Physics-Informed Machine Learning from Observations for Clouds, Convection, and Precipitation Parameterizations and Analysis

Authors

Paul Ullrich, University of California Davis

Tapio Schneider, California Institute of Technology

Da Yang, University of California Davis and Lawrence Berkeley National Laboratory

Focal Area(s)

Fusing learning from Earth observations from space and from the ground (e.g., ARM) with newly developed interpretable and generalizable physics-informed model structures to improve the parameterization of clouds and convection and advance the simulation, understanding and analysis of hydrological extreme events. This covers the prescribed foci of (2) predictive modeling using a hierarchy of models and (3) insight gleaned from complex data.

Science Challenge

Hydrological extreme events are the natural hazards causing the largest losses of property and life. Physical theory and accumulating empirical evidence show that the intensity of extreme precipitation increases in a warming climate (e.g., O'Gorman & Schneider 2009, Donat et al. 2016). Yet predictions of precisely how much more intense precipitation will become, and how this translates into local impacts such as urban flooding, have remained stubbornly uncertain (Srivastava et al. 2020; Wehner et al., 2021). Key uncertainties can be traced to longstanding biases in the representation of turbulence, convection, and clouds in climate models. Breakthroughs in this area are now within reach thanks to advances in data science and computing, potentially fused with new physics-informed model structures to achieve a new level of accuracy in modeling precipitation and related hydrological extremes.

While the potential to exploit data-driven methods is real and clear, there are two key challenges in using them in climate modeling: (1) we have to predict a climate we have not seen, for which there is no observed analog; (2) while we live in an era of big scientific data—we receive on the order of a terabyte of Earth data from space every day—this is not nearly large enough to constrain the myriad degrees of freedom of the climate system by data-driven methods alone. For example, atmospheric turbulence alone has on the order 10^{27} degrees of freedom. Thus, we need generalizable and interpretable physical models to restrict the learning from data to a subspace consistent with fundamental physics.

To advance the science and modeling of precipitation and hydrological extremes, we propose a three-legged research program that (1) advances interpretable physical model structures of turbulence, cloud, and convection parameterizations, (2) learns about unknown closure functions in these model structures from both observational data and data generated computationally in high-resolution simulations, and (3) advances understanding of precipitation and hydrological extremes through analysis

DOI: https://doi.org/10.2172/1769762
DOE BER, AI4ESP 2021 #AI4ESP1133 (https://www.ai4esp.org)

of the data-informed models obtained under (1) and (2). An outcome will be climate models that are informed by data and high-resolution simulations, yet are computationally efficient and remain viable vehicles for scientific investigation.

Rationale

Earth observations have been used to tune individual parameterizations, but the data volume used has been limited, and data are typically not employed to directly learn about the complex physical relationships that underlie parameterizations. Although it is commonly said that available data are not directly informative about small-scale processes represented by parameterizations (e.g., we have little data about turbulence in clouds), such arguments do not account for the fact that differences in how small-scale processes are represented in climate models have clear and measurable impacts on macro-scale climate and weather data. Hence, available climate and weather data are informative about small-scale processes. However, the inverse problem of learning about uncertain small-scale processes from available climate and weather data is ill-posed: multiple micro-scale processes can give rise to the same macro-scale observations. Because of the data-hungry nature of machine learning methods—namely, their success relies in large part on overparameterization—the problem is not well suited for unconstrained machine learning from observational data. Nonetheless, paired with prior information from model structures derived from the known equations of motions, methods of machine learning and data assimilation have the potential to be revolutionary in this area.

Narrative

We propose to pursue new theoretical approaches that coarse-grain the equations of motion, with clear controls on errors made in approximations, together with data science tools to learn about both systematic errors made in the coarse graining and about unknown closure functions. This dual approach avoids the potential pitfalls of both, entirely data-driven learning and overreliance on potentially misspecified physical models. We propose to pursue numerous innovative approaches. For example:

- We will exploit new algorithms for Bayesian learning, which accelerate traditional algorithms 1,000x by combining ensemble Kalman inversion with machine-learning emulators of computationally expensive climate models (e.g., Cleary et al. 2021, Dunbar et al. 2021).
- We will use tools such as learning from a dictionary of differential equation terms (Rudy et al. 2017, Zanna and Bolton 2020, Udrescu & Tegmark 2020) or direct learning of mappings in function spaces ("neural PDEs", Li et al. 2020) for data-driven discovery of unknown closure functions in parameterizations.
- We will use machine learning tools such as Gaussian process regression (Kennedy & O'Hagan 2001) and sparse learning approaches that incorporate stochastic noise (Schneider et al. 2021) to model systematic model errors in the places where the error actually occurs (e.g., within the

parameterization schemes).

We will use these data-driven approaches alongside new approaches for deriving the physical structure of parameterizations by systematic coarse graining (e.g., Cohen et al. 2020, Lopez-Gomez et al. 2020), including approaches such as the Mori-Zwanzig formalism that lead to stochastic closure models with memory terms (Palmer 2005, Lucarini et al. 2014, Wouters et al. 2016).

Put simply, we propose to harness observational data massively with machine learning and data assimilation tools to inform physics-based parameterizations of key processes in Earth's hydrological cycle (and associated surrogate models). This will allow us to obtain parameterizations that are interpretable and generalizable, thus remaining scientifically analysable, while at the same time leading to what may be a step change in their predictive capabilities. In particular, we expect that the imposition of physical constraints will enable deeper understanding of the internal structure of such systems, and thus advance our understanding of the upstream drivers of precipitation and hydrologic extremes. By training parameterizations within a climate model directly on climate statistics relevant for predictions of hydrological extremes, we expect such models to have reduced biases and uncertainties in both predicted mean and extreme fields. For example, the moments and higher-order statistics of pointwise precipitation are generally known to a high level of accuracy in observations, and they are of significant socioeconomic importance. By training a model directly on such precipitation statistics—including higher-order statistics such as the frequency with which high daily precipitation thresholds are exceeded, as prototyped in Dunbar et al. (2021)—we expect to arrive at parameterizations that capture precipitation statistics better than existing models. The resulting parameterizations can then be used directly and interchangeably in large-scale general circulation models. We expect such models to be publicly available via standard open source channels.

The models will be designed to leverage all available observational data, from reanalysis data from weather forecasting centers, direct satellite observations of clouds from platforms such as CloudSAT and CALIPSO, to in-situ observations, e.g., from ARM research facilities. The full range of available data types has not been brought to bear on the parameterization problems that are the crux for improving models and analyses of the hydrological cycle. We expect breakthroughs from harnessing orders of magnitude more data than have been used in climate modeling before.

To exploit the wealth of observational data, it is necessary to learn about parameterizations as they are embedded in a climate model, to translate small-scale processes into observable macro-scale phenomena. This can be done within the Energy Exascale Earth System Model (E3SM), or within other models, such as that under development at the Climate Modelling Alliance. In either case, the outcome will be parameterizations that can be embedded in any number of climate models. We will target atmospheric model resolutions in the range of 50-100 km—a range of resolutions within which running climate models is computationally efficient, so that they can be used as tools for scientific investigation, much like climate models such as CESM have been used for decades.

Physics-Informed Machine Learning from Observations for Clouds, Convection, and Precipitation Parameterizations and Analysis

Suggested Partners/Experts (Optional)

Data assimilation/machine learning: Andrew Stuart and Anima Anandkumar (Caltech)

Observational data: NASA Jet Propulsion Laboratory

E3SM developers

References (Optional)

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., Stuart, A.M., 2020: Calibrate, emulate, sample, *J. Comp. Phys*, **424**, 109716.

Cohen, Y., Lopez-Gomez, I., Jaruga A., He, J., Kaul, C. M., Schneider, T., 2020: Unified entrainment and detrainment closures for extended eddy-diffusivity mass-flux schemes. *J. Adv. Mod. Earth Sys.*, **12**, e2020MS002162.

Donat, M.G., Lowry, A.L., Alexander, L.V., O'Gorman, P.A. and Maher, N., 2016. More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, *6*(5), pp.508-513.

Duan, S., Ullrich, P. and Shu, L., 2020. Using Convolutional Neural Networks for Streamflow Projection in California. Front. Water 2: 28. doi: 10.3389/frwa.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., Stuart, A. M., 2020: Calibration and Uncertainty Quantification of Convective Parameters in an Idealized GCM, arxiv e-prints arXiv:2012.13262.

Kennedy, M. C., and A. O'Hagan, 2001. Bayesian calibration of computer models. *J. Roy. Statist. Soc. B*, 63:425–464.

Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, 2020: Fourier neural operator for parametric partial differential equations, arXiv:2010.08895

Lopez-Gomez, I., Cohen, Y., He, J., Jaruga, A., Schneider, T., 2020: A generalized mixing length closure for eddy-diffusivity mass-flux schemes of turbulence and convection. *J. Adv. Mod. Earth Sys.*, **12**, e2020MS002161.

Lucarini, V., R. Blender, C. Herbert, F. Ragone, S. Pascale, and J. Wouters, 2014. Mathematical and physical ideas for climate science. *Rev. Geophys.*, **52**, 809–859.

O'Gorman, P.A., T. Schneider, 2009. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proc. Natl. Acad. Sci.*, 106:14773–14777.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005. Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193.

Rudy, S. H., S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Adv.*, 3:e1602614, 2017.

Schneider, T., Stuart, A.M., Wu, J., 2021: Learning stochastic closures using ensemble Kalman inversion, *Trans. Math. Appl.*

Srivastava, A., R. Grotjahn, and P.A. Ullrich, 2020. Evaluation of historical CMIP6 model simulations of extreme precipitation over contiguous US regions. *Weather Clim. Extr.* 29, 100268, doi: 10.1016/j.wace.2020.100268.

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C.S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D. and Kodama, C., 2019. DYAMOND: The DYnamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), p.61.

Udrescu, S.-M. and M. Tegmark. AI Feynman, 2020: A physics-inspired method for symbolic regression. *Sci. Adv.*, 6:eaay2631, 2020.

Wehner, M., J. Lee, M. Risser, P.A. Ullrich, P. Gleckler, and W. Collins, 2021. Evaluation of extreme subdaily precipitation in high-resolution global climate model simulations. *Phil. Trans. Royal Soc. A (accepted).*

J. Wouters, S. I. Dolaptchiev, and V. Lucarini, 2016. Parameterization of stochastic multiscale triads. *Nonlin. Processes Geophys*., **23**, 435–445.

Zanna, L., and T. Bolton, 2020. Data‐driven equation discovery of ocean mesoscale closures, *Geophys. Res. Lett.,* 47, e2020GL088376