Author's response to anomalous reviewer 1

Thanks for the reviewer's valuable comments. Our point-by-point responses to reviewer's comments are listed below.

• *A paragraph giving some guidance on the scientific interpretation of the proposed statistics would be helpful, possibly with examples demonstrating how "good" and "not so good" agreement between two data sets looks like. It would also be helpful to explicitly point to possible issues and limitations to keep in mind when looking at complex quantities such as the skill scores proposed by the authors (equations 13 and 14).*

The VFE diagram includes three statistical variables, i.e. the vector similarity coefficient (VSC), root-mean-square length (RMSL), and root-mean-square vector difference (RMSVD). VSC was proposed in Section 2 of this study. A scientific interpretation of VSC was presented in Section 3 by discussing how the changes in vector length and angle will affect VSC. Three examples were also given in section 3.1, 3.2, and 3.3 to help readers to understand the meaning of VSC and its performance in describing the similarity of two vector fields. We have made some modifications to section 3 in the revised manuscript to improve the scientific interpretation of the VSC. In terms of RMSL and RMSVD, we did not make further interpretation because they are very easy to understand.

One possible issue of the VFE diagram is pointed out in section 5.2.3 in the revised manuscript. Namely one should be cautious when using RMSVD measures model performance because a smaller RMSVD does not necessarily indicate a better model skill. Following Taylor (2001) we proposed two model skill scores to measure model skill. Meanwhile, we also pointed out the caveat of the skill scores defined by equation 13 and 14. Namely, users should define or select appropriate skill scores based on their own applications because no skill score would be universally considered most appropriate.

• *An application of the proposed VFE diagram to evaluate the performance of models usually requires taking into account observational uncertainties as the reference data (at x=1, y=0; e.g., figure 9) are usually not the truth. This is particularly the case for quantities that have larger uncertainties than 850-hPa wind speed. I would like to see a discussion of possible issues and limitations as well as thoughts of how to deal with observational uncertainties in this type of diagram.*

We add a new section (section 6) to address how to take the observational uncertainty into account in the model evaluation. We present two approaches to show the observational uncertainty in the VFE diagram. One is Taylor's approach by showing the statistics of models relative to various observations in the VFE diagram. The other approach is comparing models and individual observational estimates with the mean of multiple observational estimates. The spread of various observational estimates can be taken as an indicator of observational uncertainty. The observational uncertainty and model statistics can be shown on the VFE diagram.

• *What are the key messages of section 3.2 (relationship of VSC with MDA)? Things need to be put into context by providing a motivation for this analysis. The statements made have to be more precise, for example, "[...] a smaller MDA generally corresponds to a larger* R_v*, and* *vice versa." does not seem to provide a lot of useful information for the application and interpretation of the VFE diagram. Again, guidelines of how to interpret the proposed statistics in a scientifically meaningful way would be helpful.*

The section 3 presents the interpretation on VSC and includes three subsections. Given the fact that VSC is determined by both vector lengths and vector directions, we present how VSC is affected by MSD and MDA in section 3.1 and 3.2, respectively. As defined by equation (10) and (11), MSD measures the mean square difference of normalized vector lengths. MDA describes the mean difference of angles of paired vectors between two vector fields. The discussion in section 3.1 and 3.2 would help readers to understand the nature of VSC and whether or not VSC can reasonably reflect the similarity of two vector fields. Section 3 is properly revised to provide the motivation of the analysis and improve the scientific interpretation of the VSC. For example, we present the motivation of section 3.2 in the revised manuscript with the sentences "In previous section, the interpretation of VSC is based on the assumption that the paired vectors have a constant included angle. In this section, we will examine how VSC is affected by the difference of included angles in a more general case". We make more precise statement in the revised manuscript, e.g. "A smaller MDA indicates smaller differences in the directions of paired vectors and hence a higher similarity between the vector fields $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$, suggesting that VSC can reasonably describe how close the vector directions between two vector fields are.". In addition the title of section 3.1 and 3.2 are changed to "Interpreting VSC based on its equation" and "Interpreting VSC based on random generated samples", respectively.

• *Section 5.1 gives an example for the application of the VFE diagram using 850-hPa wind speed from 19 CMIP5 models. There are, however, no details on the model runs used (model experiment, ensemble members, etc.). It is also not clear to me what exactly you are comparing (multi-year annual means, monthly means, etc.). I presume the models have been regridded to a common grid? If so, which grid and which interpolation method has been applied? Also, I am missing a reference for the NCEP reanalysis data used. This section needs some rewriting to make clear what exactly has been done and what is being compared here. The current description is not sufficient to reproduce any of the results shown here, which is not acceptable for a scientific paper.*

In the revised manuscript, we present detailed information about the data used, regrinding method, and what we are comparing in the text and figure captions. For example, "As an example, we assess the pattern statistics of climatological mean 850-hPa vector winds derived from the historical experiments by 19 CMIP5 models (Taylor et al., 2012) compared with the NCEP-DOE reanalysis 2 data during the period from 1979 to 2005. The evaluation was based on the monthly mean datasets from the first ensemble run of CMIP5 historical simulations and all datasets were regrided to a common grid of 2.5°×2.5°. Box averaging (bi-linear interpolation) method was used to regrid the reanalysis data and model data to a coarse (finer) resolution.". References for the NCEP reanalysis datasets and other reanalysis datasets are also presented in the revised manuscript.

• *Section 5.2.2 (statistical significance of differences): I am missing a clear definition of what the authors mean by "statistical significance of differences" in the context of 2-dim vector*

fields. The argument of ”separated groups” without further explanations or governing equations is not precise enough for a scientific paper. I have the impression that the authors are rather speculating here than presenting any scientific evidence. This leads to contradictions within the section that need to be addressed. For example, the authors claim on p. 10, l. 27-28: ”Thus, the differences between models 12, 13 [...] are likely to be significant.”. The authors contradict this statement a few lines later (p. 11, l. 1-2): ”[...] which may not be sufficient to conclude a significant difference between the three models, especially for models 12 and 13.” leaving the reader confused. Also, it is once again not clear what data have been used and what quantities are being compared (models, model experiments, ensemble members, time period, averaging, time resolution, regridding, etc.).
The title of section 5.2.2 is reworded as “statistical significance of differences in model performance” to make it precise. We present more discussions on the statistical significance of differences in model performance in section 5.2.2. As discussed in Appendix A, the VFE diagram is a generalized Taylor diagram. Both are constructed based on the cosine law. Therefore the significance of model performance presented in this study can be understood in the same way as that in Taylor (2001, section 4.1). In terms of the “contradictions” of our discussion on the significance of model performance, we were trying to make an objective interpretation. The differences between models 12, 13, and 14 are “likely” to be significant because their statistics are clearly separated from each other. However, under certain circumstance, two groups of statistics are separated from each other may not be sufficient to conclude a significant difference, e.g., the sample size is small. This is similar to the significance test for scalar variables. The difference between two group data may not be statistically significant even their mean value show a clear difference when the sample size is small. Currently the VFE diagram can only present a qualitatively evaluation on the statistical significance of differences in model performance. No quantitative evaluation is available yet. This is a drawback of our method and warrants for further study. We properly reworded our discussions in the revised manuscript. The data information is also provided in the caption of figure 8.

Specific comments

- *p. 5, l. 24, give equation numbers for RMSL*

  Done

- *p. 7, l. 8: replace ”for a certain angle” with ”by a certain angle”*

  Done

- *p. 7, l. 9: production → generation*

  Done

- *p. 7, l. 21: what do you mean by ”the performance of R_v”?*

  The sentence has been reworded as “The purpose of this analysis is to further illustrate whether Rv can well measure the similarity of two vector fields or not with observational data.”

• *p. 9, l. 8: replace "a dotted contour" with, for instance, "dotted circles"*

Done. "a dotted contour" is replaced with "dotted circles"

• *p. 9, l. 9: "line" →"lines"*

Done

• *p. 9, l. 15, insert "VFE" before "diagram"*

Done

• *p. 11, l. 30: "anomalous scalar fields" →"scalar anomaly fields"*

Done

• *p. 13, equation (A1): how did you get from line 2 to line 3? Shouldn't* $(\overline{x}_{ai} + x'_{ai})^2 = \overline{x}_{ai}{}^2 + 2\overline{x}_{ai}x'_{ai} + x'_{ai}{}^2$*? I.e., what happened to the term* $2\overline{x}_{ai}x'_{ai}$ *?*

The term $\sum_{i=1}^{N} \overline{x}_{ai}x'_{ai}$ does no appear in equation (A1) because it equals 0. Equation (A1) is written as follow:

$$L_A{}^2 = \frac{1}{N}\sum_{i=1}^{N} |\vec{A}_i|^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left((\overline{x}_a + x'_{ai})^2 + (\overline{y}_a + y'_{ai})^2\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\overline{x}_a{}^2 + \overline{y}_a{}^2) + \frac{1}{N}\sum_{i=1}^{N}(x'_{ai}{}^2 + y'_{ai}{}^2) + \frac{1}{N}\sum_{i=1}^{N}(2\overline{x}_a x'_{ai} + 2\overline{y}_a y'_{ai})$$

The third term on the right-hand side of the equation can be written as:

$$\frac{1}{N}\sum_{i=1}^{N}(2\overline{x}_a x'_{ai} + 2\overline{y}_a y'_{ai}) = \frac{2\overline{x}_a}{N}\sum_{i=1}^{N} x'_{ai} + \frac{2\overline{y}_a}{N}\sum_{i=1}^{N} y'_{ai}$$

Given $\sum_{i=1}^{N} x'_{ai} = \sum_{i=1}^{N} y'_{ai} = 0$, we have $\frac{1}{N}\sum_{i=1}^{N}(2\overline{x}_a x'_{ai} + 2\overline{y}_a y'_{ai}) = 0$

We provide more details in the derivation of equation (A1) in the revised manuscript to make it easy to understand.

• *p. 14, equation (A4): similar to eqn. (A1), how did you get from line 2 to 3?*

For the same reason, the sum of anomaly equals to 0. Thus the terms $\sum_{i=1}^{N}(\overline{x}_{ai}x'_{ai})$, $\sum_{i=1}^{N}(\overline{x}_{ai}x'_{bi})$, are removed from the equation (A4).

• *p. 19: the added value of figure 2 seems rather limited, this figure could be deleted*

Figure 2 has been deleted in the revised manuscript.

• *p. 20, "[...] and each randomly produced vector field" →"[...] and randomly generated vector fields"*

Done

• *p. 20, delete ”are” before ”included in the statistics”*

Done

• *p. 21, ”vector similar coefficients” ! ”vector similarity coefficients”*

Done

• *p. 21, ”between January and 12 months (Solid line)”: this formulation is not clear, please rephrase*

The sentence has been reworded as “(f) The vector similarity coefficients of 850-hPa vector winds between climatological mean January and 12 climatological months (Solid line)” in the revised manuscript.