







POSITION PAPER

Improving the reproducibility in geoscientific papers: lessons learned from a Hackathon in climate science

Alejandro Coca-Castro¹ , Anne Fouilloux² , Ricardo Barros Lourenço³ , Andrew McDonald^{4,5} ,
Yuhan Rao⁶  and J. Scott Hosking^{1,5} 

¹Environment and Sustainability Grand Challenge, The Alan Turing Institute, London, UK

²Simula Research Laboratory, Oslo, Norway

³School of Earth, Environment & Society, McMaster University, Hamilton, ON, Canada

⁴Department of Engineering, University of Cambridge, Cambridge, UK

⁵British Antarctic Survey, NERC, UKRI, Cambridge, UK

⁶North Carolina Institute of Climate Studies, North Carolina State University, Asheville, NC, USA

Corresponding author: Alejandro Coca-Castro; Email: acoca@turing.ac.uk

Received: 27 September 2024; **Accepted:** 27 September 2024

Keywords: climate informatics; computational research; notebooks; reproducible research; reproduction assessment

Abstract

In this paper, we explore the crucial role and challenges of computational reproducibility in geosciences, drawing insights from the Climate Informatics Reproducibility Challenge (CICR) in 2023. The competition aimed at (1) identifying common hurdles to reproduce computational climate science; and (2) creating interactive reproducible publications for selected papers of the Environmental Data Science journal. Based on lessons learned from the challenge, we emphasize the significance of open research practices, mentorship, transparency guidelines, as well as the use of technologies such as executable research objects for the reproduction of geoscientific published research. We propose a supportive framework of tools and infrastructure for evaluating reproducibility in geoscientific publications, with a case study for the climate informatics community. While the recommendations focus on future CIRCs, we expect they would be beneficial for wider umbrella of reproducibility initiatives in geosciences.

Impact Statement

This position paper discusses common challenges to reproduce computational climate science according to an online reproduction competition using interactive publications. The authors place the competition according to the state of other reproducibility initiatives in scientific communities that are naturally close to software. The proposed format of the competition is innovative by elevating the role of executable research objects and community-driven platforms to assist a collaborative computational peer review process of scholarly publications in climate and environmental data science. By proposing a supportive framework of tools and infrastructure for evaluating reproducibility for the climate informatics community, we call an action for all stakeholders involved in the geoscientific publication to strength such frameworks by implementing funding and incentives mechanisms that also support long-term maintenance of the evaluated artifacts in reproducible computational research.

1. Introduction

Reproducibility and replicability form the foundation of scientific research. These core principles not only ensure the reliability of findings but also enable scientific ideas and results to be tested, refined, and built

upon. Their pertinence extends seamlessly from large-scale studies of high-profile papers to smaller settings. For the latter, reproducibility becomes pivotal, especially for small laboratories with a limited number of technical staff. It allows new team members to construct upon previous work without excessive expenditure of their already limited resources (Leipzig et al., 2021). This scenario also underlines the importance of meticulous data management and extensive documentation in scientific research. Computational experiments play an integral role in the scientific method. They do not only support computer science research but also natural science, social science, and humanity research. One of the key challenges in computational research is to establish and maintain trust in the experiments and artifacts that are presented in published results. A fundamental aspect of building trust is the ability to reproduce those results. A computational experiment that has been developed at time t on hardware/operating system s on data d is reproducible if it can be executed at time t' on system s' on data d' that is similar to (or potentially the same as) d (Freire et al., 2012). In the context of computational research, reproducibility and replicability terms hold different meanings. Following the National Academies of Sciences, Engineering, and Medicine's (2019) report, we define reproducibility as "obtaining consistent results using the same material, data, methods and conditions of analysis". In contrast, we refer to replicability as "obtaining consistent results under new conditions with new materials, data, or methods and independently confirming original findings". These definitions are aligned to The Turing Way's (2023) matrix of computational reproducibility (Figure 1) which also includes robustness (same data, different code) and generalizable (different data and different code).

Willis and Stodden (2020) 's work claims the evolution of computational reproducibility as we understand it today has been shaped by four movements. First, early works in computer science, mathematics, and statistics focused on the review and distribution of scientific software libraries (Hopkins, 2009). Second, the "replication standard" movement in political science and economics in the 1980s and 1990s advocated for authors to share code and data for the evaluation and replication of research (Dewald et al., 1986; King, 1995). Third, the "reproducible research" movement, initiated by geoscientists (Claerbout and Karrenbach, 1992) was later adopted in statistics and signal processing. The movement emphasized a vision of a system where computational artifacts could be used to reproduce publications, including tables and figures, by "pressing a single button". Finally, the "repeatability" movement in computer science originated in the database systems community (Manolescu et al., 2008).

Fields such as political science (King, 1995; Peer et al., 2021), health sciences (Munafò et al., 2017), economics (Vilhuber, 2020), computer science (Manolescu et al., 2008; Frachtenberg, 2022), physics (Clementi and Barba, 2021) and statistics (Xiong and Cribben, 2023) have been proactive in discussing and improving upon reproducibility. For geosciences, there are some progresses for areas with a direct impact on human lives, economies, and public policy. Konkol et al. (2019) investigated the state of reproducibility in geosciences and provided a set of guidelines for authors aiming at reproducible

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Figure 1. Matrix of reproducibility (The Turing Way, 2023); made available under the Creative Commons Attribution license (CC-BY 4.0).

scientific publications. Stagge et al. (2019) revealed only 1.6% of results were fully reproducible among 360 of the 1989 articles published by six hydrology and water resources journals in 2017. An extensive survey of 347 participants from multiple fields within the earth sciences indicated a poorly documented workflow, a lack of code documentation and the availability of code and data are the major reasons for the lack of reproducibility (Reinecke et al., 2022). Bush et al. (2020) identified challenges and provided recommendations for the improvement of reproducibility and replicability in paleo-climate research and researchers using computer-simulated global climate model (GCM) experiments.

With the growing number of digital artifacts supporting scholarly publications in climate informatics and environmental data science, there is a pressing call to invest in robust frameworks and initiatives geared toward more trustworthy research in these fields. Building upon reproducibility initiatives led by Computer Science, Machine Learning, and Geographic Information Science (GIScience) communities, we convened the inaugural Climate Informatics Reproducibility Challenge immediately following the 12th Annual Conference on Climate Informatics in 2023. The competition aimed at (1) identifying common hurdles to reproduce computational climate science; and (2) creating interactive reproducible publications for selected papers of the Environmental Data Science journal. We served a set of community-driven tools and cloud-based platforms that make it easier to participate in the revision of computational research artifacts (e.g., datasets, analysis code, workflows, and the environment) and generation of reproducibility reports.

This paper seeks to (1) revisit the importance of reproducible computational research in geosciences, (2) describe initiatives and lessons learned from a recent reproduction competition organized for the Climate Informatics community, and (3) propose a supportive framework of tools and infrastructure toward the improvement of reproducibility and replicability in computational climate science. The paper is structured as follows: [Section 2](#) looks at the revision of reproducibility initiatives; [Section 3](#) focuses on describing and sharing the lessons learned from the reproduction competition; [Section 4](#) describes a framework for sustainably measuring and monitoring reproducibility and replicability in the field of climate informatics, and conclusions are examined in [Section 5](#).

2. Reproducibility initiatives

Reproducibility initiatives encompass formal activities undertaken by journal editors, conference organizers, or related stakeholders to improve the transparency and reproducibility of computational research published via their venues through the adoption of new policies, workflows, and infrastructure (Willis and Stodden, 2020). While the initiatives seem to have similar goals, they differ widely concerning policy mandates, what is reviewed, who conducts the review, and how reviewers are incentivized. We mention below some examples of initiatives led by computer science, machine learning, and GIScience communities for evaluating reproducibility and promoting trustworthy science in their computational research.

The supercomputing (SC) conference in 2015 began an optional submission for authors of accepted papers to describe further their experimental framework and results. The structure (still practiced to date) includes the submission of an artifact description (AD) appendix and a more extensive artifact evaluation (AE) appendix. The AD appendix allows us to determine whether artifacts are available, and the AE appendix provides sufficient detail to support computational reproducibility. In 2015, only a single paper responded to the initiative and became the source for the SC16 Student Cluster Competition Reproducibility Challenge and the first SC paper to display an Artifact Review and Badging (ACM) badge. By 2017 39 papers included an AD appendix. In 2019 the AD appendix became mandatory (Malik et al., 2022). The same badging system was also proposed for the remote sensing community (Frery et al., 2020), in an attempt to address the specifics for that domain, such as the enforcement of utilization of Open Data and Free/*Libre* Open Source Software (FLOSS).

The 2018 International Conference on Learning Representations (ICLR) reproducibility challenge was designed as a dedicated platform to investigate the reproducibility of papers submitted in the machine learning (ML) domain. Most participants were enrolled in graduate ML courses, and the challenge served them as the final course project. The selection of ICLR was highly motivated because the conference

submissions were automatically made available publicly on OpenReview, including during the review period. The use of the OpenReview platform allowed a wiki-like interface to provide transparency and a conversation space between authors and participants. The inaugural challenge was followed by the 2019 ICLR Reproducibility Challenge (Pineau et al., 2019), and the 2019 NeurIPS Reproducibility Challenge (Pineau et al., 2020), and many others to present with considerable improvements in the infrastructure, for example, partnership with Kaggle competitions (Sinha et al., 2023).

The Association of Geographic Information Laboratories in Europe (AGILE) initiated a series of workshops on reproducibility starting from 2017 until 2019. These workshops aimed to address the issue of reproducibility in the GIScience community. To ensure that reproducibility is measured and reported accurately, AGILE began in 2019 to produce reproducibility reviews of full papers after acceptance. The reproducibility reports and guidelines for authors and reviewers are published on the Open Science Framework. AGILE has extended the use of this framework to other communities such as GIScience. They have also taken a step further to include non-English speaking members of their communities by providing translated guidelines. This innovative approach enhances inclusivity and fosters collaboration among researchers from diverse backgrounds (Nüst et al., 2018).

The AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) promoted in 2022 a summer school, that aside from the regular training agenda, introduced what they named “trust-a-thon”, to implement explainability and interpretability of pre-trained ML models, which included challenges in models for severe storms, tropical cyclones, and space weather (McGovern et al., 2023). The participants were provided with a scientific application, curated datasets, and pre-developed code to train models for such tasks. For each challenge, it was also provided user “personas”, as examples of potential end users that would consume information about these models, and their limitations, obtained by interpretability and explainability techniques applied to those ML model outcomes. This “trust-a-thon” used extensively Jupyter Notebooks in a preset cloud environment and encouraged users to explore its interface capacities to make model outcomes more palatable to end-users, using graphics and communication that is easy to consume by these decision-making stakeholders (emergency managers, forecasters, policy-makers, among others).

3. Climate informatics reproducibility challenge

3.1. Overview

The Climate Informatics (CI) conference series attracts over 100 participants annually. Since 2015, CI started to organize hands-on hackathons collocated with annual CI conferences. These hackathons bring together community members to investigate solutions for climate challenges including the generation of synthetic night-time imagery and prediction tasks of CO₂ fluxes, arctic sea ice extent, and winter extreme rainfall, among others. Building upon the collaborative nature of these events, the 12th edition of the conference began the inaugural Climate Informatics Reproducibility Challenge in 2023 (CIRC23) to advance the scientific rigor of computational climate science. The objective was twofold (1) identifying common hurdles to reproducing computational climate science; and (2) creating interactive reproducible publications for selected papers of the EDS journal via the Environmental Data Science (EDS) book, a community-driven platform and open-source resource to host demonstrators via Jupyter notebooks (EDS book, 2023).

The CIRC23 was structured as a month-long event. Participants were required to register (see [supplement for detailed information](#)) before the start of the competition while reviewers could join until the official reviews started. When the challenge officially began, participants were assigned a team of three to four members and got their project assignments, that is, pre-approved papers from the EDS journal. The submission of a Jupyter notebook was a two-stage process: each team was expected to first submit a Minimum Viable Working notebook so that reviewers could start getting familiar with the content. Then the challenge concluded with the submission of the final peer-reviewed notebook (about 2 weeks after the first submission). The formal evaluation process was meant to improve the quality of the

submissions, and this is why we recruited independent reviewers to conduct an open review with constructive feedback on the reproducibility reports. Submissions were judged based on their adherence to open science and FAIR principles, relevance to the EDS book's objectives, reliability of results, and contribution to the overall scope of the competition. Reviewers used the guidelines to ensure objectivity and focus on the accurate reproduction of original studies, completeness of the provided information, clarity of the report, quality of code, visualizations, effective data management, and insightful comments on the original paper and results. For the entire review process, the infrastructure of EDS book was used. A fundamental part of the challenge was the provision of a JupyterHub cloud-based deployment with pre-installed scientific software in the core programming languages (Python, R and Julia). The prizes consisted of vouchers in books published by Cambridge University Press and Assessment.

While 3 out of 7 teams successfully submitted their reproducibility reports (Figure 2) of the assigned paper within the challenge duration (Domazetoski et al., 2023; Malhotra et al., 2023; Pahari et al., 2023), there was always the overarching concern of the time and resources required to achieve the challenge scope. To assess the challenge, we gathered feedback from participants and reviewers on a variety of questions (see supplement for detailed information). The time and skills required to generate reproduction studies through Jupyter notebooks are considerably higher than traditional reproducibility reports. For example, according to the CIRC23 feedback, participants spent 12–40 hours on the submission and reviewers 5–10 hours. These numbers are considerably higher than Stagge et al. (2019)'s findings claiming a range of 0.5–1 hour to conduct checklist-driven reproduction reports for papers in journals of hydrology and water resources. Within the feedback, teams and reviewers also provided recommendations for future editions of the challenge. One participant indicated it would be beneficial to have a priori information of the estimated time to run each experiment, for example, model training, and how much memory it requires. Reviewers also suggested more time between the informative sessions and the review so there is a chance to get a practice review, and more interaction with the teams.

The figure displays two side-by-side screenshots. The left screenshot shows a Jupyter notebook interface titled 'Learning the underlying physics of a simulation model of the ocean's temperature (CIRC23)'. It features a navigation menu on the right with options like 'Context', 'Purpose', 'Description', 'Highlights', 'Contributions', 'Data', 'Source code', 'Load libraries', 'Dataset', 'Download input material', 'Reproduction paper', 'Model training', 'Summary', 'Citing this Notebook', and 'Additional information'. The main content area includes a title, a 'Context' section with a paragraph about data-driven models, and a world map visualization showing ocean temperature patterns. Below the map is a source attribution: 'Source: Image generated by the NASA and taken from Common Dreams.' The right screenshot shows a Cambridge Core journal article page for 'A sensitivity analysis of a regression model of ocean temperature'. It includes the journal title 'Environmental Data Science', the publication date '30 August 2022', the authors 'Rachel Furner, Peter Haynes, Dave Munday, Brooks Paige, Daniel C. Jones and Emily Shuckburgh', and an abstract section. The abstract discusses the development of data-driven models for weather and climate predictions and the challenges of generalizability and robustness.

Figure 2. Example of an interactive reproducibility report (left) authored by Malhotra et al. (2023) published in EDS book, and the target published paper (right) authored by Furner et al. (2022), published in the EDS journal.

3.2. *Lessons learned*

Even though CICR23 was a pioneer competition for evaluating computational artifacts in climate science, we share some lessons to consider for future editions of the challenge. The lessons are transferable to competitions alike in other domains.

Strengthening open research practices and mentoring: making artifacts available requires authors to document additional materials and learn new skills and technologies in open research. Considering most earth scientists are self-taught programmers (Reinecke et al., 2022), we suggest making open research practices more accessible irrespective of the seniority level and promoting timely collaborations with research infrastructure roles (research software engineers, data wranglers, and stewards). Mentoring programs within reproducibility competitions are also a beneficial path to building capabilities and networks. For CIRC23, we found that the lack of mentorship impacted participants' interactions with networking activities offered by the program such as talks of experts (eight in total) and weekly drop-in sessions. According to post-challenge feedback, the communication between participants was a barrier, so mentors could provide a safe environment for the challenge by moderating conversations and reinforcing the scope and activities of the competition.

Supporting community standards and sustainability: the creation of standards for transparency and data sharing facilitates what artifacts and metadata should be deposited with the published research (Leipzig et al., 2021). For instance, Hydroshare, operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI), enables the sharing and publication of data and models in a citable and discoverable manner for hydrology and water resources (Horsburgh et al., 2016). The EarthCube Model Data Research Coordination Network provides guidance on what data and software elements of simulation-based research, for example, weather modeling need to be preserved and shared to meet community open science expectations, including publishers and funding agencies. In the CIRC23, 1 out of 7 target papers had all computational artefacts (data and code) citable and discoverable via Zenodo. The authors of the remaining papers stored code and data in version control platforms such as GitHub and GitLab.

Elevating the importance of executable research objects: assessing reproducibility entails a lot of work. Collaborative interactive computing can facilitate the inspection of results and peer review (Leipzig et al., 2021). We have found executable research objects, Jupyter Notebook files, in this case, informative and helpful in demonstrating results. We propose to consider them as a supporting artifact that facilitates the inspection of other artifacts (data, code, and directions to run) of published papers. This perspective aligns with existing initiatives that involve the publishing industry to make notebooks the main artifact for scientific publication (Caparelli et al., 2023) and the concept of an "executable research compendium" (Nüst et al., 2017). Adoption of good practices for writing and sharing notebooks could make research artifacts associated with published research more robust (Rule et al., 2019). While notebooks might be not optimal for sharing experiments with large datasets or complex workflows (except if there is access to infrastructure that supports this), we suggest creating a demonstration dataset and/or minimal working version with a subset of the whole study for testing reproducibility.

Advancing on social dimensions and incentives: while the technical dimensions associated with tools and infrastructure are important for computational reproducibility, expanding the peer review process and communication between the stakeholders involved in reproducible computational research also have important social and organizational dimensions (Willis and Stodden, 2020). For instance, there is an opportunity to engage and connect early career researchers (ECRs) with experts in research software and data management during the peer review process. Communication between evaluators of computational artifacts and authors of the original study is also key to improving reproducibility. A study of 613 articles in computer systems research found that reproducibility varied depending on whether communication with authors took place. When not communicating with authors, 32.1% of the experiments could be reproduced, whereas this number increased to 48.3% when communication occurred (Collberg and Proebsting, 2016). We evidenced the relevance of communication in CIRC23 as one of the teams had direct communication with the authors of the target paper and reported a missing file that was

key to completing the reproduction of the research workflow. All collaborations and communications should be transparent and be professionally rewarded and acknowledged, for example, via DOI citable reports (Stagge et al., 2019).

4. Towards an ecosystem of tools and infrastructure to support reproducible geosciences

While the format of CICR23 was innovative by incorporating interactive computing, that is, Jupyter Notebooks for generating reproducibility reports, the competition format could be maximized by its integration with a larger supportive framework of tools and infrastructure. Following the successful intake and sustainability of reproducibility initiatives in large- to medium-sized conferences such as SC and AGILE, we suggest CI conferences implement human-centred frameworks for reporting reproducibility. Inspired by Reinecke et al. (2022), Figure 3 summarises the proposed framework to strength reproducible research in CI. The submitted papers will be part of a reproducibility check in addition to the standard scientific review. To ensure equitable treatment of the diversity research spectrum, we propose to recruit enthusiasts in software and data management (at any seniority level) to conduct a technical review of the computational artifacts. The assessment will be assisted by clear guidelines adapted from AGILE and SC communities. Furthermore, we suggest hosting capacity-building activities to train authors and technical reviewers to evaluate the quality of the computational artifacts for improving the reproducibility of the published research. The papers that are fully reproducible are then targeted for a replication competition like SC conferences. In this competition, participants (mostly ECRs from university labs) will be judged for the capacity to replicate results using a different dataset or methods to potentially make a new discovery or idea. This process will be conducted through community-driven platforms such as EDS book.

It is worth mentioning the reproducibility check of the framework should involve all actors in publishing and sharing research results. For instance, Peer et al. (2022) introduce the Reproducible Research Publication Workflow which includes the evaluation of the artifacts and metadata creation as integral in the publication workflow. The proposed approach involves accumulating and updating objects and metadata along the publication process in collaboration with authors and research infrastructure roles such as data curators, stewards, and research software engineers. These conceptual approaches open opportunities to develop new infrastructure. While there is a growing market of private-led initiatives that

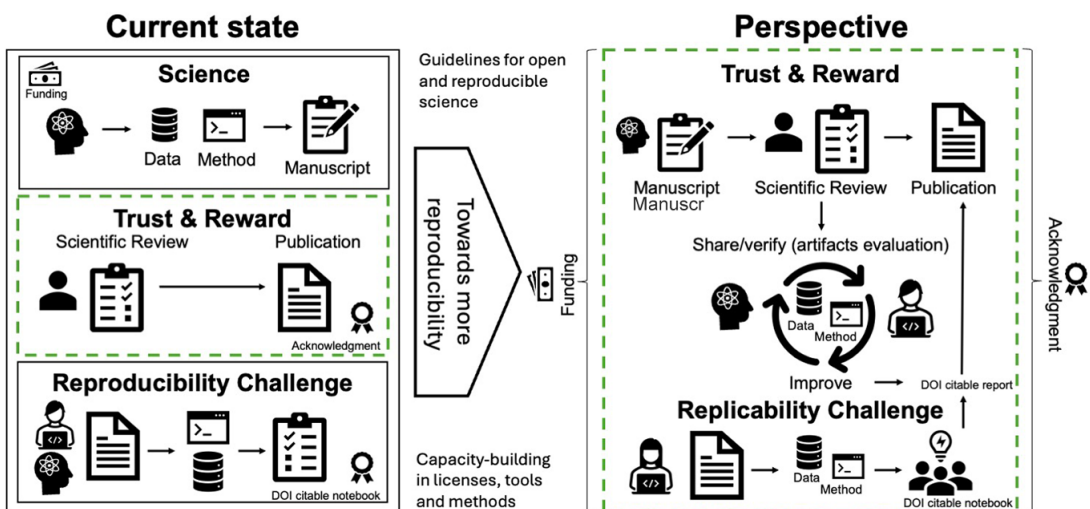


Figure 3. Our concept of what reporting of reproducibility should look like for future editions of the *Climate Informatics* conference. Adapted from the [supplementary material](#) of Reinecke et al. (2022).

have supported reproducibility in well-known journals (see for example Kousta et al. (2019)), it remains important that technology solutions are interoperable and vendor-free.

Finally, another area of work is ensuring the sustainability of the computational reproducibility of scientific results. Most reproducibility efforts are focused on point-of-manuscript-submission, however, it remains important to identify frameworks to support long-term maintenance of the evaluated artifacts (see for example Peer et al., 2021), as well implementing funding and incentive mechanisms that enable such enterprise (see Recommendations 6–3, 6–4, 6–5, and 6–6 in National Academies of Sciences, Engineering, and Medicine (2019)).

5. Conclusions

The field of geosciences faces some hurdles and complications with making computational research reproducible. Initiatives like the CIRC23 support efforts for a more open and collaborative-driven research environment. By putting theory into practice, we hope to reinforce Open Science practices in research and inspire its adoption within the wider scientific community. Our experience with the CIRC23 emphasized the importance of open research practices, mentorship programs, and standard transparency guidelines. Tools such as executable research objects were found to be highly beneficial. We propose a comprehensive and sustainable framework inclusive of clear guidelines, supportive tools, and all stakeholders and roles involved in scientific publications.

In conclusion, with the amplification of Open Science principles and by embracing shared learning and open dialogues, computational reproducibility in geosciences can enhance its position in streamlining the delivering of high-quality research that influences policymaking and strategies affecting our planet.

Open peer review. To view the open peer review materials for this article, please visit <http://doi.org/10.1017/eds.2024.35>.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/eds.2024.35>.

Acknowledgements. We thank Louisa Van Zeeland (Research Lead, The Alan Turing Institute) and the anonymous reviewer for their helpful comments and suggestions. We are also grateful with all individuals (participants, reviewers, judges, guest speakers, infrastructure, organizers and helpers) who contributed to the 2023 climate informatics reproducibility challenge. Their contributions are acknowledged in the public GitHub repository of the competition, <https://github.com/eds-book/reproducibility-challenge-2023?tab=readme-ov-file#contributors->. Special thanks to the Pangeo-EOSC project which has benefited from services and resources provided by the EGI-ACE project (funded by the European Union’s Horizon 2020 research and innovation program under Grant Agreement no. 101017567), and the C-SCALE project (funded by the European Union’s Horizon 2020 research and innovation program under grant agreement no. 101017529), with the dedicated support of CESNET. We also thank Andrew Hyde at Cambridge University Press, the publisher of Environmental Data Science, for supporting the reproducibility challenge.

Author contribution. Contributions are listed in the order of the author list. Conceptualization: A.C-C., Visualization: A.C-C., Project administration: A.F., Y.R., J.S.H., Funding acquisition: J.S.H., Writing original draft: A.C-C., Writing – review & editing: A.C-C., A.F., R.B.L., A.M., Y.R. All authors approved the final submitted draft.

Data availability statement. The public repository of the challenge is archived in Zenodo: <https://doi.org/10.5281/zenodo.10637360>. Details of the configuration of the Pangeo-EOSC cloud deployment used in the challenge are available at <https://github.com/pangeo-data/pangeo-eosc>. The notebooks submitted to the challenge are published in the EDS book website, www.edsbook.org, and archived in Zenodo (Domazetoski et al., 2023; Malhotra et al., 2023; Pahari et al., 2023).

Provenance statement. This article is part of the Climate Informatics 2024 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer review process.

Funding statement. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under the research grant EP/Y028880/1. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. A.M. was supported by the Marshall Scholarship and Gates Cambridge Scholarship during this work. R.B.L. was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery grant (RGPIN-2020-05,708), Canada Research Chairs Program (CRC2019–00139) and the McMaster University Centre for Climate Change Research Seed Fund during this work. Y.R. was supported by NOAA through the Cooperative Institute for Satellite Earth System Studies under Cooperative Agreement NA19NES4320002.

Ethical statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study countries.

Competing interest. R.B.L. was on the Editorial Board of the Environmental Data Science Journal.

References

- Bush R, Dutton A, Evans M, Loft R and Schmidt GA** (2020) Perspectives on data reproducibility and replicability in paleoclimate and climate science. *Harvard Data Science Review* 2 (4). <https://doi.org/10.1162/99608f92.00cd8f85>.
- Caprarelli G, Sedora B, Ricci M, Stall S and Giampoala M** (2023) *Notebooks now!* The future of reproducible research *Earth and Space Science* 10 (12), e2023EA003458. <https://doi.org/10.1029/2023EA003458>.
- Claerbout JF and Karrenbach M** (1992) Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*. Society of Exploration Geophysicists, pp. 601–604. <https://doi.org/10.1190/1.1822162>.
- Clementi NC and Barba LA** (2021) Reproducible validation and replication studies in nanoscale physics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379 (2197), 20200068. <https://doi.org/10.1098/rsta.2020.0068>.
- Collberg C and Proebsting TA** (2016) Repeatability in computer systems research. *Communications of the ACM* 59 (3), 62–69. <https://doi.org/10.1145/2812803>.
- Dewald WG, Thursby JG and Anderson RG** (1986) Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review*, 76(4), 587–603.
- Domazetoski V, Zúñiga-González A, Allemang O and contributors, T. E. book notebook** (2023) *Deep Learning and Variational Inversion to Quantify and Attribute Climate Change (Jupyter Notebook)* Published in the Environmental Data Science Book (v1.0.1) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.8330771>.
- EDS book** (2023) *Environmental Data Science Book: A Computational Notebook Community for Open Environmental Data Science*. (v0.2.1) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.8359648>.
- Frachtenberg E** (2022) Research artifacts and citations in computer systems papers. *PeerJ Computer Science* 8, e887. <https://doi.org/10.7717/peerj-cs.887>.
- Freire J, Bonnet P and Shasha D** (2012) Computational reproducibility: State-of-the-art, challenges, and database research opportunities. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 593–596. <https://doi.org/10.1145/2213836.2213908>.
- Frery AC, Gomez L and Medeiros AC** (2020) A badging system for reproducibility and replicability in remote sensing research. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13, 4988–4995. <https://doi.org/10.1109/JSTARS.2020.3019418>.
- Furner R, Haynes P, Munday D, Paige B, Jones DC and Shuckburgh E** (2022) A sensitivity analysis of a regression model of ocean temperature. *Environmental Data Science* 1, e11. <https://doi.org/10.1017/eds.2022.10>.
- Hopkins T** (2009) The collected algorithms of the ACM. *WIRES Computational Statistics* 1(3), 316–324. <https://doi.org/10.1002/wics.40>.
- Horsburgh JS, Morsy MM, Castronova AM, Goodall JL, Gan T, Yi H, Stealey MJ and Tarboton DG** (2016) HydroShare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain. *JAWRA Journal of the American Water Resources Association* 52(4), 873–889. <https://doi.org/10.1111/1752-1688.12363>.
- King G** (1995) Replication, replication. *PS: Political Science & Politics* 28 (3), 444–452. <https://doi.org/10.2307/420301>.
- Konkol M, Kray C and Pfeiffer M** (2019) Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science* 33 (2), 408–429. <https://doi.org/10.1080/13658816.2018.1508687>.
- Kousta S, Pastrana E and Swaminathan S** (2019) Three approaches to support reproducible research. *Science Editor* 42, 77–82.
- Leipzig J, Nüst D, Hoyt CT, Ram K and Greenberg J** (2021) The role of metadata in reproducible computational research. *Patterns* 2 (9), 100322. <https://doi.org/10.1016/j.patter.2021.100322>.
- Malhotra G, Pinto Veizaga D, Peña Velasco JE and contributors, T. E. book notebook** (2023) *Learning the Underlying Physics of a Simulation Model of the Ocean's Temperature (Jupyter Notebook)* Published in the Environmental Data Science Book (v1.0.17) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.8314669>.
- Malik T, Vahldiek-Oberwagner A, Jimenez I and Maltzahn C** (2022) Expanding the scope of artifact Evaluation at HPC conferences: experience of SC21. In *Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems*, pp. 3–9. <https://doi.org/10.1145/3526062.3536354>.
- Manolescu I, Afanasiev L, Arion A, Ditttrich J, Manegold S, Polyzotis N, Schnaitter K, Senellart P, Zoupanos S and Shasha D** (2008) The repeatability experiment of SIGMOD 2008. *ACM SIGMOD Record* 37(1), 39–45. <https://doi.org/10.1145/1374780.1374791>.
- McGovern A, Gagne DJ, Wirz CD, Ebert-Uphoff I, Bostrom A, Rao Y, Schumacher A, Flora M, Chase R, Mamalakis A, McGraw M, Lagerquist R, Redmon RJ and Peterson T** (2023) Trustworthy Artificial Intelligence for Environmental Sciences: An Innovative Approach for Summer School. *Bulletin of the American Meteorological Society*, 104(6), E1222–E1231. <https://doi.org/10.1175/BAMS-D-22-0225.1>

- Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ and Ioannidis JPA (2017) A manifesto for reproducible science. *Nature Human Behaviour* 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>.
- National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and Replicability in Science*. National Academies Press, p. 25303. <https://doi.org/10.17226/25303>.
- Nüst D, Granell C, Hofer B, Konkol M, Ostermann FO, Sileryte R and Cerutti V (2018) Reproducible research and GIScience: An evaluation using AGILE conference papers. *PeerJ* 6, e5072. <https://doi.org/10.7717/peerj.5072>.
- Nüst D, Konkol M, Pebesma E, Kray C, Schutzzeichel M, Przybytzin H and Lorenz J (2017) Opening the publication process with executable research compendia. *D-Lib Magazine* 23 (1/2). <https://doi.org/10.1045/january2017-nuest>.
- Pahari M, Bhoir R and contributors, T. E. *book notebook* (2023) *Variational Data Assimilation with Deep Prior (Jupyter Notebook)* Published in the *Environmental Data Science Book* (pre-release) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.8339299>
- Peer L, Biniössek C, Betz D and Christian T-M (2022) Reproducible research publication workflow: A canonical workflow framework and FAIR digital object approach to quality research output. *Data Intelligence* 4 (2), 306–319. https://doi.org/10.1162/dint_a_00133.
- Peer L, Orr LV and Coppock A (2021) Active maintenance: A proposal for the long-term computational reproducibility of scientific results. *PS: Political Science & Politics* 54 (3), 462–466. <https://doi.org/10.1017/S1049096521000366>.
- Pineau J, Sinha K, Fried G, Ke RN and Larochelle H (2019) *ICLR Reproducibility Challenge 2019*. Zenodo. <https://doi.org/10.5281/ZENODO.3158244>
- Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, Fox E and Larochelle H (2020) *Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)* (arXiv:2003.12206). <http://arxiv.org/abs/2003.12206>.
- Reinecke R, Trautmann T, Wagener T and Schüler K (2022) The critical need to foster computational reproducibility. *Environmental Research Letters* 17 (4), 041005. <https://doi.org/10.1088/1748-9326/ac5cf8>.
- Rule A, Birmingham A, Zuniga C, Altintas I, Huang S-C, Knight R, Moshiri N, Nguyen MH, Rosenthal SB, Pérez F and Rose PW (2019) Ten simple rules for writing and sharing computational analyses in Jupyter notebooks. *PLoS Computational Biology* 15 (7), e1007007. <https://doi.org/10.1371/journal.pcbi.1007007>.
- Sinha K, Bleeker M, Bhargav S, Forde JZ, Raparthy SC, Dodge J, Pineau J and Stojnic R (2023) *ML Reproducibility Challenge 2022*. Zenodo. <https://doi.org/10.5281/ZENODO.8200058>.
- Stage JH, Rosenberg DE, Abdallah AM, Akbar H, Attallah NA and James R (2019) Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data* 6 (1), 190030. <https://doi.org/10.1038/sdata.2019.30>.
- The Turing Way (2023) *The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research* (1.0.2) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7625728>.
- Vilhuber L (2020) Reproducibility and replicability in economics. *Harvard Data Science Review* 2 (4). <https://doi.org/10.1162/99608f92.4f6b9e67>.
- Willis C and Stodden V (2020) Trust but Verify: How to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harvard Data Science Review* 2 (4). <https://doi.org/10.1162/99608f92.25982dcf>.
- Xiong X and Cribben I (2023) The state of play of reproducibility in statistics: An empirical analysis. *The American Statistician* 77 (2), 115–126. <https://doi.org/10.1080/00031305.2022.2131625>.