

Climate Informatics[^] 2023

Reproducibility Challenge

~~April 19-21st~~ 1-31 May

~~University of Cambridge, UK~~ Everywhere Online



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**

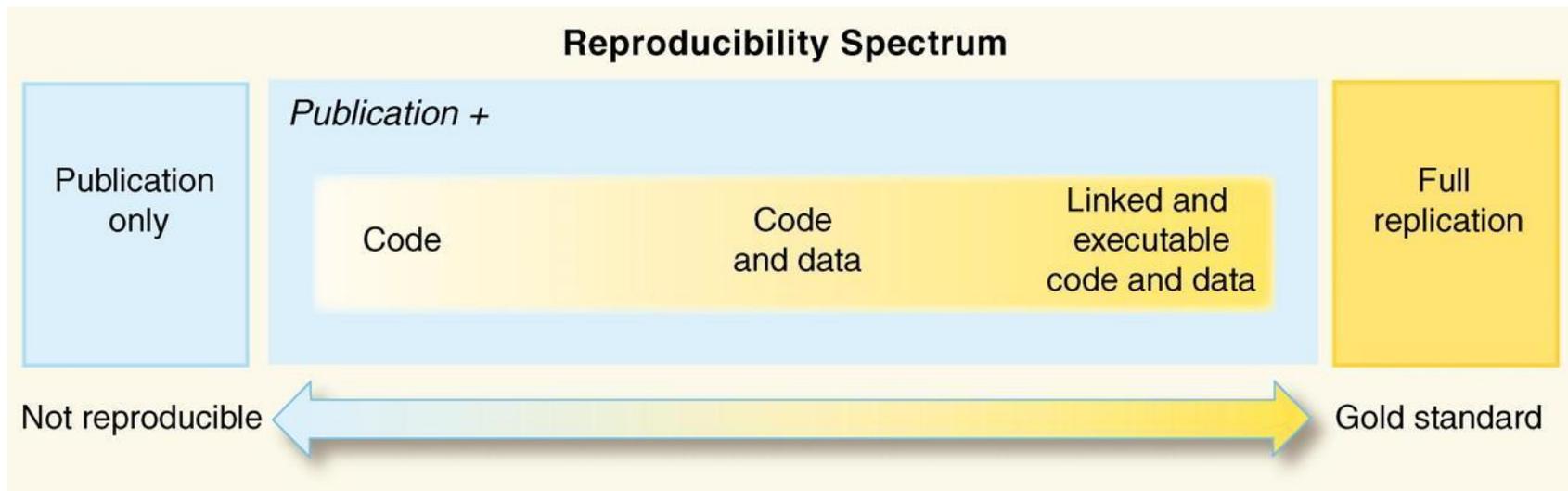
Reproducible Research

same analysis steps
on the same dataset
produces same
answer

		Data	
		Same	Different
Analysis	Same	(Reproducible)	Replicable
	Different	Robust	Generalisable



The Reproducibility Spectrum



Reproducibility Initiatives

Supercomputing

AD/AE Appendix Process & Badges



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



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:



RESCIENCE C

2018-Present

The Challenge



Scriberia 

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

*And lots of fun!
Read more*

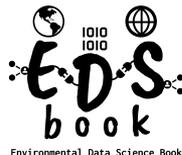


Organising Committee:

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)

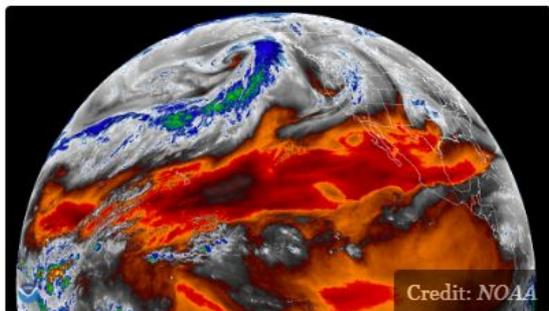


Challenge Organisers



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



Environmental Data Science

Search Environmental Data Science

Search within full text

Submit your article Announcements

✓ Access: Subscribed, Past subscription Open access



ISSN: 2634-4602 (Online)

Editor: Claire Monteleoni *University of Colorado Boulder, USA*

[Editorial board](#)

Environmental Data Science is an open access journal dedicated to the use of data-driven approaches to understand environmental processes - including climate change - and aid sustainable decision-making. The data and methodological scope is defined broadly to encompass artificial intelligence, machine learning, data mining, computer vision, econometrics and other statistical techniques.

EDS is a venue for application and methods papers, whether they relate to the geosphere (the solid earth and its processes), cryosphere (e.g. ice, snow, permafrost and tundra), biosphere (ecology), hydrosphere (oceans and fresh water, including the water cycle) or atmosphere (e.g. meteorology, climatology). It also welcomes work that shows how data science can inform societal responses to environmental problems (such as climate change, air quality, energy, natural resources and land use).

EDS promotes open data and data re-use - through data papers that describe valuable environmental data sets - and publishes shorter position papers relevant to the journal's scope.

Open access

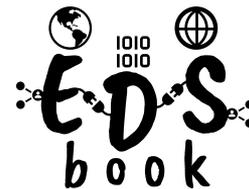
View the
editorial
board



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

@envdatascience

Environmental Data Science book



Environmental Data Science Book

Search this book...

Welcome

PREAMBLE

About EDS book

Citation and Reuse

Contribute

NOTEBOOKS

Our Notebooks

Gallery

Usage

PUBLISHING

Our Guidelines

FAQ

COMMUNITY

Our Community

AFTERWORD

Bibliography

Powered by Jupyter Book

Gallery

<p>General Exploration Standard Python</p> <p>Land Cover Data (Impact Observatory)</p>	<p>Agriculture Exploration Standard Python</p> <p>Cosmos-UK Soil Moisture (UKCEH)</p>	<p>General Preprocessing Standard Python</p> <p>Rainfall NCEP/NCAR (NOAA)</p>
<p>Millington (2022)</p> <p>license MIT launch binder render passing DOI: 10.24424/7cde-g6d5</p>	<p>Coca-Castro (2022)</p> <p>license MIT launch binder render passing DOI: 10.24424/y99k-rr74</p>	<p>Lam et al. (2022)</p> <p>license MIT launch binder render passing DOI: 10.24424/1vw8-6519</p>
<p>Polar Modelling Standard Python</p> <p>Sea ice forecasting (IceNet)</p>	<p>Forest Modelling Standard Python</p> <p>Tree crown (DetectreeRGB)</p>	<p>Wildfires Exploration Standard Python</p> <p>SEVIRI Level 1.5 (EUMESAT)</p> <p>HighLow: 44,830 HighLow: -18,800-300 High: 78,900</p>
<p>Coca-Castro (2022)</p> <p>license MIT launch binder render passing</p>	<p>Hickman (2022)</p> <p>license MIT launch binder render passing</p>	<p>Jackson (2022)</p> <p>license MIT launch binder render passing</p>

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

@ampersandmcd, CC-BY 4.0, DOI: 10.5281/zenodo.10376130

Outcomes

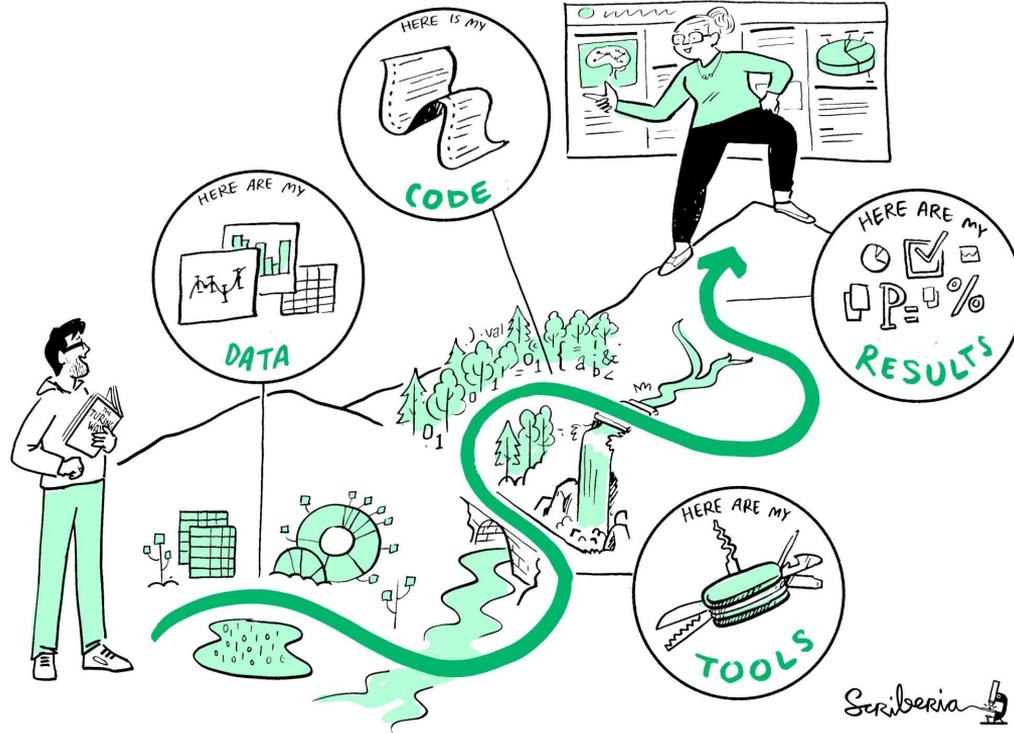


Image by Scriberia for The Turing Way, Shared under CC-BY 4.0, @turingway

@ampersandmcd, CC-BY 4.0, DOI: 10.5281/zenodo.10376130

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



Climate Informatics 2023 Reproducibility Challenge

Share-out and Teams celebration
15 June 2023 15:30-16:00 UTC+1

[Register now](#)

Hosted by:



Supported by:

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

EDS book YouTube Channel

youtube.com/@eds_book

Slides and recordings
available in
the CIRC23 website



Search

Reproducibility
Challenge
Final Team Presentation

VARIATIONAL DATA
ASSIMILATION
WITH DEEP PRIOR

MUKULIKA PAHARI RUTIKA BHOIR

25:08 / 33:27

Context

Climate change [Climate change](#)

United Nations

Climate change refers to long-term shifts in temperatures and weather patterns, mainly caused by human activities, especially the burning of fossil fuels.

[CIRC2023 Reproducibility Challenge - Share-out and Teams celebration](#)

Climate Informatics 2023 Reproducibility Challenge

Environmental Data Science book Community · 919

- 1 [CIRC2023 Reproducibility Challenge - Welcome session](#)
Environmental Data Science book Community · 13:25
- 2 [CIRC2023 Reproducibility Challenge - Technology Overview](#)
Environmental Data Science book Community · 41:20
- 3 [Talk with the Expert | Open Infrastructure, Sebastian Luna-Valero](#)
Environmental Data Science book Community · 22:12
- 4 [Talk with the Expert | Open-source software: r-spatial, Cesar Luis Aybar](#)
Environmental Data Science book Community · 32:40
- 5 [Talk with the Expert - Open-source software: Project Pythia, Brian Rose](#)
Environmental Data Science book Community · 44:10
- 6 [Talk with the Expert - Open-source software: JuliaGeo, Rafael Schouten](#)
Environmental Data Science book Community · 32:49
- 7 [Talk with the Expert - The effectiveness of Open Science badges, Lincoln Colling](#)
Environmental Data Science book Community · 18:55
- 8 [Talk with the Expert - Reproducible, Inclusive, Collaborative Data Science](#)
Environmental Data Science book Community · 21:44

[CIRC2023 Reproducibility Challenge - Share-out and Teams celebration](#)
Environmental Data Science book Community · 33:28

All From Environmental Data Science... Watched

[A 12-year-old app developer!](#)
TED · Thomas Suarez · 4:41 · 10M views · 11 years ago

[Al-J \(J\) - Matilda](#)
al-J · 46M views · 11 years ago

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

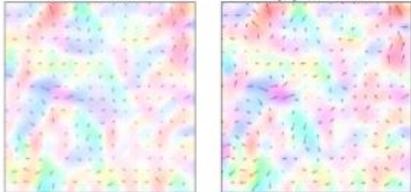
- Large dataset - time consuming to download (52 GB)
 - Zarr-optimised format in the published EDS book notebook
- The arrays created in the code are extremely huge and often cause memory issues. Require large computational resources to train the models (~100 GB).

CIRC23 submissions are online in EDS book!

Ocean **Modelling**
Special Issue **Python**

Variational data assimilation with deep prior (CIRC23)

Ground truth Deep prior 4D-Var



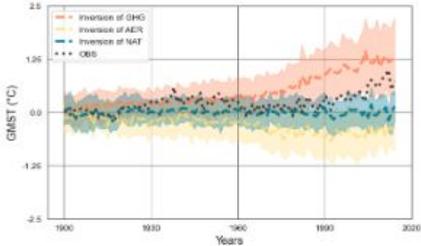
Pahari *et al.* (2023)

license MIT launch binder
render passing

DOI [10.5281/zenodo.8339299](https://doi.org/10.5281/zenodo.8339299)

General **Modelling**
Special Issue **Python**

Deep learning and variational inversion for climate science (CIRC23)



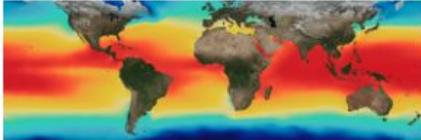
Domazetoski *et al.* (2023)

license MIT launch binder
render passing

DOI [10.5281/zenodo.8330771](https://doi.org/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://doi.org/10.5281/zenodo.8314669)

Winning Team (America)

 **Team (US-based):**

Garima Malhotra,

University of Colorado Boulder

Daniela Pinto Veizaga, University
of California Berkeley

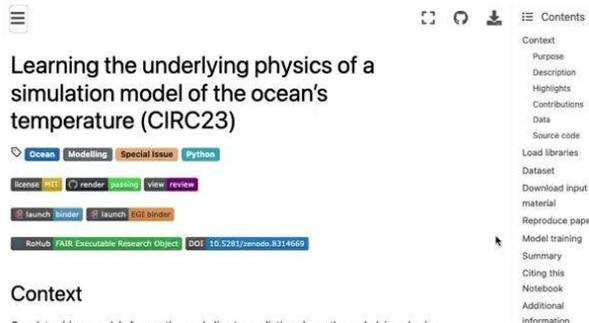
Jorge Eduardo Peña Velasco,
Claremont McKenna College

 **Reviewers:**

Rachel Furner, University
of Cambridge, UK

Oscar Bautista, World
Food Programme, Italy

Ricardo Barros Lourenço,
McMaster University,
Canada



Contents

- Context
- Purpose
- Description
- Highlights
- Contributions
- Data
- Source code
- Load libraries
- Dataset
- Download input material
- Reproduce paper
- Model training
- Summary
- Citing this Notebook
- Additional information

Ocean Modelling Special Issue Python

license MIT render jupyterlab view review

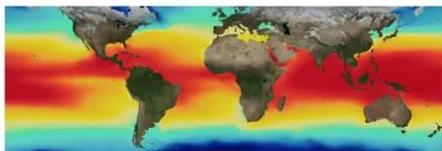
launch binder launch EOL binder

GitHub FAIR Executable Research Object DOI 10.5281/zenodo.8314669

Learning the underlying physics of a simulation model of the ocean's temperature (CIRC23)

Context

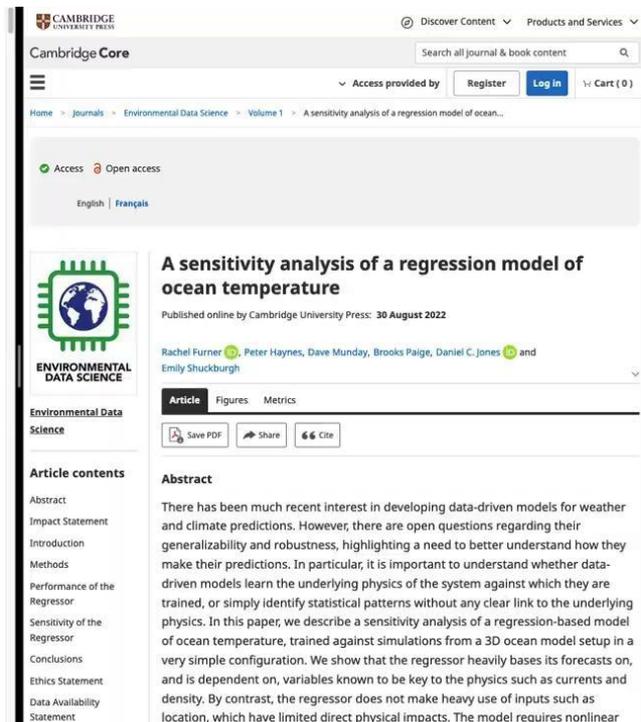
Can data-driven models for weather and climate predictions learn the underlying physics of the system against which they are trained? Or are they simply capable of identifying statistical patterns without any clear link to the underlying physics? Furner et al. (2022) run a sensitivity analysis of a regression-based ocean temperature model, trained against simulations from a 3D ocean model setup, demonstrating that regression models are capable of learning much of the physics of the underlying system.



Source: Image generated by the NASA and taken from Common Dreams.

Purpose

This notebook aims to complement the science and methodological development embedded within the original paper, using an open infrastructure that allows users to combine interactive code with text and graphical objects, translating research outputs into



Cambridge Core

Discover Content Products and Services

Search all journal & book content

Access provided by Register Log In Cart (0)

Home Journals Environmental Data Science Volume 1 A sensitivity analysis of a regression model of ocean...

Access Open access

English Français

A sensitivity analysis of a regression model of ocean temperature

Published online by Cambridge University Press: 30 August 2022

Rachel Furner, Peter Haynes, Dave Munday, Brooks Paige, Daniel C. Jones and Emily Shuckburgh

Article Figures Metrics

Save PDF Share Cite

Article contents

- Abstract
- Impact Statement
- Introduction
- Methods
- Performance of the Regressor
- Sensitivity of the Regressor
- Conclusions
- Ethics Statement
- Data Availability
- Statement

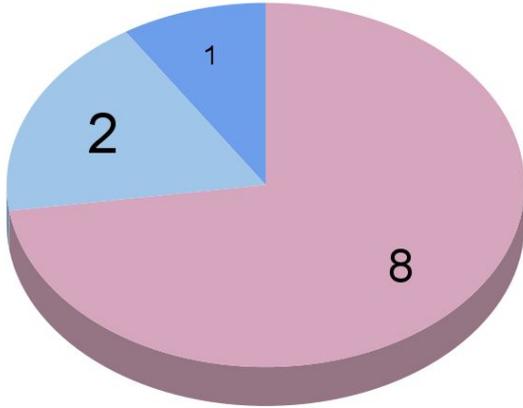
Abstract

There has been much recent interest in developing data-driven models for weather and climate predictions. However, there are open questions regarding their generalizability and robustness, highlighting a need to better understand how they make their predictions. In particular, it is important to understand whether data-driven models learn the underlying physics of the system against which they are trained, or simply identify statistical patterns without any clear link to the underlying physics. In this paper, we describe a sensitivity analysis of a regression-based model of ocean temperature, trained against simulations from a 3D ocean model setup in a very simple configuration. We show that the regressor heavily bases its forecasts on, and is dependent on, variables known to be key to the physics such as currents and density. By contrast, the regressor does not make heavy use of inputs such as location, which have limited direct physical impacts. The model requires nonlinear

Feedback

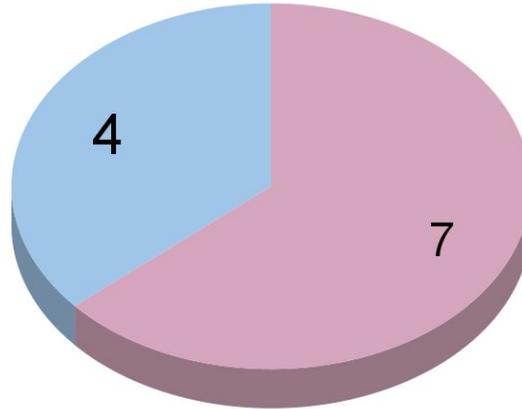
Career Stage

● Academia ● Nonprofit/NGO ● Industry



Role

● Participant ● Reviewer



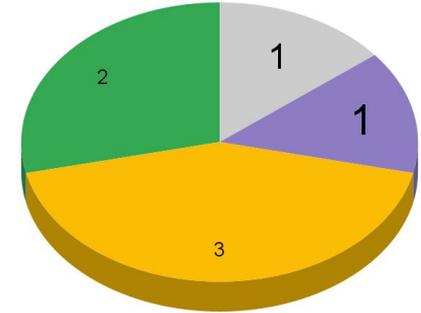
Commitment:

12 to 40 hours (participants)

5 to 10 hours (reviewers)

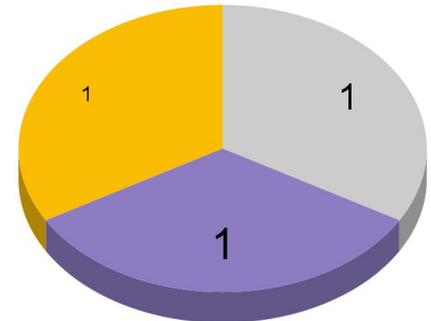
Participants - Career Stage

● Postdoc ● Master's Student ● Undergraduate Student ● Engineer

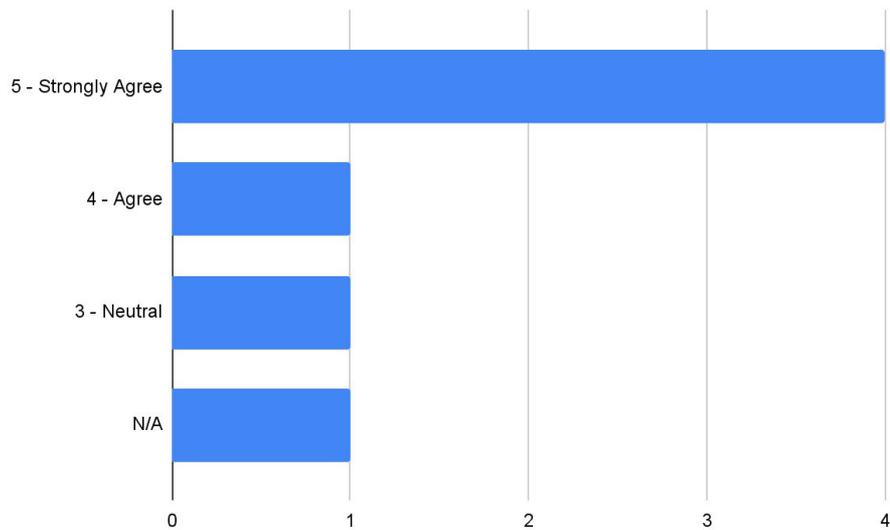


Reviewers - Career Stage

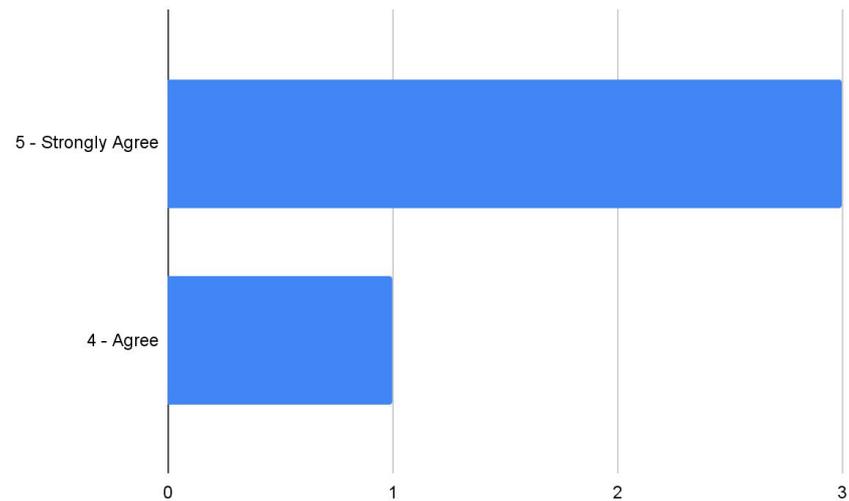
● Researcher ● Engineer ● Teaching Fellow



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.”

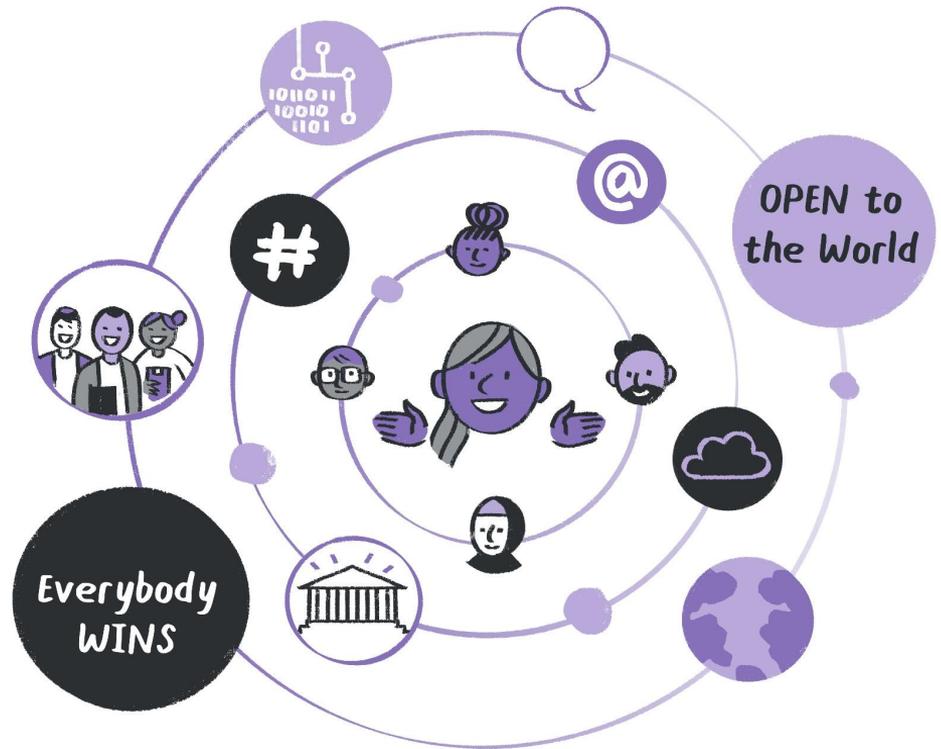
“I think the **goals of the reproducibility challenge** can be conveyed a little better. In the beginning I had a hard time understanding **why re-running** someone else's code was seen as a **scientific contribution.**”

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



COLLABORATING NOT COMPETING

Scriberia 

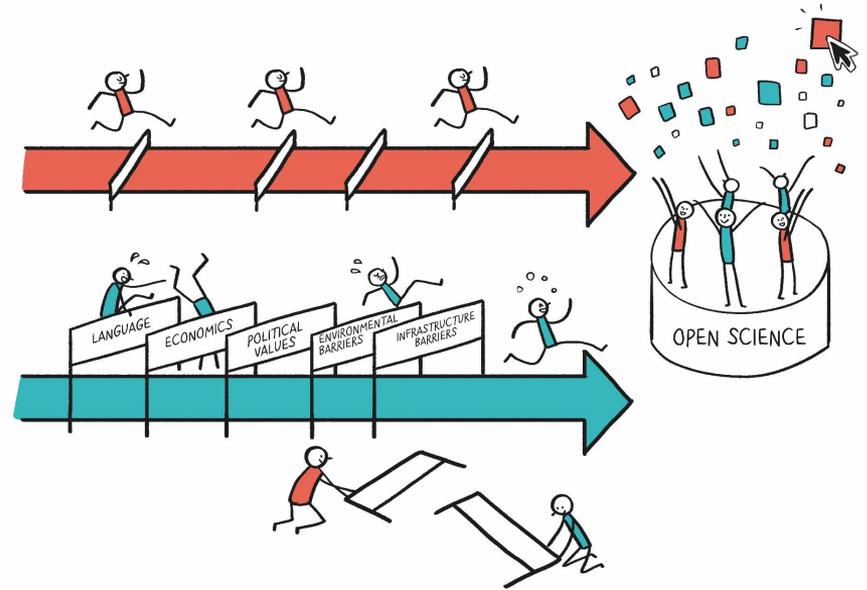
Reflections

Pluses

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

Deltas

- Format
- Timing
- Publicity



Scriberia

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

 Alejandro Coca-Castro 	 Andrew McDonald 	 Andrew Hyde 	 Andrés Camilo Zúñiga- González 	 Anne Fouilloux 	 Anne Lee Steele 	 Brian Rose 	 Caroline Arnold 
 Cesar Aybar 	 Christopher Edsall 	 Daniela Pinto Veizaga 	 Devanjan Bhattacharya 	 Garima Malhotra 	 Jean laquinta 	 Jorge Eduardo Peña Velasco 	 Lincoln Colling 
 Lukas Hejtmanek 	 Meghna Asthana 	 Mukulika Pahari 	 Nick Homer 	 Oscar Bautista 	 Owen Allemang 	 Paolo Pelucchi 	 Rutika Bhoir 
 Rachel Furner 	 Ricardo Barros Lourenço 	 Scott Hosking 	 Sebastian Luna-Valero 	 Tina Odaka 	 Viktor Domazetoski 	 Vladimír Višňovský 	 Yuhan (Douglas) Rao 

Thanks goes to these wonderful people

Participants 🧪

Reviewers 👁️

Judges 📝

Guest Speakers 🗣️

Infrastructure 🌐

Organisers 🧑‍🤝‍🧑

Helpers 🗑️ ⚠️ 😞 🚫

Supported by:

The Alan Turing
Institute



simula

Thank YOU

