Systematically Generating Hierarchies of Machine-Learning Models, from Equation Discovery to Deep Neural Networks

While the added value of machine learning (ML) for weather and climate applications is measurable, it remains challenging to explain, especially for large deep learning models. Inspired by climate model hierarchies, which use dynamical models of increasing complexity to help connect our fundamental understanding of the Earth system with operational predictions, we ask:

Given a climate process for which we have reliable data, how can we systematically generate a hierarchy of ML models, from simple analytic equations to complex neural networks?

To address this question, we choose two atmospheric science problems for which we have physically-based, analytic models with just a few tunable parameters, and deep learning algorithms whose performance was already established in previous work: Cloud cover parameterization and shortwave radiative transfer emulation. In each case, we formalize the ML-based hierarchy by working in a well-defined, two-dimensional plane: Complexity versus Performance. We choose the number of trainable parameters as a simple metric for complexity, while performance is defined using a single regression metric (e.g., the mean-squared error) calculated for the same outputs on a common validation dataset.

During this presentation, we will demonstrate how to use our data-driven hierarchies for two purposes: (1) Data–driven model development; and (2) process understanding. First, each ML model of the hierarchy occupies a well-defined (complexity, performance) position as they use the same performance metric. Models that maximize performance for a given complexity unambiguously define a Pareto frontier in (complexity) $\times$ (performance) space and can be deemed optimal. Second, optimal models on the Pareto frontier can be compared to reveal which added process/nonlinearity/regime/connectivity/etc. leads to the biggest increase in performance for a given complexity, which facilitates process understanding. For example, using sequential feature selection on simple polynomial fits, we underline the nonlinear relationship between condensate mixing ratios and cloud cover. Using a specialized type of convolutional neural network (U-net++) to emulate shortwave radiative heating, we can mostly overcome the biases of simpler models of shortwave radiation (one-stream, linear, multilayer perceptron, convolutional neural network), notably in the presence of one or more cloud layers.

To show its versatility, we apply our framework to the data-driven discovery of analytic models, which are interpretable by construction. Applying sequential feature selection to neural network models of cloud cover, we identify the five most informative features, and use them as inputs to genetic algorithms. These genetic algorithms automatically generate hundreds of candidate equations, which can be filtered using physical constraints and ranked using our (complexity) $\times$ (performance) space. Our best candidate is interpretable, achieving a coefficient of determination close to 0.95 with only 13 trainable parameters. It beats all neural networks using three features or less, the widely-used Sundqvist scheme by capturing how cloud condensate mixing ratio nonlinearly affects cloud fraction, and the Xu-Randall scheme by describing how temperature decreases cloud cover.

In summary, we can systematically build hierarchies of Pareto-optimal ML models to better understand their added value. By cleanly comparing these ML models to existing schemes, and promoting process understanding by hierarchically unveiling system complexity, we hope to improve the trustworthiness of ML models for weather and climate applications.