Specific Reviewer comments for “Copula-based synthetic data generation for machine learning emulators in weather and climate: application to a simple radiation model”

Major comments

Section 2.1 defines a general methodology for training ML models using either Observation-based training (OBT) or Emulation-based training (EBT). However, it appears that these definitions are inconsistent with the methods used in the paper. From what I can work out, the different methods are:

A. OBT (as defined): Train ML model on inputs X and outputs Y. *The source of X and Y is not defined, but is presumably observations, pseudo-observations or the like.*
A. OBT (as implemented): ***Inconsistent with A (as defined), since Y comes from a physical model fed with X like in method C (EBT). In other words, A is not implemented, only C.***

B. OBT with data generation (as defined): Train ML model on inputs X’ and outputs Y’, where X’ and Y’ is synthetic data created using a data generation model (copula) fitted to X, Y. *The source of X and Y is not defined, but is presumably observations, pseudo-observations or the like.*
B. OBT with data generation (as implemented): ***Inconsistent with B (as defined), since Y comes from a physical model fed with X like in EBT.***

C. EBT (as defined): Train ML model on inputs X and outputs Y, where Y has been generated by feeding inputs X to a physical model.
C. EBT (as implemented): ***Consistent with C (as defined), X comes from satellite data and Y comes from a toy physical model fed with X.***

D. EBT with data generation (as defined): Train ML model on inputs X’ and outputs Y’, where inputs X’ have been created using a data generation model (copula) fitted to X, and outputs Y’ come from a physical model fed with X’.
D. EBT with data generation (as implemented): **Otherwise consistent with D (as defined), but X and Y are included in training data.**

Claims to study the use of synthetic data for OBT in this paper are weak in the absence of observationally sourced outputs. If such data is unavailable or obtaining it is outside the scope of this work then my suggestion then the methodology should be reframed. Alternatively, make it clear that B is implemented in such a way to only mimic the defined method, in the absence of suitable data, and argue why this is valid.

Figure 1 is also inconsistent with the methods used. This figure should either be changed to describe the actual implementation, or a second figure with those could be made, something like this:

Baseline	Synthetic data method 1 (D)	Synthetic data method 2 (B)
X	X	X
	X	X
	generative model	physical model
	X'	Y
X	X+X'	X, Y
physical model	physical model	generative model
Y	Y+Y'	X', Y'
X, Y	(X+X'),(Y+Y')	(X+X'),(Y+Y')
prediction model	prediction model using synthetic inputs	prediction model using synthetic inputs and outputs

Minor comments:

L25: Chevallier and Krasnopolsky laid out the initial work but newer studies should be mentioned, e.g. Ukkonen et al. (2020) and Veerman et al. (2021). A lot of work has also been done recently to train NNs on data from cloud-resolving models (e.g. Brenowitz et al. 2018, Gentine et al. 2018, Rasp et al. 2018), which I'm not sure if it fits the authors definition of EBT (since X and Y has the same source) but it's important to place the present work in the context of the wider literature.

L35-46; section 2.1: another approach for "EBT with data generation", which is probably worth mentioning, is to simply sample the multidimensional input space rigorously using either random methods such as Latin hypercube sampling, or deterministic methods such as a Halton sequence. As demonstrated by Ukkonen et al. (2020), who sampled gas concentrations uniformly while keeping the dependencies of pressure and temperature intact, this can work well for difficult problems where generating outputs with physical models is cheap (EBT).

Whether it's better to train ML models on "realistic" datasets where the observed dependencies between inputs are respected, or on "dense and wide" data which sample the input space (N-dimensional hypercube) more uniformly, is to me nontrivial. On one hand, the latter may result in more generalizable models which also learn the underlying physics more effectively, assuming that the input-output mapping given by the physical model remains to some extent rooted in physics throughout this expanded input space. On the other hand, it may come at a significant cost in model complexity and computational resources. It's even possible that minimizing errors across the domain space comes at the expense of degraded performance on real datasets, regardless of model complexity, due to imperfect information. This would certainly encourage the use of approaches presented here, e.g. copulas, which respect the observed statistical distributions. (If the authors are aware of any studies on this I would be interested to know.)

L69: Before seeing the results, is strategy B a legitimate approach? It may be useful but I have not seen it being used. Fitting a simpler statistical model (copula) on observations and using it to create synthetic inputs and outputs to train a more complex statistical model (ML) seems a bit odd - does it actually extract any new information?

L85: The use of cloud optical depth instead of total optical depth for predicting longwave radiation only makes sense when the outputs come from a toy model and not observations, since clear-sky absorption is important for observed long-wave radiation. Again, due to the confusing overview of methods (2.1) the unobservant reader might think a method where the outputs come from observations is included, in which case the chosen inputs appear strange.

L90: "*We then define case A (or C) as the baseline..*" Here it first becomes apparent that both A and C are **not** used in this paper. The authors seem to redefine A to be equal to C, as they do not have observations to use as target outputs for Observation-based training (A), but then it should be made clear that strategy A is in fact not implemented.

Sections 2.3.2 - 2.3.4: For someone who was unfamiliar with copulas, this served as a good and clear introduction for the most part, but I am left confused about what kind of assumptions/parameters the "Vine-parametric" copula uses to model the dependence of two variables?

L230: *"using the source X or augmented X' data depending on the strategy (i.e. OBT or EBT).*" Again, confusing - OBT is not actually implemented. Furthermore, in Figure 1, X and X' are used in both strategy B (OBT-Augmented) and strategy D (EBT-Augmented), which is inconsistent with the highlighted sentence.

Figure 3. Perhaps a diagonal 1:1 line would aid interpretation, but this is a matter of style.

Figure 5 d). These results are good and quite interesting.

References:

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289-6298.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock?. *Geophysical Research Letters*, *45*(11), 5742-5751.

Pal, A., Mahajan, S., & Norman, M. R. (2019). Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. *Geophysical Research Letters*, *46*(11), 6069-6079.

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.

Ukkonen, P., Pincus, R., Hogan, R. J., Pagh Nielsen, K., & Kaas, E. (2020). Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002226.

Veerman, M. A., Pincus, R., Stoffer, R., van Leeuwen, C. M., Podareanu, D., & van Heerwaarden, C. C. (2021). Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200095.