DOE WHITE PAPER: Machine Learning for Surrogate Modeling of the Upper Ocean and Heat Exchange Between the Ocean and Atmosphere

Nicholas Lutsko, Bruce Cornuelle, Sarah Gille, Mathew Mazloff, Matthias Morzfeld
Scripps Institution of Oceanography, UC San Diego

I Focal Areas

This white paper primarily addresses Focal Area 2: "Predictive modeling through the use of AI techniques and AI-derived model components". A secondary focus of the white paper is Focal Area 3: "Insight gleaned from complex data (both observed and simulated) using AI, big data analytics, and other advanced methods".

II Science Challenge

The primary scientific challenge addressed in this white paper is the development of cheap and accurate surrogate models of the upper ocean and of heat exchange between the upper ocean and the atmospheric boundary layer for forecast purposes. Secondary challenges include incorporating physical constraints into machine learning surrogate modeling approaches and incorporating machine learning-based surrogate models into weather forecast and climate models.

III Rationale

A primary challenge in weather prediction is air-sea fluxes. The ocean is the energy reservoir for the Earth system: per unit area, the heat energy required to warm the top 100m of the ocean by 0.1°C would be sufficient to warm the atmosphere by 4°C. Given the immense amount of heat energy contained within the upper-ocean, air-sea fluxes between the upper-ocean and the planetary boundary layer have a profound influence on weather events and patterns, particularly extreme events. Weather forecast models often use persistent or climatologically evolving Sea Surface Temperatures (SSTs) to estimate air-sea fluxes through bulk flux parameterizations. However, this approach does not capture the complex evolution of SSTs due to upper-ocean turbulent mixing. Adding more realistic representations of the upper ocean and its interactions with the atmospheric boundary layer is a vital next step in weather prediction, particularly for improving forecasts at weekly time-scales as well as the forecast of extreme events.

At the longer time-scales of climate, there is growing evidence that SST feedbacks are important for modeling clouds, which remain the largest source of uncertainty in future climate projections. For example, *Schneider et al. (2019)* showed that with SST feedbacks included in an atmospheric Large Eddy Simulation (LES) model, changes in cloud cover led to changes in SST, which fed back on the cloud cover changes. However, Schneider et al. used a "slab" model of the upper ocean, which did not include any of the turbulent processes that determine SSTs. Thus there is a need for cheap, accurate models of upper-ocean turbulence which can be coupled to high resolution atmospheric models used to investigate cloud behavior and other atmospheric processes.

Machine learning (ML) may be able to provide the solution to these problems through the development of surrogate models of the upper ocean and the atmospheric boundary layer. These could be run at low computational cost and with limited input data. Such a strategy capitalizes on the statistics of atmospheric and oceanic patterns without requiring that the full Navier-Stokes equations be solved at high resolution. Moreover, the challenges posed in modeling upper-ocean turbulence and heat exchange with the atmospheric boundary layer share many traits with the challenges of modeling atmospheric convection and cloud formation. Recently, a number of ML strategies have been tested to improve cloud models, including the development of surrogate models that draw on ML methods to derive

DOI: https://doi.org/10.2172/1769742
DOE BER, AI4ESP 2021 #AI4ESP1084 (https://www.ai4esp.org)

compact statistical representations of cloud processes (e.g. *Rasp et al., 2018; Brenowitz and Bretherton, 2019; Yuval and O'Gorman, 2020*). We believe that similar success can be achieved in using ML to resolve the problem of air-sea fluxes in weather and climate models.

IV Narrative

i) Overview

Developing surrogate models of upper ocean turbulence involves three key steps:

- Generating training data
- Training the ML systems
- Incorporating the surrogate models into weather and climate models

Below we outline our proposed approaches to completing each of these steps.

ii) Generating training data

ML algorithms require large amounts of training data. We believe that the most effective approach for our purposes is to train on data from high resolution LES simulations of the upper ocean and of the coupled upper ocean-atmospheric boundary layer system. These simulations would provide accurate and detailed training data from a wide range of ocean states and weather conditions, which are difficult to obtain from observations. Several models appropriate for performing such simulations are currently available (e.g., MITGCM, WRF), and large-scale data to force high resolution LES simulations under a wide range of conditions can be taken from reanalyses or from regional coupled models. Thus the required modeling and observational infrastructure to perform these simulations is readily available, though the computational burden will be large.

iii) Training the ML systems

Efforts to develop surrogate models of cloud processes have shown that there can be notable trade-offs between accuracy, speed and stability in machine learning-based parameterizations (*Yuval and O'Gorman, 2020*). In order to develop surrogate models appropriate for use in forecast models we believe that a multi-pronged approach is required, in which surrogate models are simultaneously developed with several different machine-learning algorithms and architecture. These include random forest (RF) networks, and variants of convolutional neural networks (CNNs), both of which have been used by the cloud parameterization community. CNNs can potentially be run faster and more accurately than RF networks, but the predictions of RFs are averages over subsets of the training dataset and so they inherently respect physical constraints. For instance, an RF network will never predict negative precipitation rates, which has proven to be a problem when CNNs are used to parameterize cloud models. Given the complex dynamics of the upper ocean, as well as the challenges in coupling to a forecast model, we have no a priori expectations for which algorithm will provide the best performance.

It has also recently been suggested that intelligent design of machine learning architecture (i.e., the choice of input and output variables, and the order in which these are calculated) can force even CNN-based parameterizations to respect physical constraints (*Bar-Sinai et al., 2019; Beucler et al, 2019*). We believe that this is crucial for ensuring successful coupling to atmospheric forecast models, particularly by reducing drift over the medium-term. In the context of upper ocean dynamics, we would want, for example, to ensure that the turbulent energy of the upper mixed layer is consistent with the energy input by wind stress. Thus, a multi-pronged approach should also include several different machine learning architectures, closely following strategies that have demonstrated success in other contexts.

iv) Incorporating the surrogate models into weather and climate models

After settling on the most promising machine learning parameterizations, the final step is incorporating the surrogate models into forecast and climate models. This is a novel step forward, and the experiences of ML cloud models suggest that stability will be a major issue. In fact, stability may be an even larger issue for surrogate models of atmosphere-ocean heat exchange because small fluctuations of the ocean state can have large impacts on the atmosphere. For this reason, we believe that physical constraints must be strongly enforced in the development of the surrogate models and during their incorporation into weather and climate models.

The surrogate models could be incorporated into regional coupled ocean-atmosphere models developed at Scripps, for which we have extensive experience. But we also anticipate partnering with DOE model developers, and other modeling centers to incorporate the surrogate model into their weather and climate models. In either case, the hybrid machine learning-ocean-atmosphere models will be tested and benchmarked for an ensemble of events to evaluate the machine learning parameterizations.

v) Leveraging the DOE's capabilities

The steps outlined above are computationally intensive, requiring high resolution modeling, the production of large data-sets and the use of advanced machine-learning algorithms. For these reasons, bringing the DOE's exascale computing capabilities to bear on this problem would greatly enhance the likelihood of success. Data on near-surface atmospheric conditions from the DOE's Atmospheric Radiation Measurement (ARM) Facility, particularly the Eastern North Atlantic site and the former Tropical Western Pacific site would also be useful for forcing the LES simulations and for benchmarking the surrogate model of atmosphere-ocean heat exchange.

We also note that many of the steps to answer this question are closely aligned with the goals of the Data-Model Integration Scientific Grand Challenge. These include: developing and using innovative computational tools, testbeds and benchmarks; producing data at process scales to evaluate and validate models; using the results of experiments to inform model development; and developing a scalable and adaptable framework that enables improved compatibility, integration, and interoperability of a hierarchical suite of models across a range of temporal and spatial scales. Beyond the specifics of the specific scientific question addressed here, we believe that the challenges involved in developing surrogate models and incorporating them into weather and forecast models can inform a wide range of similar scientific endeavors.

vi) Incorporating FAIR principles

The code and data from the LES simulations would be made publicly available to allow for easy reproducibility, as well as to promote the active development of new ML approaches to surrogate model development. The surrogate models themselves would also be made publicly available, ideally in the form of open source software packages that could be easily and quickly incorporated into weather or climate models.

References

Bar-Sinai, Y., S. Hoyer, J. Hickey, and M. P. Brenner, 2019: Learning data-driven discretizations for partial differential equations. Proceedings of the National Academy of Sciences,116 (31),15344-15349, doi: 10.1073/pnas.1814058116

Beucler, T., M. Pritchard, S. Rasp, P. Gentine, J. Ott, and P. Baldi, 2019: Enforcing analytic constraints in neural-networks emulating physical systems. ArXiv, 1909.00912.

Brenowitz, N. D. & C. S. Bretherton, 2019: Spatially extended tests of a neural network parametrization trained by coarse-graining. J. Adv. Model. Earth Sys. 11, 2727–2744 .

Rasp, S., Pritchard, M. S. & P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. Proc. Natl. Acad. Sci. U.S.A. 115, 9684–9689.

Schneider, T., C. M. Kaul, and K. G. Pressel, 2019: Possible climate transitions from breakup of stratocumulus decks under greenhouse warming. Nature Geoscience, 12, 163–167, doi: 10.1038/s41561-019-0310-1.

Yuval, J. & P. A O'Gorman,. 2020: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*11, 3295