Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-95-AC1, 2017

Geoscientific Model Development Discussions

Interactive comment on "Multivariable Integrated Evaluation of Model Performance with the Vector Field Evaluation Diagram" *by* Zhongfeng Xu et al.

Zhongfeng Xu et al.

xuzhf@tea.ac.cn

Received and published: 28 July 2017

We would like to thank the reviewer for the valuable comments. Our point-by-point responses to the comments are detailed in the following pages.

==========

Reviewer 1

Multivariable Integrated Evaluation of Model Performance with the Vector Field Evaluation Diagram by Z. Xu, Y. Han, and C. Fu

The authors describe a method to assess the performance of climate models in simulating an arbitrary number of equally important variables based on the concept of

the vector field evaluation (VFE) introduced by Xu et al. (2016). In addition, the authors describe a method to collapse the three different metrics root mean square length (RMSL), vector field similarity coefficient, and root mean square vector deviation (RMSVD) into a single index that can be used to rank the models by overall performance for the set of given variables. 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.

General comments As an example application of the model evaluation method presented in this paper, three different temperature / precipitation datasets are averaged as reference dataset. Datasets such as the CRU 2m temperature data typically contain missing values (also over land). How are missing values being treated in this study? Also, are the grid cells weighted by their surface area? This would be needed to make sure that the skill scores are representative for the global mean values of the respective quantities. If not, the calculated average metrics would be very hard to interpret as e.g. grid cells in polar region would receive more relative weight than grid cells in low latitudes. As the method presented here suggests evaluation of global averages, I would like to see one or two sentences on this issue. Also, please be more specific on the processing of the observational data regarding e.g. missing values.

RESPONSE: The CRU dataset used is CRUTSv3.24 which was constructed to "provide full coverage of the specified continental land area, with no missing data (Harris et al., 2014)". We checked the data and confirmed that there is no missing value over the land area. The data was assigned to climatological mean value if missing data presents during the construction of CRU datasets (Harris et al., 2014). We also checked the other two pairs of temperature and precipitation datasets; there is no missing data on the continents, either. Harris et a.l, 2014: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. Int. J. Climatol. 34: 623–642.

The grid cells were not weighted by their surface area in the previous manuscript. We agree with the reviewer that area weighting should be considered here. In the revised

manuscript, we take area weighting into account to make the statistics be representative for the global mean values as the reviewer suggested. All related texts (P9, L22-23), figures, and tables are updated. Thanks for the comment!

Some parts of the manuscript are somewhat repetitive and could be shortened. For example on p.8, l. 9-10 it reads "Thus, three statistical quantities can be indicated by a single point on the VFE diagram", which is almost identical to p.10, l. 10-12: "[...] the three quantities [...] can be represented by a single point for each model on the VFE diagram." or to p. 10, l. 13-15: "Thus, each point on the VFE diagram can represent the overall performance of an individual model in terms of 3 statistical quantities [...]". I suggest to go through the text and remove such repetitions where possible.

RESPONSE: The first sentence is retained in the revised manuscript. The second sentence is reworded as "Thus, each model's performance in simulating multiple variables can be summarized by a single point that is determined by 12 statistical quantities (4 variables $\times$ 3 statistics) those derived from various scalar fields". The third sentence is deleted. We go through the whole manuscript and remove or rewrite a few sentences those read repetitive. In addition, we replace σ_ai in Eq. 13 with L_ai because the definition of σ_ai is the same as L_ai (Eq. 5).

The authors propose to include the standard deviation of the RMS values of multiple scalar fields into the VFE diagram as an additional performance measure (p. 11, l. 12-14). It remains unclear to me whether the length of the proposed additional line segments in figure 3 are σ_RMS or actually $\pm\sigma$_RMS, i.e. 2σ_RMS. Or did the authors mean variance of the RMS values (equation 23)? Please clarify.

RESPONSE: In the revised manuscript, we clarified "The length of the line segment is equal to twice the standard deviation of RMS values of multiple scalar fields" in P11,

L5 in the revised manuscript. For consistency purpose, equation 23 is rewritten in the form of standard deviation rather than variance, because standard deviation is used in Figs. 2, 4.

A proper evaluation of the performance of climate models usually requires to take into account observational uncertainties. Differences between models and observations can only be interpreted as model errors or lack of model skill if the differences are larger than the observational uncertainty. This is particularly the case for variables with a large uncertainty such as, for instance, ice water path, but also important when ranking models by performance. More and more observational datasets provide estimates of the observational uncertainty. What are the authors' thoughts about including such additional information into their calculations, in particular when calculating skill scores such as the presented multivariable integrated evaluation index (MIEI) that is then used to rank models according to their average performance skill?

RESPONSE: Thanks for the valuable comment. We further discuss how to evaluate the impact of observational uncertainties on model evaluation and ranking in the revised manuscript. To take the advantage of observational uncertainty, one can generate a number of ensemble members of observational estimates using the estimates of the observational uncertainties. Some datasets, e.g., HadCRUT4, already provided such ensemble observational estimates.

We add a new paragraph to discuss how to take observational uncertainty into account in our model evaluation methods (P13, L7-20). The new paragraph added to the revised manuscript is pasted below:

"How to take the observational uncertainties into account is of great importance in model evaluation and ranking, especially when more and more observational datasets provide estimates of the observational uncertainty. The statistics derived from each group of observational estimates are also shown in Table 1, which can roughly quan-

tify the observational uncertainties and its impact on model evaluation. Generally, the colours are clearly lighter for the statistics of individual observed variables in contrast to the modelled variables (Table 1). This indicates that the observational uncertainties are relatively small and should have less impact on the evaluation of model performance. To further quantify the impacts of observational uncertainty on ranking model performance, we calculate the MIEIs of various climate models by taking each group of observational estimates as the reference data. Three groups of observational estimates generate three groups of MIEIs. Afterwards, we calculate Spearman's rank correlation coefficient of each group of MIEIs with those derived from models and ensemble mean of multiple observational estimates. The Spearman's rank correlation coefficients are 0.996, 0.996, and 0.904, respectively, suggesting that the ranks are very close to each other no matter which group of observational estimates is used as reference data. Thus, the observational uncertainty should have less impact on ranking model performance in this case. One can use the average of Spearman's rank correlation coefficients to quantify the consistency of various ranks when a number of observational estimates are available."

Specific comments p. 1, l. 21: "[...] evaluation of model performance."

RESPONSE: Done

p. 10, l. 5: "what is meant by "summer SAT and precipitation"? Is this an average over the months June, July, August? Please be more specific.

RESPONSE: Done

p. 12, l. 20, "In comparison with the RMSVD, [...]": did you mean "In contrast to [...]"?

RESPONSE: Yes, and "In comparison with" was replaced with "In contrast to"

p. 13, l. 2, "Index"→"index"

RESPONSE: Done

p. 13, l. 5, "[...] but a larger MIEI relative to [...]": did you mean "compared to"?

RESPONSE: Yes, and "relative to" was replaced with "compared to"

p. 26, l. 3, "CMIP5 model" →"CMIP5 models"

RESPONSE: Done

p. 26, table 1: it would be interesting to add the performance of the individual observational datasets to the table as a "rough estimate of the observational uncertainties" as stated on p. 9, l. 22-23.

RESPONSE: The performance of the individual observational datasets is added to Table 1 in the revised manuscript. P13, L8-12.

p. 27, caption of figure 1, l. 4: delete "apart"

RESPONSE: Done

figure 2 is fully included in figure 3 and could be deleted

RESPONSE: Figure 2 is deleted in the revised manuscript.

p. 31, figure 5: the second level of metrics includes σRMS while the caption and the referenced equation 23 specify the variance of RMS values (σ2_RMS). Which one is correct? Is there a "2" missing?

RESPONSE: We rewrite Eq. 23 in the revised manuscript. σ_RMS, instead of σ2_RMS, is used because the line segments on Fig.3 represent twice the standard deviation of RMS values. Fig. The text and figure captions are also updated accordingly.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-95, 2017.

Table 1. Multiple statistics of CMIP5 models in simulating surface air temperature and precipitation in terms of climatological mean state and interannual variability. Tm (Pm): climatological mean surface air temperature (precipitation) in summer. Ta (Pa): temporal standard deviation of summer surface air temperature (precipitation). CMIP5 simulations and three individual groups of observational datasets are compared with the ensemble mean of three groups of SAT and precipitation data observed during the period from 1961 to 2000. RMS: the ratio of modeled to observed root mean square (RMS) values of the spatial pattern for each variable. CORR (RMSD): uncentered spatial correlation coefficient (root mean square deviation) between model and observational fields. RMSL, Rv, RMSVD measure the statistics of two vector fields, which can represent the overall statistics of all fields (Eqs. 3, 13, 16). RMSL was shown as the ratio of model simulated RMSL to the observed RMSL. RMS_stddev is the standard deviation of four RMS values, which describe the dispersion of RMS values of Tm, Pm, Ta, and Pa (Eq. 23). MIEI: multivariable integrated evaluation index (Eq. 24). Model performance is indicated by the color scale: lighter colors denote better model performance.

METRICS	RMS				RMSL	CORR				Rv	RMSD				RMSVD	RMS_std	MIEI
	Tm	Pm	Ta	Pa		Tm	Pm	Ta	Pa		Tm	Pm	Ta	Pa			
Model 1	1.01	0.99	1.11	0.97	1.02	1.00	0.92	0.95	0.92	0.94	0.09	0.41	0.35	0.39	0.34	0.05	0.34
Model 2	1.04	1.01	1.13	1.05	1.06	1.00	0.83	0.93	0.84	0.90	0.10	0.58	0.42	0.58	0.46	0.04	0.44
Model 3	1.05	0.80	1.26	0.77	0.99	1.00	0.88	0.93	0.86	0.91	0.11	0.48	0.48	0.52	0.43	0.20	0.48
Model 4	0.97	0.84	1.17	0.72	0.94	1.00	0.91	0.95	0.87	0.92	0.09	0.43	0.38	0.52	0.39	0.17	0.44
Model 5	0.99	0.94	1.43	1.19	1.15	1.00	0.86	0.95	0.84	0.90	0.10	0.51	0.58	0.64	0.50	0.19	0.51
Model 6	1.05	0.97	0.97	0.83	0.96	1.00	0.90	0.96	0.90	0.94	0.09	0.43	0.29	0.44	0.34	0.08	0.35
Model 7	1.06	1.09	1.14	1.07	1.09	1.00	0.90	0.93	0.89	0.93	0.11	0.47	0.41	0.48	0.40	0.03	0.37
Model 8	1.02	0.97	1.12	1.00	1.03	1.00	0.89	0.96	0.87	0.93	0.08	0.46	0.33	0.51	0.38	0.06	0.37
Model 9	0.92	0.91	1.01	0.66	0.88	0.99	0.86	0.93	0.87	0.91	0.14	0.51	0.37	0.54	0.42	0.13	0.46
Obs1	1.00	0.99	0.97	1.07	1.01	1.00	0.99	0.98	0.98	0.99	0.03	0.13	0.18	0.21	0.16	0.04	0.23
Obs2	0.99	1.01	0.99	1.00	1.00	1.00	0.99	0.99	0.98	0.99	0.04	0.11	0.16	0.18	0.13	0.01	0.22
Obs3	1.01	1.02	1.10	0.97	1.02	1.00	1.00	0.97	0.99	0.99	0.04	0.09	0.27	0.16	0.17	0.05	0.24

Fig. 1. Table 1

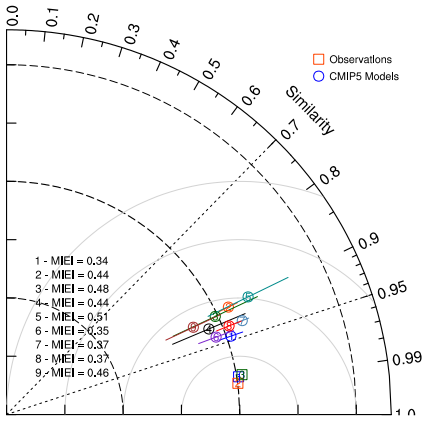

Figure 2: VFE diagram describing the normalized climatological mean SAT, precipitation, and interannual variabilities of SAT and precipitation over a land area between 60°S–60°N simulated by 9 CMIP5 models compared with three groups of SAT and precipitation data observed during the period from 1961 to 2000. The RMS length and the RMSVD have been normalized by dividing the RMS length derived from the observed data. The line segment centered at each plotted point along the azimuthal direction represents twice standard deviations of the RMS values of various fields. The value of the MIEI for each model is also shown in the diagram.

Fig. 2.

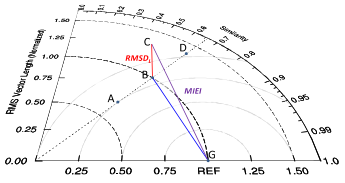

Figure 3: Schematic diagram displaying the relationship between the RMSVD, $RMSD_L$, and MIEI. The points A, B, and D represent different models. The $RMSD_L$ measures the overall difference between the modeled RMS values and the observed ones. The line segment BC is vertical with respect to the VFE diagram. The length of line segment BG is determined based on the vector field similarity, which measures the overall pattern similarity of various scalar fields relative to the observed ones. Thus, the MIEI index takes both the pattern similarity and the RMS values of various scalar fields into account.

Fig. 3.

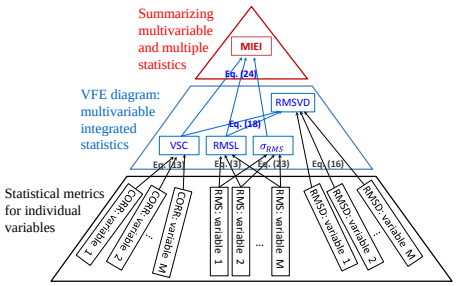

Figure 4: Pyramid chart showing the relationship between three levels of metrics. The first level of metrics, i.e., correlation coefficient (R), RMS value, and RMSD, measures the model performance in terms of individual variables. The second level of metrics, i.e., VSC, RMSL, standard deviation of RMS values (σ_{RMS}), and RMSVD, is derived from the first level of metrics and summarizes the overall performance of a climate model in simulating multiple fields. The MIEI further summarizes the VSC, RMSL, and σ^2_{RMS} into a single index to rank various climate models in terms of simulating multiple fields.

Fig. 4.