Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2020-427-RC3, 2021

Comment on gmd-2020-427

Anonymous Referee #3

Referee comment on "Copula-based synthetic data augmentation for machine-learning emulators" by David Meyer et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-427-RC3, 2021

Comments:

- It is my impression that the nomenclature Emulator-based training does not accurately reflect the principles of what is done in the setting described by the authors. Independently of how we train our ML models (using observations or simulations), both general settings provide a surrogate (an emulator) to make predictions. I believe the term "Simulation-based training (SBT)" reflects better the principles of training a surrogate model based on simulations being done by, what the authors call, physical models. This distinction, although a bit superficial, has been carefully proposed in earlier work in general scientific endeavors. See for example:
  - Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. Statistical science, 409-423.
  which has led to a more mature framework of modeling emulators based on simulation results to predict real world processes. See:
  - Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), 425-464.

- I believe that to refer to the parameters of a ML model as _weights_ lessens the generalizability of the work of the authors. This is because the term "weight" is particular to Deep Learning (DL) instances such as Artificial Neural Networks or Multilayer Linear Perceptron models and their subsequent generalizations. My issue with this is that ML is a much broader discipline than just DL. _Any_ ML used for prediction is looking for the best possible association of X (features, properties, descriptors) to Y (the target). In this sense, we are looking for the best candidate h that can achieve $Y \approx h(X)$ in some sense (for example, as measured by Mean-Square Error). We choose parametrized models due to our ability ---mainly, through iterative optimization algorithms--- to learn such approximations. Although, for an ML researcher/practitioner this is not new. Someone with no such background would not make the immediate connection in the text with w to "the best function approximation for a specific model architecture".

- This leads me to my next concern: are simulations contrasted to data? Even though, a surrogate is built upon simulation-based observations, at some point it needs to be contrasted to real world data to measure the validity of using a specific model configuration (either for the physical model or the emulator). There is work being done in this direction for climate predictions. See, for example:
  - Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2020). Calibrate, emulate, sample. Journal of Computational Physics, 424, 109716.
  as an example of a complete simulation-observation based strategy to learn surrogate models.

- The goal of learning ML models for climate applications is a hot topic in research. You can also see:
  - Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. Proceedings of the National Academy of Sciences, 115(39), 9684-9689.
  Which leads me to question what is the novelty of the proposed manuscript.

- The stochasticity of the training of an MLP can be accounted, for example, with K-fold cross validation. It should also be used as it internally provides a measure of generalization error that can help guide the selection of certain hyper-parameters (for example the optimization-related parameters). Why not use such a strategy for this manuscript?

- As a final comment, it is not clear to me the setting and the intention of the strategy presented by the authors. I believe it is not clear how to interpret these results. It seems like a particular instance of a data-augmentation strategy, with the potential benefit of preserving the observed probabilistic relationships among training data. I believe the authors does not provide clear evidence that such an augmentation achieves better results when compared to real world observations.

Minor comments:

- Typo found at line 151. It reads "...all variables a continuous..." it should be "...all variables are continuous...".