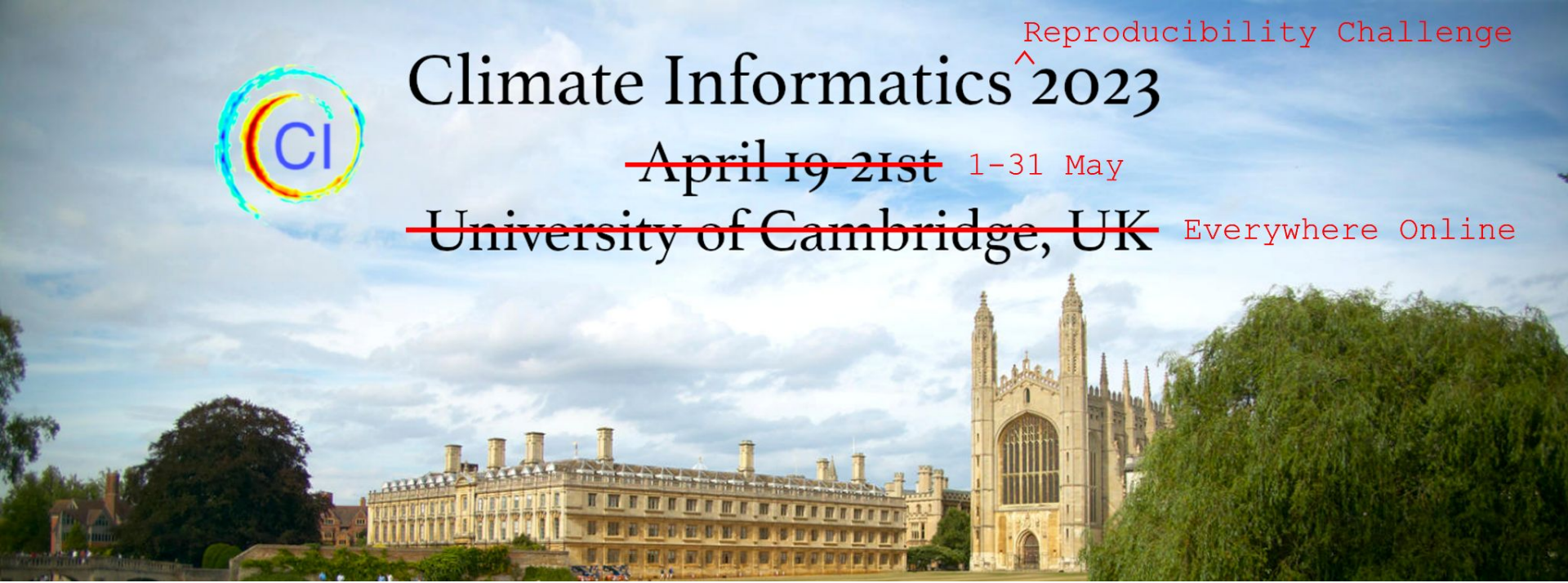


Implementing Reproducible Environmental Data Science with Open Science:

Lessons from the 1st Climate Informatics Reproducibility Challenge

Andrew McDonald*, Alejandro Coca-Castro, Anne Fouilloux, Ricardo Barros Lourenço, Andrew Hyde & Yuhan Rao

University of Cambridge | British Antarctic Survey

Outline

- **Context**
- **The Challenge**
  - Scope
  - Communities
- **Outcomes**
  - Participation
  - Submissions and demo (winning team)
  - Feedback
- **Reflections and Next steps**

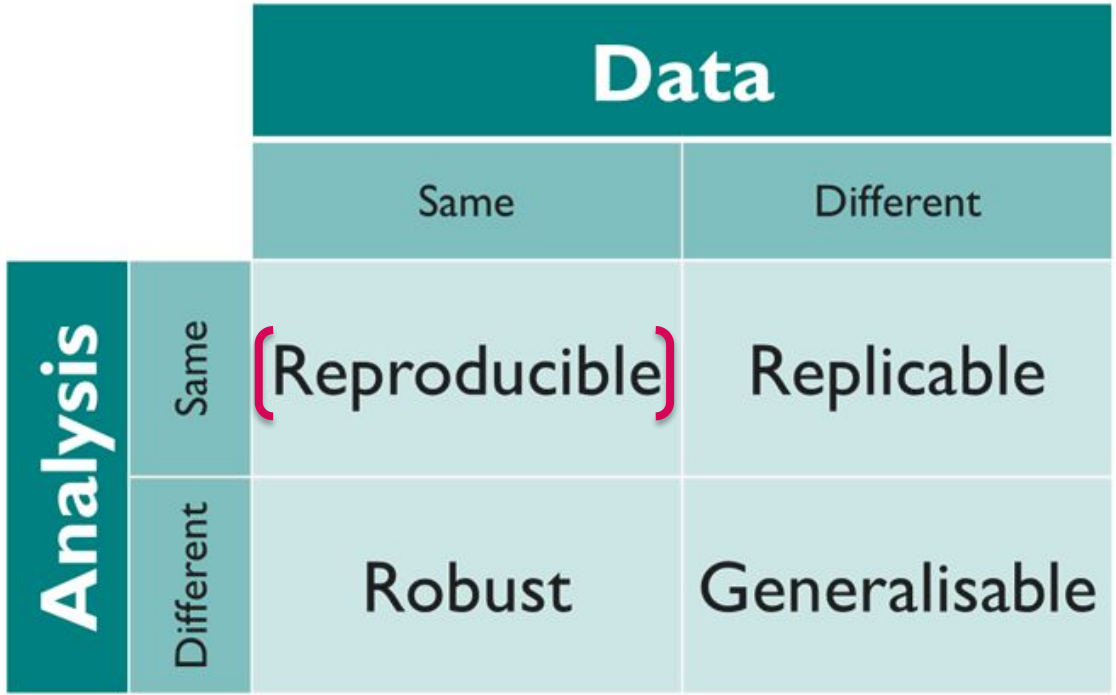
Reproducible Research

same analysis steps

on the same dataset

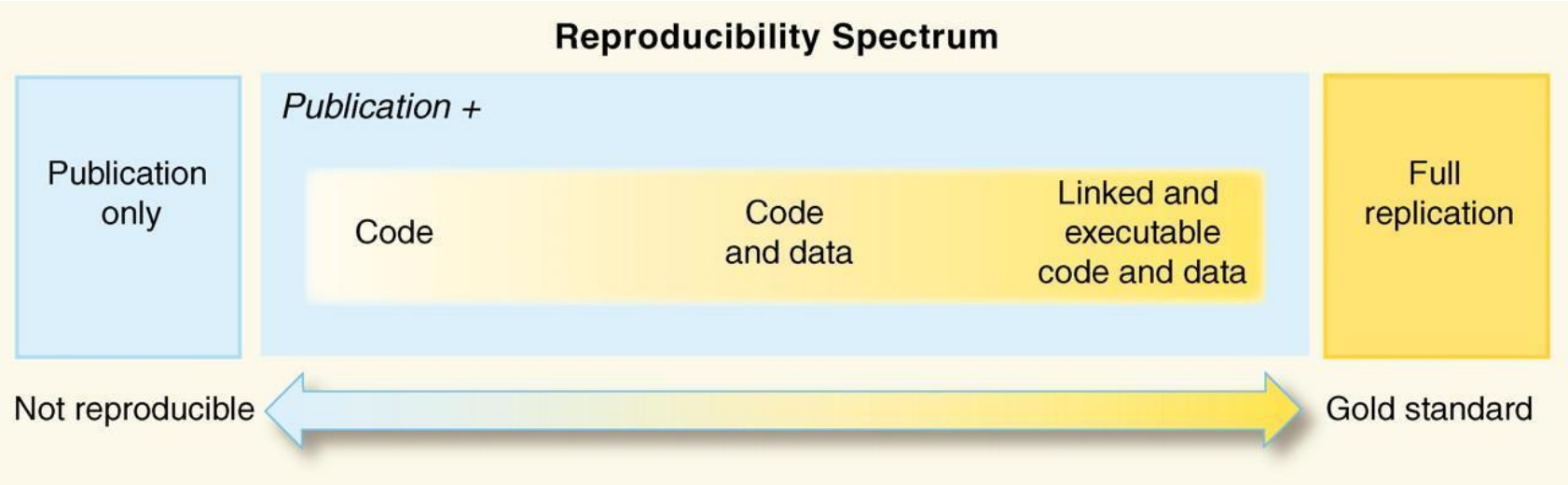
produces same

answer



https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

The Reproducibility Spectrum



Peng, 2011, doi: 10.1126/science.1213847

Reproducibility Initiatives

Supercomputing

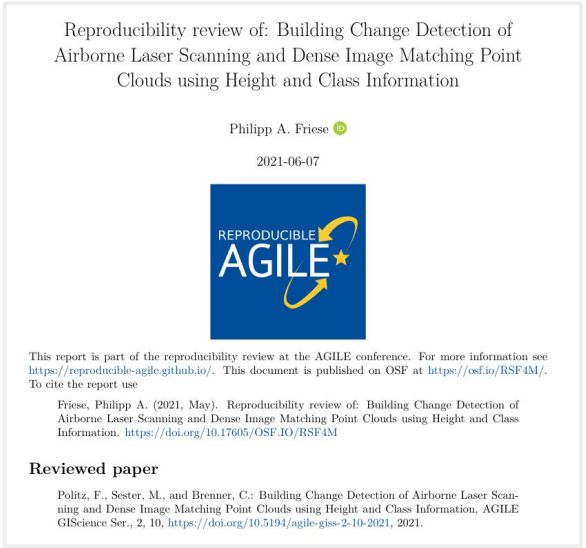
AD/AE Appendix Process & Badges

SC23 Denver, CO | i am hpc.

Results Replicated acm

2015-Present

Geographic Information



Reproducibility review of: Building Change Detection of Airborne Laser Scanning and Dense Image Matching Point Clouds using Height and Class Information

Philipp A. Friese

2021-06-07

REPRODUCIBLE AGILE

This report is part of the reproducibility review at the AGILE conference. For more information see https://reproducible-agile.github.io/. This document is published on OSF at https://osf.io/RSF4M/. To cite the report use

Friese, Philipp A. (2021, May). Reproducibility review of: Building Change Detection of Airborne Laser Scanning and Dense Image Matching Point Clouds using Height and Class Information. https://doi.org/10.17605/OSF.IO/RSF4M

Reviewed paper

Politz, F., Sester, M., and Brenner, C.: Building Change Detection of Airborne Laser Scanning and Dense Image Matching Point Clouds using Height and Class Information, AGILE GIScience Ser., 2, 10, https://doi.org/10.5194/agile-giss-2-10-2021, 2021.

2017-Present

Machine learning

ML Reproducibility Challenge 2022 Edition

for papers published in:

NEURAL INFORMATION PROCESSING SYSTEMS · ICML | 2022 Thirty-ninth International Conference on Machine Learning · ICLR · ECCV TEL AVIV 2022 · ACL 2022 · EMNLP 2022 · CVPR NEW ORLEANS · PODC 서울 2022 · TRANSACTIONS tmlr ON ML RESEARCH · JMLR · NAACL 2022 · SIGIR 2022 · Transactions of the Association for Computational Linguistics · IJCAI ECAI VIENNA 22

RESCIENCE C

2018-Present

The Challenge

Climate Informatics Reproducibility Challenge

- Launched at the CI2023 Reproducibility Panel (21 April)
- 24 participants registered from America, Europe and Asia
- Challenge dates: 1 May to 31 May
- Goal: reproduce a paper, pre-approved from the Environmental Data Science journal or participant-suggested
- Open review through notebooks hosted by the Environmental Data Science book, which are citable and DOI-tagged
- Prizes: Vouchers in Cambridge University Press books
- Communication channels: CI2023 slack, GitHub, Zoom

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Challenge Organisers

EDS book
Environmental Data Science Book

C-SCALE
Cloud infrastructure

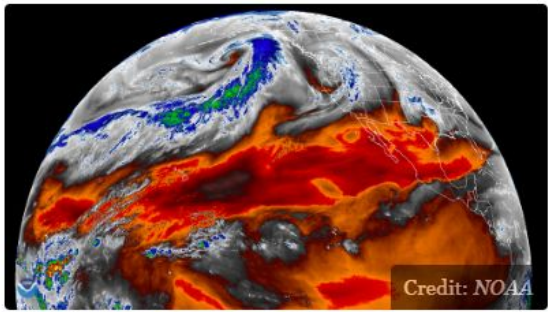
And lots of fun!
Read more

Organising Committee:

Alejandro Coca Castro
(The Alan Turing Institute)
Andrew McDonald
(University of Cambridge & British Antarctic Survey)
Andrew Hyde
(Cambridge University Press & Assessments)
Anne Fouilloux
(Simula Research Lab)

Climate Informatics

An open community at the intersection of climate science and data science.



Climate Centered

Climate Informatics is centered around the discovery of climate sciences with state-of-art methods in statistics, machine learning, and data mining.

Community Driven

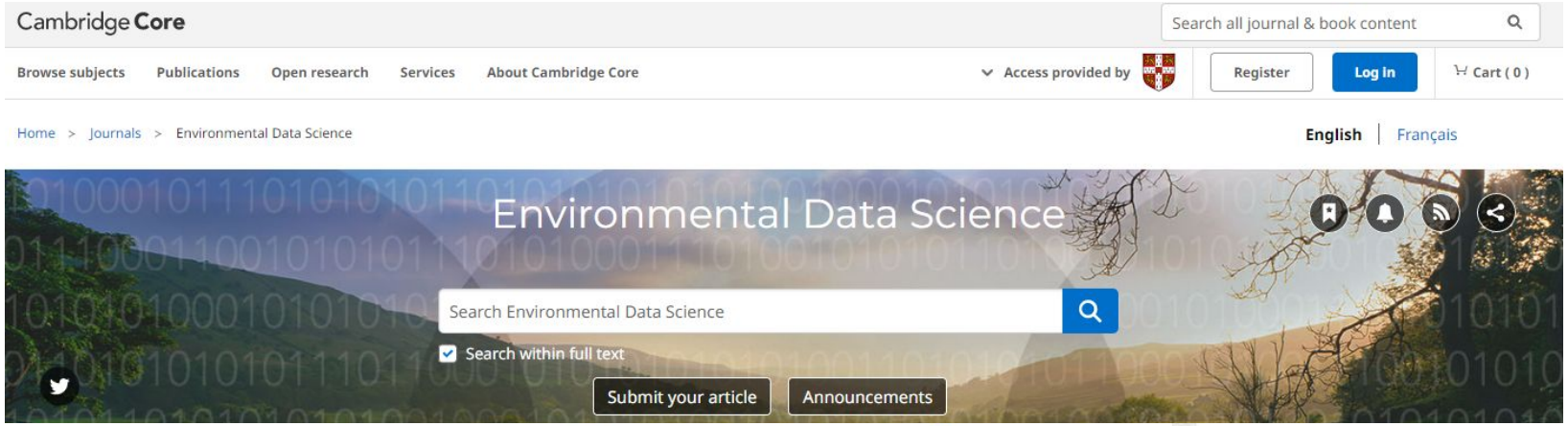
The community is driven by active members organizing annual conference series, hackathons, and other events to promote advancement in the field.

Collaborative Innovation

We promote collaborations and support early-career scientists by bringing experts across disciplines together through various events and online forums.

http://www.climateinformatics.org/

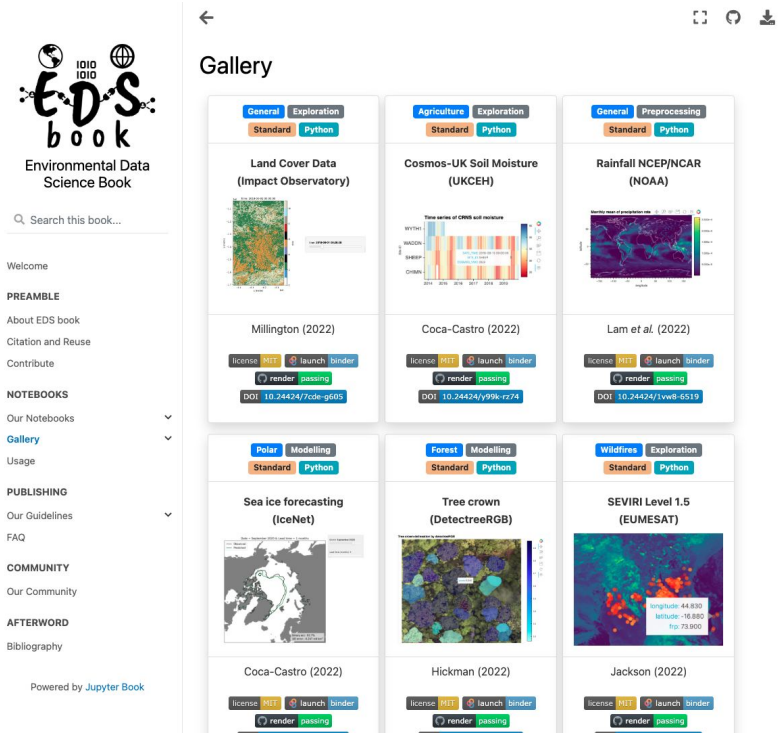
@Climformatics



https://www.cambridge.org/core/journals/environmental-data-science

@envdatascience

Environmental Data Science book



A computational notebook community for open environmental data science 🌎

A **living**, **free** and **open** online resource to **showcase** and **support** the **publication** of data, research and open-source tools for **collaborative**, **reproducible** and **transparent** **Environmental Data Science**

http://www.edsbook.org/

@eds_book

Outcomes

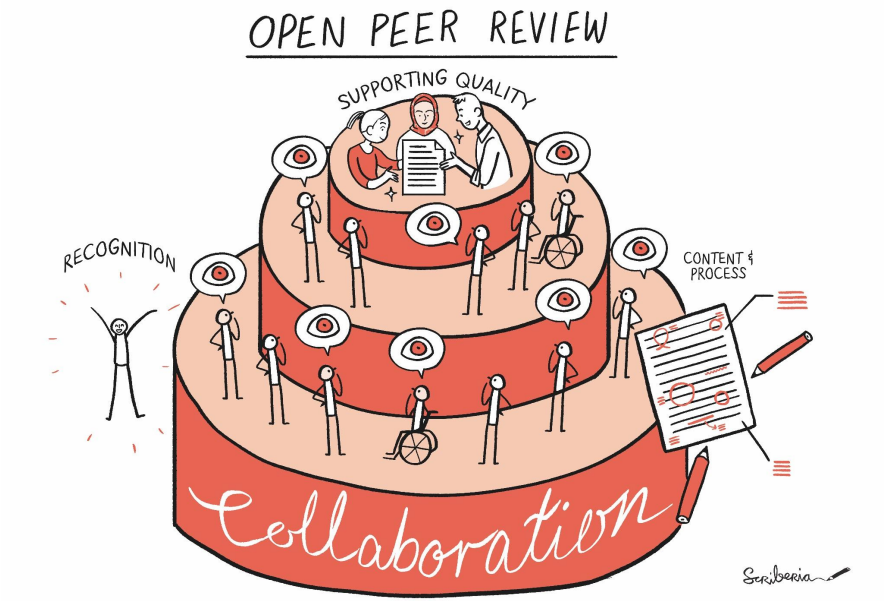
How the challenge ended...

Participation (completed)

- 3 teams (8 participants) submitted
- Vouchers of £500 (winning team) & £150 (runner-up teams)
- 9 reviewers

Support activities

- Onboarding sessions for participants and reviewers
- 4 Clinics and Coffee drop-in sessions
- 6 Talk with the Experts (Open infrastructure, Scientific Software and Open Science)
- Share-out event

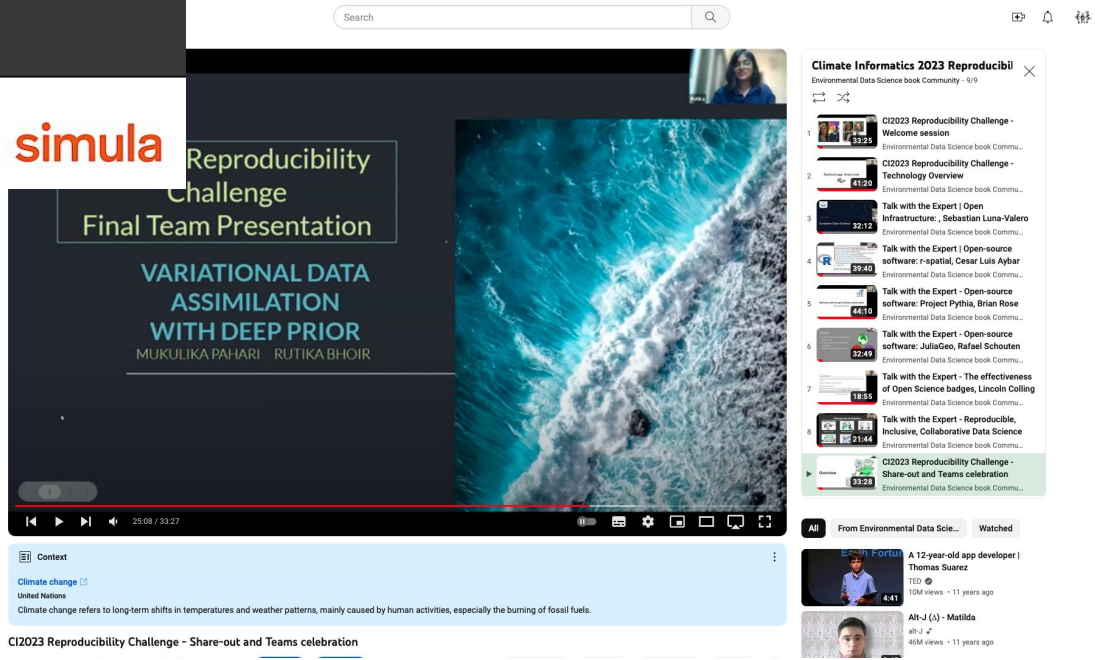


Slides and recordings
available in
the CIRC23 website

Onboarding, Talk with Experts
and Share-out videos available at:

EDS book YouTube Channel

youtube.com/@eds_book



Bottlenecks mentioned in the share-out

Paper

- Terminology
- Part of the methodology was a bit complicated to understand

Code

- Poor documentation
- Code contains additional information other than the paper

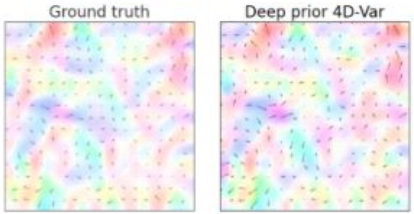
Data

- Large dataset - time consuming to download (52 GB)
    - Zarr-optimised format in the published EDS book notebook
- The arrays created in the code are extremely huge and often cause memory issues. Require large computational resources to train the models (~100 GB).

CIRC23 submissions are online in EDS book!

Ocean Modelling Special Issue Python

Variational data assimilation with deep prior (CIRC23)

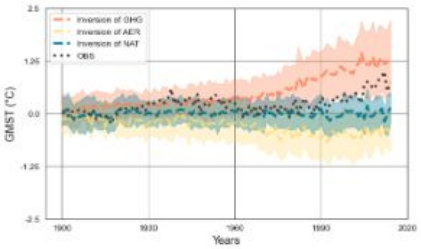


Pahari *et al.* (2023)

license MIT launch binder
render passing
DOI 10.5281/zenodo.8339299

General Modelling Special Issue Python

Deep learning and variational inversion for climate science (CIRC23)

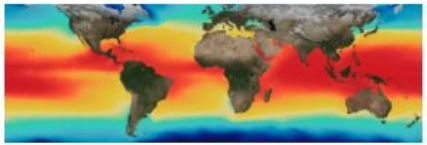


Domazetoski *et al.* (2023)

license MIT launch binder
render passing
DOI 10.5281/zenodo.8330771

Ocean Modelling Special Issue Python

Underlying physics of the ocean's temperature (CIRC23)

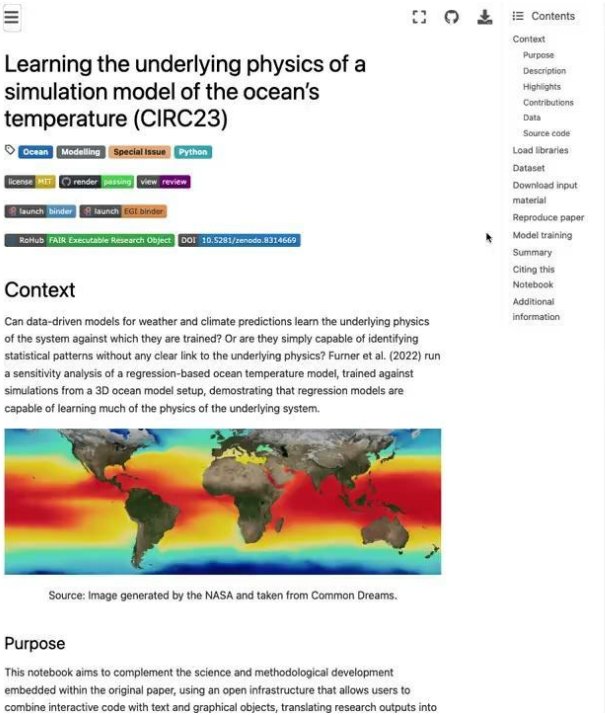


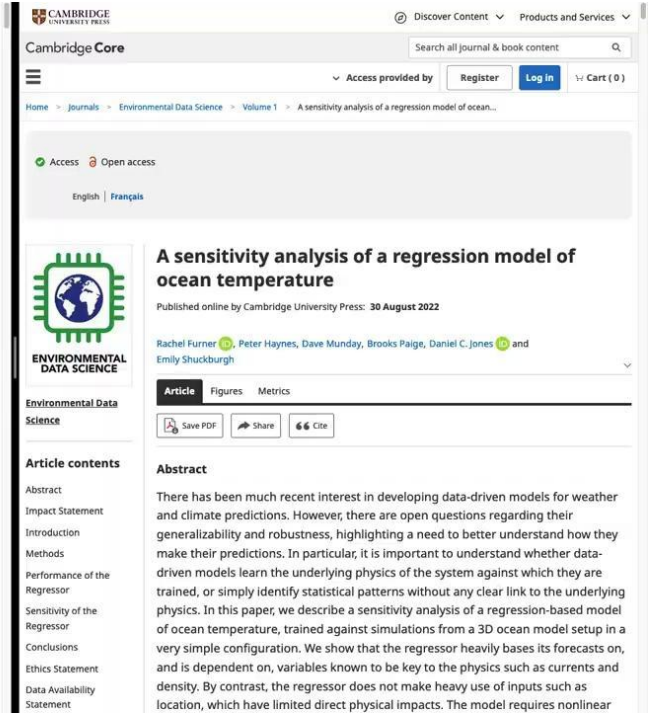
Malhotra *et al.* (2023)

license MIT launch binder
render passing
DOI 10.5281/zenodo.8314669

https://edsbook.org/notebooks/gallery/

Winning Team (America)





Team (US-based):

Garima Malhotra, University of Colorado Boulder

Daniela Pinto Veizaga, University of California Berkeley

Jorge Eduardo Peña Velasco, Claremont McKenna College

Reviewers:

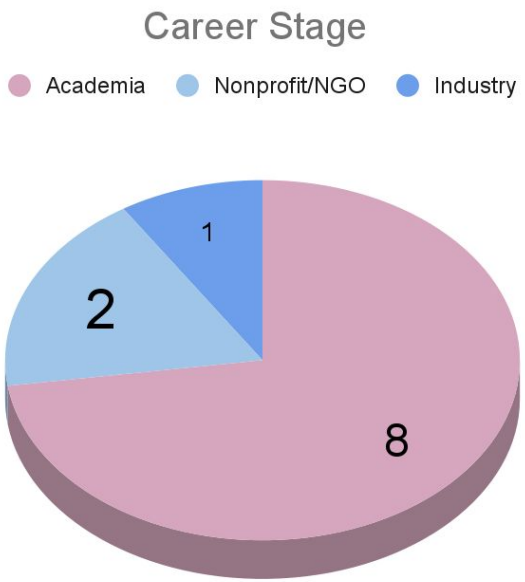
Rachel Furner, University of Cambridge, UK

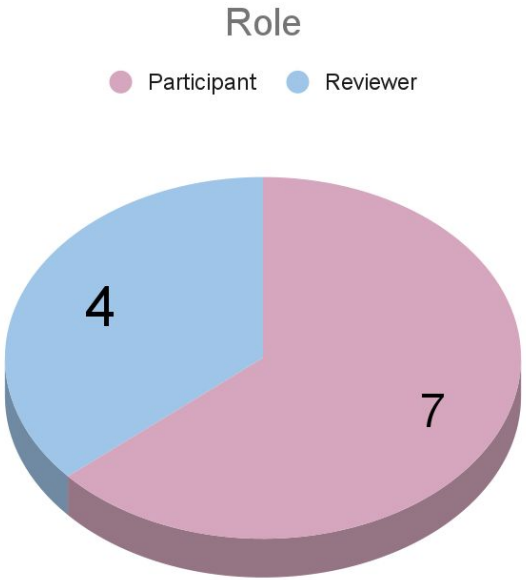
Oscar Bautista, World Food Programme, Italy

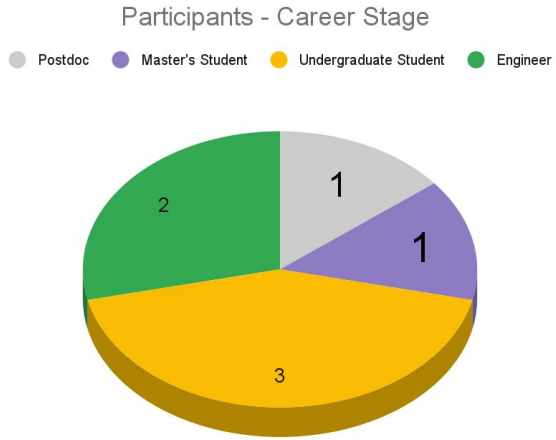
Ricardo Barros Lourenço, McMaster University, Canada

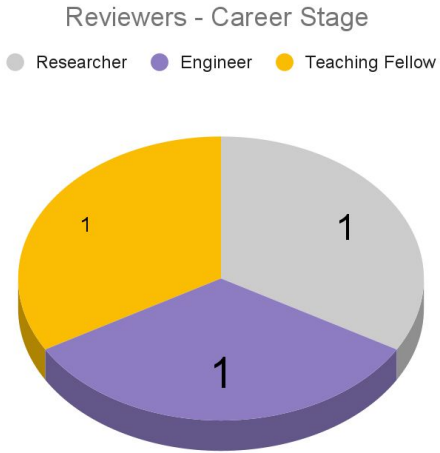
https://edsbook.org/notebooks/gallery/3286b92f-4fae-4cc6-a29e-e408bc844542/notebook

Feedback



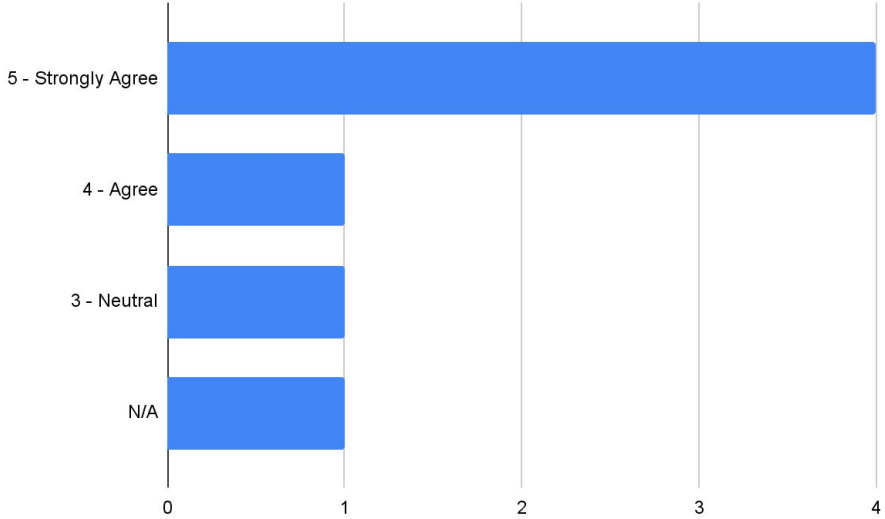




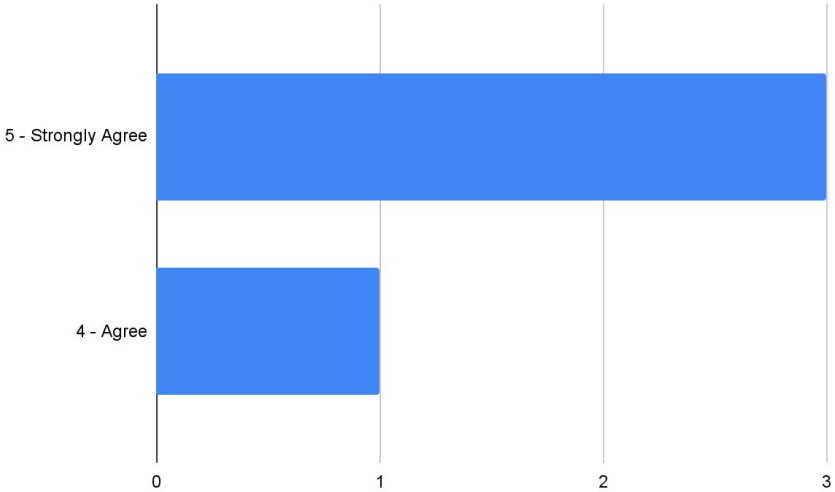


Commitment:
12 to 40 hours (participants)
5 to 10 hours (reviewers)

Overall Satisfaction



Participants



Reviewers

CIRC23 Feedback - Participants

“I felt that there was a lot of effort put into the creation of the challenge and that people were always **willing to help**. Furthermore, everything was so **well documented** that even before beginning the challenge I knew exactly how/what needed to be done.”

“The challenge was very **well organized** and the communication was very clear and prompt - by both email and slack.”

“Overall, it was a **great experience**. Thank you and kudos for running the challenge so smoothly.”

CIRC23 Feedback - Recommendations

Participants

“It would be great if the organizers can get more information from the authors about **how long does the training** of the model takes, and **how much memory** it requires.”

“I think the **goals of the reproducibility challenge** can be conveyed a little better. In the beginning I had a hard time understanding **why re-running** someone else's code was seen as a **scientific contribution**.”

Reviewers

“Maybe **more time** between the informative sessions and the review so there is a chance to get a practice review.”

“A little **more interactive** with the hackathon teams.”

Reflections & Next Steps

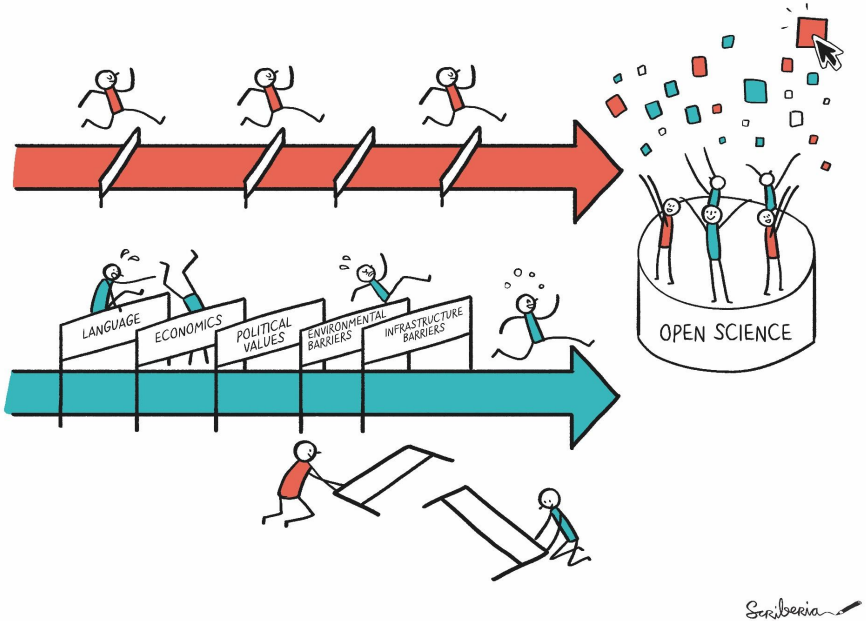
Reflections

Pluses

- Accessibility, Worldwide
- Innovative format, Interactive Reports, Everyone can try!
- Collaborations (Pangeo/C-Scale) and Engagement (Guest speakers)

Deltas

- Format
- Timing
- Publicity



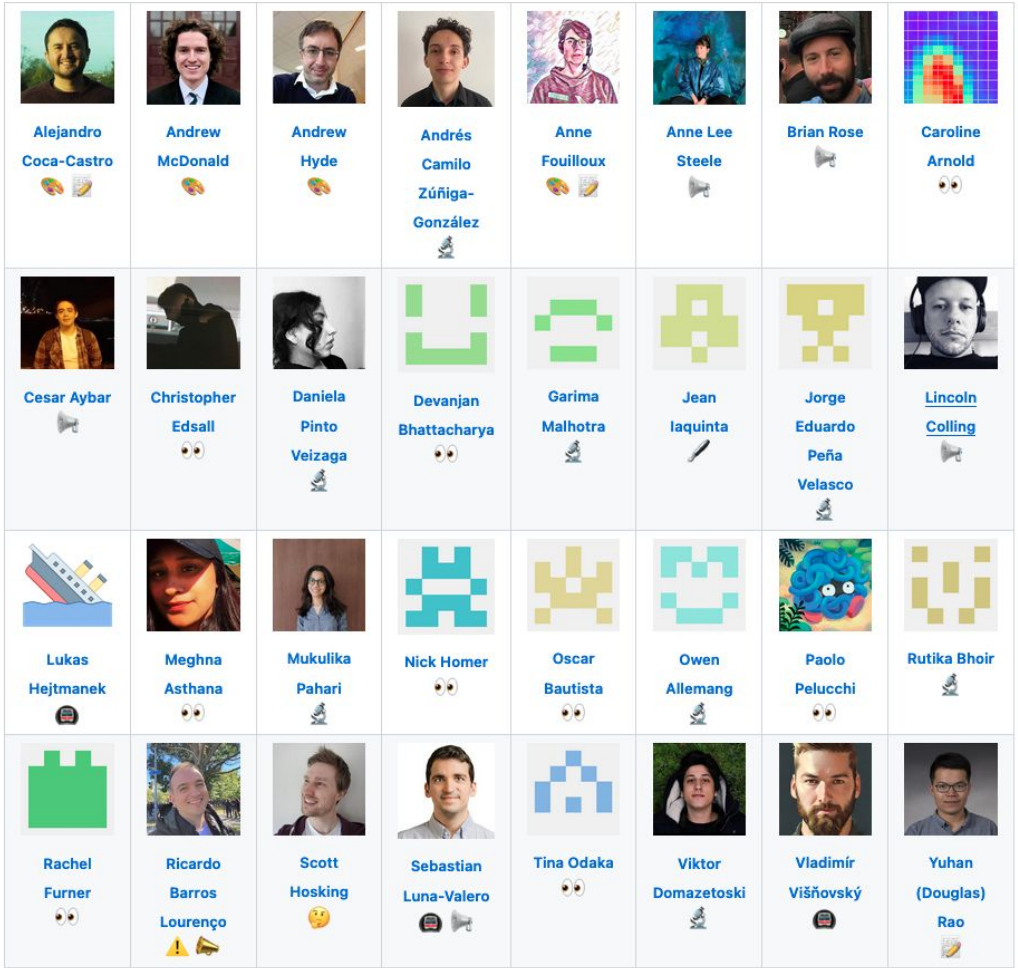
Next Steps

Short-term

- Prepare a perspective paper for EDS journal
- Produce guidelines (inspired by AGILE reproducibility)

Long-term

- Integrate the CIRC format into a portfolio of reproducibility activities for future Climate Informatics conferences
- Support other communities targeting innovative and interactive formats for reproducibility challenges



Thanks goes to these wonderful people

Participants 🔬

Reviewers 👀

Judges 📝

Guest Speakers 📢

Infrastructure 🚇

Organisers 🎨

Helpers 🖋️ ⚠️ 🤔 📣

Supported by:

The Alan Turing Institute · C-SCALE · University of Cambridge · simula

Thank YOU