Thank you very much for your careful reading and comments. Our point-by-point responses are as follows:

==========

Reviewer #1

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., (2017). 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 cannot 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 cannot 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.

Response:

Many thanks for introducing the references regarding the statistics that considered area-weighting. We will make a further discussion on this issue in the revised manuscript afterward. We agree with the reviewer that the sentence "*most previous model performance metrics did not consider spatial weight*" is inappropriate. Some statistical metrics did consider area-weighting and some were not. We will revise the sentence as "*The statistical metrics employed in Xu et al., (2016; 2017) did not consider spatial weight*". The detailed responses to the reviewer's comment are as follows:

As the reviewer pointed out that area-weighting was considered in previous statistical metrics, e.g., correlation coefficient, standard deviation (e.g., Watterson, 1996; Boer and Lambert, 2001; Masson and Knutti, 2011). These statistical metrics were designed to evaluate *scalar fields*. However, the statistical metrics employed in our previous papers (Xu et al., 2016; 2017), e.g., vector similarity coefficient (VSC), root-mean-square vector length (RMSL), and root-mean-square vector difference (RMSVD), which were devised to evaluate vector fields. To our knowledge, VSC and RMSL were *firstly defined* in our paper and the *area weight was not yet considered* (Xu et al., 2016). With these statistical metrics, we constructed a vector field evaluation (VFE) diagram, which can be regarded as a generalized Taylor diagram (Xu et al., 2016). The VFE diagram can be used to evaluate model performance in simulating vector fields or multiple variable fields with centered or uncentered statistics. In contrast, the Taylor diagram is a special case of the VFE diagram when the VFE diagram is applied to a scalar field with centered statistics. In the GMD manuscript, we redefine the VSC, RMSL, and RMSVD by taking area-weighting into account. More importantly, the three statistical metrics still satisfy the cosine law after considering area weight, which underpins the construction of the VFE or Taylor diagram. Thus, we can take area-weighting into account in the metrics that measuring vector statistics.

Previous studies, e.g., Taylor (2001), Boer and Lambert (2001), Gleckler et al. (2008), did mention or explicitly introduce area weight in the statistical metrics. However, these metrics are generally used to measure scalar fields rather than the vector fields. Thus, *the consideration of area-weighting in the definition of vector field statistics* is one of the novelty of this study relative to previous studies including

our previous studies (Taylor, 2001; Boer and Lambert, 2001;Gleckler et al., 2008; Xu et al., 2016; 2017).

Regarding the comment that "meridional grid size can be misleading in the evaluation of climate models is well known since at least Benestad (2005)", our responses are as follows: It is important to consider the effective sample size in the comparison of zonal mean between different latitudes (Benestad et al., 2011; Forland et al., 2011; Parding et al., 2019). However, in terms of model evaluation, we usually focus on the inter-comparison between various models rather than between different latitudes. All models are evaluated over the same domain and with the same horizontal resolution. Thus, the impact of meridional grid size on model evaluation should be less important after taking area-weighting into account.

The study of spatial smooth in climate model evaluation (Masson and Kuntti, 2011) is very interesting. Similar studies can also be carried out with the statistical metrics defined in our manuscript to investigate the impact of spatial smooth on the overall model performance in simulating multiple fields. This will be discussed in the section of discussion and conclusions in the revised manuscript.

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.

Response:

Thanks for the reviewer's insightful comment. We agree with the comment that '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'. This issue was discussed in our previous paper (Xu et al., 2017, page 3811, the paragraph about Eq. 21). We also discussed this issue in the section of summary and conclusion in Xu et al. (2017). For example, "*Unavoidably, the higher level of metrics (refer to the vector field evaluation or multivariable integrated evaluation metrics) loses detailed statistical information in contrast to the lower level of metrics (refer to the statistics for individual scalar field). To provide a more comprehensive evaluation of model performance, one can show the VFE diagram together with a table of statistical metrics (Table 1) or other model performance metrics as needed.*"

As the single model-evaluation index, which summarizes multiple statistics of multiple fields, can obscure detailed diagnostics, we included the statistical metrics of the individual scalar and vector variables (e.g., CORR, SD) in addition to the multivariable integrated evaluation index in the metric table (Table 1 in the GMD manuscript and Table 1 in Xu et al., 2017). Thus, the metric table can provide a more comprehensive evaluation of model performance.

On the other hand, the statistics for a scalar variable, e.g., correlation coefficient (CORR), standard deviation (SD), or root mean square difference (RMSD), may also obscure important diagnostics to a certain extent. For example, assuming we have two cases (Fig. R1), both have three time series from Model A, Model B, and observation O. In both cases, Models A and B have the same RMSD, SD, and CORR relative to observation. Thus, Models A and B will overlap with each other in the Taylor diagram. However, the time series in Models A and B show piecewise amplitude difference (Fig. R1(a)) and phase difference (Fig. R1(b)) from each other, respectively. Such errors are not captured by the statistical metrics, either. A similar issue also exists in any other statistical metrics, e.g., the Model Climate Performance Index (MCPI) defined by the average relative error of each variable (Gleckler, et al., 2008) and the Model Performance Index (Reichler et al., 2008). It is impossible to have one index that can measure or capture all errors of a model. Nonetheless, an index that can summarize the overall model performance is still very useful, especially for ranking models (Jury et al., 2014; Sidorenko et al., 2015, 2019; Rackow et al., 2019; Semmler et al., 2020). As shown in the metrics table in the manuscript,

the model with a higher multivariable integrated skill score (MISS) generally shows good performance in simulating individual variables, indicating the rationality of MISS.

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, section4 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 think as "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.

Response:

Thanks for the comments. We noticed that NCL has been kept in maintenance mode with no update since 2019. The NCL team still prepares maintenance releases containing critical bug fixes and user-contributed code. Meanwhile, the migration from NCL to Python is still underway. Lots of scientists and studies are familiar with NCL and still using NCL. NCL is still one of the most popular software in the community of climate science. The MVIETool provides users a convenient climate model evaluation tool for NCL users. Moreover, the NCL code and sample data also help readers to understand and test the method and develop their own codes with other computer languages.

On the other hand, we have started developing MVIETool scripts with Python, which is expected to be ready to use within one or two months for climate scientists as well as scientists from other disciplines.

Reference

Benestad R. E., Senan R., Balmaseda M., Ferranti L., Orsolini Y. and Melsom A.: Sensitivity of summer 2-m temperature to sea ice conditions, Tellus A: Dynamic Meteorology and Oceanography, 63:2, 324-337, DOI: 10.1111/j.1600-0870.2010.00488.x, 2011.

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

Forland E. J., Benestad R., Hanssen-Bauer I., Haugen J. E., and Skaugen T. E.: Temperature and Precipitation Development at Svalbard 1900 – 2100[J]. Advances in Meteorology, 2011(17).

Gleckler P. J., Taylor K. E., and Doutriaux C., 2008: Performance metrics for climate models, Journal of Geophysical Research Atmospheres, 2008, 113, D06104, doi: 10.1029/2007JD008972.

Jury M. W., Prein A. F., Truhetz H., and Gobiet A., 2014: Evaluation of CMIP5 Models in the Context of Dynamical Downscaling over Europe[J]. Journal of Climate, 2015, 28(14):5575-5582.

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

Parding K. M., Benestad R., Mezghani A., and Erlandsen H. B.: Statistical Projection of the North Atlantic Storm Tracks, Journal of Applied Meteorology and Climatology 58, 7; 10.1175/JAMC-D-17-0348.1, 2019.

Rackow T., Sein D., Semmler T., Danilov S., Koldunov N. V., Sidorenko D., Wang Q., and Jung T., 2018: Sensitivity of deep ocean biases to horizontal resolution in prototype CMIP6 simulations with AWI-CM1.0, Geosci. Model Dev., 12, 2635–2656, 2019, https://doi.org/10.5194/gmd-12-2635-2019.

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

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? Bull. Amer. Meteor. Soc., 89, 303–311, doi:10.1175/BAMS-89-3-303.

Semmler T., Danilov S., Gierz P., Goessling H. F., Hegewald J., Hinrichs C., Koldunov. N., Khosravi N., Mu L., Rackow T., Sein D. V., Sidorenko D., Wang Q., and Jung T., 2020: Simulations for CMIP6 With the AWI Climate Model AWI-CM-1-1, Journal of Advances in Modeling Earth Systems, 12, e2019MS002009. https://doi.org/ 10.1029/2019MS002009.

Sidorenko D., Goessling H. F., Koldunov N. V., Scholz P., Danilov S., Barbi D., et al., 2019: Evaluation of FESOM2.0 coupled to ECHAM6.3: Preindustrial and HighResMIP simulations. Journal of Advances in Modeling Earth Systems, 11, 3794 –3815. https://doi.org/10.1029/2019MS001696.

Sidorenko D., Rackow T., Jung T., Semmler T., Barbi D., Danilov S., et al., 2015: Towards multi-resolution global climate modeling with ECHAM6–FESOM. Part I: Model formulation and mean climate. Climate Dynamics, 44(3-4), 757–780. https://doi.org/10.1007/s00382-014-2290-6

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001.

Watterson, I. G.: NON-DIMENSIONAL MEASURES OF CLIMATE MODEL PERFORMANCE, Int. J. Climatol., 16(4), 379–391, 1996.

Xu, Z., Han, Y., and Fu, C.: Multivariable Integrated Evaluation of Model Performance with the Vector Field Evaluation Diagram, Geosci. Model Dev., 10, 3805–3820, 2017.

Xu, Z., Hou, Z., Han, Y., and Guo, W.: A diagram for evaluating multiple aspects of model performance in simulating vector fields, Geosci. Model Dev., 9, 4365–4380, https://doi.org/10.5194/gmd-9-4365-2016, 2016.

Figure

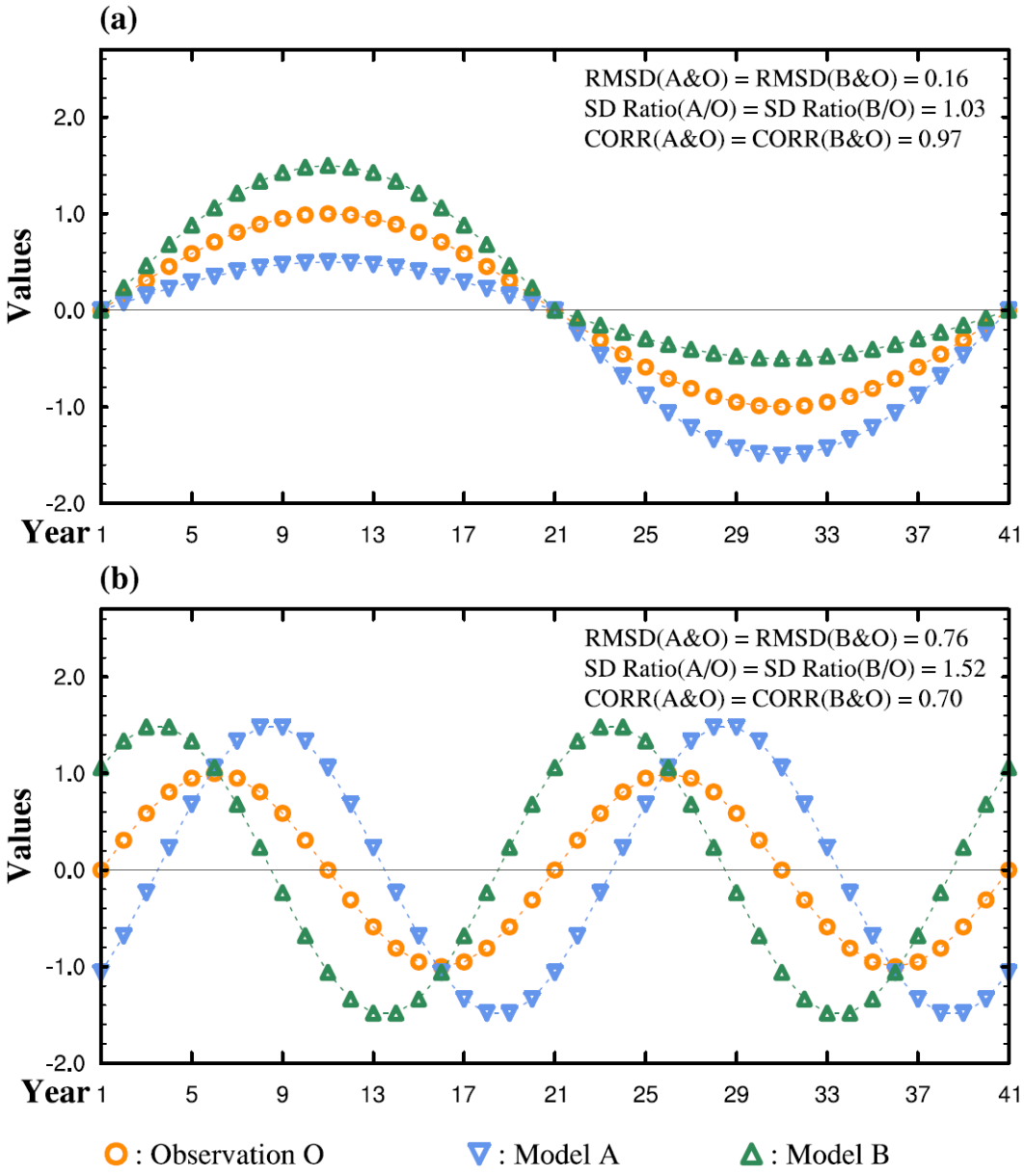

Figure R1. Two examples illustrate model errors that are not captured by the commonly used statistical metrics. Each example is composed of three time series from idealized model A (blue upper triangle), model B (green lower triangle), and observation O (orange circle), respectively. Compared to O, Model A and B show different errors, but they have the same RMSD, SD, and CORR.