Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-310-RC1, 2020

Geoscientific Model Development Discussions
Open Access

Interactive comment on "An improved multivariable integrated evaluation method and NCL code for multimodel intercomparison (MVIETool version 1.0)" *by* Meng-Zhuo Zhang et al.

Anonymous Referee #1

Received and published: 27 December 2020

I have carefully read the paper about the development of a new evaluation method for multiple fields and multimodels by Zhang et al, and, despite the fact that I think that the paper is mostly well written (language, structure, and so on), I can not recommend its acceptance in its present form.

In terms of scientific significance, I found the paper poor, since it basically uses a very simple technique (weighted average) to re-use techniques that have been published in the past.

In terms of scientific quality, I find that many references are missing, the authors do not consider techniques that have been common in climatology in the last twenty years,

presenting them as advanced. I will develop this point later in detail.

Regarding scientific reproducibility, the authors use some models as example (M1 to M10) and two reanalyses (REA1 and REA2) without mentioning the models, the reanalyses used, the periods, the experiments

For me, the main concern is related to the novelty (or lack of) of the paper. As the authors properly recognize in their section 2.1, the majority of the new methodology involved in the diagram has already been published in two papers such as Xu et al. (2016) and Xu et al., (2016). Thus, as far as I can see, and as written by the authors in the abstract, the new developments in this paper refer to:

1. The use of area-weighting by means of the use of a weighted average

2. The extension of their code to a potential combination of scalar and vector fields. Which, as explained by the authors in Figure 1, involves the change in the dimensions of the input matrix to their evaluation method.

Regarding point 1 above, the authors make what I find a very misleading statement in line 44-45 of their paper, I quote "most previous model performance metrics did not consider spatial weight". This is clearly not true. The paper by Taylor (2001) which gave rise to the idea of the Taylor diagram and which was cited by the authors, already mentions the possibility to use weighted statistics (see page 7183, lines after Eq (1) in that paper). Moreover, Boer and Lambert (2001) thoroughly cover this idea and explicitly used weights w_k in their formulation. The use of the square root of the cosine to account for the varying size of grid points in the estimation of EOFs goes back as far as North et al., (1982), at least, and is commonly used (see the description of function eofcov() in NCL, the programming language used by the authors in their implementations). Additional examples in the use of weights in the evaluation of climate models to account for different grid points can be found elsewhere such as Eq. (1) in Gleckler et al. (2008) or seminal papers in the field such as Reichler and Kim (2008). Studies can be found explicitly devoted to the analysis of the role that smoothing plays in the verification statistics (Mason and Knutti, 2011; Räisänen and Ylhäisi, 2011). The fact that meridional grid size can be misleading in the evaluation of climate models is well known since at least Benestad (2005). Thus, I think that the authors can not state that the consideration of different weight factors for different grid points to account for their different sizes as written in their paper is novel. And, by itself, the use of a weighted mean instead of a simple mean, does not seem very advanced, either. So, I can not recommend the acceptance of the paper on the basis of this being an advance in science, since this has been constantly carried out in papers during the last twenty years.

Second, the combination of multiple fields (or components of vector fields) as presented in point 2 above can also be a problem, from my point of view. As I see it, the algorithm lumps in the same indices (points in the diagram) information from different variables or components of different vector fields. Even though it might be practical to have a single model-evaluation index (point in their diagram), the fact that different variables are mixed might be obscuring important diagnostics. For instance, vector variables can show differences in the orientation of the simulated vector fields or their relative variances. I'd suggest the authors to discuss this issue by presenting (for instance) the way that two similar synthetic vector datasets behave if their error statistics are similar but they differ in the way the error statistics are distributed in the zonal and meridional directions, for instance. This would highlight the way these statistics are reflected in the diagram designed by authors. I guess that if the same amount of error is distributed in the zonal/meridional directions in two synthetic models, the authors are going to get the very same points in their diagram, but the source of the error is very different.

Finally, the authors highlight in substantial parts of their manuscript that they provide an implementation of their methodology using NCL. This is apparently an important part of their contribution, since it is stated so in the abstract, section 4 and Table 1. However, NCL has been kept in maintenance mode by NCAR

https://www.ncl.ucar.edu/open_letter_to_ncl_users.shtml since September 2019 and this is not mentioned in the manuscript.

I understand that the implementation of the technique provides a tool "ready to go" for climate scientists, but I doubt this is enough for a highly cited journal such as GMD. However, may be I am wrong and the editor thinks otherwise. For me, the difference between a rejection or a major revision is just a matter of how much the editor thinks a "ready to use" tool is a valid contribution. I am not used to the editorial policies of GMD, so that this finally ends in his/her hands.

Benestad, R. E. (2005), On latitudinal profiles of zonal means, Geophys. Res. Lett., 32, L19713, doi:10.1029/2005GL023652.

Boer, G., Lambert, S. Second-order space-time climate difference statistics. Climate Dynamics 17, 213–218 (2001). https://doi.org/10.1007/PL00013735

P. J. Gleckler, K. E. Taylor, and C. Doutriaux (2008) Performance metrics for climate models, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 113, D06104, doi:10.1029/2007JD008972

Masson, D., & Knutti, R. (2011). Spatial-Scale Dependence of Climate Model Performance in the CMIP3 Ensemble, Journal of Climate, 24(11), 2680-2692.

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling Errors in the Estimation of Empirical Orthogonal Functions, Monthly Weather Review, 110(7), 699-706.

Räisänen, J., & Ylhäisi, J. S. (2011). How Much Should Climate Model Output Be Smoothed in Space?, Journal of Climate, 24(3), 867-880.

Reichler, T., & Kim, J. (2008). How Well Do Coupled Models Simulate Today's Climate?, Bulletin of the American Meteorological Society, 89(3), 303-312.

Xu, Z. and Hou, Z. and Han, Y. and Guo, W. (2016) A diagram for evaluating multiple aspects of model performance in simulating vector fields, Geoscientific Model Development, 94365–4380 10.5194/gmd-9-4365-2016

Xu, Z. and Han, Y. and Fu, C. (2017) Multivariable integrated evaluation of model performance with the vector field evaluation diagram, Geoscientific Model Development, 10:3805–3820, doi: 10.5194/gmd-10-3805-2017

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-310, 2020.