

## Anonymous Referee #3

### General Comments

In this manuscript, authors evaluated precipitation data from 20 CMIP5 GCMs and selected four better-performing CMIP5 GCMs based on their spatial performance against observed precipitation (GPCC) during the historical period (1961-2005). To evaluate the skill of model precipitation (CMIP5 GCMs) against observed precipitation (GPCC), they used six spatial metrics (SPAEF, Goodman-Kruskal's lambda, Fractions Skill Score, Cramer's V, Mapcurves, and Kling-Gupta efficiency). Finally, they generated multi-model ensemble mean (MME) of precipitation of four selected GCMs using Random forest regression and simple mean method. The manuscript is written fairly well, and the idea of spatial assessment of CMIP5 GCMs for multi-model ensemble mean is appreciated. However, the execution of manuscript seems sloppy and hasty. There are numerous methodological, data, explanation, reporting, and citation issues in the manuscript. Thus I recommend major revisions be required before publication.

### Reply

Thank you for your highly constructive comments and suggestions on our manuscript. Your constructive comments and suggestions helped us to improve the quality of the paper. We have carefully addressed all your comments in the revision of the paper. Revised text is highlighted in red.

Major issues:

#### Comment 1

Error and unexplained parameters in the formula of matrices: I have many doubts about spatial assessment methods. Authors need to explain all six methods clearly and correctly.

**a)** In Goodman-Kruskal's lambda, how many classes you have taken in the contingency matrix? Please mention the number of classes and explain- Are these classes sufficient to explain spatial variability of rainfall or measure the matrix accurately? Did you consider only one annual map to estimate the lambda value for each model? If yes, then there may be many years those have low or high bias but not captured in the annual mean map. You need to estimate lambda value for each year or seasonal map. What is the  $\max_j$  (or  $\max_j$ )? What is the value of m and n?

#### Reply a:

Thanks for your comment. We have considered seven classes (categories) in the contingency matrix following the study by Demirel et al. (2018). We have addressed the above issues as follows.

“Goodman–Kruskal's lambda also known as Lambda coefficient ( $\lambda$ ) is used to measure the nominal/categorical association between categorical maps (Goodman and Kruskal, 1954). Lambda coefficient ( $\lambda$ ) varies between 0 and 1, where a value closer to 1 refers to a higher similarity between the map of model simulations and that of observations of  $P$ ,  $T_{max}$  and  $T_{min}$ . The Lambda ( $\lambda$ ) coefficient was calculated using Eq. (9), where  $max_j$  is the number of classes (categories) in observed and simulated maps,  $c_{ij}$  is a contingency matrix (describes the relationships between the data classes),  $i$  and  $j$  are the classes in observed and simulated maps,  $m$  represent the number of classes in observed and simulated maps respectively. In the present study, seven classes in the contingency matrix were used by following the study by Demirel et al. (2018). The “DescTools” package (Signorell, 2016) written in R programming language was employed in this study for estimating the nominal/categorical association between observed and simulated maps.

$$\lambda = \frac{\sum_{i=1}^m max_j c_{ij} - max_j \sum_{i=1}^m c_{ij}}{N - max_j \sum_{i=1}^m c_{ij}} \quad (9)$$

Regarding the calculation of Lambda value, we have calculated the Lambda value for year and seasons separately and then an average value was considered for the whole study area.

We have addressed the above issue in section 3.1 of the revised manuscript as follows.

### 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 for each year from 1961 to 2005 to mean annual, monsoon, winter, pre-monsoon, and post-monsoon precipitation, maximum and minimum temperature. Later, the GOF values of each year were temporally averaged to obtain a value for the entire study area. The details of the above spatial metrics are given below.”

**b)** In the fraction skill score, there should  $N_x \times N_y$  in the palace of N. Roberts and Lean,(2008) used  $N_x \times N_y$ . It will affect the final results. Please explain it.

#### Reply b:

Thanks for the comment, we used “verification” package (Pocernich, 2006) written in R programming language for the calculation of FSS. Verification package follows the equations used in Roberts and Lean (2008). Therefore, we have revised FSS equations as below.

### 3.2.2 Fractions Skill Score

The Fractions Skill Score (*FSS*) proposed by (Roberts and Lean, 2008) is another measure used for the assessment of spatial agreement between model simulations and observations. *FSS* varies between 0 and 1 where a value closer to 1 refers to higher agreement between observed and simulated data. In this study, *FSS* between observed and GCM simulated data was computed using Eq. (6).

$$FSS = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}} \quad (6)$$

In Eq. (6) MSE refers mean square error and is calculated using Eq. (7) and (8).

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{(n)i,j} - M_{(n)i,j}]^2 \quad (7)$$

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \left[ \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O_{(n)i,j}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} M_{(n)i,j}^2 \right] \quad (8)$$

In Eq. (7) and (8)  $N_x$  is the number of columns,  $N_y$  is the number of rows in a map (observed or simulated),  $O$  and  $M$  are observed and simulated data fractions respectively. The “verification” package (Pocernich, 2006) written in R programming language was employed in this study for estimating *FSS* values.

c) In Cramer’s  $V$ , you have taken the wrong formula. There should be  $N*(\min(m-1,n-1))$ , but you have taken  $N*(\min(m,n)-1)$ . It will also affect your final selection.

#### Reply c:

We agree that Cramér’s  $V$  is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1. Both these expression do the same action. Let’s assume  $m=34$  and  $n=32$ , thus  $\min(m,n)-1=31$  and also  $\min(m-1,n-1)=31$ . The equation of Cramer’s  $V$  used in the present study is also same as the one used in the study by Rees (2008). The text was revised as follows.

“Cramer’s  $V$  (Cramér, 1999) statistic is a Chi-square-test-based measure which is used in assessing spatial agreement between observations and model simulations (Zawadzka et al., 2015). Its value ranges between 0 and 1 and a closer the value to 1 the better the agreement. Cramer’s  $V$  is calculated using Eq. (10).

$$V = \sqrt{\frac{\chi^2}{N(\min(m,n)-1)}} \quad (10)$$

where,  $\chi^2$  is Chi-Square,  $N$  is the grand total of observations,  $m$  is the number of rows and  $n$  is the number of columns. In this exercise  $m = 42$  (number of rows of data) and  $n = 2$  (observed and modelled precipitation). The “DescTools” package (Signorell, 2016) written in R programming language was employed in this study for calculating Cramer’s  $V$  values.”

**d)** In Mapcurves method, did you classify your map in the different range of rain? If yes, how many classes you have taken? Did you calculate  $Y$  value for each month/ season/year? It should be calculated for each year (1961-2005) between model and GPCC data in the case of annual values.

**Reply d:**

Yes, we have classified our data into seven classes and calculated the map curve value for each year and each season and later an average value was considered for the whole study area. We have addressed the issue as follows.

“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 class  $X$  on the map being compared,  $B$  is the total area of a given class  $Y$  on the observed map,  $C$  is the interesting area between  $X$  and  $Y$  when the maps are overlaid and  $n$  is the number of classes in the reference map.

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

In this study the function “*mapcurves(x,y)*” available in “*sabre*” package (Nowosad and Stepinski, 2018) written in R programming language was used for estimating mapcurves values. In that function  $x$ , and  $y$  are vectors representing categorical values of historical observed data (e.g. GPCC precipitation) and categorical values of simulated data by a GCM, respectively.”

e) In Kling-Gupta efficiency, please check Demirel et al., 2018 paper. They have taken different formulas for beta and gamma.

**Reply e:**

Thanks, we rechecked the equations with the original paper related to Kling-Gupta Efficiency and found that the equations in our manuscript are correct.

f) Why did you choose these six methods? What are the limitations of each method? Please explain.

**Reply f:**

The study by Rees (2008) inspired us to test different spatial metrics in our GCM selection study. Furthermore, these metrics have been also used in other studies (Demirel et al., 2018; Koch et al., 2018; Rees, 2008). We have added a line on page 2, line 15 of introduction section as follows.

“These metrics were selected based on their recent applications in spatial performance assessment of models (Demirel et al., 2018; Koch et al., 2018; Rees, 2008).”

The limitations of these metrics are reported in Demirel et al. (2018) as follows.

“SPAEF is noted as very discriminative metric in selecting different raster maps whereas other metrics e.g. FSS, Cramer’s V (Demirel et al., 2018; Koch et al., 2018) are tolerant (less sensitive). This leads to different results in the spatial calibration of models.

**Comment 2**

Error in rating metrics formula: (P10, L10) In this formula, rank varies from 1 to 6 (n=6) but it should be 1 to 20 (model=20) for each matrix. Please explain this.

**Reply**

Sorry for the mistake; we have made the necessary correction as follows.

“The overall ranks of GCMs based on different GOFs were obtained for each season separately using Eq. (15).

$$RM = 1 - \frac{1}{nm} \sum_{i=1}^n rank_i \quad (15)$$

In Eq. (15),  $n$  refers to the number of GCMs,  $m$  refers to the number of metrics or seasons and  $i$  refers to the rank of a GCM based on  $i^{\text{th}}$  GOF. A value of RM near to 1 refers to a better GCM in terms of its ability to mimic the spatial or temporal characteristics of observations.”

### Comment 3

Pre-monsoon and Post-monsoon seasons: Why did you not consider the pre and post monsoon season for the analysis and during the overall rank. These seasons will affect significantly in the overall ranking. I recommend to estimate rank month-wise That will improve the results significantly and should not provide the same weight to each month. Here, you provided the same weight to annual, monsoon, and winter rank (during overall rank). Why?

### Reply

Thanks for your suggestion, we have revised whole analysis by considering pre-monsoon and post monsoon along with annual, monsoon and winter seasons. Besides precipitation, we also included maximum and minimum temperature for the selection of GCMs. When different seasons and climate variable were considered it significantly changed the ranks. Large uncertainties are associated in GCM outputs at monthly or finer timescales. Therefore, selections of GCMs are generally not done based on month-wise ranking. GCMs are generally ranked based on their capability of producing present-day annual and seasonal climatology. This has been mentioned in the revised manuscript as follows.

“GCMs are faltered by the uncertainty in their outputs at monthly or finer timescales such as daily or sub-daily (Xue et al., 2007; Onyutha et al., 2016) (Xue et al., 2007; Onyutha et al., 2016). Therefore, the performances of GCMs are generally evaluated according to their capability of producing present-day mean seasonal cycles, interannual variability, and spatial distribution of climatology at regional or local scales (Meher et al., 2017; Das et al., 2018) (Miao et al. 2012; Fu et al. 2013; Das et al. 2016; Meher and Das, 2017).”

### Comment 4

Inconsistency in spatial resolution: You should consider the same spatial resolution to compare the maps or data sample. In the manuscript, observation data (GPCC) are available at  $0.5^\circ$  resolution and model data are prepared at  $2^\circ$  resolution. Model data should be regridded at  $2^\circ$ .

## Reply

In order to avoid the confusion, we have added following text to section 2.2.1 of the revised manuscript as follows.

“Monthly precipitation data simulated by the 36 CMIP5 GCMs for ensemble member r1i1p1 run were extracted from the IPCC data distribution center ([http://www.ipcc-data.org/sim/gcm\\_monthly/AR5/Reference-Archive.html](http://www.ipcc-data.org/sim/gcm_monthly/AR5/Reference-Archive.html)) for period 1961-2005. The modelling centres, names of GCMs and spatial resolution of each of the selected GCMs are provided in Table 1. In order to have a common spatial resolution, precipitation, maximum and minimum temperature data obtained from different GCMs and GPCC and CRU databases were interpolated into a common  $2^{\circ}\times 2^{\circ}$  grid using bilinear interpolation.”

## Comment 5

Random Forest Method: Please explain the method and weight value.

## Reply

Thanks for the comment, we have added a new section 3.5.2 for Random Forest description 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 (*n<sub>tree</sub>*) and the number of variables randomly sampled at each split in developing the trees (*m<sub>try</sub>*). It was observed that RF performance increases with the increase in *n<sub>tree</sub>*. However, in the present study the performance was not found to increase significantly in term of root mean square error when the *n<sub>tree</sub>* was greater than 500. Therefore, *n<sub>tree</sub>* was set to 500 while the *m<sub>try</sub>* 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 6

Increase the number of CMIP5 models in the study: Authors used only 20 models for the current study and said all four RCP data available for 20 models. However, there is no use of RCP data in the analysis. Hence, they can get historical data for more than 35 CMIP5 GCMs. That will increase the scope and use of this study. I recommend they should use the maximum number of models.

#### Reply

Thanks for your suggestion. We have revised whole analysis by considering precipitation, maximum and minimum temperature data obtained from 36 CMIP5 GCMs.

#### Comment 7

Selection of better performing models should be based on at least precipitation and temperature: In the manuscript, authors used only precipitation variable to select better performing models, but there are many models under CMIP5 those have low projection skill in temperature data and high skill in precipitation. Hence, there is a possibility of the poor skill of temperature projection in the selected GCMs. Moreover, most of the studies in the hydrology and earth science commonly use precipitation and temperature variables. Therefore, they should include the temperature



variable in the analysis and select the models based on the high skill in both (Precipitation and temperature) variables.

## Reply

Following your suggestion, we have selected the GCMs based on annual, monsoon, winter, post and pre-monsoon precipitation, maximum and minimum temperature over Pakistan. The revised results are given below.

### 4.2 Evaluation and Ranking of GCMs

The SPAEF, FSS, Lambda, Cramer-V, Mapcurves, and KGE between **observed** (GPCC  $P$ , CRU  $T_{max}$  and  $T_{min}$ ) and GCMs simulated mean annual, monsoon, winter, **pre-monsoon and post-monsoon**  $P$ ,  $T_{max}$  and  $T_{min}$  of Pakistan were estimated for the period 1961 to 2005. As an example, Table 3 shows the GOF values that define the performance of each GCM in simulating GPCC mean annual precipitation. In Table 3 ranks of GCMs corresponding to each performance metric is shown within brackets. The GOF values near to 1 refer to the better performance of the GCM of interest. For example, **CESM1-CAM5** has a GOF value of **0.540** for SPAEF, and hence regarded as the best GCM in term of SPAEF, whereas CSIRO-Mk3-6-0 can be regarded as the poorest which has a GOF value of **-0.505** in term of SPAEF. The GOF values for other metrics (i.e. FSS, Lambda, Cramer-V, Mapcurves, and KGE) can be interpreted in the same manner.

**Table 3.** GOF values of GCMs obtained using different spatial metrics for mean annual precipitation.

<b>GCM</b>	<b>SPAEF (Rank)</b>	<b>FSS (Rank)</b>	<b>Lambda (Rank)</b>	<b>Cramer- V(Rank)</b>	<b>Mapcurves (Rank)</b>	<b>KGE (Rank)</b>
ACCESS1-0	0.411(7)	0.659(24)	0.143(24)	0.370(28)	0.244(29)	0.172(29)
ACCESS1-3	0.155(24)	0.712(20)	0.107(30)	0.315(34)	0.206(34)	0.310(15)
BCC-CSM1-1	0.241(21)	0.691(21)	0.143(24)	0.388(27)	0.258(27)	0.082(33)
BCC-CSM1.1(m)	0.149(25)	0.685(22)	0.214(13)	0.545(16)	0.376(16)	0.304(16)
BNU-ESM	0.185(23)	0.759(11)	0.179(18)	0.519(21)	0.349(21)	0.233(26)
CanESM2	0.250(20)	0.642(26)	0.250(6)	0.547(15)	0.378(15)	-0.443(35)
CCSM4	0.440(4)	0.798(5)	0.250(6)	0.667(4)	0.525(4)	0.420(8)
CESM1-BGC	0.439(5)	0.759(12)	0.214(13)	0.655(10)	0.508(10)	0.337(12)
CESM1-CAM5	0.540(1)	0.840(1)	0.250(6)	0.667(4)	0.525(4)	0.531(2)
CESM1-WACCM	0.430(6)	0.776(10)	0.250(6)	0.656(9)	0.510(9)	0.384(10)
CMCC-CM	-0.255(34)	0.565(33)	0.143(24)	0.496(24)	0.325(24)	0.189(28)
CMCC-CMS	-0.043(28)	0.637(28)	0.143(24)	0.369(29)	0.244(28)	0.249(22)
CNRM-CM5	0.364(12)	0.732(17)	0.250(6)	0.667(4)	0.525(4)	0.314(14)
CSIRO-Mk3-6-0	-0.505(36)	0.321(36)	0.036(36)	0.264(36)	0.179(36)	-1.837(36)
EC-EARTH	0.232(22)	0.756(13)	0.286(4)	0.759(2)	0.642(2)	0.404(9)
FGOALS-g2	0.321(13)	0.793(6)	0.179(18)	0.531(17)	0.361(17)	0.362(11)
FIO-ESM	0.281(17)	0.752(14)	0.214(13)	0.559(14)	0.391(14)	0.283(19)
GFDL-CM3	0.387(8)	0.815(4)	0.429(1)	0.782(1)	0.690(1)	0.493(3)
GFDL-ESM2G	0.307(14)	0.786(7)	0.250(6)	0.667(4)	0.525(4)	0.484(4)
GFDL-ESM2M	0.297(16)	0.778(8)	0.214(13)	0.436(26)	0.296(25)	0.458(5)
GISS-E2-H	-0.100(32)	0.616(31)	0.107(30)	0.335(33)	0.220(33)	0.245(24)
GISS-E2-R	-0.054(29)	0.616(30)	0.107(30)	0.350(31)	0.229(31)	0.236(25)
HadGEM2-AO	0.454(3)	0.740(15)	0.179(18)	0.520(20)	0.350(20)	0.315(13)
HadGEM2-CC	0.387(9)	0.683(23)	0.179(18)	0.360(30)	0.236(30)	0.222(27)
HadGEM2-ES	0.371(11)	0.721(18)	0.179(18)	0.530(18)	0.360(18)	0.277(20)
INMCM4	0.378(10)	0.777(9)	0.179(18)	0.530(18)	0.360(18)	0.422(6)
IPSL-CM5A-LR	-0.054(30)	0.634(29)	0.357(2)	0.590(12)	0.427(12)	0.117(32)
IPSL-CM5A-MR	-0.093(31)	0.548(34)	0.357(2)	0.590(12)	0.427(12)	-0.183(34)
IPSL-CM5B-LR	-0.286(35)	0.538(35)	0.107(30)	0.350(31)	0.229(31)	0.131(31)
MIROC-ESM- CHEM	0.273(18)	0.733(16)	0.214(13)	0.655(10)	0.508(10)	0.303(17)
MIROC-ESM	0.258(19)	0.720(19)	0.286(4)	0.677(3)	0.537(3)	0.29(18)
MIROC5	0.302(15)	0.828(3)	0.071(34)	0.454(25)	0.285(26)	0.420(7)
MPI-ESM-LR	-0.012(27)	0.639(27)	0.143(24)	0.517(22)	0.346(22)	0.253(21)
MPI-ESM-MR	0.041(26)	0.653(25)	0.143(24)	0.506(23)	0.335(23)	0.245(23)
MRI-CGCM3	-0.180(33)	0.572(32)	0.071(34)	0.293(35)	0.194(35)	0.169(30)
NorESM1-M	0.464(2)	0.833(2)	0.250(6)	0.667(4)	0.525(4)	0.532(1)

Table 3 shows the ranks attained by GCMs corresponding to different metrics. For example, BCC-CSM1.1 (m) attained ranks 25, 22, 13, 16, 16 and 16 for SPAEF, FSS, Lambda, Cramer-V, Mapcurves, and KGE respectively. It

was observed that **CSIRO-Mk3-6-0** is the only GCM able to secure the same rank for all metrics. However, HadGEM2-ES secured rank 18 for four metrics (i.e. FSS, Lambda, Cramer-V, Mapcurves). **Several GCMs** attained the same rank for three metrics (e.g. **BCC-CSM1.1(m)**, **CCSM4**, **CMCC-CM** and **CMCC-CMS**). Cramer-V and Mapcurve showed more or less similar ranks for GCMs. Similar results were also seen **for other seasons and variables** (not presented in the manuscript).

#### **4.3 Overall Ranks of GCMs for Precipitation, Maximum Temperature and Minimum Temperature**

The application of various evaluation metrics has yielded different ranks for the same GCM (Ahmadalipour et al., 2017;Raju et al., 2017). The ranks attained by GCMs corresponding to different metrics and seasons (**annual, monsoon, winter, pre-monsoon and post-monsoon**) were used to calculate the RM values for each GCM. The ranks of GCMs for  $P$ ,  $T_{max}$  and  $T_{min}$  are presented in Table 4 along with the RM values. As seen in Table 4, **EC-EARTH**, **BCC-CSM1.1 (m)** and **CSIRO-Mk3-6-0** were the most skillful GCMs in reproducing the spatial characteristics of  $P$ ,  $T_{max}$  and  $T_{min}$  respectively. On the other hand, **IPSL-CM5B-LR**, **CMCC-CM**, and **INMCM4** displayed the least skill in reproducing the spatial characteristics of  $P$ ,  $T_{max}$  and  $T_{min}$  respectively.

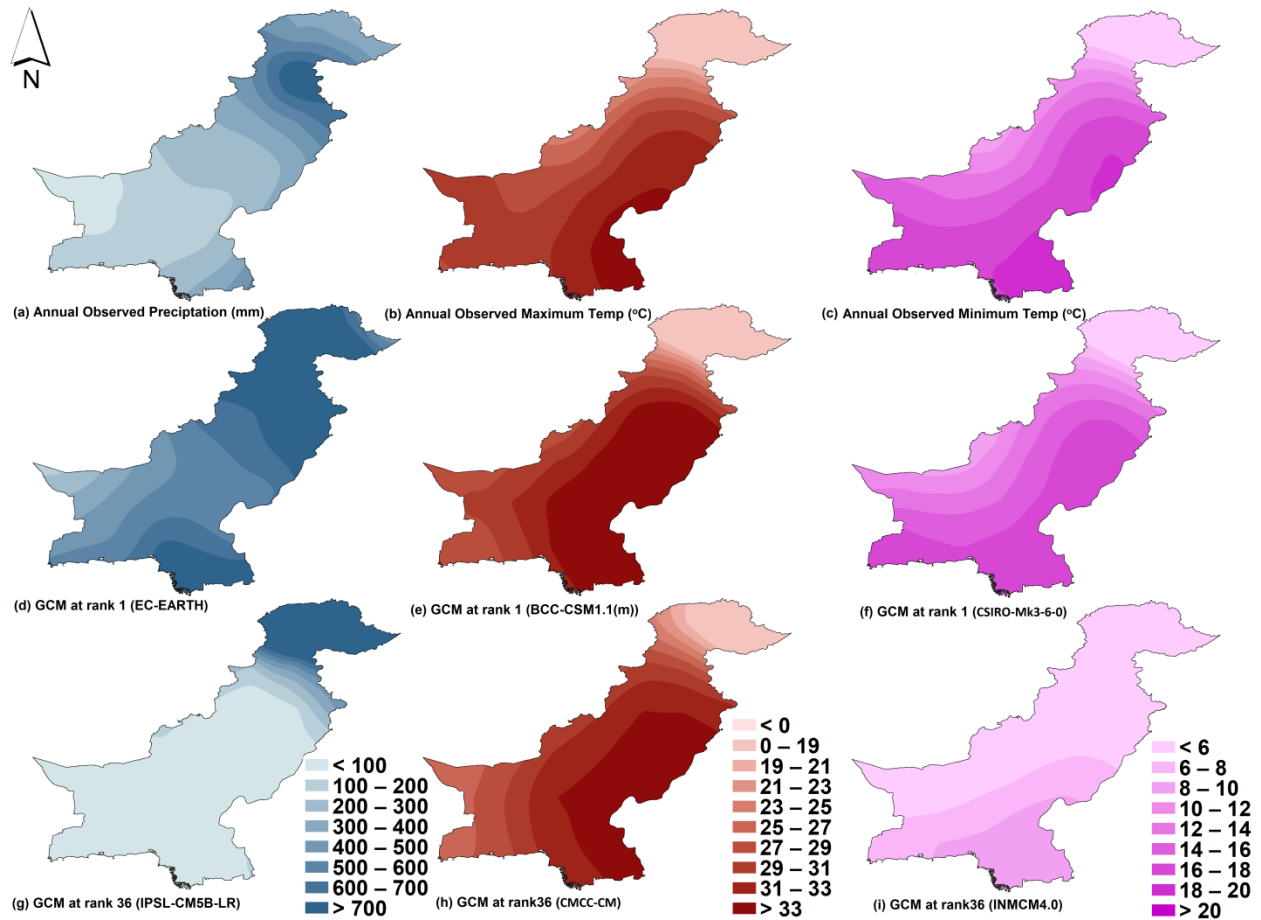
**Table 4.** Ranks of GCMs for  $P$ ,  $T_{max}$  and  $T_{min}$  based on rating metric values

GCM	$P$	Rank	GCM	$T_{max}$	Rank	GCM	$T_{min}$	Rank
EC-EARTH	0.823	1	BCC-CSM1.1(m)	0.702	1	CSIRO-Mk3-6-0	0.750	1
NorESM1-M	0.794	2	NorESM1-M	0.663	2	GFDL-ESM2G	0.720	2
GFDL-CM3	0.714	3	HadGEM2-ES	0.656	3	CMCC-CMS	0.692	3
CCSM4	0.689	4	IPSL-CM5B-LR	0.630	4	BCC-CSM1.1(m)	0.684	4
MIROC5	0.685	5	HadGEM2-AO	0.626	5	GFDL-ESM2M	0.681	5
GFDL-ESM2G	0.673	6	CMCC-CMS	0.616	6	MIROC-ESM-CHEM	0.657	6
CESM1-CAM5	0.654	7	HadGEM2-CC	0.608	7	NorESM1-M	0.656	7
HadGEM2-AO	0.651	8	FGOALS-g2	0.600	8	ACCESS1-3	0.656	8
GFDL-ESM2M	0.643	9	CSIRO-Mk3-6-0	0.594	9	MIROC-ESM	0.654	9
FGOALS-g2	0.607	10	ACCESS1-0	0.577	10	MIROC5	0.646	10
MIROC-ESM	0.589	11	IPSL-CM5A-LR	0.566	11	CCSM4	0.631	11
ACCESS1-0	0.555	12	INMCM4	0.561	12	CESM1-BGC	0.628	12
ACCESS1-3	0.555	12	GISS-E2-H	0.556	13	CESM1-CAM5	0.595	13
MIROC-ESM-CHEM	0.532	14	MIROC5	0.551	14	MRI-CGCM3	0.584	14
HadGEM2-CC	0.531	15	BNU-ESM	0.538	15	CanESM2	0.577	15
HadGEM2-ES	0.514	16	BCC-CSM1-1	0.534	16	BNU-ESM	0.569	16
BCC-CSM1-1	0.506	17	GISS-E2-R	0.532	17	FGOALS-g2	0.569	16
CESM1-WACCM	0.482	18	MPI-ESM-LR	0.532	17	MPI-ESM-MR	0.569	16
CNRM-CM5	0.480	19	FIO-ESM	0.524	19	MPI-ESM-LR	0.566	19
CESM1-BGC	0.467	20	CESM1-WACCM	0.522	20	EC-EARTH	0.506	20
INMCM4	0.464	21	ACCESS1-3	0.520	21	IPSL-CM5A-MR	0.490	21
FIO-ESM	0.462	22	GFDL-ESM2M	0.514	22	HadGEM2-ES	0.487	22
MPI-ESM-MR	0.437	23	MPI-ESM-MR	0.513	23	ACCESS1-0	0.481	23
IPSL-CM5A-LR	0.426	24	CCSM4	0.466	24	FIO-ESM	0.446	24
CanESM2	0.406	25	CESM1-BGC	0.459	25	CMCC-CM	0.428	25
MPI-ESM-LR	0.395	26	CanESM2	0.442	26	GISS-E2-R	0.418	26
BCC-CSM1.1(m)	0.394	27	MIROC-ESM	0.442	26	GISS-E2-H	0.416	27
IPSL-CM5A-MR	0.382	28	CNRM-CM5	0.434	28	HadGEM2-AO	0.416	27
CMCC-CMS	0.381	29	EC-EARTH	0.427	29	IPSL-CM5A-LR	0.416	27
MRI-CGCM3	0.381	29	MIROC-ESM-CHEM	0.427	29	BCC-CSM1-1	0.413	30
CMCC-CM	0.353	31	GFDL-ESM2G	0.416	31	HadGEM2-CC	0.413	30
BNU-ESM	0.337	32	GFDL-CM3	0.398	32	CNRM-CM5	0.361	32
GISS-E2-H	0.319	33	CESM1-CAM5	0.371	33	CESM1-WACCM	0.356	33
CSIRO-Mk3-6-0	0.273	34	IPSL-CM5A-MR	0.326	34	IPSL-CM5B-LR	0.275	34
GISS-E2-R	0.253	35	MRI-CGCM3	0.319	35	GFDL-CM3	0.231	35
IPSL-CM5B-LR	0.144	36	CMCC-CM	0.249	36	INMCM4	0.226	36

The better performance of EC-EARTH, BCC-CSM1.1 (m) and CSIRO-Mk3-6-0 in simulating  $P$ ,  $T_{max}$  and  $T_{min}$  over Indo-Pak sub-continent has also been reported in several past studies. Latif et al. (2018) reported the relatively better performance of EC-EARTH, and BCC-CSM1.1 (m) out of 36 CMIP5 GCMs in simulating precipitation over Indo-

Pakistan sub-continent based on spatial correlation. Rehman et al. (2018) conducted a study to assess the performance of CMIP5 GCMs in simulating the mean precipitation and temperature over south Asia. The study reported the better performance of EC-EARTH in simulating precipitation and CSIRO-Mk3-6-0 in simulating temperature. Khan et al. (2018) assessed the performance of 31 CMIP5 GCMs in simulating the mean precipitation and temperature over Pakistan using multiple daily gridded datasets and identified EC-EARTH as the best GCM for simulating precipitation and CSIRO-Mk3-6-0 for simulating temperature. Better performance of CSIRO-Mk3-6-0 in simulating maximum and minimum temperature is also reported in the study by (Ahmed et al., 2019c).

The spatial patterns of mean annual  $P$ ,  $T_{max}$  and  $T_{min}$  simulated by the GCMs ranked 1 and ranked 36 were compared with the spatial patterns of GPCC  $P$  and CRU  $T_{max}$  and  $T_{min}$ , and presented in Figure 3 as an example. In Figure 3 it was seen that the GCMs that attained rank 1 showed spatial patterns more or less similar to that of GPCC  $P$  and CRU  $T_{max}$  and  $T_{min}$ . On the other hand, GCMs ranked 36 (i.e. rank 36) showed large differences compared to the spatial patterns of GPCC  $P$  and CRU  $T_{max}$  and  $T_{min}$ . The Figure 3 clearly shows that GCMs which attained rank 36 under-estimated the precipitation and temperature over a large region in the study area.



**Figure 3.** Spatial patterns of GPCP precipitation (a), CRU maximum temperature (b) CRU minimum temperature (c) GCM rank 1 (d - f) and GCM rank 36 (g - i) mean annual precipitation, maximum and minimum temperature respectively.

#### 4.4 Identification of Ensemble Members

Based on the criteria mentioned in Section 3.4, ranks of each variable were estimated and then the GCMs were ranked based on the overall RM values. Table 5 shows the overall ranks of the 36 GCMs considered in this study. The four top ranked GCMs; **NorESM1-M**, **MIROC5**, **BCC-CSM1-1** and **ACCESS1-3** indicated in bold in Table 5 were designated as the members of the ensemble for  $P$ ,  $T_{max}$  and  $T_{min}$  over Pakistan.

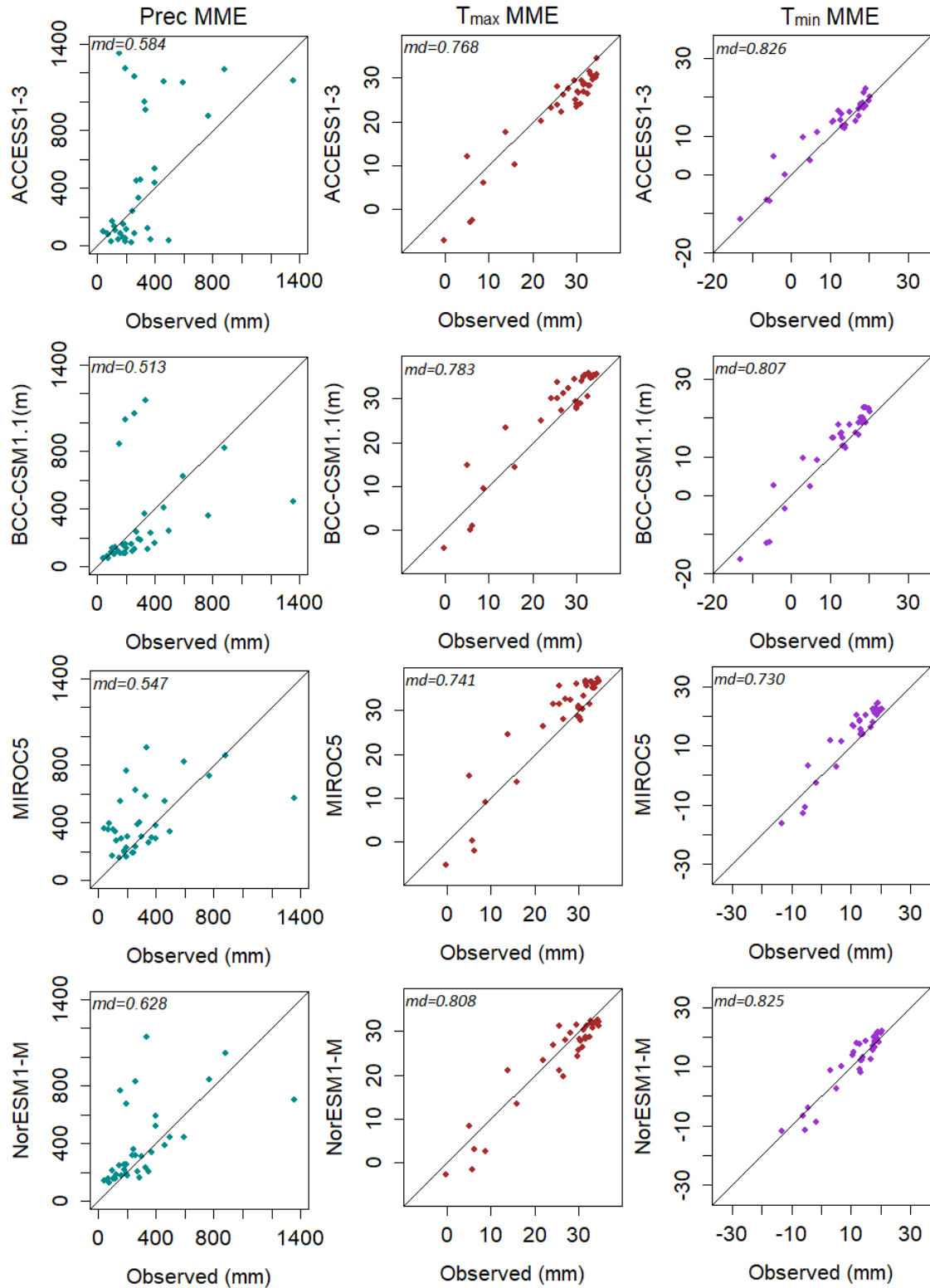
**Table 5. Overall ranks of GCMs for the identification of ensemble members**

<b>GCM</b>	<b><i>P</i> Rank</b>	<b><i>T<sub>max</sub></i> Rank</b>	<b><i>T<sub>min</sub></i> Rank</b>	<b>Overall RM Value</b>	<b>Overall Rank</b>
<b>NorESM1-M</b>	2	2	7	0.898	<b>1</b>
<b>MIROC5</b>	5	14	10	0.731	<b>2</b>
<b>BCC-CSM1-1</b>	17	16	30	0.417	<b>3</b>
<b>ACCESS1-3</b>	10	8	16	0.685	<b>4</b>
GFDL-ESM2M	9	22	5	0.667	5
CMCC-CMS	29	6	3	0.648	6
CCSM4	4	24	11	0.639	7
GFDL-ESM2G	6	31	2	0.639	8
HadGEM2-AO	8	5	27	0.630	9
FGOALS-g2	12	21	8	0.620	10
HadGEM2-ES	16	3	22	0.620	11
CSIRO-Mk3-6-0	34	9	1	0.593	12
ACCESS1-0	12	10	23	0.583	13
MIROC-ESM-CHEM	14	29	6	0.546	14
MIROC-ESM	11	26	9	0.574	15
EC-EARTH	1	29	20	0.537	16
HadGEM2-CC	15	7	30	0.519	17
CESM1-CAM5	7	33	13	0.509	18
CESM1-BGC	20	25	12	0.472	19
IPSL-CM5A-LR	24	11	27	0.426	20
MPI-ESM-LR	26	17	19	0.426	21
MPI-ESM-MR	23	23	16	0.426	22
BCC-CSM1.1(m)	27	1	4	0.704	23
BNU-ESM	32	15	16	0.417	24
FIO-ESM	22	19	24	0.398	25
CanESM2	25	26	15	0.389	26
inmcm4	21	12	36	0.361	27
GFDL-CM3	3	32	35	0.352	28
CESM1-WACCM	18	20	33	0.343	29
GISS-E2-H	33	13	27	0.324	30
IPSL-CM5B-LR	36	4	34	0.315	31
GISS-E2-R	35	17	26	0.278	32
MRI-CGCM3	29	35	14	0.278	33
CNRM-CM5	19	28	32	0.269	34
IPSL-CM5A-MR	28	34	21	0.231	35
CMCC-CM	31	36	25	0.148	36

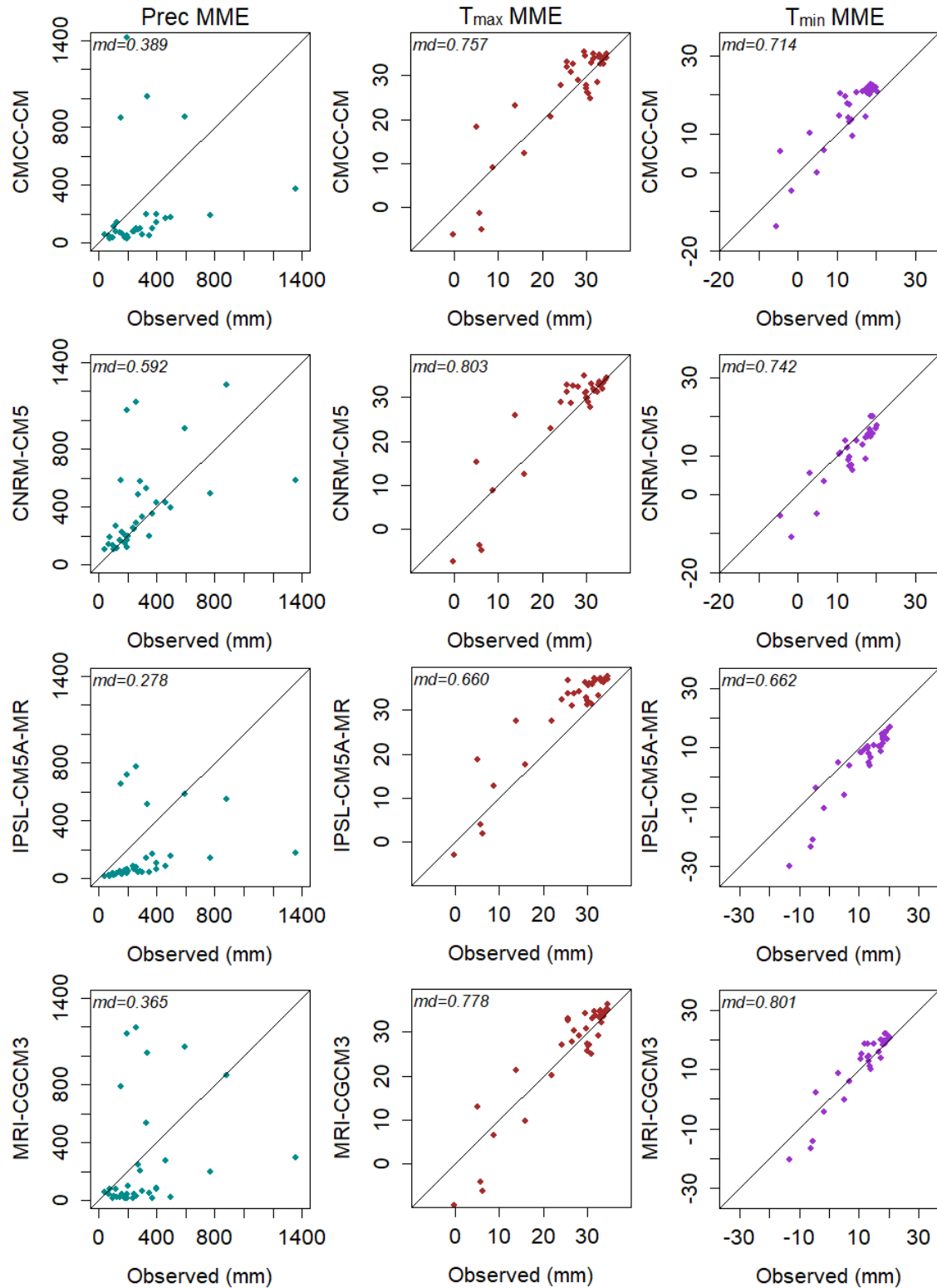
The performances of the four top ranked GCMs (i.e. GCMs ranked 1, 2, 3 and 4) and four lowest ranked GCMs (i.e. GCMs ranked 33, 34, 35, and 36) were visually evaluated using scatter plots shown in Figures 4 and 5, pertaining to mean annual *P*, *T<sub>max</sub>* and *T<sub>min</sub>* as example. In order to plot the scatter, the *P*, *T<sub>max</sub>* and *T<sub>min</sub>* simulated by each GCM

and GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$  pertaining to all grid points was averaged (spatially averaged precipitation and temperature). As expected, GCMs that attained ranks 1 to 4 showed closer agreements with the GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$  compared to that of GCMs which attained ranks 33, 34, 35, and 36. The same can also be noticed based on  $md$  values provided in each figure where top ranked GCMs showed higher  $md$  values compared to lowest ranked GCMs. The scatter plots in Figure 5 indicated that the  $P$ ,  $T_{max}$  and  $T_{min}$  simulated by the least skillful GCMs underestimated mean annual  $P$ ,  $T_{max}$  and  $T_{min}$ . Over and underestimation of  $P$ ,  $T_{max}$  and  $T_{min}$  also can be seen in the scatter plots of GCMs ranked 1, 2, 3 and 4. However, their scatter was found much aligned with the 45 degree line compared to that of GCMs ranked 33, 34, 35, and 36. Therefore, it is argued that the GCMs ranked 1, 2, 3 and 4 can be used as an ensemble for the simulation of 33, 34, 35, and 36.





**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.

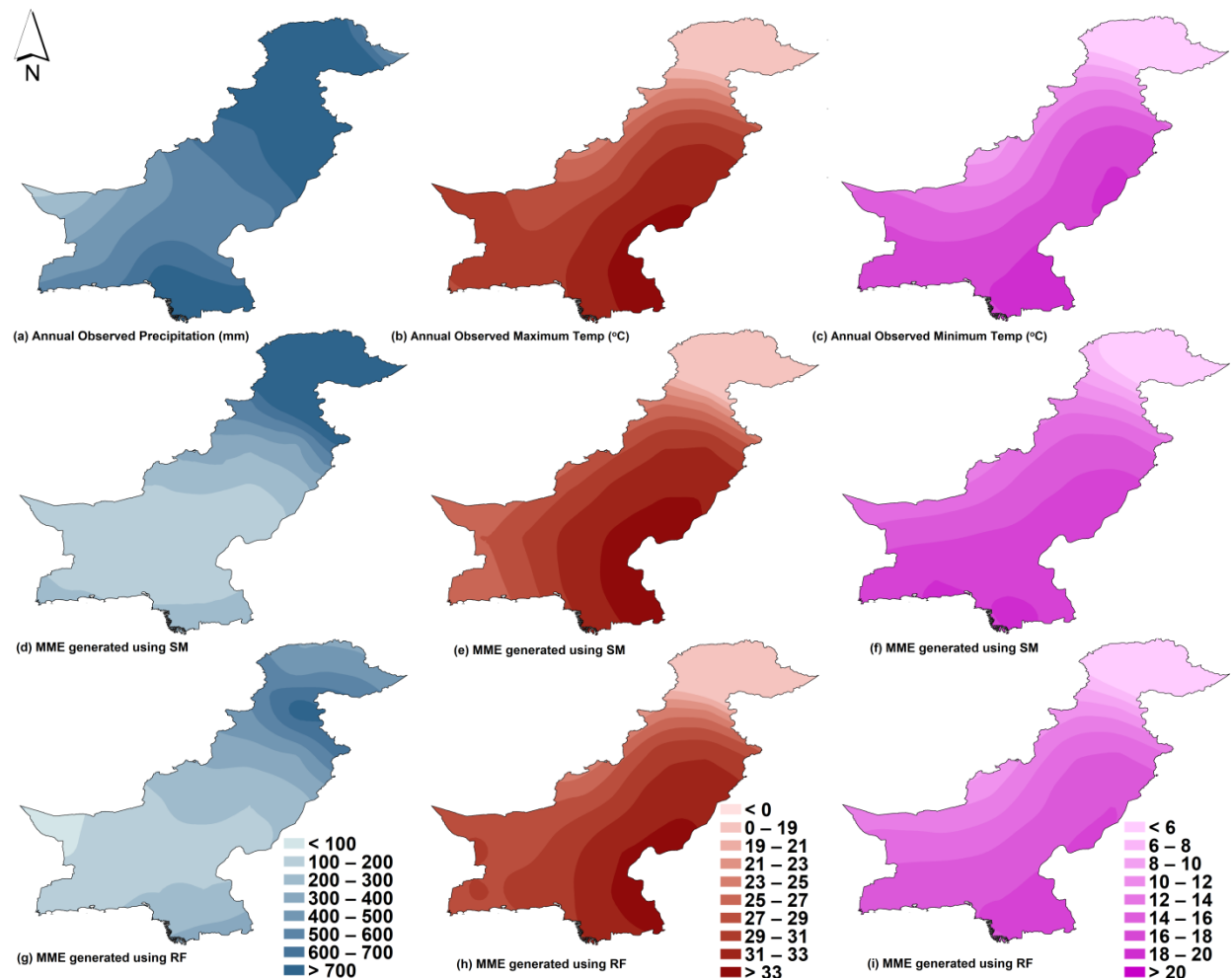
Some of the GCMs identified for the ensemble over Pakistan were found similar with GCMs that showed better performance in the neighboring countries such as India and Iran. Jena et al. (2015) used Z-value test, correlation coefficient, relative precipitation comparison test, probability function comparison, root mean square error, and Student's t-test to evaluate the performance of 20 CMIP5 GCMs in simulating Indian summer monsoon. They found that CCSM4, CESM1-CAM5, GFDL-CM3, and GFDL-ESM2G perform better compared to the other GCMs. Prasanna (2015) conducted a study to assess the performance of 12 CMIP5 GCMs using mean and coefficient of variation over South Asia (5N–35N; 65E–95E) and identified ACCESS, CNRM, HadGEM2-ES, MIROC5, CanESM, GFDL-ESM2M, GISS, MPI-ESM and NOR-ESM as better performing GCMs. Sarthi et al. (2016) evaluated the performance of 34 CMIP5 GCMs using Taylor diagram, skill score, correlation and RMSE. They found that BCC-CSM1.1(m), CCSM4, CESM1(BGC), CESM1(CAM5), CESM1(WACCM), and MPI-ESM-MR were able to better capture the Indian summer monsoon precipitation. Afshar et al. (2016) applied Nash–Sutcliffe efficiency, percent of bias, coefficient of determination, and ratio of the RMSE to the standard deviation of observations for assessing performance of precipitation simulations of 14 CMIP5 GCMs over a mountainous catchment in north-eastern Iran which borders Pakistan. They recommend GFDL-ESM2G, IPSL-CM5A-MR, MIROC-ESM, and NorESM1-M as better GCMs. Mahmood et al. (2018) used correlation coefficient, error between observed and GCM means and standard deviation, root mean square error, to assess the performance of CMIP5 GCMs in simulating precipitation over Jhelum river basin, Pakistan and found the good performance of GFDL-ESM2G, HadGEM2-ES, NorESM1-ME, CanESM2, and MIROC5. Latif et al. (2018) reported better performance of HadGEM2-AO, INM-CM4, CNRM-CM5, NorESM1-M, CCSM4 and CESM1-WACCM out of 36 GCMs in simulating precipitation over Indo-Pakistan based on partial correlation. The above findings indicated that the GCMs identified in this study for the ensemble were also found to perform well in the other studies in nearby countries/regions.

#### 4.5 Multi-model Ensemble (MME) Mean

The performances of GCM ensembles identified in Section 4.4 were validated considering two types of MME means. The MME mean of  $P$ ,  $T_{max}$  and  $T_{min}$  of the four top ranked GCMs was calculated with 1). Simple Mean (SM) and 2). Random Forest (RF). In SM, the time series of  $P$ ,  $T_{max}$  and  $T_{min}$  of the four top ranked GCMs were averaged to obtain the MME while in RF, the time series of  $P$ ,  $T_{max}$  and  $T_{min}$  of the four top ranked GCMs were considered as inputs to the RF based MME.

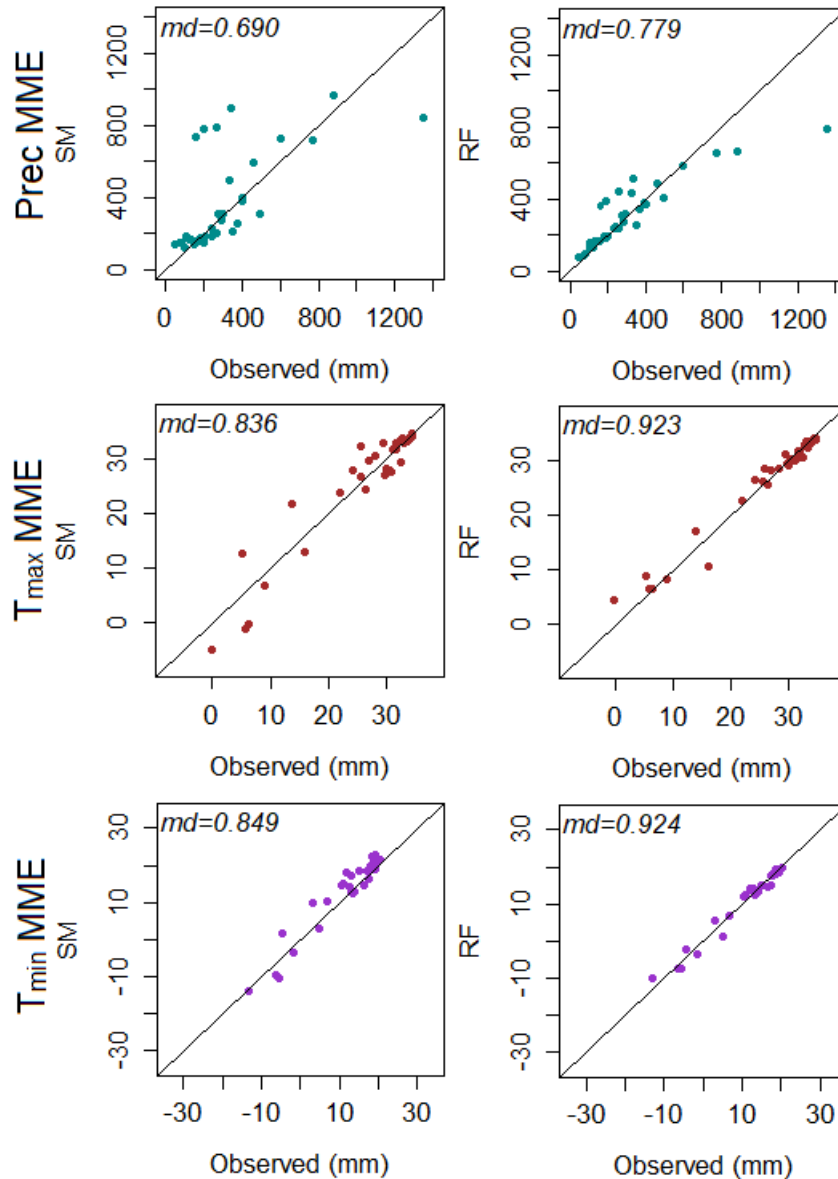
In Figure 6, the spatial patterns of  $P$ ,  $T_{max}$  and  $T_{min}$  corresponding to both MMEs derived with SM and RF were compared with that of GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$ . The spatial patterns of  $P$ ,  $T_{max}$  and  $T_{min}$  were created using Ordinary Kriging technique. Ordinary Kriging was selected as it was found to perform better than other Interpolation methods over the Pakistan (Ahmed et al., 2014). As seen in Figure 6, both MMEs captured the spatial patterns of observed  $P$ ,  $T_{max}$  and  $T_{min}$  to a good degree. However, the differences can be seen in both MMEs in replicating the spatial pattern of GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$ . The visual comparison provided in Figure 6 also indicated that RF-based MME performs better than the MME based on SM. The SM was found to underestimate

annual precipitation in the south-western and the northern regions, while the RF was found to produce spatial pattern almost identical to that of GPCC precipitation. A similar result can also be seen for maximum and minimum temperature patterns where RF showed better performance. The better performance of RF in generating MME has also been reported in several studies. Salman et al. (2018) generated MME mean for maximum and minimum temperature over Iraq using four CMIP5 GCMs and reported RF performed better compared to individual GCMs. Likewise, Wang et al. (2017b) conducted a comprehensive study to evaluate the performance of different machine learning techniques including RF, support vector machine, Bayesian model averaging and the arithmetic ensemble mean in generating MME. They considered 33 CMIP5 GCMs for precipitation and temperature over 108 station located in Australia and concluded RF and SVM can generate better MMEs compared to other techniques.



**Figure 6.** Spatial patterns of observed precipitation (a), maximum temperature (b) and minimum temperature (c), MME computed using Simple Mean (SM) (d - f) and MME computed using Random Forest (RF) (g - i) for mean annual precipitation, maximum temperature and minimum temperature respectively during 1961 to 2005.

The performance of MME ensembles was further evaluated using scatter plots shown in Figure 7. Scatter plots were developed using spatially averaged GPCCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$  and MME annual  $P$ ,  $T_{max}$  and  $T_{min}$  at all grid points for the period 1961-2005. According to scatter plots in Figure 7, RF-based MME performed significantly better compared to its counterpart SM-based MME in simulating  $P$ ,  $T_{max}$  and  $T_{min}$ .



**Figure 7.** Scatter of spatially averaged mean annual GPCCC  $P$ , 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.

Other issues:

### **Comment 1**

P2, L1 – please provide citation after several studies (related to the heatwaves, cold snaps etc.). Duffy et al. (2015) is about drought and wet spells.

### **Reply**

Thanks for your suggestion; we have added citations as given below.

“Several studies reported increase in severity and frequency of droughts (Ahmed et al., 2019a), floods (Wu et al., 2014), heatwaves (Perkins-Kirkpatrick and Gibson, 2017) and decrease in severity and frequency of cold snaps (Wang et al., 2016) in the recent years which are indicative of abrupt variations in the precipitation and temperature regimes.”

### **Comment 2**

P2, L7: please provide the correct citation. Hegerl et al.,2018 is not about the affecting hydrological cycle (that include ET, runoff, soil moisture, and precip)

### **Reply**

Sorry for the mistake, we have changed the reference as given below.

The climate modelling community has widely agreed that the sharp temperature rise in the post-industrial revolution era is significantly affecting the global hydrologic cycle (Sohoulande Djebou and Singh, 2015;Evans, 1996).

### **Comment 3**

P2, L9: should be Akhter et al., 2017

### **Reply**

Thanks, corrected as suggested.

### **Comment 4**

P2, L10: Wright et al., 2015 is about RCMs. Please provide a correct reference.

## Reply

Sorry for the mistake, we have removed the Wright et al., 2015 citation and added Pour et al., 2018 as below:

“Global Circulation Models (GCMs) are principally utilized to simulate and project climate on global scale (Pour et al., 2018; Sachindra et al., 2014).

## Comment 5

P2, L13: cite CMIP5 GCMs

## Reply

Thanks, we have added a citation to CMIP5 as shown below.

The Coupled Model Intercomparison Project Phase 5 (CMIP5) is a set of GCMs available from the IPCC AR5 (Taylor et al., 2012).

## Comment 6

P2, L14: Cited paper is not about the cmip5 and cmip3 comparison.

## Reply

Thanks, reference is replaced as below:

“GCMs showed significant improvements in climate simulations compared to its previous generation of CMIP3 models (Gao et al., 2015; Kusunoki and Arakawa, 2015).”

## Comment 7

P2, L14: more than 50 GCMs are available. Please check other papers.

## Reply

Thanks for your suggestion we have revised GCM number and citation as below:

“Currently, over 50 GCMs are available in the CMIP5 suite with different spatial resolutions (Hayhoe et al., 2017).”

### **Comment 8**

P2, L16: Ekstrom et al., 2016 is not about size and restriction on the size of the subset of GCMs.

### **Reply**

We have revised the citation as below:

“Human and computational resources pose a restriction on the size of the sub-set of GCMs used in a climate change impact assessment (Herger et al., 2018).”

### **Comment 9**

P2, L16: Salam et al., 2018a and 2018b is same

### **Reply**

Thanks. Corrected accordingly.

### **Comment 10**

P2, L17: should be 2018

### **Reply**

Thanks, corrected as suggested.

### **Comment 11**

P2, L16: cite some paper about the uncertainties in GCMs and why do we need to do ensemble mean. Please add some line about this.

### **Reply**

Thanks for the comment; we have added some references and text in relation to the above comment.

“Sa'adi et al. (2017), Salman et al. (2018), Pour et al. (2018) and Khan et al. (2018) reported that a multi-model ensemble (a sub-set) of GCMs selected considering their skills in reproducing past observed characteristics of climate can reduce the GCM associated uncertainties in climate change impact assessment.”



#### **Comment 12**

P2, L19: “prediction” (“projection”)

#### **Reply**

Thanks. Corrected as suggested.

#### **Comment 13**

P2, L22: Wang et al., 2017

#### **Reply**

Thanks. Corrected as suggested.

#### **Comment 14**

P2, L24: Wang et al., 2017

#### **Reply**

Thanks. Corrected as suggested.

#### **Comment 15**

P2, L25: Fu et al., 2018 and Dong et al., 2018 are not about the comparison between MME and individual. They are based on temperature projection.

#### **Reply**

Thanks for the comment, we have removed Fu et al., 2018 and Dong et al., 2018 and added two new references as shown below.

The SCM is relatively simple to **apply** and found to perform better than individual GCMs (Weigel et al., 2010; **Acharya et al., 2013; Wang et al., 2018**).

#### **Comment 16**

P2, L31: 2018

**Reply**

Thanks. Corrected as suggested.

**Comment 17**

P3, L15: Gleckler et al.,2008a and 2018b are same.

**Reply**

Thanks. Corrected accordingly.

**Comment 18**

P4, L1: provide citation after several studies.

**Reply**

Thanks for the suggestion; we have provided some references to support our claim.

“Overall, review of literature revealed that several studies (Khan et al., 2018;Pour et al., 2018;Salman et al., 2018;Raju et al., 2017) assessed the performance of GCMs considering several grid points over the whole study area; however they ignored the capability of GCMs to replicate the spatial patterns.

**Comment 19**

P4, L7: you used six methods. Please correct this number throughout the paper.

**Reply**

Thanks. Corrected as suggested.

**Comment 20**

P4, L11: please mention the calendar months.

**Reply**

Thanks, we have now mentioned the calendar months as shown below.

“...assessment of performance of 20 CMIP5 GCM in simulating observed annual (Jan to Dec), monsoon (Jun to Sep) and winter (Dec to Mar), pre-monsoon (Apr to May), and post-monsoon (Oct to Nov) precipitation, maximum and minimum temperature over Pakistan.”

#### Comment 21

Figure 1: should include a climate zone map also.

#### Reply

Thanks for the suggestion, we have now provided an aridity map of Pakistan separately as figure 2 adopted from the recent study by (Ahmed et al., 2019d) and added some text in study area section (2.1) as shown below.

“Pakistan is overwhelmed by arid and semi-arid climate, and displays significant climatic variations (Figure 2). Figure 2 which is based on the study by Ahmed et al. (2019d) shows that a large area of Pakistan experiences arid climate, followed by semi-arid climate, while a small area in the southwest experiences hyper-arid climate. However, a small area in the top north of the country experiences sub-humid to humid climate.

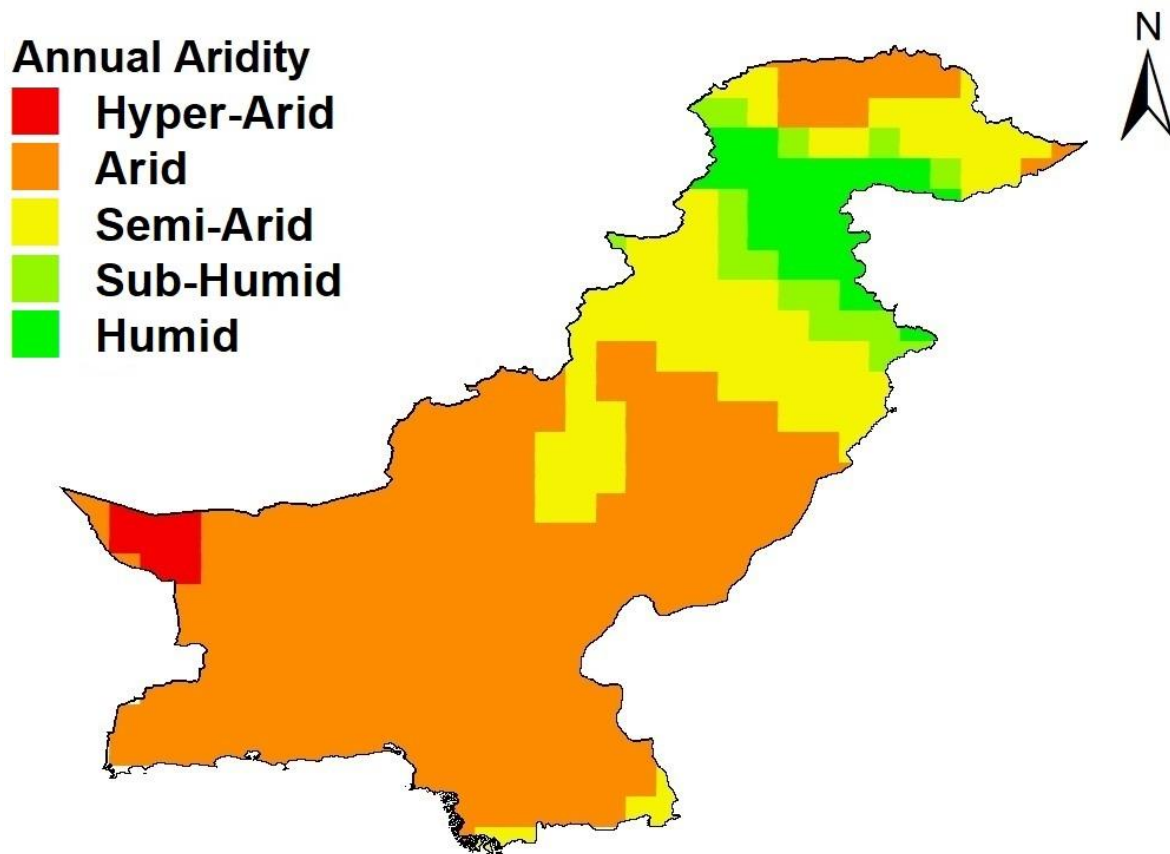


Figure 2: Aridity classification of Pakistan (adopted from (Ahmed et al., 2019d))

### Comment 22

P4, L25-29: this data conflict with the fig 3a.

### Reply

Thanks for the comment. We agree that there is some conflicting information. We checked and found that Figure 3a is prepared based on 35 grid points at a spatial resolution of  $2^{\circ} \times 2^{\circ}$  using the GPCC data for the period 1961 to 2005 while the information provided in Line 25 to 29 is based on the study by Ahmed et al. (2017) where they considered 337 grid points over Pakistan for the period 1961 to 2010. Furthermore, Ahmed et al. (2017) classified the precipitation into 10 classes while the present study classified it into seven classes (Figure 3a).

We have provided the data period as shown.

“The bulk of the summer precipitation is caused by the monsoon winds that arise from the Bay of Bengal while westerly disturbances in the Mediterranean Sea are responsible for the winter precipitation. The average precipitation in Pakistan widely varies from southwest to northern parts in the range of  $< 100$  to  $> 1000$  mm/year during 1961 to 2010. Since the country is mostly characterized by arid and semi-arid climate; the bulk of the country receives less than 500 mm/year of precipitation while only a very limited area in the north receives more than 1,000 mm/year of precipitation (Ahmed et al., 2017).”

### Comment 23

P5, L7: please provide the website link (GPCC data).

### Reply

Thanks, we have provided the weblink to GPCC data as shown below.

“In this investigation, gridded monthly precipitation data of the Global Precipitation Climatology Center (GPCC) (Schneider et al., 2013) ([dwd.de/EN/ourservices/gpcc/gpcc.html](http://dwd.de/EN/ourservices/gpcc/gpcc.html)) were used as the surrogates of observed precipitation for the period 1961-2005.”

### Comment 24

P5, L11: high correlation? Please provide the number.

### Reply

Thanks, we have provided the correlation values as shown below.

“Most importantly, GPCC precipitation data have shown correlations above 0.80 with observed precipitation over Pakistan (Ahmed et al., 2019c).”

#### Comment 25

P5, L14-20: Please mention the ensemble member that you have used in the CMIP5 GCMs.

#### Reply

Thanks, we have mentioned the ensemble member as shown below.

“Monthly precipitation data simulated by the 20 CMIP5 GCMs for ensemble member r1i1p1 run were extracted....”

#### Comment 26

P5, L15: provide a website link.

#### Reply

Thanks, a web link is provided as shown below.

“Monthly precipitation data simulated by the 20 CMIP5 GCMs were extracted from the IPCC data distribution center ([http://www.ipcc-data.org/sim/gcm\\_monthly/AR5/Reference-Archive.html](http://www.ipcc-data.org/sim/gcm_monthly/AR5/Reference-Archive.html)) for period 1961-2005.”

#### Comment 27

P6, L24: Please check the citation. In the introduction, you mentioned Demirel et al., (2018).

#### Reply

Sorry for the mistake, we have corrected it as shown below.

“SPAtial Efficiency metric (SPAEF), proposed by Demirel et al. (2018) is a robust spatial performance....”

#### Comment 28

P7, L11: Lambda (heading)

#### Reply

Thanks. Corrected as suggested.

### Comment 29

P11, L12- 25: You did not mention about the time series. Is it annual rainfall or seasonal or monthly time series? Did you check NRMSE and  $md$  between the annual time series?

### Reply

Thanks for your comment. We have revised the section 4.1 as shown below in response to your above comment.

#### “4.1 Accuracy Assessment of Gridded Precipitation Data

As a preliminary analysis, the **monthly time series** of GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$  data were validated against the **monthly time series** of observed  $P$ ,  $T_{max}$  and  $T_{min}$ . The validation was **performed** for the period 1961-2005. In the present study, two statistical metrics; Normalized Root Mean Square Error (NRMSE), and modified index of agreement ( $md$ ) were used to assess the accuracy of **monthly time series** of GPCC  $P$ , CRU  $T_{max}$  and CRU  $T_{min}$  in replicating the mean and the variability of **monthly time series** of observed  $P$ ,  $T_{max}$  and  $T_{min}$ .

The NRMSE and  $md$  values between observed  $P$  and GPCC  $P$  (pertaining to the grid point closest to the observation station), **observed  $T_{max}$  and  $T_{min}$  with CRU  $T_{max}$  and  $T_{min}$**  obtained for 17 locations in Pakistan are given in Table 2. Overall, all the stations showed low and high NRMSE and  $md$  values respectively, indicating that the accuracy of the GPCC  $P$  in replicating observed precipitation and **CRU  $T_{max}$  and CRU  $T_{min}$**  in replicating observed  $T_{max}$  and  $T_{min}$  over Pakistan is high. **Overall, NRMSE values were found in the ranges of 0.09 to 0.970 for  $P$ , 0.100 to 0.390 for  $T_{max}$ , and 0.09 to 0.470 for  $T_{min}$ . At the same time, overall,  $md$  values were found in the ranges of 0.680 to 0.960 for  $P$ , 0.810 to 0.960 for  $T_{max}$ , and 0.779 to 0.959 for  $T_{min}$ .**

**Table 2.** Validation of accuracy of GPCC precipitation using NRMSE and  $md$

Station No	Station Name	Precipitation ( $P$ )		Maximum Temperature ( $T_{max}$ )		Minimum Temperature ( $T_{min}$ )	
		NRMSE	$md$	NRMSE	$md$	NRMSE	$md$
1	Karachi	0.530	0.840	0.270	0.880	0.180	0.919
2	Pasni	0.470	0.890	0.310	0.840	0.260	0.879
3	Nawabshah	0.740	0.740	0.300	0.850	0.170	0.919
4	Padidan	0.590	0.780	0.190	0.920	0.150	0.939
5	Jacobabad	0.520	0.840	0.100	0.960	0.090	0.959
6	Dalbandin	0.090	0.960	0.140	0.940	0.230	0.889
7	Kalat	0.970	0.870	0.240	0.900	0.470	0.779
8	Sibbi	0.590	0.880	0.390	0.810	0.260	0.889
9	Bahawalnagar	0.530	0.810	0.310	0.899	0.270	0.881
10	Quetta	0.750	0.760	0.240	0.890	0.120	0.949
11	Multan	0.730	0.740	0.120	0.950	0.120	0.949
12	Faisalabad	0.700	0.740	0.210	0.900	0.170	0.919
13	Lahore	0.710	0.700	0.140	0.940	0.110	0.959
14	Sargodha	0.790	0.680	0.160	0.930	0.170	0.919
15	Mianwali	0.720	0.750	0.240	0.890	0.120	0.949
16	Islamabad	0.450	0.840	0.160	0.930	0.190	0.909
17	Peshawar	0.690	0.720	0.190	0.920	0.110	0.949

**Comment 30**

No need for figure 2. You can remove the figure 2 and include the rank in table 3 in brackets.

**Reply**

Thanks for your suggestion, we have removed Figure 2 and included ranks in Table 3 in brackets as shown below.

**Table 3.** GOF values and ranks of GCMs obtained using different spatial metrics for mean annual precipitation

<b>GCM</b>	<b>SPAEF (Rank)</b>	<b>FSS (Rank)</b>	<b>Lambda (Rank)</b>	<b>Cramer- V(Rank)</b>	<b>Mapcurves (Rank)</b>	<b>KGE (Rank)</b>
ACCESS1-0	0.411(7)	0.659(24)	0.143(24)	0.370(28)	0.244(29)	0.172(29)
ACCESS1-3	0.155(24)	0.712(20)	0.107(30)	0.315(34)	0.206(34)	0.310(15)
BCC-CSM1-1	0.241(21)	0.691(21)	0.143(24)	0.388(27)	0.258(27)	0.082(33)
BCC-CSM1.1(m)	0.149(25)	0.685(22)	0.214(13)	0.545(16)	0.376(16)	0.304(16)
BNU-ESM	0.185(23)	0.759(11)	0.179(18)	0.519(21)	0.349(21)	0.233(26)
CanESM2	0.250(20)	0.642(26)	0.250(6)	0.547(15)	0.378(15)	-0.443(35)
CCSM4	0.440(4)	0.798(5)	0.250(6)	0.667(4)	0.525(4)	0.420(8)
CESM1-BGC	0.439(5)	0.759(12)	0.214(13)	0.655(10)	0.508(10)	0.337(12)
CESM1-CAM5	0.540(1)	0.840(1)	0.250(6)	0.667(4)	0.525(4)	0.531(2)
CESM1-WACCM	0.430(6)	0.776(10)	0.250(6)	0.656(9)	0.510(9)	0.384(10)
CMCC-CM	-0.255(34)	0.565(33)	0.143(24)	0.496(24)	0.325(24)	0.189(28)
CMCC-CMS	-0.043(28)	0.637(28)	0.143(24)	0.369(29)	0.244(28)	0.249(22)
CNRM-CM5	0.364(12)	0.732(17)	0.250(6)	0.667(4)	0.525(4)	0.314(14)
CSIRO-Mk3-6-0	-0.505(36)	0.321(36)	0.036(36)	0.264(36)	0.179(36)	-1.837(36)
EC-EARTH	0.232(22)	0.756(13)	0.286(4)	0.759(2)	0.642(2)	0.404(9)
FGOALS-g2	0.321(13)	0.793(6)	0.179(18)	0.531(17)	0.361(17)	0.362(11)
FIO-ESM	0.281(17)	0.752(14)	0.214(13)	0.559(14)	0.391(14)	0.283(19)
GFDL-CM3	0.387(8)	0.815(4)	0.429(1)	0.782(1)	0.690(1)	0.493(3)
GFDL-ESM2G	0.307(14)	0.786(7)	0.250(6)	0.667(4)	0.525(4)	0.484(4)
GFDL-ESM2M	0.297(16)	0.778(8)	0.214(13)	0.436(26)	0.296(25)	0.458(5)
GISS-E2-H	-0.100(32)	0.616(31)	0.107(30)	0.335(33)	0.220(33)	0.245(24)
GISS-E2-R	-0.054(29)	0.616(30)	0.107(30)	0.350(31)	0.229(31)	0.236(25)
HadGEM2-AO	0.454(3)	0.740(15)	0.179(18)	0.520(20)	0.350(20)	0.315(13)
HadGEM2-CC	0.387(9)	0.683(23)	0.179(18)	0.360(30)	0.236(30)	0.222(27)
HadGEM2-ES	0.371(11)	0.721(18)	0.179(18)	0.530(18)	0.360(18)	0.277(20)
INMCM4	0.378(10)	0.777(9)	0.179(18)	0.530(18)	0.360(18)	0.422(6)
IPSL-CM5A-LR	-0.054(30)	0.634(29)	0.357(2)	0.590(12)	0.427(12)	0.117(32)
IPSL-CM5A-MR	-0.093(31)	0.548(34)	0.357(2)	0.590(12)	0.427(12)	-0.183(34)
IPSL-CM5B-LR	-0.286(35)	0.538(35)	0.107(30)	0.350(31)	0.229(31)	0.131(31)
MIROC-ESM- CHEM	0.273(18)	0.733(16)	0.214(13)	0.655(10)	0.508(10)	0.303(17)
MIROC-ESM	0.258(19)	0.720(19)	0.286(4)	0.677(3)	0.537(3)	0.29(18)
MIROC5	0.302(15)	0.828(3)	0.071(34)	0.454(25)	0.285(26)	0.420(7)
MPI-ESM-LR	-0.012(27)	0.639(27)	0.143(24)	0.517(22)	0.346(22)	0.253(21)
MPI-ESM-MR	0.041(26)	0.653(25)	0.143(24)	0.506(23)	0.335(23)	0.245(23)
MRI-CGCM3	-0.180(33)	0.572(32)	0.071(34)	0.293(35)	0.194(35)	0.169(30)
NorESM1-M	0.464(2)	0.833(2)	0.250(6)	0.667(4)	0.525(4)	0.532(1)

- *The numbers in the brackets are obtained ranks*



## References

- Acharya, N., Singh, A., Mohanty, U. C., Nair, A., and Chattopadhyay, S.: Performance of general circulation models and their ensembles for the prediction of drought indices over India during summer monsoon, *Nat. Hazards*, 66, 851-871, 10.1007/s11069-012-0531-8, 2013.
- Afshar, A. A., Hasanzadeh, Y., Besalatpour, A. A., and Pourreza-Bilondi, M.: Climate change forecasting in a mountainous data scarce watershed using CMIP5 models under representative concentration pathways, *Theor. Appl. Climatol.*, 129, 683-699, 10.1007/s00704-016-1908-5, 2016.
- Ahmadalipour, A., Rana, A., Moradkhani, H., and Sharma, A.: Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis, *Theor. Appl. Climatol.*, 128, 71-87, 10.1007/s00704-015-1695-4, 2017.
- Ahmed, K., Shahid, S., and Harun, S. B.: Spatial interpolation of climatic variables in a predominantly arid region with complex topography, *Environment Systems and Decisions*, 34, 555-563, 2014.
- Ahmed, K., Shahid, S., Chung, E.-S., Ismail, T., and Wang, X.-J.: Spatial distribution of secular trends in annual and seasonal precipitation over Pakistan, *Climate Research*, 74, 95-107, 2017.
- Ahmed, K., Shahid, S., Chung, E.-S., Wang, X.-j., and Harun, S. B.: Climate Change Uncertainties in Seasonal Drought Severity-Area-Frequency Curves: Case of Arid Region of Pakistan, *J. Hydrol.*, <https://doi.org/10.1016/j.jhydrol.2019.01.019>, 2019a.
- Ahmed, K., Shahid, S., Nawaz, N., and Khan, N.: Modeling climate change impacts on precipitation in arid regions of Pakistan: a non-local model output statistics downscaling approach, *Theor. Appl. Climatol.*, 137, 1347-1364, 10.1007/s00704-018-2672-5, 2019b.
- Ahmed, K., Shahid, S., Sachindra, D. A., Nawaz, N., and Chung, E.-S.: Fidelity assessment of general circulation model simulated precipitation and temperature over Pakistan using a feature selection method, *J. Hydrol.*, 573, 281-298, <https://doi.org/10.1016/j.jhydrol.2019.03.092>, 2019c.
- Ahmed, K., Shahid, S., Wang, X., Nawaz, N., and Khan, N.: Spatiotemporal changes in aridity of Pakistan during 1901–2016, *Hydrol. Earth Syst. Sci.*, 23, 3081-3096, 10.5194/hess-23-3081-2019, 2019d.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5-32, 10.1023/A:1010933404324, 2001.
- Cramér, H.: *Mathematical methods of statistics (PMS-9)*, Princeton university press, 1999.
- Das, L., Dutta, M., Mezghani, A., and Benestad, R. E.: Use of observed temperature statistics in ranking CMIP5 model performance over the Western Himalayan Region of India, *Int. J. Climatol.*, 38, 554-570, 2018.
- Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, *Hydrol. Earth Syst. Sci.*, 22, 1299-1315, 2018.
- Evans, T. E.: *The effects of changes in the world hydrological cycle on availability of water resources, Global Climate Change and Agricultural Production: Direct and Indirect Effects of Changing Hydrological, Pedological and Plant Physiological Processes*, 1996.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M.: Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning, *Agr. Forest Meteorol.*, 264, 1-15, 2019.

Gao, Y., Wang, H., and Jiang, D.: An intercomparison of CMIP5 and CMIP3 models for interannual variability of summer precipitation in Pan-Asian monsoon region, *Int. J. Climatol.*, 35, 3770-3780, 2015.

Goodman, L. A., and Kruskal, W. H.: Measures of association for cross classifications, *Journal of the American statistical association*, 49, 732-764, 1954.

Hargrove, W. W., Hoffman, F. M., and Hessburg, P. F.: Mapcurves: a quantitative method for comparing categorical maps, *J. Geog. Syst.*, 8, 187, 2006.

Hayhoe, K., Edmonds, J., Kopp, R., LeGrande, A., Sanderson, B., Wehner, M., and Wuebbles, D.: Climate models, scenarios, and projections, 2017.

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J.: Spatial downscaling of precipitation using adaptable random forests, *Water Resour. Res.*, 52, 8217-8237, 2016.

Herger, N., Abramowitz, G., Knutti, R., Angéilil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, *Earth System Dynamics*, 9, 135-151, 2018.

Jena, P., Azad, S., and Rajeevan, M. N.: Statistical selection of the optimum models in the CMIP5 dataset for climate change projections of Indian monsoon rainfall, *Climate*, 3, 858-875, 2015.

Khan, N., Shahid, S., Ahmed, K., Ismail, T., Nawaz, N., and Son, M.: Performance Assessment of General Circulation Model in Simulating Daily Precipitation and Temperature Using Multiple Gridded Datasets, *Water*, 10, 1793, 2018.

Koch, J., Demirel, M. C., and Stisen, S.: The SPAtial Efficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models, *Geoscientific Model Development*, 11, 1873-1886, 2018.

Kusunoki, S., and Arakawa, O.: Are CMIP5 Models Better than CMIP3 Models in Simulating Precipitation over East Asia?, *J. Clim.*, 28, 5601-5621, 10.1175/JCLI-D-14-00585.1, 2015.

Latif, M., Hannachi, A., and Syed, F.: Analysis of rainfall trends over Indo-Pakistan summer monsoon and related dynamics based on CMIP5 climate model simulations, *Int. J. Climatol.*, 38, e577-e595, 2018.

Mahmood, R., Jia, S., Tripathi, N., and Shrestha, S.: Precipitation Extended Linear Scaling Method for Correcting GCM Precipitation and Its Evaluation and Implication in the Transboundary Jhelum River Basin, *Atmosphere*, 9, 160, 2018.

Meher, J. K., Das, L., Akhter, J., Benestad, R. E., and Mezghani, A.: Performance of CMIP3 and CMIP5 GCMs to simulate observed rainfall characteristics over the Western Himalayan region, *J. Clim.*, 30, 7777-7799, 2017.

Nowosad, J., and Stepinski, T. F.: Spatial association between regionalizations using the information-theoretical V-measure, *International Journal of Geographical Information Science*, 32, 2386-2401, 2018.

Onyutha, C., Tabari, H., Rutkowska, A., Nyeko-Ogiramoi, P., and Willems, P.: Comparison of different statistical downscaling methods for climate change rainfall projections over the Lake Victoria basin considering CMIP3 and CMIP5, *Journal of hydro-environment research*, 12, 31-45, 2016.

Perkins-Kirkpatrick, S. E., and Gibson, P. B.: Changes in regional heatwave characteristics as a function of increasing global temperature, *Scientific Reports*, 7, 12256, 10.1038/s41598-017-12520-2, 2017.

Pocernich, M. M.: The verification package, available at [cran.r-project.org](http://cran.r-project.org), 2006.

Pour, S. H., Shahid, S., Chung, E.-S., and Wang, X.-J.: Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh, *Atmos. Res.*, 213, 149-162, <https://doi.org/10.1016/j.atmosres.2018.06.006>, 2018.

Prasanna, V.: Regional climate change scenarios over South Asia in the CMIP5 coupled climate model simulations, *Meteorol. Atmos. Phys.*, 127, 561-578, 2015.

Raju, K. S., Sonali, P., and Kumar, D. N.: Ranking of CMIP5-based global climate models for India using compromise programming, *Theor. Appl. Climatol.*, 128, 563-574, 2017.

Rees, W.: Comparing the spatial content of thematic maps, *Int. J. Remote Sens.*, 29, 3833-3844, 2008.

Rehman, N., Adnan, M., and Ali, S.: Assessment of CMIP5 climate models over South Asia and climate change projections over Pakistan under representative concentration pathways, *International Journal of Global Warming*, 16, 381-415, 2018.

Roberts, N. M., and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, *Monthly Weather Review*, 136, 78-97, 2008.

Sa'adi, Z., Shahid, S., Chung, E.-S., and bin Ismail, T.: Projection of spatial and temporal changes of rainfall in Sarawak of Borneo Island using statistical downscaling of CMIP5 models, *Atmos. Res.*, 197, 446-460, 2017.

Sachindra, D., Huang, F., Barton, A., and Perera, B.: Statistical downscaling of general circulation model outputs to precipitation—part 2: bias-correction and future projections, *Int. J. Climatol.*, 34, 3282-3303, 2014.

Salman, S. A., Shahid, S., Ismail, T., Ahmed, K., and Wang, X.-J.: Selection of climate models for projection of spatiotemporal changes in temperature of Iraq with uncertainties, *Atmos. Res.*, 213, 509-522, <https://doi.org/10.1016/j.atmosres.2018.07.008>, 2018.

Sarathi, P. P., Kumar, P., and Ghosh, S.: Possible future rainfall over Gangetic Plains (GP), India, in multi-model simulations of CMIP3 and CMIP5, *Theor. Appl. Climatol.*, 124, 691-701, 10.1007/s00704-015-1447-5, 2016.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCP's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, *Theor. Appl. Climatol.*, 115, 15-40, 10.1007/s00704-013-0860-x, 2013.

Signorell, A.: DescTools: Tools for descriptive statistics, R package version 0.99, 18, 2016.

Sohoulane Djebou, D., and Singh, V.: Impact of climate change on the hydrologic cycle and implications for society, *Environ Soc Psychol*, 1, 9-16, 2015.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, 93, 485-498, 2012.

Wang, B., Zheng, L., Liu, D. L., Ji, F., Clark, A., and Yu, Q.: Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia, *Int. J. Climatol.*, 0, doi:10.1002/joc.5705, 2017a.

Wang, B., Zheng, L., Liu, D. L., Ji, F., Clark, A., and Yu, Q.: Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia, *Int. J. Climatol.*, 38, 4891-4902, 2018.

Wang, X., Yang, T., Li, X., Shi, P., and Zhou, X.: Spatio-temporal changes of precipitation and temperature over the Pearl River basin based on CMIP5 multi-model ensemble, *Stoch. Environ. Res. Risk Assess.*, 31, 1077-1089, 10.1007/s00477-016-1286-7, 2017b.

Wang, Y., Shi, L., Zanobetti, A., and Schwartz, J. D.: Estimating and projecting the effect of cold waves on mortality in 209 US cities, *Environment international*, 94, 141-149, 2016.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of model weighting in multimodel climate projections, *J. Clim.*, 23, 4175-4191, 2010.

Wu, C., Huang, G., Yu, H., Chen, Z., and Ma, J.: Impact of Climate Change on Reservoir Flood Control in the Upstream Area of the Beijiang River Basin, South China, *J. Hydrometeorol.*, 15, 2203-2218, 10.1175/jhm-d-13-0181.1, 2014.

Xue, Y., Vasic, R., Janjic, Z., Mesinger, F., and Mitchell, K. E.: Assessment of dynamic downscaling of the continental US regional climate using the Eta/SSiB regional climate model, *J. Clim.*, 20, 4172-4193, 2007.

Zawadzka, J., Mayr, T., Bellamy, P., and Corstanje, R.: Comparing physiographic maps with different categorisations, *Geomorphology*, 231, 94-100, 2015.