Machine learning techniques for convection parameterization.

Summary

A number of impressive results in the literature establish machine learning as a powerful tool for improving convective parameterization. However, it is early days, and we should be caution about considering certain problems solved. Greater understanding between the convection and machine learning communities in future will improves the quality of machine learning parameterization and accelerate the rate of progress.

Discussion points

In a wide-ranging introduction, we discussed the general application of machine learning in the atmospheric sciences. A distinction was drawn between those processes for which we know the equations but cannot afford to run a process-based simulation in weather or climate models, and those processes for which we do not know the equations. Where we do know the equations, it is possible in principle to run process-based simulations to train our machine learning models for optimal results and to validate them. If we do not extrapolate beyond the training data when running simulations, machine learning-based parameterizations can be very successful.

Parameterization of radiation in weather forecasting has been accelerated using machine learning for several decades, for example. Some of the group felt, however, that even greater rewards might be possible where we do not know the equations, because machine learning algorithms are extremely good at finding patterns in data empirically. Another broad distinction was made between the weather and climate problems. Whereas possible 5 day weather forecasts for Exeter might be a subset of historical observations, and should therefore by-and-large exist within a good model training dataset, the problem of 21st century climate change means extrapolating to a state that is very little observed. One perspective is that the weather model may simply be replaced in its entirety with a machine learning algorithm that produces weather forecasts given the present state of the Earth system. All agreed that this is not a viable solution for climate at this point.

A number of publications in which machine learning representation has been trained to produce a representation of convection that can be placed in a coarse general circulation model as a substitution for conventional parameterization now exist. The substitution has been done for convective parameterization, all atmospheric physics parameterizations and using high resolution simulations of the atmosphere. Although impressive, it was felt that these representations tend to have problems with numerical stability, again especially where the machine learning algorithm is forced to represent conditions not found in its

training dataset. Machine learning and statistical techniques generally produce poor outcomes when extrapolated outside their training inputs, putting machine learning at a large disadvantage compared with conventional physically-motivated parameterizations when extrapolation is necessary. It was conceded that this is likely to remain a problem. However, the group discussed a number of ways this problem might be ameliorated. One solution was that the machine learning algorithm could be "trapped" within a physical parameterization structure that only allowed machine learning to adjust the values of unknown parameters for specific input cases, providing an optimally tuned simulation, but never straying too far from what convection modellers would deem physically possible.

The strength of machine learning in identifying patterns is useful not only for building parameterizations from data but also for understanding those patterns. Statistical techniques are now available that throw light on what neural networks and other machine learning algorithm actually do to represent convection or other processes. The potential overlap with the Convection Playground group was noted. That group intends to build linear algebra representations of convection from impulse-response tests on a large variety of convection schemes and high-resolution simulations within one model framework. It was agreed that this is an excellent goal, but the chosen methodology might not be optimal for discovering the convective processes most important in the real environment. Certainly, representing impulses on individual GCM levels with a LES would be a very large use of computational resources.

Collaboration with the machine learning community could be improved by adopting the techniques they use when introducing problems for solution. Problems are often introduced to the community as "challenges" with a clear goal that competing partcipants aim to solve as accurately and efficiently as possible, often with the award of a prize for the most successful solution. There may also be a benefit if convection researchers design simpler theoretical problems for machine learning rather than assuming that machine learning can solve all aspects of convection in one iteration -- just as is done in the development of conventional parameterizations. In addition to specific challenges, well-documented datasets specifically designed for training machine learning algorithms would be useful for making progress. In the case of a process like convection where the relevant equations are known, high-resolution simulations could provide a clean dataset with few caveats, ideal for researchers expert in machine learning and less expert in the processes of convection. With these goals in mind, Leif Denby has established the website http://ai4environment.io/to give atmospheric and machine learning researchers a place to interact and solve problems. We hope that this can become a place where researchers with different expertise can collect to make progress on weather and climate model parameterization.