Earth and Space Science

COMMENTARY
10.1029/2021EA002108

Special Section:
The Power of Many: Opportunities and Challenges of Integrated, Coordinated, Open, and Networked (ICON) Science to Advance Geosciences

Key Points:

- Networks across communities, with Coordinated data and information modeling practices, improve scientific outcomes for all involved
- Integrated, Coordinated, and Open data requires sustainable support to create and maintain infrastructure for interdisciplinary Networks
- Integrated and Coordinated use of data in machine learning calls for Open benchmark datasets shared across Networks for improved outcomes

Correspondence to:
D. J. Hills and J. E. Damerow,
denise.j.hills@gmail.com;
JoanDamerow@lbl.gov

Citation:
Hills, D. J., Damerow, J. E., Ahmmed, B., Catolico, N., Chakraborty, S., Coward, C. M., et al. (2022). Earth and Space Science Informatics perspectives on Integrated, Coordinated, Open, Networked (ICON) science. *Earth and Space Science*, *9*, e2021EA002108. https://doi.org/10.1029/2021EA002108

Received 4 NOV 2021
Accepted 18 FEB 2022

Earth and Space Science Informatics Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science

D. J. Hills[1,2], J. E. Damerow[3], B. Ahmmed[4], N. Catolico[5], S. Chakraborty[6], C. M. Coward[7], R. Crystal-Ornelas[3], W. D. Duncan[3], L. N. Goparaju[8], C. Lin[9], Z. Liu[6,10], M. K. Mudunuru[11], Y. Rao[12], R. J. Rovetto[13,14], Z. Sun[10], B. P. Whitehead[15], L. Wyborn[16], and T. Yao[6,17]

[1]Geological Survey of Alabama, Tuscaloosa, AL, USA, [2]Ronin Institute for Independent Scholarship, Tuscaloosa, AL, USA, [3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [4]Los Alamos National Laboratory, Los Alamos, NM, USA, [5]Battelle, National Ecological Observatory Network, Boulder, CO, USA, [6]NASA's Goddard Space Flight Center, Greenbelt, MD, USA, [7]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, [8]Vindhyan Ecology and Natural History Foundation, Mirzapur, India, [9]Atkinson Center for Sustainability and Department of Information Science, Cornell University, Ithaca, NY, USA, [10]George Mason University, Fairfax, VA, USA, [11]Pacific Northwest National Laboratory, Richland, WA, USA, [12]North Carolina Institute for Climate Studies, North Carolina State University, Asheville, NC, USA, [13]Center for Orbital Debris Education & Research, University of Maryland, College Park, MD, USA, [14]Independent, New York, NY, USA, [15]Manaaki Whenua – Landcare Research, Palmerston North, New Zealand, [16]Australian National University, Canberra, ACT, Australia, [17]Science Systems and Applications, Inc., Lanham, MD, USA

Abstract This article is composed of three independent commentaries about the state of Integrated, Coordinated, Open, Networked (ICON) principles (Goldman, et al., 2021b, https://doi.org/10.1029/2021EO153180) in Earth and Space Science Informatics (ESSI) and includes discussion on the opportunities and challenges of adopting them. Each commentary focuses on a different topic: (Section 2) Global collaboration, cyberinfrastructure, and data sharing; (Section 3) Machine learning for multiscale modeling; (Section 4) Aerial and satellite remote sensing for advancing Earth system model development by integrating field and ancillary data. ESSI addresses data management practices, computation and analysis, and hardware and software infrastructure. Our role in ICON science therefore involves collaborative work to assess, design, implement, and promote practices and tools that enable effective data management, discovery, integration, and reuse for interdisciplinary work in Earth and space science disciplines. Networks of diverse people with expertise across Earth, space, and data science disciplines are essential for efficient and ethical exchanges of findable, accessible, interoperable, and reusable (FAIR) research products and practices. Our challenge is then to coordinate the development of standards, curation practices, and tools that enable integrating and reusing multiple data types, software, multi-scale models, and machine learning approaches across disciplines in a way that is as open and/or FAIR as ethically possible. This is a major endeavor that could greatly increase the pace and potential of interdisciplinary scientific discovery.

Plain Language Summary We present commentaries on the state of "Integrated, Coordinated, Open, Networked (ICON) principles" in Earth and Space Science Informatics. ICON principles are meant to improve the research experience for all. Ultimately, data standardized according to community conventions and formats lead to more effective and efficient collaboration, data discovery, integration, and analyses. Data standards, tools, and machine learning developed using ICON principles enhance our understanding of Earth processes. Using ICON principles improves model results and efficacy, fosters interdisciplinary research, and provides a framework by which non-experts can confidently contribute volunteered data and findings. Standardized data also provides reliable common resources to help train and benchmark machine learning algorithms. When networked communities work together to standardize and share data openly, the resulting web of research products is more readily findable, accessible, interoperable, and reusable. Ongoing support is crucial to develop and sustain the people, systems, and tools necessary to embrace ICON principles in Earth and Space Science Informatics now and in the future.

1. Introduction

Integrated, Coordinated, Open, Networked (ICON) science aims to enhance synthesis, increase resource efficiency, and create transferable knowledge (Goldman, et al., 2021b). This article belongs to a collection of

Author Contributions:

Conceptualization: D. J. Hills, J. E. Damerow, B. Ahmmed, N. Catolico, S. Chakraborty, R. Crystal-Ornelas, W. D. Duncan, L. N. Goparaju, C. Lin, Z. Liu, M. K. Mudunuru, Y. Rao, R. J. Rovetto, Z. Sun, B. P. Whitehead, L. Wyborn, T. Yao

Writing – original draft: D. J. Hills, J. E. Damerow, B. Ahmmed, N. Catolico, S. Chakraborty, C. M. Coward, R. Crystal-Ornelas, W. D. Duncan, L. N. Goparaju, C. Lin, Z. Liu, M. K. Mudunuru, Y. Rao, Z. Sun, B. P. Whitehead, L. Wyborn, T. Yao

Writing – review & editing: D. J. Hills, J. E. Damerow, N. Catolico, S. Chakraborty, C. M. Coward, R. Crystal-Ornelas, W. D. Duncan, L. N. Goparaju, C. Lin, Z. Liu, M. K. Mudunuru, Y. Rao, R. J. Rovetto, Z. Sun, B. P. Whitehead, L. Wyborn, T. Yao

commentaries (Goldman et al., 2021a) spanning geoscience on the state and future of ICON science. Earth and Space Science Informatics (ESSI) encompasses a broad field that addresses data management practices, computation and analysis, and hardware and software infrastructure. ESSI's role in ICON science therefore involves collaborative work to assess, design, implement, and promote practices and tools that enable effective data management, discovery, integration, and reuse for interdisciplinary work in Earth and space science (ESS) disciplines. In this series of commentaries, we examine the current state, challenges, and opportunities of ICON science through the lenses of global collaboration, cyberinfrastructure, and data sharing (Section 2); machine learning and multiscale modeling (Section 3); and remote sensing for advancing Earth system models (ESM) development by integrating field and ancillary data (Section 4).

2. Global Collaboration, Cyberinfrastructure, and Data Sharing

2.1. Current Status

Global collaboration across disciplines is essential to the development and implementation of data/metadata standards and cyberinfrastructures. Thus, many organizations have emerged to facilitate such collaboration, for example, Research Data Alliance, World Data System, Earth Science Information Partners. These organizations have produced numerous active groups involved in Earth, space and environmental science data and research, and developed many data tools and services, for example, Earth, Space and Environmental Sciences Data Vocabulary Repositories. Research is more efficient with Networked data practices and cyberinfrastructures that support scientific discovery. Yet, there is still a large disconnect and lack of Coordination across many informatics communities and the broader communities we aim to support.

Research teams often lack sufficient resources (e.g., appropriate cyberinfrastructure, expert data/software personnel, financial allotment) to effectively manage, standardize, and publish high-quality data (Mons, 2020). This hinders data from being Open and/or Findable, Accessible, Interoperable, and Reusable (FAIR; Wilkinson et al., 2016). Further, specific criteria to implement the FAIR Guiding Principles (Gries et al., 2019; Jones et al., 2019) inevitably vary across disciplines and data types as inconsistencies in interpretations of the principles have grown (e.g., Kinkade & Shepherd, 2021; Mons et al., 2017; Stall et al., 2019). Importantly, FAIR does not mean Open; data can be Open without being FAIR, and vice versa (see What is the difference between “FAIR data” and “Open data” if there is one?). Thus, even if the data cannot be fully Open, it is still possible for the science itself to be Open, or at least transparent.

Supporting ESS research requires assessing, designing, building, and maintaining cyberinfrastructures (e.g., data repositories/archives, application programming interfaces (APIs), visualization tools, search interfaces) that are often organized around a particular data type, discipline, or organization (e.g., Pertzold et al., 2019). Ever-increasing volumes of open data and tools now allow us to ask science questions that synthesize data and knowledge across scientific disciplines from globally distributed resources, thus expanding the impact of funded research (e.g., Michener, 2015; Rosenberg et al., 2019). More successful Networked data sharing efforts (e.g., Global Biodiversity Information Facility, Ameriflux, Consortium of Universities for the Advancement of Hydrologic Science, Inc., Long-Term Ecological Research Network, National Ecological Observatory Network, Deep Carbon Observatory, HydroShare) have been driven by (a) demand for and funding to support a specific data type (Barrett et al., 2012; Novick et al., 2018; Robertson et al., 2014); (b) reporting standards that enable global data search and integration (e.g., Wieczorek et al., 2012; Yilmaz et al., 2011); and (c) associated user-friendly tools (Clark et al., 2016; Robertson et al., 2014).

2.2. Challenges and Opportunities

Most cyberinfrastructures lack the resources for Integration and Coordination necessary for broader interdisciplinary work, including guidance and leading practices; domain semantics; technical, data, methodological, and instrumentation standards; workflow management; training; and sustainable technical and financial support. These deficits hinder the availability of Open data that could foster machine actionable, interdisciplinary scientific discovery. While existing standards and practices may address similar concepts, they are not fully interoperable or Integrated within and across relevant disciplines. Valuable resources are spent developing/updating translators, or disciplinary standards are simply disconnected and inefficient for interdisciplinary users. Coordination is needed to implement standards for effective interdisciplinary data discovery and exchange. A major challenge to Coordination involves a lack of consistent and transparent protocols (e.g., data and code production, processing methods) across interdisciplinary teams. Further, informatics initiatives and working groups (e.g., RDA, ESIP) are primarily volunteer-based without appropriate recognition or funding that would accelerate and improve this work. These combined factors create barriers to Open and FAIR data.

Replicable and transparent research that reflects ICON principles requires sustainable investment in cyberinfrastructure to improve interoperability and Integration. Global high-level Coordination across organizations is needed to bridge siled efforts across disciplines, organizations, and/or countries. A commitment to community engagement is needed to bring together input across disciplines, understand data management challenges and needs, and promote the adoption of shared practices. Making data as Open and/or FAIR as ethically possible requires key advocates who facilitate Networked collaboration.

Data users, code contributors, and tool developers should align with established standards or community practices. We can encourage practices that promote ICON principles, such as Open publication of study plans (e.g., PLOS ONE study proposals), data production and processing protocols (e.g., Common Workflow Language), and software code. We must continually evaluate how to Coordinate and Integrate across existing cyberinfrastructure from local to global scales, which involves iterative rounds of engagement; education and outreach; and feedback across data providers, tool and service creators, and scientists who use ESS data and services. Coordinating Networks across disciplines will involve technical approaches to connect related data (e.g., globally unique and resolvable persistent identifiers (PIDs), APIs, ontologies, geospatial standards) and promoting widespread adoption of community standards that improve scientific outcomes and benefit all participants in the Network. Coordination is also key to shifting legacy cyberinfrastructure and data to be more ICON-aligned.

3. Machine Learning for Multiscale Modeling

3.1. Current Status

Over the past decade, artificial intelligence approaches, including machine learning (AI/ML), have revolutionized scientific discovery across disciplines, including Earth and space science informaticsinformatics (Maskey, Alemohannad, et al., 2020). The AI/ML revolution, driven by a wealth of Open data and rapid technological development in computational cyberinfrastructure, has led to more processing power and greater Networking between cyberinfrastructure as well as data generators and data users which allows unprecedented resource and data sharing. There are many success stories demonstrating how AI/ML has been used to address challenging issues in ESS, for example, extreme weather prediction (Maskey, Ramachandran, et al., 2020; Pradhan et al., 2018; Wimmers et al., 2019), land use/land cover change monitoring (Hansen et al., 2013), Earth system modeling (Reichstein et al., 2019), endangered species identification (Allen et al., 2021), spatial downscaling of climate models and satellite observations (López López et al., 2018; Vandal et al., 2019), space weather forecasting (Wintoft et al., 2017), and lunar and planetary landform classification (Palafox et al., 2017; Silburt et al., 2019). Various funding agencies worldwide have recently released their strategic plans and guidelines to expand the investment in AI/ML research which will further its adoption within informatics for at least the next decade to accelerate scientific discovery and address pressing societal issues such as combatting climate change, facilitating the energy transition, and ensuring food security.

3.2. Challenges and Opportunities

To accelerate this adoption, the ESS community needs to collectively address several key challenges to make AI/ML in ESS more efficient and ICON-aligned. Most AI/ML applications in ESS are ad hoc research that lacks system-wide Coordination and are time-consuming. There are little AI-ready data (e.g., cleaned, harmonized, formatted, well documented) that can be efficiently Integrated across domains or applications and few recommended practices on proper model development and documentation (Maskey, Alemohammad, et al., 2020). As the capacity and application scope of AI/ML heavily depends on patterns in training data, it should be as representative as possible. These requirements for big training datasets have led to calls for libraries of Open and FAIR benchmark datasets (WILDS, Koh et al., 2020; Radiant Earth Foundation; Rasp et al., 2020) related to questions within ESS (Crystal-Ornelas et al., 2021).

AI-ready training datasets and standardized AI/ML model development practices would enable the ESS community to collaboratively develop open AI/ML applications at scale. However, there are no current community-recommended practices on how to properly develop, document, and share the AI/ML applications that track provenance and enable reproducibility (Sun et al., 2020). Increased connection through cloud computing (Gorelick et al., 2017; Mayer-Schönberger & Cukier, 2013) allows sharing data and models in the cloud, enabling Networked researchers around the world access to these resources without being limited by local computing power. However, despite recent progress, work needs to be done to make cloud computing more accessible. Increased Openness in the exchange of data handling practices allows sharing common workflows while handling large datasets. Integration across disciplines could be improved by: (a) including physics in ML models (Jia et al., 2019; Raissi et al., 2019), (b) leveraging ML exploratory tools (Montavon et al., 2017; Ying et al., 2019), and (c) better mechanism for codevelopment between domain experts and AI/ML developers. Coordination via automated workflows would improve development efficiency (e.g., auto-sklearn, AutoKeras) (He et al., 2021). To improve AI engineering efficiency and reduce data collection and processing costs, developers may also use data augmentation methods such as mixup (Zhang et al., 2017) to fill in the missing data and enhance data quality (Alexandrov & Vesselinov, 2014; Vesselinov et al., 2018).

The ability to readily interpret and generalize AI/ML models are also major concerns for the ESS community (McGovern et al., 2019; Toms et al., 2020). To address complex questions in ESS systems, we need to better understand why AI/ML models perform in a certain way, their consistency with domain knowledge, and how models developed using a specific set of data can adjust dynamically to shifts in ESS data. To address these concerns, the ESS community should establish benchmark tasks with Open and standardized data and a Coordinated evaluation framework to enhance future development. Licensing approaches are still evolving, highlighting the need for increased Coordination on policy and ethics considerations. Ethical awareness, conduct, and responsibility in AI/ML and related activities are essential to the practice of principled research; while beyond the scope of this paper, some particular concerns include misleading results due to biased training data; cognitive biases in general; and incorrect annotation, classification or characterization of data. AI/ML applications heavily rely on input data, thus the ESS community needs to establish Coordinated standards that clarify the impact of input data quality on downstream applications to ensure trustworthiness. These community standards should Integrate both domain sciences and social sciences.

4. Aerial and Satellite Remote Sensing for Advancing Earth System Model Development by Integrating Field and Ancillary Data

4.1. Current Status

Remote sensing technology combined with field and ancillary data (e.g., field measurements, other imagery; Acton, 1996) provides a compelling example of how dedicated resources supporting ICON science and advanced AI/ML technologies have transformed the development of ESMs as they have advanced from aerial imagery of the early nineteenth century (Necsoiu et al., 2013) to the present-day's Google Earth Engine (Gorelick et al., 2017) and Unmanned Aerial Vehicles (Singh & Frazier, 2018). Most publicly-funded remote sensing datasets are Open and hosted on public repositories (e.g., government-sponsored repositories, Github, Zenodo). In addition, this data is distributed through Coordinated standards between government agencies across the globe (Alameh, 2020). Integration of remote sensing technology with independent field measurements and high spatial resolution satellite imagery has been essential for ESM validation. This also includes estimating derived data products (e.g., from satellites) accuracy and quantifying uncertainty (Strahler et al., 2006). Crowdsourcing and citizen science have further advanced the integration of remote sensing with field data (e.g., RaspberryShake, Khan et al., 2018; Saralioglu & Gungor, 2020; Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems [WHONDRS], Stegen & Goldman, 2018), resulting in broader Networked efforts that benefit researchers and a wide variety of data users. Many agencies in the US and Europe have made some or all of their data Open to all users internationally. Some examples, associated cyberinfrastructure, and tools are included in an associated github repository.

4.2. Challenges and Opportunities

Two primary challenges that the ESM community still faces are limited global data collection and inadequate cyberinfrastructure. Despite advances in sensors, crowdsourcing, and citizen science (e.g., RaspberryShake, WHONDRS), collecting and hosting high-quality global data present immense challenges. For example, RaspberryShake has collected more than 30 TB of seismographic data over the past decade but lacks the necessary cyberinfrastructure to reliably and sustainably store it.

Recent progress in AI/ML has improved available data to represent Earth system processes (e.g., thermal, land physics and hydrology, radiation, atmospheric ocean circulation) in ESMs (Rasp et al., 2018). ML, in particular, requires massive datasets to represent processes at both normal and extreme events (e.g., hurricanes, wildfires); however, extreme event data are rare due to the unique challenges faced during collection. Thus, the concept of crowdsourcing data collection, using Coordinated methods (e.g., RaspberryShake, WHONDRS) on extreme events, is an attractive option that improves Networked research.

There has been a Coordinated effort from US and European agencies to develop cyberinfrastructure that improves and increases access to data to enhance predictions and understanding of various Earth system processes. For example, the European Space Agency Sentinel data products are recently available in the Copernicus Data and Information Access Service cloud environments. In addition, the US Geological Survey Landsat satellite data inventory has been Open to the public since 2008 and has been in the cloud since 2020 (U.S. Geological Survey, 2008). Furthermore, the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) have adopted a strategic vision to leverage cloud computing and operate multiple components of their data systems in a retail cloud environment. This calls for action to identify the opportunities to improve policy and strategy planning across various countries to make satellite data products accessible to all users in open data portals. In addition, automated quality assurance of satellite observations is needed to support global, regional, or local data services. Coordinated across international agencies, a standard open data cyberinfrastructure will help to assure ESM data from multiple sources (national, regional, governments, academia, and the private sector) are available and easily Integrated into open-source platforms and networks.

Coordination would help international agencies and organizations build a standard open data cyberinfrastructure to ensure that Earth science data are free, Open, and easily Integrated into ESMs. We also need next-generation sensors and satellites which provide more fine resolution data to increase the accuracy of ESMs. For example, the joint NASA-Indian Space Research Organization (ISRO) Synthetic Aperture Radar (SAR) (NISAR) mission is anticipated to provide Open radar data with a spatial resolution of less than a centimeter to Integrate into ESM for studying the Earth's features and processes. The role of AI/ML needs to be expanded to fill the gaps of remote sensing data.

5. Concluding Remarks

Earth and space science research facilitated by modern informatics techniques that follow the ICON principles enables data synthesis, increases resource efficiency, and creates knowledge that transcends individual systems (Goldman, et al., 2021b). ESSI can work to ensure that diverse scientists have user-friendly resources to contribute and use data that follows community conventions. Such collections of Open and/or FAIR data, shared across Networks for mutual benefit, are critical to appropriately train AI/ML, which furthers Integration and Coordination in Earth and space science informatics. Cross-community Networks improve scientific outcomes for all involved. Communities must work together to share data openly using community standards, to produce Open and/or FAIR data that enables data synthesis and can revolutionize fields of research (e.g., Kelling et al., 2009). Ongoing, sustainable support is vital to create and maintain the cyberinfrastructure and human resources necessary for Integrated, Coordinated, and Open and/or FAIR data (as much ethically as possible) for interdisciplinary Networks.

Data Availability Statement

No data was used for this commentary.

Acknowledgments

D. J. Hills, J. E. Damerow, N. Catolico, C. M. Coward, W. D. Duncan, Z. Liu, R. J. Rovetto, B. P. Whitehead, and L. Wyborn authored Section 2 "Global collaboration, cyberinfrastructure, and data sharing." R. Crystsl-Ornelas, S. Chakraborty, B. Ahmmed, C. Lin, Y. Rao, T. Yao, and Z. Sun authored Section 3 "Machine learning for multiscale modeling." L. N. Goparaju, M. K. Mudunuru, and TY authored Section 4 "Aerial and satellite remote sensing for advancing Earth system model development by integrating field and ancillary data." J. E. Damerow and R. Crystal-Ornelas were funded by the ESS-DIVE repository and W. D. Duncan by the National Microbiome Data Collaborative, both by the U.S. DOE's Office of Science Biological and Environmental Research under contract number DE-AC02-05CH11231. N. Catolico was supported by NEON, a program sponsored by the NSF and operated under cooperative agreement by Battelle Memorial Institute. S. Chakraborty was supported by an appointment to the NASA Postdoctoral Program at NASA Goddard Space Flight Center, administered by Universities Space Research Association under contract with NASA. C. Lin was supported by an appointment as a postdoctoral fellow at the Cornell Atkinson Center for Sustainability, and an affiliation with the Department of Information Science. Z. Liu was supported by Cooperative Geoinformation Research with the NASA GSFC Earth Sciences Data and Information Services Center (GES DISC). M. K. Mudunuru was supported by the U.S. DOE-SC, SFA at PNNL. Y. Rao was supported by NOAA through the Cooperative Institute for Satellite Earth System Studies under Cooperative Agreement NA19NES4320002. B. P. Whitehead was supported by the Ministry of Business Innovation and Employment (MBIE) Infrastructure Platform. T. Yao was supported by NASA Applied Sciences Disasters Program and NASA's LANCE system, part of NASA's EOSDIS. C. M. Coward's research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). The views and opinions of authors expressed herein do not necessarily state or reflect those of the US Government or any international agency thereof.

References

Acton, C. H. (1996). Ancillary data services of NASA's navigation and Ancillary information facility. *Planetary and Space Science*, *44*(1), 65–70. https://doi.org/10.1016/0032-0633(95)00107-7

Alameh, N. (2020). A future of location data integration. *Geo: GeoConnexion International Magazine*, *19*(6), 18–19. Retrieved from https://www.geoconnexion.com/publication-articles/a-future-of-location-data-integration

Alexandrov, B. S., & Vesselinov, V. V. (2014). Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. *Water Resources Research*, *50*(9), 7332–7347. https://doi.org/10.1002/2013wr015037

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., et al. (2021). A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Frontiers in Marine Science*, *8*, 165. https://doi.org/10.3389/fmars.2021.607321

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Research*, *40*(D1), D57–D63. https://doi.org/10.1093/nar/gkr1163

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Crystal-Ornelas, R., Varadharajan, C., Christianson, D., Damerow, J., Weierbach, H., Robles, E., et al. (2021). *A library of AI-assisted FAIR water cycle and related disturbance datasets to enable model training, parameterization and validation*. Office of Scientific and Technical Information (OSTI). https://doi.org/10.2172/1769646

Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A., & Stegen, J. C. (2021a). Integrated, Coordinated, Open, and Networked (ICON) science to advance the geosciences: Introduction and synthesis of a special collection of commentary articles. *Earth and Space Science Open Archive*. https://doi.org/10.1002/essoar.10508554.1

Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A., Stegen, J. C., & Fox, P. (2021b). Special collection on open collaboration across geosciences (Vol. *102*). *Eos*. https://doi.org/10.1029/2021eo153180

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Gries, C., Servilla, M., O'Brien, M., Vanderbilt, K., Smith, C., Costa, D., & Grossman-Clarke, S. (2019). Achieving FAIR data principles at the environmental data initiative, the US-LTER data repository. *Biodiversity Information Science and Standards*, *3*, e37047. https://doi.org/10.3897/biss.3.37047

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850–853. https://doi.org/10.1126/science.1244693

He, X., Zhao, K., & Chu, X. (2021). AutoML: A Survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM international conference on Data Mining*. Society for Industrial and Applied Mathematics. (SDM) (pp. 558–566). https://doi.org/10.1137/1.9781611975673.63

Jones, M. B., Slaughter, P., & Habermann, T. (2019). *Quantifying FAIR: Automated metadata improvement and guidance in the DataONE repository network*. https://doi.org/10.5281/zenodo.3408466

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity Studies. *BioScience*, *59*(7), 613–620. https://doi.org/10.1525/bio.2009.59.7.12

Khan, A., Denton, P., Stevenson, J., & Bossu, R. (2018). Engaging citizen seismologists worldwide. *Astronomy and Geophysics*, *59*(4), 4.15–4.18. https://doi.org/10.1093/astrogeo/aty190

Kinkade, D., & Shepherd, A. (2021). Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. *Geoscience Data Journal*. gdj3.120. https://doi.org/10.1002/gdj3.120

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2020). *WILDS: A benchmark of in-the-Wild distribution shifts. arXiv [cs.LG]*. Retrieved from https://arxiv.org/abs/2012.07421

López López, P., Immerzeel, W. W., Rodríguez Sandoval, E. A., Sterk, G., & Schellekens, J. (2018). Spatial downscaling of satellite-based precipitation and its impact on discharge simulations in the Magdalena river basin in Colombia. *Frontiers of Earth Science in China*, *6*, 68. https://doi.org/10.3389/feart.2018.00068

Maskey, M., Alemohammad, H., Murphy, K. J., & Ramachandran, R. (2020). Advancing AI for Earth Science: A Data Systems Perspective (Vol. *101*). *Eos*. https://doi.org/10.1029/2020eo151245

Maskey, M., Ramachandran, R., Ramasubramanian, M., Gurung, I., Freitag, B., Kaulfus, A., et al. (2020). Deepti: Deep-Learning-Based tropical cyclone intensity estimation system. In *IEEE journal of selected topics in applied Earth observations and remote sensing*. https://doi.org/10.1109/jstars.2020.3011907

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. Retrieved from https://play.google.com/store/books/details?id=uy4lh-WEhhIC

McGovern, A., Lagerquist, R., Gagne, D. J., Eli Jergensen, G., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*(7796), 491. https://doi.org/10.1038/d41586-020-00505-7

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/isu-170824

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Necsoiu, M., Dinwiddie, C. L., Walter, G. R., Larsen, A., & Stothoff, S. A. (2013). Multi-temporal image analysis of historical aerial photographs and recent satellite imagery reveals evolution of water body surface area and polygonal terrain morphology in Kobuk Valley National Park, Alaska. *Environmental Research Letters*, *8*(2), 025007. https://doi.org/10.1088/1748-9326/8/2/025007

Novick, K. A., Biederman, J. A., Desai, A. R., Litvak, M. E., Moore, D. J. P., Scott, R. L., & Torn, M. S. (2018). The AmeriFlux network: A coalition of the willing. *Agricultural and Forest Meteorology*, *249*, 444–456. https://doi.org/10.1016/j.agrformet.2017.10.009

Palafox, L. F., Hamilton, C. W., Scheidt, S. P., & Alvarez, A. M. (2017). Automated detection of geological landforms on mars using convolutional neural networks. *Computers & Geosciences*, *101*, 48–56. https://doi.org/10.1016/j.cageo.2016.12.015

Petzold, A., Asmi, A., Vermeulen, A., Pappalardo, G., Bailo, D., Schaap, D., et al. (2019). ENVRI-FAIR - Interoperable Environmental FAIR Data and Services for Society, Innovation and Research. *2019 15th International Conference on eScience (eScience)* (pp. 277–280). https://doi.org/10.1109/eScience.2019.00038

Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2018). Tropical cyclone intensity estimation using a deep convolutional neural network. In *IEEE transactions on image processing*. https://doi.org/10.1109/tip.2017.2766358

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020ms002203

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PloS One*, *9*(8), e102623. https://doi.org/10.1371/journal.pone.0102623

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., et al. (2019). Decline of the north American avifauna. *Science*, eaaw1313. https://doi.org/10.1126/science.aaw1313

Saralioglu, E., & Gungor, O. (2020). Crowdsourcing in remote sensing: A review of applications and future directions. *IEEE Geoscience and Remote Sensing Magazine*, *8*(4), 89–110. https://doi.org/10.1109/mgrs.2020.2975132

Silburt, A., Ali-Dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., et al. (2019). Lunar crater identification via deep learning. *Icarus*, *317*, 27–38. https://doi.org/10.1016/j.icarus.2018.06.022

Singh, K. K., & Frazier, A. E. (2018). A meta-analysis and review of unmanned aircraft system (UAS) imagery for terrestrial applications. *International Journal of Remote Sensing*, *39*(15–16), 5078–5098. https://doi.org/10.1080/01431161.2017.1420941

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., et al. (2019). Make scientific data FAIR. *Nature*, *570*(7759), 27. https://doi.org/10.1038/d41586-019-01720-7

Stegen, J. C., & Goldman, A. E. (2018). WHONDRS: A community resource for studying Dynamic River corridors. *mSystems*, *3*(5), e00151–18. https://doi.org/10.1128/msystems.00151-18

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps* (Publication EUR 22156 EN). European Commission, Joint Research Center. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/52730469-6bc9-47a9-b486-5e2662629976

Sun, Z., Di, L., Burgess, A., Tullis, J. A., & Magill, A. B. (2020). Geoweaver: Advanced cyberinfrastructure for managing hybrid geoscientific AI workflows. *ISPRS International Journal of Geo-Information*, *9*(2), 119. https://doi.org/10.3390/ijgi9020119

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9). https://doi.org/10.1029/2019ms002002

U.S. Geological Survey. (2008). *Imagery for everyone: Timeline set to release entire USGS Landsat archive at No charge*. Retrieved from https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGStechann-20080421-landsat-imagery-release.pdf

Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theoretical and Applied Climatology*, *137*(1–2), 557–570. https://doi.org/10.1007/s00704-018-2613-3

Vesselinov, V. V., Alexandrov, B. S., & O'Malley, D. (2018). Contaminant source identification using semi-supervised machine learning. *Journal of Contaminant Hydrology*, *212*, 134–142. https://doi.org/10.1016/j.jconhyd.2017.11.002

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PloS One*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

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. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wimmers, A., Velden, C., & Cossuth, J. H. (2019). Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, *147*(6), 2261–2282. https://doi.org/10.1175/mwr-d-18-0391.1

Wintoft, P., Wik, M., Matzka, J., & Shprits, Y. (2017). Forecasting Kp from solar wind data: Input parameter study using 3-hour averages and 3-hour range values. *Journal of Space Weather and Space Climate*, *7*, A29. https://doi.org/10.1051/swsc/2017027

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420. https://doi.org/10.1038/nbt.1823

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, *32*, 9240–9251. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/32265580

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1710.09412