Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-310-RC2, 2021

Geoscientific Model Development Discussions

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 #2

Received and published: 20 January 2021

The authors describe the extension of a method to assess the performance of climate models in simulating scalar or vector fields based on the concept of the vector field evaluation (VFE) first introduced by Xu et al. (2016). In addition, the authors describe a method summarize a model's ability to simulate multiple variables by introducing the multivariable integrated skill score (MISS).

The manuscript is generally well written and I suggest minor revisions to the manuscript before publication in Geoscientific Model Development addressing the points given below. I do not agree with reviewer #1 that the paper does not provide enough novelty for publication in GMD. In my opinion, extending the widely used Taylor

diagrams to include area weighting and proposing a new integrated measure of a model's performance across variables while giving the user the possibility to adjust the relative importance of RMS and VSC is very welcome. Compiling all metrics into a tool for model evaluation and making it available to other users is worth publishing this description of the tool and the methods used.

I do think, however, that the descriptions of the methods and the tool itself lack some detail. I also have the impression that the MVIETool does not live up to its full potential and could strongly benefit from implementing the routines into the framework of existing model evaluation software. I see this, however, as a potential future pathway and not as a prerequisite for publication. More detail is given in the general comments below.

General comments

- Regridding and masking are important processing steps that are not explained in enough detail. For example, it is not clear to me whether all variables from the same source (model or observations) have to be on the same grid (horizontal and vertical). Simply referring to external software such as CDO is not enough and a concrete example should be given (e.g. to reproduce the figures shown). The same is true for masking of missing values. How is this done? For example, is a mask generated for each time step and each dataset? Are all datasets used to create a common mask that is then applied to all datasets or is the mask created separately for each model-observations pair? If masks are generated separately for each model-observations pair, what does this mean for the comparability across different models? It should become clear how data have to be

preprocessed and which implications this might have on comparing across different models and/or observational datasets. I recommend adding some discussion on this issue.

- Considering observational uncertainties in model evaluation is of fundamental importance. The approach taken here by using the average of possibly available multiple observational datasets as reference data seems very basic. What effect does this averaging have on the skill scores? I would expect this kind of averaging to reduce the spatial and/or temporal variability of the reference data compared to the individual observational datasets and thus have an impact on the skill scores. Also, are there ways to include possibly available uncertainty information on a per pixel basis (e.g. standard error provided with some ESA CCI satellite datasets)? At least a brief discussion on thoughts on this topic should be added.

- Is there a way to visualize observational uncertainty e.g. in the VFE diagrams in addition to showing individual observational datasets against the reference dataset, e.g. by shading the area representing the uncertainty range in the diagrams?

- A weighting factor F has been introduced but is not discussed. In l. 430, the authors state that "The factor F in cMISS and uMISS is 2". What is the reasoning for this choice? What is recommended to users wanting to apply the MVIETool? Maybe give some examples for specific applications.

- How is the grid cell area calculated that is used as weighting factor? Again, the statement (l. 189/190) that "If users want to consider area weighting in the statistics, the variables should be saved with the coordinate information (e.g., time, latitude, and longitude)" does not provide enough detail. How do the coordinates have to be defined? Is following the CF standard sufficient? Do the coordinates have to follow the CMOR conventions? Can area files ("fx" files in CMIP) be provided e.g. for irregular grids or is the analysis limited to regular grids?

- It is also not clear to me how time series of variables are handled. For example, additional information on (possibly) selecting a user specified time range is needed. Are attributes of the time coordinate such as calendar taken into account when calculating time means (e.g. number of days per month)? Is temporal interpolation done if the time resolution of two datasets does not match?

- I was missing an overview (e.g. a table) on which models, experiments, years, time resolution, etc. of the model data and which reanalysis datasets have been used to create the example figures. This makes it impossible to reproduce the examples as an independent check (i.e. downloading the data yourself, applying the preprocessing steps and running the MVIETool) that the software is working as expected.

- I have the impression that the MVIETool would benefit substantially from taking advantage of the infrastructure of existing model evaluation tools such as, for instance, the ESMValTool (Righi et al., Geosci. Model Dev., 2020). Such tools provide the possibility to preprocess all datasets in a consistent way regarding checking of input data, horizontal and vertical regridding, masking, time selection, vertical level selection, etc. I would like to encourage the authors to add some discussion on such a step as a possible outlook to the summary section.

- It becomes increasingly more important to provide traceable and reproducible results. For model evaluation, this usually means providing a provenance record of the input data, used software, configuration, processing steps, etc. Is anything like this planned for the MVIETool? Again, I feel that the MVIETool could strongly benefit from taking advantage of the infrastructure of existing model evaluation tools that are already capable of providing provenance records.

- Are there plans to continue development of MVIETool? I would recommend to add an outlook and thoughts about possible future directions to the summary section.

Specific comments

- l. 16, ”MIEI” has not been defined yet
- l. 44, ”most previous model performance metrics did not consider spatial weight”: while this statement is true for the three examples mentioned, this is not the case for many other metrics and therefore needs rephrasing
- l. 65, ”... by dividing the corresponding rms value of the observation ...’: it is not entirely clear to me what is divided by what, please consider rephrasing; if the original variable is divided by the rms of the observations, are all data on the same spatial and temporal grid? If not, what are the effects of this? What is the effect of averaging possibly available multiple observational datasets (see also general comments)?
- l. 188: ”variable” → ”variables”
- l. 196/197: What is meant by ”put in parenthesis”? Where does this have to be done (source code, namelist, etc.)? Please be more specific.
- l. 215: What is meant by ”standardize the missing points”? Does this mean a common mask is created from all missing grid cells/time steps in all datasets (models + reference) across all variables? Please give more details on how masking is done (see also general comments).
- l. 233: ”one piece of observational data” → e.g. ”one observational dataset”?
- l. 234: ”written in a new NetCDF file” → ”written to a new NetCDF file”
- l.258/258: What is meant by ”better model performance” and ”worse model performance”? Better of worse relative compared to what? Please be more specific and if possible more quantitative.

- l. 272: ”relative” → ”compared”

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