Machine Learning Climate Model Dynamics: Offline versus Online Performance

Noah D. Brenowitz,[*] **Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa**
W. Andre Perkins, Oliver Watt-Meyer, Christopher S. Bretherton
Vulcan, Inc.
Seattle, WA
{noahb, brianh, spencerc, annak, jeremym, andrep, oliwm, chrisbr}@vulcan.com

Abstract

Climate models are complicated software systems that approximate atmospheric and oceanic fluid mechanics at a coarse spatial resolution. Typical climate forecasts only explicitly resolve processes larger than $100\,\mathrm{km}$ and approximate any process occurring below this scale (e.g. thunderstorms) using so-called parametrizations. Machine learning could improve upon the accuracy of some traditional physical parametrizations by learning from so-called global cloud-resolving models. We compare the performance of two machine learning models, random forests (RF) and neural networks (NNs), at parametrizing the aggregate effect of moist physics in a $3\,\mathrm{km}$ resolution global simulation with an atmospheric model. The NN outperforms the RF when evaluated offline on a testing dataset. However, when the ML models are coupled to an atmospheric model run at $200\,\mathrm{km}$ resolution, the NN-assisted simulation crashes with 7 days, while the RF-assisted simulations remain stable. Both runs produce more accurate weather forecasts than a baseline configuration, but globally averaged climate variables drift over longer timescales.

1 Introduction

Machine learning has the potential to improve the accuracy of climate models but should take advantage of our existing physical knowledge. Climate and weather models represent the motions of the Earth's atmosphere as a system of discretized ordinary differential equations (ODEs). Some of these terms are well-known from first principles such as the Navier-Stokes equations and radiative transfer. However, climate models typically have a horizontal grid-size of $100\,\mathrm{km}$ (Intergovernmental Panel on Climate Change, 2014) and cannot resolve the dominant physical scales of some atmospheric processes, especially turbulence, cumulus convection, and cloud-radiation interactions; these must be parametrized. We call those processes that involve the formation of clouds and precipitation the 'moist physics'. New simulations called global cloud resolving models (GCRMs), which explicitly resolve key aspects of global cloud fields associated with precipitating cumulus convection by using horizontal grid scales of less than 5 km (Satoh et al., 2019), are currently too expensive for long climate simulations but contain a wealth of information about how clouds interact with the large-scale—a valuable training dataset for inexpensive machine learning (ML) parametrizations.

Sub-grid-scale parametrization is a function approximation problem. Let the ODEs describing a climate model can be divided into two components as follows

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{g}_i(\mathbf{x}, t) + \mathbf{f}(\mathbf{x}_i; \theta). \tag{1}$$

The known physics $\mathbf{g}$ account for large-scale atmospheric fluid mechanics as well as some set of known physics, and are a function of the non-local state $\mathbf{x}$. The vector $\mathbf{x}$ represents the state of the

---

[*]Corresponding author

global atmosphere (e.g. temperature, humidity, and wind). Because the atmosphere mixes rapidly in the vertical direction, parametrized physics $\mathbf{f}$ are typically assumed to depend only on some ML parameters θ and the *atmospheric column* overlying a single horizontal grid cell i. We denote this horizontally local state with $\mathbf{x}_i \in \mathbb{R}^m$ where m is the number of vertical grid points times the number of 3D fields input to the parametrization. In this work, f will include sources and sinks of humidity and heat due to moist physical processes.

Moist physics are conventionally handled by a suite of human-devised parametrizations encorporating physical constraints and empirical knowledge. These typically assume an analytical sub-grid-scale cloud model in statistical equilibrium with the large-scale environment (Arakawa, 2004), but recent work proposes parameterizing moist physics using machine learning models trained from either higher-fidelity simulations (Rasp et al., 2018; Brenowitz and Bretherton, 2018, 2019; Yuval and O'Gorman, 2020; Krasnopolsky et al., 2010) or reanalysis data (McGibbon and Bretherton, 2019).

These ML schemes are trained offline, as a supervised learning problem, where the inputs $\mathbf{x}_i$ and outputs $\mathbf{f}$ are taken from a pre-computed dataset. Because this training fails to account for feedbacks between $\mathbf{g}_i$ and $\mathbf{f}$, offline accuracy does not translate immediately into online accuracy when the ML is coupled to the fluid dynamics solver and used to simulate the weather or climate. For example, offline training will often yield a numerically unstable scheme that causes an online simulation to crash within a matter of days (Brenowitz and Bretherton, 2019; Brenowitz et al., 2020). The reasons behind this are starting to emerge. Brenowitz et al. (2020) demonstrated with formal stability analysis that this instability is related to the linearized behavior of a neural network (NN) when coupled to idealized wave dynamics. Also, feedback loops in coupled simulations can easily generate out-of-sample inputs unlike any in the training sample. On the other hand, random forests (RFs) are more stable online, likely because they can only predict averages of observed samples (O'Gorman and Dwyer, 2018; Yuval and O'Gorman, 2020).

Prior work has focused on idealized aqua-planet configurations over a global ocean with fixed surface temperature. While this is a useful prototyping configuration, ML parametrizations need to be accurate with realistic geography and topography to be used in real-world forecast models. Therefore, this manuscript has two goals: 1) demonstrate the feasibility of ML parametrization on a more realistic atmospheric model and 2) compare the offline and online performance of RFs and NNs for this problem.

We first introduce the atmospheric model we are trying to improve and which we use to generate the training data in Section 2. Then, we explain the random forest and NN training processes and ML formulation. We share online and offline performance results in Section 3 and conclude in Section 4.

2 Methods

2.1 The FV3GFS Atmospheric model

We use the FV3GFS (Harris et al., 2020) atmospheric model to generate both fine-resolution training data and evaluate the accuracy at a coarser resolution. FV3GFS solves the three-dimensional Euler equations over a spherical geometry discretized on a cubed-sphere grid (Putman and Lin, 2007).

To generate our training data, we run the FV3GFS with an approximate horizontal grid of $3\,\mathrm{km}$. At this resolution, FV3GFS can resolve the dominant motions due to deep convection Satoh et al. (2019), so this model is run with only the microphysics scheme, radiative transfer, shallow, convection, and a boundary layer turbulence scheme (Harris et al., 2020). Since this model has its own set of biases, we include an additional Newtonian relaxation term to the temperature, pressure, and wind fields which nudges this simulation towards a an observationally-derived gridded data product called a reanalysis with a time scale of 1 day. We perform a 40 day run with this configuration at a NOAA super-computing facility and save the full 3D state of the atmosphere and process-wise tendency information, horizontally block-averaged to $200\,\mathrm{km}$ resolution, every 15 minutes.

The baseline model we are hoping to improve is based on the FV3GFS model run at a $200\,\mathrm{km}$ resolution. This scale is typical for climate models, but cannot resolve many moist atmospheric processes. We, therefore, assume that the known physics $\mathbf{g}$ only includes advection together with the standard parametrizations of clear-sky radiation and turbulence, a configuration we call Clouds-off. The ML will predict the remaining processes $\mathbf{f}$. We compare ML-assisted runs with a baseline configuration (All-physics) with the standard human-designed moist physics parametrizations.

2.2 Machine learning models

For simplicity, the ML models will only predict sources of temperature and humidity, letting the clouds-off physics handle frictional processes. We compute the terms-to-be-parametrized $\mathbf{f}$ as budget residual of (1), so that $\mathbf{f}_i = d\mathbf{x_i}/dt - \mathbf{g_i}$. The total tendency $d\mathbf{x_i}/dt$ is the sum of all the physical-process tendencies in the fine-resolution data and the convergence of vertical temperature and humidity fluxes. If initialized with the coarsened state of the fine-resolution model, the coarse-resolution model will develop strong transients damaging the estimated known physics $\mathbf{g}_i$ (Lynch, 2008). To generate smooth estimates of $\mathbf{g}_i$ and $\mathbf{x}_i$, we nudge a coarse clouds-off simulation towards the fine-resolution data with a 3-hour nudging timescale for temperature, humidity, pressure, and the winds. The inputs features $\mathbf{x}_i$ include the vertical profiles of temperature and humidity along with the cosine of the solar zenith angle, the surface elevation, and the land-sea mask. The nudging time-scale is a regularization parameter; longer time-scales will give smoother, but more biased estimates, of $\mathbf{x}_i$ and $\mathbf{g}_i$.

We compare the online and offline performance of RFs and NNs on this problem. The training data consist of 130 randomly drawn snapshots from (August 5 through 31) and testing data are 72 snapshots from (September 1 through 7). Each snapshot contains $13\,824$ spatial samples. Both the RF and NN are trained to minimize the mean-squared-error scaled by the inverse standard deviation of each output feature. A RF is fit with a maximum depth of 13 and 13 ensemble members (one per batch of 10 timesteps). A two layer NN, with 128 nodes per layer and ReLU activation, is fit with the Adam optimizer (learning rate 0.001) with inputs normalized by the standard deviation and mean computed over a single batch. A mini-batch size of 512 is used for NN training and 8 passes through the training data (epochs) are completed.

3 Results

The RF and NN have comparable accuracy offline on the testing data (cf. Figure 1). The RF has a worse coefficient of determination (R^2) than the NN does at all pressure levels of the output. On the other hand, the global and time average of the NN is more biased than the RF, likely because this problem features highly non-Gaussian outliers that distort the MSE-based loss function of the NN. Both the RF and NN predict too much global average heating in for pressure levels between $1000\,\mathrm{mb}$ to $400\,\mathrm{mb}$.

To test the online accuracy, we couple the NN and RF to the $200\,\mathrm{km}$ atmospheric model run with clouds-off known physics, and perform 10-day hindcast simulations initialized at 0 UTC on August 5, 2016 and compare them to an all-physics simulation (cf. Figure 2). Even though the NN outperformed the RF offline, the NN simulation crashes after around 7 days, while the RF simulation successfully completes the 10-day simulation. This confirms findings that random forests are often stable for this problem (O'Gorman and Dwyer, 2018) while neural network are not (Brenowitz et al., 2020).

We evaluate the predictive skill for each ML method at each forecast lead time using the root-mean squared error (RMSE) and global average of precipitable water (PW) and $500\,\mathrm{mb}$ geopotential height (Z500) compared to the verification high-resolution training data across all grid columns around the globe. Overall, the global average of Z500 for the NN remains closer to the verification until it crashes, and it has the highest forecast skill of the three simulations at predicting Z500. The RF has the lowest RMSE and smallest bias for PW for a few days, but eventually starts to dry out. Overall, the baseline model is more robust and has less systematic drift in PW and Z500, but the ML-assisted runs have better skill (lower RMSE) for forecasts for up to 5 days.

4 Conclusion

We have compared how well a random forest and a simple neural network perform when they replace the human-designed moist-physical parametrizations of a coarse resolution model. Unlike past studies, we have trained the RF and NN with the exact same training data, a global cloud-resolving simulation with an approximate resolution of 3km. To our knowledge, this is the first such clean comparison in the literature.[2] Moreover, this simulation is significantly more complex that the idealized datasets used in past work because it includes a realistic land surface and topography.

---

[2]Since the initial submission of pre-print, Yuval et al. (2020) have also performed a similar comparison.
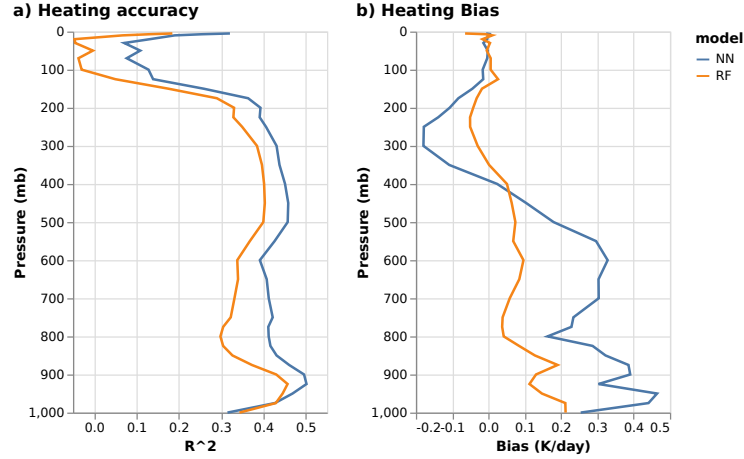
Figure 1: Offline accuracy of ML-predicted temperature source (i.e. heating) for the NN and RF. (a) area-weighted R^2 scores and (b) area-weighted averages computed over the testing times. The predictions and truth are interpolated to fixed pressure levels before computing the metrics.
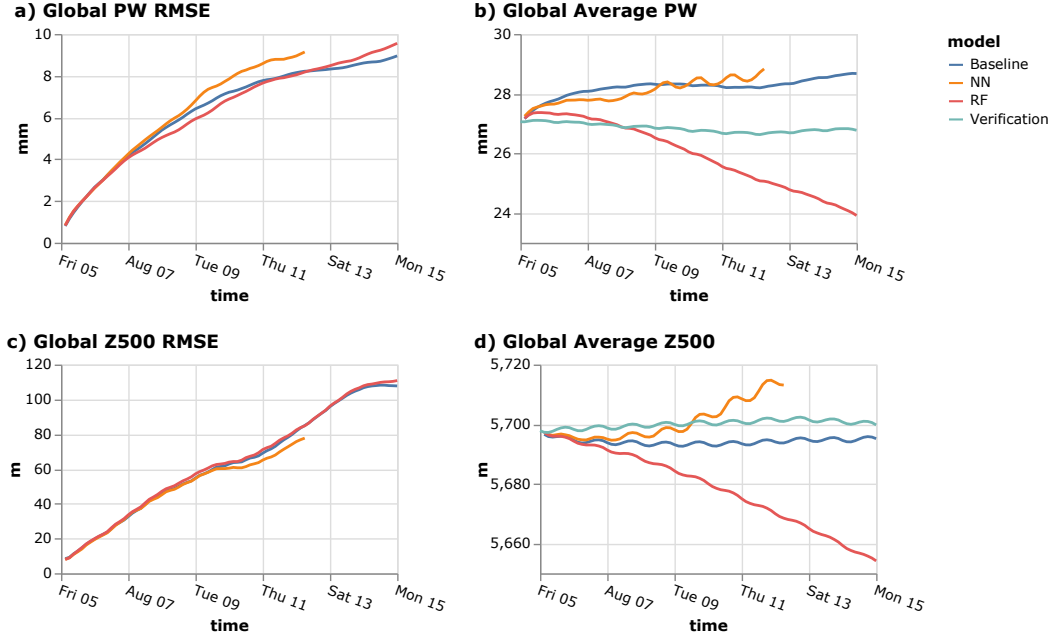

Figure 2: Online skill of the NN and RF. Compares the forecast accuracy in terms of RMSE for precipitable water (a) and $500\,\mathrm{mb}$ geo-potential height (c). The respective global averages are shown in (b) and (d).

The NN is more accurate offline than the RF, but is not numerically stable online (i.e. when coupled to known physics). Because the training process does not account for feedbacks between the ML and the known physics, online simulations can quickly produce samples unlike any seen in the training data. Because RFs predict outputs within the convex hull of their training data, they are likely more robust than NNs when forced to extrapolate to such new samples. Future work should focus on finding an offline input-output prediction problem that translates to good online performance.

The global average of the NN predictions are also more biased offline, possibly because its training procedure is less robust to extreme rainfall events in the training data. This offline bias could possibly be addressed using robust loss functions (e.g. Huber loss or mean absolute error).

Broader Impact

If successfully expanded upon, this work promises to improve the physical models we use to forecast weather and climate with machine learning. In particular, our goal is to improve accuracy of precipitation forecasts with these ML moist physics parametrizations. Such forecasts, with quantified uncertainty, of precipitation trends and extremes in a changing climate will allow policymakers and the general public to make better-informed decisions about climate impacts on many aspects of society and the natural world.

Acknowledgments and Disclosure of Funding

References

Akio Arakawa. The cumulus parameterization problem: Past, present, and future. *J. Clim.*, 17(13): 2493–2525, July 2004. ISSN 0894-8755. doi: 10.1175/1520-0442(2004)017<2493:RATCPP>2.0. CO;2.

Noah D Brenowitz and Christopher S Bretherton. Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, 17:2493, June 2018. ISSN 0094-8276. doi: 10.1029/2018GL078510.

Noah D Brenowitz and Christopher S Bretherton. Spatially extended tests of a neural network parametrization trained by coarse-graining. *J. Adv. Model. Earth Syst.*, July 2019. ISSN 1942-2466, 1942-2466. doi: 10.1029/2019MS001711.

Noah D Brenowitz, Tom Beucler, Michael Pritchard, and Christopher S Bretherton. Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of Atmospheric Sciences*, Early online release, 2020. doi: 10.1175/JAS-D-20-0082.1.

Lucas Harris, Linjiong Zhou, Shian-Jiann Lin, Jan-Huey Chen, Xi Chen, Kun Gao, Matthew Morin, Shannon Rees, Y Qiang Sun, Mingjing Tong, Baoqiang Xiang, Morris Bender, Rusty Benson, Kai-Yuan Cheng, Spencer Clark, Oliver Elbert, Andrew Hazelton, Jacob Huff, Alex Kaltenbaugh, Zhi Liang, Timothy Marchok, and Hyeyum Hailey Shin. GFDL SHiELD: A unified system for Weather-to-Seasonal prediction. September 2020.

Intergovernmental Panel on Climate Change. *Climate Change 2013 - The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, March 2014. ISBN 9781107661820. doi: 10.1017/CBO9781107415324.

V M Krasnopolsky, M S Fox-Rabinovitz, and A A Belochitski. Development of neural network convection parameterizations for numerical climate and weather prediction models using cloud resolving model simulations. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2010. doi: 10.1109/IJCNN.2010.5596766.

Peter Lynch. The origins of computer weather prediction and climate modeling. *J. Comput. Phys.*, 227(7):3431–3444, March 2008. ISSN 0021-9991. doi: 10.1016/j.jcp.2007.02.034.

J McGibbon and C S Bretherton. Single-Column emulation of reanalysis of the northeast pacific marine boundary layer. *Geophys. Res. Lett.*, July 2019. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL083646.

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. *J. Adv. Model. Earth Syst.*, 10(10):2548–2563, October 2018. ISSN 1942-2466. doi: 10.1029/2018MS001351.

William M Putman and Shian-Jiann Lin. Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, 227(1):55–78, November 2007. ISSN 0021-9991. doi: 10.1016/j.jcp.2007.07.022.

Stephan Rasp, Michael S Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. U. S. A.*, 115(39):9684–9689, September 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1810286115.

Masaki Satoh, Bjorn Stevens, Falko Judt, Marat Khairoutdinov, Shian-Jiann Lin, William M Putman, and Peter Düben. Global Cloud-Resolving models. *Current Climate Change Reports*, 5(3): 172–184, September 2019. ISSN 2198-6061. doi: 10.1007/s40641-019-00131-0.

Janni Yuval and Paul A O'Gorman. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nat. Commun.*, 11(1):3295, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17142-3.

Janni Yuval, Paul A O'Gorman, and Chris N Hill. Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. October 2020.