Anonymous Referee #2

Comment

The authors have evaluated precipitation simulation of 20 CMIP5 GCMs for Pakistan, and developed multi-model ensemble mean at annual and seasonal timescales. The topic is relevant to the journal and the findings are interesting. The authors have shown the application of random forests for MME, which is somehow novel. However, the study needs substantial improvement in explaining the methods. The details of some of the methods are missing and should be further explained. Please find more detailed comments in the following:

P refers to page number and L is the line number (please consider using continuous line numbering in future publications)

Reply

Thank you very much for your highly constructive comments on our manuscript. We have addressed all your major concerns and minor comments carefully in the revised manuscript. We hope that you will find the revised paper suitable for publication.

Comment 1

P1, L13: “number metrics” » “number of metrics”

Reply

Thanks. Corrected as suggested.

Comment 2

P1 L14: “very little attention has been given to spatial performance of GCMs” » Better to rephrase this sentence, since several studies have considered both spatial and temporal characteristics for evaluating GCMs.

Reply

Thanks for your suggestion, we have rephrased it accordingly.

Comment 3

P2, L4: “land and ocean temperature” » “land and ocean surface air temperature”

Reply

Thanks. Corrected as suggested.

Comment 4

P2, L26: “better GCMs are assigned higher weightages” » “higher weights are assigned to better GCMs”

Reply

Thanks. Corrected as suggested.

Comment 5

P2, L31: “is climate change modelling” » “in climate change modelling”

Reply

Thanks. Corrected as suggested.

Comment 6

P3, L7: “such as such as” » please remove the redundant “such as”.

Reply

Thanks. Corrected as suggested.

Comment 7

P5, L25: “five : : : measures” » six measures are introduced here. Please revise the number.

Reply

Thanks. Corrected as suggested.

Comment 8

P6, L4-5: did you apply these measures on each grid? Or are they applied on temporally averaged data? For instance, how is KGE calculated? Please explain.

Reply

Thanks for your comment. All spatial metrics were applied to each grid point separately and then by spatially averaging them a single value for each spatial metric was obtained. For example, first, KGE was applied to 35 grid points (shown in Figure 1) individually, and then these KGE values obtained for each of these grid points (altogether 35 values) were averaged to obtain a single value representative of the whole study area.

We have addresses the issue in section 3.2 of revised manuscript as below:

“3.2 GCM Performance Assessment

SPAtial EFficiency, Fractions Skill Score, Goodman–Kruskal's lambda, Cramer’s V, Mapcurves, and Kling-Gupta efficiency were individually applied to mean annual, monsoon, winter, pre-monsoon, and post-monsoon precipitation, maximum and minimum temperature for each year from 1961 to 2005 of. Later, the GOF values corresponding to each year were temporally averaged to obtain a value for the entire study area.”

Comment 9

P5, L8-9: “comprehensive rating metric” » what does this indicate? How were the ranks of GCMs (from different measures) combined? Do you mean like averaging the ranks? If so, “comprehensive” is misleading and it is better to be revised.

Reply

Thanks for your comment. Comprehensive rating metric is an index used in several studies (Chen et al., 2011; Jiang et al., 2015; Jiang et al., 2012) to combine ranks of a GCM obtained using different metrics or/and considering different climate variables or/and seasons into one single overall rank. In other words, comprehensive rating metric

helps to obtain an overall rank from different ranks obtained using different metrics or/and considering different climate variables or/and seasons. In the application of rating metric ranks obtained using different metrics or/and considering different climate variables or/and seasons not averaged to obtain an overall rank for a given GCM. In this study, application of rating metric involved summing the ranks of a GCM corresponding to different seasons (i.e. annual, monsoon and winter) were aggregated and normalized by the refers to the number of GCMs ($m = 36$) x the number of metrics ($n = 6$) as shown in Eq 15. The details of the rating matric are given in section 3.3.

We agree that word comprehensive rating metric is somewhat misleading, but we have followed the original manuscript where they used the term "Comprehensive Rating Metric".

Comment 10

P6, L28: Eq (1) seems to be the equation for KGE. Please make sure to provide the equation for SPAEF here.

Reply

Sorry for this mistake, this should be SPAEF instead of KGE. We have changed it accordingly.

Comment 11

Section 3.1.1: I am still not sure how the measure is calculated? Is it applied to each grid (2◦x2◦), and then maybe the spatial mean value of KGE is considered? Or, did you take the long-term average of precipitation and then calculate the KGE for a few grids?

Reply

Thanks for your comment. In calculating KGE and other metrics, first all monthly data (i.e. GPCC and GCM) of different spatial resolutions (see Table 1 for resolutions) were used to derive data for annual, monsoon, winter, pre-monsoon and post-monsoon seasons corresponding to a common grid with a spatial resolution of 2◦x2◦. This grid contained 35 points as shown in Figure 1. Then, the means of GPCC and GCM precipitation for the each year for the period 1961 to 2005 were calculated for each grid point for each season and each variable. Later, the GOF values of each year were temporally averaged to obtained a value for the entire study area. We have discusses this in section 3.2 as shown below.

"3.2 GCM Performance Assessment

SPAtial EFficiency, Fractions Skill Score, Goodman–Kruskal's lambda, Cramer's V, Mapcurves, and Kling-Gupta efficiency were individually applied to mean annual, monsoon, winter, pre-monsoon, and post-monsoon

precipitation, maximum and minimum temperature for each year from 1961 to 2005 of. Later, the GOF values corresponding to each year were temporally averaged to obtain a value for the entire study area."

Comment 12

Section 3.1.5 is not clearly explained. A, B, and C need more explanation. What do you mean by "total area of historical and GCM simulated maps"? Is this the coverage area? If so, then both GCM and obs have the same number of grids and the areas should be identical in all cases. In addition, what does "the degree of intersection (for C)" refer to? How can one quantify such thing? Is there a function for calculating it? These need to be clearly explained.

Reply

Thanks for your comment. Total area refers to the historical and GCM simulated maps area. Yes, observed and GCM maps have same number of grid points (i.e. 42) in all cases, C is the intersecting area, and mapcurve function is used in R to calculate GOF values. In order to avoid the confusion, we have revised the explanation as follows.

3.2.5 Mapcurves

Mapcurves is another statistical measure, developed by Hargrove et al. (2006) for the measurement of similarity between categorical maps. Mapcurves quantifies the degree of concordance between two maps. The value of Mapcurves can vary from 0 to 1 (perfect agreement). In the present study, the degree of concordance between the historical observed *P*, T_{max} and T_{min} map and each of the GCM simulated *P*, T_{max} and T_{min} maps was determined using Eq. (11) where, MC_X refers the Mapcurves value, A is the total area of a given category *X* on the map being compared (i.e. map of a GCM simulated variable), B is the total area of a given category Y on the map of observations, C is the overlapping area between *X* and *Y* when the maps are overlaid and *n* is the number of classes in the reference map (e.g. map of observations).

$$MC_X = \sum_{Y=1}^{n} \left[\left(\frac{C}{A} \cdot \frac{C}{B} \right) \right] \quad (11)$$

In this study the function "mapcurves(x,y)" available in "sabre" package (Nowosad and Stepinski, 2018) written in R programing language was used for estimating mapcurves values. In the above function x, and y are numerical vectors, which represent categorical values of observed precipitation (i.e. GPCC precipitation) or map of observed precipitation and categorical values of GCM simulated precipitation, respectively.

Comment 13

P11, l2: “can be reduce” » “can be reduced”

Reply

Thanks. Corrected as suggested.

Comment 14

Section 3.4: There is no explanation about the details of the random forest method. How many trees were included? What are the inputs to the model (time series of 4 selected GCMs?)? Is the model applied separately for each grid, or did you employ a consistent model for the entire study domain? How long is the training and testing periods? How did you evaluate the performance of the random forest outputs?

Reply

Thanks for the comment; we have added new section 3.5.2 containing information related to Random Forest as shown below.

“**3.5.2 Random Forest (RF)**

Random Forest (RF) algorithm (Breiman, 2001) was used in the calculation of the mean time series of P, T_{max} and T_{min} corresponding to an MME of four top ranked GCMs. RF is a relatively new machine learning algorithm widely used in modelling non-linear relationships between predictors and predictands (Ahmed et al., 2019b). RF algorithm is found to perform well with spatial data sets and less prone to over-fitting (Folberth et al., 2019). Most importantly Folberth et al. (2019) reported that RF is less sensitive to multivariate correlation.

RF is an ensemble technique where regression is done using multiple decision trees. RF algorithm uses the following steps in regression.

1. A bootstrap resampling method is used to select sample sets from training data.
2. Classification And Regression Tree (CART) technique is used to develop unpruned trees using the bootstrap sample.
3. A large number of trees are developed with the samples selected repetitively from training data so that all training data have equal probability of selection.
4. A regression model is fitted for all the trees and the performance of each tree is assessed.
5. Ensemble prediction is estimated by averaging the predictions of all trees which is considered as the final prediction.

Wang et al. (2017a) and He et al. (2016) reported that the performance of RF varies with the number of trees (*ntree*) and the number of variables randomly sampled at each split in developing the trees (*mtry*). It was observed that RF performance increases with the increase in *ntree*. However, in the present study the performance was not found to increase significantly in term of root mean square error when the *ntree* was greater than 500. Therefore, *ntree* was set to 500 while the *mtry* was set to $p/3$ where p is the number of variables (i.e. GCMs) used for developing RF-based MME.

The MME prediction can be improved by assigning larger weight to the GCMs which show better performance (Sa'adi et al., 2017). RF regression models developed using historical P, T_{max} and T_{min} simulations of GCMs as independent variable and historical observed P, T_{max} and T_{min} as dependent variable provide weights to the GCMs according to their ability to simulate historical observed P, T_{max} and T_{min}.

The "Random Forest" package written in R programming language was employed in this study for developing RF-based MMEs. RF-based MMEs were calibrated with the first 70% of the data and validated with the rest of the data.

Comment 15

Figure 4 caption: Does the figure show long-term average values for different grids? Or, does it show spatial mean precipitation in various years. Please clarify it in the caption and the text, and mention the period for it as well.

Reply

Thanks for the comment. The Figures 4, 5 and 7 show long-term average values of precipitation for the period 1961 to 2005 over different grids. Each figure has 35 points representing the long-term average values for the whole study area. We have revised the captions as follows.

"Figure 4. Scatter of spatially averaged annual P, T_{max} and T_{min}, of four top ranked GCMs against GPCC P, CRU T_{max} and CRU T_{min} for the period 1961 to 2005.

Figure 5. Scatter of spatially averaged annual P, T_{max} and T_{min} of four lowest ranked GCMs against GPCC P, CRU T_{max} and CRU T_{min} for the period 1961 to 2005.

Figure 7. Scatter of spatially averaged mean annual GPCC P, CRU T_{max} and CRU T_{min} MME of four top ranked GCMs against P, CRU T_{max} and CRU T_{min} using Simple Mean (SM) and Random Forest (RF) for the period 1961 to 2005."