EAM-HLR: Enhancing the low-resolution E3SM Atmosphere Model with an ML model of high-low-resolution residual in convective processes

Authors/Affiliations:

Andrew M. Bradley (Computational Mathematics, Sandia National Labs. (SNL))
Oksana Guba (Computational Science, SNL)
Kathryn Maupin (Optimization and Uncertainty Quantification, SNL)
Sivasankaran Rajamanickam (Scalable Algorithms, SNL)

Focal Area(s). (2) Augment the low-resolution (LR) E3SM Atmosphere Model (EAM) with a machine learning (ML) column model of the residual in convective processes with respect to high-resolution (HR) and ultra-HR (UHR) simulation and observational data.

Science Challenge. Substantially improve LR EAM climatology to increase multi-decadal water cycle (WC) predictability.

Rationale. Long-time climate and large-ensemble simulation campaigns are essential to quantitatively predict and bound uncertainty of climatological, including WC-related, changes in the next several decades. Despite continued substantial increases in computational power, computational and discretization scaling arguments show that HR and especially uniform UHR are orders of magnitude too expensive to provide these simulations. Thus, LR models will continue to have fundamental importance in answering questions critical to the WC. E3SM has a new UHR capability, SCREAM, and there are a number of other UHR data sources. We propose to leverage short, focused, HR and UHR simulations and other UHR data sources to improve the climatology of the LR model and thus its predictive skill.

Narrative. We propose three key ideas.

KI 1. Use E3SM and ARM current and future capabilities to generate training data sets for column physics, with a focus in this white paper on convective processes.

KI 2. Use the training data set to build a column ML model of the residual between LR column convective parameterizations and UHR and HR coarse-grained column data; we call this the High-Low resolution Residual model (HLR). In particular, we propose to supplement, not replace, existing column parameterizations. The HLR model's input will likely include column neighbors for additional context and robustness.

KI 3. Add HLR as an extra parameterization to the LR EAM model, with the hypothesis that the resulting EAM-HLR model will produce climatology that better matches HR and UHR climatology but at the substantially lower cost of LR EAM.

We focus on building a model for the *residual* in the column parameterizations—as opposed to full ML column parameterizations or for 3D processes—for a number of reasons. First, we hypothesize that a residual model will provide a stable and accurate LR EAM as quickly as possible—note that the trivial residual model of 0 recovers LR EAM exactly—thus focusing the project on increasing accuracy rather than trying to stabilize an ML-enhanced EAM. Second, an ML model of column physics that builds on a training data set of column inputs and outputs avoids the training and generalization problems that 3D ML models have. Third, a byproduct of this approach might be a more detailed understanding of where and why current column parameterizations are failing, leading to improvements in these parameterizations. HLR can naturally adapt to these improvements by retraining. Thus, HLR leverages previous, current, and future work on non-ML column parameterizations and might guide their future development.

DOI: https://doi.org/10.2172/1769697
DOE BER, AI4ESP 2021 #AI4ESP1015 (https://www.ai4esp.org)

In the remainder, we divide the proposed ideas into seven tasks and components (TC).

TC 1: High-resolution data. Assembling high-quality training data is the most important part of using ML in a practical application. Success will benefit not just this paper's project but any other focused on improving column parameterizations.

The first set of sources are HR and UHR simulation data, respectively from the EAM 1/4° model and the SCREAM 3.25km model. Our approach leverages these valuable HR and UHR models while minimizing the time these expensive models must simulate. In addition, we can create regionally refined mesh (RRM) and possibly small-planet configurations to create UHR training data at lower cost than a full-planet, uniform-resolution model. As suggested in [PB19], E3SM-MMF is another potential source of UHR training data. E3SM-MMF parameterizes convection using an LES model in each column.

UHR simulations are insufficient to capture processes such as boundary layer turbulence that have length scales of 100m or less. Thus, three additional sources of training data are important: first, targeted standalone periodic-planar-domain LES simulations in important convective regimes; second, ARM and other field data; third, high-resolution weather data. The DOE ARM program provides a number of value-added products focused on specific convective regimes and other atmospheric phenomena. For example, the LASSO data set includes an LES model and high-frequency observational data related to convection and cloud processes [GJVL+20]. The CACTI campaign provides data on cloud formation and dynamics, in particular, for orographic clouds and deep convective systems [NMVF21]. The PCCP project provides high-frequency 3D data of clouds' geometries and velocities [RÖ18].

UHR data will be expensive to generate, even a decade from now; thus, the ability to run short, focused UHR simulations is important. The methods of experimental design [FM08, LN95] and active design [Set09] are relevant: Use some metric of information content to probe the regime in which the next high-fidelity model should be run.

TC 2: Low-resolution EAM. The LR EAM is the foundation of the final product. It is much faster than higher-resolution simulations. LR (1°) EAM uses very roughly 32× fewer resources than uniform HR (1/4°) EAM and 12000× fewer than uniform UHR (3.25km SCREAM). At 1° global-model resolution, EAM-MMF with a 2D 64-column CRM is approximately 3.6× slower than EAM on a per-node basis with EAM-MMF running on a Summit node (2 P9 + 6 V100) and EAM running on a ∼13× less powerful Cori-KNL node.

TC 3: Residual inputs and outputs. LR EAM has coupled column models to parameterize convection: a version of the Zhang-McFarlane scheme for deep convection and CLUBB for shallow convection. In addition, a number of new schemes and modifications of current ones are being considered for EAMv3. Traditionally, EAM maps input column data to output column data. However, some modified schemes map column data in a neighborhood of the primary column to output for the primary column [YFX+17]. This approach may increase the robustness of ML-based models by providing a wider view of the input drivers to a process. A number of efforts have focused on developing ML models to replace column parameterizations, e.g. [MB19, BB19, BB18, BBPB20, OD18]. It appears that multiple efforts, e.g. [BHM+20, BBPB20], find that online use of an ML-based parameterization is not always robust. We believe these same researchers are now trying to model a discrepancy or residual between parameterized LR simulations and resolved UHR ones [BBWM+20, WMBB+20], but this literature has not developed yet. Here, we focus on building a residual model to accompany existing parameterizations. First, we must identify a composition

$\boldsymbol{w} = p(\boldsymbol{u})$ of EAM parameterized processes p_i that represent convection. A training data element in the training data set is a pair $(\boldsymbol{u}_{\mathrm{HR}}, \boldsymbol{w}_{\mathrm{HR}})$ in which $\boldsymbol{u}_{\mathrm{HR}}$ is a valid input to p and $\boldsymbol{w}_{\mathrm{HR}}$ is a valid output. The residual to model is $r(\boldsymbol{u}_{\mathrm{HR}}) \equiv \boldsymbol{w}_{\mathrm{HR}} - p(\boldsymbol{u}_{\mathrm{HR}})$.

TC 4: Collect and downscale (U)HR data. Each element $(\boldsymbol{u}_{\mathrm{HR}}, \boldsymbol{w}_{\mathrm{HR}})$ is obtained from a (U)HR data set. $\boldsymbol{u}_{\mathrm{HR}}$ and $\boldsymbol{w}_{\mathrm{HR}}$ are downscaled (coarsened) in space and time from a 4D volume of a (U)HR simulation. These 4D volumes are not available in standard output because saving this much data is orders of magnitude too prohibitive. Thus, an important task in building the training data set is developing software embedded in the model that selects and downscales 4D space-time volumes and writes just the downscaled data to file.

TC 5: Physics-informed ML models of residual. Given a high-quality training data set, it will be easy to try many different ML models of the residual, especially by using tools such as PyTorch and TensorFlow and because the ML model is a column and not 3D model. A number of successes in ML (e.g. protein folding, sequencing and taxanomic tasks in genomics, natural language processing, image classification, text and image generation) are based on identification of a key problem and development of training data sets, followed by rapid multi-group, multi-effort development of a variety of ML methods to address these. For these reasons, we do not describe specific ML models but rather focus on primary objectives.

Most important is building known physical constraints into the model. Physical constraints can be thought of as infinitely large data sets with no or well-characterized uncertainty, and thus every physical constraint one knows should be applied to the model. In the overall column model $\boldsymbol{w} = \hat{p}(p(\boldsymbol{u}), r(\boldsymbol{u}))$, the existing parameterizations p presumably obey a number of conservation and microphysical constraints. $\hat{p}$ must, as well. Some of these can be imposed as a post-processing step; hence $\hat{p}$ is a general function of p and r rather than just a sum of the two. Others should be applied to r according to either of the roughly two broad methods of physics-informed ML models, constraints on architecture and imposing algebraic constraints when optimizing model weights.

Also important may be to consider a hierarchy of complexity of residual models. An important part of this white paper's conception of the problem enables using very simple residual models because we start with a physically valid overall model of convection.

TC 6: EAM-HLR. The final product is EAM-HLR, LR EAM enhanced with a residual model of convection. Our focus on a column residual ML model will yield a stable, long-running EAM-HLR model with better climatology than standard EAM early in the project; later work can then focus on improving accuracy rather than debugging stability problems.

TC 7: Verification and validation. Verification will assure the column physics including the ML-model residual obey physical constraints. Validation data sets may include reanalysis-nudged LR and HR simulations, ARM field data, and standard climatology observational data sets, e.g. GCPC, MERRA, ERA, CERES. Primary validation will not be against left-out training data, which describes column-level processes, but rather against global and regional climatology data, the quantities of primary interest for making decisions.

GC targets and FAIR principles. Our proposal addresses the integrated WC scientific GC (EESSD Strategic Plan 4.1) and data-model integration GC (EESSD SP 4.5); TC 1, 4 address parts of goals 1, 3, and TC 1, 4, 6 address parts of goal 4. EAM-HLR development will occur in an E3SM fork that will ultimately be merged to the E3SM source repository. TC 3, 5 development will occur in an E3SM-Project repository focused on ML for E3SM simulations. Training data and simulations used to generate these will be archived.

Suggested Partners/Experts.
Mike Pritchard, Prof., Univ. of California, Irvine
Chris Bretherton, Prof., Univ. of Washington, Seattle
Noah Brenowitz, Senior Machine Learning Scientist for Climate Modeling, Vulcan, Inc.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2021-1561 O.

References

[BB18] Noah D. Brenowitz and Christopher S. Bretherton. Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45(12):6289–6298, 2018.

[BB19] Noah D. Brenowitz and Christopher S. Bretherton. Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, 11(8):2728–2744, 2019.

[BBPB20] Noah D. Brenowitz, Tom Beucler, Michael Pritchard, and Christopher S. Bretherton. Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, 77(12):4357–4375, 2020.

[BBWM+20] Christopher Stephen Bretherton, Noah Brenowitz, Oliver Watt-Meyer, Walter Andre Perkins, Jeremy McGibbon, Anna Kwa, Brian M. Henn, Spencer Clark, and Lucas Harris. Improving the forecast skill of a climate model using machine learning trained on a global 3 km simulation. In *AGU Fall Meeting 2020*. AGU, 2020.

[BHM+20] Noah D. Brenowitz, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Oliver Watt-Meyer, and Christopher S. Bretherton. Machine learning climate model dynamics: Offline versus online performance. *arXiv preprint arXiv:2011.03081*, 2020.

[FM08] Gaia Franceschini and Sandro Macchietto. Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, 63(19):4846–4872, 2008.

[GJVL+20] William I. Gustafson Jr, Andrew M. Vogelmann, Zhijin Li, Xiaoping Cheng, Kyle K. Dumas, Satoshi Endo, Karen L. Johnson, Bhargavi Krishna, Tami Fairless, and Heng Xiao. The Large-Eddy Simulation (LES) Atmospheric Radiation Measurement (ARM) Symbiotic Simulation and Observation (LASSO) activity for continental shallow convection. *Bulletin of the American Meteorological Society*, 101(4):E462–E479, 2020.

[LN95] Shuangzhe Liu and Heinz Neudecker. A V-optimal design for Scheffé's polynomial model. *Statistics & probability letters*, 23(3):253–258, 1995.

[MB19] Jeremy McGibbon and Christopher S. Bretherton. Single-column emulation of reanalysis of the Northeast Pacific marine boundary layer. *Geophysical Research Letters*, 46(16):10053–10060, 2019.

[NMVF21] T Connor Nelson, James Marquis, Adam Varble, and Katja Friedrich. Radiosonde observations of environments supporting deep moist convection initiation during RELAMPAGO-CACTI. *Monthly Weather Review*, 149(1):289–309, 2021.

[OD18] Paul A. O'Gorman and John G. Dwyer. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 2018.

[PB19] Mike Pritchard and Tom Beucler. Outsourcing sub-grid cloud physics to neural networks. *https://e3sm.org/outsourcing-sub-grid-cloud-physics-to-neural-networks*, 2019.

[RÖ18] David M. Romps and Ruşen Öktem. Observing clouds in 4D with multiview stereophotogrammetry. *Bulletin of the American Meteorological Society*, 99(12):2575–2586, 2018.

[Set09] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[WMBB+20] Oliver Watt-Meyer, Noah Brenowitz, Christopher Stephen Bretherton, Spencer Clark, Brian M Henn, Anna Kwa, Jeremy McGibbon, Andre Perkins, and Lucas Harris. Correcting weather models by learning nudging tendencies from hindcast simulations. In *AGU Fall Meeting 2020*. AGU, 2020.

[YFX+17] Yuxing Yun, Jiwen Fan, Heng Xiao, Guang J. Zhang, Steven J. Ghan, Kuan-Man Xu, Po-Lun Ma, and William I. Gustafson Jr. Assessing the resolution adaptability of the Zhang-McFarlane cumulus parameterization with spatial and temporal averaging. *Journal of Advances in Modeling Earth Systems*, 9(7):2753–2770, 2017.