Using Decision Trees for Data-Driven Process Studies, Conditional Bias Correction, and Model Physics Improvements

Augustin Vintzileos[(1)], Benjamin A Cash[(2)], and James L. Kinter III[(2)]

(1) University of Maryland – ESSIC
(2) George Mason University - COLA

A major limiting factor in numerical weather prediction and climate simulation and projection is the quality of the representation of unresolved physical processes in the atmosphere. This shortcoming traditionally has been addressed with process-based model development; however, there have been recent rapid advances in the application of data-driven approaches such as machine learning (ML). There is a gap between these two classes of approach in that process-oriented models do not necessarily take the best advantage of the prodigious volume of observations, and data-driven approaches are difficult to interpret physically.

Decision trees are the most explainable ML methodology. The use of decision trees inducted based on the modeler's or forecaster's expertise has been developed at ECMWF (the ecPoint) for (a) generating subgrid cumulative distribution functions and (b) bias correction. Our work builds on the ecPoint method by automating decision tree induction. Instead of inducting the trees using human expertise, the experts interpret the set of rules generated automatically. Error distribution at the leaves of the tree can be used for conditional bias correction. This set of rules also divulges atmospheric conditions for which further fine-tuning of parameterizations is necessary. This paper explores the third possibility offered by decision trees: data-driven process studies. We use the tendency of hourly 2-meter temperature from ERA5 to induct decision trees to demonstrate their utility for studying heat waves.