RESEARCH PAPER

WHICH CLIMATE MODEL EVALUATION METHODS CAN CONSISTENTLY SELECT SKILLFUL MODELS FROM THE CMIP6 ENSEMBLE?

Natalia V. Gnatiuk[1*], Iuliia V. Radchenko[1], Richard Davy[2], Jiechen Zhao[3,4], Leonid P. Bobylev[1]

[1]Nansen International Environmental and Remote Sensing Centre, 14-Liniya V.O. 7, St. Petersburg, 199034, Russia

[2]Nansen Environmental and Remote Sensing Center, Jahnebakken 3, Bergen, 5006, Norway

[3]Qingdao Innovation and Development Base of Harbin Engineering University, Sansha Road 1777, Qingdao, 266000, China

[4]First Institute of Oceanography, MNR & Decade Collaborative Center on Ocean-Climate Nexus and Coordination, Xianxialing Road 6, Qingdao, 266000, China

***Corresponding author:** gnatiuk.n@gmail.com

Received: October 29th 2024 / Accepted: May 15th 2025 / Published: June 30th 2025

https://doi.org/10.24057/2071-9388-2025-3694

ABSTRACT. When considering the possible use of climate model data, it is necessary to choose which model is most appropriate to use. There are many methods for evaluating and selecting climate models in the literature, but there is no established consensus on which method is the most robust for determining model skill. In this article, we tested seven widely used methods for evaluating climate models in the Arctic using CMIP6 surface air temperature data: a single statistical metric method (root mean square error, spatial trends), a single skill score method (Taylor skill score, probability density function), a combination of several statistical metric methods (Taylor diagram, interannual variability skill score, comprehensive rating metric, etc.), and a multiple statistical criteria method (percentile-based approach). To evaluate their consistency, each method was applied to two periods: 1951-1980 and 1981-2010. For each method, the models were ranked and classified into three quality groups (very good, satisfactory, unsatisfactory). The comparison of methods was performed by comparing the differences in the average values of the normalized statistical measures, the differences in the model ranks, and the definition of the model quality groups. For each method, an optimal set of models corresponding to the top 25% was selected. One of the main objectives of the study was to compare the ability of the methods to identify the best model for the selected ensemble, regardless of the time period (i.e., without sensitivity to natural variability). The results suggest a preference for methods using root mean square error and a percentile-based approach.

KEYWORDS: climate model, evaluation method, CMIP6, sub-ensemble, climate model selection

CITATION: Gnatiuk N. V., Radchenko I. V., Davy R., Zhao J., Bobylev L. P. (2025). Which Climate Model Evaluation Methods Can Consistently Select Skillful Models From The Cmip6 Ensemble? Geography, Environment, Sustainability, 2 (18), 126-149
https://doi.org/10.24057/2071-9388-2025-3694

ACKNOWLEDGEMENTS: The study was funded by the Russian Science Foundation (RSF) grant No. 23-77-01106, https://rscf.ru/en/project/23-77-01106/.

Conflict of interests: The authors reported no potential conflict of interests.

INTRODUCTION

At present, climate models are the most valuable tool in projecting future climate under different scenarios (Taylor et al. 2012; Stocker et al. 2014; Otero et al. 2018). Over the years, global climate models (GCMs) have been continuously developed by different modeling centers, incorporating diverse parameterizations. Consequently, future climate projections from these models can vary significantly, leading to substantial uncertainty in projected changes of climate variables (Knutti et al. 2010; Stocker et al. 2014). In addition to this considerable projection uncertainty, individual models often have significant biases. Therefore, a multi-model ensemble of GCM simulations is commonly employed in research, as it has been shown that ensemble means tend to cancel out individual model biases - i.e., the ensemble mean of a large group of models generally outperforms any single model in most cases (Gleckler et al. 2008; Knutti et al. 2010; Raju and Kumar 2020).

However, in the IPCC's Sixth Assessment Report (AR6), climate models were assigned weightings when assessing future projections for the first time. This was due to the "hot model" problem: a subset of CMIP6 models exhibited climate sensitivities outside the range estimated from multiple lines of evidence (Sherwood et al. 2020). This issue has prompted efforts within the community to establish robust model selection criteria (Hausfather et al. 2022). Research indicates that selecting skillful GCMs can reduce uncertainty in ensemble projections compared to using the full set of dozens of models (Herger et al. 2018; Gnatiuk et al. 2020). As a result, various model selection criteria and

evaluation methods have been introduced in the literature. A categorization mind map of these methods is provided in Appendix A.2.

One approach to reducing uncertainty is to select only those models that perform well in historical simulations when compared to observations. The primary assumption underpinning this approach is that model skill in historical simulations is a reliable predictor of model performance in future climate projections; that is, models that are skillful in historical simulations are also likely to be skillful in their response to forcing. Numerous studies have demonstrated that the response to forcing is sensitive to baseline climatology, which supports this assumption (Caballero and Huber 2013). However, if there is a substantial change in the relative importance of different processes shaping regional climate, then this assumption – that historical skill predicts future model performance – may not hold. Additionally, given the considerable internal variability within models, it is essential that the evaluation of model skill be conducted over a long period (several decades) to ensure a fair assessment of model performance (Jain et al. 2023). Furthermore, it is crucial that the criteria for model selection are not overly sensitive to the phase of natural variability within the climate system.

There is no consensus among researchers on best practice for climate model evaluation and selection (Knutti et al. 2010; Ahmadalipour et al. 2017; Herger et al. 2018; Calvin et al. 2023). This lack of agreement has resulted in a diverse range of approaches; for example, some studies employ only a single statistical metric for GCM evaluation (e.g., Walsh et al. 2008; Macadam et al. 2010; Sillmann et al. 2013; Agosta et al. 2015), while others utilize multiple statistical metrics (e.g., McMahon et al. 2015; Aghakhani Afshar et al. 2017; Ruan et al. 2019). For instance, the near-surface air temperature simulations from 17 GCMs were analyzed using just one statistical metric – root mean square error (RMSE) – and models were ranked from the lowest to the highest RMSE values (Reifen and Toumi 2009; Macadam et al. 2010). Herger et al. (2018) selected an optimal subset of 38 GCMs for surface air temperature and precipitation based on RMSE. RMSE was also employed for inter-model comparison and evaluation by Sillmann et al. (2013) and Zhou et al. (2014). Other statistical metrics have been used to assess GCM accuracy as well – for example, Kumar et al. (2013) evaluated 19 GCMs based on trends in temperature and precipitation across continental areas. Maxino et al. (2008) and Perkins et al. (2007) proposed a skill score that measures the common area between the probability density functions (PDFs) of modeled and observed data.

Many studies evaluating the accuracy of GCMs utilize the Taylor diagram, which combines three statistical criteria – standard deviation (STD), RMSE, and correlation coefficient (r) (Taylor 2001). This diagram is summarized into a single metric – the Taylor skill score (Taylor 2001; Inoue and Ueda 2011; Ogata et al. 2014; Sharmila et al. 2015; Kadel et al. 2018; Yang et al. 2020). It should be noted that even when employing the same evaluation methods, researchers often apply different thresholds for sub-ensemble selection, as there is frequently no clear criterion defining a model as "good" or "bad" within these evaluation frameworks. For example, the Taylor skill score has been utilized by Sharmila et al. (2015), Kadel et al. (2018), and Yang et al. (2020), but with varying thresholds.

Some studies evaluating GCMs have employed multiple statistical metrics. For example, McMahon et al. (2015) assessed 23 GCMs using RMSE, the Nash-Sutcliffe Efficiency coefficient (NSE), and the coefficient of determination (r^2) for temperature and precipitation patterns. Kumar et al. (2015) considered bias, trend analysis, and Taylor diagrams to evaluate simulations of extreme winds from 15 GCMs across 22 regions. Aghakhani Afshar et al. (2017) evaluated 14 GCMs for precipitation using four statistical criteria: r^2, NSE, percent of bias (PBIAS), and the ratio of root mean square error to the standard deviation of measured data (CPI). Furthermore, Aghakhani Afshar et al. (2017) categorized the statistical metrics into four groups – very good, good, satisfactory, and unsatisfactory – using a threshold criterion, ultimately selecting models with scores between 75% and 100%, which were ranked as the very good group. Jiang et al. (2015) evaluated 31 GCMs for total precipitation and three indices (the fraction of total rainfall from events exceeding the long-term 95th percentile, precipitation intensity, and maximum consecutive dry days) over China, employing a Taylor diagram and the Interannual Variability Skill Score (IVS; Chen et al. 2011). They further ranked the models using a Comprehensive Rating Metric. You et al. (2018) applied a similar analysis and model selection process as Jiang et al. (2015), but additionally analyzed trends and IVS for both the sub-ensemble and full ensemble across 16 temperature indices. This comprehensive rating metric has been utilized in many studies for model ranking, including those by Jiang et al. (2015), You et al. (2018), Rao et al. (2019), Ahmed et al. (2019, 2020), and Cai et al. (2021).

Other researchers have employed more complex methods involving multiple (up to seven) statistical criteria for evaluating the reliability of GCMs (e.g., Fu et al. 2013; Rupp et al. 2013; Ruan et al. 2019; Jia et al. 2019; Gnatiuk et al. 2020). These studies used different combinations of variables – such as air temperature, precipitation, wind speed, shortwave radiation, and nutrients – and a varying number of GCMs, ranging from 11 to 41. To compare the models, researchers ranked them based on their total scores and selected the relatively best GCMs: for example, 8 out of 41 models – around 20% (Rupp et al. 2013), the top 25% of models (Ruan et al. 2019; Gnatiuk et al. 2020), or the top 30% (Jia et al. 2019). All of these authors suggest the use of a method that incorporates multiple statistical criteria rather than relying on a single metric.

Furthermore, there is no universally accepted method for climate model evaluation and selection. A significant challenge is the trade-off associated with ensemble size: the stricter the filtering of non-skillful models, the smaller the ensemble becomes, which can increase the influence of individual model biases in the ensemble mean projections.

Any model selection criterion should also be robust – that is, it should consistently identify similar models as skillful across different time periods used for model evaluation. This task presents particular difficulties given the large natural variability and internal variability within models. Consequently, there has been a shift toward evaluating models based on processes rather than states (Eyring et al. 2019).

In summary, various methods are employed to evaluate and select appropriate GCMs for specific research questions (Herger et al. 2018; Raju and Kumar 2020; Calvin et al. 2023). Chai and Draxler (2014) recommend using several statistical metrics, including RMSE and mean absolute error (MAE). Raju and Kumar (2020) suggest considering the statistical metrics in a category-wise manner – for example, one metric for error, one for correlation, and one for skill score – and then computing the overall weight of the metrics, for instance, using a rating method. Fu et al. (2013), Ruan et al. (2019), Jia et al. (2019), and Gnatiuk et al. (2020) propose the application of multiple criteria for model evaluation.

Currently, there is no comprehensive comparison of these different methods for evaluating climate models; therefore, this study aims to fill this gap. We tested several widely used approaches, ranging from a single statistical metric to multiple metrics: (i) RMSE, (ii) spatial trends, (iii) TSS, (iv) PDF, (v) Taylor diagram, IVS, MR, (vi) Taylor diagram, MAE, trend, and (vii) a percentile-based method for evaluating CMIP6 surface air temperature (SAT) in the Arctic. To compare these methods with each other, we ranked GCMs according to their performance for each approach and selected the top 25%. Additionally, we conducted the analysis over two different periods – 1951-1980 and 1981-2010 – to assess the consistency of these methods.

MATERIALS AND METHODS

Study area and data

In this paper, we compare model evaluation methods based on surface air temperature in the Arctic (60-90° N). Simulations of historical surface air temperatures from 25 GCMs from the Coupled Model Intercomparison Project Phase 6 (CMIP6) were obtained from the Earth System Grid Federation portal[1]. Information about these models is provided in Table A.1. The list of model names is presented in Fig. 1. The model data were compared to observations from the Berkeley Earth database (Rohde et al. 2013; Rohde and Hausfather 2020). The Berkeley Earth database is a comprehensive global land-ocean temperature record that integrates monthly land temperature data from over 40,000 weather stations with sea surface temperature data from HadSST3 (Hadley Centre Sea Surface Temperature dataset, version 3). Using kriging-based spatial interpolation, it provides extensive spatial coverage for the period spanning from 1850 to the present. It offers average temperatures in 1° x 1° latitude-longitude grid cells for each month.

We selected seven of the most frequently used methods. Table A.3 summarizes the published works employing these methods. Each method is described in detail below:

i) Method of model comparison by ***root mean square error***

Root mean square error (*RMSE*) is a commonly used statistic to quantify differences between two fields of data (Eq. 1):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Tm_i - To_i\right)^2}{n}} \quad (1)$$

where Tm_i is the temperature from the model and To_i is the temperature from observations at time step, *i*, with *n* representing the number of measurements in the time series. The smaller the RMSE, the better the agreement between the two data fields.

ii) Method of model comparison by ***trends***

Analysis of spatial trends using statistics such as the correlation coefficient *(r)*, standard deviation *(STD)*, and mean value ($\overline{T}$) is proposed by Kumar et al. (2013). The correlation coefficient r is calculated as (Yang et al. 2020) (Eq. 2):

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(To_i - \overline{To}\right)\times\left(Tm_i - \overline{Tm}\right)}{STDo \times STDm} \quad (2)$$

where Tm_i is the temperature from a given model and To_i is the temperature from observations at time step *i*, $\overline{Tm}$ is the average temperature of the model, while $\overline{To}$ is the average temperature of the observations, *n* represents the number of observations in the time series, *STDm* is the standard deviation of the model temperatures, and *STDo* is the standard deviation of the observed temperatures.

STD is calculated as (Eq. 3):

$$STD = \sqrt{\frac{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)}{n-1}} \quad (3)$$

where T_i is the temperature at time step *i*, $\overline{T}$ is the mean temperature, and *n* is the number of data points in the time series.

Trend (*Tr*) was calculated using the least squares method (Eq. 4):

$$Tr = \frac{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)\times\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2} \quad (4)$$

where T_i is the temperature at time step *i*, $\overline{T}$ is the average temperature, Y_i is the time step of the time series, is the mean of the time series, *n* is the number of data points in the series.

Each model was ranked based on its ability to represent observations across three metrics: r, the difference between the mean values of models and observations, and the difference between the STD of models and observations.

iii) Method of model comparison by ***Taylor skill score***

The Taylor skill score (TSS) is calculated using r and STD statistical metrics (Taylor 2001) (Eq. 5):

$$TSS = \frac{4\times(1+r)^4}{\left(STDN + \frac{1}{STDN}\right)^2 \times (1+r)^4} \quad (5)$$

where *r* is the correlation coefficient and *STDN* is the ratio of the model's and observations' *STD*. The *TSS* is bounded between 0 and 1, with 1 indicating a perfect fit between the model and observations.

iv) Method of model comparison by S_{score}

The skill score based on PDFs (S_{score}) was proposed by Perkins et al. (2007) as a robust metric for evaluating and ranking climate models because it is less affected by observational errors than the mean value and standard deviation. The S_{score} measures the common area between the PDFs of observed and model data. The formula for S_{score} is expressed as (Eq. 6):

$$S_{score} = \sum_{1}^{n} minimum\left(Z_m, Z_o\right) \quad (6)$$

where *n* is the number of temperature bins, Z_m is the frequency of model data values within the corresponding bin, and Z_o is the frequency of observed data values within that bin. Therefore, the S_{score} ranges between 0 and 1, with a score of 1 indicating a perfect match between the observed and model distributions and a score of 0 indicating no overlap between them. When applying this skill score metric, we used bins that are 1°C wide.

v) Comprehensive rating metric based on Taylor diagram and interannual variability skill score

The Comprehensive Rating Metric (*MR*) was proposed by Jiang et al. (2015) (Eq. 7):

$$MR = 1 - \frac{1}{nm}\sum_{i=1}^{n} rank_i \quad (7)$$

where *n* is the total number of parameters, *m* is the number of models, and rank is the position of a given model based on its performance (with 1 being the best model). The metric ranges between 0 and 1, with higher *MR* values indicating better model skill. The *MR* was calculated for each of the metrics used in the Taylor diagram (*r, STDN, RMSE*), and an average *MR* value was derived for each metric.

Additionally, *MR* was calculated for the Interannual Variability Skill Score (*IVS*; Chen et al. 2011) (Eq. 8):

$$IVS = \left(\frac{STD_m}{STD_o} - \frac{STD_o}{STD_m} \right)^2 \quad (8)$$

where *STDm* is the standard deviation from the model and *STDo* is the standard deviation of the observations. Smaller *IVS* values indicate that the simulated variability more closely matches the observed variability.

vi) Method of model comparison by ***Taylor Diagram, bias and trend***

Kumar et al. (2015) proposed to evaluate models using Taylor diagram, bias, and trend statistics. Bias (B) is calculated as the mean of the differences between the model and observations (Eq. 9):

$$B = \frac{\sum_{i=1}^{n} T_{m_i} - T_{o_i}}{n} \quad (9)$$

where *n* is the number of observations in the time series.

vii) Percentile-based method

A percentile score-based model ranking method introduced by Gnatiuk et al. (2020) includes the analysis of the mean spatially averaged climatology of the annual cycle, interannual variability of parameters using r, RMSE, STD, the Climate Prediction Index (CPI – ratio of RMSE to the STD of the observations) (Agosta et al. 2015), as well as the spatial trends and biases (at each grid point) between the model data and reanalysis/observations to illustrate how temperature varies across the study area. The range of the statistical indices for each model was divided into four categories: 0-25% – very good, 25-50% – good, 50-75% – satisfactory, and 75-100% – unsatisfactory. Each category was assigned a score from 3 to 0, respectively. For correlation, the scoring was reversed. These scores were then summed for each model to obtain a total skill score. Based on this total skill score, the top 25% of GCMs were selected as an optimal ensemble.

Approach for comparing model evaluation methods

SAT in the Arctic was analyzed for two periods, 1951-1980 and 1981-2010, to evaluate the consistency of model evaluation methods. We assume that if the models selected by each individual method are consistent across these two periods, then the method can reliably identify skillful models regardless of potential inconsistencies arising from different phases of natural and internal variability. For each period, the statistical metrics were normalized to a scale from 0 to 1 (with 1 indicating perfect performance). The original statistical metrics prior to normalization are provided in the Appendices (Fig. A.10-A.13, Table A.14-A.17). Models were ranked based on their ability to simulate SAT; if multiple metrics were used, their mean value was employed for the final ranking. Following ranking, the top 25% (in this case, six models) were selected as the optimal model ensemble for each period (Aghakhani Afshar et al. 2017; Ruan et al. 2019; Gnatiuk et al. 2020). Additionally, three quality groups (QGs) of models were distinguished based on their rankings: the first 25% of the models were classified as very good (QG I), the last 25% as unsatisfactory (QG III), and the remaining 50% as satisfactory (QG II).

The consistency of the model evaluation methods was analyzed using mean absolute differences of normalized statistical metrics and mean absolute differences of ranks between two time periods of all models. The consistency of a method is considered better when the specified absolute differences are closer to zero. Furthermore, the study examined whether a model belonged to the same quality group in both periods. If this was the case, the model's ranking was defined as consistent. Finally, we summarized the percentage of models that were consistent according to each method. These values were then used to compare the evaluation methods.

RESULTS

Model evaluation according to each considered method

i) Method of model comparison by ***root mean square error*** (RMSE)

The results of the normalized values of RMSE along with the assigned model ranks and quality groups for two periods are presented in Fig. 1. Based on the ranking results, the following models were selected for the sub-ensemble (the best models are highlighted in green and belong to the first quality group, QG I):

- for the period 1951-1980: ACCESS-ESM1-5, AWI-CM-1-1-MR, CESM2-WACCM, GFDL-ESM4, MPI-ESM1-2-LR and NorESM2-LM;
- for the period 1981-2010: ACCESS-ESM1-5, CESM2-WACCM, EC-Earth3-Veg, FIO-ESM-2-0, MPI-ESM1-2-HR, and MPI-ESM1-2-LR.

In Fig. 1, the three far right columns display the evaluation of the method's consistency. The intermodel mean difference in normalized RMSE between 1981-2010 and 1951-1980 is 0.06; the mean difference in model ranks is 3.0; and the results of the consistency assessment based on the model classified into the same quality group in both periods are 60%. Additionally, during both periods, 3 models out of 25 were classified as QG I, 8 as QG II, and 4 as QG III.

It should be noted that in this case, RMSE normalization involved converting values so that 1 indicates perfect performance and 0 indicates poor performance, allowing for comparison across all methods using the mean of the statistic. For original RMSE values, see Fig. A.10.

A similar ranking of models and assignment of each model to a quality group was performed for the other six model evaluation methods. Figures analogous to Fig. 1 for these additional evaluation methods are presented in Figs. A.4-A.9. The results of the comparison of the model evaluation methods are summarized in Table 1.

Intercomparison of the model evaluation methods

The summarized results of the consistency assessment for all methods are presented in Table 1 and Fig. 2. The first column in Table 1 indicates the number of models that were included in the selected sub-ensemble across both periods. The value in parentheses represents the percentage of the possible six sub-ensemble models, corresponding to the top 25% of the full ensemble. For example, for method iii, no models were included in the selected sub-ensemble in either period. The highest consistency was observed for methods i and vii, where three and four models,

№	Model acronym	Period 1951-1980			Period 1981-2010			RMSE diff.	Model rank diff.	Consistency (QG)
		RMSE	Model rank	Quality group	RMSE	Model rank	Quality group			
1	ACCESS-CM2	0.62	20	III	0.54	24	III	0.08	4	yes
2	ACCESS-ESM1-5	0.96	3	I	0.98	4	I	0.03	1	yes
3	AWI-CM-1-1-MR	0.94	5	I	0.88	10	II	0.06	5	no
4	BCC-CSM2-MR	0.81	12	II	0.81	14	II	0.00	2	yes
5	CAMS-CSM1-0	0.62	19	II	0.56	23	III	0.06	4	no
6	CanESM5	0.90	10	II	0.97	7	II	0.07	3	yes
7	CESM2-WACCM	0.95	4	I	0.97	5	I	0.01	1	yes
8	CIESM	0.53	23	III	0.58	22	III	0.06	1	yes
9	EC-Earth3	0.55	22	III	0.78	16	II	0.24	6	no
10	EC-Earth3-Veg	0.75	16	II	0.99	3	I	0.24	13	no
11	FGOALS-f3-L	0.46	24	III	0.60	21	III	0.14	3	yes
12	FGOALS-g3	0.00	25	III	0.00	25	III	0.00	0	yes
13	FIO-ESM-2-0	0.90	8	II	1.00	1	I	0.10	7	no
14	GFDL-ESM4	0.98	2	I	0.94	8	II	0.04	6	no
15	INM-CM4-8	0.77	15	II	0.70	19	II	0.06	4	yes
16	INM-CM5-0	0.90	9	II	0.88	11	II	0.02	2	yes
17	IPSL-CM6A-LR	0.89	11	II	0.86	12	II	0.02	1	yes
18	KACE-1-0-G	0.61	21	III	0.71	18	II	0.10	3	no
19	MIROC6	0.80	13	II	0.82	13	II	0.02	0	yes
20	MPI-ESM1-2-HR	0.92	7	II	0.97	6	I	0.05	1	no
21	MPI-ESM1-2-LR	1.00	1	I	0.99	2	I	0.01	1	yes
22	MRI-ESM2-0	0.79	14	II	0.78	15	II	0.01	1	yes
23	NESM3	0.67	18	II	0.65	20	III	0.02	2	no
24	NorESM2-LM	0.94	6	I	0.92	9	II	0.01	3	no
25	NorESM2-MM	0.73	17	II	0.73	17	II	0.00	0	yes
							Mean	0.06	3.0	yes - 60%

- very good (I) - satisfactory (II) - unsatisfactory (II) - QUALITY GROUPS (QG)

I (3) II (8) III (4) / got into different groups (10) - number of models belong to each QG in two periods

Fig. 1. Results of normalized RMSE, model ranks, quality groups and consistency assessment for 25 models for the periods 1951-1980 and 1981-2010 for SAT over the Arctic

respectively, appeared in the selected sub-ensembles across both periods. The difference in mean normalized statistical metrics between the two periods (where 0 indicates perfect agreement) ranges from 0.06 (method i) to 0.36 (method iii). The most effective methods are (i) RMSE, (vi) Taylor diagram, bias and trend, and (vii) the percentile-based method. The difference in rank values between the two periods (where 0 indicates perfect agreement) varies from 3.0 (method i) to 8.9 (method ii). The best-performing methods are (i) RMSE, (vii) the percentile-based approach, and (iv) the S_{score} method. The last column in Table 1 and Fig. 2 (left) shows the percentage of models classified into the same quality group across both periods. The lowest consistency based on quality groups was observed for method iii at 40%, while the highest was for method vii at 72%. Thus, among the seven evaluated methods, the most effective model evaluation techniques are (vii) the percentile-based method and (i) RMSE.

Fig. 3 illustrates how many times each model was identified as the best model and included in the sub-ensemble (top 25%), as well as how many times it was defined as unsatisfactory (worst 25%) across 14 ranking cases (7 methods × 2 periods). The most frequently selected models for the sub-ensemble are ACCESS-ESM1-5, AWI-CM-1-1-MR, EC-Earth3-Veg, GFDL-ESM4, and MPI-ESM1-2-LR. Conversely, the GCMs most commonly identified as unsatisfactory are CAMS-CSM1-0, CIESM, FGOALS-g3, INM-CM4-8, and NESM3. Climate models such as AWI-CM-1-1-MR, BCC-CSM2-MR, INM-CM5-0, and MPI-ESM1-2-LR were never classified among the worst 25%. Similarly, models like CAMS-CSM1-0 and MIROC6 were never ranked in the top 25%.

Interannual variability of SAT over the Arctic for the periods 1951-1980 (left) and 1981-2010 (right), for observations, the full ensemble, and sub-ensembles using the seven model assessment methods, are presented in Figs. 4 and 5. In general, all sub-ensembles exhibit an interannual variability distribution pattern similar to that of the observations but with some errors (larger or smaller). The averaging of the full ensemble significantly smooths the interannual temperature amplitude. Overall, the SAT of the full ensemble and sub-ensembles based on methods iii and iv underestimate observations during the period 1951-1980; sub-ensembles based on methods i, ii, v, and vii better reproduce the SAT. During the period 1981-2010, models selected based on methods iii and ii underestimate SAT compared to observations; however, methods i, v, vi, and vii better reproduce the interannual variability of SAT.

Boxplots of the annual SAT over the Arctic for the periods 1951-1980 and 1981-2010 are shown in Fig. 5. The dashed line indicates the mean value of observations. In the first period, only the mean values for the sub-ensemble selected by method vii are close to the mean of observations. For other methods, deviations from the observational mean generally range from 0.5 to 1 degree Celsius. In the second period, the shape of the distribution

Table 1. Results of selected sub-ensembles using seven methods and an assessment of their consistency

Method	Same selected models	Mean Diff. in value	Mean Diff. in rank	Models belong to each QG in two period simultaneously	Consistency (QG)
i	3 (50%)	0.06	3.0	I (3) II (8) III (4) / got into different groups (10)	60%
ii	2 (33%)	0.18	8.9	I (2) II (10) III (1) / got into different groups (12)	52%
iii	0	0.36	8.3	I (0) II (7) III (3) / got into different groups (15)	40%
iv	2 (33%)	0.31	5.7	I (2) II (7) III (3) / got into different groups (13)	48%
v	1 (17%)	0.32	8.5	I (1) II (8) III (2) / got into different groups (14)	44%
vi	2 (33%)	0.12	6.7	I (2) II (7) III (3) / got into different groups (13)	52%
vii	4 (67%)	0.15	4.4	I (4) II (10) III (4) / got into different groups (7)	72%

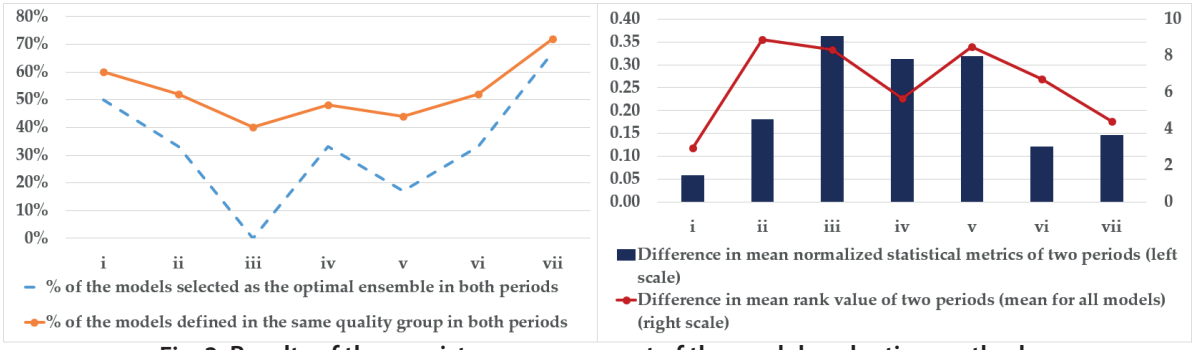

Fig. 2. Results of the consistency assessment of the model evaluation methods

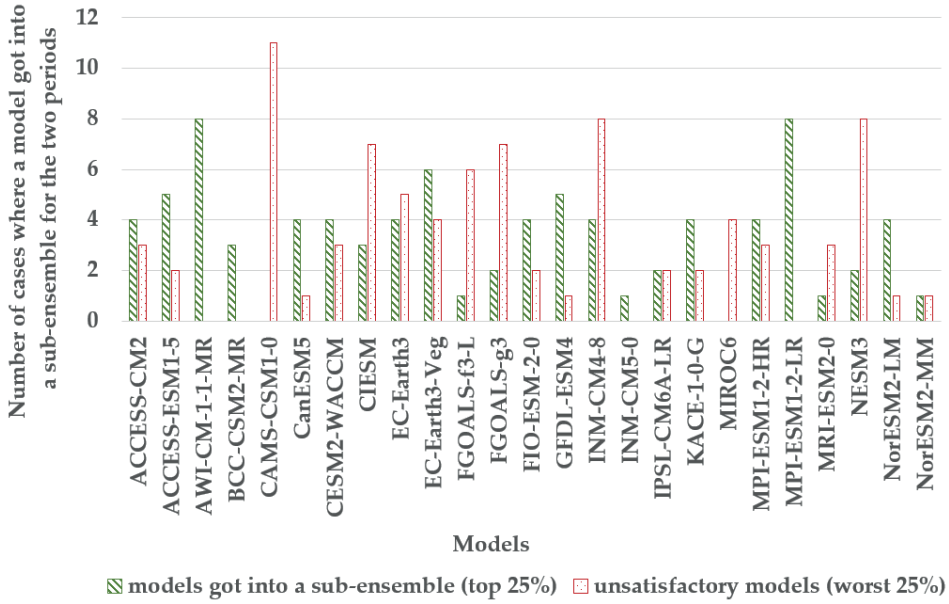

Fig. 3. Number of cases where a model entered a sub-ensemble (top 25% of models – in green) and where a model was defined as unsatisfactory (bottom 25% of models – in red) for two periods. The maximum number of cases is 14 (7 methods × 2 periods)

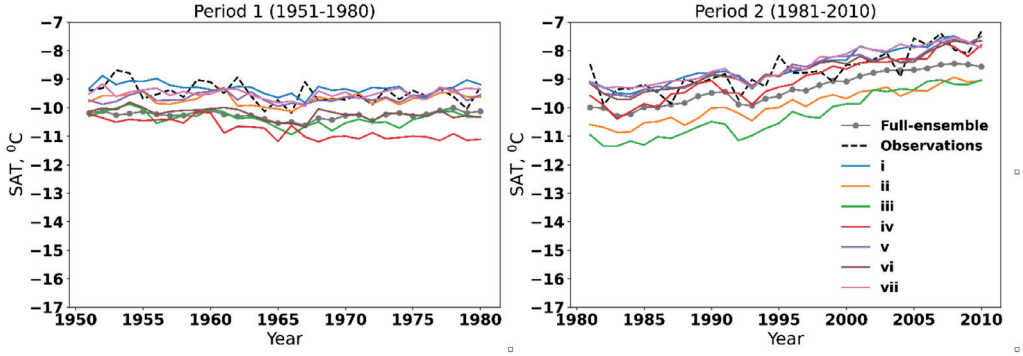

Fig. 4. Results of the annual SAT over the Arctic for observations, the full ensemble, and selected sub-ensembles using seven model assessment methods for 1951-1980 and 1981-2010

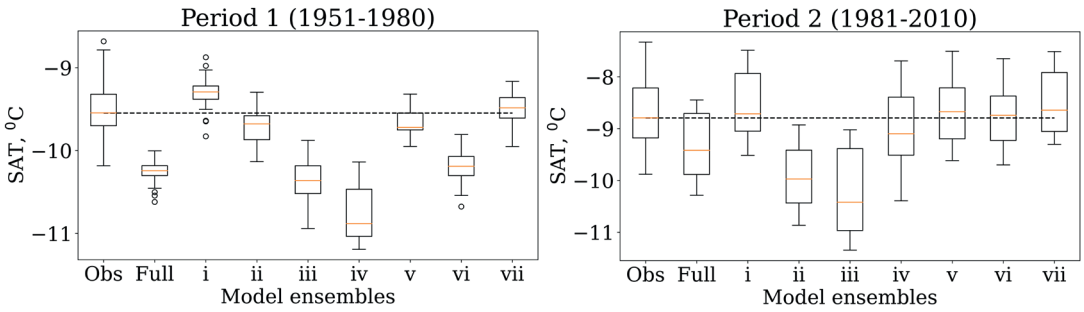

Fig. 5. Boxplots of the annual SAT over the Arctic for observations, the full ensemble, and selected sub-ensembles using the seven model assessment methods for 1951-1980 and 1981-2010

for all methods is closer to that of observations, except for methods ii and iii, which significantly underestimate them.

Regarding the annual cycle (Fig. 6), greater agreement is watched during the warm period and less during the cold period between the different sub-ensembles and observations. The exception is method iii, which underestimates SAT in all months. However, for some methods (e.g., iv and vii), the agreement is better in winter than in summer and autumn.

Fig. 7 shows the results of GCMs ranking by seven methods for the two periods. Such a comparison allows us to evaluate the consistency of the methods. We observe considerable disagreement in the results, with some methods ranking certain models as among the best, while others classify them as among the worst – for example, ACCESS-CM2, CIESM, EC-Earth3, EC-Earth3-Veg, and KACE-1-0-G. However, some models are consistently classified as either "good" or "bad" by most methods – for instance, AWI-CM-1-1-MR, CASM-CSM1-0, and MPI-ESM1-2-LR. From Fig. 7, we can clearly see the consistency of each method; for example, method (i) identifies the ACCESS-CM2 model as unsatisfactory in both periods – showing high consistency. Conversely, method (ii) classifies the same model as very good across both periods. However, when analyzing these two methods (i and ii) across all models, it becomes evident that the consistency of method (ii) is significantly lower and it more frequently assigns incorrect categories to models.

DISCUSSION AND CONCLUSIONS

Seven different model evaluation methods for the selection of a sub-ensemble were tested for CMIP6 SAT over the Arctic. All model evaluation methods were analyzed for two periods, 1951-1980 and 1981-2010, to assess their consistency. Specifically, differences in mean values, model rankings, and the matching of assigned quality groups were examined. The ability of a model evaluation method to identify the climate model as superior, regardless of the time (e.g., warming or cooling), confirms its robustness. For each evaluation method, a sub-ensemble comprising the top 25% of models was selected based on ranking, which was 6 out of 25 GCMs.

The intercomparison results indicate superior performance for the methods (i) root mean square error and (vii) the percentile-based approach. The models selected for the sub-ensemble under the (i) root mean square error method for both periods are ACCESS-ESM1-5, CESM2-WACCM, and MPI-ESM1-2-LR. Similarly, under the (vii) the percentile-based approach, the models selected for both periods are ACCESS-ESM1-5, AWI-CM-1-1-MR, GFDL-ESM4, and MPI-ESM1-2-LR. The most frequently included models - appearing more than 4 times out of 14 - under the tested evaluation methods across both periods (1951-1980 and 1981-2010) are ACCESS-ESM1-5, AWI-CM-1-1-MR, EC-Earth3-Veg, GFDL-ESM4, and MPI-ESM1-2-LR. The GCMs most commonly identified as unsatisfactory include CAMS-CSM1-0, CIESM, FGOALS-g3, INM-CM4-8, and NESM3.

Considering that the simulated distribution of interannual variability is comparable to observations, albeit with some systematic errors (Fig. 2), we recommend applying bias correction to the CMIP6 temperature data.

In general, comparing the results obtained here with those from other studies is challenging due to the use of different sets of model input data (e.g., 22, 25, 30, 35 models, etc.), various study regions, and differing meteorological parameters. Furthermore, most studies employ only one method of model evaluation and selection without comparing it to alternative approaches. Given the wide range of such methods, comparing their effectiveness for evaluation, ranking, and selecting climate models remains an important task.

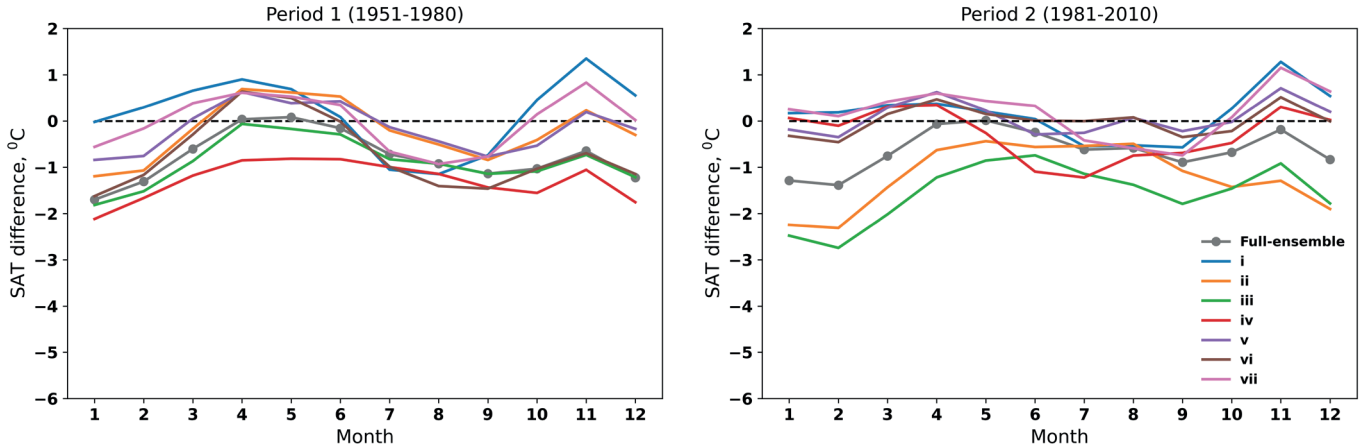

Fig. 6. Difference between the multi-year monthly SAT of observations, the full ensemble, and selected sub-ensembles using the seven methods for the periods 1951-1980 and 1981-2010 (closer to the dashed line indicates closer agreement with observations)

No	Models/methods	Period 1951-1980							Period 1981-2010						
		i	ii	iii	iv	v	vi	vii	i	ii	iii	iv	v	vi	vii
1	ACCESS-CM2	20	3	2	24	8	6	16	24	4	19	19	15	18	18
2	ACCESS-ESM1-5	3	24	22	19	9	18	4	4	11	5	13	17	13	4
3	AWI-CM-1-1-MR	5	13	16	11	1	3	3	10	9	2	10	1	3	5
4	BCC-CSM2-MR	12	9	11	8	15	5	11	14	16	4	6	12	10	11
5	CAMS-CSM1-0	19	20	24	22	17	23	21	23	18	25	22	25	25	23
6	CanESM5	10	5	12	20	12	8	12	7	3	18	8	4	1	8
7	CESM2-WACCM	4	19	20	7	18	14	6	5	10	8	5	23	16	24
8	CIESM	23	2	5	17	2	12	22	22	24	22	21	16	21	7
9	EC-Earth3	22	12	1	2	6	7	23	16	25	12	4	22	23	19
10	EC-Earth3-Veg	16	25	21	9	25	25	19	3	2	9	1	6	2	1
11	FGOALS-f3-L	24	10	13	21	19	20	24	21	1	13	15	9	14	20
12	FGOALS-g3	25	8	10	4	10	24	25	25	19	3	23	18	24	25
13	FIO-ESM-2-0	8	21	18	12	22	17	5	1	5	16	2	13	9	9
14	GFDL-ESM4	2	11	7	1	16	4	1	8	13	21	14	7	11	6
15	INM-CM4-8	15	1	4	25	4	2	20	19	22	24	20	24	22	21
16	INM-CM5-0	9	18	19	16	5	13	10	11	12	11	18	8	8	10
17	IPSL-CM6A-LR	11	15	17	23	14	16	7	12	14	7	24	2	4	16
18	KACE-1-0-G	21	22	9	14	11	10	17	18	6	1	7	5	5	14
19	MIROC6	13	17	25	10	13	21	15	13	7	23	25	19	19	15
20	MPI-ESM1-2-HR	7	6	23	13	3	19	8	6	21	20	17	10	12	3
21	MPI-ESM1-2-LR	1	4	3	18	7	1	2	2	15	10	16	11	6	2
22	MRI-ESM2-0	14	16	14	3	20	15	13	15	20	15	11	20	17	17
23	NESM3	18	23	8	6	24	22	18	20	23	6	9	21	20	22
24	NorESM2-LM	6	14	15	5	21	9	14	9	17	17	3	3	7	12
25	NorESM2-MM	17	7	6	15	23	11	9	17	8	14	12	14	15	13

- very good - satisfactory - unsatisfactory

Fig. 7. Ranking of the models based on the seven model evaluation methods for the periods 1951-1980 and 1981-2010

Among all the methods for estimating model skill examined in this article, we recommend using the following for temperature data: (i) root mean square error and (vii) a percentile-based method, as they produce the most consistent results. It should be noted that we did not assess the sensitivity of the choice of evaluation method to different variables; therefore, our findings should be applied with caution to other meteorological parameters. For example, methods based solely on RMSE may not be suitable for precipitation. When selecting climate models for other meteorological variables, more comprehensive approaches are likely to be more robust than those relying on a single statistical parameter.

REFERENCES

Aghakhani Afshar A., Hasanzadeh Y., Besalatpour A.A., and Pourreza-Bilondi M. (2017). Climate change forecasting in a mountainous data scarce watershed using CMIP5 models under representative concentration pathways. Theoretical and Applied Climatology, 129, 683-699, DOI: 10.1007/s00704-016-1908-5.

Agosta C., Fettweis X., and Datta R. (2015). Evaluation of the CMIP5 models in the aim of regional modelling of the Antarctic surface mass balance. Cryosphere, 9, 2311-2321, DOI: 10.5194/tc-9-2311-2015.

Ahmadalipour A., Rana A., Moradkhani H., Sharma A. (2017). Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis. Theoretical and Applied Climatology, 128, 71-87, DOI: 10.1007/s00704-015-1695-4.

Ahmed K., Sachindra D., Shahid S., et al. (2020). Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmospheric Research, 236, 104806, DOI: 10.1016/j.atmosres.2019.104806.

Ahmed K., Sachindra D.A., Shahid S., et al. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. Hydrology and Earth System Sciences, 23,4803-4824, DOI: 10.5194/hess-23-4803-2019.

Caballero R., Huber M. (2013). State-dependent climate sensitivity in past warm climates and its implications for future climate projections. Proceedings of the National Academy of Sciences, 110, 14162-14167, DOI: doi.org/10.1073/pnas.1303365110.

Cai Z., You Q., Wu F., et al. (2021). Arctic Warming Revealed by Multiple CMIP6 Models: Evaluation of Historical Simulations and Quantification of Future Projection Uncertainties. Journal of Climate, 34, 4871-4892, DOI: 10.1175/JCLI-D-20-0791.1.

Calvin K., Dasgupta D., Krinner G., et al. (2023). IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.

Chai T. and Draxler R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247-1250, DOI: 10.5194/gmd-7-1247-2014.

Chen W., Jiang Z., and Li L. (2011). Probabilistic Projections of Climate Change over China under the SRES A1B Scenario Using 28 AOGCMs. Journal of Climate, 24, 4741-4756, DOI: 10.1175/2011JCLI4102.1.

Eyring V., Cox P.M., Flato G.M., et al. (2019). Taking climate model evaluation to the next level. Nature Climate Change, 9, 102-110, DOI: 10.1038/s41558-018-0355-y.

Fu G., Liu Z., Charles S.P., et al. (2013). A score-based method for assessing the performance of GCMs: A case study of southeastern Australia. Journal of Geophysical Research-Atmospheres, 118, 4154-4167, DOI: doi.org/10.1002/jgrd.50269.

Gleckler P.J., Taylor K.E., and Doutriaux C. (2008). Performance metrics for climate models. Journal of Geophysical Research-Atmospheres, 113, 1-20, DOI: 10.1029/2007JD008972.

Gnatiuk N., Radchenko I., Davy R., et al. (2020). Simulation of factors affecting Emiliania huxleyi blooms in Arctic and sub-Arctic seas by CMIP5 climate models: model validation and selection. Biogeosciences, 17, 1199-1212, DOI: 10.5194/bg-17-1199-2020.

Hausfather Z., Marvel K., Schmidt G.A., et al. (2022). Climate simulations: recognize the 'hot model' problem. Nature, 605, 26-29, DOI: 10.1038/d41586-022-01192-2.

Herger N., Abramowitz G., Knutti R., et al. (2018). Selecting a climate model subset to optimise key ensemble properties. Earth System Dynamics, 9, 135-151, DOI: 10.5194/esd-9-135-2018.

Inoue T., and Ueda H. (2011). Delay of the First Transition of Asian Summer Monsoon under Global Warming Condition. SOLA, 7, 81-84, DOI: 10.2151/sola.2011-021.

Jain S., Scaife A.A., Shepherd T.G., et al. (2023). Importance of internal variability for climate model assessment. npj Climate and Atmospheric Science, 6, 68, DOI: 10.1038/s41612-023-00389-0.

Jia K., Ruan Y., Yang Y., and You Z. (2019). Assessment of CMIP5 GCM Simulation Performance for Temperature Projection in the Tibetan Plateau. Earth and Space Science, 6, 2362-2378, DOI: 10.1029/2019EA000962.

Jiang Z., Li W., Xu J., and Li L. (2015). Extreme Precipitation Indices over China in CMIP5 Models. Part I: Model Evaluation. Journal of Climate, 28, 8603-8619, DOI: 10.1175/JCLI-D-15-0099.1.

Kadel I., Yamazaki T., Iwasaki T., and Abdillah M. (2018). Projection of future monsoon precipitation over the central Himalayas by CMIP5 models under warming scenarios. Climate Research, 75, 1-21, DOI: 10.3354/cr01497.

Knutti R., Furrer R., Tebaldi C., et al. (2010). Challenges in Combining Projections from Multiple Climate Models. Journal of Climate, 23, 2739-2758, DOI: 10.1175/2009JCLI3361.1.

Kumar D., Mishra V., and Ganguly A.R. (2015). Evaluating wind extremes in CMIP5 climate models. Climate Dynamics, 45, 441-453, DOI: 10.1007/s00382-014-2306-2.

Kumar S., Merwade V., Kinter J.L., and Niyogi D. (2013). Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations. Journal of Climate, 26, 4168-4185, DOI: 10.1175/JCLI-D-12-00259.1.

Macadam I., Pitman A.J., Whetton P.H., and Abramowitz G. (2010). Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments. Geophysical Research Letters, 37, 16704, DOI: 10.1029/2010GL043877.

Maxino C.C., McAvaney B.J., Pitman A.J., and Perkins S.E. (2008). Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. International Journal of Climatology, 28, 1097-1112, DOI: 10.1002/joc.1612.

McMahon T.A., Peel M.C., and Karoly D.J. (2015). Assessment of precipitation and temperature data from CMIP3 global climate models for hydrologic simulation. Hydrology and Earth System Sciences, 19, 361-377, DOI: 10.5194/hess-19-361-2015.

Ogata T., Ueda H., Inoue T., et al. (2014). Projected Future Changes in the Asian Monsoon: A Comparison of CMIP3 and CMIP5 Model Results. Journal of the Meteorological Society of Japan, 92, 207-225, DOI: 10.2151/jmsj.2014-302.

Otero N., Sillmann J., and Butler T. (2018). Assessment of an extended version of the Jenkinson-Collison classification on CMIP5 models over Europe. Climate Dynamics, 50, 1559-1579, DOI: 10.1007/s00382-017-3705-y.

Perkins S.E., Pitman A.J., Holbrook N.J., and McAneney J. (2007). Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. Journal of Climate, 20, 4356-4376, DOI: 10.1175/JCLI4253.1.

Raju K.S., and Kumar D.N. (2020). Review of approaches for selection and ensembling of GCMs. Journal of Water and Climate Change, 11, 577-599, DOI: 10.2166/wcc.2020.128.

Rao X., Lu X., and Dong W. (2019). Evaluation and Projection of Extreme Precipitation over Northern China in CMIP5 Models. Atmosphere, 10, 691, DOI: 10.3390/atmos10110691.

Reifen C., and Toumi R. (2009). Climate projections: Past performance no guarantee of future skill? Geophysical Research Letters, 36, DOI: 10.1029/2009GL038082.

Rohde R., Muller R.A., Jacobsen R., et al. (2013). A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. Geoinformatics and Geostatistics: An Overview, 1:1, DOI: 10.4172/2327-4581.1000101.

Rohde R.A., and Hausfather Z. (2020). The Berkeley Earth Land/Ocean Temperature Record. Earth System Science Data, 12, 3469-3479, DOI: 10.5194/essd-12-3469-2020.

Ruan Y., Liu Z., Wang R., and Yao Z. (2019). Assessing the Performance of CMIP5 GCMs for Projection of Future Temperature Change over the Lower Mekong Basin. Atmosphere, 10, 93, DOI: 10.3390/atmos10020093.

Rupp D.E., Abatzoglou J.T., Hegewisch K.C., and Mote P.W. (2013). Evaluation of CMIP5 20 th century climate simulations for the Pacific Northwest USA. Journal of Geophysical Research-Atmospheres, 118, 884-894, DOI: 10.1002/jgrd.50843.

Sharmila S., Joseph S., Sahai A., et al. (2015). Future projection of Indian summer monsoon variability under climate change scenario: An assessment from CMIP5 climate models. Global and Planetary Change, 124, 62-78, DOI: 10.1016/j.gloplacha.2014.11.004.

Sherwood S.C., Webb M.J., Annan J.D., et al. (2020). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. Reviews of Geophysics, 58, e2019RG000678, DOI: 10.1029/2019RG000678.

Sillmann J., Kharin V.V., Zhang X., et al. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. Journal of Geophysical Research-Atmospheres, 118, 1716-1733, DOI: 10.1002/jgrd.50203.

Stocker T.F., Qin D., Plattner G.-K., et al. (2014). Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Taylor K.E. (2001). Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research-Atmospheres, 106, 7183-7192, DOI: 10.1029/2000JD900719.

Taylor K.E., Stouffer R.J., and Meehl G.A. (2012). An Overview of CMIP5 and the Experiment Design. Bulletin of the American Meteorological Society, 93, 485-498, DOI: 10.1175/BAMS-D-11-00094.1.

Walsh J.E., Chapman W.L., Romanovsky V., et al. (2008). Global Climate Model Performance over Alaska and Greenland. Journal of Climate, 21, 6156-6174, DOI: 10.1175/2008JCLI2163.1.

Yang X., Yu X., Wang Y., et al. (2020). The Optimal Multimodel Ensemble of Bias-Corrected CMIP5 Climate Models over China. Journal of Hydrometeorology, 21, 845-863, DOI: 10.1175/JHM-D-19-0141.1.

You Q., Jiang Z., Wang D., et al. (2018). Simulation of temperature extremes in the Tibetan Plateau from CMIP5 models and comparison with gridded observations. Climate Dynamics, 51, 355-369, DOI: 10.1007/s00382-017-3928-y.

Zhou B., Wen Q.H., Xu Y., et al. (2014). Projected Changes in Temperature and Precipitation Extremes in China by the CMIP5 Multimodel Ensembles. Journal of Climate, 27, 6591-6611, DOI: 10.1175/JCLI-D-13-00761.1.

APPENDICES

Table A.1. CMIP6 models used for the evaluation of surface air temperature in the Arctic

ID	Model acronym	Modeling center (acronym, full name, city and country)	Resolution (° lon × ° lat)
1	ACCESS-CM2	ARCCSS (Australian Research Council Centre of Excellence for Climate System Science), CSIRO (Commonwealth Scientific and Industrial Research Organization, Aspendale, Victoria, Australia)	1.875×1.25
2	ACCESS-ESM1-5	Commonwealth Scientific and Industrial Research Organization, Aspendale, Victoria, Australia	1.875×1.25
3	AWI-CM-1-1-MR	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany	0.938
4	BCC-CSM2-MR	Beijing Climate Center, Beijing, China	1.125
5	CAMS-CSM1-0	Chinese Academy of Meteorological Sciences, Beijing, China	1.125
6	CanESM5	Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, Canada	2.8125
7	CESM2-WACCM	National Center for Atmospheric Research, Climate and Global Dynamics Laboratory, Boulder, USA	1.25×0.94
8	CIESM	Department of Earth System Science, Tsinghua University, Beijing, China	1.25
9	EC-Earth3	European Union, Mailing address: EC-Earth consortium, Rossby Center, Swedish Meteorological and Hydrological Institute/SMHI, SE-601 76 Norrkoping, Sweden	0.703
10	EC-Earth3-Veg		
11	FGOALS-f3-L	Chinese Academy of Sciences, Beijing, China	1.25×1
12	FGOALS-g3		2
13	FIO-ESM-2-0	FIO (First Institute of Oceanography, Ministry of Natural Resources, Qingdao, China), QNLM (Qingdao National Laboratory for Marine Science and Technology, Qingdao, China)	1.25×0.94
14	GFDL-ESM4	National Oceanic and Atmospheric Administration, Geophysical Fluid Dynamics Laboratory, Princeton, USA	1.25
15	INM-CM4-8	Institute for Numerical Mathematics, Russian Academy of Science, Moscow, Russia	2×1.5
16	INM-CM5-0		
17	IPSL-CM6A-LR	Institut Pierre Simon Laplace, Paris, France	2×1.268
18	KACE-1-0-G	National Institute of Meteorological Sciences/Korea Meteorological Administration, Climate Research Division, Seogwipo-si, Jejudo, Republic of Korea	1.875×1.25
19	MIROC6	JAMSTEC (Japan Agency for Marine-Earth Science and Technology, Kanagawa, Japan), AORI (Atmosphere and Ocean Research Institute, The University of Tokyo, Chiba, Japan), NIES (National Institute for Environmental Studies, Ibaraki, Japan), and R-CCS (RIKEN Center for Computational Science, Hyogo, Japan)	1.406
20	MPI-ESM1-2-HR	Max Planck Institute for Meteorology, Hamburg Deutscher Wetterdienst, Offenbach am Main Deutsches Klimarechenzentrum, Hamburg, Germany	0.938
21	MPI-ESM1-2-LR	Max Planck Institute for Meteorology, Hamburg 20146, Germany; Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany	1.875
22	MRI-ESM2-0	Meteorological Research Institute, Tsukuba, Ibaraki Japan	1.125
23	NESM3	Nanjing University of Information Science and Technology, Nanjing, China	1.875
24	NorESM2-LM	NorESM Climate modeling Consortium consisting of CICERO (Center for International Climate and Environmental Research, Oslo), MET-Norway (Norwegian Meteorological Institute, Oslo), NERSC (Nansen Environmental and Remote Sensing Center, Bergen), NILU (Norwegian Institute for Air Research, Kjeller), UiB (University of Bergen, Bergen), UiO (University of Oslo, Oslo) and UNI (Uni Research, Bergen), Norway.	2.5×1.89
25	NorESM2-MM		1.25×0.94

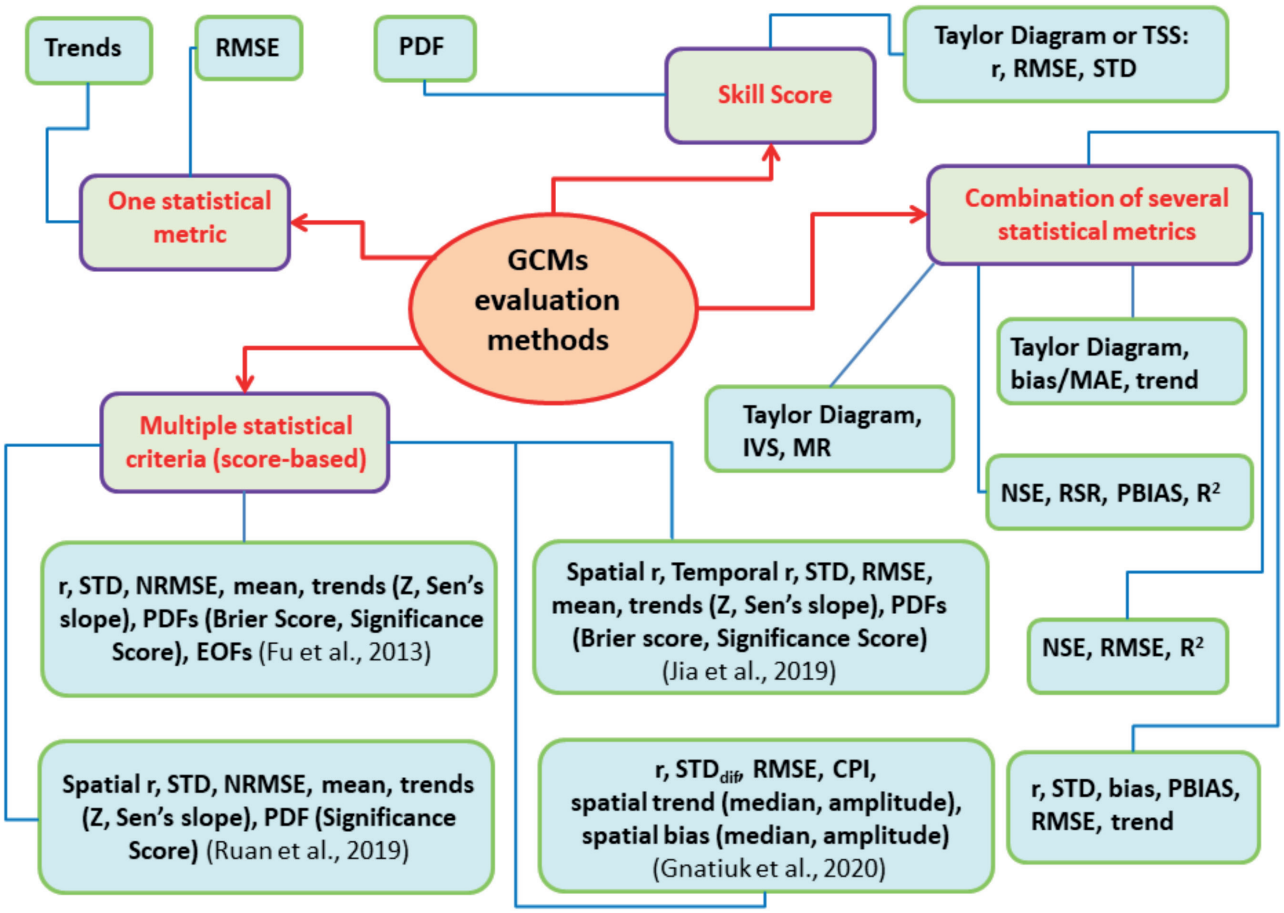

Fig. A.2. Mind map of methods for GCMs' evaluation found in the literature

Table A.3. Climate models selection methods in the literature (SLP – sea level pressure, H500 – geopotential height at 500 hPa, T500 – the temperature at 500 hPa, SST – sea surface temperature, SAT – surface air temperature, P – precipitation, SIC – sea-ice concentration, WS – wind speed, pCO_2 – dissolved CO_2 partial pressure, S – salinity, OCS – ocean surface current speed, SDSR – surface downwelling shortwave radiation)

#	Statistical metric	Reference	GCMs considered / selected	Variable considered	Ranking method
1	Root mean square error (RMSE)	1) (Walsh et al. 2008)	15/4	SLP, SAT, P	Based on the sum of ranks for all variables
		2) (Reifen and Toumi 2009)	17/ -	SAT	Ranking
		3) Macadam et al. 2010)	17/ -	SAT	Ranking
		4) (Sillmann et al. 2013)	18-31/-	Temp. and Precip. Indices (in total 27)	Ranking
		5) (Zhou et al. 2014)	24/total ensemble	Temp. and Precip. Indices (in total 20)	Ranking
		6) (Herger et al. 2018)	38/8 for SAT, 12 for P	SAT, P	Gurobi Optimization
2	Trends	1) (Kumar et al. 2013) (mean, spatial standard deviation, spatial correlation)	79 model runs from 19 GCMs/ -	SAT, P	
		2) (Saha et al. 2014) (temporal trends)	42/ -	Indian Summer Monsoon Rainfall	
3	Probability density function (PDF) – Skill score (SS)	1) (Perkins et al. 2007)	10/6; 13/10; 14/3	Maximum and minimum SAT, P	Ranked SS for each variable, select >0.8
		2) (Maxino et al. 2008)	9/3	Maximum and minimum SAT, P	Ranked averaged SS, select >0.8
		3) (Anandhi and Nanjundiah 2015)	19/5	P	Average rank of GCMs using SS
		4) (Sun et al. 2015)	14/ -	SAT (maximum, mean, minimum), P	SS
		5) (Bannister et al. 2017)	47/5	SAT (maximum, mean, minimum)	Ranking based on aggregated SS
		6) (Anandhi et al. 2019)	20/5	SAT (mean, max, min) P, WS	Ranking based on averaged SS
4	Taylor Diagram or Taylor Skill Score (TSS)	1) (Inoue and Ueda 2011) (Taylor's Skill Score)	19/total ensemble	300 hPa temp, 850 hPa zonal wind	
		2) (Ogata et al. 2014) (Taylor's Skill Score)	20 (cmip3)/20 24 (cmip5)/24	850 hPa zonal wind, P SST	
		3) (Sharmila et al. 2015)	20/4	Indian summer monsoon, precip.	threshold criteria (correlation ≥ 0.5 and normalized SD 0.8-1.2)
		4) (Kadel et al. 2018)	38/6	Summer Monsoon Season, precip.	threshold criteria (r≥ 0.6, std 0.5-1.5, rmse ≤ 1.0)
		5) (Ahmed et al. 2020)	36/18	SAT (maximum, minimum), P	based on TSS, Comprehensive Rating Metric (MR)
		6) (Yang et al. 2020)	25/9 for precip., 25/11 for temp.	SAT, P	based on threshold – 0.49
		7) (Wang et al. 2016)	28/7	SAT	based on TSS

5	Taylor Diagrams, Interannual Variability Skill Score, Comprehensive Rating Metric	1)(Jiang et al. 2015) 2)(Cai et al. 2021)	31/5 22/6	P, P intensity, Max consec. dry days SAT	Total complex rating metric Total complex rating metric
		3)(Rao et al. 2019)	32/5	total extreme P, max consecutive five days of P and wet days >10 mm	Total complex rating metric
		4)(You et al. 2018)	17/4	16 temperature extreme indices	Total complex rating metric
6	Combinations of several metrics	1)(Kumar et al. 2015): bias, trend analysis, and Taylor Diagrams	15/ -	Maxima WS	Four groups
		2)(Aghakhani Afshar et al. 2017): R2, NSE, PBIAS, RSR	14/4	P	Ranking based on the total sum of the ranks by NSE
		3)(McMahon et al. 2015): RMSE, NSE, R2	23/5	SAT, P	Ranking based on metrics
		4)(Ongoma et al. 2019): r, STD, bias, PBIAS, RMSE, trend	22/8	P	Ranking based on the overall score calculated using an individual score for each GCM, error and variable
7	Score-based methods using multiple statistical metrics	1) (Fu et al. 2013): r, STD, NRME, mean, trends (Z, Sen's slope), PDF (Brier Score, Significance Score), empirical orthogonal functions	25/ -	SAT, P, mean SLP	Ranking based on the overall score calculated using an individual score for each GCM and error
		2) (Jia et al. 2019): mean, temporal r, STD, r spatial, RMSE, PDF, trends (Z and Slope)	33/top-30% (10)	P	Ranking based on the overall score calculated using an individual score for each GCM and error
		3) (Ruan et al. 2019): mean, STD, RMSE, r spatial, PDF (Significance Score), trends (Z and Sen's slope)	34/top-25% (9)	SAT	Ranking based on the overall score for each variable and sea individually
		4) (Gnatiuk et al. 2020): r, RMSE, STD, CPI spatial trends, spatial bias	11-30/individual sub-set (top-25%) varies from 3 to 11	pCO_2, pH, NO_3, PO_4, SI, SST, S, OCS, 10m WS, SDSR	

№	Model acronym	Period 1951-1980						Period 1981-2010						Mean value diff.	Model rank diff.	Consistency (QG)
		r	Mean diff.	STD diff.	Mean value	Model rank	Quality group	r	Mean diff.	STD diff.	Mean value	Model rank	Quality group			
1	ACCESS-CM2	0.71	0.78	0.98	0.82	3	I	0.73	0.87	0.94	0.84	4	I	0.02	1	yes
2	ACCESS-ESM1-5	0.25	0.38	0.63	0.42	24	III	0.84	0.51	0.88	0.74	11	II	0.32	13	no
3	AWI-CM-1-1-MR	0.51	0.84	0.61	0.66	13	II	0.50	0.88	0.94	0.78	9	II	0.12	4	yes
4	BCC-CSM2-MR	0.73	0.84	0.58	0.72	9	II	0.24	0.83	0.95	0.67	16	II	0.05	7	yes
5	CAMS-CSM1-0	0.51	0.53	0.76	0.60	20	III	0.74	0.53	0.74	0.67	18	II	0.07	2	no
6	CanESM5	0.90	0.74	0.70	0.78	5	I	0.87	0.95	0.93	0.92	3	I	0.14	2	yes
7	CESM2-WACCM	0.00	0.85	0.98	0.61	19	II	0.64	0.76	0.86	0.75	10	II	0.14	9	yes
8	CIESM	0.68	0.93	0.94	0.85	2	I	0.15	0.41	0.80	0.45	24	III	0.39	22	no
9	EC-Earth3	0.31	0.84	0.94	0.70	12	II	0.60	0.00	0.00	0.20	25	III	0.50	13	no
10	EC-Earth3-Veg	0.22	0.00	0.44	0.22	25	III	1.00	0.91	0.89	0.93	2	I	0.71	23	no
11	FGOALS-f3-L	0.67	0.56	0.92	0.72	10	II	0.96	0.95	0.91	0.94	1	I	0.22	9	no
12	FGOALS-g3	0.69	0.92	0.62	0.74	8	II	0.41	0.71	0.88	0.67	19	II	0.08	11	yes
13	FIO-ESM-2-0	0.93	0.52	0.22	0.56	21	III	0.80	0.86	0.87	0.84	5	I	0.29	16	no
14	GFDL-ESM4	0.79	0.67	0.65	0.70	11	II	0.46	0.85	0.87	0.73	13	II	0.03	2	yes
15	INM-CM4-8	0.86	0.82	0.90	0.86	1	I	0.08	0.57	0.96	0.53	22	III	0.33	21	no
16	INM-CM5-0	0.21	0.73	0.90	0.61	18	II	0.61	0.90	0.71	0.74	12	II	0.12	6	yes
17	IPSL-CM6A-LR	0.60	0.59	0.73	0.64	15	II	0.32	0.96	0.89	0.72	14	II	0.08	1	yes
18	KACE-1-0-G	0.01	0.97	0.44	0.47	22	III	0.67	0.89	0.86	0.81	6	I	0.34	16	no
19	MIROC6	0.32	0.58	0.99	0.63	17	II	0.76	0.70	0.91	0.79	7	II	0.16	10	yes
20	MPI-ESM1-2-HR	0.81	0.46	0.98	0.75	6	I	0.00	0.73	0.89	0.54	21	III	0.21	15	no
21	MPI-ESM1-2-LR	0.43	0.97	0.99	0.80	4	I	0.38	0.89	0.79	0.69	15	II	0.11	11	no
22	MRI-ESM2-0	0.70	0.71	0.49	0.63	16	II	0.37	0.77	0.81	0.65	20	III	0.02	4	no
23	NESM3	1.00	0.31	0.00	0.44	23	III	0.36	0.25	0.81	0.47	23	III	0.04	0	yes
24	NorESM2-LM	0.05	0.91	0.99	0.65	14	II	0.23	0.93	0.86	0.67	17	II	0.02	3	yes
25	NorESM2-MM	0.54	0.89	0.82	0.75	7	II	0.74	0.72	0.88	0.78	8	II	0.03	1	yes
													Mean	0.18	8.9	yes-52%

- very good (I) - satisfactory (II) - unsatisfactory (III) - QUALITY GROUPS (QG)

I (2) II (10) III (1) / got into different groups (12) - number of models belong to each QG in two periods

Fig. A.4. Results of normalized spatial trends analysis, model rank and consistency assessment for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (ii) of model comparison by trends (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980			Period 1981-2010			TSS diff.	Model rank diff.	Consistency (QG)
		TSS	Model rank	Quality group	TSS	Model rank	Quality group			
1	ACCESS-CM2	0.99	2	I	0.49	19	II	0.50	17	no
2	ACCESS-ESM1-5	0.03	22	III	0.77	5	I	0.73	17	no
3	AWI-CM-1-1-MR	0.32	16	II	0.88	2	I	0.56	14	no
4	BCC-CSM2-MR	0.52	11	II	0.78	4	I	0.26	7	no
5	CAMS-CSM1-0	0.03	24	III	0.00	25	III	0.03	1	yes
6	CanESM5	0.51	12	II	0.52	18	II	0.01	6	yes
7	CESM2-WACCM	0.16	20	III	0.70	8	II	0.54	12	no
8	CIESM	0.92	5	I	0.29	22	III	0.63	17	no
9	EC-Earth3	1.00	1	I	0.58	12	II	0.42	11	no
10	EC-Earth3-Veg	0.05	21	III	0.66	9	II	0.60	12	no
11	FGOALS-f3-L	0.51	13	II	0.57	13	II	0.07	0	yes
12	FGOALS-g3	0.58	10	II	0.82	3	I	0.24	7	no
13	FIO-ESM-2-0	0.21	18	II	0.53	16	II	0.32	2	yes
14	GFDL-ESM4	0.65	7	II	0.33	21	III	0.33	14	no
15	INM-CM4-8	0.92	4	I	0.04	24	III	0.88	20	no
16	INM-CM5-0	0.19	19	II	0.59	11	II	0.40	8	yes
17	IPSL-CM6A-LR	0.26	17	II	0.71	7	II	0.46	10	yes
18	KACE-1-0-G	0.59	9	II	1.00	1	I	0.41	8	no
19	MIROC6	0.00	25	III	0.26	23	III	0.26	2	yes
20	MPI-ESM1-2-HR	0.03	23	III	0.46	20	III	0.43	3	yes
21	MPI-ESM1-2-LR	0.93	3	I	0.64	10	II	0.29	7	no
22	MRI-ESM2-0	0.40	14	II	0.53	15	II	0.13	1	yes
23	NESM3	0.65	8	II	0.72	6	I	0.07	2	no
24	NorESM2-LM	0.38	15	II	0.53	17	II	0.15	2	yes
25	NorESM2-MM	0.88	6	I	0.55	14	II	0.33	8	no
							Mean	0.36	8.3	yes-40%

- very good (I) - satisfactory (II) - unsatisfactory (III) - QUALITY GROUPS (QG)

I (0) II (7) III (3) / got into different groups (15) - number of models belong to each QG in two periods

Fig. A.5. Results of normalized TSS, model rank and consistency assessment for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (iii) of model comparison by Taylor skill score (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980			Period 1981-2010			S_{score} diff.	Model rank diff.	Consistency (QG)
		S_{score}	Model rank	Quality group	S_{score}	Model rank	Quality group			
1	ACCESS-CM2	0.07	24	III	0.42	19	II	0.35	5	no
2	ACCESS-ESM1-5	0.32	19	II	0.83	13	II	0.51	6	yes
3	AWI-CM-1-1-MR	0.66	11	II	0.87	10	II	0.21	1	yes
4	BCC-CSM2-MR	0.73	8	II	0.89	6	I	0.16	2	no
5	CAMS-CSM1-0	0.08	22	III	0.34	22	III	0.26	0	yes
6	CanESM5	0.32	20	III	0.88	8	II	0.56	12	no
7	CESM2-WACCM	0.77	7	II	0.91	5	I	0.15	2	no
8	CIESM	0.34	17	II	0.36	21	III	0.02	4	no
9	EC-Earth3	0.93	2	I	1.00	4	I	0.07	2	yes
10	EC-Earth3-Veg	0.72	9	II	1.00	1	I	0.28	8	no
11	FGOALS-f3-L	0.08	21	III	0.80	15	II	0.71	6	no
12	FGOALS-g3	0.85	4	I	0.32	23	III	0.52	19	no
13	FIO-ESM-2-0	0.65	12	II	1.00	2	I	0.35	10	no
14	GFDL-ESM4	1.00	1	I	0.83	14	II	0.17	13	no
15	INM-CM4-8	0.00	25	III	0.42	20	III	0.42	5	yes
16	INM-CM5-0	0.37	16	II	0.76	18	II	0.39	2	yes
17	IPSL-CM6A-LR	0.07	23	III	0.00	24	III	0.07	1	yes
18	KACE-1-0-G	0.41	14	II	0.88	7	II	0.47	7	yes
19	MIROC6	0.71	10	II	0.00	25	III	0.71	15	no
20	MPI-ESM1-2-HR	0.64	13	II	0.76	17	II	0.13	4	yes
21	MPI-ESM1-2-LR	0.33	18	II	0.77	16	II	0.44	2	yes
22	MRI-ESM2-0	0.91	3	I	0.87	11	II	0.04	8	no
23	NESM3	0.77	6	I	0.87	9	II	0.11	3	no
24	NorESM2-LM	0.77	5	I	1.00	3	I	0.23	2	yes
25	NorESM2-MM	0.39	15	II	0.85	12	II	0.46	3	yes
							Mean	0.31	5.7	yes-48%

- very good (I) - satisfactory (II) unsatisfactory (III) - **QUALITY GROUPS (QG)**

I (2) II (7) III (3) / got into different groups (13) - number of models belong to each QG in two periods

Fig. A.6. Results of normalized Sscore, model rank and consistency assessment for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (iv) of model comparison by Sscore (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980					Period 1981-2010					MR_{mean} diff.	Model rank diff.	Consistency (QG)
		MR_{IVS}	MR_{Taylor}	MR_{mean}	Model rank	Quality group	MR_{IVS}	MR_{Taylor}	MR_{mean}	Model rank	Quality group			
1	ACCESS-CM2	0.71	0.71	0.79	8	II	0.58	0.30	0.45	15	II	0.34	7	yes
2	ACCESS-ESM1-5	0.67	0.62	0.72	9	II	0.13	0.75	0.38	17	II	0.35	8	yes
3	AWI-CM-1-1-MR	0.92	0.87	1.00	1	I	1.00	1.00	1.00	1	I	0.00	0	yes
4	BCC-CSM2-MR	0.33	0.51	0.45	15	II	0.50	0.66	0.56	12	II	0.11	3	yes
5	CAMS-CSM1-0	0.50	0.16	0.41	17	II	0.04	0.00	0.00	25	III	0.41	8	no
6	CanESM5	0.46	0.56	0.55	12	II	0.88	0.68	0.79	4	I	0.23	8	no
7	CESM2-WACCM	0.25	0.44	0.36	18	II	0.25	0.17	0.20	23	III	0.17	5	no
8	CIESM	1.00	0.69	0.98	2	I	0.17	0.77	0.41	16	II	0.57	14	no
9	EC-Earth3	0.83	0.67	0.86	6	I	0.00	0.51	0.20	22	III	0.66	16	no
10	EC-Earth3-Veg	0.00	0.00	0.00	25	III	0.67	0.89	0.75	6	I	0.75	19	no
11	FGOALS-f3-L	0.38	0.22	0.35	19	II	0.79	0.55	0.68	9	II	0.33	10	yes
12	FGOALS-g3	0.75	0.40	0.69	10	II	0.33	0.40	0.34	18	II	0.34	8	yes
13	FIO-ESM-2-0	0.17	0.33	0.26	22	III	0.42	0.74	0.54	13	II	0.28	9	no
14	GFDL-ESM4	0.21	0.71	0.45	16	II	0.83	0.66	0.75	7	II	0.31	9	yes
15	INM-CM4-8	0.79	0.80	0.89	4	I	0.21	0.17	0.17	24	III	0.71	20	no
16	INM-CM5-0	0.88	0.64	0.88	5	I	0.71	0.74	0.71	8	II	0.16	3	no
17	IPSL-CM6A-LR	0.42	0.42	0.47	14	II	0.96	0.89	0.93	2	I	0.46	12	no
18	KACE-1-0-G	0.54	0.49	0.58	11	II	0.75	0.83	0.78	5	I	0.20	6	no
19	MIROC6	0.58	0.27	0.51	13	II	0.38	0.34	0.34	19	II	0.17	6	yes
20	MPI-ESM1-2-HR	0.96	0.62	0.92	3	I	0.63	0.64	0.62	10	II	0.30	7	no
21	MPI-ESM1-2-LR	0.63	1.00	0.86	7	II	0.46	0.85	0.61	11	II	0.24	4	yes
22	MRI-ESM2-0	0.29	0.33	0.34	20	III	0.29	0.42	0.33	20	III	0.01	0	yes
23	NESM3	0.04	0.42	0.21	24	III	0.08	0.43	0.21	21	III	0.00	3	yes
24	NorESM2-LM	0.13	0.44	0.28	21	III	0.92	0.77	0.85	3	I	0.58	18	no
25	NorESM2-MM	0.08	0.44	0.25	23	III	0.54	0.47	0.50	14	II	0.25	9	no
											Mean	0.32	8.5	yes-44%

- very good (II) - satisfactory (II) - unsatisfactory (III) - **QUALITY GROUPS (QG)**

I (1) II (8) III (2) / got into different groups (14) - number of models belong to each QG in two periods

Fig. A.7. Results of normalized MR for Taylor and IVS metrics for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (v) based on Taylor diagram and interannual variability skill score (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980								Period 1981-2010								Mean value diff.	Model rank diff.	Consistency (QG)
		RMSE	r	|dif_std|	|dif_T|	|B|	Mean value	Model rank	Quality group	RMSE	r	|dif_std|	|dif_T|	|B|	Mean value	Model rank	Quality group			
1	ACCESS-CM2	0.51	0.94	0.79	0.78	0.67	0.74	6	I	0.44	0.59	0.88	0.87	0.61	0.68	18	II	0.06	12	no
2	ACCESS-ESM1-5	0.79	0.11	0.76	0.38	0.96	0.60	18	II	0.79	0.93	0.55	0.51	0.96	0.75	13	II	0.15	5	yes
3	AWI-CM-1-1-MR	0.78	0.52	0.85	0.84	0.84	0.77	3	I	0.71	0.91	1.00	0.88	0.78	0.86	3	I	0.09	0	yes
4	BCC-CSM2-MR	0.67	0.73	0.58	0.84	0.88	0.74	5	I	0.65	0.86	0.81	0.83	0.87	0.80	10	II	0.06	5	no
5	CAMS-CSM1-0	0.52	0.09	0.69	0.53	0.73	0.51	23	III	0.45	0.06	0.54	0.53	0.68	0.45	25	III	0.06	2	yes
6	CanESM5	0.74	0.70	0.62	0.74	0.86	0.73	8	II	0.78	0.83	0.98	0.95	0.93	0.89	1	I	0.16	7	no
7	CESM2-WACCM	0.79	0.38	0.57	0.85	0.88	0.69	14	II	0.78	0.40	0.67	0.76	0.86	0.70	16	II	0.01	2	yes
8	CIESM	0.44	0.89	0.85	0.93	0.44	0.71	12	II	0.47	0.61	0.74	0.41	0.45	0.54	21	III	0.17	9	no
9	EC-Earth3	0.45	0.93	0.81	0.83	0.65	0.74	7	II	0.63	1.00	0.00	0.00	0.83	0.49	23	III	0.24	16	no
10	EC-Earth3-Veg	0.62	0.31	0.00	0.00	0.82	0.35	25	III	0.80	0.74	0.89	0.91	0.95	0.86	2	I	0.51	23	no
11	FGOALS-f3-L	0.38	0.71	0.61	0.56	0.59	0.57	20	III	0.49	0.66	0.94	0.95	0.69	0.75	14	II	0.18	6	no
12	FGOALS-g3	0.00	0.72	0.79	0.92	0.00	0.48	24	III	0.00	0.92	0.69	0.71	0.00	0.46	24	III	0.02	0	yes
13	FIO-ESM-2-0	0.75	0.46	0.52	0.52	0.93	0.63	17	II	0.81	0.66	0.78	0.86	0.91	0.80	9	II	0.17	8	yes
14	GFDL-ESM4	0.81	0.83	0.52	0.68	0.97	0.76	4	I	0.76	0.40	0.97	0.85	0.91	0.78	11	II	0.02	7	no
15	INM-CM4-8	0.64	0.90	0.81	0.82	0.71	0.78	2	I	0.57	0.00	0.75	0.57	0.67	0.51	22	III	0.27	20	no
16	INM-CM5-0	0.74	0.38	0.83	0.73	0.86	0.71	13	II	0.71	0.68	0.91	0.90	0.84	0.81	8	II	0.10	5	yes
17	IPSL-CM6A-LR	0.73	0.49	0.61	0.59	0.80	0.65	16	II	0.70	0.78	0.98	0.96	0.73	0.83	4	I	0.18	12	no
18	KACE-1-0-G	0.50	0.74	0.75	0.97	0.70	0.73	10	II	0.57	1.00	0.91	0.89	0.77	0.83	5	I	0.10	5	no
19	MIROC6	0.67	0.00	0.74	0.58	0.83	0.56	21	III	0.67	0.36	0.77	0.70	0.83	0.66	19	II	0.10	2	no
20	MPI-ESM1-2-HR	0.76	0.09	0.84	0.46	0.81	0.59	19	II	0.78	0.55	0.88	0.73	0.87	0.76	12	II	0.17	7	yes
21	MPI-ESM1-2-LR	0.83	0.91	0.74	0.97	0.92	0.87	1	I	0.80	0.75	0.81	0.89	0.86	0.82	6	I	0.05	5	yes
22	MRI-ESM2-0	0.66	0.63	0.58	0.72	0.74	0.67	15	II	0.63	0.67	0.68	0.77	0.72	0.70	17	II	0.03	2	yes
23	NESM3	0.55	1.00	0.02	0.32	0.76	0.53	22	III	0.52	0.98	0.31	0.25	0.74	0.56	20	III	0.03	2	yes
24	NorESM2-LM	0.77	0.64	0.50	0.92	0.83	0.73	9	II	0.75	0.62	0.98	0.93	0.79	0.81	7	II	0.08	2	yes
25	NorESM2-MM	0.61	1.00	0.38	0.90	0.76	0.73	11	II	0.59	0.65	0.87	0.72	0.75	0.72	15	II	0.01	4	yes
																	Mean	0.12	6.7	yes-52%

- very good (I) - satisfactory (II) - unsatisfactory (III) - QUALITY GROUPS (QG)

I (2) II (7) III (3) / got into different groups (13) - number of models belong to each QG in two period

Fig. A.8. Results of normalized Taylor Diagram statistics, bias and trend differences for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (vi) of model comparison by Taylor Diagram, bias and trend (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980			Period 1981-2010			Total score diff.	Model rank diff.	Consistency (QG)
		Total score	Model rank	Quality group	Total score	Model rank	Quality group			
1	ACCESS-CM2	0.58	16	II	0.41	18	II	0.17	2	yes
2	ACCESS-ESM1-5	0.89	4	I	0.86	4	I	0.03	0	yes
3	AWI-CM-1-1-MR	0.95	3	I	0.86	5	I	0.08	2	yes
4	BCC-CSM2-MR	0.68	11	II	0.73	11	II	0.04	0	yes
5	CAMS-CSM1-0	0.47	21	III	0.27	23	III	0.20	2	yes
6	CanESM5	0.68	12	II	0.82	8	II	0.13	4	yes
7	CESM2-WACCM	0.84	6	I	0.23	24	III	0.61	18	no
8	CIESM	0.47	22	III	0.86	7	II	0.39	15	no
9	EC-Earth3	0.47	23	III	0.41	19	II	0.06	4	no
10	EC-Earth3-Veg	0.53	19	II	1.00	1	I	0.47	18	no
11	FGOALS-f3-L	0.21	24	III	0.36	20	III	0.15	4	yes
12	FGOALS-g3	0.00	25	III	0.00	25	III	0.00	0	yes
13	FIO-ESM-2-0	0.89	5	I	0.82	9	II	0.08	4	no
14	GFDL-ESM4	1.00	1	I	0.86	6	I	0.14	5	yes
15	INM-CM4-8	0.53	20	III	0.36	21	III	0.16	1	yes
16	INM-CM5-0	0.74	10	II	0.77	10	II	0.04	0	yes
17	IPSL-CM6A-LR	0.79	7	II	0.64	16	II	0.15	9	yes
18	KACE-1-0-G	0.58	17	II	0.68	14	II	0.10	3	yes
19	MIROC6	0.63	15	II	0.68	15	II	0.05	0	yes
20	MPI-ESM1-2-HR	0.79	8	II	0.91	3	I	0.12	5	no
21	MPI-ESM1-2-LR	1.00	2	I	0.95	2	I	0.05	0	yes
22	MRI-ESM2-0	0.68	13	II	0.64	17	II	0.05	4	yes
23	NESM3	0.58	18	II	0.32	22	III	0.26	4	no
24	NorESM2-LM	0.68	14	II	0.73	12	II	0.04	2	yes
25	NorESM2-MM	0.79	9	II	0.73	13	II	0.06	4	yes
							Mean	0.15	4.4	yes-72%

- very good (I) - satisfactory (II) - unsatisfactory (II) - QUALITY GROUPS (QG)

I (4) II (10) III (4) / got into different groups (7) - number of models belong to each QG in two periods

Fig. A.9. Results of normalized values of total scores for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 for method (vii) – Percentile-based method (see 2.2 Methods for model evaluation compared in the study)

№	Model acronym	Period 1951-1980			Period 1981-2010		
		Method i	Method iii	Method iv	Method i	Method iii	Method iv
		RMSE	TSS	S_{score}	RMSE	TSS	S_{score}
1	ACCESS-CM2	2.3	0.85	0.06	2.7	0.59	0.36
2	ACCESS-ESM1-5	0.7	0.07	0.26	0.6	0.81	0.60
3	AWI-CM-1-1-MR	0.8	0.31	0.53	1.1	0.90	0.63
4	BCC-CSM2-MR	1.4	0.47	0.58	1.4	0.82	0.64
5	CAMS-CSM1-0	2.2	0.07	0.07	2.6	0.20	0.32
6	CanESM5	1.0	0.45	0.26	0.7	0.61	0.63
7	CESM2-WACCM	0.7	0.17	0.61	0.7	0.76	0.65
8	CIESM	2.7	0.79	0.27	2.5	0.43	0.33
9	EC-Earth3	2.6	0.85	0.74	1.5	0.66	0.70
10	EC-Earth3-Veg	1.6	0.09	0.57	0.6	0.72	0.70
11	FGOALS-f3-L	3.0	0.45	0.08	2.4	0.66	0.58
12	FGOALS-g3	5.1	0.52	0.67	5.2	0.85	0.31
13	FIO-ESM-2-0	0.9	0.21	0.51	0.5	0.62	0.70
14	GFDL-ESM4	0.6	0.57	0.79	0.8	0.46	0.60
15	INM-CM4-8	1.6	0.79	0.01	1.9	0.23	0.36
16	INM-CM5-0	1.0	0.20	0.30	1.1	0.67	0.56
17	IPSL-CM6A-LR	1.0	0.25	0.07	1.2	0.76	0.12
18	KACE-1-0-G	2.3	0.52	0.33	1.9	0.99	0.63
19	MIROC6	1.4	0.05	0.56	1.4	0.41	0.12
20	MPI-ESM1-2-HR	0.9	0.07	0.51	0.7	0.56	0.56
21	MPI-ESM1-2-LR	0.5	0.80	0.27	0.6	0.71	0.57
22	MRI-ESM2-0	1.4	0.37	0.72	1.5	0.62	0.62
23	NESM3	2.0	0.57	0.61	2.2	0.77	0.63
24	NorESM2-LM	0.8	0.35	0.61	0.9	0.62	0.70
25	NorESM2-MM	1.7	0.76	0.31	1.8	0.64	0.62

- very good - satisfactory - unsatisfactory

Fig. A.10. Results of RMSE (method i), TSS (method iii) and S_{score} (method iv) for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010

№	Model acronym	Period 1951-1980			Period 1981-2010		
		r	Mean diff.	STD diff.	r	Mean diff.	STD diff.
1	ACCESS-CM2	0.20	0.014	0.001	0.36	0.009	0.002
2	ACCESS-ESM1-5	-0.20	0.044	0.011	0.47	0.037	0.005
3	AWI-CM-1-1-MR	0.03	0.010	0.011	0.16	0.008	0.002
4	BCC-CSM2-MR	0.22	0.010	0.012	-0.07	0.012	0.001
5	CAMS-CSM1-0	0.02	0.033	0.007	0.38	0.036	0.012
6	CanESM5	0.37	0.017	0.009	0.49	0.003	0.002
7	CESM2-WACCM	-0.42	0.010	0.000	0.29	0.017	0.006
8	CIESM	0.18	0.004	0.002	-0.16	0.045	0.008
9	EC-Earth3	-0.15	0.011	0.002	0.25	0.077	0.046
10	EC-Earth3-Veg	-0.23	0.071	0.016	0.61	0.006	0.004
11	FGOALS-f3-L	0.17	0.030	0.002	0.58	0.003	0.003
12	FGOALS-g3	0.18	0.005	0.011	0.08	0.021	0.005
13	FIO-ESM-2-0	0.40	0.033	0.023	0.43	0.009	0.005
14	GFDL-ESM4	0.28	0.023	0.010	0.13	0.010	0.005
15	INM-CM4-8	0.33	0.012	0.003	-0.22	0.033	0.001
16	INM-CM5-0	-0.24	0.019	0.003	0.26	0.007	0.013
17	IPSL-CM6A-LR	0.10	0.028	0.008	-0.01	0.001	0.004
18	KACE-1-0-G	-0.42	0.001	0.017	0.31	0.007	0.006
19	MIROC6	-0.14	0.029	0.000	0.39	0.022	0.003
20	MPI-ESM1-2-HR	0.29	0.038	0.000	-0.29	0.019	0.004
21	MPI-ESM1-2-LR	-0.04	0.001	0.000	0.05	0.007	0.009
22	MRI-ESM2-0	0.20	0.020	0.015	0.04	0.016	0.008
23	NESM3	0.46	0.049	0.030	0.03	0.058	0.008
24	NorESM2-LM	-0.38	0.005	0.000	-0.08	0.004	0.006
25	NorESM2-MM	0.06	0.007	0.005	0.37	0.021	0.005

- very good - satisfactory - unsatisfactory

Fig. A.11. Results of spatial trends statistics (r, Mean difference, STD difference) for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 - method ii

№	Model acronym	Period 1951-1980			Period 1981-2010		
		MR_{IVS}	MR_{Taylor}	MR_{mean}	MR_{IVS}	MR_{Taylor}	MR_{mean}
1	ACCESS-CM2	0.68	0.60	0.64	0.56	0.28	0.42
2	ACCESS-ESM1-5	0.64	0.55	0.59	0.12	0.60	0.36
3	AWI-CM-1-1-MR	0.88	0.69	0.79	0.96	0.77	0.87
4	BCC-CSM2-MR	0.32	0.48	0.40	0.48	0.53	0.51
5	CAMS-CSM1-0	0.48	0.27	0.37	0.04	0.07	0.05
6	CanESM5	0.44	0.51	0.47	0.84	0.55	0.69
7	CESM2-WACCM	0.24	0.44	0.34	0.24	0.19	0.21
8	CIESM	0.96	0.59	0.77	0.16	0.61	0.39
9	EC-Earth3	0.80	0.57	0.69	0.00	0.43	0.21
10	EC-Earth3-Veg	0.00	0.17	0.09	0.64	0.69	0.67
11	FGOALS-f3-L	0.36	0.31	0.33	0.76	0.45	0.61
12	FGOALS-g3	0.72	0.41	0.57	0.32	0.35	0.33
13	FIO-ESM-2-0	0.16	0.37	0.27	0.40	0.59	0.49
14	GFDL-ESM4	0.20	0.60	0.40	0.80	0.53	0.67
15	INM-CM4-8	0.76	0.65	0.71	0.20	0.19	0.19
16	INM-CM5-0	0.84	0.56	0.70	0.68	0.59	0.63
17	IPSL-CM6A-LR	0.40	0.43	0.41	0.92	0.69	0.81
18	KACE-1-0-G	0.52	0.47	0.49	0.72	0.65	0.69
19	MIROC6	0.56	0.33	0.45	0.36	0.31	0.33
20	MPI-ESM1-2-HR	0.92	0.55	0.73	0.60	0.52	0.56
21	MPI-ESM1-2-LR	0.60	0.77	0.69	0.44	0.67	0.55
22	MRI-ESM2-0	0.28	0.37	0.33	0.28	0.36	0.32
23	NESM3	0.04	0.43	0.23	0.08	0.37	0.23
24	NorESM2-LM	0.12	0.44	0.28	0.88	0.61	0.75
25	NorESM2-MM	0.08	0.44	0.26	0.52	0.40	0.46

- very good - satisfactory - unsatisfactory

Fig. A.12. Results of MR for Taylor and IVS metrics for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 - method v

№	Model acronym	Period 1951-1980					Period 1981-2010				
		RMSE	r	|dif_std|	|dif_T|	|B|	RMSE	r	|dif_std|	|dif_T|	|B|
1	ACCESS-CM2	2.26	0.33	0.06	0.01	2.13	2.66	0.59	0.08	0.01	2.51
2	ACCESS-ESM1-5	0.69	-0.28	0.08	0.04	0.18	0.63	0.77	0.29	0.04	0.12
3	AWI-CM-1-1-MR	0.76	0.02	0.04	0.01	0.94	1.09	0.76	0.00	0.01	1.39
4	BCC-CSM2-MR	1.36	0.17	0.16	0.01	0.67	1.43	0.73	0.12	0.01	0.77
5	CAMS-CSM1-0	2.24	-0.29	0.11	0.03	1.73	2.58	0.31	0.30	0.04	2.02
6	CanESM5	0.96	0.16	0.14	0.02	0.86	0.70	0.72	0.01	0.00	0.39
7	CESM2-WACCM	0.69	-0.08	0.17	0.01	0.72	0.69	0.49	0.21	0.02	0.80
8	CIESM	2.69	0.29	0.04	0.00	3.65	2.46	0.60	0.17	0.04	3.59
9	EC-Earth3	2.60	0.32	0.05	0.01	2.26	1.55	0.81	0.64	0.08	1.00
10	EC-Earth3-Veg	1.64	-0.13	0.44	0.07	1.08	0.60	0.67	0.07	0.01	0.24
11	FGOALS-f3-L	3.00	0.16	0.15	0.03	2.67	2.37	0.63	0.04	0.00	2.00
12	FGOALS-g3	5.14	0.17	0.07	0.01	6.62	5.16	0.76	0.20	0.02	6.68
13	FIO-ESM-2-0	0.94	-0.02	0.20	0.03	0.39	0.54	0.63	0.14	0.01	0.49
14	GFDL-ESM4	0.58	0.25	0.19	0.02	0.11	0.82	0.49	0.02	0.01	0.48
15	INM-CM4-8	1.56	0.30	0.05	0.01	1.82	1.91	0.28	0.16	0.03	2.14
16	INM-CM5-0	0.96	-0.08	0.04	0.02	0.81	1.09	0.64	0.06	0.01	0.96
17	IPSL-CM6A-LR	1.01	0.00	0.15	0.03	1.23	1.17	0.69	0.01	0.00	1.72
18	KACE-1-0-G	2.32	0.19	0.08	0.00	1.88	1.89	0.81	0.06	0.01	1.41
19	MIROC6	1.40	-0.36	0.09	0.03	1.05	1.36	0.47	0.15	0.02	1.03
20	MPI-ESM1-2-HR	0.86	-0.30	0.04	0.04	1.15	0.69	0.57	0.08	0.02	0.79
21	MPI-ESM1-2-LR	0.48	0.31	0.09	0.00	0.46	0.58	0.67	0.12	0.01	0.80
22	MRI-ESM2-0	1.45	0.11	0.17	0.02	1.63	1.54	0.63	0.21	0.02	1.79
23	NESM3	2.03	0.37	0.44	0.05	1.51	2.18	0.80	0.44	0.06	1.64
24	NorESM2-LM	0.78	0.11	0.20	0.00	1.03	0.89	0.60	0.01	0.00	1.32
25	NorESM2-MM	1.72	0.37	0.26	0.01	1.47	1.78	0.62	0.09	0.02	1.60

- very good - satisfactory - unsatisfactory

Fig. A.13. Results of Taylor Diagram statistics, bias and trend differences for 25 models over the Arctic for the period 1951-1980 and for the period 1981-2010 - method vi

Table A.14. Results of the CMIP6 model performance for SAT in the Arctic over the period 1951-1980 using the percentile-based method - method vii (RMSE - root-mean-square error, ^{0}C; r – correlation coefficient between models and reanalysis; CPI – climate prediction index; |dif_std| – modulus of standard deviation difference (model minus observations), ^{0}C; |Trm| - modulus of spatial trend mean difference (model minus observations), ^{0}C yr^{-1}; |Tra| - modulus of spatial trend amplitude difference (model minus observations), ^{0}C yr^{-1}; |Bm| - modulus of spatial bias mean difference (model minus observations), ^{0}C; |Ba| – modulus of spatial biases amplitude difference (model minus observations), ^{0}C)

ID	Model acronym	Seasonal variability (averaged over the territory)				Interannual variability (averaged over the territory)				Spatial variability			
		rmsd	r	CPI	|dif_std|	rmsd	r	CPI	|dif_std|	|Trm|	|Tra|	|Brm|	|Bra|
1	ACCESS-CM2	2.49	1.00	0.23	0.70	2.26	0.33	6.30	0.06	0.02	0.25	2.13	21.69
2	ACCESS-ESM1-5	0.99	1.00	0.09	0.20	0.69	-0.28	1.93	0.08	0.06	0.05	0.18	28.78
3	AWI-CM-1-1-MR	0.97	1.00	0.09	0.36	0.76	0.02	2.13	0.04	0.03	0.17	0.94	22.00
4	BCC-CSM2-MR	1.74	0.99	0.16	0.48	1.36	0.17	3.79	0.16	0.03	0.06	0.67	30.32
5	CAMS-CSM1-0	3.07	0.99	0.28	1.11	2.24	-0.29	6.24	0.11	0.05	0.06	1.73	25.97
6	CanESM5	1.30	1.00	0.12	0.35	0.96	0.16	2.68	0.14	0.03	0.16	0.86	28.05
7	CESM2-WACCM	1.17	1.00	0.11	0.92	0.69	-0.08	1.93	0.17	0.02	0.26	0.72	20.88
8	CIESM	2.98	0.99	0.27	0.27	2.69	0.29	7.49	0.04	0.01	0.28	3.65	19.95
9	EC-Earth3	2.72	1.00	0.25	0.13	2.60	0.32	7.24	0.05	0.02	0.23	2.26	26.14
10	EC-Earth3-Veg	1.58	1.00	0.14	0.28	1.64	-0.13	4.57	0.44	0.11	0.04	1.08	25.89
11	FGOALS-f3-L	3.32	1.00	0.30	1.45	3.00	0.16	8.36	0.15	0.05	0.26	2.67	25.17
12	FGOALS-g3	5.54	1.00	0.50	2.01	5.14	0.17	14.33	0.07	0.01	0.04	6.62	34.33
13	FIO-ESM-2-0	1.03	1.00	0.09	0.60	0.94	-0.02	2.62	0.20	0.05	0.11	0.39	21.31
14	GFDL-ESM4	0.77	1.00	0.07	0.18	0.58	0.25	1.63	0.19	0.02	0.10	0.11	21.85
15	INM-CM4-8	2.08	0.99	0.19	0.86	1.56	0.30	4.35	0.05	0.03	0.22	1.82	27.31
16	INM-CM5-0	1.35	1.00	0.12	0.37	0.96	-0.08	2.67	0.04	0.04	0.24	0.81	27.01
17	IPSL-CM6A-LR	1.13	1.00	0.10	0.54	1.01	0.00	2.82	0.15	0.04	0.21	1.23	23.00
18	KACE-1-0-G	2.64	1.00	0.24	1.07	2.32	0.19	6.47	0.08	0.01	0.08	1.88	23.25
19	MIROC6	1.42	1.00	0.13	0.23	1.40	-0.36	3.90	0.09	0.04	0.22	1.05	24.03
20	MPI-ESM1-2-HR	1.06	1.00	0.10	0.53	0.86	-0.30	2.41	0.04	0.05	0.23	1.15	20.04
21	MPI-ESM1-2-LR	0.67	1.00	0.06	0.36	0.48	0.31	1.35	0.09	0.01	0.20	0.46	23.88
22	MRI-ESM2-0	1.48	1.00	0.13	0.55	1.45	0.11	4.04	0.17	0.02	0.11	1.63	22.89
23	NESM3	2.36	1.00	0.21	0.95	2.03	0.37	5.67	0.44	0.07	0.01	1.51	25.80
24	NorESM2-LM	1.40	1.00	0.13	1.14	0.78	0.11	2.18	0.20	0.00	0.22	1.03	22.82
25	NorESM2-MM	1.72	1.00	0.16	0.00	1.72	0.37	4.80	0.26	0.00	0.11	1.47	21.40
	maximum	5.54	1.00	0.50	2.01	5.14	1.00	14.33	0.44	0.11	0.28	6.62	34.33
	75%	3.65	0.75	0.33	1.50	3.49	0.75	9.74	0.31	0.08	0.21	4.88	30.73
	50%	2.44	0.50	0.22	1.00	2.33	0.50	6.49	0.20	0.05	0.14	3.25	27.14
	25%	1.22	0.25	0.11	0.50	1.16	0.25	3.25	0.10	0.03	0.07	1.63	23.54
	minimum	0.67	0.00	0.06	0.00	0.48	0.00	1.35	0.04	0.00	0.01	0.11	19.95

Table A.15. Results of the CMIP6 model performance for SAT in the Arctic over the period 1981-2010 using the percentile-based method - method vii (RMSE - root-mean-square error, ^{0}C; r – correlation coefficient between models and reanalysis; CPI – climate prediction index; |dif_std| – modulus of standard deviation difference (model minus observations), ^{0}C; |Trm| - modulus of spatial trend mean difference (model minus observations), ^{0}C yr^{-1}; |Tra| - modulus of spatial trend amplitude difference (model minus observations), ^{0}C yr^{-1}; |Bm| - modulus of spatial bias mean difference (model minus observations), ^{0}C; |Ba| – modulus of spatial biases amplitude difference (model minus observations), ^{0}C)

ID	Model acronym	Seasonal variability (averaged over the territory)				Interannual variability (averaged over the territory)				Spatial variability			
		rmsd	r	CPI	|dif_std|	rmsd	r	CPI	|dif_std|	|Trm|	|Tra|	|Brm|	|Bra|
1	ACCESS-CM2	2.86	1.00	0.27	0.94	2.66	0.59	3.90	0.08	0.02	0.05	2.51	58.06
2	ACCESS-ESM1-5	0.94	1.00	0.09	0.28	0.63	0.77	0.93	0.29	0.04	0.09	0.12	55.90
3	AWI-CM-1-1-MR	1.36	1.00	0.13	0.56	1.09	0.76	1.60	0.00	0.00	0.06	1.39	60.97
4	BCC-CSM2-MR	1.74	1.00	0.16	0.50	1.43	0.73	2.10	0.12	0.00	0.02	0.77	60.91
5	CAMS-CSM1-0	3.31	1.00	0.31	1.23	2.58	0.31	3.79	0.30	0.05	0.00	2.02	59.10
6	CanESM5	0.73	1.00	0.07	0.24	0.70	0.72	1.03	0.01	0.01	0.02	0.39	67.82
7	CESM2-WACCM	1.18	1.00	0.11	0.83	0.69	0.49	1.00	0.21	0.01	0.02	0.80	61.63
8	CIESM	2.77	0.99	0.26	0.31	2.46	0.60	3.61	0.17	0.05	0.05	3.59	62.73
9	EC-Earth3	1.43	1.00	0.13	0.14	1.55	0.81	2.27	0.64	0.07	0.30	1.00	59.23
10	EC-Earth3-Veg	0.69	1.00	0.06	0.39	0.60	0.67	0.88	0.07	0.00	0.03	0.24	51.53
11	FGOALS-f3-L	2.65	1.00	0.25	1.20	2.37	0.63	3.48	0.04	0.00	0.05	2.00	71.32
12	FGOALS-g3	5.55	1.00	0.51	2.00	5.16	0.76	7.57	0.20	0.01	0.11	6.68	68.98
13	FIO-ESM-2-0	0.56	1.00	0.05	0.48	0.54	0.63	0.80	0.14	0.02	0.08	0.49	65.64
14	GFDL-ESM4	0.88	1.00	0.08	0.31	0.82	0.49	1.20	0.02	0.02	0.03	0.48	57.97
15	INM-CM4-8	2.27	1.00	0.21	0.97	1.91	0.28	2.80	0.16	0.04	0.06	2.14	70.04
16	INM-CM5-0	1.39	1.00	0.13	0.45	1.09	0.64	1.60	0.06	0.02	0.06	0.96	66.01
17	IPSL-CM6A-LR	1.35	1.00	0.13	0.49	1.17	0.69	1.72	0.01	0.01	0.04	1.72	65.80
18	KACE-1-0-G	2.15	1.00	0.20	0.96	1.89	0.81	2.77	0.06	0.01	0.01	1.41	62.17
19	MIROC6	1.41	1.00	0.13	0.22	1.36	0.47	2.00	0.15	0.03	0.04	1.03	57.84
20	MPI-ESM1-2-HR	0.82	1.00	0.08	0.36	0.69	0.57	1.01	0.08	0.03	0.03	0.79	60.29
21	MPI-ESM1-2-LR	0.77	1.00	0.07	0.51	0.58	0.67	0.85	0.12	0.01	0.04	0.80	55.22
22	MRI-ESM2-0	1.53	1.00	0.14	0.58	1.54	0.63	2.26	0.21	0.01	0.05	1.79	60.60
23	NESM3	2.57	0.99	0.24	1.00	2.18	0.80	3.19	0.44	0.06	0.09	1.64	62.79
24	NorESM2-LM	1.53	1.00	0.14	1.19	0.89	0.60	1.31	0.01	0.02	0.04	1.32	59.30
25	NorESM2-MM	1.77	1.00	0.16	0.13	1.78	0.62	2.61	0.09	0.03	0.03	1.60	60.45
	maximum	5.55	1.00	0.51	2.00	5.16	1.00	7.57	0.64	0.07	0.30	6.68	71.32
	75%	3.74	0.75	0.35	1.40	3.46	0.75	5.08	0.48	0.05	0.23	4.92	66.37
	50%	2.50	0.50	0.23	0.93	2.31	0.50	3.39	0.32	0.04	0.15	3.28	61.43
	25%	1.25	0.25	0.12	0.47	1.15	0.25	1.69	0.16	0.02	0.08	1.64	56.48
	minimum	0.56	0.00	0.05	0.13	0.54	0.00	0.80	0.00	0.00	0.00	0.12	51.53

№	Model acronym	Seasonal variablity				Interannual variability				Spatial variability				Total score	Rank
		rmsd	r	CPI	dif_std	rmsd	r	CPI	dif_std	|Trm|	|Tra|	|Bm|	|Ba|		
1	ACCESS-CM2	1	3	1	2	2	1	2	3	3	0	2	3	23	16
2	ACCESS-ESM1-5	3	3	3	3	3	0	3	3	1	3	3	1	29	4
3	AWI-CM-1-1-MR	3	3	3	3	3	0	3	3	2	1	3	3	30	3
4	BCC-CSM2-MR	2	3	2	3	2	0	2	2	2	3	3	1	25	11
5	CAMS-CSM1-0	1	3	1	1	2	0	2	2	2	3	2	2	21	21
6	CanESM5	2	3	2	3	3	0	3	2	2	1	3	1	25	12
7	CESM2-WACCM	3	3	3	2	3	0	3	2	3	0	3	3	28	6
8	CIESM	1	3	1	3	1	1	1	3	3	0	1	3	21	22
9	EC-Earth3	1	3	1	3	1	1	1	3	3	0	2	2	21	23
10	EC-Earth3-Veg	2	3	2	3	2	0	2	0	0	3	3	2	22	19
11	FGOALS-f3-L	1	3	1	1	1	0	1	2	2	0	2	2	16	24
12	FGOALS-g3	0	3	0	0	0	0	0	3	3	3	0	0	12	25
13	FIO-ESM-2-0	3	3	3	2	3	0	3	2	2	2	3	3	29	5
14	GFDL-ESM4	3	3	3	3	3	0	3	2	3	2	3	3	31	1
15	INM-CM4-8	2	3	2	2	2	1	2	3	2	0	2	1	22	20
16	INM-CM5-0	2	3	2	3	3	0	3	3	2	0	3	2	26	10
17	IPSL-CM6A-LR	3	3	3	2	3	0	3	2	2	0	3	3	27	7
18	KACE-1-0-G	1	3	1	1	2	0	2	3	3	2	2	3	23	17
19	MIROC6	2	3	2	3	2	0	2	3	2	0	3	2	24	15
20	MPI-ESM1-2-HR	3	3	3	2	3	0	3	3	1	0	3	3	27	8
21	MPI-ESM1-2-LR	3	3	3	3	3	1	3	3	3	1	3	2	31	2
22	MRI-ESM2-0	2	3	2	2	2	0	2	2	3	2	2	3	25	13
23	NESM3	2	3	2	2	2	1	2	0	1	3	3	2	23	18
24	NorESM2-LM	2	3	2	1	3	0	3	2	3	0	3	3	25	14
25	NorESM2-MM	2	3	2	3	2	1	2	1	3	2	3	3	27	9

Fig. A.16. Results of the percentile-based method with the final model score over the period 1951-1980 - method vii (Green color denotes a very good group, yellow – good, orange – satisfactory, and red – unsatisfactory group)

№	Model acronym	Seasonal variablity				Interannual variability				Spatial variability				Total score	Rank
		rmsd	r	CPI	dif_std	rmsd	r	CPI	dif_std	|Trm|	|Tra|	|Bm|	|Ba|		
1	ACCESS-CM2	1	3	1	1	1	2	1	3	2	3	2	2	22	18
2	ACCESS-ESM1-5	3	3	3	3	3	3	3	2	1	2	3	3	32	4
3	AWI-CM-1-1-MR	2	3	2	2	3	3	3	3	3	3	3	2	32	5
4	BCC-CSM2-MR	2	3	2	2	2	2	2	3	3	3	3	2	29	11
5	CAMS-CSM1-0	1	3	1	1	1	1	1	2	1	3	2	2	19	23
6	CanESM5	3	3	3	2	3	2	3	2	3	3	3	1	31	8
7	CESM2-WACCM	1	3	1	3	1	1	1	2	0	3	1	1	18	24
8	CIESM	3	3	3	3	3	2	3	3	3	3	3	0	32	6
9	EC-Earth3	2	3	2	3	2	3	2	0	0	0	3	2	22	19
10	EC-Earth3-Veg	3	3	3	3	3	2	3	3	3	3	3	3	35	1
11	FGOALS-f3-L	1	3	1	1	1	2	1	3	3	3	2	0	21	20
12	FGOALS-g3	0	3	0	0	0	3	0	2	3	2	0	0	13	25
13	FIO-ESM-2-0	3	3	3	2	3	2	3	3	3	2	3	1	31	9
14	GFDL-ESM4	3	3	3	3	3	1	3	3	2	3	3	2	32	7
15	INM-CM4-8	2	3	2	1	2	1	2	2	1	3	2	0	21	21
16	INM-CM5-0	2	3	2	3	3	2	3	3	2	3	3	1	30	10
17	IPSL-CM6A-LR	2	3	2	2	2	2	2	3	3	3	2	1	27	16
18	KACE-1-0-G	2	3	2	1	2	3	2	3	3	3	3	1	28	14
19	MIROC6	2	3	2	3	2	1	2	3	2	3	3	2	28	15
20	MPI-ESM1-2-HR	3	3	3	3	3	2	3	3	2	3	3	2	33	3
21	MPI-ESM1-2-LR	3	3	3	2	3	2	3	3	3	3	3	3	34	2
22	MRI-ESM2-0	2	3	2	2	2	2	2	2	3	3	2	2	27	17
23	NESM3	1	3	1	1	2	3	2	1	0	2	3	1	20	22
24	NorESM2-LM	2	3	2	1	3	2	3	3	2	3	3	2	29	12
25	NorESM2-MM	2	3	2	3	2	2	2	3	2	3	3	2	29	13

Fig. A.17. Results of the percentile-based method with the final model score over the period 1981-2010 - method vii (Green color denotes a very good group, yellow – good, orange – satisfactory, and red – unsatisfactory group)

A.18 REFERENCES TO TABLE A2

Aghakhani Afshar A., Hasanzadeh Y., Besalatpour A.A., and Pourreza-Bilondi M. (2017). Climate change forecasting in a mountainous data scarce watershed using CMIP5 models under representative concentration pathways. Theoretical and Applied Climatology, 129, 683-699, DOI:10.1007/s00704-016-1908-5.

Ahmed K., Sachindra D., Shahid S., et al. (2020). Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmospheric Research, 236, 104806, DOI:10.1016/j.atmosres.2019.104806.

Anandhi A. and Nanjundiah R.S. (2015). Performance evaluation of AR4 Climate Models in simulating daily precipitation over the Indian region using skill scores. Theoretical and Applied Climatology, 119, 551-566, DOI:10.1007/s00704-013-1043-5.

Anandhi A., Pierson D.C., and Frei A. (2019). Evaluation of Climate Model Performance for Water Supply Studies: Case Study for New York City. Journal of Water Resources Planning and Management, 145, 06019006, DOI:10.1061/(ASCE)WR.1943-5452.0001054.

Bannister D., Herzog M., Graf H.-F., et al. (2017). An Assessment of Recent and Future Temperature Change over the Sichuan Basin, China, Using CMIP5 Climate Models. Journal of Climate, 30, 6701-6722, DOI:10.1175/JCLI-D-16-0536.1.

Cai Z., You Q., Wu F., et al. (2021). Arctic Warming Revealed by Multiple CMIP6 Models: Evaluation of Historical Simulations and Quantification of Future Projection Uncertainties. Journal of Climate, 34, 4871-4892. DOI:10.1175/JCLI-D-20-0791.1.

Fu G., Liu Z., Charles S.P., et al. (2013). A score-based method for assessing the performance of GCMs: A case study of southeastern Australia. Journal of Geophysical Research-Atmospheres, 118, 4154-4167, DOI:10.1002/jgrd.50269.

Gnatiuk N., Radchenko I., Davy R., et al. (2020). Simulation of factors affecting Emiliania huxleyi blooms in Arctic and sub-Arctic seas by CMIP5 climate models: model validation and selection. Biogeosciences, 17, 1199-1212, DOI:10.5194/bg-17-1199-2020.

Herger N., Abramowitz G., Knutti R., et al. (2018). Selecting a climate model subset to optimise key ensemble properties. Earth System Dynamics, 9, 135-151, DOI:10.5194/esd-9-135-2018.

Inoue T. and Ueda H. (2011). Delay of the First Transition of Asian Summer Monsoon under Global Warming Condition. SOLA, 7, 81-84, DOI:10.2151/sola.2011-021.

Jia K., Ruan Y., Yang Y., and You Z. (2019). Assessment of CMIP5 GCM Simulation Performance for Temperature Projection in the Tibetan Plateau. Earth and Space Science, 6, 2362-2378, DOI:10.1029/2019EA000962.

Jiang Z., Li W., Xu J., and Li L. (2015). Extreme Precipitation Indices over China in CMIP5 Models. Part I: Model Evaluation. Journal of Climate, 28, 8603-8619, DOI:10.1175/JCLI-D-15-0099.1.

Kadel I., Yamazaki T., Iwasaki T., and Abdillah M. (2018). Projection of future monsoon precipitation over the central Himalayas by CMIP5 models under warming scenarios. Climate Research, 75, 1-21, DOI:10.3354/cr01497.

Kumar D., Mishra V., and Ganguly A.R. (2015). Evaluating wind extremes in CMIP5 climate models. Climate Dynamics, 45, 441-453, DOI:10.1007/s00382-014-2306-2.

Kumar S., Merwade V., Kinter J.L., and Niyogi D. (2013). Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations. Journal of Climate, 26, 4168-4185, DOI:10.1175/JCLI-D-12-00259.1.

Macadam I., Pitman A.J., Whetton P.H., and Abramowitz G. (2010). Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments. Geophysical Research Letters, 37, 16704, DOI:10.1029/2010GL043877.

Maxino C.C., McAvaney B.J., Pitman A.J., and Perkins S.E. (2008). Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. International Journal of Climatology, 28, 1097-1112, DOI:10.1002/joc.1612.

McMahon T.A., Peel M.C., and Karoly D.J. (2015). Assessment of precipitation and temperature data from CMIP3 global climate models for hydrologic simulation. Hydrology and Earth System Sciences, 19, 361-377, DOI:10.5194/hess-19-361-2015.

Ogata T., Ueda H., Inoue T., et al. (2014). Projected Future Changes in the Asian Monsoon: A Comparison of CMIP3 and CMIP5 Model Results. Journal of the Meteorological Society of Japan, 92, 207-225, DOI:10.2151/jmsj.2014-302.

Ongoma V., Chen H., Gao C. (2019). Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial East Africa. Theoretical and Applied Climatology, 135, 893-910, DOI:10.1007/s00704-018-2392-x.

Perkins S.E., Pitman A.J., Holbrook N.J., and McAneney J. (2007). Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. Journal of Climate, 20, 4356-4376, DOI:10.1175/JCLI4253.1.

Rao X., Lu X., and Dong W. (2019). Evaluation and projection of extreme precipitation over northern China in CMIP5 models. Atmosphere, 10, 691, DOI:10.3390/atmos10110691.

Reifen C. and Toumi R. (2009). Climate projections: Past performance no guarantee of future skill? Geophysical Research Letters, 36, DOI:10.1029/2009GL038082.

Ruan Y., Liu Z., Wang R., and Yao Z. (2019). Assessing the Performance of CMIP5 GCMs for Projection of Future Temperature Change over the Lower Mekong Basin. Atmosphere, 10, 93, DOI:10.3390/atmos10020093.

Saha A., Ghosh S., Sahana A.S., and Rao E.P. (2014). Failure of CMIP5 climate models in simulating post-1950 decreasing trend of Indian monsoon. Geophysical Research Letters, 41, 7323-7330, DOI:10.1002/2014GL061573.

Sharmila S., Joseph S., Sahai A., et al. (2015). Future projection of Indian summer monsoon variability under climate change scenario: An assessment from CMIP5 climate models. Global and Planetary Change, 124, 62-78, DOI:10.1016/j.gloplacha.2014.11.004.

Sillmann J., Kharin V.V., Zwiers F.W., et al. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. Journal of Geophysical Research-Atmospheres, 118, 2473-2493, DOI:10.1002/JGRD.50188.

Sun Q., Miao C., and Duan Q. (2015). Comparative analysis of CMIP3 and CMIP5 global climate models for simulating the daily mean, maximum, and minimum temperatures and daily precipitation over China. Journal of Geophysical Research-Atmospheres, 120, 4806-4824, DOI:10.1002/2014JD022994.

Walsh J.E., Chapman W.L., Romanovsky V., et al. (2008). Global Climate Model Performance over Alaska and Greenland. Journal of Climate, 21, 6156-6174, DOI:10.1175/2008JCLI2163.1.

Wang B., Liu D.L., Macadam I., et al. (2016). Multi-model ensemble projections of future extreme temperature change using a statistical downscaling method in south eastern Australia. Climatic Change, 138, 85-98, DOI:10.1007/s10584-016-1726-x.

Yang X., Yu X., Wang Y., et al. (2020). The Optimal Multimodel Ensemble of Bias-Corrected CMIP5 Climate Models over China. Journal of Hydrometeorology, 21, 845-863, DOI:10.1175/JHM-D-19-0141.1.

You Q., Jiang Z., Wang D., et al. (2018). Simulation of temperature extremes in the Tibetan Plateau from CMIP5 models and comparison with gridded observations. Climate Dynamics, 51, 355-369, DOI:10.1007/s00382-017-3928-y.

Zhou B., Wen Q.H., Xu Y., et al (2014). Projected Changes in Temperature and Precipitation Extremes in China by the CMIP5 Multimodel Ensembles. Journal of Climate, 27, 6591-6611, DOI:10.1175/JCLI-D-13-00761.1.