PLOS Climate

OPINION

“The work of thought”–The machine learning revolution can be a revolution for our understanding of the Earth System

Joshua Oldham-Dorrington[1,2]*, **Julian Quinting**[3¤], **Stefan Sobolowski**[1,2]

1 Geophysical Institute (GFI), University of Bergen, Bergen, Norway, **2** Bjerknes Centre for Climate Research, Bergen, Norway, **3** Institute of Meteorology and Climate Research, Troposphere Research (IMKTRO), Karlsruhe Institute of Technology, Karlsruhe, Germany,

¤ Current Address: Institute of Geophysics and Meteorology, University of Cologne, Cologne, Germany.
* joshua.dorrington@uib.no

OPEN ACCESS

Citation: Oldham-Dorrington J, Quinting J, Sobolowski S (2025) “The work of thought”–The machine learning revolution can be a revolution for our understanding of the Earth System. PLOS Clim 4(9): e0000710. https://doi.org/10.1371/journal.pclm.0000710

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

Published: September 17, 2025

Funding: JOD received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101151904. JQ received funding from the European Research Council (ERC) under grant agreement no. 101077260 (ASPIRE). SS received funding from the European Union’s

Upon seeing one of the world’s first photographs in 1840 the French master of romantic realism, Paul Delaroche, declaimed that “Painting is dead!”. Instead, the following decades saw the emergence of impressionism—Monet’s waterlilies, Van Gogh’s starry night—and later, the rise of Picasso and cubism. Art, freed from the expectation of perfect reproduction, flourished. Later still, the fabulously detailed works of photo- and hyperrealism have demonstrated that ultimately there is no incompatibility between the worlds of canvas and film (or indeed of SD cards).

Earth system science is currently undergoing a revolution no less momentous than the invention of photography. Rapidly advancing machine learning (ML) methods now outperform traditional weather forecasts [1], and can accurately emulate km-scale rainfall [2], aerosol evolution [3] and ocean temperatures [4], all at a fraction of the traditional cost. Foundation models [5], aim to go further still: unifying Earth’s components and scales into a single framework. The increased efficiency of ML architectures is levelling the global playing field, allowing state-of-the-art trained models to be run with consumer hardware. There is little doubt that our field will look very different in the coming years.

Different is, of course, not necessarily better. Could a shift towards “data” come at the expense of genuine knowledge? By dedicating more time and training to statistics and software engineering, will the domain expertise and hard-won theory that underpin our fields wither away, harming our ability to reason causally and build hypotheses? We believe, in fact, that the opposite is the case.

The power of ML to predict and simulate, cheaply and flexibly, has enormous potential to support a boom in physical understanding, while ML’s fundamentally statistical basis *requires process-based understanding to critically validate.* There is therefore a continuous feedback loop between ML development and theory (Fig 1). Traditionally, data and theory are used to derive models; now we may also interrogate data-driven models to build theory –and evaluate model realism in the process. This opportunity requires community development of refined diagnostics to identify state- and process-dependent errors, suitable for large datasets. Such careful

Horizon 2020 research innovation programme under grant agreement no. 101081555 (IMPETUS4CHANGE).

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

process-based evaluation [e.g., 6] is key for ML systems to be trusted, as data-driven models can contain subtle cancelling errors and are not guaranteed to faithfully reflect known physics [7]. The transparently non-physical nature of ML models actually offers conceptual clarity compared to large, parameterised models that may be less 'physical' than they appear [8]. This may spur renewed pursuit of conceptual models for phenomena we can (somewhat) simulate but do not understand well (e.g., convective organisation and mid-latitude blocking).

Physical models themselves will remain essential, but perhaps in new contexts: synthetic training data, boundary conditions for climate downscaling, and integration into hybrid models. In these contexts, the ability to represent diverse climate states and to faithfully represent slow modes of variability and teleconnections becomes more important than matching observed trends or achieving middle-of-the-pack climate sensitivity. A shift of prediction towards data-driven methods can provide physical modellers with breathing room to address cancelling errors and implausible tuning, to explore counterfactual or paleoclimates, and to build comprehensible, simplified models.

We describe a total rebalancing of our field: precise representation handled by open-source ML models developed by computer scientists, generalisable knowledge obtained by domain scientists using ML *alongside* a physical model hierarchy. The value of domain separation and the uptake of flexible, scientist-friendly modelling approaches has steadily grown [e.g., 9] and the current juncture provides a perfect opportunity to finalise this transition. Untangling the technical and scientific aspects of modern modelling will allow for efficient progress, but these two paths must remain deeply collaborative and interconnected in order to fully realise the positive feedback loops we envision.

More fundamental than *how* we run our models, however, is *what* we want them to do. Until now the primary focus of ML advancement has, understandably, been on prediction problems. Now, as ML models start to surpass their physical equivalents, we argue that long-term progress will be best served by a focus on generalisation. Our observational record captures only a sparsely observed slice of our planet's atmospheric and oceanic variability: we know the extremes of our current climate only imperfectly, let alone those to come. Under such data poverty, only models able to generalise robustly and reliably beyond their training data are of value. This is of course well-acknowledged, and many ML papers include at least one unseen extreme event or future climate they were able to approximately represent.

It is our opinion that a far more systematic effort to understand and quantify ML generalisability is needed. Core-AI theory places few constraints on model performance out-of-sample, so if we desire robustly generalisable models the solution must come from *our* domain knowledge. The open availability of the ERA5 reanalysis dataset [10] and the WeatherBench benchmark [11] were essential in enabling the current revolution, allowing computer scientists with little or no domain knowledge to develop models that targeted the requirements of modern prediction systems. We advocate for community development of generalisability benchmarks – standardised sets of training data, out-of-sample tests and

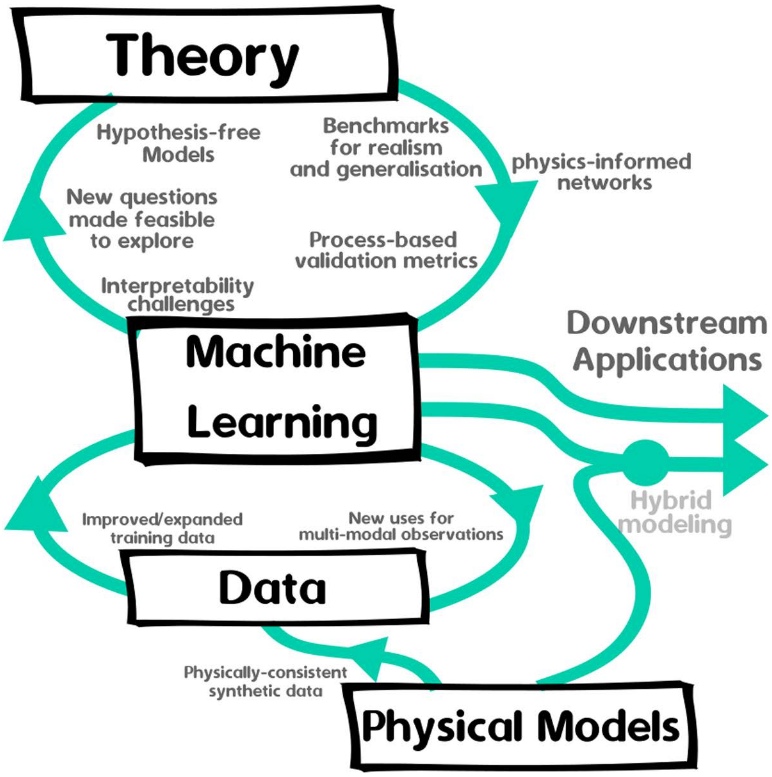
Theory
Hypothesis-free Models
Benchmarks for realism and generalisation
physics-informed networks
New questions made feasible to explore
Process-based validation metrics
Interpretability challenges
Machine Learning
Downstream Applications
Hybrid modeling
Improved/expanded training data
New uses for multi-modal observations
Data
Physically-consistent synthetic data
Physical Models

Fig 1. A (non-exhaustive) schematic highlighting new ways in which developments in one area of earth system science are stimulating advancements in other areas (green arrows). Advances in machine learning form a positive feedback loop with advances in data and fundamental theory, both dependent on them and supportive of them.

https://doi.org/10.1371/journal.pclm.0000710.g001

process-based skill assessments that can stress-test future models. Such tests should be challenging, representing the idealised capacities of a perfect model, akin to the ARC benchmark used to assess progress towards human-level AGI [12]. We might, as recent studies have pioneered, assess forecast models on their ability to reproduce simple dynamical instabilities [13], chaotic error growth [14], or to produce unseen hurricanes [15]. Regional climate emulators might be assessed for energy and mass conservation, on paleoclimates, across domains, or under unseen internal variability. Such tests will:

a) Quantify the physical robustness of current models so they can be used confidently.

b) Clarify which inductive leaps ML models can and cannot make.

c) Set ambitious targets for computer-scientists, driving development of skilful, trustworthy ML models.

This is not the vision of Delaroche's death knell, but perhaps of his contemporary, A.J. Wiertz:

> *"Let it not be thought [photography] kills art. No, it only kills the work of patience, and pays homage to the work of thought".*

The differences between science and the arts are many, but we believe many in our field will recognise the pre-eminence of *thought*—of imagination and theory-building—as a scientific ideal; one which stands to flourish as machine learning amplifies the power of human understanding.

Author contributions

Conceptualization: Joshua Dorrington, Julian Quinting, Stefan Sobolowski.

Visualization: Joshua Dorrington, Julian Quinting.

Writing – original draft: Joshua Dorrington.

Writing – review & editing: Joshua Dorrington, Julian Quinting, Stefan Sobolowski.

References

1. Lang S, Alexe M, Chantry M, Dramsch J, Pinault F, Raoult B, et al. AIFS- ECMWF's data-driven forecasting system. arXiv. 2025. http://arxiv.org/abs/2406.01465
2. Mardani M, Brenowitz N, Cohen Y, Pathak J, Chen C-Y, Liu C-C, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. Commun Earth Environ. 2025;6(1):124. https://doi.org/10.1038/s43247-025-02042-5
3. Dewey M, Hansson HC, Watson-Parris D, Samset BH, Wilcox LJ, Lewinschal A, et al. AeroGP: Machine Learning How Aerosols Impact Regional Climate.
4. Dheeshjith S, Subel A, Gupta S, Adcroft A, Fernandez-Granda C, Busecke J, et al. Transfer Learning for Emulating Ocean Climate Variability across CO2 Forcing. arXiv. 2024. https://doi.org/10.48550/arXiv.2405.18585
5. Bodnar C, Bruinsma WP, Lucic A, Stanley M, Vaughan A, Brandstetter J, et al. A Foundation Model for the Earth System. arXiv. 2024. https://doi.org/10.48550/arXiv.2405.13063
6. Müller SK, Pichelli E, Coppola E, Berthou S, Brienen S, Caillaud C, et al. The climate change response of alpine-mediterranean heavy precipitation events. Clim Dyn. 2024;62(1):165–86. https://doi.org/10.1007/s00382-023-06901-9
7. Bonavita M. On Some Limitations of Current Machine Learning Weather Prediction Models. Geophysical Research Letters. 2024;51(12):e2023GL107377. https://doi.org/10.1029/2023gl107377
8. Palmer TN. A personal perspective on modelling the climate system. Proc Math Phys Eng Sci. 2016;472(2188):20150772. https://doi.org/10.1098/rspa.2015.0772 PMID: 27274686
9. Klöwer M, Gelbrecht M, Hotta D, Willmert J, Silvestri S, Wagner GL, et al. SpeedyWeather.jl: Reinventing atmospheric general circulation models towards interactivity and extensibility. JOSS. 2024;9(98):6323. https://doi.org/10.21105/joss.06323
10. Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, et al. The ERA5 global reanalysis. Quart J Royal Meteoro Soc. 2020;146(730):1999–2049. https://doi.org/10.1002/qj.3803
11. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. J Adv Model Earth Syst. 2020;12(11):e2020MS002203. https://doi.org/10.1029/2020ms002203
12. Chollet F, Knoop M, Kamradt G, Landers B. ARC prize 2024: technical report. 2025. https://doi.org/10.48550/arXiv.2412.04604
13. Hakim GJ, Masanam S. Dynamical Tests of a Deep Learning Weather Prediction Model. Artificial Intelligence for the Earth Systems. 2024;3(3). https://doi.org/10.1175/aies-d-23-0090.1
14. Selz T, Craig GC. Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?. Geophysical Research Letters. 2023;50(20):e2023GL105747. https://doi.org/10.1029/2023gl105747
15. Sun YQ, Hassanzadeh P, Zand M, Chattopadhyay A, Weare J, Abbot DS. Can AI weather models predict out-of-distribution gray swan tropical cyclones? Proc Natl Acad Sci U S A. 2025;122(21):e2420914122. https://doi.org/10.1073/pnas.2420914122 PMID: 40392853