Thanks very much for the insightful comments. The comments are very helpful not only for improving this manuscript but also for our future study. Our point-by-point responses are as follows:

==========

Reviewer #2

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.

Response:

Thanks for the reviewer's comment. We will add more detailed explanations on the regridding and masking of missing values in the revised manuscript. Our responses to these questions are as follows:

The MVIETool requires all the variables on the same grid for datasets of all models and observations. We will give an example to illustrate how to regrid data with CDO in an updated pdf file (graphic_guide_MVIETool.pdf). The regridding can be done with one command line of CDO. For example, the following CDO command can interpolate input data (time,lev,lat,lon) to a 1.25°×1.25° longitude-latitude grid and 14 vertical pressure levels:

```
cdo remapbil, r288x145 –intlevel,100000,925000,85000,70000,60000,
50000,40000,30000,25000,20000,15000,10000,7000,5000 input.nc
output.nc
```

In terms of masking, the present MVIETool generates a common mask for each model-observation pair. The mask is the same for all variables in this pair of data. If

more than one observational data are available, the tool will generate a common mask for the missing grids first for the observational datasets before the evaluation. In doing so, we can take full advantage of the model output. However, it makes the evaluation less comparable between different models. In the revised manuscript, we will generate a common mask for all models and observation as a default option.

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?

Response:

Thanks for the reviewer's comments about observation uncertainties in model evaluation. The average of multiple observation datasets may reduce the spatial and/or temporal variability of the reference data compared to the individual observational datasets. This impact can be roughly estimated by the individual observational datasets in the VFE diagram. For example, the cRMSL is slightly greater than 1 for REA1 (point 11) and REA2 (point 12) in Fig. 7 of the manuscript, which indicates that the average of two reanalysis data leads to a slight reduction in spatial variability. However, this reduction is very small and its impacts on the evaluation should be neglectable. If the cRMSDs of individual observational data are clearly greater than 1 for individual observation, one should not use the average of multiple observation datasets as reference. Currently, users can use one of the observational data as reference. This issue will be discussed in the revised manuscript. We are also considering other ways to estimate the observational uncertainty. A preliminary idea is as follows: Assuming we have K observational dataset. Based on

K observational data, we can compute the VFE statistics K times and get K points in the VFE diagram for one model. Afterwards, we create a shaded patch using the K points. The area of the patch can represent the impact of observational uncertainty on the evaluation. The approach may only work well when the VFE diagram only show a few models. Otherwise, it would be hard to discriminate one model from others if many shaded patches overlap together.

In terms of the observational data that already has an uncertainty estimation (standard deviation in each grid point), we will add a horizontal bar centered at reference point on the x-axis. The length the bar represents the mean standard deviation of the observation, which can roughly represent the mean spread of observational datasets.

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.

Response:

Thanks for the comment. MISS is defined based on the MIEI (Eq. 7a in the manuscript) proposed by Xu (2017). MISS is equal to MIEI when factor *F* is 2. MIEI represents the length of line segment CG in Figure 3 in Xu et al., (2017). MIEI (MISS) has a geometric meaning when *F* is equal to 2. Meanwhile, the MISS is more sensitive to the changes in VSC than RMSL when *F* is equal to 2. As climate models can often reasonably reproduce the pattern similarity, it is getting harder to improve pattern correlation when the correlation coefficient is greater than 0.8 or higher. Thus, it is usually more desirable to give VSC more weight. F is a flexible factor; user can modify its value based on the applications. We will make more discussion on the factor *F* in the revised paper.

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?

Response:

Thanks for the comment about the weighting factor in MVIETool. At present, the tool can only deal with the area weighting for regular grids and the area weighting is calculated by the tool with the equation $sin(lat+d_{lat})- sin(lat-d_{lat})$, where d_{lat} is the grid distance in latitude. Hence, variables should first be defined with dimension names and assign the coordinate variables (referring to *http://www.ncl.ucar.edu/Document/Language/cv.shtml*). The CF standard is sufficient. The requirement for coordinate information will be explained in the revised paper. In addition to determining the area weight, the coordinate information is also used to select sub-regions for a regional evaluation. One has to regrid all data (model and reanalysis) into a common resolution before the computation of statistical metrics. Thus, one can regrid all data (in a regular or irregular grid) to a regular grid with coordinate information.

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?

Response:

Thanks for the comment. The MVIETool assumes all input variables on the same grid and with the same coordinate variables, including time, latitude and longitude grid. Otherwise, the processing will break out and report error. In terms of the selection of time range, the tool provides two options: First, one can specify the time range with strings in the format: "YYYYMM", "YYYYYMMDD", or "YYYYMMDDHH", where YYYY is the year, MM is the month, DD is the day, and HH is the hour, e.g., "198101:199012". In this case, the coordinate variable assigned to the time dimension of input variables should have a "calendar" attribution (referring to *http://www.ncl.ucar.edu/Document/Functions/Built-in/cd_calendar.shtml*). Second,

one can specify the time range with the time steps, e.g., "1:10". We will illustrate how to setting time coordinate in the pdf file (graphic_guide_MVIETool.pdf) as well.

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.

Response:

Thanks for the comment. The reason we did not give the model name in the manuscript is that the ranks of the models' performance depend on the variables, seasons, and regions evaluated. The model showing good (or poor) performance does not necessarily mean a good (or poor) performance for other variables, seasons and regions. We show the examples to illustrate the methods rather than evaluate specific models. We can present the model name, institution and horizontal resolution of 10 CMIP5 models used in the revised manuscript (attached in Table.R1 below) if it is necessary. We used the monthly mean datasets derived from the first ensemble run of historical experiments during the period from 1961 to 2000 (L244).

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.

Response:

Thanks for the reviewer's constructive comments. We will carefully consider these suggestions in our future work. Currently, the MVIETool only provides some basic functions to calculate the relevant statistics and generate figures. We will try to take advantage of the infrastructure of existing model evaluation tools for further improvement. Our follow-up work intends to devise a significance test method for the difference between two vector fields (and two MISSs). If we can make significant progress, we will incorporate these significant tests into the MVIETool in the future. An outlook of plans and future directions of the MVIETool will be added to the summary section in the revised paper.

L. 16: "MIEI" has not been defined yet

Response:

The MIEI will be defined in the revised manuscript.

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

Response:

Thanks for the comment. We will revise the sentence as "The statistical metrics employed in Xu et al., (2016; 2017) did not consider spatial weight".

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)?

Response:

Thanks for the comment. The MVIETool requires all input data on the same spatial and temporal grid. We will rephrase the sentence as “We need to normalize

each variable derived from the model by dividing the rms value of the corresponding variable derived from observation”

L. 188: "variable" →"variables"

Response:

We will replace “variable” with “variables” in the revised paper.

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.

Response:

The MVIETool treats the variables in parenthesis as a vector field, e.g. (ua, va). The variables in the parenthesis represent the different components of a vector field. These variable names are specified in the `namelist` part of the tool. This will be clarified in the revised manuscript.

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).

Response:

Yes, it means generating a common mask for the missing grid. We will reword the sentence and explain more details on the generation of masks in the revised manuscript.

L. 233: "one piece of observational data"→ e.g. "one observational dataset"?

Response:

We will make the replacement in the revised paper.

L. 234: "written in a new NetCDF file" →"written to a new NetCDF file"

Response:

We will make the replacement in the revised paper.

L.258/258: What is meant by "better model performance" and "worse model performance"? Better or worse relative compared to what? Please be more specific and if possible more quantitative.

Response:

Thanks for the comment. "better model performance" and "worse model performance" are determined based on the rank of model performance metrics. The color represents the value of the model performance metrics. This will be clarified in the revised manuscript.

L. 272: "relative" →"compared"

Response:

We will replace “relative” with “compared” in the revised paper.

Reference

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.

Table

Table R1. Model names, institution and horizontal resolution for 10 CMIP5 models (M1–M10) used in the paper.

	Model	Institution	Horizontal resolution
M1	BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University(China)	2.81° × 2.81°
M2	CCSM4	NCAR (National Center for Atmospheric Research) Boulder(USA)	1.25°× 0.94°
M3	CNRM-CM5	Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation	1.41° × 1.41°

		Avancees en Calcul Scientifique(France)	
M4	BCC-CSM1-1	Beijing Climate Center, China Meteorological Administration(China)	2.81° × 2.81°
M5	FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences; and CESS, Tsinghua University(China)	2.81° × 3.05°
M6	GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory(USA)	2.5° × 2.0°
M7	GISS-E2-H	NASA Goddard Institute for Space Studies(USA)	2.5° × 2.0°
M8	MIROC4h	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology(Japan)	0. 56° × 0.56°
M9	MIROC-ESM-CHEM	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology(Japan)	2.81° × 2.79°
M10	inmcm4	Institute for Numerical Mathematics(Russia)	2.0° × 1.5°