

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

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

<sup>1</sup>[Geological Survey of Alabama](#), Tuscaloosa, AL, USA.

<sup>2</sup>[Ronin Institute for Independent Scholarship](#), Tuscaloosa, AL, USA.

<sup>3</sup>[Lawrence Berkeley National Laboratory](#), Berkeley, CA, USA.

<sup>4</sup>[Los Alamos National Laboratory](#), Los Alamos, NM, USA.

<sup>5</sup>[Battelle, National Ecological Observatory Network](#), CO, USA.

<sup>6</sup>[NASA's Goddard Space Flight Center](#), Greenbelt, MD, USA.

<sup>7</sup>[Jet Propulsion Laboratory](#), California Institute of Technology, Pasadena, CA, USA.

<sup>8</sup>Vindhyan Ecology and Natural History Foundation, U.P. India.

<sup>9</sup>Atkinson Center for Sustainability and Department of Information Science, [Cornell University](#), Ithaca, NY, USA.

<sup>10</sup>[George Mason University](#), Fairfax, VA, USA.

<sup>11</sup>[Pacific Northwest National Laboratory](#), Richland, WA, USA.

<sup>12</sup>North Carolina Institute for Climate Studies, [North Carolina State University](#), Asheville, NC, USA.

<sup>13</sup>Center for Orbital Debris Education & Research, [University of Maryland](#), MD, USA.

<sup>14</sup>Independent, New York, NY, USA.

<sup>15</sup>[Manaaki Whenua – Landcare Research](#), Palmerston North, New Zealand

<sup>16</sup>[Australian National University](#), Canberra, ACT, Australia

<sup>17</sup>[Science Systems and Applications, Inc.](#), Lanham, MD, USA.

Corresponding authors: Denise Hills ([denise.j.hills@gmail.com](mailto:denise.j.hills@gmail.com)); Joan Damerow ([JoanDamerow@lbl.gov](mailto:JoanDamerow@lbl.gov))

## 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.

## 35 **Abstract**

36 This article is composed of three independent commentaries about the state of ICON principles  
37 (Goldman et al., 2021) in Earth and Space Science Informatics (ESSI) and includes discussion on  
38 the opportunities and challenges of adopting them. Each commentary focuses on a different  
39 topic: **(Section 2)** Global collaboration, cyberinfrastructure, and data sharing; **(Section 3)**  
40 Machine learning and multiscale modeling; **(Section 4)** Remote sensing for advancing Earth  
41 system model development by integrating field and ancillary data. ESSI addresses data  
42 management practices, computation and analysis, and hardware and software infrastructure. Our  
43 role in ICON science therefore involves collaborative work to assess, design, implement, and  
44 promote practices and tools that enable effective data management, discovery, integration, and  
45 reuse for interdisciplinary work in Earth and space science disciplines. Networks of diverse  
46 people with expertise across Earth, space, and data science disciplines are essential for efficient  
47 and ethical exchanges of FAIR research products and practices. Our challenge is then to  
48 coordinate the development of standards, curation practices, and tools that enable integrating and  
49 reusing multiple data types, software, multi-scale models, and machine learning approaches  
50 across disciplines in a way that is as open and/or FAIR as ethically possible. This is a major  
51 endeavor that could greatly increase the pace and potential of interdisciplinary scientific  
52 discovery.

## 53 **Plain Language Summary**

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

## 68 **1 Introduction**

69 Integrated, Coordinated, Open, Networked (ICON) science aims to enhance synthesis,  
70 increase resource efficiency, and create transferable knowledge (Goldman et al., 2021). This  
71 article belongs to a collection of commentaries ([Leadership Team et al., 2021](#)) spanning  
72 geoscience on the state and future of ICON science. Earth and Space Science Informatics (ESSI)  
73 encompasses a broad field that addresses data management practices, computation and analysis,  
74 and hardware and software infrastructure. ESSI's role in ICON science therefore involves  
75 collaborative work to assess, design, implement, and promote practices and tools that enable  
76 effective data management, discovery, integration, and reuse for interdisciplinary work in Earth  
77 and space science (ESS) disciplines. In this series of commentaries, we examine the current state,  
78 challenges, and opportunities of ICON science through the lenses of global collaboration,

79 [cyberinfrastructure](#), and data sharing (Section 2); machine learning and multiscale modeling  
80 (Section 3); and remote sensing for advancing Earth system models (ESM) development by  
81 integrating field and ancillary data (Section 4).

## 82 **2 Global collaboration, cyberinfrastructure, and data sharing**

### 83 2.1 Current state and challenges

84 Global collaboration across disciplines is essential to the development and  
85 implementation of data/metadata standards and cyberinfrastructures. Thus, many organizations  
86 have emerged to facilitate such collaboration, e.g., [Research Data Alliance](#), [World Data System](#),  
87 [OneGeology](#), [Earth Science Information Partners](#). These organizations have produced numerous  
88 active [groups involved in Earth, space and environmental science data and research](#), and  
89 developed many data tools and services, e.g. [Earth, Space and Environmental Sciences Data](#)  
90 [Vocabulary Repositories](#). Research is more efficient with **Networked** data practices and  
91 cyberinfrastructures that support scientific discovery. Yet, there is still a large disconnect and  
92 lack of **Coordination** across many informatics communities and the broader communities we  
93 aim to support.

94 Research teams often lack sufficient resources (e.g., appropriate cyberinfrastructure,  
95 expert data/software personnel, financial allotment) to effectively manage, standardize, and  
96 publish high-quality data (Mons, 2020). This hinders data from being **Open and/or Findable**,  
97 **Accessible, Interoperable, and Reusable** (FAIR; Wilkinson et al., 2016). Further, specific  
98 criteria to make data FAIR (Gries et al., 2019; Jones et al., 2019) inevitably vary across  
99 disciplines and data types. Because there are no widely accepted standards to evaluate FAIR-  
100 ness, data may be miscaterogized (e.g., Kinkade & Shepherd, 2021; Mons et al., 2017; Stall et  
101 al., 2019). Importantly, FAIR does not mean **Open**; data can be **Open** without being FAIR, and  
102 *vice versa* (see [What is the difference between “FAIR data” and “Open data” if there is one?](#)).

103 Supporting ESS research requires assessing, designing, building, and maintaining  
104 cyberinfrastructures (e.g., data repositories/archives, application programming interfaces (APIs),  
105 visualization tools, search interfaces) that are often organized around a particular data type,  
106 discipline, or organization. Interoperability issues are then minimized using bespoke or *ad hoc*  
107 conventions within that particular community (e.g., [Deep Carbon Observatory](#), [HydroShare](#),  
108 [Long-Term Ecological Research Network](#), [National Ecological Observatory Network](#)). However,  
109 most cyberinfrastructures lack the resources for **Integration** and **Coordination** necessary for  
110 interdisciplinary work, including guidance and leading practices; domain semantics; technical,  
111 data, methodological, and instrumentation standards; workflow management; training; and  
112 sustainable technical and financial support. These deficits hinder **Open** data that fosters machine  
113 actionable, interdisciplinary scientific discovery.

114 While existing standards and practices may address similar concepts, they are not fully  
115 interoperable or **Integrated** within and across relevant disciplines. Valuable resources are spent  
116 developing/updating translators, or disciplinary standards are simply disconnected and inefficient  
117 for interdisciplinary users. **Coordination** is needed to implement standards for effective  
118 interdisciplinary data discovery and exchange. A major limitation to **Coordination** involves a  
119 lack of consistent and transparent protocols (e.g., data and code production, processing methods)  
120 across interdisciplinary teams that limits reuse and replication. These combined factors create  
121 barriers to **Open and FAIR** data.

122 Ever-increasing volumes of open data and tools now allow us to ask science questions  
123 that synthesize data and knowledge across scientific disciplines from globally distributed  
124 resources, thus expanding the impact of funded research (e.g., Michener, 2015; Rosenberg et al.,  
125 2019). More successful **Networked** data sharing efforts (e.g., [Global Biodiversity Information](#)  
126 [Facility](#), [Ameriflux](#), [Consortium of Universities for the Advancement of Hydrologic Science,](#)  
127 [Inc.](#)) have been driven by 1) demand for a specific data type (Barrett et al., 2012; Novick et al.,  
128 2018; Robertson et al., 2014); 2) reporting standards that enable global data search and  
129 integration (e.g., Wieczorek et al., 2012; Yilmaz et al., 2011); and 3) associated user-friendly  
130 tools (Clark et al., 2016; Robertson et al., 2014).

## 131 2.2. Opportunities and moving forward

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

139 Data users, code contributors, and tool developers should align with established standards  
140 or community practices. We can encourage practices that promote ICON principles, such as  
141 **Open** publication of study plans (e.g., [PLOS ONE study proposals](#)), data production and  
142 processing protocols (e.g., [Common Workflow Language](#)), and software code. We must  
143 continually evaluate how to **Coordinate** and **Integrate** across existing cyberinfrastructure from  
144 local to global scales, which involves iterative rounds of engagement; education and outreach;  
145 and feedback across data providers, tool and service creators, and scientists who use ESS data  
146 and services. **Coordinating Networks** across disciplines will involve technical approaches to  
147 connect related data (e.g., PIDs, APIs, ontologies, geospatial standards) and promoting  
148 widespread adoption of community standards that improve scientific outcomes and benefit all  
149 participants in the network.

## 150 3 Machine learning and multiscale modeling

### 151 3.1 Success and current status of AI/ML

152 Over the past decade, artificial intelligence approaches, including machine learning  
153 (AI/ML), have revolutionized scientific discovery across disciplines, including ESSI (Maskey et  
154 al., 2020a). The AI/ML revolution, driven by a wealth of **Open** data and rapid technological  
155 development in computational cyberinfrastructure, has led to more processing power and greater  
156 **Networking** which allows unprecedented resource and data sharing. There are many success  
157 stories demonstrating how AI/ML has been used to address challenging issues in ESS, e.g.,  
158 extreme weather prediction (Maskey et al., 2020b; Pradhan et al., 2018; Wimmers et al., 2019),  
159 land use/land cover change monitoring (Hansen et al., 2013), earth system modeling (Reichstein  
160 et al., 2019), endangered species identification (Allen et al., 2021), spatial downscaling of  
161 climate models and satellite observations (López López et al., 2018; Vandal et al., 2019), space  
162 weather forecasting (Wintoft et al., 2017), and lunar and planetary landform classification  
163 (Palafox et al., 2017; Silburt et al., 2019). Various funding agencies worldwide have recently

164 released their strategic plans and guidelines to expand the investment in AI/ML research which  
165 will further its adoption within ESSI for at least the next decade.

### 166 3.2 Common challenges in AI/ML

167 To accelerate this adoption, the ESS community needs to collectively address three key  
168 challenges. First, most AI/ML applications in ESS are *ad hoc* research that lacks system-wide  
169 **Coordination** and is time-consuming. There are little AI-ready data (e.g., cleaned, harmonized,  
170 formatted, well understood) that can be efficiently **Integrated** across domains or applications  
171 and few recommended practices on proper model development and documentation (Maskey et  
172 al., 2020a). Thus, amplifying the value of AI/ML in ESS requires an ecosystem including AI-  
173 ready training datasets and standardized model development practices. This ecosystem would  
174 enable the ESS community to collaboratively develop open AI/ML applications at scale. A  
175 second challenge is related to the wealth of **Open** data in ESS. Currently, there are no  
176 community-recommended practices on how to properly develop, document, and share the AI/ML  
177 applications that track provenance and enable reproducibility (Sun et al., 2020). Third, the  
178 explainability and generalizability of AI/ML models are also major concerns for the ESS  
179 community (McGovern et al., 2019; Toms et al., 2020). To address complex questions in ESS  
180 systems, we need to better understand why AI/ML models perform in a certain way, their  
181 consistency with domain knowledge, and how models developed using a specific set of data can  
182 adjust dynamically to shifts in ESS data. Additionally, ethical awareness, conduct, and  
183 responsibility in AI/ML and related activities are essential to the practice of principled research.

### 184 3.3 Opportunities and moving forward

185 We identify five opportunities where researchers may focus their efforts to make ESS  
186 AI/ML more efficient. One opportunity relates to big data in ESS. Because the capacity and  
187 application scope of AI/ML heavily depends on patterns in training data, it should be as  
188 representative as possible. The requirements for big training datasets have led to calls for  
189 libraries of **Open** and FAIR benchmark datasets ([WILDS](#), Koh et al., 2020; [Radiant Earth](#)  
190 [Foundation](#); Rasp et al., 2020) related to questions within ESS (Crystal-Ornelas et al., 2021). A  
191 second opportunity is increased **Networking** through cloud computing (Gorelick et al., 2017;  
192 Mayer-Schönberger & Cukier, 2013). By sharing data and models in the cloud, researchers  
193 around the world can access these resources without being limited by local computing power.  
194 More work needs to be done to make cloud computing more accessible for ESS despite recent  
195 progress. Increased **Openness** in the exchange of data handling practices to allow sharing  
196 common workflows while handling large datasets is a third opportunity. A fourth opportunity is  
197 to improve interpretability through **Integration** across disciplines by: (1) including physics in  
198 ML models (Jia et al., 2019; Raissi et al., 2019), (2) leveraging machine learning exploratory  
199 tools (Montavon et al., 2017; Ying et al., 2019), and (3) involving domain experts into AI/ML  
200 pipelines. A final opportunity for growth is to automate workflows to improve the development  
201 efficiency (e.g., auto-sklearn, AutoKeras) (He et al., 2021). To improve AI engineering  
202 efficiency and reduce data collection and processing costs, modelers may also use data  
203 augmentation methods such as mixup (Zhang et al., 2017) to fill in the missing data and enhance  
204 data quality (Alexandrov & Vesselinov, 2014; Vesselinov et al., 2018). We emphasize that these  
205 opportunities for ESS to inform and apply AI/ML models is not exhaustive; rather it is a starting  
206 point for exploring how ICON science can benefit the future of this rapidly growing field within  
207 ESS.

## 208 **4 Remote sensing for advancing Earth system model development by integrating field and** 209 **ancillary data**

### 210 4.1 Current Status

211 Remote sensing technology combined with field and ancillary data (e.g., field  
212 measurements, other imagery; Acton, 1996) has transformed the development of ESMs as they  
213 have advanced from aerial imagery of the early nineteenth century (Necsoiu et al., 2013) to the  
214 present-day's Google Earth Engine (Gorelick et al., 2017) and Unmanned Aerial Vehicles (Singh  
215 & Frazier, 2018). Most publicly-funded remote sensing datasets are **Open** and hosted on public  
216 repositories (e.g., government-sponsored repositories, Github, Zenodo). In addition, this data is  
217 collected through **Coordinated** standards between government agencies across the globe  
218 (Alameh, 2020). **Integration** of remote sensing technology with independent field measurements  
219 and high spatial resolution satellite imagery has been essential for ESM validation. This also  
220 includes estimating derived data products (e.g., from satellites) accuracy and quantifying  
221 uncertainty (Strahler et al., 2006). Crowdsourcing and citizen science have further advanced the  
222 integration of remote sensing with field data (e.g., [RaspberryShake](#), Khan et al., 2018; Saralioglu  
223 & Gungor, 2020; Worldwide Hydrobiogeochemistry Observation Network for Dynamic River  
224 Systems [[WHONDRS](#)], Stegen & Goldman, 2018), resulting in broader **Networked** efforts that  
225 benefit researchers and a wide variety of data users. We note that agencies in the US and Europe  
226 have open-sourced their data to all users internationally. Some popular open data sources,  
227 associated cyberinfrastructure, and tools are included in [an associated github repository](#).

### 228 4.2 Challenges and call to action

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

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

242 There has been a **Coordinated** effort from US and European agencies to develop  
243 cyberinfrastructure that improves and increases access to data to enhance predictions and  
244 understanding of various Earth system processes. For example, the European Space Agency  
245 Sentinel data products are recently available in the [Copernicus Data and Information Access](#)  
246 [Service](#) cloud environments. In addition, the US Geological Survey Landsat satellite data  
247 inventory has been open to the public since 2008 and has been in the cloud since 2020 (U.S.  
248 Geological Survey, 2008). Furthermore, the National Aeronautics and Space Administration  
249 (NASA) and the National Oceanic and Atmospheric Administration (NOAA) have adopted a  
250 strategic vision to leverage cloud computing and operate multiple components of their data

251 systems in a retail cloud environment. This calls for action to identify the opportunities to  
252 improve policy and strategy planning across various countries to make satellite data products  
253 accessible to all users in open data portals. In addition, automated quality assurance of satellite  
254 observations is needed to support global, regional, or local data services. **Coordinated** across  
255 international agencies, a standard open data cyberinfrastructure will help to assure ESM data  
256 from multiple sources (national, regional, governments, academia, and the private sector) are  
257 available and easily **Integrated** into open-source platforms and networks.

#### 258 4.3 Opportunities and moving forward

259 First, close coordination would help international agencies and organizations build a  
260 standard open data cyberinfrastructure to ensure that earth science data are free, open, and easily  
261 integrated into ESMs. Second, we need next-generation sensors and satellites which provide  
262 more fine resolution data to increase the accuracy of ESMs. For example, the joint NASA-Indian  
263 Space Research Organization (ISRO) Synthetic Aperture Radar (SAR) ([NISAR](#)) mission is  
264 anticipated to provide fine-scale resolution radar data with a spatial resolution of less than a  
265 centimeter to study the earth's features and processes. Third, the role of AI/ML needs to be  
266 expanded to plug in the gaps of remote sensing data.

#### 267 5 Concluding remarks

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

#### 280 Acknowledgments

281 DJH, JED, NC, CC, WDD, ZL, RJR, BPW, and LW authored section 2 'Global  
282 collaboration, cyberinfrastructure, and data sharing.' RCO, SC, BA, CL, YR, TYC, and ZS  
283 authored section 3 'Machine learning and multiscale modeling.' LNG, MKM, and TY authored  
284 section 4 'Remote sensing for advancing Earth system model development by integrating field  
285 and ancillary data.'

286 Sky Bristol (USGS) was instrumental in early discussions, particularly of cost-benefit  
287 analysis.

288 JED and RCO were funded by the ESS-DIVE repository and WDD by the National  
289 Microbiome Data Collaborative, both by the U.S. DOE's Office of Science Biological and  
290 Environmental Research under contract number DE-AC02-05CH11231. NC was supported by  
291 NEON, a program sponsored by the NSF and operated under cooperative agreement by Battelle

292 Memorial Institute. SC was supported by an appointment to the NASA Postdoctoral Program at  
293 NASA Goddard Space Flight Center, administered by Universities Space Research Association  
294 under contract with NASA. CL was supported by an appointment as a postdoctoral fellow at the  
295 Cornell Atkinson Center for Sustainability, and an affiliation with the Department of Information  
296 Science. ZL was supported by Cooperative Geoinformation Research with the NASA GSFC  
297 Earth Sciences Data and Information Services Center (GES DISC). MKM was supported by the  
298 U.S. DOE-SC, SFA at PNNL. YR was supported by NOAA through the Cooperative Institute for  
299 Satellite Earth System Studies under Cooperative Agreement NA19NES4320002. BPW was  
300 supported by the Ministry of Business Innovation and Employment (MBIE) Infrastructure  
301 Platform. TY was supported by [NASA Applied Sciences Disasters Program](#) and NASA's  
302 [LANCE](#) system, part of NASA's EOSDIS.

303 The views and opinions of authors expressed herein do not necessarily state or reflect  
304 those of the US Government or any international agency thereof.

### 305 **References**

- 306 Acton, C. H. (1996). Ancillary data services of NASA's Navigation and Ancillary Information  
307 Facility. *Planetary and Space Science*, 44(1), 65–70. [https://doi.org/10.1016/0032-](https://doi.org/10.1016/0032-0633(95)00107-7)  
308 [0633\(95\)00107-7](https://doi.org/10.1016/0032-0633(95)00107-7)
- 309 Alameh, N. (2020). A future of location data integration. *Geo: GeoConnexion International*  
310 *Magazine*, 19(6), 18–19. Retrieved from [https://www.geoconnexion.com/publication-articles/a-](https://www.geoconnexion.com/publication-articles/a-future-of-location-data-integration)  
311 [future-of-location-data-integration](https://www.geoconnexion.com/publication-articles/a-future-of-location-data-integration)
- 312 Alexandrov, B. S., & Vesselinov, V. V. (2014). Blind source separation for groundwater  
313 pressure analysis based on nonnegative matrix factorization. *Water Resources Research*, 50(9),  
314 7332–7347. <https://doi.org/10.1002/2013wr015037>
- 315 Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., et al. (2021). A  
316 Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse,  
317 Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, 8, 165.  
318 <https://doi.org/10.3389/fmars.2021.607321>
- 319 Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al.  
320 (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of  
321 metadata. *Nucleic Acids Research*, 40(D1), D57–D63. <https://doi.org/10.1093/nar/gkr1163>

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

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

328 Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A., Stegen, J. C., & Fox,  
329 P. (2021). Special collection on open collaboration across geosciences. *Eos*, 102.  
330 <https://doi.org/10.1029/2021EO153180>

331 Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google  
332 Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*,  
333 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>

334 Gries, C., Servilla, M., O’Brien, M., Vanderbilt, K., Smith, C., Costa, D., & Grossman-Clarke, S.  
335 (2019). Achieving FAIR Data Principles at the Environmental Data Initiative, the US-LTER  
336 Data Repository. *Biodiversity Information Science and Standards*, 3, e37047.  
337 <https://doi.org/10.3897/biss.3.37047>

338 Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al.  
339 (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160),  
340 850–853. <https://doi.org/10.1126/science.1244693>

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

343 Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics  
344 Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature

- 345 Profiles. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)* (pp.  
346 558–566). Society for Industrial and Applied Mathematics.  
347 <https://doi.org/10.1137/1.9781611975673.63>
- 348 Jones, M. B., Slaughter, P., & Habermann, T. (2019). *Quantifying FAIR: automated metadata*  
349 *improvement and guidance in the DataONE repository network*.  
350 <https://doi.org/10.5281/zenodo.3408466>
- 351 Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G.  
352 (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *Bioscience*, 59(7),  
353 613–620. <https://doi.org/10.1525/bio.2009.59.7.12>
- 354 Khan, A., Denton, P., Stevenson, J., & Bossu, R. (2018). Engaging citizen seismologists  
355 worldwide. *Astronomy & Geophysics*, 59(4), 4.15–4.18. <https://doi.org/10.1093/astrogeo/aty190>
- 356 Kinkade, D., & Shepherd, A. (2021). Geoscience data publication: Practices and perspectives on  
357 enabling the FAIR guiding principles. *Geoscience Data Journal*, (gdj3.120).  
358 <https://doi.org/10.1002/gdj3.120>
- 359 Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2020).  
360 WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv [cs.LG]*. Retrieved from  
361 <https://arxiv.org/abs/2012.07421>
- 362 López López, P., Immerzeel, W. W., Rodríguez Sandoval, E. A., Sterk, G., & Schellekens, J.  
363 (2018). Spatial Downscaling of Satellite-Based Precipitation and Its Impact on Discharge  
364 Simulations in the Magdalena River Basin in Colombia. *Frontiers of Earth Science in China*, 6,  
365 68. <https://doi.org/10.3389/feart.2018.00068>
- 366 Maskey, M., Alemohammad, H., Murphy, K. J., & Ramachandran, R. (2020a). Advancing AI for  
367 Earth Science: A data systems perspective. *Eos*, 101. <https://doi.org/10.1029/2020EO151245>

- 368 Maskey, M., Ramachandran, R., Ramasubramanian, M., Gurung, I., Freitag, B., Kaulfus, A., et  
369 al. (2020b). Deepti: Deep-Learning-Based Tropical Cyclone Intensity Estimation System. *IEEE*  
370 *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.  
371 <https://doi.org/10.1109/jstars.2020.3011907>
- 372 Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how*  
373 *We Live, Work, and Think*. Houghton Mifflin Harcourt. Retrieved from  
374 <https://play.google.com/store/books/details?id=uy4lh-WEhhIC>
- 375 McGovern, A., Lagerquist, R., Gagne, D. J., Eli Jergensen, G., Elmore, K. L., Homeyer, C. R., &  
376 Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical  
377 Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11),  
378 2175–2199. <https://doi.org/10.1175/BAMS-D-18-0195.1>
- 379 Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, 29, 33–44.  
380 <https://doi.org/10.1016/j.ecoinf.2015.06.010>
- 381 Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, 578(7796),  
382 491–491. <https://doi.org/10.1038/d41586-020-00505-7>
- 383 Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M.  
384 D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the  
385 European Open Science Cloud. *Information Services & Use*, 37(1), 49–56.  
386 <https://doi.org/10.3233/isu-170824>
- 387 Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining  
388 nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65,  
389 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>

- 390 Necsoiu, M., Dinwiddie, C. L., Walter, G. R., Larsen, A., & Stothoff, S. A. (2013). Multi-  
391 temporal image analysis of historical aerial photographs and recent satellite imagery reveals  
392 evolution of water body surface area and polygonal terrain morphology in Kobuk Valley  
393 National Park, Alaska. *Environmental Research Letters*, 8(2), 025007.  
394 <https://doi.org/10.1088/1748-9326/8/2/025007>
- 395 Novick, K. A., Biederman, J. A., Desai, A. R., Litvak, M. E., Moore, D. J. P., Scott, R. L., &  
396 Torn, M. S. (2018). The AmeriFlux network: A coalition of the willing. *Agricultural and Forest*  
397 *Meteorology*, 249, 444–456. <https://doi.org/10.1016/j.agrformet.2017.10.009>
- 398 Palafox, L. F., Hamilton, C. W., Scheidt, S. P., & Alvarez, A. M. (2017). Automated detection of  
399 geological landforms on Mars using Convolutional Neural Networks. *Computers & Geosciences*,  
400 101, 48–56. <https://doi.org/10.1016/j.cageo.2016.12.015>
- 401 Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2018). Tropical  
402 Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Transactions*  
403 *on Image Processing*. <https://doi.org/10.1109/tip.2017.2766358>
- 404 Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A  
405 deep learning framework for solving forward and inverse problems involving nonlinear partial  
406 differential equations. *Journal of Computational Physics*, 378, 686–707.  
407 <https://doi.org/10.1016/j.jcp.2018.10.045>
- 408 Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in  
409 climate models. *Proceedings of the National Academy of Sciences of the United States of*  
410 *America*, 115(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115>

- 411 Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020).  
412 WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances*  
413 *in Modeling Earth Systems*, 12(11). <https://doi.org/10.1029/2020ms002203>
- 414 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat.  
415 (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*,  
416 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- 417 Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The  
418 GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the  
419 internet. *PloS One*, 9(8), e102623. <https://doi.org/10.1371/journal.pone.0102623>
- 420 Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., et al.  
421 (2019). Decline of the North American avifauna. *Science*, eaaw1313.  
422 <https://doi.org/10.1126/science.aaw1313>
- 423 Saralioglu, E., & Gungor, O. (2020). Crowdsourcing in Remote Sensing: A Review of  
424 Applications and Future Directions. *IEEE Geoscience and Remote Sensing Magazine*, 8(4), 89–  
425 110. <https://doi.org/10.1109/MGRS.2020.2975132>
- 426 Silburt, A., Ali-Dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., et al. (2019). Lunar crater  
427 identification via deep learning. *Icarus*, 317, 27–38. <https://doi.org/10.1016/j.icarus.2018.06.022>
- 428 Singh, K. K., & Frazier, A. E. (2018). A meta-analysis and review of unmanned aircraft system  
429 (UAS) imagery for terrestrial applications. *International Journal of Remote Sensing*, 39(15-16),  
430 5078–5098. <https://doi.org/10.1080/01431161.2017.1420941>
- 431 Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al.  
432 (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment*  
433 *of global land cover maps* (Publication EUR 22156 EN). European Commission, Joint Research

434 Center. Retrieved from [https://op.europa.eu/en/publication-detail/-/publication/52730469-6bc9-](https://op.europa.eu/en/publication-detail/-/publication/52730469-6bc9-47a9-b486-5e2662629976)  
435 [47a9-b486-5e2662629976](https://op.europa.eu/en/publication-detail/-/publication/52730469-6bc9-47a9-b486-5e2662629976)

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

438 Stegen, J. C., & Goldman, A. E. (2018). WHONDRS: a Community Resource for Studying  
439 Dynamic River Corridors. *mSystems*, 3(5), e00151–18. [https://doi.org/10.1128/mSystems.00151-](https://doi.org/10.1128/mSystems.00151-18)  
440 [18](https://doi.org/10.1128/mSystems.00151-18)

441 Sun, Z., Di, L., Burgess, A., Tullis, J. A., & Magill, A. B. (2020). Geoweaver: Advanced  
442 Cyberinfrastructure for Managing Hybrid Geoscientific AI Workflows. *ISPRS International*  
443 *Journal of Geo-Information*, 9(2), 119. <https://doi.org/10.3390/ijgi9020119>

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

447 U.S. Geological Survey. (2008). *Imagery for Everyone: Timeline Set to Release Entire USGS*  
448 *Landsat Archive at No Charge*. Retrieved from [https://prd-wret.s3.us-west-](https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGSStechann-20080421-landsat-imagery-release.pdf)  
449 [2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGSStechann-20080421-](https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGSStechann-20080421-landsat-imagery-release.pdf)  
450 [landsat-imagery-release.pdf](https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGSStechann-20080421-landsat-imagery-release.pdf)

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

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

- 457 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012).  
458 Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PloS One*, 7(1),  
459 e29715. <https://doi.org/10.1371/journal.pone.0029715>
- 460 Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al.  
461 (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific*  
462 *Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- 463 Wimmers, A., Velden, C., & Cossuth, J. H. (2019). Using Deep Learning to Estimate Tropical  
464 Cyclone Intensity from Satellite Passive Microwave Imagery. *Monthly Weather Review*, 147(6),  
465 2261–2282. <https://doi.org/10.1175/MWR-D-18-0391.1>
- 466 Wintoft, P., Wik, M., Matzka, J., & Shprits, Y. (2017). Forecasting Kp from solar wind data:  
467 input parameter study using 3-hour averages and 3-hour range values. *Journal of Space Weather*  
468 *and Space Climate*, 7, A29. <https://doi.org/10.1051/swsc/2017027>
- 469 Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011).  
470 Minimum information about a marker gene sequence (MIMARKS) and minimum information  
471 about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29(5), 415–420.  
472 <https://doi.org/10.1038/nbt.1823>
- 473 Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating  
474 Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*,  
475 32, 9240–9251. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/32265580>
- 476 Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond Empirical Risk  
477 Minimization. *arXiv [cs.LG]*. Retrieved from <http://arxiv.org/abs/1710.09412>