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

Denise Hills<sup>1,1,1,1,1</sup>, Joan Damerow<sup>2,2,2,2,2</sup>, Bulbul Ahmmed<sup>3,3,3,3,3</sup>, Nora Catolico<sup>4,4,4,4,4</sup>, Srija Chakraborty<sup>5,5,5,5,5</sup>, Caroline Coward<sup>6,6,6,6,6</sup>, Robert Crystal-Ornelas<sup>2,2,2,2,2</sup>, William Duncan<sup>7,7,7,7,7</sup>, Laxmi Goparaju<sup>8,8,8,8,8</sup>, Cindy Lin<sup>9,9,9,9,9</sup>, Zhong Liu<sup>10,10,10,10,10</sup>, Maruti Mudunuru<sup>11,11,11,11,11</sup>, Yuhan Rao<sup>12,12,12,12,12</sup>, Robert Rovetto<sup>13,13,13,13,13</sup>, Ziheng Sun<sup>10,10,10,10,10</sup>, Brandon Whitehead<sup>14,14,14,14</sup>, Lesley Wyborn<sup>15,15,15,15,15</sup>, and Tian Yao<sup>16,16,16,16</sup>

<sup>1</sup>Geological Survey of Alabama <sup>2</sup>Lawrence Berkelev National Laboratory <sup>3</sup>Los Alamos National Laboratory <sup>4</sup>Battelle, National Ecological Observatory Network <sup>5</sup>NASA's Goddard Space Flight Center <sup>6</sup>NASA Jet Propulsion Laboratory <sup>7</sup>Lawrence Berkelev National Laboratory, National Center for Ontological Research <sup>8</sup>Vindhyan Ecology and Natural History Foundation, U.P. India <sup>9</sup>Cornell University <sup>10</sup>George Mason University <sup>11</sup>Pacific Northwest National Laboratory <sup>12</sup>North Carolina Institute for Climate Studies <sup>13</sup>Center for Orbital Debris Education & Research, University of Maryland <sup>14</sup>Manaaki Whenua - Landcare Research, Minnesota State University Mankato, Minnesota State University Mankato <sup>15</sup>Australian National University <sup>16</sup>Science Systems and Applications, Inc

November 30, 2022

#### Abstract

This article is composed of three independent commentaries about the state of ICON principles (Goldman et al., 2021) 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 and multiscale modeling; (Section 4) 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 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.

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

3 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.

- 4 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>6,10</sup>, M. K. Mudunuru<sup>11</sup>, Y.
- 5 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>
- 6 <sup>1</sup><u>Geological Survey of Alabama</u>, Tuscaloosa, AL, USA.
- 7 <sup>2</sup><u>Ronin Institute for Independent Scholarship</u>, Tuscaloosa, AL, USA.
- 8 <sup>3</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
- 9 <sup>4</sup>Los Alamos National Laboratory, Los Alamos, NM, USA.
- 10 <sup>5</sup>Battelle, <u>National Ecological Observatory Network</u>, CO, USA.
- 11 <sup>6</sup>NASA's Goddard Space Flight Center, Greenbelt, MD, USA.
- 12 <sup>7</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA.
- 13 <sup>8</sup>Vindhyan Ecology and Natural History Foundation, U.P. India.
- <sup>9</sup>Atkinson Center for Sustainability and Department of Information Science, <u>Cornell University</u>,
- 15 Ithaca, NY, USA.
- 16 <sup>10</sup>George Mason University, Fairfax, VA, USA.
- 17 <sup>11</sup>Pacific Northwest National Laboratory, Richland, WA, USA.
- 18 <sup>12</sup>North Carolina Institute for Climate Studies, <u>North Carolina State University</u>, Asheville, NC,
- 19 USA.
- 20 <sup>13</sup>Center for Orbital Debris Education & Research, <u>University of Maryland</u>, MD, USA.
- 21 <sup>14</sup>Independent, New York, NY, USA.
- 22 <sup>15</sup><u>Manaaki Whenua Landcare Research</u>, Palmerston North, New Zealand
- 23 <sup>16</sup><u>Australian National University</u>, Canberra, ACT, Australia
- 24 <sup>17</sup><u>Science Systems and Applications, Inc.</u>, Lanham, MD, USA.
- 25 Corresponding authors: Denise Hills (<u>denise.j.hills@gmail.com</u>); Joan Damerow
- 26 (JoanDamerow@lbl.gov)
- 27 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.
- 34

#### 35 Abstract

- 36 This article is composed of three independent commentaries about the state of ICON principles
- 37 (Goldman et al., 2021a) in Earth and Space Science Informatics (ESSI) and includes discussion
- 38 on 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 for multiscale modeling; (Section 4) Aerial and satellite remote sensing for
- 41 advancing Earth system model development by integrating field and ancillary data. ESSI
- 42 addresses data management practices, computation and analysis, and hardware and software
- 43 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
- 46 of diverse people with expertise across Earth, space, and data science disciplines are essential for
- 47 efficient 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
- (FAIK). Ungoing support is crucial to develop and sustain the people, systems, and tools
- 66 necessary to embrace 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., 2021a). This 71 article belongs to a collection of commentaries (Goldman, et al., 2021b) spanning geoscience on 72 the state and future of ICON science. Earth and Space Science Informatics (ESSI) encompasses a 73 broad field that addresses data management practices, computation and analysis, and hardware 74 and software infrastructure. ESSI's role in ICON science therefore involves collaborative work 75 to assess, design, implement, and promote practices and tools that enable effective data 76 management, discovery, integration, and reuse for interdisciplinary work in Earth and space 77 science (ESS) disciplines. In this series of commentaries, we examine the current state,
- refere (ESS) disciplines. In this series of commentaries, we examine the current state,
   challenges, and opportunities of ICON science through the lenses of global collaboration,

- 79 <u>cyberinfrastructure</u>, 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 status

84 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, e.g., Research Data Alliance, World Data System,

have emerged to facilitate such collaboration, e.g., <u>Research Data Alliance</u>, <u>World Data System</u>,
 <u>Earth Science Information Partners</u>. These organizations have produced numerous active groups

involved in Earth, space and environmental science data and research, and developed many data

89 tools and services, e.g. Earth, Space and Environmental Sciences Data Vocabulary Repositories.

90 Research is more efficient with **Networked** data practices and cyberinfrastructures that support

91 scientific discovery. Yet, there is still a large disconnect and lack of **Coordination** across many

92 informatics communities and the broader communities we aim to support.

93 Research teams often lack sufficient resources (e.g., appropriate cyberinfrastructure,

94 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

96 Accessible, Interoperable, and Reusable (FAIR; withinson et al., 2016). Further, specific 97 criteria to implement the FAIR Guiding Principles (Gries et al., 2019; Jones et al., 2019)

97 chieffa to implement the FAIR Guiding Frinciples (Gries et al., 2019, Jones et al., 2019)
 98 inevitably vary across disciplines and data types as inconsistencies in interpretations of the

99 principles have grown (e.g., Kinkade & Shepherd, 2021; Mons et al., 2017; Stall et al., 2019).

100 Importantly, FAIR does not mean **Open**; data can be **Open** without being FAIR, and *vice versa* 

101 (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

103 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

111 (e.g., <u>Global Biodiversity Information Facility</u>, <u>Ameriflux</u>, <u>Consortium of Universities for the</u>

Advancement of Hydrologic Science, Inc., Long-Term Ecological Research Network, National
 Ecological Observatory Network, Deep Carbon Observatory, HydroShare) have been driven by

(1) demand for and funding to support a specific data type (Barrett et al., 2012; Novick et al.,

2018; Robertson et al., 2014); (2) reporting standards that enable global data search and

116 integration (e.g., Wieczorek et al., 2012; Yilmaz et al., 2011); and (3) associated user-friendly

- 117 tools (Clark et al., 2016; Robertson et al., 2014).
- 118 2.2 Challenges and opportunities
- 119 Most cyberinfrastructures lack the resources for **Integration** and **Coordination**
- 120 necessary for broader interdisciplinary work, including guidance and leading practices; domain

121 semantics; technical, data, methodological, and instrumentation standards; workflow

- 122 management; training; and sustainable technical and financial support. These deficits hinder the
- 123 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
- 126 spent developing/updating translators, or disciplinary standards are simply disconnected and 127 inefficient for interdisciplinary users. **Coordination** is needed to implement standards for
- 128 effective interdisciplinary data discovery and exchange. A major challenge to **Coordination**
- 129 involves a lack of consistent and transparent protocols (e.g., data and code production,
- 130 processing methods) across interdisciplinary teams. Further, informatics initiatives and working
- 131 groups (e.g. RDA, ESIP) are primarily volunteer-based without appropriate recognition or
- 132 funding that would accelerate and improve this work. These combined factors create barriers to
- 133 **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 siloed efforts across disciplines,
- 137 organizations, and/or countries. A commitment to community engagement is needed to bring
- 138 together input across disciplines, understand data management challenges and needs, and
- 139 promote the adoption of shared practices. Making data as **Open and/or FAIR** as ethically
- 140 possible requires key advocates who facilitate **Networked** collaboration.
- Data users, code contributors, and tool developers should align with established standards 141 142 or community practices. We can encourage practices that promote ICON principles, such as 143 **Open** publication of study plans (e.g., PLOS ONE study proposals), data production and 144 processing protocols (e.g., Common Workflow Language), and software code. We must 145 continually evaluate how to Coordinate and Integrate across existing cyberinfrastructure from 146 local to global scales, which involves iterative rounds of engagement; education and outreach; 147 and feedback across data providers, tool and service creators, and scientists who use ESS data 148 and services. Coordinating Networks across disciplines will involve technical approaches to 149 connect related data (e.g., globally unique and resolvable persistent identifiers (PIDs), APIs, 150 ontologies, geospatial standards) and promoting widespread adoption of community standards 151 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.

### 153 **3 Machine learning for multiscale modeling**

154 3.1 Current status

155 Over the past decade, artificial intelligence approaches, including machine learning

- 156 (AI/ML), have revolutionized scientific discovery across disciplines, including Earth and space
- science informatics informatics (Maskey, Alemohannad, et al., 2020). The AI/ML revolution,
- driven by a wealth of **Open** data and rapid technological development in computational
- 159 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
- 161 resource and data snaring. There are many success stories demonstrating now AI/ML has 162 used to address challenging issues in ESS, e.g., extreme weather prediction (Maskey,
- Ramachandran, et al., 2020; Pradhan et al., 2018; Wimmers et al., 2019), land use/land cover

164 change monitoring (Hansen et al., 2013), Earth system modeling (Reichstein et al., 2019),

165 endangered species identification (Allen et al., 2021), spatial downscaling of climate models and

166 satellite observations (López López et al., 2018; Vandal et al., 2019), space weather forecasting

167 (Wintoft et al., 2017), and lunar and planetary landform classification (Palafox et al., 2017;

168 Silburt et al., 2019). Various funding agencies worldwide have recently released their strategic

169 plans and guidelines to expand the investment in AI/ML research which will further its adoption 170 within informatics for at least the next decade to accelerate scientific discovery and address

170 within informatics for at least the next decade to accelerate scientific discovery and address 171 pressing societal issues such as combatting climate change, facilitating the energy transition, and

172 ensuring food security.

#### 173 3.2 Challenges and opportunities

174 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 175 176 in ESS are *ad hoc* research that lacks system-wide **Coordination** and are time-consuming. There 177 are little AI-ready data (e.g., cleaned, harmonized, formatted, well documented) that can be 178 efficiently Integrated across domains or applications and few recommended practices on proper 179 model development and documentation (Maskey, Alemohammad, et al., 2020). As the capacity 180 and application scope of AI/ML heavily depends on patterns in training data, it should be as 181 representative as possible. These requirements for big training datasets have led to calls for 182 libraries of Open and FAIR benchmark datasets (WILDS, Koh et al., 2020; Radiant Earth

183 Foundation; Rasp et al., 2020) related to questions within ESS (Crystal-Ornelas et al., 2021).

184 AI-ready training datasets and standardized AI/ML model development practices would 185 enable the ESS community to collaboratively develop open AI/ML applications at scale. 186 However, there are no current community-recommended practices on how to properly develop, 187 document, and share the AI/ML applications that track provenance and enable reproducibility 188 (Sun et al., 2020). Increased connection through cloud computing (Gorelick et al., 2017; Mayer-189 Schönberger & Cukier, 2013) allows sharing data and models in the cloud, enabling Networked 190 researchers around the world access to these resources without being limited by local computing 191 power. However, despite recent progress, work needs to be done to make cloud computing more 192 accessible. Increased **Openness** in the exchange of data handling practices allows sharing 193 common workflows while handling large datasets. Integration across disciplines could be 194 improved by: (1) including physics in ML models (Jia et al., 2019; Raissi et al., 2019), (2) 195 leveraging ML exploratory tools (Montavon et al., 2017; Ying et al., 2019), and (3) better 196 mechanism for codevelopment between domain experts and AI/ML developers. Coordination 197 via automated workflows would improve development efficiency (e.g., auto-sklearn, AutoKeras) 198 (He et al., 2021). To improve AI engineering efficiency and reduce data collection and 199 processing costs, developers may also use data augmentation methods such as mixup (Zhang et 200 al., 2017) to fill in the missing data and enhance data quality (Alexandrov & Vesselinov, 2014; 201 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

- 208 framework to enhance future development. Licensing approaches are still evolving, highlighting
- 209 the need for increased **Coordination** on policy and ethics considerations. Ethical awareness,
- 210 conduct, and responsibility in AI/ML and related activities are essential to the practice of
- 211 principled research; while beyond the scope of this paper, some particular concerns include
- 212 misleading results due to biased training data; cognitive biases in general; and incorrect
- 213 annotation, classification or characterization of data. AI/ML applications heavily rely on input 214 data, thus the ESS community needs to establish **Coordinated** standards that clarify the impact
- 214 data, thus the ESS community needs to establish **Coordinated** standards that clarify the impact 215 of input data quality on downstream applications to ensure trustworthiness. These community
- 216 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

219 4.1 Current status

220 Remote sensing technology combined with field and ancillary data (e.g., field 221 measurements, other imagery; Acton, 1996) provides a compelling example of how dedicated 222 resources supporting ICON science and advanced AI/ML technologies have transformed the 223 development of ESMs as they have advanced from aerial imagery of the early nineteenth century 224 (Necsoiu et al., 2013) to the present-day's Google Earth Engine (Gorelick et al., 2017) and 225 Unmanned Aerial Vehicles (Singh & Frazier, 2018). Most publicly-funded remote sensing 226 datasets are **Open** and hosted on public repositories (e.g., government-sponsored repositories, 227 Github, Zenodo). In addition, this data is distributed through Coordinated standards between 228 government agencies across the globe (Alameh, 2020). Integration of remote sensing 229 technology with independent field measurements and high spatial resolution satellite imagery has 230 been essential for ESM validation. This also includes estimating derived data products (e.g., 231 from satellites) accuracy and quantifying uncertainty (Strahler et al., 2006). Crowdsourcing and 232 citizen science have further advanced the integration of remote sensing with field data (e.g., 233 RaspberryShake, Khan et al., 2018; Saralioglu & Gungor, 2020; Worldwide 234 Hydrobiogeochemistry Observation Network for Dynamic River Systems [WHONDRS], Stegen 235 & Goldman, 2018), resulting in broader Networked efforts that benefit researchers and a wide 236 variety of data users. Many agencies in the US and Europe have made some or all of their data 237 **Open** to all users internationally. Some examples, associated cyberinfrastructure, and tools are

- 238 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 30TB 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.

253 There has been a **Coordinated** effort from US and European agencies to develop 254 cyberinfrastructure that improves and increases access to data to enhance predictions and 255 understanding of various Earth system processes. For example, the European Space Agency 256 Sentinel data products are recently available in the Copernicus Data and Information Access 257 Service cloud environments. In addition, the US Geological Survey Landsat satellite data 258 inventory has been Open to the public since 2008 and has been in the cloud since 2020 (U.S. 259 Geological Survey, 2008). Furthermore, the National Aeronautics and Space Administration 260 (NASA) and the National Oceanic and Atmospheric Administration (NOAA) have adopted a 261 strategic vision to leverage cloud computing and operate multiple components of their data 262 systems in a retail cloud environment. This calls for action to identify the opportunities to 263 improve policy and strategy planning across various countries to make satellite data products 264 accessible to all users in open data portals. In addition, automated quality assurance of satellite 265 observations is needed to support global, regional, or local data services. Coordinated across 266 international agencies, a standard open data cyberinfrastructure will help to assure ESM data 267 from multiple sources (national, regional, governments, academia, and the private sector) are 268 available and easily Integrated into open-source platforms and networks.

269 **Coordination** would help international agencies and organizations build a standard open 270 data cyberinfrastructure to ensure that Earth science data are free, **Open**, and easily **Integrated** 271 into ESMs. We also need next-generation sensors and satellites which provide more fine 272 resolution data to increase the accuracy of ESMs. For example, the joint NASA-Indian Space 273 Research Organization (ISRO) Synthetic Aperture Radar (SAR) (NISAR) mission is anticipated 274 to provide **Open** radar data with a spatial resolution of less than a centimeter to **Integrate** into 275 ESM for studying the Earth's features and processes. The role of AI/ML needs to be expanded to 276 fill the gaps of remote sensing data.

#### 277 5 Concluding remarks

278 Earth and space science research facilitated by modern informatics techniques that follow 279 the ICON principles enables data synthesis, increases resource efficiency, and creates knowledge 280 that transcends individual systems (Goldman et al., 2021a). ESSI can work to ensure that diverse 281 scientists have user-friendly resources to contribute and use data that follows community 282 conventions. Such collections of **Open and/or FAIR** data, shared across **Networks** for mutual 283 benefit, are critical to appropriately train AI/ML, which furthers Integration and Coordination 284 in Earth and space science informatics. Cross-community Networks improve scientific outcomes 285 for all involved. Communities must work together to share data openly using community 286 standards, to produce Open and/or FAIR data that enables data synthesis and can revolutionize 287 fields of research (e.g., Kelling et al., 2009). Ongoing, sustainable support is vital to create and 288 maintain the cyberinfrastructure and human resources necessary for Integrated, Coordinated, 289 and **Open and/or FAIR** data (as much ethically as possible) for interdisciplinary **Networks**.

- 290 Acknowledgments
- DJH, JED, NC, CC, WDD, ZL, RJR, BPW, and LW authored section 2 'Global
  collaboration, cyberinfrastructure, and data sharing.' RCO, SC, BA, CL, YR, TYC, and ZS

authored section 3 'Machine learning for multiscale modeling.' LNG, MKM, and TY authored
 section 4 'Aerial and satellite remote sensing for advancing Earth system model development by

295 integrating field and ancillary data.'

Sky Bristol (USGS) was instrumental in early discussions, particularly of cost-benefitanalysis.

298 JED and RCO were funded by the ESS-DIVE repository and WDD by the National 299 Microbiome Data Collaborative, both by the U.S. DOE's Office of Science Biological and 300 Environmental Research under contract number DE-AC02-05CH11231. NC was supported by 301 NEON, a program sponsored by the NSF and operated under cooperative agreement by Battelle 302 Memorial Institute. SC was supported by an appointment to the NASA Postdoctoral Program at 303 NASA Goddard Space Flight Center, administered by Universities Space Research Association 304 under contract with NASA. CL was supported by an appointment as a postdoctoral fellow at the 305 Cornell Atkinson Center for Sustainability, and an affiliation with the Department of Information 306 Science. ZL was supported by Cooperative Geoinformation Research with the NASA GSFC 307 Earth Sciences Data and Information Services Center (GES DISC). MKM was supported by the 308 U.S. DOE-SC, SFA at PNNL. YR was supported by NOAA through the Cooperative Institute for 309 Satellite Earth System Studies under Cooperative Agreement NA19NES4320002. BPW was 310 supported by the Ministry of Business Innovation and Employment (MBIE) Infrastructure 311 Platform. TY was supported by NASA Applied Sciences Disasters Program and NASA's 312 LANCE system, part of NASA's EOSDIS. CC's research was carried out at the Jet Propulsion 313 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.
- 317 No data was used for this commentary.
- 318

#### 319 References

- 320 Acton, C. H. (1996). Ancillary data services of NASA's Navigation and Ancillary Information
- 321 Facility. Planetary and Space Science, 44(1), 65–70. https://doi.org/10.1016/0032-
- 322 <u>0633(95)00107-7</u>
- 323 Alameh, N. (2020). A future of location data integration. Geo: GeoConnexion International
- 324 Magazine, 19(6), 18–19. Retrieved from https://www.geoconnexion.com/publication-articles/a-
- 325 <u>future-of-location-data-integration</u>

- 326 Alexandrov, B. S., & Vesselinov, V. V. (2014). Blind source separation for groundwater
- 327 pressure analysis based on nonnegative matrix factorization. Water Resources Research, 50(9),
- 328 7332–7347. <u>https://doi.org/10.1002/2013wr015037</u>
- 329 Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., et al. (2021). A
- 330 Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse,
- 331 Long-Term Passive Acoustic Dataset. Frontiers in Marine Science, 8, 165.
- 332 <u>https://doi.org/10.3389/fmars.2021.607321</u>
- 333 Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al.
- 334 (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of
- 335 metadata. Nucleic Acids Research, 40(D1), D57–D63. <u>https://doi.org/10.1093/nar/gkr1163</u>
- 336 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank.
- 337 Nucleic Acids Research, 44(D1), D67–D72. <u>https://doi.org/10.1093/nar/gkv1276</u>
- 338 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
- 340 enable model training, parameterization and validation. Office of Scientific and Technical
- 341 Information (OSTI). <u>https://doi.org/10.2172/1769646</u>
- 342 Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A., Stegen, J. C., & Fox,
- 343 P. (2021a). Special collection on open collaboration across geosciences. *Eos*, 102.
- 344 <u>https://doi.org/10.1029/2021EO153180</u>
- 345 Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A., & Stegen, J. C.
- 346 (2021b). Integrated, Coordinated, Open, and Networked (ICON) science to advance the
- 347 geosciences: Introduction and synthesis of a special collection of commentary articles. *Earth and*
- 348 Space Science Open Archive. <u>https://doi.org/10.1002/essoar.10508554.1</u>

- 349 Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google
- 350 Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*,
- 351 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031
- 352 Gries, C., Servilla, M., O'Brien, M., Vanderbilt, K., Smith, C., Costa, D., & Grossman-Clarke, S.
- 353 (2019). Achieving FAIR Data Principles at the Environmental Data Initiative, the US-LTER
- 354 Data Repository. *Biodiversity Information Science and Standards*, *3*, e37047.
- 355 <u>https://doi.org/10.3897/biss.3.37047</u>
- 356 Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al.
- 357 (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160),
- 358 850–853. <u>https://doi.org/10.1126/science.1244693</u>
- 359 He, X., Zhao, K., & Chu, X. (2021). AutoML: A Survey of the state-of-the-art. Knowledge-
- 360 Based Systems, 212, 106622. <u>https://doi.org/10.1016/j.knosys.2020.106622</u>
- 361 Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics
- 362 Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature
- 363 Profiles. In Proceedings of the 2019 SIAM International Conference on Data Mining (SDM) (pp.
- 364 558–566). Society for Industrial and Applied Mathematics.
- 365 <u>https://doi.org/10.1137/1.9781611975673.63</u>
- Jones, M. B., Slaughter, P., & Habermann, T. (2019). Quantifying FAIR: automated metadata
- *improvement and guidance in the DataONE repository network.*
- 368 <u>https://doi.org/10.5281/zenodo.3408466</u>
- 369 Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G.
- 370 (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *Bioscience*, 59(7),
- 371 613–620. <u>https://doi.org/10.1525/bio.2009.59.7.12</u>

- 372 Khan, A., Denton, P., Stevenson, J., & Bossu, R. (2018). Engaging citizen seismologists
- 373 worldwide. Astronomy & Geophysics, 59(4), 4.15–4.18. <u>https://doi.org/10.1093/astrogeo/aty190</u>
- 374 Kinkade, D., & Shepherd, A. (2021). Geoscience data publication: Practices and perspectives on
- 375 enabling the FAIR guiding principles. *Geoscience Data Journal*, (gdj3.120).
- 376 <u>https://doi.org/10.1002/gdj3.120</u>
- 377 Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2020).
- 378 WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv [cs.LG]. Retrieved from
- 379 <u>https://arxiv.org/abs/2012.07421</u>
- 380 López López, P., Immerzeel, W. W., Rodríguez Sandoval, E. A., Sterk, G., & Schellekens, J.
- 381 (2018). Spatial Downscaling of Satellite-Based Precipitation and Its Impact on Discharge
- 382 Simulations in the Magdalena River Basin in Colombia. Frontiers of Earth Science in China, 6,
- 383 68. <u>https://doi.org/10.3389/feart.2018.00068</u>
- 384 Maskey, M., Alemohammad, H., Murphy, K. J., & Ramachandran, R. (2020). Advancing AI for
- 385 Earth Science: A data systems perspective. *Eos*, *101*. <u>https://doi.org/10.1029/2020EO151245</u>
- 386 Maskey, M., Ramachandran, R., Ramasubramanian, M., Gurung, I., Freitag, B., Kaulfus, A., et
- al. (2020). Deepti: Deep-Learning-Based Tropical Cyclone Intensity Estimation System. IEEE
- *Journal of Selected Topics in Applied Earth Observations and Remote Sensing.*
- 389 <u>https://doi.org/10.1109/jstars.2020.3011907</u>
- 390 Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution that Will Transform how
- 391 We Live, Work, and Think. Houghton Mifflin Harcourt. Retrieved from
- 392 <u>https://play.google.com/store/books/details?id=uy4lh-WEhhIC</u>
- 393 McGovern, A., Lagerquist, R., Gagne, D. J., Eli Jergensen, G., Elmore, K. L., Homeyer, C. R., &
- 394 Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical

- 395 Implications of Machine Learning. Bulletin of the American Meteorological Society, 100(11),
- 396 2175–2199. <u>https://doi.org/10.1175/BAMS-D-18-0195.1</u>
- 397 Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, 29, 33–44.
- 398 <u>https://doi.org/10.1016/j.ecoinf.2015.06.010</u>
- 399 Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, 578(7796),
- 400 491–491. <u>https://doi.org/10.1038/d41586-020-00505-7</u>
- 401 Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M.
- 402 D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the
- 403 European Open Science Cloud. Information Services & Use, 37(1), 49–56.
- 404 <u>https://doi.org/10.3233/isu-170824</u>
- 405 Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining
- 406 nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65,
- 407 211–222. <u>https://doi.org/10.1016/j.patcog.2016.11.008</u>
- 408 Necsoiu, M., Dinwiddie, C. L., Walter, G. R., Larsen, A., & Stothoff, S. A. (2013). Multi-
- 409 temporal image analysis of historical aerial photographs and recent satellite imagery reveals
- 410 evolution of water body surface area and polygonal terrain morphology in Kobuk Valley
- 411 National Park, Alaska. *Environmental Research Letters*, 8(2), 025007.
- 412 https://doi.org/10.1088/1748-9326/8/2/025007
- 413 Novick, K. A., Biederman, J. A., Desai, A. R., Litvak, M. E., Moore, D. J. P., Scott, R. L., &
- 414 Torn, M. S. (2018). The AmeriFlux network: A coalition of the willing. Agricultural and Forest
- 415 *Meteorology*, 249, 444–456. <u>https://doi.org/10.1016/j.agrformet.2017.10.009</u>

- 416 Palafox, L. F., Hamilton, C. W., Scheidt, S. P., & Alvarez, A. M. (2017). Automated detection of
- 417 geological landforms on Mars using Convolutional Neural Networks. Computers & Geosciences,
- 418 *101*, 48–56. <u>https://doi.org/10.1016/j.cageo.2016.12.015</u>
- 419 Petzold, A., Asmi, A., Vermeulen, A., Pappalardo, G., Bailo, D., Schaap, D., et al. (2019).
- 420 ENVRI-FAIR Interoperable Environmental FAIR Data and Services for Society, Innovation
- 421 and Research. 2019 15th International Conference on eScience (eScience), 277-280.
- 422 <u>https://doi.org/10.1109/eScience.2019.00038</u>
- 423 Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2018). Tropical
- 424 Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Transactions*
- 425 on Image Processing. <u>https://doi.org/10.1109/tip.2017.2766358</u>
- 426 Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A
- 427 deep learning framework for solving forward and inverse problems involving nonlinear partial
- 428 differential equations. *Journal of Computational Physics*, 378, 686–707.
- 429 <u>https://doi.org/10.1016/j.jcp.2018.10.045</u>
- 430 Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in
- 431 climate models. Proceedings of the National Academy of Sciences of the United States of
- 432 *America*, *115*(39), 9684–9689. <u>https://doi.org/10.1073/pnas.1810286115</u>
- 433 Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020).
- 434 WeatherBench: A benchmark data set for data-driven weather forecasting. Journal of Advances
- 435 *in Modeling Earth Systems*, *12*(11). <u>https://doi.org/10.1029/2020ms002203</u>
- 436 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat.
- 437 (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*,
- 438 566(7743), 195–204. <u>https://doi.org/10.1038/s41586-019-0912-1</u>

- 439 Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The
- 440 GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the
- 441 internet. *PloS One*, *9*(8), e102623. <u>https://doi.org/10.1371/journal.pone.0102623</u>
- 442 Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., et al.
- 443 (2019). Decline of the North American avifauna. *Science*, eaaw1313.
- 444 https://doi.org/10.1126/science.aaw1313
- 445 Saralioglu, E., & Gungor, O. (2020). Crowdsourcing in Remote Sensing: A Review of
- 446 Applications and Future Directions. *IEEE Geoscience and Remote Sensing Magazine*, 8(4), 89–
- 447 110. <u>https://doi.org/10.1109/MGRS.2020.2975132</u>
- 448 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
- 450 Singh, K. K., & Frazier, A. E. (2018). A meta-analysis and review of unmanned aircraft system
- 451 (UAS) imagery for terrestrial applications. International Journal of Remote Sensing, 39(15-16),
- 452 5078–5098. <u>https://doi.org/10.1080/01431161.2017.1420941</u>
- 453 Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al.
- 454 (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment
- 455 of global land cover maps (Publication EUR 22156 EN). European Commission, Joint Research
- 456 Center. Retrieved from <u>https://op.europa.eu/en/publication-detail/-/publication/52730469-6bc9-</u>
- 457 <u>47a9-b486-5e2662629976</u>
- 458 Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., et al. (2019).
- 459 Make scientific data FAIR. *Nature*, 570(7759), 27. <u>https://doi.org/10.1038/d41586-019-01720-7</u>

- 460 Stegen, J. C., & Goldman, A. E. (2018). WHONDRS: a Community Resource for Studying
- 461 Dynamic River Corridors. *mSystems*, 3(5), e00151–18. <u>https://doi.org/10.1128/mSystems.00151-</u>
- 462 <u>18</u>
- 463 Sun, Z., Di, L., Burgess, A., Tullis, J. A., & Magill, A. B. (2020). Geoweaver: Advanced
- 464 Cyberinfrastructure for Managing Hybrid Geoscientific AI Workflows. ISPRS International
- 465 Journal of Geo-Information, 9(2), 119. <u>https://doi.org/10.3390/ijgi9020119</u>
- 466 Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks
- 467 for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling*
- 468 *Earth Systems*, *12*(9). <u>https://doi.org/10.1029/2019ms002002</u>
- 469 U.S. Geological Survey. (2008). Imagery for Everyone: Timeline Set to Release Entire USGS
- 470 *Landsat Archive at No Charge*. Retrieved from <u>https://prd-wret.s3.us-west-</u>
- 471 <u>2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGStechann-20080421-</u>
- 472 <u>landsat-imagery-release.pdf</u>
- 473 Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods
- 474 for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied*
- 475 *Climatology*, *137*(1-2), 557–570. <u>https://doi.org/10.1007/s00704-018-2613-3</u>
- 476 Vesselinov, V. V., Alexandrov, B. S., & O'Malley, D. (2018). Contaminant source identification
- 477 using semi-supervised machine learning. *Journal of Contaminant Hydrology*, 212, 134–142.
- 478 <u>https://doi.org/10.1016/j.jconhyd.2017.11.002</u>
- 479 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012).
- 480 Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PloS One*, 7(1),
- 481 e29715. <u>https://doi.org/10.1371/journal.pone.0029715</u>

- 482 Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al.
- 483 (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific
- 484 Data, 3, 160018. <u>https://doi.org/10.1038/sdata.2016.18</u>
- 485 Wimmers, A., Velden, C., & Cossuth, J. H. (2019). Using Deep Learning to Estimate Tropical
- 486 Cyclone Intensity from Satellite Passive Microwave Imagery. Monthly Weather Review, 147(6),
- 487 2261–2282. <u>https://doi.org/10.1175/MWR-D-18-0391.1</u>
- 488 Wintoft, P., Wik, M., Matzka, J., & Shprits, Y. (2017). Forecasting Kp from solar wind data:
- 489 input parameter study using 3-hour averages and 3-hour range values. Journal of Space Weather
- 490 and Space Climate, 7, A29. <u>https://doi.org/10.1051/swsc/2017027</u>
- 491 Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011).
- 492 Minimum information about a marker gene sequence (MIMARKS) and minimum information
- 493 about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29(5), 415–420.
- 494 <u>https://doi.org/10.1038/nbt.1823</u>
- 495 Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating
- 496 Explanations for Graph Neural Networks. Advances in Neural Information Processing Systems,
- 497 32, 9240–9251. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/32265580
- 498 Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond Empirical Risk
- 499 Minimization. *arXiv* [cs.LG]. Retrieved from <u>http://arxiv.org/abs/1710.09412</u>