**Title: Big data in Earth Science: Emerging Practice and Promise**

**Authors:** Tiffany C. Vance[1]*†, Thomas Huang[2]†, Kevin A. Butler[3]†

**Affiliations:**

[1] NOAA/US Integrated Ocean Observing System (IOOS); Silver Spring, MD 20910, USA.

[2] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA.

[3] Esri; Redlands, CA 92373, USA.

*Corresponding author. Email: tiffany.c.vance@noaa.gov

†These authors contributed equally to this work.

**One-Sentence Summary:** New opportunities to enhance our understanding of a complex Earth.

**Abstract:** Improvements in the number and resolution of Earth and satellite-based sensors coupled with finer-resolution models have resulted in an explosion in the volume of Earth science data. This data-rich environment is changing the practice of Earth science, extending it beyond discovery and applied science to new realms. This review highlights recent big data applications in three sub-disciplines; hydrology, oceanography, and atmospheric science. We illustrate how big data relates to contemporary challenges in science; replicability and reproducibility, and the transition from raw data to information products. Big data provides unprecedented opportunities to enhance our understanding of the Earth's complex patterns and interactions. The emergence of digital twins enables us to learn from the past, understand the current state, and improve the accuracy of future predictions.

**Main Text:**

## Introduction

Over the past decade, improvements in the number and resolution of Earth and satellite-based sensors and finer-resolution climate and meteorological models have resulted in an explosion in the volume of Earth science data. These data have enabled scientists to explore complex phenomena and Earth processes at unprecedented scales and resolutions and fueled the development of new tools and techniques for data analysis, visualization, and interpretation. As a result, some have argued that big data represents a fourth leg of scientific research, distinct from theory, experimentation, and computation (*1*). However, despite its transformative impacts on Earth science, a clear and concise

definition of big data is elusive (*2*). Big data are traditionally defined by five V's - Volume, Velocity, Veracity, Variety, and Value. These describe data that are large, arrive quickly, may be of mixed reliability or accuracy, are in a number of formats, and have high value. Used as early as the 1990s, the term big data described how improvements in sensors, models, data management, and computing resources made it possible to gather and analyze ever larger datasets (*3*). The term has and will continue to co-evolve with the ever-changing advancements in technology and datafication (*4*) of our natural and human systems. Big data can also be thought of as data that are **deep** - there are simply a large number of measurements, **wide** - to understand a problem data from a wide variety of sources or sensors are needed and data may be collected over large spatial or temporal domains, and **unruly** - the data, for example data from social media, are not in a consistent format. The definition of these terms will always be relative and will evolve as technology changes, especially with improvements in computing power and on board processing of data. As an example of deep, starting in the 1970s, the World Ocean Circulation Experiment (WOCE) hydrographic surveys took 24 to 36 salinity, temperature, and dissolved oxygen samples from the surface to bottom every 30 nautical miles (~55km) (http://woceatlas.tamu.edu/printed/SOA_WOCE.html). ARGO floats now collect these variables and additional measurements using unmanned floats and collect 400 or more profiles a day from 3800 floats across the entire ocean (https://argo.ucsd.edu/about/). Wide data include projects such as the regional cabled array on the Juan de Fuca Ridge ( https://interactiveoceans.washington.edu/about/regional-cabled-array/} which has deployed an array of 150 sensors across much of a 500 km by 150 km tectonic plate. Data can be transmitted at up to 240gb/s and two way communication allows real time control of instruments. Unruly data include a variety of crowd sourced data, such as weather conditions described in social media posts as opposed to data provided by trained weather spotters.

For this review, we define Big Earth Data as massive (relative to commonly available datasets) amounts of diverse, complex, and continuously accumulating data generated from heterogeneous sources that require advanced (relative to commonly available computing resources) and potentially novel computational and analytical tools to extract meaningful insights and knowledge about the Earth system.

Earth science is a vast and complex field that studies the Earth's physical, chemical, and biological systems and their interactions. Using big data to make advances in the earth sciences is critical. At a basic level, the earth is large and processes of interest can cover large areas. Systems are dynamic and it may require a long time series to detect patterns of interest. Understanding the earth as an integrated system leads to the need to look at all sorts of phenomena - and hence the need for wide data. For example, understanding the effects of sea level rise requires understanding a multivariate system of both oceanic and terrestrial variables.

Due to this immense scope, it is impossible to review the impacts of big data on the entire discipline in a single article or study. This review will focus on three sub-disciplines: hydrography, oceanography, and atmospheric science. We illustrate how big data relates to contemporary challenges in science; replicability and reproducibility, the transition from raw data to information products, and the emergence of Digital Twins.

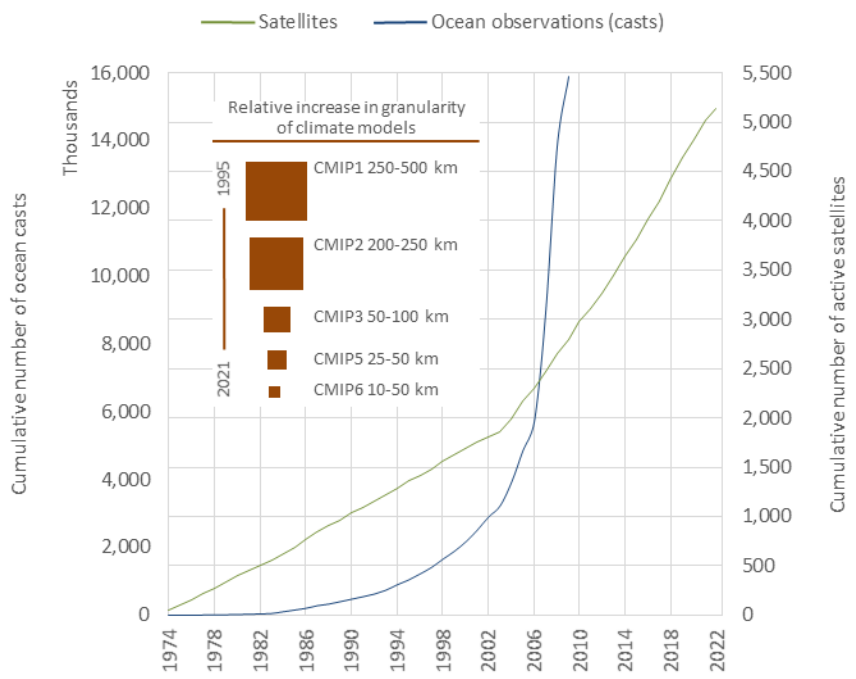


**Fig. 1. Increases in the spatial granularity of climate models and the number of Earth and space-based sensors**. The right axis shows the increase (green) in the number of active satellites beginning with the launch of Oscar 7 in 1974. The left axis shows the rapid increase (blue) in the number of ocean observations (casts) archived in the NOAA/IODE World Ocean Database (WOD). WOD data before 1974 are not shown. The inset depicts the rapid increase in the granularity of a representative climate model, Coupled Model Intercomparison Project (CMIP).

## How did we get here, who actually uses big data, and how?

The development of new sensors able to gather more data for longer time periods and the ability to measure totally new variables has contributed to ever larger earth science datasets (Figure 1). Drifting buoys and floats are widely deployed and better communications allow buoys to send data to shore in near real time (*5*). Autonomous vehicles go on extended missions gathering data and tagged animals send back streams of environmental and behavioral data. Passive acoustic data show how animals use sound and the impacts of anthropogenic sounds (*6*). Models cover larger areas at finer spatial scales and can be run for longer durations due to improvements in computing such as device scaling, high-performance and cloud computing generally (*7, 8*). The greater acceptance of citizen and crowdsourced data has increased the amount of unruly data that can contribute to research on climate change and other phenomena (9). Crowd sourced data such as personal weather stations have expanded to include bathymetric data, data

from watches and body worn sensors, social media posts about weather and oceanic phenomena, and sensors on trucks reporting weather and road conditions. New analytical methods such as artificial intelligence, neural networks, and machine learning make use of these big datasets to study everything from predicting seafloor carbon (10) to seismology (*11*).

The need for big data was understood as early as Schnase (*12*) who argued that the big questions and projects in climate science, such as the Intergovernmental Panel on Climate Change (IPCC), are tackling problems that require big data for solutions. More recently, big data are being used to support research towards the UN Sustainable Development Goals (SDGs) such as climate action (SDG 13) and life below water (SDG 14) (*13*). Authors have provided a comprehensive review of the applications of big data in geophysics (*14*), biology (*15*), and for the use of big data and AI (*16*). The European Marine Board's *Future Science Brief on Big Data in Marine Science* (https://tinyurl.com/2wxv5kvw) identifies continued sensor development, infrastructure for data collection, processing, and archiving, near real time data transmission and long term funding as core recommendations.

Determining who is actually using big data can be a challenge. Ideally, papers using big data would formally cite data DOIs - both to enable tracking of data usage and as a way to associate datasets, and researchers responsible for curating and publishing them, with their use and citation in traditional peer-reviewed publications (17) . Another way to extract trends is to look at the use of data from large repositories and project archives. The creation of world data centers amid the International Geophysical Year gave a real push to creating big datasets (*18*). The TOGA/TAO/TRITON buoy array maintains a list of over 1000 publications citing their data since the array's inception in 1986 [https://www.pmel.noaa.gov/gtmba/tao-journal-publications] . NCAR's Research Data Archive maintains a similar listing for their datasets at https://rda.ucar.edu/resources/metrics/ which also includes over 1000 citations for their datasets. NOAA's Open Data Dissemination Program has assembled a list of papers citing their data which shows 56 publications using their 14 most popular datasets. [update with public link]. The Ocean Observatories Initiative lists 333 papers using data from their cabled arrays https://ooipublications.whoi.edu/biblio. While the NASA archives do not list citations, the Physical Oceanography Distributed Active Archive (po.daac) does provide data on the volume of downloads. In 2022, 11TB/day were downloaded and the archive grew by 2TB a day https://www.earthdata.nasa.gov/eosdis/system-performance-and-metrics/eosdis-annual-metrics-reports .

Tools for discovering, analyzing, and visualizing big data are rapidly developing, with both commercial and open-source solutions available. Platforms like Google's GEE (Google Earth Engine), a cloud-based geospatial analysis platform, Microsoft's Planetary Computer, a dedicated computing platform for Earth science data analysis, and Esri's Living Atlas of the World, a comprehensive collection of geospatial datasets and applications, reduce the challenges faced by Earth scientists transitioning to data science roles (*19, 20*). Open-source solutions like the Apache Science Data Analytic Platform (SDAP) offer an alternative approach. SDAP (21) is an open-source solution that harmonizes access and analysis of Earth science data (satellite, model, and in-situ),

supporting various domains including sea level rise, gravity, marine science, wildfire, air quality, greenhouse gas, flood, land cover, surface topography and vegetation, and ocean science. The Pangeo project (https://pangeo.io/) is "a community platform for Big Data geoscience" that includes software, documentation, infrastructure and a community of practice. It supports open science and focuses on open source tools including xarray for working with multidimensional data and Iris for analyzing and visualizing oceanographic and meteorological data. These platforms provide access to extensive geospatial data and computational capabilities. However, it is important to critically review these platforms, considering limitations such as inadequate data curation, guaranteed data availability, data quality and vendor lock-in.

**Surface hydrology**

Surface waters in the form of rivers, lakes, reservoirs, streams, and wetlands are critical to human survival yet are increasingly scarce due to water quantity and water quality issues (*22*). Surface water presents unique data challenges due to the amount of water on Earth and its ephemeral nature. Flowing surface waters can range in width from a few meters to many kilometers, and their extent can change across time scales ranging from minutes to decades. Big data have enabled more accurate and complete delineations of surface water, a more detailed understanding of surface water dynamics, and informed models that more accurately quantify terrestrial water budgets. A necessary first step in understanding surface waters is delineating their spatial extent and generating a digital hydrographic map (*23*). MERIT Hydro (http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_Hydro/) and HydroSHEDs (https://www.hydrosheds.org/) are commonly used digital hydrographic products derived from 90 and 30-meter digital elevation models, respectively. Both products suffer from unavoidable data gaps and inconsistencies in the underlying data sources.

The HydroSHED product is being re-engineered based on improved DEMs generated from data from the TanDEM-X mission. The TanDEM-X dataset, which has a resolution of 12 meters at the equator, serves as the foundation for HydroSHEDS v2. This enhanced Global DEM incorporates advanced pre-processing techniques to preserve the high-resolution details of the DEM. These techniques include "an infill of invalid and unreliable elevation values, an automatic coastline delineation refined with manual corrections, an AI-based water detection algorithm, and a modification of elevation data in urban and vegetated areas for improved evaluation of the flow of water." Additionally, the hydrologically pre-conditioned DEM and the water body mask derived from the TanDEM-X dataset undergo further processing with "refined hydrological optimization and correction algorithms to derive flow direction and flow accumulation maps" (24). The promise is a globally consistent, high-resolution digital hydrographic map for the globe.

Surface waters are dynamic and generate costly and deadly floods. Combining a spatially granular (10-meter resolution) map of the Height Above the Nearest Discharge (HAND) for the conterminous United States and streamflow predictions for approximately 2.7 million river reaches has produced rapid continental scale flood prediction maps. Surface

waters are also essential to Earth's natural hydrological system and influence the available water supply for domestic, agricultural, and industrial use. Integrating multiple big datasets such as soil maps, precipitation, surface and groundwater extent, croplands, population density, and GDP. National Geographic Explorer Marc Bierkens has modeled and mapped the difference between the supply and demand of water globally (https://worldwatermap.nationalgeographic.org/).

**A Three-Dimensional (3D) ocean**

Understanding the ocean as a three-dimensional system has long been a goal of researchers. While satellites can provide a synoptic view, they don't see into the ocean depths. Models can represent the full 3D ocean but they rely on observations both as inputs and to verify results. Big data can be the result of long term campaigns, such the California Cooperative Oceanic Fisheries Investigations (CalCOFI) which has been observing the California Current ecosystem since 1949 [https://calcofi.org/]. The desire to better understand El Nino and its effects on climate led to the deployment of the TOGA/TAO arrays of buoys in the tropical Pacific starting in 1985 (*25*). This is an early example of gathering detailed and extensive observations resulting in big data. While the data are carefully formatted and quality controlled, they are wide in time and space and deep in that they measure a large number of variables. TAO data have been used to study many phenomena in the tropical Pacific, for example to better understand El Nino and its effects, look at thermal structures in the ocean, look at variations in wind and rain in the tropical Pacific, and conduct basin wide studies of sea surface temperature.
More recently, new technologies such as ARGO floats, surface and underwater autonomous vehicles and the Ocean Observatories Initiative (OOI) gather and transmit large amounts of data, often in real time, to provide a three dimensional view of previously data poor regions (Figure 2). These data have supported discoveries in everything from cross-shelf transport of water (26) to sensing whale calls (27) and ship traffic (28) to studying heat flux from hydrothermal vents (19), and seismology (11).

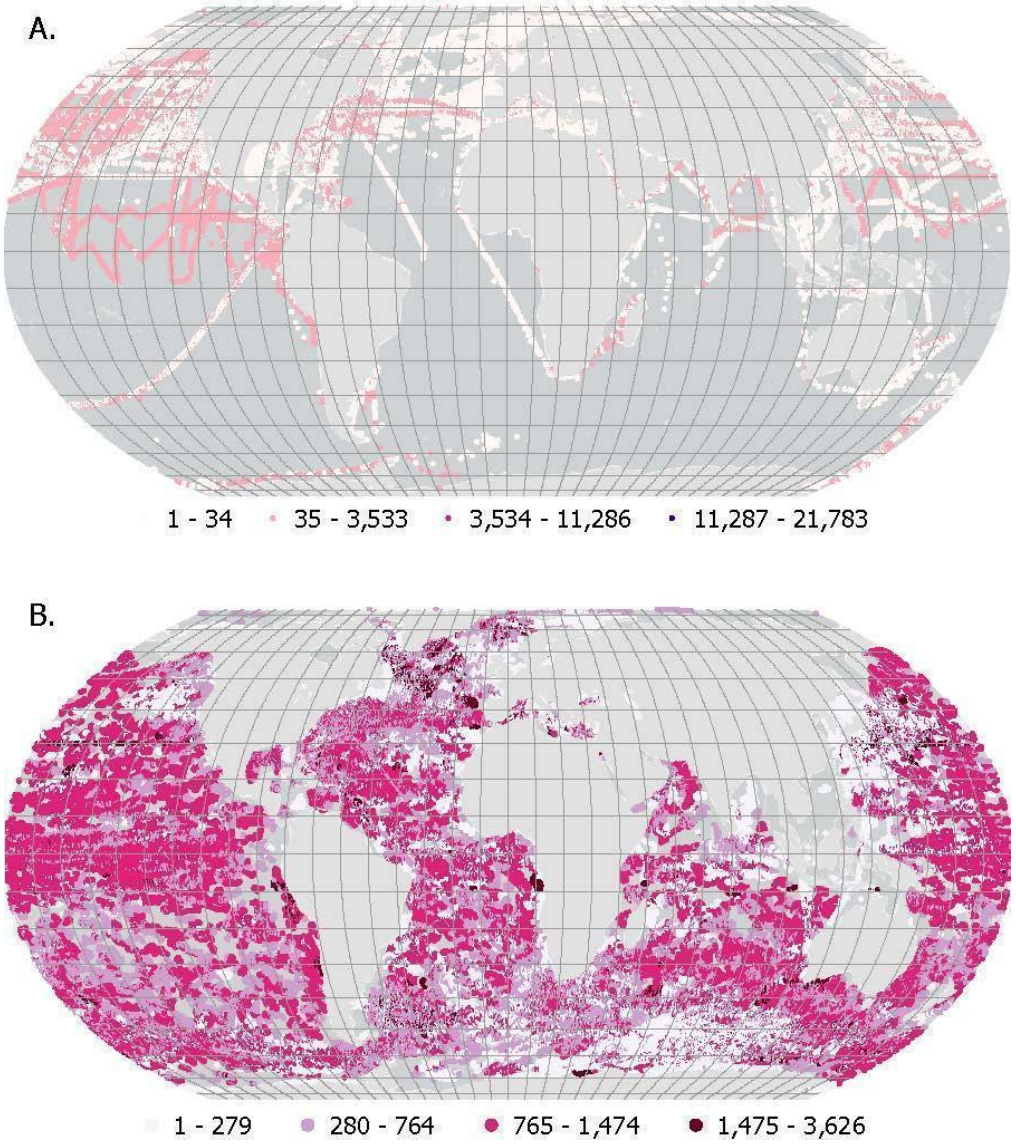


**Figure 2 – Increase in the volume of ocean temperature measurements.** Panel A shows the number of temperature observations at all depths collected in 1995. An emphasis on near coast data collection and oversampling along ship routes is apparent. Panel B shows the number of temperature observations collected in 2022 and demonstrates the rapid expansion of ocean observing systems such as the Argo float network. Data source: NOAA/IODE World Ocean Database (WOD).

Models of the ocean (29)have become more detailed, can cover larger areas and increasingly include assimilation of observed data (30) . Models are both the product of big data supported boundary conditions including detailed bathymetric grids and forcing data from river flows and wind patterns, and the producer of big data outputs. Improvements in models have supported more detailed mapping of ocean circulation patterns, better understanding of the transport of nutrients, better measurements of the heat inputs to hurricanes and improved modeling of oil spills, and the dispersion of larval fish. These models have contributed to improved forecasting of hurricanes, the siting of marine

protected areas, oil spill response and improved understanding of marine ecosystems. As model outputs become ever larger, there are questions about exactly what outputs should be archived and what can better be recreated for specific needs (31).

In biology, researchers have been able to better understand the growth and extent of harmful algal blooms (HABs) via large datasets of images of plankton. In a modern version of the Continuous Plankton Recorder (32), the Imaging Flow Cytobot (IFCB) collects large numbers of images of plankton. The images are classified using neural network techniques (33) and have led to better understanding of patterns of plankton distributions and the health of plankton populations. Environmental DNA or eDNA measurements can rapidly gather large amounts of data on the species present in a location, the presence of undetected species, genetic relationships between populations and the general biodiversity of an area or region (34).

**Atmospheric modeling: global climate models**

Global climate models (GCM) are complex mathematical representations of Earth's climate system that incorporate the physical, chemical, and biological processes governing the atmosphere, oceans, land surfaces, and ice (35). Since their inception in the early 1960s, global climate models have evolved significantly in terms of complexity, resolution, and the inclusion of additional components such as land surfaces, sea ice, and vegetation. The Coupled Model Intercomparison Project (CMIP) generates a widely used set of coordinated climate model experiments conducted under the auspices of the World Climate Research Programme (WCRP) (36). The project's latest iteration, CMIP6 (Coupled Model Intercomparison Project Phase 6), represents a global collaborative effort to enhance our understanding of Earth's climate system. The improved resolution and complexity of modeling the dynamics of Earth's climate have led to several advancements in our understanding of our climate.

Big data-based analyses provide new insights into climate processes in ocean-climate interactions, severe weather, and wind potential. Marine biogeochemistry plays a crucial role in climate by regulating the exchange of CO2 with the atmosphere. Improved CMIP6 models more accurately model oxygen levels at ocean depths of 150 m, explicitly showing decreased errors in the Southern Ocean compared to CMIP5. This provides an improved representation of deep ocean ventilation and a more precise characterization of biogeochemical properties in outcropping water masses (37). The enhanced horizontal resolution of historical runs of the CMIP6 models more accurately reproduced midlatitude storm tracks and reduced the equatorward bias present in previous models (*38*). Exploring the potential impacts of future climates on European wind resources, Carvalho et al. (*39*) showed a substantial decline in wind resources for most of Europe by the end of the century under a high emissions/population growth/energy demand/heavily fossil fuels reliant scenario.

Global climate models enable quantification and prediction of the consequences of a warming planet on migration patterns and socio-economic stability. In a review of the impacts of climate change on migration, Kaczan & Orgill-Meyer, (40) explore the impacts of climate shocks such as drought, floods, and extreme temperatures on migration. They found that climate-induced migration is (1) not limited to poorer households, (2) long-distance domestic moves are more prevalent than local or international moves, (3) slow-onset changes like droughts drive increased migration more than rapid-onset changes like floods, and (4) the severity of climate shocks impacts migration in a non-linear manner, influenced by the dominance of either capability or vulnerability channels. The growing frequency and intensity of severe weather events can have significant economic impacts on agricultural activities and threaten local and global food security (*33*). Using climate data from the ERA5 (https://www.ecmwf.int/en/research/climate-reanalysis) atmospheric re-analysis model (2000 through August 2018) and 82,000 crop yield reports, the authors could explain 65% of historical yield anomalies using machine learning (41).

While climate change is often viewed as a gradual and long-term process, extensive collections of observed and modeled climate data point to potential abrupt changes or tipping points in climate that may be irreversible. The IPCC defines tipping points as "critical thresholds in a system that, when exceeded, can lead to a significant change in the state of the system, often with an understanding that the change is irreversible" (42, p. 262). In addition to the loss of the Amazonian rainforest and melting of the Greenland ice sheet, Lenton et al., (43) point to potential changes in the frequency of fires in boreal forests, reductions in Arctic sea ice, large-scale die-offs of coral reefs, and that these individual consequences of climate tipping points could combine via feedback loops into a global cascade of events.

**Replicability and reproducibility**

To increase credibility, improve generalizability, and facilitate the reuse of scientific results, the Earth science community has renewed efforts to make geospatial research more replicable and reproducible (*36*). Reproducibility is achieved when the original data and computational methods produce the same scientific results, whereas replicability occurs when the same scientific conclusions are reached using new data (44). Big Earth Data presents unique challenges to promoting replicability and reproducibility. Researchers proposed a five-star guide (46) to enhance replicability and reproducibility in geospatial research. Successive levels of the five-star guide require increasing metadata and data-sharing granularity. The sheer volume of Big Earth Data can make converting and storing research data to open scientific data formats time-consuming and financially challenging (47). In recognition of these challenges, governmental data providers are partnering with commercial infrastructure-as-a-service (IaaS) providers to document and host massive collections of satellite, earth-based, maritime-based, and modeled analysis-ready data (https://www.noaa.gov/information-technology/open-data-dissemination, https://www.earthdata.nasa.gov/eosdis/cloud-evolution) These massive data collections

proximate to massive compute resources hold great promise for enhancing the sharing, reproducibility, replicability and collaborative nature of research. These concepts are foundational to operationalizing the larger goal of Open Science – “the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity” (https://www.nature.com/articles/d41586-023-00019-y)

**Earth information products**

Earth science is well rooted in the foundational scientific method and, albeit with potential augmentation of the process through data science (48), will remain so for the foreseeable future. This process's primary form of scientific output has historically been peer-reviewed publications. However, big data can provide additional opportunities for scholarly communication, the "system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use" (49). Data collection, generation, organization, and management, particularly of big data, are increasingly recognized as scientific output. Peer-reviewed data papers or publications that review and document the collection methods, data cleaning, and quality assurance are now common (50). While data availability is essential for replicability and reproducibility, big data's volume, variety, and veracity can make its use intractable. As a result, a pattern of producing Earth Information Products is emerging. These are synthesized, structured, and organized presentations of scientific data, findings, and research outcomes, in a format that is accessible, informative, and useful inside and outside of the scientific community. Earth information products seek to "turn data into information, and information into insight" (attributed to C. Fiorina) with the ultimate goal of making science actionable (51). Earth information products can be data-centric, focusing primarily on providing the analytical results of big data workflows and have been used to share ecological modeling results for the land (52), ocean (53), and coasts (54). Alternatively, earth information products can be more narrative in form and integrate interactive maps with text and multimedia content to facilitate scientific communication by summarizing the research context, workflows, and primary results[1].

**Digital Twins**

A rapidly increasing area of research and discovery based upon big data is Digital Twins (DT). Digital Twins of the Earth System are intended to establish highly accurate digital representations of the Earth system so that they can improve our understanding of the impacts of climate change and extreme weather events, and potentially help us better assess potential socio-economic and health impacts. The DT concept has proven to be effective in various commercial sectors (55, 56). Figure 3 provides the high-level

[1] See UNISECO [https://uniseco-project.eu/assets/content/resources/02-deliverables/UNISECO-D3.3.pdf] for an extensive list of examples in the context of social-ecological systems.

representation of a DT for the Earth System as an integrated information system where each oval contributes to the overall DT goals. The NASA Advanced Information Systems Technology (AIST) program's website (https://esto.nasa.gov/aist) summarizes the primary goals of DT of the Earth.

- To provide a continuous and accurate representation of the systems as they change over time.
- To mirror the Earth science system through advanced analysis, Artificial Intelligence, and state-of-the-art models to help predict the Earth's response to various phenomena.
- To provide tools for scenario-based predictions to recommend possible actions the researchers and/or decision-maker might take, which could involve possible mitigation options as well as acquiring additional data and analysis.

DT presents multidisciplinary, multivariate data challenges, that is, the five Vs of big data. The expectation for DT of the Earth is to mirror the Earth Science System to not only understand the current condition of our environment or climate, but also to automatically analyze changes in our environment and autonomously acquire new data in order to improve its prediction and forecast (57). For a DT of the Earth System to be useful, it must accurately represent the interactions or forcing between the subsystems. The accuracy of a DT is highly reliant on the quality of the data and the analysis it incorporates. Artificial intelligence (AI) plays a vital role in the overall DT architecture. Our rapidly growing collections of observations and model data require a DT to be smarter about when and what data and analysis to include. For a DT solution to be sustainable, its architecture needs to be able to address multiple big data challenges.

- Multivariate analysis. Almost all Earth science research requires working with multiple measurements in order to understand their interactions and causes and effects. It could involve various remote sensing and in-situ measurements. Depending on the type of analysis (i.e., global and regional), it would also require different resolutions of data. Working with large collections of data requires new methods in managing data to promote parallel computing of the data.
- Assimilation and numerical models. The GCM section presented the complexity of developing and running large-scale numerical model simulations. DT requires ongoing updates of the model runs with the latest state of Earth system to generate the most accurate forecast.
- Advanced AI. DT requires capabilities to identify the relevant data, analysis, and model results to incorporate according to the user-scenario and the science the scenario requires. DT requires advanced orchestration support from both cost and operation optimization points of view.

There are various funded DT for Earth development efforts. AIST's Integrated Digital Earth Analysis System (IDEAS) [https://ideas-digitaltwin.jpl.nasa.gov/] led by NASA JPL (58)is designing a reusable Earth System Digital Twins (ESDT) framework using flood [https://ideas-digitaltwin.jpl.nasa.gov/hydrology/] and air quality [https://ideas-digitaltwin.jpl.nasa.gov/airquality/] as its use cases (59). The Space for

Climate Observatory (SCO)'s FloodDAM [https://www.spaceclimateobservatory.org/flooddam-garonne] led by CNES is establishing a solution to help predict and analyze flood events in France (60) . NASA's IDEAS-based hydrology ESDT and SCO's FloodDAM are working jointly to establish a federated DT solution for both US and the French regions. The European Space Agency's (ESA) Destination Earth (DestinE) [https://digital-strategy.ec.europa.eu/en/policies/destination-earth] (61) is being developed by three entities, the European Center for Medium-Range Weather Forecasts (ECMWF), the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), and ESA. The EU's Iliad Digital Twins of the Ocean [https://www.ocean-twin.eu/digital-twins] is a consortium chartered to develop a portfolio of ocean-released DT.

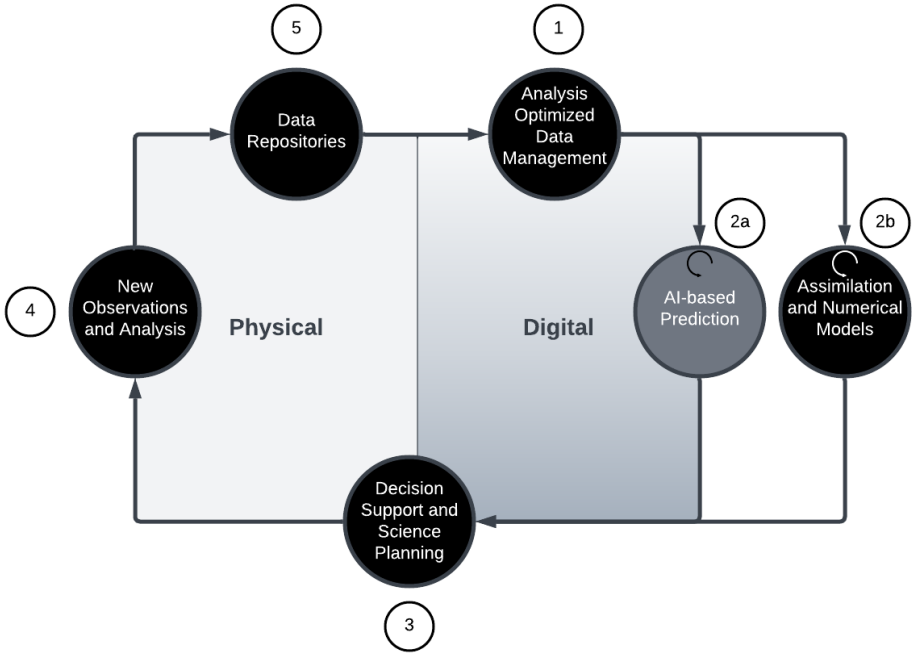


**Figure 3:** High-level representation of the elements of a Digital Twin for the Earth system. The cyclic framework signifies the continuous flow of information within the digital twins that bridge the physical and the digital representations. The arrows illustrate the general information flow between the elements. 1) The harmonized, analysis-optimized data management solution for fast access and analysis. 2a) AI plays an essential role in digital twins, including identifying the relevant data, analysis, and numerical models, as well as resource management. AI formalizes the process to learn from the past to improve the accuracy of future predictions. AI-based models require ongoing training and continuous validation. 2b) Advanced, physics-based models are essential to forecast the environment's reactions. Like AI-based models, numerical models require continuous assimilation and validation. 3) One of the promises of digital twins of the Earth is to deliver actionable predictions. Actions could include mitigation recommendations, new observations, and analysis to improve our understanding. 4) New observations could include re-tasking Earth-observing instruments, deploying unmanned vehicles, acquiring data from in-situ sensors, and on-demand value-added product generations, etc. 5) The acquired or processed data is made available to the digital twin to be incorporated for improving analysis and predictions.

## Conclusions

This review focused on three sub-disciplines of Earth science: hydrography, oceanography, and climate science, to illustrate the profound and ongoing impacts of big data on the broader discipline. Big data have facilitated advancements in understanding surface water, such as more accurate delineations of riparian networks, improved flood predictions, and comprehensive water supply and demand assessments. In oceanography, big data has enabled researchers to achieve a deeper understanding of the three-dimensional nature of the ocean, leading to discoveries in areas ranging from cross-shelf water transport to whale behavior. Global climate models' improved spatial and temporal granularity has resulted in insights into climate processes and interactions. The new insights enabled through big data come at a cost. The size of big data presents challenges to scientific replicability and reproducibility and data sharing and management. Ensuring these datasets meet the requirements of the FAIR/CARE[2] principles and bridge the gap between data and information is a continuing effort. Traditional incentives for researchers to publish papers need to be supplemented by investing time in properly curating datasets. These challenges are outweighed by the promise big data brings for a deeper, more comprehensive understanding of our Earth through environmental information products and digital twins. As sensors and modeling continue to advance, coupled with parallel progress in computing resources, the concept of big data will constantly evolve. Nevertheless, the potential of big data to enrich our comprehension of Earth as a complex, interrelated, and dynamic system while supporting Open Science remains steadfastly promising.

## References and Notes:


1. M.Y. Vardi, Science Has Only Two Legs. *Communications of the ACM* **58**, 5-5 (2015).
2. M. Favaretto, E. De Clercq, C.O. Schneble, B.S. Elger, What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE* **15**, e0228987 (2020). https://doi.org/10.1371/journal.pone.0228987
3. F. X. Diebold, On the origin(s) and development of the term 'big data.'. *Business Economics* **47**, 85-91 (2012).
4. K. Cukier, V. Mayer-Schoenberger, The rise of big data: How it's changing the way we think about the world. *Foreign Aff* **92**, 28 (2013).
5. D. Roemmich et al., The Argo Program: Observing the Global Ocean with Profiling Floats. *Oceanography* ***22***, 34–43. (2009). http://www.jstor.org/stable/24860957


[2] Findable, Accessible, Interoperable and Reusable (FAIR) and Collective Benefit, Authority to Control, Responsibility and Ethics (CARE)

6. C.C. Wall et al., The Next Wave of Passive Acoustic Data Management: How Centralized Access Can Enhance Science. *Frontiers in Marine Science* **8** https://doi.org/10.3389/fmars.2021.703682 (2021).
7. T.C. Vance et al., From the Oceans to the Cloud: Opportunities and Challenges for Data, Models, Computation and Workflows. *Frontiers in Marine Science* **6** https://doi.org/10.3389/fmars.2019.00211 (2019).
8. J. Ang and D. Mountain, New Horizons for High-Performance Computing. *Computer*, **55**, 156–162 (2022). doi: 10.1109/MC.2022.3200859
9. A, Karagiannopoulou, A.Tsertou, G. Tsimiklis, A. Amditis, Data Fusion in Earth Observation and the Role of Citizen as a Sensor: A Scoping Review of Applications, Methods and Future Trends. *Remote Sens.* **14** https://doi.org/10.3390/rs14051263 (2022).
10. T.R. Lee, W.T. Wood, B.J. Phrampus, A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Global Biogeochemical Cycles*, **33**, 37– 46 (2019). https://doi.org/10.1029/2018GB005992
11. S.J. Arrowsmith et al., Big Data Seismology. *Reviews of Geophysics* **60**, e2021RG000769 (2022). https://doi. org/10.1029/2021RG000769
12. J.L. Schnase et al., Big Data Challenges in Climate Science. *IEEE Geosci Remote Sens Mag.* **4**, 10-22 (2016). doi: 10.1109/MGRS.2015.2514192.
13. H. Hassani, X. Huang, S. MacFeely, S. M.R. Entezarian, Big Data and the United Nations Sustainable Development Goals (UN SDGs) at a Glance. *Big Data Cogn. Comput.* https://doi.org/ 10.3390/bdcc5030028 (2021).
14. A.D. Gvishiani, M.N. Dobrovolsky, B.V. Dzeranov, Big Data in Geophysics and Other Earth Sciences. *Izv., Phys. Solid Earth* **58**, 1–29 (2022).
15. E.Aronova, K.S. Baker, N. Oreskes, Big science and big data in biology: from the international geophysical year through the international biological program to the long term ecological research (LTER) Network, 1957—Present. *Hist Stud Nat Sci* **40**, 183-224 (2010).
16. E. Verdu, Y.V. Nieto, N. Saleem, Big data and artificial intelligence in earth science: recent progress and future advancements. *Acta Geophysica* **71,** 1373-1375. (2023).
17. Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11. (2014) https://doi.org/10.25490/a97f-egyk
18. F.L. Korsmo, The Origins and Principles of the World Data Center System. *Data Science Journal* 8. IGY55-IGY65. 10.2481/dsj.SS_IGY-011 (2010).
19. L.M. Bouwer, D. Dransch, R. Ruhnke, D. Rechid, S. Frickenhaus, J. Greinert, Eds, *Integrating Data Science and Earth Science. Springer Briefs in Earth System Sciences*. (Springer, 2022). https://doi.org/10.1007/978-3-030-99546-1_1
20. T. Huang, T.C. Vance, C. Lynnes, Eds., *Big Data Analytics in Earth, Atmospheric, and Ocean Sciences* (American Geophysical Union 2022).
21. T. Huang et al., "Open Source Exploratory Analysis of Big Earth Data With NEXUS," in *Big Data Analytics in Earth, Atmospheric, and Ocean Sciences* (American Geophysical Union, 2022), pp.115-136.
22. M.T.H. van Vliet et al., Global water scarcity including surface water quality and expansions of clean water technologies. *Environ. Res. Lett.* 16 024020 (2012)

23. B. Lehner, M.L. Messager, M.C. Korver, S. Linke, Global hydro-environmental lake characteristics at high spatial resolution. *Sci Data* ***9***, https://doi.org/10.1038/s41597-022-01425-z (2022).
24. L. Warmedinger et al., Improved hydrologic conditioning of the TanDEM-X dataset for HydroSHEDS v2, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-15182, https://doi.org/10.5194/egusphere-egu23-15182, 2023.
25. S. P. Hayes, L.J. Mangum, J. Picaut, A. Sumi, K. Takeuchi, TOGA TAO: A moored array for real-time measurements in the tropical Pacific Ocean. *Bulletin of the American Meteorological Society* **72**, 339–347 (1991).
26. W. Zhang et al., Cross-shelf exchange associated with a shelf-water streamer at the Mid-Atlantic Bight shelf edge, Progress in Oceanography **210**, p. 102931, (2023).
27. W. Wilcock, S. Abadi, B.Lipovsky, Distributed acoustic sensing recordings of low-frequency whale calls and ship noise offshore Central Oregon, JASA Express Letters **3** p. 026002, (2023)
28. R. Bochenek, J. Austin, J-M. Dunaway, T. Vance. "Developing Big Data Infrastructure for Analyzing AIS Vessel Tracking Data on a Global Scale" in Big Data Analytics in Earth, Atmospheric, and Ocean Sciences (American Geophysical Union, 2022), pp. 273-292.
29. B. Fox-Kemper et al. Challenges and Prospects in Ocean Circulation Models. *Front. Mar. Sci.* **6:65**. doi: 10.3389/fmars.2019.00065 (2019)
30. J. Wilkin et al., Advancing coastal ocean modelling, analysis, and prediction for the US Integrated Ocean Observing System, *Journal of Operational Oceanography,* **10:2**, 115-126, DOI: 10.1080/1755876X.2017.1322026 (2017)
31. G.L. Mullendore, M.S. Mayernik, D.C. Schuster, Open Science Expectations for Simulation-Based Research. *Front. Clim.* **3**:763420. (2021) doi: 10.3389/fclim.2021.763420 DOI=10.3389/fmars.2019.00211
32. P.C. Reid, J.M. Colebrook, J.B.L. Matthews, J. Aiken, The Continuous Plankton Recorder: concepts and history, from Plankton Indicator to undulating recorders. *Progress in Oceanography* **58**, 117-173 (2003). https://doi.org/10.1016/j.pocean.2003.08.002.
33. P. González, A.Castaño, E.E. Peacock, J. Díez, J.J. Del Coz, H. M. Sosik, Automatic plankton quantification using deep features, *Journal of Plankton Research*, **41**, 449–463 (2019). https://doi.org/10.1093/plankt/fbz023
34. P. Suarez-Bregua et al., Environmental DNA (eDNA) for monitoring marine mammals: Challenges and opportunities. *Frontiers in Marine Science* **9** https://doi.org/10.3389/fmars.2022.987774 (2022).
35. Arias, P.A et al., Technical Summary. in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change,* V. Masson-Delmotte et al. Eds.. (Cambridge University Press 2012), pp 33−144. doi:10.1017/9781009157896.002
36. V. Eyring et al., Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958 (20. https://doi.org/10.5194/gmd-9-1937-2016.

37. R. Séférian et al., Tracking Improvement in Simulated Marine Biogeochemistry Between CMIP5 and CMIP6. *Curr Clim Change Rep* **6**, 95–119 (2020). https://doi.org/10.1007/s40641-020-00160-0
38. M.D. Priestley et al., An overview of the extratropical storm tracks in CMIP6 historical simulations. *Journal of Climate*, **33**, 6315-6343 (2020).
**39.** D. Carvalho, A. Rocha, X. Costoya, M. deCastro, M. Gómez-Gesteira, Wind energy resource over Europe under CMIP6 future climate projections: What changes from CMIP5 to CMIP6, *Renewable and Sustainable Energy Reviews* **151** https://doi.org/10.1016/j.rser.2021.111594, (2021).
40. D.J. Kaczan, J. Orgill-Meyer, The impact of climate change on migration: a synthesis of recent empirical insights. *Climatic Change* **158**, 281–300 (2020). https://doi.org/10.1007/s10584-019-02560-0
41. D. Beillouin, B. Schauberger, A. Bastos, P. Ciais, D. Makowski, Impact of extreme weather conditions on European crop production in 2018. *Philosophical Transactions of the Royal Society B* **375** https://doi.org/10.1098/rstb.2019.0510 (2020).
42. IPCC, *Global Warming of 1.5°C.An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* [Masson-Delmotte, V. et al. (eds.)]. (2018)
43. T.M. Lenton et al.,Climate tipping points—too risky to bet against. *Nature* **575**, 592-595 (2019).
44. P. Kedron, W. Li, S. Fotheringham, M. Goodchild, Reproducibility and replicability: opportunities and challenges for geospatial research, *International Journal of Geographical Information Science* DOI: 10.1080/13658816.2020.1802032 (2020).
45. National Academies of Sciences, Engineering, and Medicine (NASEM), *Reproducibility and replicability in science*. (The National Academies Press, 2019) doi:10.17226/25303
46. J.P. Wilson et al. , A five-star guide for achieving replicability and reproducibility when working with GIS software and algorithms. *Annals of the American Association of Geographers* **111**, 1311-1317 (2021).
47. H. Guo et al., Big Earth Data science: an information framework for a sustainable planet, *International Journal of Digital Earth*, **13:7**, 743-767, (2020) DOI: 10.1080/17538947.2020.1743785
48. D. Ezer, K. Whitaker, Data science for the scientific life cycle. *Elife* **8**, e43979 (2019).
49. ACRL Scholarly Communication Committee. SCHOLARLY COMMUNICATION: Principles and strategies for the reform of scholarly communication: Issues related to the formal system of scholarly communication. *College & Research Libraries News*, **64**, 526-547 (2003).
50. M. Fenner et al. , A data citation roadmap for scholarly data repositories. *Sci Data.* **10** doi: 10.1038/s41597-019-0031-8 (2019)
51. J.C. Arnott, C.J. Kirchhoff, R.M. Meyer, A. M. Meadow, A.T. Bednarek, Sponsoring actionable science: what public science funders can do to advance

sustainability and the social contract for science. *Current Opinion in Environmental Sustainability*, **42**, 38-44 (2020).
52. R. Sayre et al., A *New Map of Global Ecological Land Units — An Ecophysiographic Stratification Approach*. Association of American Geographers, (2014).
53. R.G. Sayre et al., A Three-Dimensional Mapping of the Ocean Based on Environmental Data. *Oceanography*, **30**, 90–103 (2017). http://www.jstor.org/stable/24897845
54. R. Sayre et al. A new 30 meter resolution global shoreline vector and associated global islands database for the development of standardized ecological coastal units, *Journal of Operational Oceanography*, **12** (sup2), S47-S56 (2019). DOI: 10.1080/1755876X.2018.1529714
55. L. James, "Digital twins will revolutionise healthcare: Digital twin technology has the potential to transform healthcare in a variety of ways – improving the diagnosis and treatment of patients, streamlining preventative care and facilitating new approaches for hospital planning," *Engineering & Technology* **16**, 50-53 (2012) doi: 10.1049/et.2021.0210.
56. M. Farsi, A. Daneshkhah, A. Hosseinian-Far, H. Jahankhani, *Digital Twin Technologies and Smart Cities* (Springer (2020) doi: 10.1007/978-3-030-18732-3
57. A. Fuller, Z. Fan, C. Day, C. Barlow, Digital Twin: Enabling Technologies, Challenges and Open Research, *IEEE Access* **8** doi:10.1109/ACCESS.2020.2998358 (2020).
58. T. Huang et al., Big Data Smart: Federated Earth System Digital Twins, International Geoscience and Remote Sensing Symposium (IGARSS), (2023)
59. T. Huang et al., Applications of Open-Source Digital Twins Framework for Wildfire and Air Quality, International Geoscience and Remote Sensing Symposium (IGARSS), (2023), Pasadena, CA.
60. R. Rodriguez-Suquet et al., R., The SCO-FloodDAM Project Toward a Digital Twin for Flood Detection, Prediction and Flood Risk Assessments," International Geoscience and Remote Sensing Symposium (IGARSS), 2023, Pasadena, CA.
61. P. Bauer, , B.Stevens, W.Hazeleger. A Digital Twin of Earth for the Green Transition. *Nature Climate Change* **11 80-83**, (2021) 80-83. https://doi.org/10.1038/s41558-021-00986-y.

**Acknowledgments:**

Thank you to Micah Wengren, Ronald Doel, and three anonymous reviewers for helpful comments and suggestions. Thank you to Brian Voss of the NOAA Central Library and Katelyn Szura of NOAA's Open Data Dissemination Program for bibliographic research and locating papers citing or using big data.

**Funding:** N/A

**Author contributions:** All authors contributed equally.

**Competing interests:** Authors declare that they have no competing interests.

**Data and materials availability:** N/A