Ethics in climate AI: From theory to practice

Viviana Acquaviva , Elizabeth A. Barnes, David John Gagne II, Galen A. McKinley, Savannah Thais

Published: August 2, 2024 • https://doi.org/10.1371/journal.pclm.0000465

Citation: Acquaviva V, Barnes EA, Gagne DJ II, McKinley GA, Thais S (2024) Ethics in climate AI: From theory to practice. PLOS Clim 3(8): e0000465. https://doi.org/10.1371/journal.pclm.0000465

Editor: Jamie Males, PLOS Climate, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Published: August 2, 2024

Funding: VA acknowledges support from a PIVOT fellowship grant of the Simons Foundation (Award #981849). VA and GAM acknowledge funding from the National Science Foundation through the LEAP STC (Award #2019625). DJG acknowledges support from the National Science Foundation under Grant No. ICER-2019758 and Cooperative Agreement AGS-1852977. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Climate science, and climate artificial intelligence (AI) in particular, cannot be disconnected from ethical societal issues, such as resource access, conservation, and public health. An apparently apolitical choice—for example, treating all data points used to train an AI model equally—can result in models that are more accurate in regions where the density and quality of data is higher; these often coincide with the northern and western areas of the world (e.g., [1, 2]).

Inequity in the access to data and computational resources exacerbates gaps between communities in understanding climate change impacts and acting towards mitigation and adaptation, often in ways that are detrimental to those who are most affected (e.g., [3, 4]). While these issues are not exclusive to AI, widespread opacity in the development and functioning of AI models, presentation of AI model outcomes, and the rapid evolution of the AI field further increase the inequality in power and agency among differently resourced parties.

This creates an opportunity for climate scientists to rethink the role of ethics in their approach to research. There are many ways in which climate scientists can interact with society. Here we focus on the process of scientific research, identifying some good practices for building trustworthy and responsible models and then providing some resources.

In creating and training models, we encourage researchers to recognize that science cannot claim to be purely “objective”, and that the choice of priors, data, and metrics all carry biases (e.g., [5]). Resolving or eliminating them is not realistic, as the interpretation of a “better” model or result is highly dependent on the user’s specific goal. Hence, it is crucial to be open and specific about the assumptions made, the algorithms and hyperparameters used, and the evaluation metrics and processes, and ideally to also make data and code available, following the principles of reproducible science (e.g., [6]).

In evaluating and presenting the performance of statistical or machine learning models, considering possible failure modes can be a useful lens through which to examine systems’ behaviors. Good starting points are the taxonomy proposed by [7], which considers flaws in design, implementation, and communication, and, specifically for climate science, the list compiled by [8]. Failure modes that are particularly relevant for AI models from the climate domain include robustness under distribution shifts; for example, for models trained on historical data, it is difficult to predict how they will perform in the unseen conditions brought about by climate change. Good practices induced by consideration of failure modes may include efforts to quantify expected model performance through simulations, transparency in defining the anticipated range of applicability of the model, and considering how existing physics/climate knowledge can be applied in defining an evaluation strategy and in validating the generalization properties of the model. A related practice is to consider how the functionality lens changes and what failure modes become relevant when looking at the whole “phase space” of a model, intended as the array of different breakdowns of data, such as geographical areas, time periods, or user groups, as well as the broader socio-technical context in which a model can be used beyond the use case for which it was developed.

In creating and distributing data sets, a term we use to indicate both observational data and other products such as outputs of Earth System Models, we begin by recognizing that we typically don’t hold power on who will use the data and how the data will be used. We can, however, aim to provide clear documentation on what dataset or numerical model creators thought of as desired and pertinent usage, as well as undesired, inappropriate, or wrong usage. Additionally, we can aim to be as transparent as possible in language and jargon in order to make climate information more widely accessible to a broad range of users and decision makers; see, e.g., [9] for example of potential pitfalls when datasets are used across domains. A possible pipeline to consider when creating data sets is provided by the Datasheets for Earth Science Datasets [10].

When possible, we propose to prefer AI models that are intrinsically interpretable (sometimes defined as those built from pieces whose behavior is interpretable, such as decision trees or linear models) or explainable (sometimes defined as those whose behavior and decisions can be explained *post hoc*, for example, by means of a simpler surrogate model; see e.g., [11]).

Interpretability and explainability can protect us from dangerous failure modes by creating insights into the ability of our model to generalize beyond the training domain, allowing us to check whether the explanation is consistent with our understanding of a physical system, and by allowing access to a wider and more inclusive user base, for example those who are not fully trained in or don't have the resources to run more complex models, such as deep learning methods. However, applying explainability tools or using interpretable methods is not always possible or practical. We would like to propose a wider interpretation of explainability, which includes interrogating the system by exploring its responses to a wide range of inputs, including distribution shifts and adversarial attacks. In addition, we can perform physically informed perturbations of the system that act on the spatial, temporal or phenomenological scales of interest to assess their impact. By studying the behavior of AI systems under varied circumstances, we can gain a more comprehensive understanding that is a proxy for explainability.

An important reflection on encouraging the use of broadly-defined explainable models, as well as the need for clear documentation mentioned above, is that creating well-documented and explainable models can be a much slower process than using potentially more powerful, but black-box, methods. Generating physically meaningful features that can improve performance, implementing frameworks to enhance explainability, and writing documentation are time-consuming processes.

Hence, rewarding the creation of such models requires a culture shift. Our current incentive structure is not built to reward this approach. Most often, the threshold for publication is set by an improvement in model performance typically measured by model accuracy on a pre-chosen data set. Furthermore, the measure of academic "success" (and in return, the chances to secure academic positions or grant funding, or to win awards) is heavily influenced by the number of papers and conference papers, creating pressure to publish more but not necessarily to curate impact. We can, however, challenge and change the status quo, by leading by example in our research, fostering a culture of explainable work in our labs, and appropriately rewarding well-documented, explainable, robust, and broadly impactful work when evaluating papers, proposals, job applications, and nominations for awards.

Besides being ethical, this culture shift in climate data science is made crucial by the fact that many private companies, including Google Deep Mind, NVIDIA, and Microsoft, are now racing ahead of academic centers in building complex machine learning-based climate and weather models (e.g., [12]). With very few exceptions, the computational resources available to universities won't match those offered by tech giants, but we can and must retain the important role of creating generalizable insights and providing, with each of our papers and products, a foundation for others to improve on our work. If we collaborate or work for these private companies, we should also remain aware of the potential for them to limit access to the information we produce.

Finally, we would like to emphasize that ethical AI and science are intertwined, interdependent topics. We should begin to include ethics in the climate science discourse, starting in the classroom. It is important to have these conversations in undergraduate and graduate courses and to integrate relevant resources in the climate science curricula, as well as at climate science conferences through special sessions.

Starting points for such conversations include:

- The lectures of the Trustworthy Artificial Intelligence for Environmental Science 2022 Summer School [13];
- AI Ethics Education for Scientists [14];
- AI Ethics and Fundamental Physics [15].

The benefits of applying principles of ethical AI in scientific inquiry and to reward explainability, robustness, and trustworthiness are many: making more information available to more people, empowering communities affected by climate change, making climate research more impactful and accessible, and increasing the level of interaction and trust between climate scientists and society. This comes at the price of a culture shift where scientists are called to build models and data sets according to certain principles, to change the system of incentives in order to reward slower but more robust, explainable, and transparent work, and to engage in discussing and teaching the ethical aspects of climate data science as early as possible.

Acknowledgments

We are thankful to the organizers of the Climate Informatics 2024 conference, where many of these ideas were originally discussed, and to our collaborators within the LEAP, M^{2}LInES, and AI2ES teams.

References

1. Kull DW, Riishojgaard LP, Eyre J, Varley RA. The Value of Surface-based Meteorological Observation Data. World Bank Group. 2021. Available from: http://documents.worldbank.org/curated/en/192461614151036836/The-Value-of-Surface-based-Meteorological-Observation-Data

2. Gloege L, McKinley GA, Landschützer P, Fay AR, Frolicher TL, Fyfe JC et al. Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability. Global Biogeochem. Cycles 2021, 1–14.
View Article • Google Scholar

3. UNEP Copenhagen Climate Centre, Technology Transfer for Climate Mitigation and Adaptation, Policy Brief, 2023. Available from: https://unepccc.org/wp-content/uploads/2023/06/tech-transfer-policy-brief-oecd.pdf

4. Balogun AL, Marks D, Sharma R, Shekhar H, Balmes C, Maheng D et al. Assessing the Potentials of Digitalization as a Tool for Climate Change Adaptation and Sustainable Development in Urban Centres. Sustainable Cities and Society 2019. 101888.

View Article • Google Scholar

5. Reiss J, Sprenger J. Scientific Objectivity. *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Zalta Edward N. (ed.). Available from: https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/
6. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie Du Sert N et al. A manifesto for reproducible science. *Nat Hum Behav* 2017. Available from: https://www.nature.com/articles/s41562-016-0021 pmid:33954258
View Article • PubMed/NCBI • Google Scholar
7. Raji ID, Kumar EI, Horowitz A, Selbst A. The Fallacy of AI Functionality. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). 2022. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158
8. McGovern A, Ebert-Uphoff I, Gagne DJ, Bostrom A. Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science.* 2022;1:e6.
View Article • Google Scholar
9. Auffhammer M, Hsiang S, Schlenker W, Sobel A. Using Weather Data and Climate Model Output in Economic Analyses of Climate Change. Review of Environmental Economics and Policy 2013.
View Article • Google Scholar
10. https://github.com/dmhuehol/Datasheets-for-Earth-Science-Datasets
11. Roscher R, Bohn B, Duarte MF, Garcke J, Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access 2020*, vol. 8, pp. 42200–42216, 2020,
View Article • Google Scholar
12. Price I, Sanchez-Gonzalez A, Alet F, Ewalds T, El-Kadi A, Stott J, et al. GenCast: Diffusion-based ensemble forecasting for medium-range weather. arXiv preprint arXiv:2312.15796. 2023 Dec 25.
View Article • Google Scholar
13. https://www.cisl.ucar.edu/events/tai4es-2022-summer-school
14. https://openreview.net/forum?id=3awrGYl7YD
15. https://indico.nikhef.nl/event/4875/sessions/1763/#20240502