



# A new Monte Carlo Feature Selection (MCFS) algorithm-based weighting scheme for multi-model ensemble of precipitation

Abdul Baseer<sup>1</sup> · Zulfiqar Ali<sup>1</sup> · Maryam Ilyas<sup>1</sup> · Mahrukh Yousaf<sup>1</sup>

Received: 6 July 2023 / Accepted: 7 September 2023 / Published online: 15 September 2023  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2023

## Abstract

Changes in patterns of meteorological parameters, like precipitations, temperature, wind, etc., are causing significant increases in various extreme events. And these extreme events, i.e., floods, heatwaves, hurricanes, droughts, etc., lead to a shortage of water resources, crop failures, wildfires, and economic losses. However, Global Circulation Models (GCMs) are considered the most important tools for quantifying climate change. Therefore, we selected 20 different GCMs of precipitation in our research, as the frequency of extreme events, like drought and flood, is highly related to changes in precipitation patterns. However, this research introduced a new weighting scheme — MCFSAWS-Ensemble: Monte Carlo Feature Selection Adaptive Weighting Scheme to Ensemble multiple GCMs, whereas, Monte Carlo Feature Selection (MCFS) is one of the most popular algorithms for discovering important variables. However, the proposed weighting scheme (MCFSAWS-Ensemble) is mainly based on two sources. Initially, it evaluates the prior performance of each GCM model to define their relative importance using MCFS. Then, it computes value by value difference between the observed and simulated model. In addition, the application of this paper is based on the monthly time series data of precipitation in the Tibet Plateau region of China. In addition, we used twenty GCMs from the Coupled Model Intercomparison Project Phase 6 (CMIP6) to analyze the implications of the MCFSAWS-Ensemble. Further, we compared the performance of the MCFSAWS-Ensemble scheme with Simple Model Averaging (SMA) through Mean Average Error (MAE) and correlation statistics. The results of this research indicate that the proposed weighting scheme (MCFSAWS-Ensemble) is more accurate than the SMA approach. Consequently, we recommend the use of advanced machine learning algorithms such as MCFS for making accurate multi-model ensembles.

## 1 Introduction

In recent decades, global warming has emerged as one of the most challenging problems (Smith et al. 2018; Mokhov 2022), with human-induced heat-trapping gases such as  $CO_2$ ,  $CH_4$ , and  $N_2O$  significantly contributing to the rise in global temperatures (Liu et al. 2022a; Yasmin et al. 2022). Consequently, various adverse effects on forestry (Sperry et al. 2019), agriculture (Baldos et al. 2019), hydrology (Wine

and Davison 2019), Physiology of fishes in water (Alfonso et al. 2021), livestock (Lacetera, 2019), and climate have been observed due to the escalating levels of these greenhouse gases (Rajak 2021). Global warming also disrupts the natural cycle of meteorological variables (Dou et al. 2022; Zhang et al. 2021), intensifying evaporation and leading to localized storms and droughts. This amplified water cycle gives rise to extreme weather conditions such as floods and droughts (Duan and Duan 2020; Oh et al. 2020; Çakmak et al. 2021; Wei et al. 2021; Çakmak and Acar 2022). Moreover, Mare et al. (2018) found that the number of fatalities resulting from precipitation-related natural disasters consistently exceeds those caused by all other global incidents. Consequently, it is crucial to monitor and forecast precipitation characteristics with greater accuracy to ensure environmental sustainability (Perović et al. 2021).

Numerous methods have been developed in the literature to forecast global climate changes in different regions and time periods (Russo et al. 2022). GCMs are mathematical models that are built upon biological and physical laws, playing a crucial

✉ Zulfiqar Ali  
zulfiqar.stat@pu.edu.pk

Abdul Baseer  
1560113546941@pu.edu.pk

Maryam Ilyas  
maryam.stat@pu.edu.pk

Mahrukh Yousaf  
3520229327396@pu.edu.pk

<sup>1</sup> College of Statistical Sciences, University of the Punjab, Lahore, Pakistan

role in understanding global climate variations and predicting future conditions (Try et al. 2022; Hamed et al. 2022). Whereas, CMIP is a collaborative project of scientists that standardizes the use of GCMs and analyzes the outcomes generated by these models. For instance, Wu et al. (2020) employed climate models from CMIP6 to forecast changes in wind speed, while Yue et al. (2021) utilized GCMs from phase 6 of CMIP to analyze and project changes in precipitation and temperatures. There are various sources of uncertainties associated with climate models (Davies-Barnard et al. 2022; Lovenduski et al. 2016; Zheng et al. 2021). To enhance the robustness of climate projections, multiple models are often combined using Multiple Model Ensemble (MME) approach (Kim et al. 2020).

MME is a technique that combines various models to provide a more comprehensive and reliable understanding. However, different weighted and unweighted strategies exist in literature to ensemble multiple models (Mudryk et al. 2020). The unweighted ensemble methods combine all GCMs by simple average (Raju and Kumar 2020). For example, Xu et al. (2019), Liu et al. (2022), and Dong et al. (2021) used the SMA approach to create an ensemble, whereas, weighted MME schemes assign unequal weights to each model (Kim et al. 2020; Mudryk et al. 2020). These weights reflect the prior performance of the model (Jose et al. 2022; Morim et al. 2020). Therefore, several researchers developed different weighting schemes to ensemble the data (Knutti et al. 2019; Sanderson et al. 2015). For example, Bayesian Model Averaging (BMA) is a prominent ensemble method that assigns weights through the posterior probability distributions of models (Ombadi et al. 2021; Raftery et al. 2005; Zhang et al. 2016). Wootten et al. (2020) ensemble multiple models of CMIP6 by using BMA to forecast precipitation. Some other weighting strategies include cue weighting strategy (Otterbring et al. 2022), Flow-based weighting scheme (Dong et al. 2021), and Copula-based Bayesian Model Averaging (CBMA) (Ehteram et al. 2022; Seifi et al. 2022). Weighted ensemble projections are generally considered more robust compared to non-weighted ensemble techniques (Xu et al. 2022a; Scafetta 2022).

To the best of our knowledge, the extent to which extreme values influence GCM predictions is not yet fully understood in many MME weighting approaches. Moreover, no existing weighting scheme ranked models based on their relevant features for better accuracy. Hence, the proposed weighting scheme evaluates the value-to-value variation between observed and simulated models to reduce the influence of extreme values. Further, this research incorporates Monte Carlo Feature Selection (MCFS) (Dramiński et al. 2008) to identify the relative importance (RI) of GCMs in comparison to observed data. The choice to utilize MCFS is based on its superiority over other algorithms in terms of improved performance in handling data-related issues such as multicollinearity, capturing nonlinear interactions, and providing flexibility in model selection.

MCFS is a machine learning tool, and in contemporary times, machine learning applications are playing essential roles in forecasting and classification. This is due to fact of availability of large and high dimension data from various fields such as healthcare, social media, online education, and environmental sciences (Nematzadeh et al. 2019; Zhou et al. 2022). The high dimensional data set may overfit or underfit the results due to redundant or irrelevant features (Alirezanejad et al. 2020). Therefore, feature selection is necessary not only for effectively handling several variables but also for the selection of relevant features for accurate modeling and prediction (Tadist et al. 2019).

In the literature, there are several machine learning-based algorithms of machine learning (Hasan and Bao 2021). There are two main types of feature selection techniques that exist in literature, i.e., filter and wrapper techniques. The wrapper technique is further divided into wrapped methods and embedded methods (Li et al. 2021a). The wrapper methods use machine learning algorithms, and the embedded ones use techniques such as RIDGE and LASSO regression (Alirezanejad et al. 2020; Zhou et al. 2022). On the other hand, the filter approaches don't use any models. They have their own computationally efficient techniques (Alhakami et al. 2019).

Overall, the objective of this research is to incorporate the implications of the MCFS algorithm with some mathematical formulations to improve the ensemble of multiple GCMs. As MCFS algorithm alone only gives a single value weight to models by evaluating the whole prior performance, the research proposes a new weighting scheme called MCFSAWS-Ensemble that considers the value-to-value variation and provides a weight to each value of model. Hence, by using the MCFSAWS-Ensemble, policymakers can make better policies for a sustainable environment and can reduce the impact of extreme events on society. For application purposes, MCFSAWS-Ensemble focused on the precipitation data of the Tibet Plateau region of China and considered 20 GCM simulations from CMIP6.

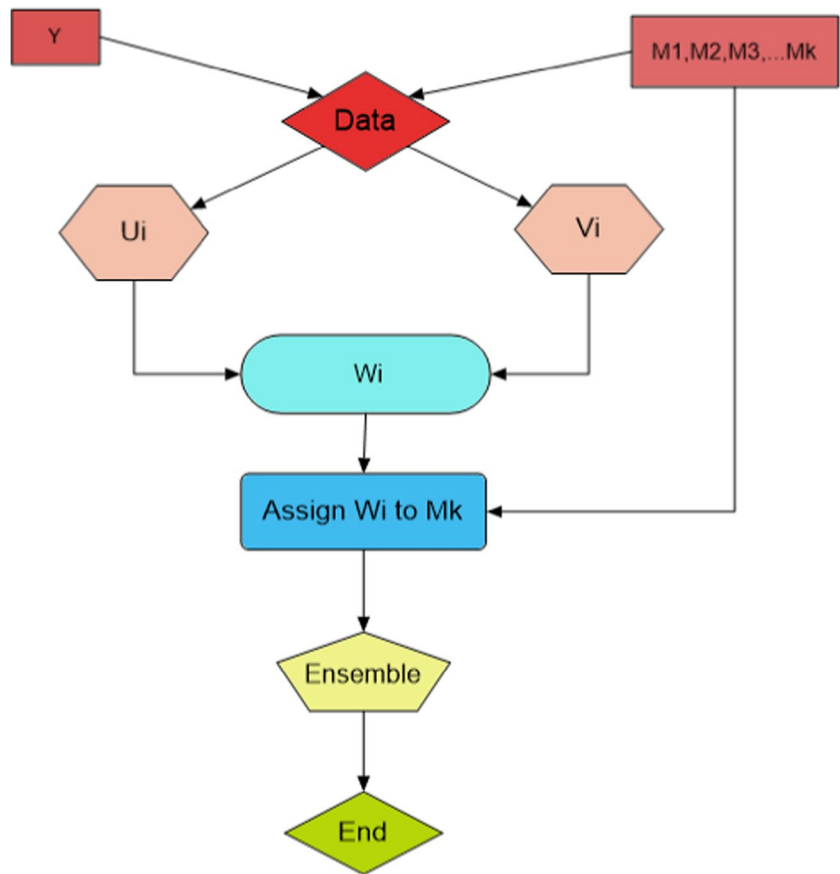
The remainder of this paper is organized as follows: [Section 2](#) presents the existing and proposed methods. [Section 3](#) provides a description of the study area and data. [Section 4](#) discusses the results. Finally, [Section 5](#) presents the conclusions [Fig. 1](#).

## 2 Methods

### 2.1 Monte Carlo Feature Selection

Monte Carlo Feature Selection (MCFS) is a variables/features ranking algorithm of machine learning developed by Dramiński et al. (2008). The MCFS algorithm consists of three main steps. The first step involves estimating the importance of features. In the second step, validation is performed to

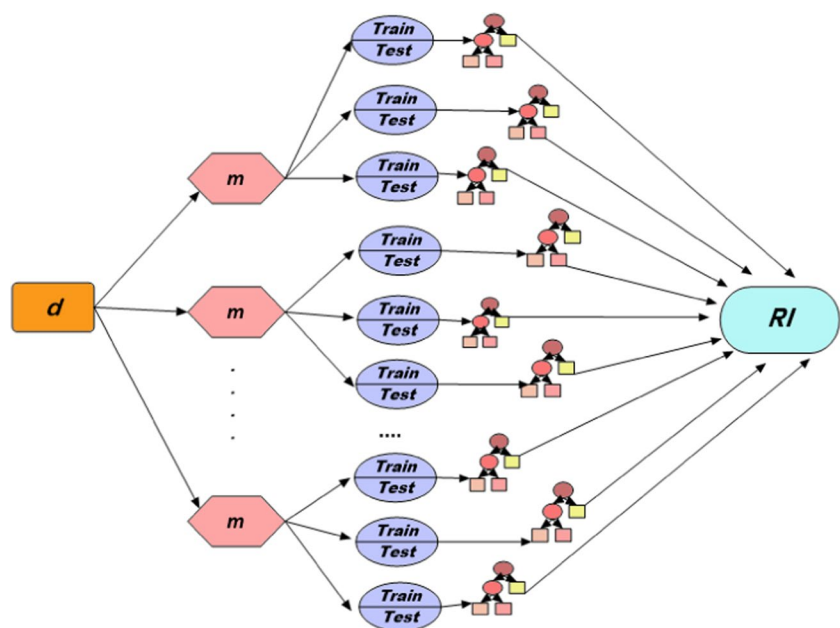
**Fig. 1** Flow chart of the proposed weighting scheme



evaluate the performance of the selected features. Finally, the third and last step confirms the features and identifies the most important ones then assigns weights to them according to their significance. The visual representation of the MCFS

algorithm is shown in Fig. 2. In summation, MCFS identifies the informative and non-informative features and ranks them, accordingly (Li et al. 2020). It uses mathematical models to depict the variations between the features in terms of

**Fig. 2** Flowchart of the MCFS algorithm



probability distributions (Niaz et al. 2020). Several iterations of a random sample procedure are performed to derive the subsets of features (Tadist et al. 2019). Different subsets are chosen at each iteration. The outcomes are used to estimate the probability that each feature is relevant to the classification task. The main purpose of the MCFS algorithm is to estimate the ranking of several features by making a thousand trees for selected random subsets (Yan et al. 2019). For a sample of size  $n$ , consider a classification problem with  $c$  classes (Dramiński et al. 2010). The weighted accuracy (WAcc) of a tree is used to determine the classification skill of the tree on a test set. The mathematical formulation of WAcc is mentioned below.

$$\text{WAcc} = \frac{1}{c} \sum_{j=1}^c \frac{n_{jj}}{n_{j1} + n_{j2} + \dots + n_{jc}} \quad (1)$$

WAcc highlights the relative significance of a specific feature. The classification of the number of samples from the  $j^{\text{th}}$  class to the  $k^{\text{th}}$  class is  $n_{jk}$  ( $j, k = 1, 2, 3, \dots, c; \sum_{jk} n_{jk} = n$ ). Therefore, WAcc is just the average of true positive rates of all the classes (Equation (1)). This weight accuracy is used to determine the relative importance ( $RI_{jk}$ ) of a feature ( $jk$ ) (Equation (2)). The mathematical formulation  $RI_{jk}$  is given below.

$$RI_{jk} = \sum_{i=1}^{s,t} \text{WAcc}_i^u GR(n_{jk}(i)) \left( \frac{\text{no.in } n_{jk}(i)}{\text{no.in } i} \right)^v \quad (2)$$

In the above equation,  $s, t$  is the total number of trees. The nodes of  $i^{\text{th}}$  tree for feature  $jk$  are  $n_{jk}(i)$ . The gain ratio for node  $n_{jk}(i)$  is  $GR(n_{jk}(i))$ . Whereas,  $\text{no. in } n_{jk}(i)$  is the number of samples in the node  $n_{jk}(i)$ . On the other hand, the number of samples in  $i^{\text{th}}$  root of the tree is  $(\text{no. in } i)$ , and  $v$  and  $u$  are positive real numbers.

## 2.2 The proposed weighting scheme — MCFAWS-Ensemble: Monte Carlo Feature Selection Adaptive Weighting Scheme for GCMs Ensemble

This section presents the mathematical structure of the proposed ensemble weighting scheme and its flow chart given in Fig. 1. The proposed weighting scheme consists of the hybridization of two types of weights derived from two types of sources. The first sources quantify the relative importance of simulated data. While the second source extracts weights by quantifying value by value difference of GCMs with observed data. We hypothesize that the first sourced weights will reflect the relative importance of each climate model in a multi-model ensemble, whereas, the second sourced weights ensure the diminishment of the impact of outliers in model aggregation.

Mathematically, consider the multivariate precipitation time series of observed and simulated data from multiple climate models at a single grid point, denoted as,  $R = [Y, M_1, M_2, M_3, \dots, M_k]$ . Here,  $Y$  is the observed precipitation data of a particular grid point and  $M_1, M_2, M_3, \dots, M_k$  are the temporal vectors of the simulated data. To ensemble multiple models under various future scenarios, this research suggests the following steps for the derivation of weights against each model.

### 2.2.1 Source 1 — Implication of model importance using MCFS

This source computes the overall importance of each model while considering their prior performance. We considered the Relative Importance Score (RIS) ( $V_i$ ) as the first source weight for each model, computed using Equation (2). In this study, the RIS is calculated using the `rmcfs` library of the R software. This score serves as the initial weight for each simulated model. In this paper, we denote the RIS as the first source weighting using  $V_i$ .

### 2.2.2 Source 2 — Real-time value by value base extraction of weights under exponential transformation

This source provides a set of equations that extracts weights for each model by transforming the deviations among each value of the observed ( $Y$ ) that and simulated data ( $M_i$ ). The transformation is made in such a way that the nearest values of climate models ( $M_i$ ) to observed data ( $Y$ ) receive high weight and vice versa. Mathematically, firstly we suggest taking the absolute difference between the observed and each model value using the following equation.

$$d_i = |Y - M_i| \quad (3)$$

Secondly, the following equation exponential the differences computed by Equation (3). The main purpose of exponentiation of the difference is to explore the impact of extreme values and outliers.

$$z_i = e^{d_i} \quad (4)$$

Thirdly, the following transformation provides a set of indices that describes the closeness of simulated to observed data.

$$p_i = 1 - \frac{z_i}{q} \quad (5)$$

In the above equation,  $q = \sum_{i=1}^k z_i$ .

The main objective of this transformation is to assign weights to each GCM according to its distance to the observed data set. This transformation allows higher weight to lower distance value and vice versa.

Then, we standardized the weights computed from Equation (5) by the following mathematical equation.

$$U_i = \frac{P_i}{\sum_{i=1}^k P_i} \tag{6}$$

In the above equation,  $U_i$  are the standardized weights and  $\sum_{i=1}^k U_i = 1$ . Further, we hybridize the weights combining the initial ( $V_i$ ) and standardized weights ( $U_i$ ) of  $i^{\text{th}}$  GCM through simple average (Equation (7)).

$$b_i = \frac{V_i + U_i}{2} \tag{7}$$

$$W_i = \frac{b_i}{\sum_{i=1}^k b_i} \tag{8}$$

All the  $k$  time series ( $i = 1, 2, 3 \dots k$ ) are iterated through Equations (3), (4), (5), and (6). Further, we aggregated data of multiple GCMs using the proposed weighting scheme that accounts for unequal weights in the multi-model ensemble. Mathematically,

$$X_t = \sum_{i=1}^k W_{it} M_{it} \tag{9}$$

In the above equation,  $X_t$  denoted aggregated data of multiple GCMs under the proposed ensemble scheme.

### 2.3 Comparative methods and measures

In this article, we used two comparative statistical measures, namely Mean Absolute Error (MAE) and Pearson correlation coefficient, to assess the appropriateness and efficiency of MCFSAWS-Ensemble. These measures were then compared with Simple Models Averaging (SMA) method.

SMA is a statistical method that combines models by giving each model equal weights (e.g., Dey et al. 2022; Zeng et al. 2022). It is simple and easy to implement. Therefore, it is widely used to combine multiple models. This simple average of  $k$  climate models is presented in Equation (10). Here,  $M_j(t)$  is the output of  $j^{\text{th}}$  GCM at time  $t$ .

$$SMA = \frac{1}{k} \sum_{j=1}^k M_j(t) \tag{10}$$

MAE is a relative performance measure and is widely used in literature. For example, Xu et al. (2022b), Chen et al. (2022), and Niu et al. (2023) have used this index as a performance assessment criterion. The mathematical formulation of MAE is presented below.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \rho_i| \tag{11}$$

In the above equation,  $n$  is the sample size.  $Y_i$  is the observed data and  $\rho_i$  is the estimated data, whereas, the Pearson correlation coefficient is a statistical tool that is used to determine the linear relationships between variables (Rungskunroch et al. 2022). Several researchers have used correlation coefficients in their studies. For example, Varney et al. (2022) used correlation to compare the performance of the proposed index. Its value ranges from  $(-1, 1)$ . The correlation ( $r$ ) between the observed data ( $Y$ ) and the estimated data ( $\rho$ ) is presented in Equation (12).

$$r = \frac{Cov(Y, \rho)}{\sigma_Y \sigma_\rho} \tag{12}$$

In the above equation,  $\sigma_Y$  is the standard deviation of  $Y$  and  $\sigma_\rho$  is the standard deviation of  $\rho$ .

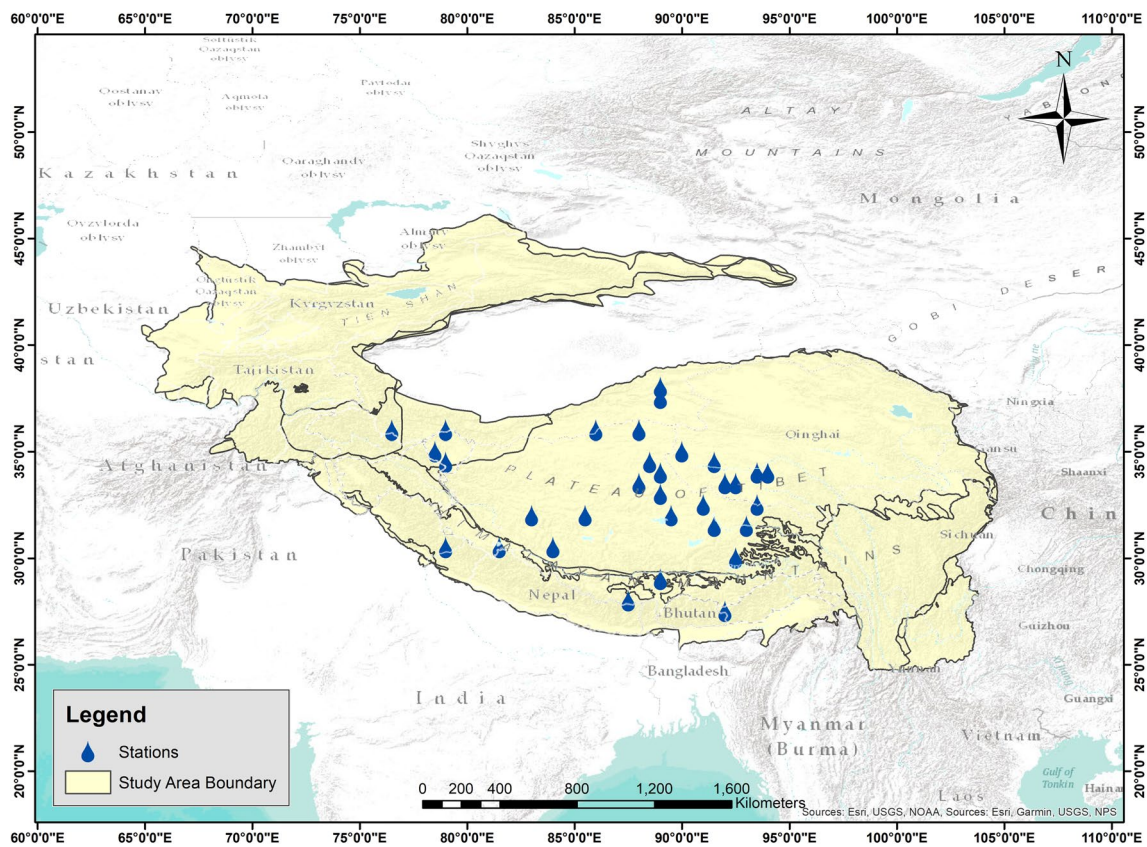
## 3 Application

Tibetan is the region of China that has covered more than 2.5 million  $\text{km}^2$  of the world (Chen et al. 2022a). The Tibet Plateau (or Himalayan Plateau) is located in central Asia, with an average elevation of almost 4000 m (Zhang et al. 2022). It is also known as the ‘‘Roof of the World’’ because it is the source of many Asian rivers. The ecology and climate of Asian countries mostly depend upon the Tibet plateau (Wang et al. 2022). The boundary of the Tibet plateau touches the southwest Himalayas and northeast Kunlun and Aljin mountains (Chen et al. 2022). Its border is identified as being above the 2500-m contour line (Zhang et al. 2022). In our study, we considered the monthly precipitation data of 32 randomly selected stations on the Tibetan plateau (Fig. 3). Gridded CN05.1 observational data set of precipitation on a resolution of  $0.5^\circ \times 0.5^\circ$  is considered as the observed data. Therefore, simulated model data is also re-gridded to a standard resolution of  $0.5^\circ \times 0.5^\circ$ . Moreover, the unit of data is a millimeter per month. However, the lack of precipitation and rising temperatures in this region were the reasons for choosing it. These extreme events can signal drought or flood Li et al. 2021b. We used historical precipitation time series data from the Tibet Plateau region of China from 1961 to 2014. Moreover, we used 20 models of CMIP6 for future prediction. Table 1 is describing the model’s name, modeling center, and resolution of each selected GCM model.

## 4 Results and discussion

### 4.1 Implication of MCFSAWS-Ensemble

This section presents the results associated with the execution of MCFSAWS-Ensemble. In this paper, we provide a numerical and graphical description of the RMCF



**Fig. 3** Spatial distribution of the selected locations

execution for one random grid point. The remaining results are archived in the authors' gallery.

Table 2 provides the cutoff value and size of variables associated with different algorithms based on the importance and non-importance scores of each variable under RMCFs. In this research, we employed permutation methods as a cutoff method. Using this method, we observed that all the GCMs are considered important. Table 3 presents the ranking of various GCMs based on their RI scores under the RMCFs algorithm. The RI scores indicate the importance or performance of each model in simulating climate conditions compared to observed data. The models are listed in descending order, with CanESM5 achieving the highest RI score of 0.122969, followed by CanESM5. CanOE at 0.102068, and ACCESS.ESM1.5 at 0.085126. These top-ranked models are considered to have higher fidelity in replicating the observed climate patterns. As we move down the ranking, the RI scores decrease, indicating relatively lower performance in simulating climate conditions. The models at the bottom of the ranking, such as GFDL.ESM4, INM.CM5.0, HadGEM3.GC31.LL, INM.CM4.8, and EC.Earth3.Veg, have lower RI scores ranging from 0.002701 to 0.001403. The ranking of the GCMs based on their RI scores is significant for model ensemble

construction. Models with higher RI scores are generally more reliable and accurate in capturing the observed climate behavior. Therefore, when forming a model ensemble, the higher-ranked models would typically be given more weight or importance due to their superior performance, while the lower-ranked models may be assigned lesser weight or excluded from the ensemble altogether.

Figure 4 displays the relative importance of each GCM, with the horizontal axis representing the model's RI value ranging from 0 to 1 and the vertical axis showing the model names. The CanESM5 model stands out with the largest RI bar compared to the other models. In Fig. 5, the Interdependency Discover (ID) of each GCM is graphically presented, with the size and color indicating the strength of interdependency. The vibrant color of the model point signifies high dependence and importance for other models, while lighter colors represent lower dependence, and the larger arrow size represents a higher correlation between models.

After assessing the Relative Importance (RI), we calculated the point-to-point differences and transformed them according to the method described in source 2 of the MCFSAWS-Ensemble scheme. Subsequently, we hybridized these two sources by combining their weighted and standardized values.

**Table 1** Description of CMIP6 models

Sr. no	Model name	Modeling center	Resolution (longitude×latitude)
1	ACCESS-CM2	Commonwealth Scientific and Industrial Research Organization, Australia	1.875°×1.25°
2	ACCESS-ESM1-5	Commonwealth Scientific and Industrial Research Organization, Australia	1.875°×1.2143°
3	AWI-CM-1-1-MR	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany	0.9375°×0.9375°
4	BCC-CSM2-MR	Beijing Climate Center and China Meteorological Administration, China	1.125°×1.125°
5	CanESM5	Canadian Centre for Climate Modeling and Analysis, Canada	2.8125°×2.8125°
6	CanESM5-CanOE	Canadian Centre for Climate Modeling and Analysis, Canada	2.8125°×2.8125°
7	CNRM-CM6-1	National Centre for Meteorological Research and European Centre for Research and Advanced Training in Scientific Computation, France	1.40625°×1.40625°
8	EC-Earth3-Veg	EC-Earth consortium, Europe	0.703125°×0.703125°
9	FGOALS-g3	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences, China	2°×2.25°
10	GFDL-ESM4	National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory, United States	1.25°×1°
11	HadGEM3-GC31-LL	Met Office Hadley Centre, UK	1.875°×1.25°
12	INM-CM4-8	Institute for Numerical Mathematics, Russia	2°×1.5°
13	INM-CM5-0	Institute for Numerical Mathematics, Russia	2°×1.5°
14	IPSL-CM6A-LR	Institute Pierre Simon Laplace, France	2.5°×1.25874°
15	MIROC6	University of Tokyo, National Institute for Environmental Studies and Japan Agency for Marine-Earth Science and Technology, Japan	1.40625°×1.40625°
16	MIROC-ES2L	University of Tokyo, National Institute for Environmental Studies and Japan Agency for Marine-Earth Science and Technology, Japan	2.8125°×2.8125°
17	MPI-ESM1-2-HR	Max Planck Institute for Meteorology, Germany	0.9375°×0.9375°
18	MPI-ESM1-2-LR	Max Planck Institute for Meteorology, Germany	1.875°×1.875°
19	MRI-ESM2-0	Meteorological Research Institute, Japan	1.125°×1.125°
20	NESM3	Nanjing University of Information Science and Technology, China	1.875°×1.875°

**Table 2** Cutoff value based on its importance and non-importance scores under RMCFs algorithm

Method	Mini RI	Size	Mini ID
Critical Angle	0.03988	8	Na
K-means	0.085126	3	Na
Permutations	6.00e-05	20	0.39134
Mean	0.013368	10	Na

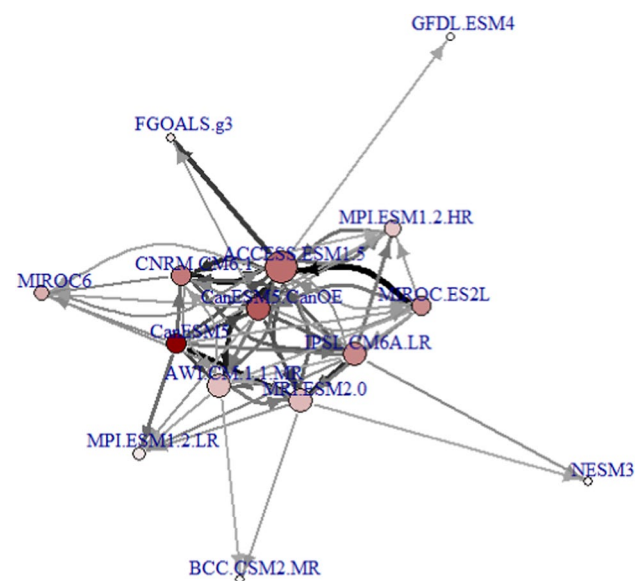
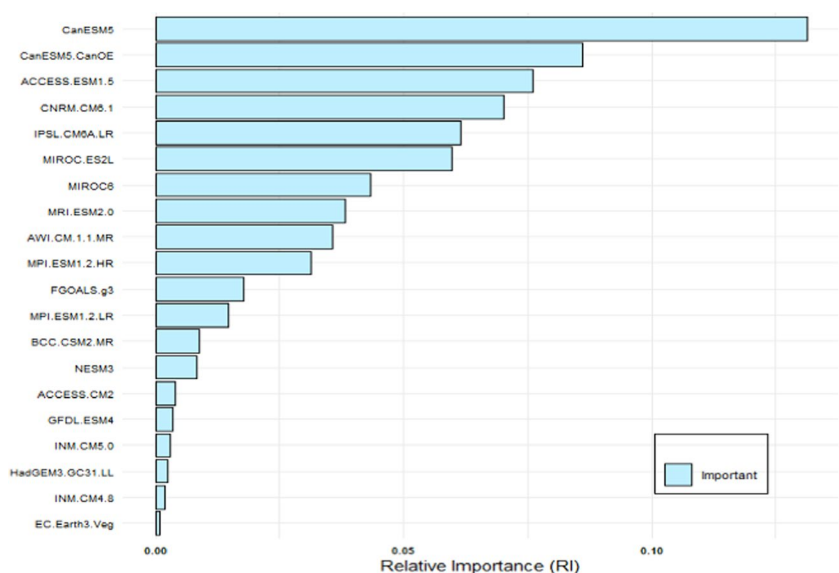
The proposed weights for a single spatial location of time series data are depicted in Fig. 6. The horizontal axis represents the time period from 1961 to 2014, while the vertical axis displays the values of the proposed weighting scheme. Different colors are used to represent various models. It was observed that the weights of each model varied over time, indicating that none of the models remained consistently important. This outcome aligns with the desired results since as time and parameters change; the importance of models should also change in the aggregation process. Figure 7 presents significant deviations in the temporal behavior of aggregated data between the MCFSAWS-Ensemble and SMA approaches.

The next subsection evaluates the validity of the proposed weighting scheme. We utilize Mean Absolute Error (MAE) and correlation measures to assess the superiority of

**Table 3** Relative importance (RI) of various GCMs simulation under RMCFs algorithm

Position	Models	RI
1	CanESM5	0.122969
2	CanESM5.CanOE	0.102068
3	ACCESS.ESM1.5	0.085126
4	CNRM.CM6.1	0.059593
5	IPSL.CM6A.LR	0.053427
6	MIROC.ES2L	0.050333
7	MIROC6	0.042260
8	MRI.ESM2.0	0.039880
9	AWI.CM.1.1.MR	0.036514
10	MPI.ESM1.2.HR	0.034515
11	FGOALS.g3	0.013368
12	MPI.ESM1.2.LR	0.011371
13	BCC.CSM2.MR	0.011167
14	NESM3	0.006307
15	ACCESS.CM2	0.003912
16	GFDL.ESM4	0.003021
17	INM.CM5.0	0.002809
18	HadGEM3.GC31.LL	0.002701
19	INM.CM4.8	0.002380
20	EC.Earth3.Veg	0.001403

**Fig. 4** Relative importance (RI) of various climate simulations of GCMs at one random point



**Fig. 5** Interdependency discovery plot among GCMs model under MCFS

MCFSAWS-Ensemble over SAM approaches. These results are based on all selected grid points.

**4.2 Validation**

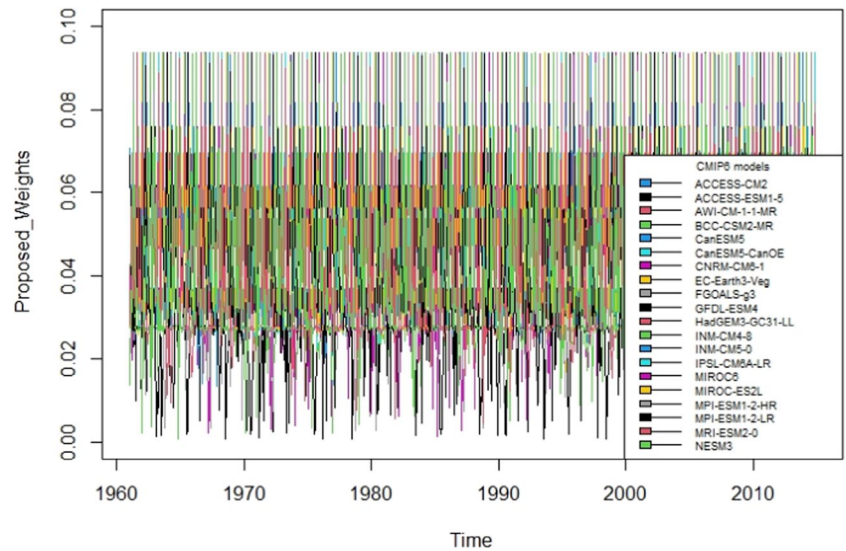
This section presents results associated with the validation of the proposed weighting scheme. Table 4 presents a comparison of the MAE and correlation values for the proposed MCFSAWS-Ensemble scheme and the SMA scheme. These statistics infer the superiority of the proposed procedure over the SMA scheme, particularly when considering unequal

weights. The MAE values indicate the average magnitude of the differences between the simulated precipitation and the observed values. Under the MCFSAWS-Ensemble scheme, the minimum MAE is 0.5037328, while the SMA scheme has a slightly higher minimum MAE of 0.5180123. This suggests that the proposed weighting scheme performs marginally better in terms of minimizing errors at the lowest level. Moving to the quartiles, the first quartile MAE for the MCFSAWS-Ensemble scheme is 1.0350234, while the SMA scheme exhibits a slightly higher first quartile MAE of 1.1211800. This trend continues for the median and mean MAE values, where the MCFSAWS-Ensemble scheme demonstrates lower values (1.4950217 and 1.6426240, respectively) compared to the SMA scheme (1.3675169 and 1.6925493, respectively). These results suggest that the proposed weighting scheme generally outperforms the SMA scheme in terms of achieving lower MAE values. Regarding correlation, higher values indicate a stronger positive relationship between the simulated precipitation and the observed values. The MCFSAWS-Ensemble scheme shows a minimum correlation of  $-0.54468$ , while the SMA scheme has a slightly better minimum correlation of  $-0.462500$ . However, when considering the quartiles, median, and mean correlation values, the MCFSAWS-Ensemble scheme consistently exhibits higher values (ranging from 0.608719 to 0.813871) compared to the SMA scheme (ranging from 0.584670 to 0.742601). This suggests that the proposed weighting scheme generally provides a more favorable correlation with the observed precipitation data.

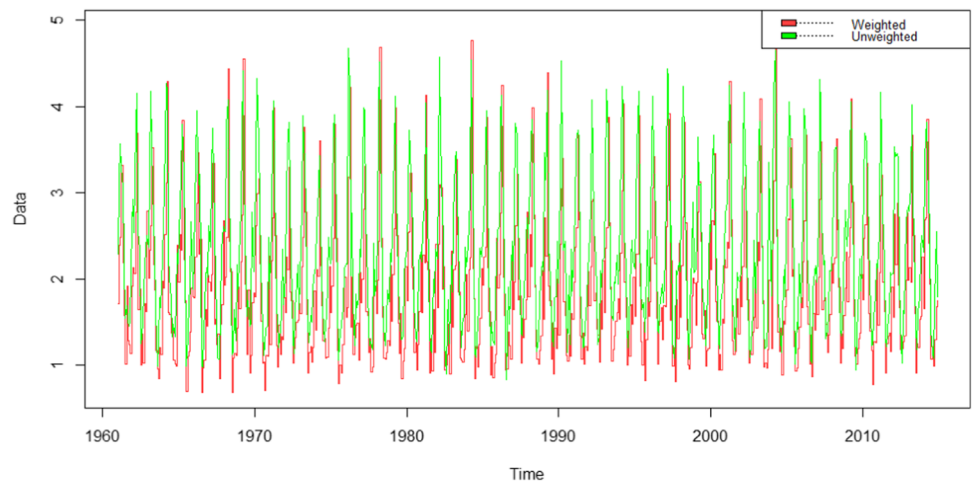
In summary, the table indicates that the MCFSAWS-Ensemble scheme, with its unequal weighting approach, performs better than the SMA scheme in terms of achieving lower MAE values and higher correlations. These findings highlight the efficiency and superiority of the proposed



**Fig. 6** Weights of various GCMs model at one random grid point



**Fig. 7** Temporal behavior of ensemble data under MCF-SAWS-Ensemble and SMA scheme at one random grid point



**Table 4** Overall description of comparative statistics of MAE and correlation under proposed MCFSAWS-Ensemble and SMA Schemes in all the grid points

Statistics	MCFSAWS-Ensemble scheme		SMA scheme	
	MAE	Correlation	MAE	Correlation
Min.	0.5037328	-0.54468	0.5180123	-0.462500
1st Qu.	1.0350234	0.608719	1.1211800	0.584670
Median	1.4950217	0.750442	1.3675169	0.680660
Mean	1.6426240	0.647631	1.6925493	0.598336
3rd Qu.	1.8444847	0.813871	1.8990142	0.742601
Max.	5.1499249	0.905341	6.3541390	0.851145

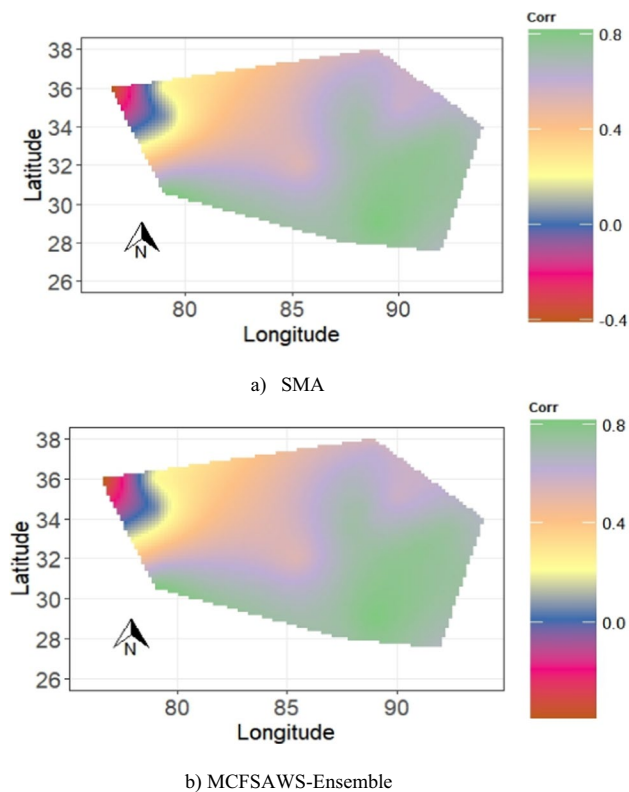
procedure for climate simulation of precipitation in a multi-model ensemble of GCMs.

Figure 7 shows the temporal behavior of ensemble data indicating the consistency of two schemes at a specific grid point, while Fig. 8 illustrates the spatial distribution of

correlation coefficients between the data from the MCFSAWS-Ensemble scheme and the SMA scheme with observed data. We observe that the proposed weighting scheme is consistent with SMA and there is no contradiction in any single grid points. This inference supports and enhances the understanding of the comparative analysis provided in Table 4.

### 5 Conclusion

Continuous monitoring and projection of climate change are compulsory to care about future global health. Therefore, prioritizing the monitoring and projection of climate change is crucial to promoting global health and ensuring a sustainable future for all. Multi-model ensemble approach for GCMs is important for climate change assessments, as it can help to account for uncertainties and provide a more comprehensive understanding of the potential impacts of climate change. In this paper, we proposed a new weighting scheme



**Fig. 8** Spatial distribution of correlation coefficient of data MCFSAWS-Ensemble and SMA scheme with observed data

— MCFSAWS-Ensemble: Monte Carlo Feature Selection Adaptive Weighting Scheme for GCMs Ensemble. Unlike, the SAM scheme, the proposed weighting scheme has the potential to reduce the effects of extreme values. For application, we used the simulated time series data of precipitation data from 20 GCMs of CMIP phase 6 at the Tibetan Plateau. Values of quality measures indicate that the proposed unequal weighting scheme (MCFSAWS-Ensemble) performed better as compared to the SMA approach. MCFSAWS-Ensemble successfully reduced the effect of extreme values in the time of precipitation data of the Tibetan Plateau. The suggested weights under the proposed scheme can be cogitated to combine CMIP6 data of future scenarios. In summation, the proposed weighting scheme can help to aggregate multiple, which can improve our understanding of the climate system and extreme events like drought and flood. In future research, the same research framework can be extended to aggregate other important variables like temperature, humidity, evaporation, etc. These findings contribute in several ways to our understanding of multi-model ensembles and provide a basis for accurate assessment of climate change and its impact. The potential limitation of the study is that we only used precipitation data. In future research, other meteorological parameter such as temperature, humidity, and wind speed can be incorporated to enhance the accuracy of drought assessment.

**Author contribution** Abdul Baseer and Zulfiqar Ali conceived the presented idea. Abdul Baseer developed the theory and performed the computations. Maryam Ilyas verified the analytical methods and computations. Mahrukh Yousaf revised the manuscript and addressed all the technical questions raised by the reviewers. All authors discussed the results and contributed to the final manuscript.

**Funding** The current research is a part of a funded research project awarded by the University of the Punjab Lahore, Pakistan (2022). Therefore, the authors are thankful to the project awarding institution.

**Data availability** The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

- Alfonso S, Gesto M, Sadoul B (2021) Temperature increase and its effects on fish stress physiology in the context of global warming. *J Fish Biol* 98(6):1496–1508
- Alhakami W, ALharbi A, Bourouis S, Alroobaea R, Bouguila N (2019) Network anomaly intrusion detection using a nonparametric Bayesian approach and feature selection. *IEEE Access* 7:52181–52190
- Alirezanejad M, Enayatifar R, Motameni H, Nematzadeh H (2020) Heuristic filter feature selection methods for medical datasets. *Genomics* 112(2):1173–1181
- Baldos UL, Hertel TW, Moore FC (2019) Understanding the spatial distribution of welfare impacts of global warming on agriculture and its drivers. *Am J Agric Econ* 101(5):1455–1472
- Çakmak EE, Acar S (2022) The nexus between economic growth, renewable energy and ecological footprint: an empirical evidence from most oil-producing countries. *J Clean Prod* 352:131548
- Çakmak S, Demir T, Canpolat E, Aytaç AS (2021) Evaluation of the effects of precipitation and flow characteristics on suspended sediment transport in mountain-type Mediterranean climate; Korkuteli Stream sample, Antalya, Turkey. *Arab J Geosci* 14(19):1–17
- Chen H, Li X, Wu Y, Zuo L, Lu M, Zhou Y (2022a) Compressive strength prediction of high-strength concrete using long short-term memory and machine learning algorithms. *Buildings* 12(3):302
- Chen R, Duan K, Shang W, Shi P, Meng Y, Zhang Z (2022) Increase in seasonal precipitation over the Tibetan Plateau in the 21st century projected using CMIP6 models. *Atmos Res* 277:106306
- Davies-Barnard T, Zaehle S, Friedlingstein P (2022) Assessment of the impacts of biological nitrogen fixation structural uncertainty in CMIP6 earth system models. *Biogeosciences* 19(14):3491–3503
- Dey A, Sahoo DP, Kumar R, Remesan R (2022) A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin. *Int J Climatol* 42(16):9215–9236
- Dong F, Javed A, Saber A, Neumann A, Arnillas CA, Kaltenecker G, Arhonditsis G (2021) A flow-weighted ensemble strategy to assess the impacts of climate change on watershed hydrology. *J Hydrol* 594:125898
- Dou X, Ma X, Zhao C, Li J, Yan Y, Zhu J (2022) Risk assessment of soil erosion in Central Asia under global warming. *Catena* 212:106056

- Dramiński M, Kierczak M, Koronacki J, Komorowski J (2010) Monte Carlo feature selection and interdependency discovery in supervised classification. *Advances in Machine Learning II: Dedicated to the Memory of Professor Ryszard S. Michalski* 2:371–385
- Dramiński M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J (2008) Monte Carlo feature selection for supervised classification. *Bioinformatics* 24(1):110–117
- Duan Q, Duan A (2020) The energy and water cycles under climate change. *Natl Sci Rev* 7(3):553–557
- Ehteram M, Kalantari Z, Ferreira CS, Chau KW, Emami SMK (2022) Prediction of future groundwater levels under representative concentration pathway scenarios using an inclusive multiple model coupled with artificial neural networks. *J Water Clim Change* 13(10):3620–3643
- Hamed MM, Nashwan MS, Shahid S (2022) A novel selection method of CMIP6 GCMs for robust climate projection. *Int J Climatol* 42(8):4258–4272
- Hasan N, Bao Y (2021) Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol* 11:49–62
- Jose DM, Vincent AM, Dwarakish GS (2022) Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. *Sci Rep* 12(1):1–25
- Kim YH, Min SK, Zhang X, Sillmann J, Sandstad M (2020) Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather Clim Extrem* 29:100269
- Knutti R, Baumberger C, Adorn GH (2019) Uncertainty quantification using multiple models—prospects and challenges. In: Beisbart C, Saam N (eds) *Computer simulation validation. Simulation foundations, methods and applications*. Springer. [https://doi.org/10.1007/978-3-319-70766-2\\_34](https://doi.org/10.1007/978-3-319-70766-2_34)
- Lacetera N (2019) Impact of climate change on animal health and welfare. *Anim Front* 9(1):26–31
- Li J, Lu L, Zhang YH, Xu Y, Liu M, Feng K et al (2020) Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther* 27(1):56–69
- Li L, Zhu C, Zhang R, Liu B (2021a) Roles of the Tibetan Plateau vortices in the record Meiyu rainfall in 2020. *Atmos Sci Lett* 22(3):e1017
- Li Y, Stroe DI, Cheng Y, Sheng H, Sui X, Teodorescu R (2021b) On the feature selection for battery state of health estimation based on charging–discharging profiles. *J Energy Storage* 33:102122
- Liu F, Xu C, Long Y, Yin G, Wang H (2022) Assessment of CMIP6 model performance for air temperature in the arid region of north-west China and subregions. *Atmosphere* 13(3):454
- Liu H, Wang H, Li N, Shao J, Zhou X, van Groenigen KJ, Thakur MP (2022a) Phenological mismatches between above-and below-ground plant responses to climate warming. *Nat Clim Change* 12(1):97–102
- Lovenduski NS, McKinley GA, Fay AR, Lindsay K, Long MC (2016) Partitioning uncertainty in ocean carbon uptake projections: Internal variability, emission scenario, and model structure. *Glob Biogeochem Cycles* 30(9):1276–1287
- Mare F, Bahta YT, Van Niekerk W (2018) The impact of drought on commercial livestock farmers in South Africa. *Dev Pract* 28(7):884–898
- Mokhov II (2022) Climate change: causes, risks, consequences, and problems of adaptation and regulation. *Her Russ Acad Sci* 92(1):1–11
- Morim J, Trenham C, Hemer M, Wang XL, Mori N, Casas-Prat M et al (2020) A global ensemble of ocean wave climate projections from CMIP5-driven models. *Sci Data* 7(1):105
- Mudryk L, Santolaria-Otín M, Krinner G, Ménégos M, Derksen C, Brutel-Vuilmet C et al (2020) Historical Northern Hemisphere snow cover trends and projected changes in the CMIP6 multi-model ensemble. *Cryosphere* 14(7):2495–2514
- Nematzadeh H, Enayatifar R, Mahmud M, Akbari E (2019) Frequency based feature selection method using whale algorithm. *Genomics* 111(6):1946–1955
- Niaz R, Almanjahie IM, Ali Z, Faisal M, Hussain I (2020) A novel framework for selecting informative meteorological stations using Monte Carlo Feature Selection (MCFS) algorithm. *Adv Meteorol* 2020:1–13
- Niu C, Jian S, Liu S, Liu C, Hu C (2023) Comparative study of reference evapotranspiration estimation models based on machine learning algorithm: a case study of Zhengzhou City. *Hydrol Res* 54(8):945–964
- Oh SG, Sushama L, Teufel B (2020) Arctic precipitation and surface wind speed associated with cyclones in a changing climate. *Clim Dyn* 55(11):3067–3085
- Ombadi M, Nguyen P, Sorooshian S, Hsu KL (2021) Retrospective analysis and Bayesian model averaging of CMIP6 precipitation in the Nile River Basin. *J Hydrometeorol* 22(1):217–229
- Otterbring T, Samuelsson P, Arsenovic J, Elbæk CT, Folwarczny M (2022) Shortsighted sales or long-lasting loyalty? The impact of salesperson-customer proximity on consumer responses and the beauty of bodily boundaries. *Eur J Mark* 57(7):1854–1885
- Perović V, Kadović R, Đurđević V, Pavlović D, Pavlović M, Čakmak D et al (2021) Major drivers of land degradation risk in Western Serbia: current trends and future scenarios. *Ecol Indic* 123:107377
- Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon Weather Rev* 133(5):1155–1174
- Rajak J (2021) A preliminary review on impact of climate change and our environment with reference to global warming. *Int J Environ Sci* 10:11–14
- Raju KS, Kumar DN (2020) Review of approaches for selection and ensembling of GCMs. *J Water Clim Change* 11(3):577–599
- Rungskunroch P, Shen ZJ, Kaewunruen S (2022) Benchmarking socio-economic impacts of high-speed rail networks using K-nearest neighbour and Pearson's correlation coefficient techniques through computational model-based analysis. *Appl Sci* 12(3):1520
- Russo MA, Carvalho D, Martins N, Monteiro A (2022) Forecasting the inevitable: a review on the impacts of climate change on renewable energy resources. *Sustain Energy Technol Assess* 52:102283
- Sanderson BM, Knutti R, Caldwell P (2015) Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J Clim* 28(13):5150–5170
- Scafetta N (2022) CMIP6 GCM ensemble members versus global surface temperatures. *Clim Dyn* 1-30:3091–3120. <https://doi.org/10.1007/s00382-022-06493-w>
- Seifi A, Ehteram M, Soroush F, Haghighi AT (2022) Multi-model ensemble prediction of pan evaporation based on the Copula Bayesian Model Averaging approach. *Engineering Applications of Artificial Intelligence* 114:105124
- Smith DM, Scaife AA, Hawkins E, Bilbao R, Boer GJ, Caian M et al (2018) Predicted chance that global warming will temporarily exceed 1.5 C. *Geophys Res Lett* 45(21):11–895
- Sperry JS, Venturas MD, Todd HN, Trugman AT, Anderegg WR, Wang Y, Tai X (2019) The impact of rising CO<sub>2</sub> and acclimation on the response of US forests to global warming. *Proc Natl Acad Sci* 116(51):25734–25744
- Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A (2019) Feature selection methods and genomic big data: a systematic review. *J Big Data* 6(1):1–24
- Try S, Tanaka S, Tanaka K, Sayama T, Khujanazarov T, Oeurng C (2022) Comparison of CMIP5 and CMIP6 GCM performance for flood projections in the Mekong River Basin. *J Hydrol Reg Stud* 40:101035

- Varney RM, Chadburn SE, Burke EJ, Cox PM (2022) Evaluation of soil carbon simulation in CMIP6 Earth system models. *Biogeosciences* 19(19):4671–4704
- Wang W, Zhang P, Garziona CN, Liu C, Zhang Z, Pang J et al (2022) Pulsed rise and growth of the Tibetan Plateau to its northern margin since ca. 30 Ma. *Proc Natl Acad Sci* 119(8):e2120364119
- Wang Z, Afgan MS, Gu W, Song Y, Wang Y, Hou Z et al (2021) Recent advances in laser-induced breakdown spectroscopy quantification: from fundamental understanding to data processing. *TrAC Trends Anal Chem* 143:116385
- Wei Q, Xu J, Liao L, Yu Y, Liu W, Zhou J, Ding Y (2021) Indicators for evaluating trends of air humidification in arid regions under circumstance of climate change: relative humidity (RH) vs. Actual water vapour pressure (ea). *Ecol Ind* 121:107043
- Wine ML, Davison JH (2019) Untangling global change impacts on hydrological processes: resisting climatization. *Hydrol Process* 33(15):2148–2155
- Wooten AM, Massoud EC, Sengupta A, Waliser DE, Lee H (2020) The effect of statistical downscaling on the weighting of multi-model ensembles of precipitation. *Climate* 8(12):138
- Wu J, Shi Y, Xu Y (2020) Evaluation and projection of surface wind speed over China based on CMIP6 GCMs. *J Geophys Res Atmos* 125(22):e2020JD033611
- Xu D, Ivanov VY, Kim J, Fatichi S (2019) On the use of observations in assessment of multi-model climate ensemble. *Stoch Environ Res Risk Assess* 33(11–12):1923–1937
- Xu J, Zhang X, Zhang W, Hou N, Feng C, Yang S et al (2022a) Assessment of surface downward longwave radiation in CMIP6 with comparison to observations and CMIP5. *Atmos Res* 270:106056
- Xu Y, Zhang H, Yang F, Tong L, Yan D, Yang Y et al (2022b) State of charge estimation of supercapacitors based on multi-innovation unscented Kalman filter under a wide temperature range. *Int J Energy Res* 46(12):16716–16735
- Yan C, Liang J, Zhao M, Zhang X, Zhang T, Li H (2019) A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. *Anal Chim Acta* 1080:35–42
- Yasmin N, Jamuda M, Panda AK, Samal K, Nayak JK (2022) Emission of greenhouse gases (GHGs) during composting and vermicomposting: measurement, mitigation, and perspectives. *Energy Nexus* 7:100092
- Yue Y, Yan D, Yue Q, Ji G, Wang Z (2021) Future changes in precipitation and temperature over the Yangtze River Basin in China based on CMIP6 GCMs. *Atmos Res* 264:105828
- Zeng W, Jin S, Liu W, Qian C, Luo P, Ouyang W, Wang X (2022) Not all tokens are equal: human-centric visual analysis via token clustering transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 11101–11111
- Zhang X, Hua L, Jiang D (2022) Assessment of CMIP6 model performance for temperature and precipitation in Xinjiang, China. *Atmos Ocean Sci Lett* 15(2):100128
- Zhang X, Xiong Z, Zhang X, Shi Y, Liu J, Shao Q, Yan X (2016) Using multi-model ensembles to improve the simulated effects of land use/cover change on temperature: a case study over northeast China. *Clim Dyn* 46:765–778
- Zhang Y, Yang X, Chen C (2021) Substantial decrease in concurrent meteorological droughts and consecutive cold events in Huai River Basin, China. *Int J Climatol* 41(13):6065–6083
- Zheng Z, Zhao L, Oleson KW (2021) Large model structural uncertainty in global projections of urban heat waves. *Nat Commun* 12(1):3736
- Zhou H, Wang X, Zhu R (2022) Feature selection based on mutual information with trend and uncertainty. *Clim Dyn* 1–21. <https://doi.org/10.1007/s00382-022-06518-4>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.