

Geosci. Model Dev. Discuss., referee comment RC2
<https://doi.org/10.5194/gmd-2020-427-RC2>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on gmd-2020-427

Anonymous Referee #2

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-RC2>, 2021

The authors have written an interesting paper on the use of copulas for synthetic data generation for ML emulators in weather and climate applications. This is a valuable and, as far as I know, novel contribution to the ML in weather and climate modelling field. The paper is well-written and for the most part, clear and concise.

Besides some improvements that could be made with regards to literature review, I found a larger issue with regards to the methodology, which appears inconsistent in the way it is defined and implemented. The authors initially appear to use both observation-based training (OBT) and emulation-based training (EBT), but strictly speaking no OBT is done as the outputs come from a physical model (just like in the authors' definition of EBT). Perhaps the implementation of method B (OBT with data generation) is intended only to mimic the method as it is defined, but then this should be made explicit and the validity of this approach discussed.

While some readers may find no problem with the conceptualization and presentation, since it is open to interpretation, I found it nonintuitive. That being said, this a publishable paper if the clarity is improved. I attach my specific comments as a PDF supplement.

Please also note the supplement to this comment:

<https://gmd.copernicus.org/preprints/gmd-2020-427/gmd-2020-427-RC2-supplement.pdf>