Machine Learning for Multiscale Systems: From Turbulence to Climate Prediction

By Dhruv Balwada and Laure Zanna

Continued improvements in the prediction of our planet's future state—whether it be via weather reports on the order of days to weeks or climate forecasts for the coming decades—have been one of the great revolutions of the last century. This foretelling power relies on our ability to solve mathematical equations that describe the dynamics of natural systems. Such equations combine first principles with empiricism and often represent conservation and/or thermodynamic laws. These equations also comprise a climate model, which is composed of interacting parts that represent distinct components like the ocean, atmosphere, ice, or land. The behavior of each component is highly nonlinear and turbulent, and the coupling results in emergent variability that is more complex than the sum of its parts (e.g., the dominant non-seasonal mode of variability on Earth that is called the El Niño-Southern Oscillation).

Climate as a Set of Coupled Multiscale Systems

The complexity and turbulent behavior of natural systems often stems from interactions between a wide range of spatial and temporal scales. For example, oceanic flows are frequently dominated by chaotic vortices that are approximately 100 to 200 kilometers (km) in diameter; the size of these structures is not a direct reflection of the forcing scales (e.g., atmospheric winds or solar heating that vary on scales larger than 1,000 km) or the instability (e.g., baroclinic instability with scales of 10 to 50 km). Instead, their size is a result of the natural propensity of vortices in flows with large aspect ratios to merge and form larger vortices. In the atmosphere, clouds develop from microscale processes and impact the Earth's albedo and thus the planetary-scale energy balance. In fact, three researchers received the 2021 Nobel Prize in Physics[1] for their work towards understanding and modeling these complex nonlinear multiscale systems.

Identifying accurate and representative solutions to such multiscale systems requires the resolution of an extensive range of scales—from millimeters to thousands of km—which is impossible for any modern computer and likely implausible for any computational system that will arise in the near future. Climate scientists therefore solve the equations for the (resolved) scales that are computationally feasible and most useful for decision-making purposes, while also parameterizing the impacts of the unresolved scales (known as the closure problem in fluid dynamics).

We can conceptually solidify this parameterization challenge by considering the partial differential equations that describe the turbulent flows in the ocean or atmosphere:

$$\frac{\partial Y}{\partial t} = -\nabla \cdot (\mathbf{u}Y) + F.$$

Here, $\mathbf{u}$ signifies the velocity vector, Y corresponds to quantities like momentum (velocity) or tracers (e.g., temperature), and F represents forcings, sources, sinks, pressure gradients, dissipation, and so forth. The flux—i.e., the multiplicative term $(\mathbf{u}Y)$ on the right—usually encompasses the multiscale interactions that emerge in turbulent flows. However, solving these equations on a finite grid fails to resolve the scales that are close to or smaller than the grid resolution. We can thus mathematically represent the true solution (Y) as $Y = \underline{Y} + Y'$, where Y is written as a sum of the resolved $(\underline{Y})$ and unresolved (Y') components respectively. For the sake of simplicity, we assume that this decomposition is akin to a low-pass filtering, which we accomplish with a Reynolds operator. Under this conceptual decomposition, the flux in the equation for the resolved scales becomes $\mathbf{u}\underline{Y} = \mathbf{u}\underline{Y} + \mathbf{u}'\underline{Y'}$, with contributions from resolved components $(\mathbf{u}\underline{Y})$ and unresolved or unknown small-scale components $(\underline{\mathbf{u}'Y'})$. The parameterization (or closure) problem refers to the estimation of the unresolved contribution $(\underline{\mathbf{u}'Y'})$ solely as a function of resolved variables.

Researchers often frame the parameterization of unresolved scales as the estimation of a dependence between the small scales' impact on the large scales as a function of the large scales. They have traditionally achieved this through purely physics-based approaches, which combine semi-empirical techniques with intelligent guesswork. For example, G.I. Taylor's seminal 1922 study parameterized the impact of small-scale turbulent motions on the large-scale dispersion of a passive tracer—like smoke from a chimney—as an *eddy* diffusion [5]. In this formulation, one would have to empirically determine the parameter—the eddy diffusivity—which would be many orders of magnitude larger than the molecular diffusivity of air. Scientists have employed similar reasoning to develop a number of parameterizations that are currently used in modern climate models, and often utilize observations of turbulence in the natural world to constrain the structure and parameters in these schemes [1, 4]. In these scenarios, additional sophistication may account for physical constraints and the parameters (like eddy diffusivity) might depend on large-scale variables with tunable coefficients. While these parameterizations form the backbone of modern-day climate models, inaccuracies in the parameterizations' structural forms—or even the parameters themselves—result in biases or systematic errors in the solutions.

[1] https://www.nobelprize.org/prizes/physics/2021/summary

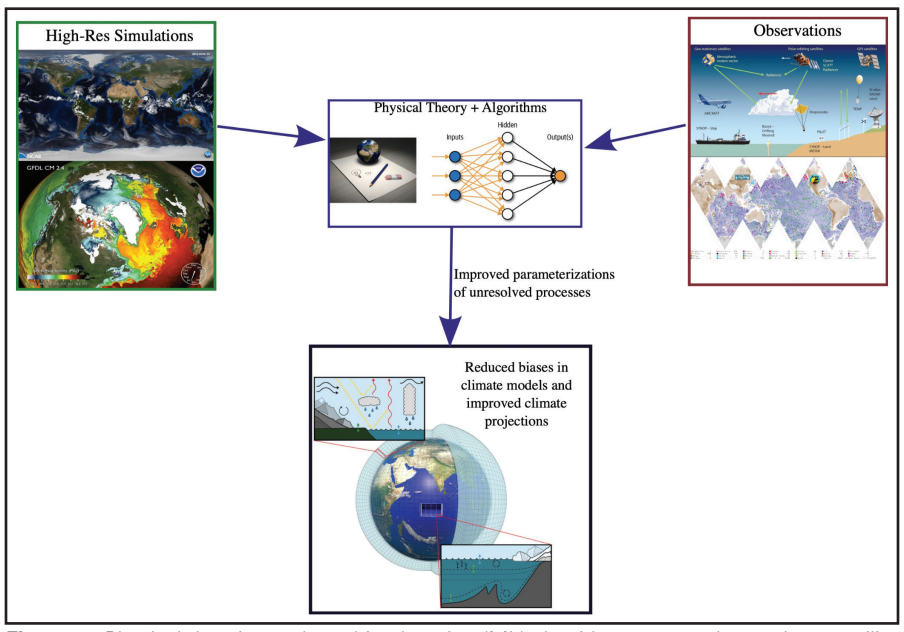

Figure 1. *Physical theories and machine learning (ML) algorithms can work together to utilize information from high-resolution simulations and observations in the pursuit of better parameterizations. Such parameterizations can potentially reduce biases in climate models and improve climate projections. Figure adapted from [2] and [7].*

Machine Learning as a Potential Avenue for Parameterization Improvement

An alternative route to parameterization involves using a statistical/machine learning (ML) algorithm for regression to determine the functional form of the small-scale impacts on the large scales, rather than scientists prescribing it themselves. Doing so requires the availability of data about the small scales, which we can procure through limited high-resolution simulations and observations with resolved small scales. It also necessitates computational technologies that can handle these large datasets, like graphics processing unit clusters on computational clouds — which are fortunately becoming more accessible.

The simplest data-driven approach that directly learns functional dependence assumes little to no prior knowledge and uses traditional "out of the box" ML algorithms, including neural networks (NNs), convolutional neural networks (CNNs), gaussian processes, and random forests — all of which have shown potential in many domains. For example, deep NNs—which include an increasingly large number of layers and trainable parameters—have displayed a high degree of skill in image recognition and game play. These purely data-driven approaches have also exhibited promise for climate science and are now leading rapid research advancements in "physics-aware" ML methods that combine data-driven approaches with physical knowledge (see Figure 1).

Broadly speaking, scientists are currently investigating three approaches for physics-aware ML methods. The first category is parameter estimation, which addresses parameterizations in which the structure is based on physical principles and the ML algorithm estimates some unknown free coefficients. In the second category, the loss function that trains the ML algorithm has a penalty or regularization term that incorporates certain known physical constraints. And in the third category, the structure of the neural network or another ML algorithm is modified to preserve some known symmetries or conservation properties of the system.

Recent work [3, 6] has demonstrated the promise of these physics-aware ML approaches for parameterizations of subgrid momentum and heat fluxes that arise in ocean turbulence (see Figure 2). Researchers used two physics-aware approaches that incorporated the known physical constraints into the architecture of the ML algorithm. The first approach utilized a CNN with a modified final layer that included physical conservation laws, and the second approach employed relevance vector machines that discover equations by combining basis functions that were selected based on physical knowledge about the problem. Both approaches showed superior skill over traditional parameterizations that are purely physics-based; they produced better pointwise predictions of sub-grid fluxes (offline evaluation) and the evolution of their parameterized coarse scale model more closely agreed with the high-resolution model (online evaluation). Along with other studies, this work has provided an exciting proof of concept and a potential way to accelerate improvements in climate models with data-driven ML techniques.

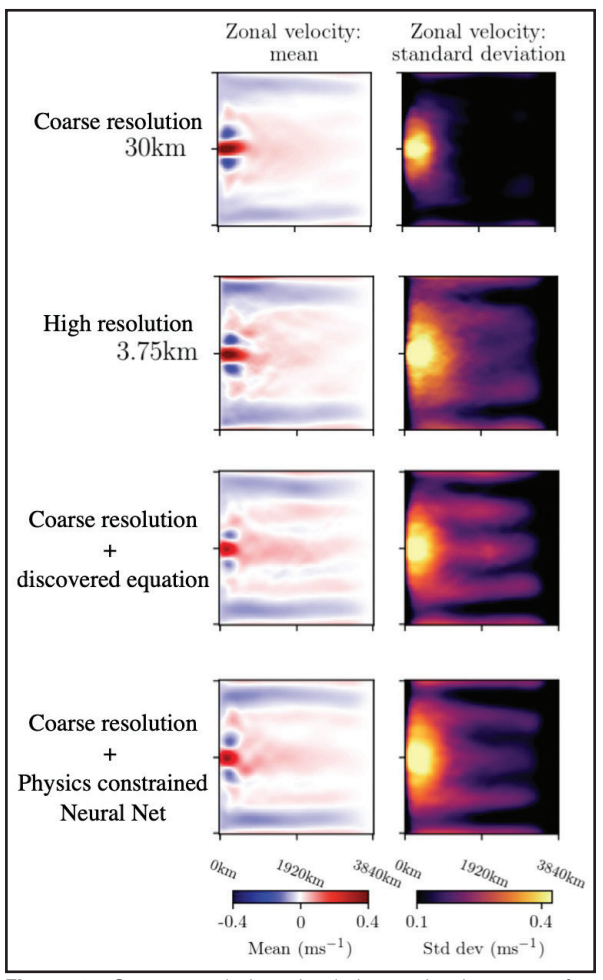

Figure 2. *Coarse-resolution simulations miss important features that high-resolution simulations can capture. Adding physics-aided, machine learning-based parameterizations to the coarse-resolution models significantly improves simulation skill. Figure adapted from [7].*

Is a Major Upgrade to Climate Models on the Horizon?

ML's arrival in recent years has promised accelerated advancement in many areas of science, including the understanding and parameterization of turbulent processes in climate models. Initial attempts to utilize these technologies have hinted at the possibility of potential breakthroughs that could provide a major upgrade to the current generation of climate models, ultimately reducing bias, improving the skill of predictions, and hopefully translating to better resource management and preparedness for the future. Only time and research will tell whether this promise comes to fruition.

Many exciting challenges related to implementation, generalization, and interpretation are located on the path towards these goals. The aforementioned progressions have galvanized the climate science community, as evidenced by the formation of the Multiscale Machine Learning In Coupled Earth System Modeling[2] (M^2LInES) international collaborative team and many other centers and institutes. Researchers are addressing multiple specific questions that advance critical thinking and move progress forward. For instance, how do we generalize (extrapolate) to regimes that are not part of the training data (for example, a future climate with warmer temperatures)? Do we have to use ML primarily as a black box, or can it be interpretable and aid in scientific discovery? What are the best ways to learn from both high-resolution simulations and observational noisy and/or sparse datasets? How should we approach the practical challenges that accompany the combination of disparate

[2] https://m2lines.github.io

technologies and scientific domains? How can we make these scientific advancements more reproducible and accessible?

We currently find ourselves at the beginning of an era with more questions than answers, similar to the early days of numerical fluid dynamics in the 20th century. While some of these problems will initially be solved with brute force empiricism, the development of a fundamental understanding that is based on a solid theoretical, mathematical and numerical foundation will prove essential for long-term success.

References

[1] Balwada, D., LaCasce, J.H., Speer, K.G., & Ferrari, R. (2021). Relative dispersion in the Antarctic circumpolar current. *J. Phys. Oceanog.*, *51*(2), 553-574.

[2] Christensen, H., & Zanna, L. (2022). Parametrisation in weather and climate models. Invited submission to *Oxford research encyclopedia of climate science*. Under review.

[3] Guillaumin, A.P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Sys.*, *13*(9), e2021MS002534.

[4] Roach, C.J., Balwada, D., & Speer, K. (2018). Global observations of horizontal mixing from Argo float and surface drifter trajectories. *J. Geophys. Res.: Oceans*, *123*(7), 4560-4575.

[5] Taylor, G.I. (1922). Diffusion by continuous movements. *Proc. London Math. Soc.*, *s2-20*(1), 196-212.

[6] Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophys. Res. Lett.*, *47*(17), e2020GL088376.

[7] Zanna, L., & Bolton, T. (2021). Deep learning of unresolved turbulent ocean processes in climate models. In G. Camps-Valls, D. Tuia, X.X. Zhu, & M. Reichstein (Eds.), *Deep learning for the earth sciences* (pp. 298-306). New York, NY: John Wiley & Sons, Inc.

Dhruv Balwada is an associate research scientist at the Lamont-Doherty Earth Observatory of Columbia University. He uses ocean observations and numerical simulations to understand the operation of ocean turbulence at scales of 1-100 kilometers and its impact on ocean circulation and climate. Laure Zanna is a professor in mathematics and atmosphere/ocean science at New York University's Courant Institute of Mathematical Sciences. Her research focuses on the ocean's role in climate on local and global scales through the analysis of observations and a hierarchy of simulations.