for different user groups?

Usability

Data format: is it standard/non-standard? Are there interoperable formats? Is the available metadata adequate and in a usable form? Are the data and metadata sufficiently documented? Is there any need for specific knowledge to use the data? Are there online capabilities available, such as visualizations or a product user guide?

TABLE 1 | Maturity scale applied for the assessment of the GTN-G datasets [modified after Peng et al. (2015, 2019)].

Maturity scale / GTN-G data sets (across)	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Key components (below)	***Ad-hoc* not managed**	**Minimal managed limited**	**Intermediate managed defined partially implemented**	**Advanced managed well-defined fully implemented**	**Optimal level 4 + measured controlled, audit**
Preservability: (The state of being preservable)	Any data storage location data only (no metadata)	Non-designated repository data storage redundant limited archiving metadata	Designated repository (e.g., CoreTrustSeal) data storage redundant community-standard archiving metadata limited archiving standards apply	Designated repository (e.g., CoreTrustSeal) data storage redundant community-standard archiving metadata community archiving standards apply	Designated repository (e.g., CoreTrustSeal) data storage redundant community-standard archiving metadata community archiving standards apply archiving process monitored and audited planned future data archiving
Accessibility: (The state of being searchable and accessible publicly)	Data not publicly available data access person-to-person	Data is publicly available direct file download possible data searchable online (on dataset level)	Data is publicly available direct file download possible non-standard data service provided limited data server performance data searchable online (on granule/file level) limited search metrics	Data is publicly available direct file download possible community-standard data service provided enhanced data server performance data searchable online (on granule/file level) community search metrics internal dissemination report	Data is publicly available direct file download possible community-standard data service provided enhanced data server performance data searchable online (on granule/file level) community search metrics dissemination report available online planned future data accessibility
Usability: (The state of being easy to use)	Specific knowledge required no documentation online	Non-standard data format limited documentation online	Community-standard data format (incl. Metadata) documentation online	Community-standard data format (incl. Metadata) documentation online basic data characterization online	Community-standard data format (incl. Metadata) documentation online enhanced data characterization online (e.g., visualization) community metrics of data characterization online external ranking
Production sustainability: (The state of data production being sustainable and extendable)	*Ad-hoc* initiative no deliverables existing	Short-term initiative individual commitment by PI's	Medium-term initiative institutional commitment	Long-term initiative (program) institutional commitment product improvement process in place	Long-term initiative (program) national or international commitment planned product improvement
Data quality assurance (DQA): (The state of data quality being assured)	DQA procedure unknown or inexistent	DQA procedure random DQA procedure not defined and documented	DQA procedure defined, documented and partially implemented	DQA procedure well-documented, fully implemented and available online limited DQA metadata available	DQA procedure well-documented, fully implemented, available online, monitored and reported community-standard DQA metadata available external review of DQA
Data quality control/monitoring: (The state of data quality being controlled and monitored)	No quality monitoring of data and metadata	Limited monitoring of data and metadata	Regular monitoring of data and metadata, not automatic	Fully automatic monitoring of data and metadata following community standards consistency checks	Fully automatic monitoring of data and metadata following community standards consistency checks provider/user feedback in place planned future data quality control
Data quality assessment: (The state of data quality being assessed)	Method and theoretical basis assessed	Method and theoretical basis assessed research product assessed	Method and theoretical basis assessed research product assessed operational product assessed	Method and theoretical basis assessed research product assessed operational product assessed quality metadata assessed	Method and theoretical basis assessed research product assessed operational product assessed quality metadata assessed assessments performed on recurring basis and following community standards external evaluation

(Continued)

TABLE 1 | Continued

Maturity scale / GTN-G data sets (across)	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Key components (below)	***Ad-hoc* not managed**	**Minimal managed limited**	**Intermediate managed defined partially implemented**	**Advanced managed well-defined fully implemented**	**Optimal level 4 + measured controlled, audit**
Transparency/traceability: (The state of being transparent, trackable, and traceable)	Limited product information available knowledge transfer person-to-person	Product information available in the literature	Product information available	Product information online Unique Object Identifier (OID) assigned Digital Object Identifier (DOI) assigned	Complete data provenance online Unique Object Identifier (OID) assigned Digital Object Identifier (DOI) assigned planned future data traceability
Data integrity: (The state of data integrity being verifiable)	Unknown or no data ingest integrity check	Data ingest integrity verifiable	Data ingest integrity verifiable data archive integrity verifiable	Data ingest integrity verifiable data archive integrity verifiable data access integrity verifiable	Data ingest integrity verifiable data archive integrity verifiable data access integrity verifiable data authenticity verifiable (hash codes, digital signatures) performance of data integrity checks

Production Sustainability

To what extent is there a commitment and stewardship, from individuals (e.g., Principal Investigators) to organizations/services (e.g., WGMS)? What is the rating of the dataset, ranging from *ad-hoc* initiatives (with or without deliverables) to long-term programs secured through national or international funding?

Data Quality Assurance

Is a DQA procedure implemented? Is the procedure manual or automated? Is there sufficient documentation of the DQA? Are there any reports about the DQA, according to community standards and with external review?

Data Quality Control/Monitoring

Is the data quality controlled and monitored based on a regular sampling and analysis? Is there a systematic and/or an automatic procedure? Are there regular consistency checks following community standards? Are provider and user feedback mechanisms in place?

Data Quality Assessment

Are there quality reports for methods and results? Is there sufficient metadata about quality assessment? Is there an assessment on a recurring basis? Is there an external evaluation?

Transparency/Traceability

Is there (online) product information available? Is the data provenance sufficiently documented and are there related operational algorithms? Are the data governance mechanisms online available? Is all information important for reproducibility available?

Data Integrity

Are there integrity checks? How do they perform, are they verifiable? Integrity checks should address the following: ingestion of data, data archiving, data access, data authenticity. Is there a monitoring and reporting of the performance of data integrity checks?

RESULTS: PERFORMANCE OF THE GTN-G DATASETS

Table 2 summarizes the performance of the evaluated GTN-G datasets. A score from 1 to 5 was assigned for each key component, which is explained with the comments given in **Table 1**.

Fluctuations of Glaciers

The FoG database performs between advanced and optimal. It is a designated repository for glacier fluctuations data following standards of the glaciological community and key standards regarding archiving quality and security. Data are accessible through different channels. Data quality assurance (DQA) is manually ensured, but not automatically enforced. Data integrity checks are not fully automatic. Each version of the dataset is identified by its own DOI and the provenance of the data is documented in detail both in the metadata and the database itself.

TABLE 2 | Summary matrix with the performance of the seven datasets available within GTN-G.

Maturity scale / GTN-G data sets (across)	FoG	WGI	GLIMS	RGI	GlaThiDa	GMC	GPC
Preservability: (The state of being preservable)	4	4	3	3	4	2	4
Accessibility: (The state of being searchable and accessible publicly)	4	3	4	4	3	2	3
Usability: (The state of being easy to use)	5	4	4	4	4	2	3
Production sustainability: (The state of data production being sustainable and extendable)	5	3	3	3	3	1	4
Data quality assurance (DQA): (The state of data quality being assured)	3	3	2	2	3	1	2
Data quality control/monitoring: (The state of data quality being controlled and monitored)	3	2	3	3	4	2	2
Data quality assessment: (The state of data quality being assessed)	5	2	2	2	2	1	2
Transparency/traceability: (The state of being transparent, trackable, and traceable)	5	4	2	2	4	3	3
Data integrity: (The state of data integrity being verifiable)	3	3	3	3	2	1	3

A score from 1 to 5 is given for each key component. The traffic-light colors give an additional hint to the maturity of the single datasets: green colors: good performance, yellow: medium performance, orange: limited performance.

TABLE 3 | List of all GTN-G datasets with their URL (Uniform Resource Locator) and citation.

Dataset	URL	Citation
Fluctuations of Glaciers (FoG)	https://dx.doi.org/10.5904/wgms-fog-2021-05	WGMS, 2021a: Fluctuations of Glaciers Database. World Glacier Monitoring Service, Zurich, Switzerland. doi: 10.5904/wgms-fog-2021-05
World Glacier Inventory (WGI)	https://nsidc.org/data/glacier_inventory/index.html	WGMS, and National Snow and Ice Data Center (2012) World Glacier Inventory. Compiled and made available by the World Glacier Monitoring Service, Zurich, Switzerland, and the National Snow and Ice Data Center, Boulder CO, USA. Digital Media
Global Land Ice Measurements from Space (GLIMS) Initiative	https://www.glims.org	GLIMS and National Snow and Ice Data Center (2021): Global Land Ice Measurements from Space glacier database. Compiled and made available by the international GLIMS community and the National Snow and Ice Data Center, Boulder CO, U.S.A. doi: 10.7265/N5V98602
Randolph Glacier Inventory (RGI)	https://www.glims.org/RGI/index.html	RGI Consortium (2017): Randolph Glacier Inventory – A Dataset of Global Glacier Outlines: Version 6.0: Technical Report, Global Land Ice Measurements from Space, Colorado, USA. Digital Media. https://doi.org/10.7265/N5-RGI-60
Glacier Thickness Database (GlaThiDa)	https://www.gtn-g.ch/data_catalog_glathida/	GlaThiDa Consortium (2020): Glacier Thickness Database 3.1.0. World Glacier Monitoring Service, Zurich, Switzerland. doi: 10.5904/wgms-glathida-2020-10
Glacier Map Collection (GMC)	https://wgms.ch/products_fog_maps/	WGMS (2018): Glacier Map Collection (GMC), World Glacier Monitoring Service, Zurich, Switzerland. doi: 10.5904/wgms-maps-2018-02
Glacier Photograph Collection (GPC)	https://nsidc.org/data/glacier_photo/	National Snow and Ice Data Center (2021): Glacier Photograph Collection, Version 1. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. https://doi.org/10.7265/N5/NSIDC-GPC-2009-12

For this advanced performance, the FoG database was already certified as trustworthy repository by CoreTrustSeal in 2019.

World Glacier Inventory

The WGI dataset performs between intermediate and advanced. It is a well-managed dataset with clearly defined aims and purposes. Lower scores stem from the data quality control/monitoring (limited monitoring of data and metadata) as well as the data quality assessment (assessment is performed of the research product, but not of the dataset itself). As this dataset represents a snapshot from 1989, with an update from 2012, data curation is currently non-existent. Hence, the overall performance is low.

Global Land Ice Measurements From Space

The GLIMS database performs with an overall score of intermediate. While the database is accessible through an

enhanced data server with provided dissemination metrics according to community standards, a clear backdrop is the largely voluntary basis of data provision by scientists from all over the world. This leads to data contributions happening by chance from research projects, with a wide range of interpretation of glacier extent and limited control on data quality and uncertainty assessment. Although several guidelines exist for the community (Raup and Khalsa, 2007; Paul et al., 2009), various quality checks have to be performed before data ingest. This includes file formats, completeness of metadata, topologic errors, location errors, outline quality, etc. Despite automated tools being available for parts of this work, this still requires effort, especially for less standard data formatting.

Randolph Glacier Inventory

The dataset performs similar to the GLIMS database, with an overall score of intermediate. Data quality assurance (DQA) takes place on an *ad-hoc* basis, is not systematic, and is dependent on the data provider. Standard checks for data quality control are implemented but not documented. Data transparency is low, because product info can only be found in the literature that documents the submitted data (but it is available and citable).

Glacier Thickness Database

The dataset achieves an overall performance between intermediate and advanced. There is a designated repository for the dataset (WGMS), which stores all versions and their metadata, and ensures public and direct access to the files. Data and metadata are machine-readable following community standards. Data quality checks are performed systematically and automatically, and are all documented either in the metadata or in the source code. A lower performance stems from relatively low production sustainability, as there is only a medium-term commitment from the data repository to further develop the dataset. Only cursory data integrity checks are performed, but all changes to the files are tracked in a version-controlled (git) repository.

Glacier Map Collection

This collection performs between minimal and *ad-hoc* for management of the dataset, which currently consists of *ad-hoc* initiatives (though regular inclusion of new maps in former times). Data are accessible online, but there is only limited documentation of the data itself. DQA procedures are random and only the method and its theoretical basis are assessed. There are no data integrity checks performed.

Glacier Photograph Collection

The collection is preserved at a designated repository (NSIDC) with well-formed dataset and file-level metadata, following high community-archiving standards. There is a direct and public access of the data, with some search metrics provided. Long-term commitment by the data repository ensures the production sustainability. DQA procedures are performed but not defined or documented. Data product information is available in a user guide and data integrity checks are in place. This leads to an overall performance of intermediate (advanced in a few criteria).

DISCUSSION OF OPPORTUNITIES AND CHALLENGES

The maturity matrix approach (Peng et al., 2015) applied in this study allows a clear and comprehensive assessment of the individual glacier datasets, as well as a cross-comparison to other datasets. Similar assessment schemes for maturity matrices are available (Bates and Privette, 2012; Albani and Maggio, 2020; CEOS, 2020), often with very similar parameters as they are predominantly applied in environmental sciences. For example, the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) uses the maturity matrix to assess the maturity of climate data records and the development of Essential Climate Variables (ECVs). EUMETSAT applies the systematic approach by Bates and Privette (2012) to assess if the data record generation follows best practices in the areas of science, information preservation, and usage of the data. This approach was also used when preparing the Copernicus Climate Change Service (C3S) and assessing the needs for full access to standardized climate change data. In this C3S context, the FoG and RGI glacier datasets were also evaluated regarding the availability and quality of metadata, user documentation, uncertainty characterization, public access, and usage. A comparison of the C3S assessment with the outcome of this study reveals a congruent performance.

The assessment of glacier datasets showed that most of the datasets perform on an intermediate level. Given the individual significance of the datasets, the most important ones, when it comes to basic data on glacier distribution and glacier changes, are managed on a long-term perspective, but have only limited funding.

Historical Development

The current state of the GTN-G datasets can largely be explained through their historical development. The glacier fluctuation dataset (FoG) traces back to the end of the 19th century, when the worldwide coordination of glacier monitoring was initiated. With time, the uninterrupted continuation of the data collection has become a strong argument to further institutionalize the collection of glacier data. This led to the formation of the Permanent Service on the Fluctuation of Glaciers (PSFG) in 1967, under the umbrella of international auspice organizations, and later in 1986 to the formation of the WGMS. The commitment of the coordinators of this network as well as the dedication of many investigators and collaborators in turn helped to emphasize the achievements and positive reception of the services. Different challenges that emerged during that time had to be tackled, and different needs from data users, data producers, and international organizations have to be satisfied by the international data centers. As a consequence, this development is also reflected in the GTN-datasets as presented of today.

This can be seen in several examples. First, FoG emerged from simple length change measurements and later on included *in situ,* geodetic and point mass balances. Hence, the dataset has become more comprehensive, but this also needed more effort for maintenance and continued support. Second, GLIMS developed

from glacier outlines for individual regions to a globaly near-complete and partly multi-temporal glacier inventory, forming the base for the RGI that reached completeness with a different data model independently and was in turn ingested into GLIMS to get it spatially complete. To maintain the RGI for the long-term, it is now also provided via GLIMS and the NSIDC. These examples show the independent development of each dataset while maintaining links to the other GTN-G datasets, such as GPC or GlaThiDa, which are steadily increasing in data richness.

We note that GTN-G has developed in a research environment and, hence, has never reached the support levels of operational monitoring networks such as within the WMO with its national meteorological services. In summary, there is a long history in glacier monitoring and so far, a good job was done, but with the increasing amount of data, the challenges and requirements (from different users) have changed and need to be tackled. This is only possible with data curation and stewardship.

Data Curation and Stewardship

The increasing amount of data from different sources pushes most storage systems to their capacity limits and require regular expansion of the hardware, including mirror sites. To ensure fast and long-term data access, constant updates of the software are required as well. Further, the increasing demand for direct access to most up-to-date information is a common desire of many data users. Therefore, data feeds, checks and updates must be carried out continuously. In parallel, proper dataset versioning is needed to guarantee traceability. Most of these challenges come with technical demands of increasing complexity.

While the NSIDC offers a rather large infrastructure for data repositories but has limited funding for active (hands-on) data curation, the WGMS is a small service that has its strengths in data analysis with a strong focus on one specific database but limited capacities to host additional datasets. Hence, for the different operational bodies, individual data curation strategies need to be set up, evaluated and revised on a regular basis and the responsible database manager(s) need to run consistent procedures of data archiving, access and quality checks. Regular training and exchange with colleagues from the glacier community would also be an advantage to take up current challenges quickly and become responsible data stewards. Following these procedures will professionalize the repositories, strengthen the data services, and serve the community of data providers and users optimally.

The best-practice measures mentioned above of course come at a price. In addition to upkeep of technical equipment (hardware, software), science officers and database managers need to be trained technically and substantively to ensure a qualified data processing chain. To bolster support for adequate technical equipment and staff training for the different data services, support is needed from higher-level agencies or international organizations. They are the only ones that can commit and contribute to the data services for the long run. Therefore, lobbyists are needed who communicate the recent shortcomings and challenges to the responsible decision makers. In the case of GTN-G, this task could be taken on by the Advisory Board, as they know the glacier community sufficiently well and have the right contacts to international organizations.

Funding Situation

From the assessment, we noted a direct relationship between the scores of the datasets and the respective funding available to maintain it. Funding often comes from research projects that cover at most the next few years. In these cases, a long-term perspective is lacking, since follow-up projects that would ensure a direct continuation are often not guaranteed, or even dismissed due to the "lack of innovation." Hence, existing structures first need to be sustained for a more long-term operation.

Within GTN-G, the funding situation currently looks as follows: the only dataset with dedicated long-term funding for data management is the FoG dataset (with 3 FTE (full time equivalent), Swiss GCOS 2021-24; C3S 312b 2020-21). In addition, the RGI runs on short-term funding (1 FTE, C3S 312b 2020-21), as does GLIMS (0.5 FTE, NASA Distributed Active Archive Center funding). The other datasets are updated on a more voluntary or *ad-hoc* basis without dedicated funding, although the WGI and the GPC are minimally maintained with support from the NOAA Cooperative Agreement with CIRES, NA17OAR4320101. In the future, *ad-hoc* data compilations, such as GlaThiDa, will be easier to fund, as they can build on existing structures and can be linked to scientific projects or sponsored by societies such as IACS. On the other hand, long-term monitoring necessitates a long-term commitment, which is more difficult to secure funding for. Running trustworthy repositories needs long-term security and perspective. Dedicated support and long-term commitment for certified data repositories build the basis for the successful democratization of data.

In the field of glacier data, this balancing act has so far been successfully achieved through joint collaboration between data repository institutions, data providers, and data users. However, the money spent on the data provider and user side for creating and working with the datasets (generally, scientific projects) is several orders of magnitude larger than the funds available for data curation. Hence, international organizations as well as national authorities must offer support and take responsibility on both sides. Most challenges can only be overcome in a financially safe and secure setting for data services and with the help of international standardization, as, for example, provided by the CoreTrustSeal.

For the GLIMS glacier database, the funding situation is too low to elevate its maturity score. GLIMS was started and maintained for some years on short-term (3–5 year) project funds, but has recently been folded into the NSIDC DAAC, funded by NASA. Current GLIMS activities are being performed mainly by one person at a 40% engagement, with other software developers contributing on occasion. The database has some issues that need to be improved to reach a higher standard, but without sufficient and sustained funding of the required experts this is difficult, or too slow. Given the importance of this database for many other multi-million-dollar projects, the limited funding available for keeping the database healthy and growing is more than shameful. We acknowledge, however, that this is also a result of the historical development up to the current explosion

of available datasets and recently changed user demands and possibilities (e.g., cloud computing).

Future Ambitions

In respectful view of the historical developments and the awareness of the recent challenges, the individual GTN-G data services need to take urgent actions. Minimum actions are required to simply keep the state-of-the-art. Far-reaching measures must be taken to secure the future of data services and their benefits for the entire community and to serve different stakeholders–experts, policy makers, the interested public and journalists.

As mentioned in Data Curation and Stewardship, individual data curation strategies are needed for the operational bodies of GTN-G. Trained database managers will have to organize and monitor the implementation of this strategy, and run consistent procedures for data archiving and to perform access and quality checks on a regular basis. In addition, different outreach products need to be compiled for different levels of data users; while direct data access is suitable for experts, decision makers need well-condensed policy briefs and journalists often request individual mentoring. By providing these services, the management of the repositories will be professionalized and ready to serve the entire community.

To address these future ambitions, problems of technical equipment and the hiring and long-term retention of qualified personnel must be tackled. Both aspects are required for proper data curation and dissemination of glacier datasets. Hence, in the future GTN-G has to find long-term funding to run all datasets in a mature and sustainable way and serve the community with FAIR and trustworthy glacier data of the best quality.

CONCLUSIONS

Dedicated support and long-term commitment for certified data repositories build the basis for the successful democratization of data. In the field of glacier data, this balancing act has so far been achieved through joint collaboration between data repository institutions, data providers, and data users. From the comparison of seven glacier datasets (**Table 3**) available within the Global Terrestrial Network for Glaciers (GTN-G) we conclude:

- The current state of the GTN-G datasets can largely be explained through historical development, reflecting different needs from stakeholders incl. users.
- The GTN-G datasets have been developed in a research environment, hence long-term data curation and stewardship are absolutely necessary.
- Currently, datasets that are managed based on a mid- to long-term funding (e.g., the FoG dataset) have the highest maturity.
- Urgent action has to be taken to keep the state-of-the-art and individual data curation strategies need to be implemented and tailored for each operational body, considering the context (e.g., funding situation; project funds vs. long-term funding).
- These strategies need to be evaluated, revised, and adapted on a regular basis, which can be ensured through the GTN-G Advisory Board.
- Data curation requires constant updates of software to meet technical demands of increasing complexity and to provide direct access to most up-to-date information, which in turn needs proper data versioning.
- International standardization such as provided for example by the CoreTrustSeal contributes to a secure setting for the data services.
- Technical equipment, hiring professional staff and long-term retention of qualified personnel is key to offer the different services and to serve the entire community.

Most challenges can only be overcome in a financially safe and secure setting for data services. However, the money spent on the data provider and user side for creating and working with the datasets is several orders of magnitude larger than the funds available for data curation. Considering the importance of glacier data to answer numerous key environmental and societal questions (from water availability to global sea-level rise), this bias needs to be adjusted.

DATA AVAILABILITY STATEMENT

All data are available on the website of the Global Terrestrial Network for Glaciers (GTN-G; www.gtn-g.org) or on the specific websites (see **Table 3**).

AUTHOR CONTRIBUTIONS

IG-R, SN, and MZ conceived the study, assessed the FoG database, assessed the WGI, and assessed the GMC. BR, FP, and MZ assessed the GLIMS database and assessed the RGI. EW, MZ, and IG-R assessed the GlaThiDa. AW, FF, and SN assessed the GPC. IG-R and SN wrote the paper and produced the figures. All authors studied and commented on the selected methodology, reviewed all assessments, and commented on and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research has been supported by the Federal Office of Meteorology and Climatology MeteoSwiss within the framework of the Global Climate Observing System (GCOS) Switzerland. FP acknowledges additional funding from the ESA project Glaciers_cci (4000127593/19/I-NB). AW and FF acknowledge support from NOAA Cooperative Agreement with CIRES, NA17OAR4320101. BR acknowledges support for Global Land Ice Measurements from Space from National Aeronautics and Space Administration under the National Snow and Ice Data Center and Distributed Active Archive Center.

REFERENCES

Albani, M., and Maggio, I. (2020). Long time data series and data stewardship reference model, *Big Earth Data* 4, 353–366. doi: 10.1080/20964471.2020.1800893

Allison, I., Fierz, C., Hock, R., Mackintosh, A., Kaser, G., and Nussbaumer, S. U. (2019). IACS: past, present, and future of the international association of cryospheric sciences. *Hist. Geo Space Sci.* 10, 97–107. doi: 10.5194/hgss-10-97-2019

Bates, J. J., and Privette, J. L. (2012). A maturity model for assessing the completeness of climate data records, Eos trans. *AGU* 93, 441. doi: 10.1029/2012EO440006

Bayr, K. J., Hall, D. K., and Kovalick, W. M. (1994). Observations on glaciers in the eastern Austrian Alps using satellite data. *Int. J. Remote Sens.* 15, 1733–1742. doi: 10.1080/01431169408954205

Bolch, T., Menounos, B., and Wheate, R. (2010). Landsat-based inventory of glaciers in western Canada, 1985–2005. *Remote Sens. Environ.* 114, 127–137. doi: 10.1016/j.rse.2009.08.015

CEOS. (2020). *WGISS Data Management and Stewardship Maturity Matrix*. Version 1.3, Available online at: https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS%20Data%20Management%20and%20Stewardship%20Maturity%20Matrix.pdf (accessed May 30, 2022).

Cogley, J. G. (2009). A more complete version of the world glacier inventory. *Ann. Glaciol.* 50, 32–38. doi: 10.3189/172756410790595859

Farinotti, D., Brinkerhoff, D. J., Clarke, G. K. C., Fürst, J. J., Frey, H., Gantayat, P., et al. (2017). How accurate are estimates of glacier ice thickness? Results from ITMIX, the ice thickness models intercomparison experiment. *The Cryosphere* 11, 949–970 doi: 10.5194/tc-11-949-2017

Farinotti, D., Huss, M., Fürst, J. J., Landmann, J., Machguth, H., Maussion, F., et al. (2019). A consensus estimate for the ice thickness distribution of all glaciers on Earth. *Nat. Geosci.* 12, 168–173. doi: 10.1038/s41561-019-0300-3

Field, W. O. (1975). *Mountain Glaciers of the Northern Hemisphere*. CRREL, Hanover, Vol. 1, 698p, Vol. 2, 932p and an Atlas with 49 plates.

Frey, H., Paul, F., and Strozzi, T. (2012). Compilation of a glacier inventory for the western Himalayas from satellite data: methods, challenges, and results. *Remote Sens. Environ.* 124, 832–843. doi: 10.1016/j.rse.2012.06.020

Gärtner-Roer, I., Naegeli, K., Huss, M., Knecht, T., Machguth, H., and Zemp, M. (2014). A database of worldwide glacier thickness observations. *Glob. Planet. Change* 122, 330–344. doi: 10.1016/j.gloplacha.2014.09.003

Gärtner-Roer, I., Nussbaumer, S. U., Hüsler, F., and Zemp, M. (2019). Worldwide assessment of national glacier monitoring and future perspectives. *Mt. Res. Dev.* 39, A1–A11. doi: 10.1659/MRD-JOURNAL-D-19-00021.1

GlaThiDa Consortium (2020). *Glacier Thickness Database 3.1.0*. Zurich, Switzerland: World Glacier Monitoring Service. doi: 10.5904/wgms-glathida-2020-10

GLIMS Consortium (2005). *GLIMS Glacier Database, Version 1*. Boulder, CO: NASA National Snow and Ice Data Center Distributed Active Archive Center.

Guo, W., Liu, S., Xu, J., Wu, L., Shangguan, D., Yao, X., et al. (2015). The second Chinese glacier inventory: data, methods and results. *J. Glaciol.* 61, 357–372. doi: 10.3189/2015JoG14J209

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., et al. (2021). Accelerated global glacier mass loss in the early twenty-first century. *Nature*. 59, 726–731. doi: 10.1038/s41586-021-03436-z

IPCC (2021). "Climate change 2021: the physical science basis," in *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, eds V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge, UK; New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157896

MacGregor, J. A., Studinger, M., Arnold, E., Leuschen, C. J., Rodríguez-Morales, F., and Paden, J. D. (2021). Brief communication: an empirical relation between center frequency and measured thickness for radar sounding of temperate glaciers. *The Cryosphere* 15, 2569–2574. doi: 10.5194/tc-15-2569-2021

Mercer, J. H. (1967). *Southern Hemisphere Glacier Atlas. American Geographical Society, US Army Natick Laboratories Technical Report,* Natick, 325p.

Müller, F. (1978). *Instructions for the Compilation and Assemblage of Data for a World Glacier Inventory*. Supplement: Identification/glacier number. Temporary Technical Secretariat for the World Glacier Inventory.Zurich: Swiss Federal Institute of Technology.

Müller, F., Caflisch, T., and Müller, G. (1977). *Instructions for Compilation and Assemblage of Data for a World Glacier Inventory*. Temporary Technical Secretariat for the World Glacier Inventory. Zurich: Swiss Federal Institute of Technology.

National Snow and Ice Data Center (2021). *Glacier Photograph Collection, Version 1*. Boulder, CO: NSIDC: National Snow and Ice Data Center.

National Snow and Ice Data Center (2021) *Glacier Photograph Collection, Version 1*. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. doi: 10.7265/N5/NSIDCGPC-2009-12

Nussbaumer, S. U., Hoelzle, M., Hüsler, F., Huggel, C., Salzmann, N., and and, M., Zemp (2017). Glacier monitoring and capacity building: important ingredients for sustainable mountain development. *Mt. Res. Dev.* 37, 141–152. doi: 10.1659/MRD-JOURNAL-D-15-00038.1

Paul, F., Barry, R. G., Cogley, J. G., Frey, H., Haeberli, W., Ohmura, A., et al. (2009). Recommendations for the compilation of glacier inventory data from digital sources. *Ann. Glaciol.* 50, 119–126. doi: 10.3189/172756410790595778

Paul, F., and Kääb, A. (2005). Perspectives on the production of a glacier inventory from multispectral satellite data in Arctic Canada: cumberland Peninsula, Baffin Island. *Ann. Glaciol.* 42, 59–66. doi: 10.3189/1727564057818 13087

Paul, F., Kääb, A., Maisch, M., Kellenberger, T., and Haeberli, W. (2002). The new remote-sensing-derived Swiss glacier inventory: I. Methods. *Ann. Glaciol.* 34, 355–361. doi: 10.3189/172756402781817941

Paul, F., Winsvold, S. H., Kääb, A., Nagler, T., and Schwaizer, G. (2016). Glacier remote sensing using Sentinel-2. Part II: Mapping glacier extents and surface facies, and comparison to landsat 8. *Remote Sens.* 8, 575, doi: 10.3390/rs8070575

Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., and Ansari, S. (2015). A unified framework for measuring stewardship practices applied to digital environmental data s. *Data Sci. J.* 13, 231–252. doi: 10.2481/dsj.14-049

Peng, G., Wright, W., Baddour, O., Lief, C., and the SMMCD Work Group (2019). *The Guidance Booklet on the WMO-Wide Stewardship Maturity Matrix for Climate Data*. Figshare.

Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., et al. (2014). The randolph glacier inventory: a globally complete inventory of glaciers. *J. Glaciol.* 60, 537–552. doi: 10.3189/2014JoG13J176

Pospiech, M., and Felden, C. (2012). "Big data – A state-of-the-art," in *AMCIS (Americas Conference on Information Systems) 2012 Proceedings.* Available online at: https://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22

Rastner, P., Bolch, T., Mölg, N., Machguth, H., Le Bris, R., and Paul, F. (2012). The first complete inventory of the local glaciers and ice caps on Greenland. *The Cryosphere* 6, 1483–1495. doi: 10.5194/tc-6-1483-2012

Raup, B., and Khalsa, S. J. S. (2007). *GLIMS Analysis Tutorial. Global Land Ice Measurements from Space (GLIMS)*, 15p. Available online at: https://www.glims.org/MapsAndDocs/assets/GLIMS_Analysis_Tutorial_a4.pdf

Raup, B., Racoviteanu, A., Khalsa, S. J. S., Helm, C., Armstrong, R., and Arnaud, Y. (2007). The GLIMS geospatial glacier database: a new tool for studying glacier change. *Glob. Planet. Change* 56, 101–110. doi: 10.1016/j.gloplacha.2006.07.018

Raup, B. H., Kieffer, H. H., Hare, T. M., and Kargel, J. S. (2000). Generation of data acquisition requests for the ASTER satellite instrument for monitoring a globally distributed target: glaciers. *IEEE Trans. Geosci. Remote Sens.* 38, 1105–1112. doi: 10.1109/36.841989

RGI Consortium (2017). Randolph Glacier Inventory – a data set of global glacier outlines: version 6.0. Technical report. *Global Land Ice Measurements from Space (GLIMS),* Boulder, CO, Digital Media.

Scherler, K. E. (1983). *Guidelines for Preliminary Glacier Inventories*. GEMS, UNEP, UNESCO, ICSI, ETH-Z. Temporary Technical Secretariat for the World Glacier Inventory. Zurich: Swiss Federal Institute of Technology.

UNESCO/IASH. (1970). *Perennial Ice and Snow Masses*. A guide for compilation and assemblage of data for a world inventory. Technical Papers in Hydrology No. 1. Paris: United Nations Educational, Scientific and Cultural Organization.

Vaughan, D. G., Comiso, J. C., Allison, I., Carrasco, J., Kaser, G., Kwok, R., et al. (2013). *Observations: Cryosphere. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker, T. F., Qin, D., Plattner,

G. K., Tignor, M., Allen S. K., Boschung J., Nauels, A., Xia, Y., Bex, V., and Midgley P. M (Cambridge University Press, Cambridge, and New York, NY).

Welty, E., Zemp, M., Navarro, F., Huss, M., Fürst, J. J., Gärtner-Roer, I., et al. (2020). Worldwide version-controlled database of glacier thickness observations. *Earth Syst. Sci. Data* 12, 3039–3055. doi: 10.5194/essd-12-3039-2020

WGMS (1989). *World Glacier Inventory – Status 1988*, eds Haeberli, W., Bösch, H., Scherler, K., Østrem, G. and Wallén, C. C. IAHS(ICSI)/UNEP/UNESCO, World Glacier Monitoring Service, Zurich, Switzerland, 458.

WGMS (1998). Into the second century of worldwide glacier monitoring: prospects and strategies, eds Haeberli, W., Hoelzle, M, and S. Suter. *Studies and Reports in Hydrology*. Paris: UNESCO Publishing, 227.

WGMS (2018). *Glacier Map Collection (GMC)*. Zurich, Switzerland: World Glacier Monitoring Service. doi: 10.5904/wgms-maps-2018-02

WGMS (2021a). *Fluctuations of Glaciers Database*. Zurich: World Glacier Monitoring Service.

WGMS (2021b). Global Glacier Change Bulletin No. 4 (2018-2019). Zemp, M., Nussbaumer, S.U., Gärtner-Roer, I., Bannwart, J., Paul, F., and Hoelzle, M, eds *ISC(WDS)/IUGG(IACS)/UNEP/UNESCO/WMO*. Zurich: World Glacier Monitoring Service, 278.

WGMS, and National Snow and Ice Data Center (2012). *World Glacier Inventory, Version 1. [Indicate Subset Used]*. Boulder, CO: NSIDC: National Snow and Ice Data Center.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi: 10.1038/sdata.2016.18

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., and Woodcock, C. E. (2012). Opening the archive: how free data has enabled the science and monitoring promise of Landsat. *Remote Sens. Environ.* 122, 2–10. doi: 10.1016/j.rse.2012.01.010

Zemp, M. (2011). *The Monitoring of Glaciers at Local, Mountain, and Global Scale. Schriftenreihe Physische Geographie; Glaziologie und Geomorphodynamik 65. University of Zurich, Faculty of Science*, 72p.

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

The reviewer LT declared a past collaboration with several of the authors IG-R, SN, FP, and MZ to the handling Editor.

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

frontiers | Frontiers in Climate

TYPE Perspective
PUBLISHED 25 August 2022
DOI 10.3389/fclim.2022.958533

OPEN ACCESS

EDITED BY
Tiffany C. Vance,
U.S. Integrated Ocean Observing System, United States

REVIEWED BY
Ted Habermann,
Metadata Game Changers, United States

*CORRESPONDENCE
Rosalie R. Rossi
rosalie.rossi@tamucc.edu

SPECIALTY SECTION
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

RECEIVED 31 May 2022
ACCEPTED 29 July 2022
PUBLISHED 25 August 2022

CITATION
Rossi RR, LeBel DA and Gibeaut J (2022) Growing pains of a data repository: GRIIDC's evolution from environmental disaster rapid response to promoting FAIR data. *Front. Clim.* 4:958533. doi: 10.3389/fclim.2022.958533

Growing pains of a data repository: GRIIDC's evolution from environmental disaster rapid response to promoting FAIR data

Rosalie R. Rossi*, Deborah A. LeBel and James Gibeaut

Harte Research Institute for Gulf of Mexico Studies, Texas A&M University — Corpus Christi, Corpus Christi, TX, United States

GRIIDC is a multidisciplinary data repository created in the aftermath of the Deepwater Horizon oil spill. Development of the repository occurred even as researchers collected post-spill data, and as a result, the data management system initially focused on the ingestion of data and metadata. Data sharing was not as prevalent as it is currently, and many researchers were not familiar with data sharing and data organization best practices. Implementation of data management planning, submission, citation, and distribution features required many iterations and occurred while GRIIDC was assisting researchers with managing their rapid response data. From this challenging beginning, over the decade since the Deepwater Horizon oil spill, GRIIDC has improved the data management system and the training of researchers, which has enhanced the ease of submission and quality of data submitted. The GRIIDC system has also evolved to prioritize the implementation of FAIR data principles to ensure the data are findable, accessible, interoperable, and reusable. All data are issued digital object identifiers (DOIs) through DataCite and are findable via GRIIDC's data search page, DataONE, and Google Dataset Search. Each dataset has a landing page where the data and metadata can be accessed. GRIIDC is continuously striving to add FAIR principles to the system. Although there are still many challenges including quality of data and metadata received, funding limitations, and program priorities, GRIIDC must always continue to improve its ability to meet user needs while implementing FAIR data principles.

KEYWORDS

data sharing, data management plan (DMP), FAIR data, multidisciplinary data repository, data citation, data discoverability

Introduction

The Deepwater Horizon (DWH) offshore drilling rig operated by BP, located 50 miles off the coast of Louisiana, experienced a blowout on 20 April 2010 resulting in an explosion that killed 11 workers, released an estimated 4.9 million barrels of oil (McNutt et al., 2011), and sank the rig. Approximately 2.1 million gallons of dispersant were released both at the surface and wellhead, the first time a dispersant was applied to the

water column (Kujawinski et al., 2011). A disaster this large mitigated with new methods required immediate research to study the potential effects of oil and dispersant on the environment. Although previous oceanographic research had been performed in the Gulf of Mexico, the information collected proved insufficient for this spill (Shepherd et al., 2016). Data for determining effects of oil on species (Bjorndal et al., 2011) and assessing the effects of the deep-water application of dispersants were lacking (Kujawinski et al., 2011).

On 24 May 2010, while the well was still releasing oil, BP committed $500 million dollars over a 10-year period "to fund an independent research program designed to study the impact of the oil spill and its associated response on the environment and public health in the Gulf of Mexico." This program, the Gulf of Mexico Research Initiative (GoMRI), would be independent of BP's control and administered by the Gulf of Mexico Alliance (GOMA). A Master Research Agreement (MRA) between GOMA and BP stated that GoMRI-funded data should be submitted to a "Research Database" and "that all data shall be fully accessible and posted thereto with minimum time delay." The research database formed was the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). GRIIDC would be based out of the Harte Research Institute for Gulf of Mexico Studies (HRI) at Texas A&M University—Corpus Christi as HRI's vision and mission to support a sustainable Gulf of Mexico aligned nicely with that of GoMRI.

Developing a data repository in parallel with initial data collection presented several challenges. Time was a critical issue as a team of software developers was building the system while other GRIIDC personnel were working with researchers to help them organize and submit their data. Another barrier was that in 2010, data sharing and data management best practices were only just being developed. Some researchers were not familiar with or resisted data sharing. Other researchers did not identify their work as data, applying a traditional model of a physical sample collected in the field and analyzed in the laboratory. Still others valued only a publication as a product with impact, not recognizing the benefits of data sharing to the researcher and the general scientific community, including higher citation rates (Piwowar et al., 2007). A final challenge was the breadth of the research being undertaken in the aftermath of the DWH disaster. This included data collection in environmental, ecological, and sociological/public health sectors.

GRIIDC did have the benefit of an advisory committee which included members of its future research board and a number of principal investigators from the GoMRI research consortia. During initial GRIIDC planning meetings in 2011, data management topics discussed included data management plans, metadata standards, digital object identifiers (DOIs), data citations, data types to accept, levels of processed data to store, and best practices. The majority of these are features of a good data management plan. It is obvious when reviewing meeting notes that GoMRI and GRIIDC had already made a clear commitment to adopting best data and metadata practices as set by funding agencies such as National Science Foundation and National Oceanic and Atmospheric Administration, including interoperability, persistent DOIs, and promoting a different, open culture for data sharing.

In 2016, FAIR data principles were published, codifying principles which are finable, accessible, interoperable, and reusable (Wilkinson et al., 2016). GRIIDC had already established several FAIR data principles, including data management planning and issuing DOIs, and continues to learn and apply those principles in software development and data curation practices. In the 11 years since the formation of GRIIDC, the data management system has evolved to mitigate submission barriers for researchers and grow with the data sharing movement as best practices advanced.

GRIIDC has developed easy-to-use and intuitive submission and search interfaces, created useful management tools, crafted curation standards, and trained researchers, resulting in the submission of more useful and well-documented data that meets funding deadlines and adheres to FAIR data principles. The following sections present the principles GRIIDC initially identified as critical: data management, data and metadata submission, citation, and distribution.

Data management planning

A data management plan (DMP) template was one of the first items prioritized as GRIIDC needed to collect information about the data to be ingested to help determine repository development needs (see Figure 1 for a timeline of events). A DMP is a document that describes what data will be collected or generated and how those data will be organized, stored, documented, and backed up throughout the entirety of the research project. GoMRI research consortia were required to complete the DMP template and submit to GRIIDC via email at the beginning of a funding cycle to plan for data submission. At the beginning of the program, researchers were not familiar with DMPs or the concept of sharing data and needed guidance to develop these documents. GRIIDC reviewed all GoMRI proposals to help determine what data were to be collected and worked with researchers to develop and understand the importance of DMPs. GRIIDC has updated the DMP template through the years, adding more fields to account for the wide variety of data types GRIIDC receives (Figure 2). More specific details are obtained for each data type such as research cruise, field work, environmental lab analysis, microcosms/mesocosms, modeling, mapping, social surveys, images, and video. Researchers can utilize these resources for any project as many funding organizations now require DMPs when submitting proposals.

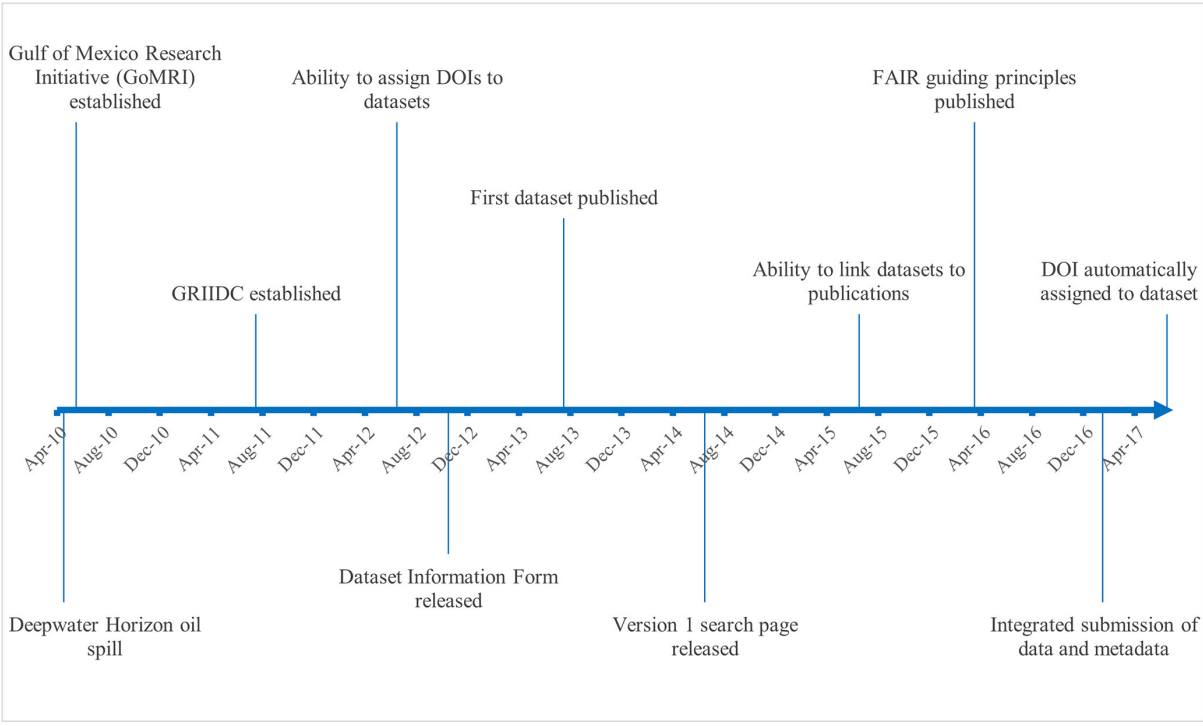


FIGURE 1
GRIIDC timeline of events.

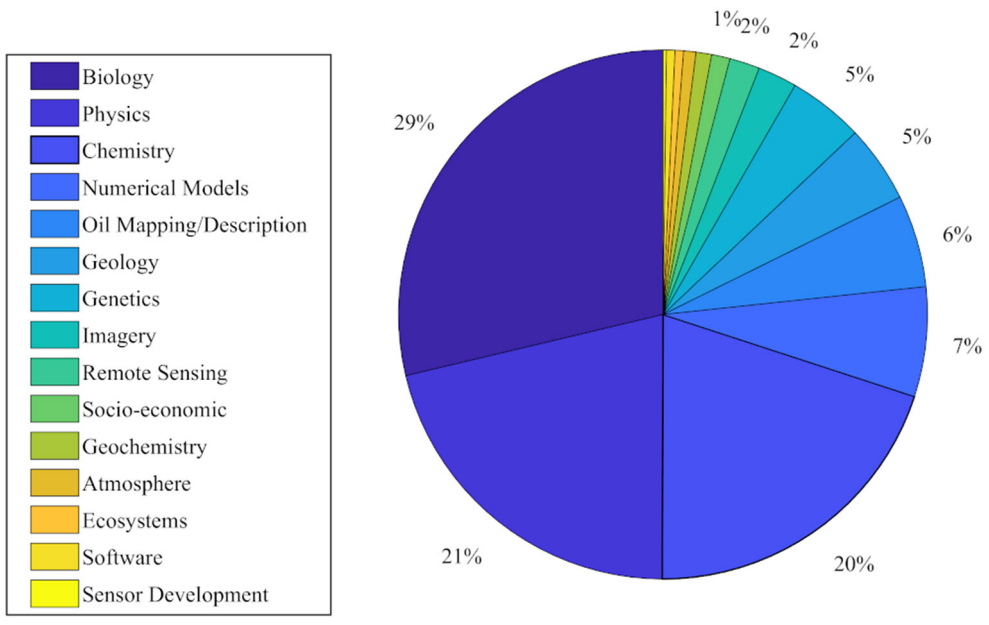


FIGURE 2
Distribution by discipline of the 3,086 GoMRI-funded datasets. Imagery and model datasets typically have a second classification indicating subject matter. Classes not labeled with a percentage comprise <1% of the total GoMRI-funded datasets.

An important advance that GRIIDC made in data management planning was the development of the Dataset Information Form (DIF), which initiates metadata collection for a dataset expected to be developed. Although a DMP for the project has important information on the project level, GRIIDC determined that more detail on specific datasets to be submitted was needed to initiate tracking (Gibeaut, 2016) and to organize dataset submissions. The DIF also helps GRIIDC prepare to ingest the data. The DIF is implemented through an online tool that GRIIDC developed, and it is integrated into the data submission workflow on the GRIIDC website. The DIF collects basic metadata such as title, abstract, data parameters and units, size of dataset, estimated data sampling period, and spatial extent. It also provides the opportunity for a researcher to indicate if the data are already located at a national data archive or if they are governed under the Institutional Review Board (IRB) or Health Insurance Portability and Accountability Act (HIPAA). When researchers are ready to submit data, the submission form is pre-filled with information provided in the DIF, thereby reducing work. GRIIDC's dataset monitoring page displays the status of a dataset through the data management workflow allowing submitters, managers, journals, and funding organizations to monitor its status. Requiring data management planning prepares a researcher for the data management lifecycle and provides a document to describe how data will be FAIR.

Submitting data and metadata

Gathering information about GoMRI-funded projects and data that were collected before GRIIDC was well established was difficult as most researchers had never prepared to share data before. GRIIDC recognized that the data submission process would need to be straightforward to accommodate researchers' various levels of technical experience, time, and patience. However, with data already being collected, a submission interface would need to be developed quickly. The first interface included a "registration" page where users could upload data and metadata. GRIIDC developed a metadata editor with which users created ISO 19115-2 metadata xml files. Users had to save the file locally and then submit the xml file to the GRIIDC system. GRIIDC encountered issues with this process as researchers would submit the data but not the metadata, causing delays in the review of the dataset or prohibiting publishing an incomplete dataset. Additionally, the submission interface could only accept a single file, requiring users to create an archive for multi-file datasets. GRIIDC would have to mitigate issues with corrupt archives and files that could not be opened.

Following user feedback and software development improvements, GRIIDC has developed an easy-to-use dataset submission form that integrates metadata and data submission into one interface (Figure 1). The form is pre-filled with information previously collected in the DIF. Users simply enter metadata such as abstract, keywords, data parameters and units, methods, spatial extent, and other descriptive information. An ISO 19115-2 compliant metadata file is automatically generated from this information and also includes other attributes such as suggested citation, data usage license, and distribution information. GRIIDC has added these fields to ensure data are findable, interoperable, and reusable. GRIIDC provides metadata in a human-readable format along with the ISO-19115-2 xml version, allowing access to users with different levels of technicality (Gries et al., 2018). Once the metadata is provided, a user can submit the data by direct upload. If data are large (over 25 gigabytes), the researcher may transfer the data via SFTP, GridFTP, Globus, or an external hard drive. If data are already located at a national data archive, a user can provide the DOI URL for the data at that location. Providing multiple methods for data submission allows researchers to choose the best option for upload given the size of their data, connection quality, location of data, and technical experience.

Due to GRIIDC's unique beginning in which researchers were studying various effects of the Deepwater Horizon oil spill, a wide range of data types were submitted to the repository including biology, chemistry, physical oceanography, sociology, political science, and public health (Figure 2). The varied documentation and metadata presented another challenge for GRIIDC. To provide more information to researchers, GRIIDC to date has created 12 guidance documents that describe recommendations for each data type. These are constantly evolving as data standards are continuously being developed and improved. For example, in 2018, to complement the required metadata and facilitate submission of data to the National Centers for Environmental Information (NCEI), GRIIDC requested researchers submitting data acquired on research vessels complete a cruise data documentation template. This template provides supplemental information, including cruise platform, dates, chief scientist, and cruise designation. This allows identification of related data housed at other data repositories such as Rolling Deck to Repository (R2R) and NCEI and assists in obtaining additional documentation such as cruise reports.

Data citation

GRIIDC determined at the beginning of the program that assigning DOIs was a vital component of the data submission process to make sure data were findable and reusable (Figure 1). The University of California's California Digital Library EZID service was initially used to create DOIs for GRIIDC datasets. GRIIDC developed a DOI request form that users would submit as a separate process from data submission. The DOI at EZID

would automatically have an "unavailable" status, meaning that the DOI would resolve to a tombstone page with the citation's metadata and reason for not being available. GRIIDC personnel would review the request; once the dataset had passed the data package review process, the DOI would be changed to "public" and would resolve to a dataset landing page. The researcher would then have to return to the registration page and enter the DOI to include it as part of the dataset. This process required multiple steps from the user and GRIIDC personnel. Additionally, it did not ensure that all datasets were assigned DOIs as it relied on the user to request one. In 2017, GRIIDC integrated DOI assignment with the dataset submission process and switched to DataCite for DOI minting services. Upon submission of a dataset, a DOI is assigned which automatically displays on the dataset landing page where the data can be downloaded, as well as a map displaying the spatial extent (if applicable), author information, a suggested citation, number of files, file size, file format, and the collected metadata. The DOI will not resolve to the landing page if the dataset has not completed the data package review process or if there is an embargo on the dataset. Automating this process has ensured that each GRIIDC dataset is assigned a DOI and eliminates additional steps for the user and GRIIDC personnel.

Displaying a DOI on a dataset landing page upon data submission facilitates the user providing the DOI to journals that require data be made publicly available. The dataset landing page contains a suggested citation, which makes it convenient for users of the data to properly cite the resource. Citation provides credit to the researcher, helps in data access and findability, and can track impact (Ball and Duke, 2015). Also found on the dataset landing page is a link to associated publications. GRIIDC has linked 1,358 publications to GRIIDC datasets. Pairing the linking of dataset to publication and referencing the dataset DOI within its associated publication maximizes the findability and impact of the data.

Distributing data

Data can be found and downloaded using GRIIDC's search page. In keeping with the rapid response nature of GRIIDC's origin, the search functionality was originally quite minimal, returning a simple listing of datasets. Improvements were made with new software releases. Users can now enter advanced search terms and narrow down to specific fields such as dataset title, abstract, author, or theme keywords. Facets can be used to further filter results by dataset status, funding organizations, and research groups. Data may be downloaded by anyone with no requirement of a GRIIDC account. Improvements to the user interface in 2021 allow a dataset to be downloaded in its entirety as a zip file or as individual files. Upon download, a SHA256 checksum hash is calculated for compressed files to confirm transfer integrity.

Reflecting GRIIDC's commitment to FAIR data principles and long-term data archival, GRIIDC data is also available from additional sources. Increased discoverability of data is provided by participation in the Data Observation Network for Earth (DataONE) where metadata of GRIIDC datasets can be found. GRIIDC also submits GoMRI-funded oceanographic data to NCEI for long-term archival. The use of standardized National Oceanographic Data Center (NODC) vocabulary terms or the National Aeronautics and Space Administration's (NASA) Global Change Master Directory (GCMD) vocabulary terms for data types and instruments enhances data discovery.

GRIIDC is currently improving an Environmental Research Division Data Access Program (ERDDAP) server, initially developed in 2015, to further serve its oceanographic data (hydrographic data, current measurements, underway sensor measurements, and drifter/float trajectories). An ERDDAP server provides additional search functionality and online map and graph creation. It also provides the ability to download data in a single format of the user's choice, adding flexibility and reducing the extraction/translation/load (ETL) burden.

Discussion

GRIIDC has a unique origin story as a data repository. Due to the urgency of its initial development and the rapidly evolving climate of data sharing, GRIIDC has faced challenges since its inception. As GRIIDC was at the forefront of the data sharing movement (Gibeaut, 2016), data standards were still being developed and researchers' knowledge of what constitutes data, data organization, and data sharing data was limited. However, involving an advisory committee during developmental stages of the program helped to address these challenges and develop data management best practices that would set the program up for success well into the future. The data sharing culture has vastly changed since the origination of the GoMRI program. Many funding agencies and journals now require that data be shared, and researchers are accepting the numerous benefits of sharing data: open data can be used to discover errors, create new questions, or be combined with other data (McNutt et al., 2016). GRIIDC has prepared researchers for success in this data sharing culture as they have been trained in data organization and management and are now familiar with submitting data and creating descriptive metadata.

GRIIDC is always striving to support FAIR data practices and contribute to the ever-growing collection of open data. GRIIDC now hosts data not only from GoMRI but also from the Florida RESTORE Act Centers of Excellence Program; the Mississippi Based Center of Excellence; the Harte Research Institute; the National Academies of Sciences, Engineering, and Medicine Gulf Research Program; as well as others.

While its inception was based on an environmental disaster, GRIIDC has come a long way, developing a data repository that strives to follow the FAIR data principles and will continue to ensure a data and information legacy for the Gulf of Mexico.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RR was the primary author of the manuscript. DL wrote sections of the manuscript, provided edits, and created figures. JG provided edits and feedback. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by a grant from the Gulf of Mexico Research Initiative.

Acknowledgments

We would like to acknowledge the Gulf of Mexico Research Initiative (GoMRI), the GoMRI Research Board Data Committee, researchers, and research consortia data managers for their assistance and feedback over the years that contributed to the development and improvement of GRIIDC. We would also like to acknowledge the Harte Research Institute for Gulf of Mexico Studies at Texas A&M University—Corpus Christi for their support.

Conflict of interest

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

Publisher's note

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

References

Ball, A., and Duke, M. (2015). *How to cite datasets and link to publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. [online]*. Available online at: https://www.dcc.ac.uk/guidance/how-guides/cite-datasets (accessed March 25, 2022).

Bjorndal, K. A., Bowen, B. W., Chaloupka, M., Crowder, L. B., Heppell, S. S., Jones, C. M., et al. (2011). Better science needed for restoration in the Gulf of Mexico. *Science*. 331, 537–538. doi: 10.1126/science.1199935

Gibeaut, J. (2016). Enabling data sharing through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). *Oceanography*. 29, 33–37. doi: 10.5670/oceanog.2016.59

Gries, C., Budden, A., Laney, C., O'Brien, M., Servilla, M., Sheldon, W., et al. (2018). Facilitating and improving environmental research data repository interoperability. *Data Science J.* 17, 22. doi: 10.5334/dsj-2018-022

Kujawinski, E. B., Kido Soule, M. C., Valentine, D. L., Boysen, A. K., Longnecker, K., and Redmond, M. C. (2011). Fate of dispersants associated with the Deepwater Horizon oil spill. *Environ. Sci. Technol.* 45, 1298–1306. doi: 10.1021/es103838p

McNutt, M., Camilli, R., Guthrie, G., Hsieh, P., Labson, V., Lehr, B., et al. (2011). "Assessment of flow rate estimates for the Deepwater Horizon / Macondo well oil spill," in *Flow Rate Technical Group report to the National Incident Command, Interagency Solutions Group*, March 10, 2011.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., and King, J. L. (2016). Liberating field science samples and data. *Science*. 351, 1024–1026. doi: 10.1126/science.aad7048

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308. doi: 10.1371/journal.pone.0000308

Shepherd, J., Benoit, D. S., Halanych, K. M., Carron, M., Shaw, R., and Wilson, C. (2016). Introduction to the special issue: an overview of the Gulf of Mexico Research Initiative. *Oceanography*. 29, 26–32. doi: 10.5670/oceanog.2016.58

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*. 3, 160018. doi: 10.1038/sdata.2016.18 Erratum in: (2019) *Sci. Data*. 6, 6.

frontiers | Frontiers in Climate

TYPE Perspective
PUBLISHED 06 October 2022
DOI 10.3389/fclim.2022.1011194

OPEN ACCESS

EDITED BY
Tiffany C. Vance,
U.S. Integrated Ocean Observing System, United States

REVIEWED BY
Karyn DeCino,
CSS Inc., United States

*CORRESPONDENCE
Avery B. Paxton
avery.paxton@noaa.gov

SPECIALTY SECTION
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

RECEIVED 03 August 2022
ACCEPTED 13 September 2022
PUBLISHED 06 October 2022

CITATION
Paxton AB, Ebert EF, Casserley TR and Taylor JC (2022) Intuitively visualizing spatial data from biogeographic assessments: A 3-dimensional case study on remotely sensing historic shipwrecks and associated marine life. *Front. Clim.* 4:1011194. doi: 10.3389/fclim.2022.1011194

Intuitively visualizing spatial data from biogeographic assessments: A 3-dimensional case study on remotely sensing historic shipwrecks and associated marine life

Avery B. Paxton[1]*, Erik F. Ebert[1,2], Tane R. Casserley[3] and J. Christopher Taylor[1]

[1]National Centers for Coastal Ocean Science, National Ocean Service, National Oceanic and Atmospheric Administration, Beaufort, NC, United States, [2]CSS-Inc., Fairfax, VA, United States, [3]Monitor National Marine Sanctuary, Office of National Marine Sanctuaries, National Ocean Service, National Oceanic and Atmospheric Administration, Newport News, VA, United States

Biogeographic assessments aim to determine spatial and temporal distributions of organisms and habitats to help inform resource management decisions. In marine systems, rapid technological advances in sensors employed for biogeographic assessments allow scientists to collect unprecedented volumes of data, yet it remains challenging to visually and intuitively convey these sometimes massive spatial or temporal data as actionable information in geographically relevant maps or virtual models. Here, we provide a case study demonstrating an approach to bridge this data visualization gap by displaying coastal ocean data in a 3D, interactive online format. Our case study documents a workflow that provides resource managers, stakeholders, and the general public with a platform for direct exploration of and interaction with 3D data from hydrographically mapping shipwrecks and marine life on the continental shelf of North Carolina, USA. We simultaneously mapped shipwrecks and their associated fish using echosounders. A multibeam echosounder collected high-resolution multibeam bathymetry of the shipwrecks and detected the broad extent of fish schools. A calibrated splitbeam echosounder detected individual fish and fish schools. After processing the echosounder data, we built an interactive, online 3D data visualization web application complemented by multimedia and story text using ESRI geographic information systems. The freely available visual environment, called "Living Shipwrecks 3D," allows direct engagement with the biogeographic assessment data in a customizable format. We anticipate that additional interactive 3D data applications can be constructed using a similar workflow allowing seamless exploration of complex spatial data used in biogeographic assessments.

KEYWORDS

biogeographic assessment, data visualization, echosounder, online spatial application, habitat mapping, shipwreck, water-column acoustics

Introduction

Biogeographic assessments aim to quantify spatial and temporal relationships between organisms and their habitats to inform spatial planning decisions (Caldow et al., 2015). Complex spatial data streams resulting from biogeographic assessments, however, are challenging to communicate and translate into accessible formats that can inform resource management decisions and foster stakeholder engagement (Caldow et al., 2015). This challenge is especially pronounced in marine ecosystems, largely stemming from rapid technological innovations that enable scientists to more quickly and efficiently collect larger and more complex data at fine resolution and over expanded spatial and temporal scales (Porter et al., 2009). For example, active acoustics surveys, such as using echosounders to map seafloor habitats and detect biological organisms, including fish and plankton, can generate more than 2GB of data per minute during acquisition, and passive acoustic monitoring of marine soundscapes and soniferous organisms can accrete data at rates exceeding multiple GB of data per minute. Optical sensors, such as 4K and low-light video used to visually characterize ecosystems, can collect over several GB of imagery per minute, and photogrammetric [e.g., structure from motion (SfM)] imagery of seabed habitats and associated sessile organisms can breach 1 GB of imagery per m^2. Advances in marine robotics have allowed vehicles, such as autonomous underwater vehicles (Morris et al., 2014), autonomous surface vehicles (Ludvigsen et al., 2018), and uncrewed aerial vehicles (Ridge and Johnston, 2020), to be outfitted with acoustic and optical sensors further expand the reach and endurance to continuously collect data over broader spatial and temporal scales, amplifying the amount of data collected in marine ecosystems that require visualization and translation for biogeographic assessments.

Myriad approaches have been developed to more effectively convey highly quantitative, large, spatial data for resource managers and stakeholders by displaying these data within geographic information systems (GIS), often manifested through data or mapping portals and decision-support tools. These applications provide platforms that can integrate ecological, social, and economic information. For example, "Marine Cadstre" (https://www.marinecadastre.gov/), a government agency-supported data portal within the USA, provides spatial data to support resource management decisions, including offshore energy planning. As part of Marine Cadastre, a tool called OceanReports (https://www.marinecadastre.gov/oceanreports/) can output spatial characterizations and high-level spatial planning analyses of coastal ocean areas to further facilitate planning decisions. Formal decision-support tools, like the "Barbuda Blue Halo" (https://www.seasketch.org/), integrate multilevel survey information (e.g., habitat classifications, biological organism occurrence), allowing direct stakeholder interaction and exploration of the data. Decision-support tools come in many different forms to facilitate different aspects of spatial planning, as in the case of "Coexist" that merges simulation models and stakeholder consultations within an online framework aimed toward sustainably integrating aquaculture and fisheries in Europe (https://www.coexistproject.eu/). While these data portals and decision-support tools provide pathways for constituents to interact directly with and explore data, the tools do not always provide data in a visually intuitive, easy to understand manner. In fact, a recent review of decision-support tools for marine spatial planning concluded that future tools could benefit from expanded avenues for stakeholder engagement with data (Pinarbaşi et al., 2017), and another synthesis concluded that dramatic improvements are required when sharing data to the public (e.g., accessible, translated, effectively communicated) to foster a more transparent, integrated, and successful resource management process (Caldow et al., 2015).

Here, we present a case study detailing a novel approach for sharing complex, spatial data from biogeographic assessments in a three-dimensional (3D), interactive online format. The goal of our case study was to characterize and visualize cultural and ecological resources within and around the USA's first federally-designated National Marine Sanctuary, Monitor National Marine Sanctuary, to assess these resources. We also developed quantitative metrics for hypothesis-driven research on the ecological function of these resources (Paxton et al., 2019), but in this paper we focus on the 3D visualization of these complex data as a path toward disseminating and translating key spatial data to support resource management decisions and stakeholder engagement. Below we share our workflow and use it to illustrate how this visualization method can be applied to other coastal ecosystems, allowing seamless exploration of complex coastal spatial data stemming from biogeographic assessments.

Visualization approach

Overview

We simultaneously mapped shipwrecks and their associated fish on 19 historical shipwrecks off North Carolina, USA (Figure 1). The shipwrecks included the Civil War ironclad vessel, USS *Monitor,* which sank in 1862 and was later designated as the USA's first national marine sanctuary in 1975, as well as shipwrecks on the outer continental shelf of North Carolina (https://monitor.noaa.gov/). These surrounding shipwrecks include three from the World War I time period, two from the mid-1920's, and thirteen from World War II. The shipwrecks rest in waters ranging from 17 m (*Ashkhabad*) to 231 m deep (SS *Bluefields*). Each shipwreck was selected based on its historical significance, and some were also selected because

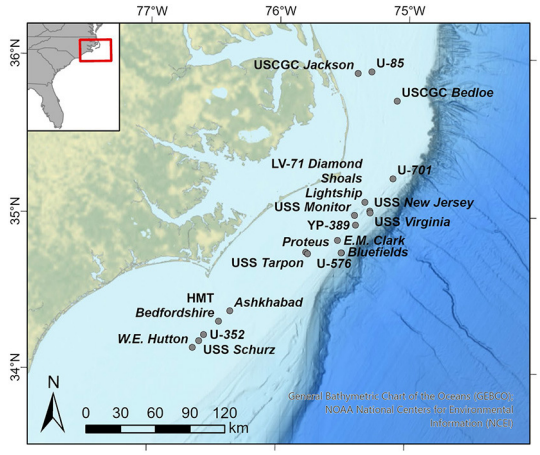


FIGURE 1
Locations of shipwrecks for which biogeographic assessments were conducted. Splitbeam echosounder and multibeam echosounder data from each shipwreck was displayed in three-dimensions within the "Living Shipwrecks 3D" visualization tool. Inset shows broader geographic context. Basemap credit: General Bathymetric Chart of the Oceans, NOAA National Centers for Environmental Information.

they had not yet been assessed and data were required to support resource management decisions.

We surveyed each shipwreck using a suite of scientific echosounders, including a multibeam echosounder and splitbeam echosounders. We first collected high-resolution multibeam bathymetry imagery of each shipwreck. Using the resulting bathymetry, we then designed additional surveys to detect fish associated with the shipwrecks. These fish surveys were conducted using splitbeam echosounders and the watercolumn data from the multibeam echosounder and were designed in a grid survey pattern, with orthogonal along-shipwreck and across-shipwreck survey lines. Survey line spacing was determined based on the size of the shipwreck from the multibeam bathymetry imagery to enable adequate spatial coverage for fish detections.

Multibeam bathymetry

The multibeam echosounders (Reson 7125 and Kongsberg EM2040) collected multibeam bathymetry of each shipwreck at fine resolution (<1 m x 1 m cell size); the exact resolution was selected to provide optimal coverage based on the depth and anticipated shipwreck size. We corrected multibeam bathymetry data for changes in the speed of sound throughout the water column, tidal influence, static draft, latency, roll, pitch, yaw, and sensor offsets during post-acquisition processing (NOAA OCS, 2021). To display these data visually within a GIS framework, we imported the bathymetry elevation of each shipwreck as ground layers into a scene rendered within ESRI ArcGIS Pro version 2.4.0 (ESRI, 2020) and imported the corresponding geotiff of the bathymetry imagery into the ArcGIS Pro scene, as well.

Splitbeam echosounder

We detected fish associated with the shipwrecks using splitbeam echosounders. The splitbeam echosounders (Kongsberg Simrad EK60 with 7° beam angle) emitted sound pulses downwards into the water column at three frequencies and corresponding pulse lengths (38 kHz—0.256 μs, 120 kHz—0.128 μs, and 200 kHz—0.128 μs). Splitbeam ping emissions were triggered by multibeam pings to reduce interference among the echosounders. The hull-mounted transducers were calibrated for backscatter response using a tungsten carbide sphere (Demer et al., 2015). Following data acquisition, we processed raw echogram data within Echoview version 10.0 (Echoview Software Pty Ltd, 2020) to identify and characterize individual fish and schools of fish around the shipwrecks. We focused on the 120 kHz echosounder because data from this frequency were most commonly used by the authors in other studies for detecting fish across the varying shipwreck depths.

To detect individual fish, we applied a target detection and fish tracking algorithm that classifies sequential acoustic targets as discrete fish. Data for tracked individual fish were exported from Echoview with their corresponding latitude, longitude, depth, and target strength. These data were read into R version 3.5 (R Core Team, 2020) using a custom written script and exported as a shapefile. The shapefile was imported into ArcGIS Pro with the "Feature to 3D by Attribute" geoprocessing tool within the "3D Analyst" toolbox and displayed using at the identified geographic location and depth using a selected 3D symbology, where colored spheres sized proportionally to the mean target strength represent individual fish.

We applied a SHAPES school detection algorithm (Barange, 1994) to detect schools of fish and calculate geometric metrics associated with the schools, such as school thickness, school length, school perimeter, and school area. Data for fish schools were exported from Echoview with their corresponding centroid latitude, centroid longitude, centroid depth, and geometries (thickness, length, area, perimeter—all corrected for beam geometry). Similar to the workflow described for individual fish, we then read the exported data into R, exported the data from R as a shapefile, imported the shapefile into ArcGIS Pro to display the schools at the appropriate geographic coordinates and depth, and set 3D symbology where spheres represent fish schools. Sphere height was proportional to the corrected fish school thickness, whereas sphere width was proportional to the corrected fish school length. The presentation of schools in this way simplifies the shape of often irregular fish schools,

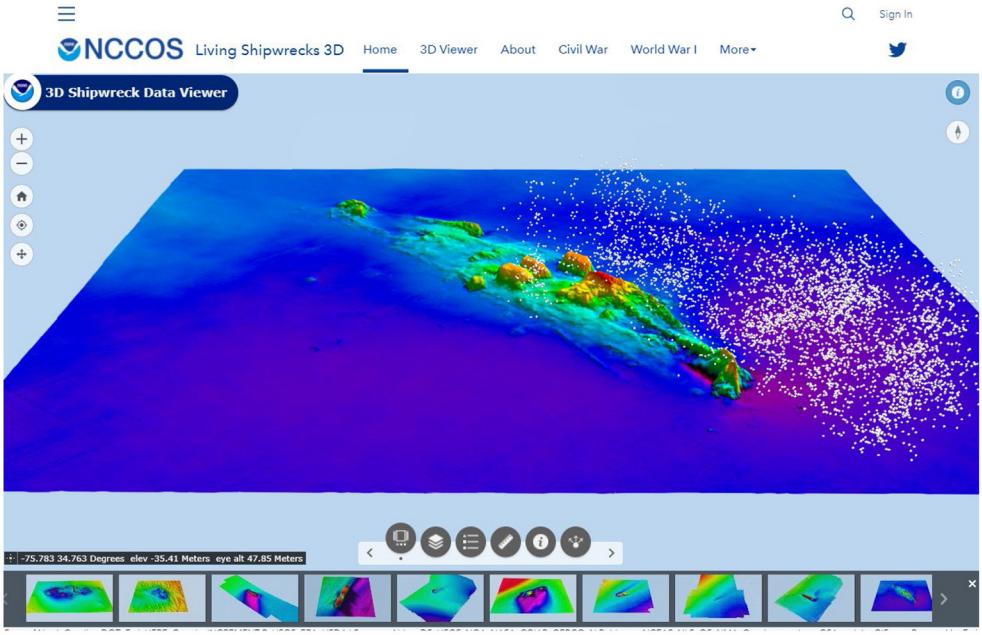


FIGURE 2
"Living Shipwrecks 3D" is an online interactive tool that displays habitat mapping data collected from acoustic surveys over shipwrecks, as well as photographs and videos. The tool combines bathymetry maps with maps of fish (white circles) detected with echosounders.

but provides a standard presentation of relative size and extent across the seascape in the 3D visualization.

Multibeam watercolumn

The multibeam echosounder used to acquire multibeam bathymetry also collected watercolumn data. We used these watercolumn data to detect the across-ship path extent of fish schools associated with the shipwrecks. In comparison to the narrow (7°) beam width of the splitbeam echosounder, the broader (~130°) beam width of the multibeam echosounder permitted fish school detection of a larger area of the watercolumn around shipwrecks. Raw multibeam data were processed within Echoview to detect fish targets comprising a fish school. Data rates for the multibeam echosounder require significant computing and graphical resources. Therefore, for each shipwreck, we selected segments of transects that contained fish schools detected from the splitbeam echosounder data and then applied a multibeam target detection algorithm to subsets of ping transmissions in the data files, yielding a "cloud" of targets constituting the fish school. These identified multibeam fish targets representing the school were exported from Echoview by multibeam ping. For each ping, fish target values, including the target range, mean, major axis angle, and minor axis angle, were provided.

The multibeam fish target data were then read into R, where we performed geometric corrections accounting for ship position and motion to compute the position of each target in geographic space (latitude, longitude, depth). These processed data with a corresponding latitude, longitude, and depth for each target in the school were exported from R as a shapefile. The shapefile was imported into ArcGIS Pro, as per the splitbeam fish data described above, and set to the appropriate 3D symbology, where standard sized spheres represented fish targets—we did not vary sphere size by attribute because the multibeam system is uncalibrated and backscatter values are affected by numerous factors not limited to fish size and angular orientation relative to the acoustic beam. We also applied a convex hull to the multibeam fish targets within ArcPro, which allowed us to quantify the volume of the school, as well as the school width, thickness, and length. Ultimately, these schooling fish targets from the wider angle multibeam fan convey the broader spatial extent of the same fish schools that were originally detected and visualized in a narrower slice of the watercolumn using the splitbeam echosounder.

Data visualization

To visualize the multibeam bathymetry, splitbeam detected individual and schooling fish, and multibeam fish school extents, we next developed an online 3D tool using ArcGIS software products (Figure 2 and Supplementary Video S1). We exported each layer from the ArcGIS Pro scene (bathymetry, ground bathymetry, imagery, splitbeam individual fish, splitbeam fish

schools, multibeam fish school extent) to the online NOAA Geoplatform using the "Share as Web Layer" tool. Once the layers were uploaded to the Geoplatform, we created a Web Scene to depict all layers in 3D using the following steps. First, we added a background ground layer to the Web Scene for ocean bathymetry that displays the broader geographic study area on a 3D ocean topography map ("TopoBathy 3D") (ESRI, 2020). Second, we added the ground elevation layers for each shipwreck that display the bathymetry-derived shipwreck elevations in 3D. Third, we added the bathymetry-derived geotiff imagery of the shipwreck, which drapes over the ground elevation layer, providing a visual representation of the shipwreck. Fourth, we added the fish detection layers (splitbeam individual fish, splitbeam fish schools, multibeam fish school extent), which displayed in 3D around the ground elevation layers and accompanying geotiffs. We then imported the Web Scene into a Web Application. By pulling the Web Scene into a Web Application, we could customize the user interface by adding menus, navigation options, and styling to facilitate constituent exploration of and interaction with the multiple data streams.

Once our data were compiled into the customized Web Application, we created an Arc Hub site. Arc Hub is an online ESRI software product that allows creation of customized web-page content using a GUI interface (ESRI, 2020). By using Arc Hub, we created a Hub Site called "Living Shipwrecks 3D" where we could combine visual media and story text with the data from the Web Scene and resulting Web Application (Figure 2 and Supplementary Video S1). The beauty of Arc Hub is that by building Hub Pages within the Hub Site, we can organize information into intuitive manners. For example, we created Hub Pages specific to shipwrecks from certain time periods to facilitate interaction with these data by stakeholders interested in history. The visual media that we added to the Hub Site included photographs and videos.

Conclusions

The "Living Shipwrecks 3D" visualization tool that we developed allows resource managers and stakeholders to directly access and engage with data in a way that best meets their needs. Resource managers can use the tool to understand the spatial extent and arrangement of shipwrecks and the spatial distributions of fish reliant upon the shipwrecks. For example, managers can measure the vertical height of shipwrecks from habitat mapping data and relate the vertical height to fish abundance and biomass. Information gained from interacting with remote sensing data can help inform resource management decisions on how to best ensure that shipwrecks remain special places within the seascape. Stakeholders, including those with an interest in ecology and history, can learn more about how shipwrecks function as habitat for marine life using the tool. For instance, recreational divers can use the tool to understand the layout of shipwrecks that they may visit for recreational dives.

We anticipate that additional interactive 3D data tools can be constructed using a similar workflow allowing seamless exploration of complex coastal spatial data used in biogeographic assessments. These tools can help overcome inherent challenges of visualizing and translating complex spatial datasets into formats that can be interpreted by diverse stakeholders and into actionable information to guide resource management decisions. Pursuits to develop similar visualization tools can help democratize data access.

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: Data are archived and publicly available as: JT, AP, and EE. 2020. Living Shipwrecks 3D: Water column data in Southeast Atlantic, 2016-10-29 to 2018-08-21 (NCEI Accession 0215765). NOAA National Centers for Environmental Information Dataset. doi: 10.25921/y18j-8h61. The Living Shipwrecks 3D website is publicly available at https://3d-shipwreck-data-viewer-noaa.hub.arcgis.com/.

Author contributions

All authors conceptualized this research and reviewed and edited the manuscript. JT and TC acquired funding. EE processed multibeam bathymetry and splitbeam data. AP processed multibeam water-column data, created the 3D visualization tool, and drafted the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank the NOAA National Centers for Coastal Ocean Science and Monitor National Marine Sanctuary for supporting this research. We thank the officers and crew of the NOAA ship Nancy Foster for supporting missions to collect data. AP was supported by CSS under NOAA/NCCOS Contract #EA113C17BA0062 during part of the study.

Conflict of interest

Author EE was employed by CSS-Inc. Author AP was employed by CSS-Inc. during part of the study.

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

Publisher's note

Author disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of the US Government, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fclim.2022.1011194/full#supplementary-material

SUPPLEMENTARY VIDEO S1
Tour of the "Living Shipwrecks 3D" online tool. The video tour of the tool showcases data associated with two shipwrecks: USS *Tarpon* and *Proteus*.

References

Barange, M. (1994). Acoustic identification, classification and structure of biological patchiness on the edge of the Agulhas Bank and its relation to frontal features. *South Afr. J. Marine Sci.* 14, 333–347. doi: 10.2989/025776194784286969

Caldow, C., Monaco, M. E., Pittman, S. J., Kendall, M. S., Goedeke, T. L., Menza, C., et al. (2015). Biogeographic assessments: a framework for information synthesis in marine spatial planning. *Marine Policy* 51, 423–432. doi: 10.1016/j.marpol.2014.07.023

Demer, D. A., Berger, L., Bernasconi, M., Bethke, E., Boswell, K., Chu, D., et al. (2015). Calibration of Acoustic Instruments. *ICES Cooperative Research Report No. 326*. p. 133. doi: 10.25607/OBP-185

Echoview Software Pty Ltd (2020). *Echoview Software*. Hobart, Australia: Echoview Software Pty Ltd.

ESRI (2020). *ArcGIS Pro*. *Redlands, California, USA*. Redlands, California, United States: Environmental Systems Research Institute. Environmental Systems Research Institute.

Ludvigsen, M., Berge, J., Geoffroy, M., Cohen, J. H., De La Torre, P. R., Nornes, S. M., et al. (2018). Use of an Autonomous Surface Vehicle reveals small-scale diel vertical migrations of zooplankton and susceptibility to light pollution under low solar irradiance. *Sci. Adv.* 4, eaap9887. doi: 10.1126/sciadv.aap9887

Morris, K. J., Bett, B. J., Durden, J. M., Huvenne, V. A. I., Milligan, R., Jones, D. O. B., et al. (2014). A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. *Limnol. Oceangraph. Method.* 12, 795–809. doi: 10.4319/lom.2014.12.795

NOAA OCS (2021). *Hydrographic Survey Specifications and Deliverables*. NOAA Office of Coast Survey, Hydrographic Surveys Division.

Paxton, A. B., Taylor, J. C., Peterson, C. H., Fegley, S. R., and Rosman, J. H. (2019). Consistent spatial patterns in multiple trophic levels occur around artificial habitats. *Marine Ecol. Prog. Series* 611, 189–202. doi: 10.3354/meps12865

Pinarbaşi, K., Galparsoro, I., Borja, Á., Stelzenmüller, V., Ehler, C. N., and, Gimpel, A. (2017). Decision support tools in marine spatial planning: present applications, gaps and future perspectives. *Marine Policy* 83, 83–91. doi: 10.1016/j.marpol.2017.05.031

Porter, J. H., Nagy, E., Kratz, T. K., Hanson, P., Collins, S. L., Arzberger, P., et al. (2009). New eyes on the world: advanced sensors for ecology. *BioScience* 59, 385–397. doi: 10.1525/bio.2009.59.5.6

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ridge, J. T., and Johnston, D. W. (2020). Unoccupied aircraft systems (UAS) for marine ecosystem restoration. *Front. Marine Sci.* 7, 438. doi: 10.3389/fmars.2020.00438

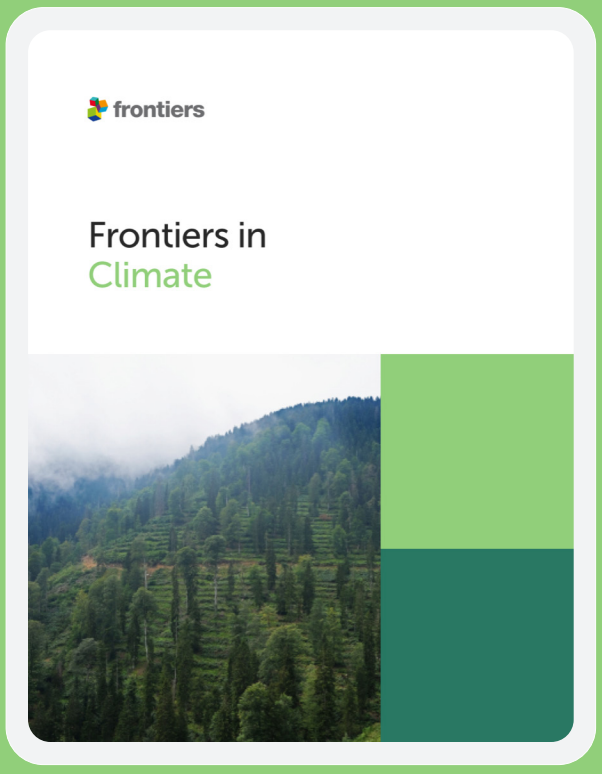
Frontiers in Climate

Explores solutions which can help humanity mitigate and adapt to climate change

Explores scientific advances in climate research, focusing on mitigation, adaptation, emissions and modelling. It shares research that contributes to climate policy and economic drivers, addressing societal challenges created by climate change.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact