What aspect of model performance is the most relevant to skillful future projection on a regional scale?

Tong Li, Xuebin Zhang, & Zhihong Jiang

2024

Original citation:

Li, T., Zhang, X., & Jiang, Z. (2024). What aspect of model performance is the most relevant to skillful future projection on a regional scale? *Journal of Climate*, 37(5), 1567–1580. https://doi.org/10.1175/JCLI-D-23-0312.1

What Aspect of Model Performance is the Most Relevant to Skillful Future Projection on a Regional Scale?

TONG LI,[a,b,c] XUEBIN ZHANG,[c] AND ZHIHONG JIANG[a,b]

[a] *Key Laboratory of Meteorological Disaster of Ministry of Education (KLME), Nanjing University of Information Science and Technology, Nanjing, China*
[b] *Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disaster, Nanjing University of Information Science and Technology, Nanjing, China*
[c] *Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada*

(Manuscript received 31 May 2023, in final form 25 October 2023, accepted 10 December 2023)

ABSTRACT: Weighting models according to their performance has been used to produce multimodel climate change projections. But the added value of model weighting for future projection is not always examined. Here we apply an imperfect model framework to evaluate the added value of model weighting in projecting summer temperature changes over China. Members of large-ensemble simulations by three climate models of different climate sensitivities are used as pseudo-observations for the past and the future. Performance of the models participating in the phase 6 of the Coupled Model Intercomparison Project (CMIP6) are evaluated against the pseudo-observations based on simulated historical climatology and trends in global, regional, and local temperatures to determine the model weights for future projection. The weighted projections are then compared with the pseudo-observations in the future period. We find that regional trend as a metric of model performance yields generally better skill for future projection, while past climatology as performance metric does not lead to a significant improvement to projection. Trend at the grid-box scale is also not a good performance indicator as small-scale trend is highly uncertain. For the model weighting to be effective, the metric for evaluating the model's performance must be relatable to future changes, with the response signal separable from internal variability. Projected summer warming based on model weighting is similar to that of unweighted projection but the 5th–95th-percentile uncertainty range of the weighted projection is 38% smaller with the reduction mainly in the upper bound, with the largest reduction appearing in southeast China.

KEYWORDS: Climate change; Uncertainty; Ensembles; Climate models; Trends

1. Introduction

The Intergovernmental Panel on Climate Change in its Sixth Assessment Working Group II Report stated that "human-induced climate change, including more frequent and intense extreme events, has caused widespread adverse impacts and related losses and damages to nature and people" (IPCC 2022, p. 9). Climate change adaptation planning requires future climate change projections along with the quantification of associated uncertainty. Global climate models (GCMs) and Earth system models (ESMs) have played a crucial role in producing such projections. Simulations provided by GCMs and ESMs participating in successive phases of the Couple Model Intercomparison Project (CMIP) such as the latest phase 6 (CMIP6) driven by various emissions scenarios have provided a range of plausible future climate projections (Eyring et al. 2016; IPCC 2021). Their proper synthesis can provide a coherent projection.

Traditionally, a "democratic" approach (i.e., each model being given equal weight) has been used to synthesize multimodel projections. Projections by multimodel ensembles synthesized with this approach are more robust than those based on simulations by a single model (Eyring et al. 2019; Knutti 2010; Tebaldi and Knutti 2007). Different models can have different levels of complexity and as well as different approaches to the treatment of the same physical processes such as cloud and radiation. Because of this, models are not all equally skillful in simulating past climates. For this reason, efforts have been made to give different weights to the projections by individual models based on models' performance in a hope to produce more reliable future projection, approaches including the rank-based weighting scheme and the reliability ensemble averaging (REA) method, among others (Chen et al. 2011; Giorgi and Mearns 2003; Li et al. 2021). Existing multimodel ensembles such as those produced through the CMIPs are "ensembles of opportunity" and are not designed to systematically explore plausible model structures and epistemic uncertainty (Knutti 2010; Sanderson et al. 2015; Shiogama et al. 2022). Some models share components, making them not completely independent (Boé 2018). This aspect needs to be considered when synthesizing multimodel ensembles as well. Knutti et al. (2017) proposed the Climate Model Weighting by Independence and Performance (ClimWIP) scheme to take both model performance and independence into consideration when producing future projections. The method is explicit and has been widely used to project future

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-23-0312.s1.

Corresponding author: Zhihong Jiang, zhjiang@nuist.edu.cn

DOI: 10.1175/JCLI-D-23-0312.1

changes for a range of variables on global scale and for different regions, including for example global mean temperature (Liang et al. 2020), Arctic sea ice (Knutti et al. 2017), European temperature and precipitation (Brunner et al. 2019), and Chinese mean and extreme precipitation (Li et al. 2021).

The use of performance indicators to weight models generally involves two related assumptions: 1) confidence in a model is lower if the model simulates past climate less well and thus shall have lower weight, and 2) future projection produced with a model that better simulates past climate is more reliable (Knutti et al. 2013; Hall et al. 2019; Shiogama et al. 2022). While the first assumption is reasonable, the validity of the second assumption is not obvious. It is not always possible to test the validity of the second assumption because the future is not known. Thus, the performance of the models used in a weighting approach has been usually measured by comparing simulated past and present climates with the observed historical climate (Abramowitz et al. 2019; Bishop and Abramowitz 2012; Tebaldi and Knutti 2007). But it is unclear if the performance based on historical climate holds for future climate when making future projections (Knutti et al. 2010; Tebaldi and Knutti 2007). Different metrics have been used to evaluate models, resulting in differences in the level of model performance and thus different weighting schemes for the same set of models. For example, two dominant metrics—past climatology and past trend of a variable over a region—are both used in evaluating models and assigning model weights for future projection (Brunner et al. 2019; Liang et al. 2020). A model that simulates historical climatology the best may not simulate the historical trend equally well. Similarly, a model simulating a global scale trend well may not simulate trend for a region of interest well. It thus can be challenging to select a metric as the most suitable for a specific purpose (Knutti 2010). It can also be challenging to select the appropriate spatial scale to evaluate a model for the purpose of future projection. Evaluation on smaller spatial scale would be more affected by natural internal variability. Yet, evaluation on too large scale may not fully capture regionally important processes, such as the East Asian summer monsoon or feedbacks unique to the region.

The so-called "imperfect model test" or "model-as-truth test" provides an approach to estimate the skill for future projections. This approach uses one simulation that includes both past and future climates, conducted by a climate model, as pseudo "observed" past and future climates. The simulations of other climate models for the past climate are calibrated with the pseudo past climate, and those for the future climate are used to produce the future projection along with the calibration. As the pseudo future is "observed," it can then be used to compare with the projection to evaluate the skill of the calibration. Previous studies have used the concept of imperfect model test by applying it to individual runs from multiple models separately and then synthesizing the results across all available runs (Brunner et al. 2020; Herger et al. 2019). As simulations involved in these earlier studies typically had only a few runs by individual models, it became difficult to separate the inference of internal variability from structural differences among the models, and thus difficulties in interpreting the evaluation results (Frankcombe et al. 2018; Suarez-Gutierrez et al. 2021). In this regard, large-ensemble initial condition simulations have a unique advantage by providing many pseudo-observations (Deser et al. 2020; Milinski et al. 2020).

Here we conduct imperfect model tests with model performance being evaluated by two metrics—climatology and long-term trends—and on various spatial scales. This will enable exploring the effect of the use of different metrics on projection skill and identifying a more suitable spatial scale on which the models' performance should be evaluated for skillful future projection. To demonstrate the utility of our approach, we focus on summer mean temperature over China as high summer mean temperature is associated with a larger number and more severe summer heatwaves (Sun et al. 2014). The remainder of this paper is organized as follows: section 2 provides a detailed description of the datasets and methodology used in this study. The following section 3 shows the main results of the skill assessment and future projections. Finally, section 4 provides general conclusions and discussion.

2. Data and methods

a. Data used

1) CMIP6 SIMULATIONS

We use 204 simulations carried out by 25 models that participated in CMIP6. Table S1 in the online supplemental material summarizes the essential properties of all models and members. Among these, members from three large ensembles CanESM5 (50 members), EC-Earth3 (18 members), and MIROC6 (50 members) are used as the pseudo-observations for establishing model weighting schemes and for verification of the projection under the imperfect model test framework. These three models are selected for two reasons: 1) sufficient samples to estimate model response to external forcing and spread caused by internal variability and 2) a large range of climate sensitivity of the models, with climate sensitivity lying in the upper (CanESM5), the middle (EC-Earth3), and the bottom (MIROC6) of available CMIP6 models (as represented by historical trend and future warming in Figs. S1 and S2).

Monthly temperature data from the simulations forced by observed historical forcing and future emission scenario Shared Socioeconomic Pathway 5–8.5 (SSP5-8.5; O'Neill et al. 2014) are used. Historical simulations over 1971–2014 are used for model evaluation since the warming trend during this period is proven to be dominated by greenhouse gases (Liang et al. 2020; Tokarska et al. 2020). We focus on projected changes in the mid-twenty-first century (2041–60) relative to the 1995–2014 baseline. Model data come with different spatial resolutions, and they are interpolated onto a common $2.5° \times 2.5°$ grid using bilinear interpolation.

2) OBSERVATIONAL DATA

The observational gridded temperature dataset CN05.1 (Wu and Gao 2013) is used to evaluate the models' performance after we have identified the most relevant model weighting scheme. The monthly gridded dataset covers 1961–2015,

with a spatial resolution of 0.25° × 0.25°. We use the data from 1971 to 2014 to align with the model simulations, and regrid them to a resolution of 2.5° × 2.5° for model evaluation.

b. Imperfect model test framework

To investigate the effectiveness of a weighting scheme based on historical data for future projection, we conducted a series of model-as-truth tests within the framework of the imperfect model test. This process involves two steps: estimating models' weights based on their performance in simulating historical climate as well as independence among the models, and then evaluating the skill of the weighed projection by comparing it with the "pseudo" future observations. The components contributing to the imperfect model test are detailed below, including 1) the candidate metrics used for estimating models' performance, 2) the weighting scheme adopted in this work, 3) the candidate spatial scales of trend for regional projection, 4) the approach for synthesizing projections when using multiple models and runs, and 5) the three-aspect skill scores for projection assessment.

1) METRICS FOR ESTIMATING DISTANCE

The models' performance and independence are assessed by a distance measure based on suitably constructed metrics. As we focus on summer mean temperature, all metrics are calculated from this variable. The metrics under consideration include the following: (i) spatial distribution of summer temperature climatology (on a 2.5° × 2.5° grid, referred to as the climatology metric below), (ii) trend in global or regional mean summer temperature (referred to as the trend metric), (iii) spatial distribution of trend on a 2.5° × 2.5° grid (referred to as the trend pattern metric), and (iv) the combination of the climatology and trend metrics that is referred to as the composite metric. Trends are estimated based on nonparametric Sen's slope estimator, with the consideration of lag-1 autocorrelation (Zhang et al. 2000). For the composite metric, the distance between two models or between a model and the observation is the average of the (i) climatology distance and (ii) regional trend distance.

2) WEIGHTING METHOD CLIMWIP

We follow the ClimWIP approach for determining models' weights. This method was proposed by Knutti et al. (2017) based on Sanderson et al. (2015) and has been used in many studies (e.g., Amos et al. 2020; Liang et al. 2020; Merrifield et al. 2020). The basic idea is that a model that agrees more poorly with observations and that also largely duplicates existing models gets less weight (Knutti et al. 2017). The weight w_i is assigned to the model i according to the following equation:

$$w_i = e^{-(D_i/\sigma_d)^2} \Bigg/ \left[1 + \sum_{j \neq i}^{M} e^{-(S_{ij}/\sigma_s)^2}\right], \tag{1}$$

where D_i is the distance between the model i and the observation, and S_{ij} is the distance between the model i and model j. For all the metrics used, the independence distance S_{ij} is always computed based on spatial distribution of climatology following Merrifield et al. (2020), as the root-mean-square difference of climatological values across all grids in China region. But D_i is computed differently according to the metrics being used. When climatology or trend pattern metric is used, the distance D_i is the root-mean-square difference of climatological or trend values for all grids within the spatial domain, respectively. When the trend metric is used, the distance D_i is the absolute difference of the trends. For both D_i and S_{ij}, the raw distances are normalized separately by dividing the raw distances by their respective medium values. Figure S3 shows the model–model distances S_{ij} normalized by its median. When a model has multiple runs, we use the ensemble mean to compute the weights. This has the advantage of reducing the influence of internal variability, in particular when trend metric is used.

In Eq. (1), M is the number of models participating weighting (24 in the imperfect model tests), and σ_d and σ_s are shape parameters that represent the strength of performance of individual models and independence among models. A larger σ_d leads to more equal weighting among models, and a larger σ_s means models are treated to be more dependent. The appendix provides more details about these parameters and their estimation.

3) SPATIAL SCALE FOR MODEL EVALUATION

When a model's performance is evaluated on different spatial scales, the results can be different. As we will show later, the use of model trend as a performance metric improves the projection skills. For this reason, we pay close attention to the influence of spatial scale when the trend metric is used. Specifically, we consider four spatial scales of the trend metric: (i) global: trend in global mean summer temperature (referred to as the global trend metric); (ii) regional: trends in China mean summer temperature (referred to as the regional trend metric, which is the same as the trend metric mentioned earlier); (iii) subregional: trends in East China mean summer temperature and in West China mean summer temperature, with the 105°E [see Li et al. (2021) separating East China from West China (referred to as the subregional trend metric)]; and (iv) grid box: trends in summer mean temperature at individual grids (referred to as the grid trend metric). When model weighting is determined on a spatial large scale in the cases of (i) to (iii), each model is assigned one set of weights that applies to all grids within the domain. In the case of (iv), each grid box has its own set of weights.

4) SYNTHESIZING PROJECTION FROM MULTIPLE MODELS

We compare the main distributional characteristics including mean and spread of the multimodel weighted projection with the "known" future projection by the large ensembles. The spread in the large ensemble represents the influence of internal variability. To construct a weighted mixture distribution from the multiple model weighted projection, Herger et al. (2019) fitted future projections from subset models into separate probability distributions and then sampled the distributions so that the number of samples generated from subset models is proportional to weights. As the number of runs

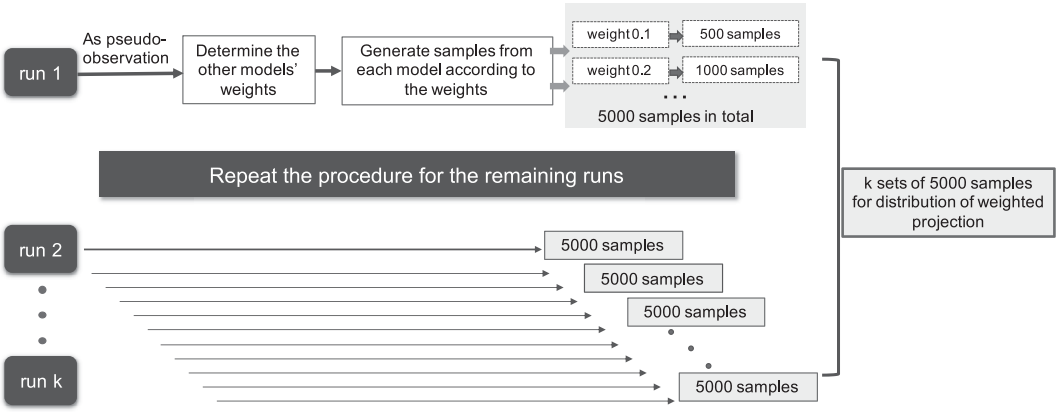

FIG. 1. Schematic diagram of the entire process of the sampling procedure used to generate mixture projection distributions.

from CMIP6 models can be quite limited, instead of fitting distributions for individual models separately, we sample the model projection directly. Figure 1 illustrates the sampling procedure. To describe our procedure, in the following, we use CanESM5 large ensemble as an example.

We initially take the first run of CanSEM5 large ensemble as pseudo-observation, then compare its historical simulation with those simulated by other 24 CMIP6 models to determine their respective weights using the ClimWIP method. Following this, we sample the available projection runs of the CMIP6 models (with replacement) for 5000 times, ensuring that the proportion of the sample from an individual model is equal to its weight. This procedure is repeated for the remaining 49 model runs in turn. Once completing the sampling process, we obtain 50×5000 samples, which are then used to estimate the distributional characteristics. These estimates are finally compared with the future projection produced by the 50 ensemble members of CanESM5. It should be noted that when producing future projections for summer mean temperature averaged over China, we sample the national mean values. When producing future projections for individual grid boxes, we sample the two-dimensional spatial maps of projected changes by individual model runs to maintain the spatial structure of temperature changes.

5) SKILL SCORES

The skill of the multimodel weighted projection against the unweighted projection is assessed on three aspects: (i) bias, (ii) difference in the width of the distribution, and (iii) the similarity between probability distributions. We examine whether weighted projection improves upon unweighted projection against the "known" future as simulated by the large ensembles. In all cases, a positive skill score indicates an improvement by the weighted ensemble projection.

(i) Bias

This compares the absolute bias between the median values of multimodel ensemble projection against the "known" projection by the large ensembles. The skill score is defined as following:

$$\text{Bias skill score} = |\text{Bias}_{\text{unweighted}}| - |\text{Bias}_{\text{weighted}}|. \tag{2}$$

(ii) Width

The difference between the 5th percentile and the 95th percentile of the projection is used to represent the uncertainty range ($\text{Value}_{95\text{th}} - \text{Value}_{5\text{th}}$). We calculate the absolute values of the width difference between the multimodel ensemble projection against the "known" projection by the large ensembles first, expressed as $|\Delta\text{Width}|$. We then compute the width skill according to the following equation:

$$\text{Width skill score} = |\Delta\text{Width}_{\text{unweighted}}| - |\Delta\text{Width}_{\text{weighted}}|. \tag{3}$$

(iii) Similarity between probability distributions

This measures how similar two probability density functions (PDFs) are. Perkins et al. (2007) proposed the use of the area where two PDFs overlap. A larger area of overlap indicates better agreement between the two PDFs. A perfect match of the PDF would give a value of one. The calculation involves dividing the PDFs into multiple bins and counting the number of occurrences in each bin. The smaller value of the occurrence from the two PDFs represents the portion of the overlap. Mathematically, this is expressed as

$$\text{S-score} = \sum_{1}^{n} \text{minimum}(Z_s, Z_o), \tag{4}$$

where n is the number of bins for which we use 50. The terms Z_s and Z_o represent the frequency in each bin from the weighted/unweighted projection and from the observation, respectively. This statistic has an advantage over other statistics used for comparing two distributions such as the statistic used in the K–S test as it is more robust against sampling errors and the number of bins used in computing the statistic. The

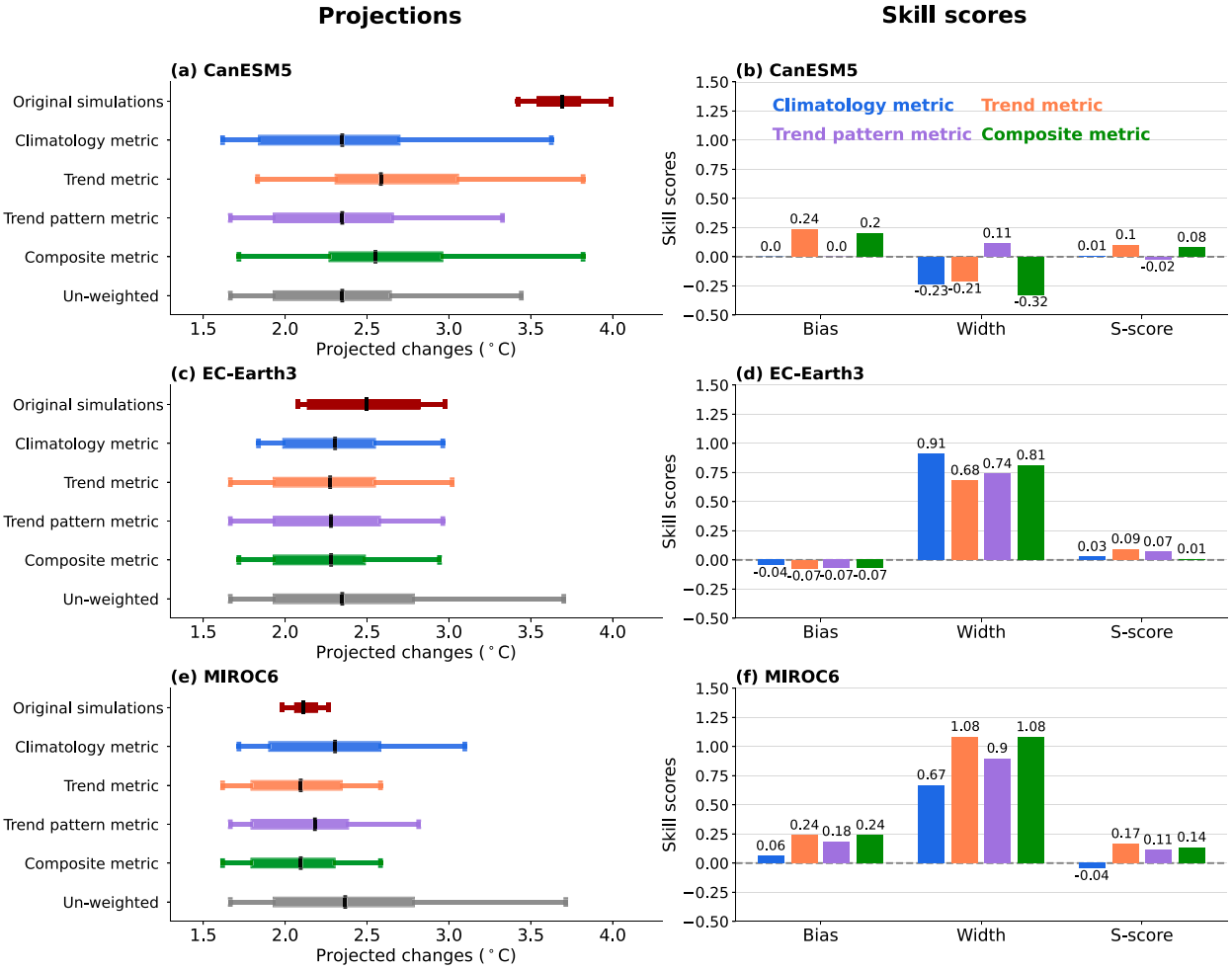

FIG. 2. Multimodel projections and their target projections for changes in China's summer mean temperature (°C) during 2041–60. (left) The median (black ticks), the 25th–75th percentiles (boxes), and the 5th–95th percentiles (whiskers) with pseudo-observations produced by three different models. (right) The skill scores of weighted projections. In all cases, a positive skill score indicates an improvement by the weighted ensemble projection.

bin size n will of course influence the S-score but as long as n is the same across calculations the final conclusion about the model performance will not be impacted. The S-score skill score is computed according to the following equation:

$$\text{S-score skill score} = \text{S-score}_{\text{weighted}} - \text{S-score}_{\text{unweighted}}. \quad (5)$$

c. *Influence of sampling uncertainty in observed trends on future projection*

As we will show later, our imperfect model analyses indicate that the regional mean temperature trend is the most skillful metric for weighting models and as a result, we will use the observed regional mean temperature trend to weight CMIP6 models for future projection. The trend estimated from the observation is subject to sampling uncertainty (or internal climate variability), and this sampling uncertainty is accounted for when weighting the models. To do so, we assume the temperature time series contains a linear trend and a red noise that follows a first-order autoregressive process. Red noise is often used when assessing statistical significance of trends (e.g., von Storch 1995; von Storch and Zwiers 1999; Zhang et al. 2000). We then use the method of Zhang et al. (2000) to estimate the trend and its uncertainty. We found the best estimate of the summer mean temperature trend over China during 1971–2014 to be 0.27°C decade^{-1}, with a standard deviation of 0.043°C decade^{-1}.

The best estimate and the standard deviation of the trend are used to specify a Gaussian distribution with which we generated a set of 100 trend values as possible observed trend. We then use these trend values to produce constrained future projections, as we did for the three large ensemble models.

3. Results

a. *Skills of the regional climatology and trend metric*

The left panel of Fig. 2 presents the projected summer mean temperature over China by the middle of the twenty-first century with simulations of the large ensembles by the three models

as pseudo-observations. Multimodel ensemble projections are constructed by using equal weighting (i.e., unweighted) or optimal weighting according to the four different metrics. As expected, different metrics consider different aspects, resulting in different projections. However, the regional trend metric enhances the consistency between ensemble projections and the pseudo-observations (of the future) to a greater extent than the climatology or trend pattern metrics.

While climatology metric has been widely used, as a default metric, to evaluate models' performance, our results show that the weighted projection based on the performance in reproducing historical climatology alone does not necessarily lead to an improvement on projection especially regarding bias and distribution similarity. The bias skill scores of the climatology metric are the smallest when warming in the pseudo-observations is much higher or lower than that in the simulations by the CMIP6 multimodel ensemble simulations, such as that simulated by CanESM5 or by MIROC6. This indicates that climatology metric has little capability to reduce model bias in warming, possibly because present-day climate conditions are not strongly correlated to the magnitude of warming or climate sensitivity (Herger et al. 2018; Knutti et al. 2010; Sanderson et al. 2017). This is also shown in Herger et al. (2018): "there is very little improvement (hardly any RMSE improvement) to be gained by constraining the climatology in terms of out-of-sample skill" (p. 146).

Weighted projections based on the performance of reproducing past trend are generally more accurate, with better agreement in the magnitude of projected changes and higher skill scores in bias and S-score. The improvement over unweighted projection is especially clear when simulations by the high sensitivity model CanESM5 or low sensitivity model MIROC6 are used as pseudo-observations. When simulations of CanESM5 are pseudo-observations, the unweighted CMIP6 ensemble projection could not reproduce the large magnitude of warming simulated by CanESM5. The projection has a cold bias of 1.34°C. The weighted ensemble based on trend reduces the bias to about 1.1°C, leading to positive skill scores of bias and S-score of 0.24 and 0.1, respectively. This indicates that while model weighting can reduce bias, there is a limit in the improvement when the bias is large. In the case of MIROC6 as pseudo-observation, the weighted projection shifts the value downward when compared to the unweighted projection, with the median value closer to the target median value and a considerable reduction in the uncertainty range. These indicate that the trend metric, which has a high correlation with future warming, has the effect of reducing bias and uncertainty.

In contrast, when the trend pattern (spatial distribution of trend) is used as a metric, it offers little improvement to the projection, far less than that offered by the regional trend metric. This is due to the fact that local trends are more affected by natural variability, which renders the spatial distribution of trends of little use for model evaluation.

The composite metric has been used to provide a comprehensive evaluation of models and to avoid overconfidence in model weighting (Lorenz et al. 2018; Merrifield et al. 2020). As half of the metric is not highly related to future warming and the other half is directly connected to it, the effect of the metric on enhancing projections lies somewhere between the effects of its two components.

We note that the weighting does reduce uncertainty range regardless of the metrics being used when simulations by EC-Earth3 or MIROC6 models are used as pseudo-observations. We also note that weighting models does not always reduce uncertainty when compared with unweighted projection in the case where simulations by CanESM5 are used as pseudo-observations. This is to a large extent due to the fact that the warming rate of CanESM5 is more an outlier compared with other models. Given that the observed change is not such an outlier when compared with model simulations, we expect real-world applications of model weighting to reduce uncertainty in projection.

As many applications require local-scale projection, we now present various skill measures at the grid box scale. Figures 3–5 show the skill scores computed at the grid box level when simulations by the three models are used as pseudo-observations. Overall, these skill scores resemble those computed for the national mean temperatures. When the simulations by CanESM5 and MIROC6 are the pseudo-observations (Figs. 3 and 5), weighting models based on regional trend and composite metrics show substantial reduction in bias and improvement in matching the probability distribution as indicated by mostly positive S-score. By comparison, there was less pronounced effect across the whole region when the climatology metric was used to weight models, indicating that better performance in simulating present-day climatology does not guarantee better future projections. The use of the trend pattern metric for model weighting offers some improvement regarding the bias and S-score, but the improvement is smaller than the regional trend metric. When the simulations by EC-Earth3 are the pseudo-observations (Fig. 4), weighting the models based on any metric does not affect the bias or the S-score, but the width of uncertainty range is reduced.

The skill scores are not uniform over the space. For instance, when performance in reproducing the regional trend is used to weight models and the simulations of CanESM5 are pseudo-observations, notably better skill scores can be seen in the northeastern region while the scores in the lower reach of the Yellow River Basin can be close to zero or even negative (Fig. 3j). In the MICRO6 case, negative skill scores can be seen in the Tibetan Plateau area and parts of Northwest China (Figs. 5j,l). These grid box skill scores should be interpreted in the context that projection on local scale is inherently more uncertain.

b. Proper spatial scale of the trend metric

Having identified that the trend metric is more effective for model weighting, we now look at the connection between the spatial scales of the trend and corresponding skill scores of the weighted projection. Figure 6 displays the results when simulations by CanESM5 are the pseudo-observations and when temperature trends over the globe, over China, over West and East China, or at the gridbox scale are used as the model's performance metric. In general, trends on different spatial scales as metrics for model weighting do improve the

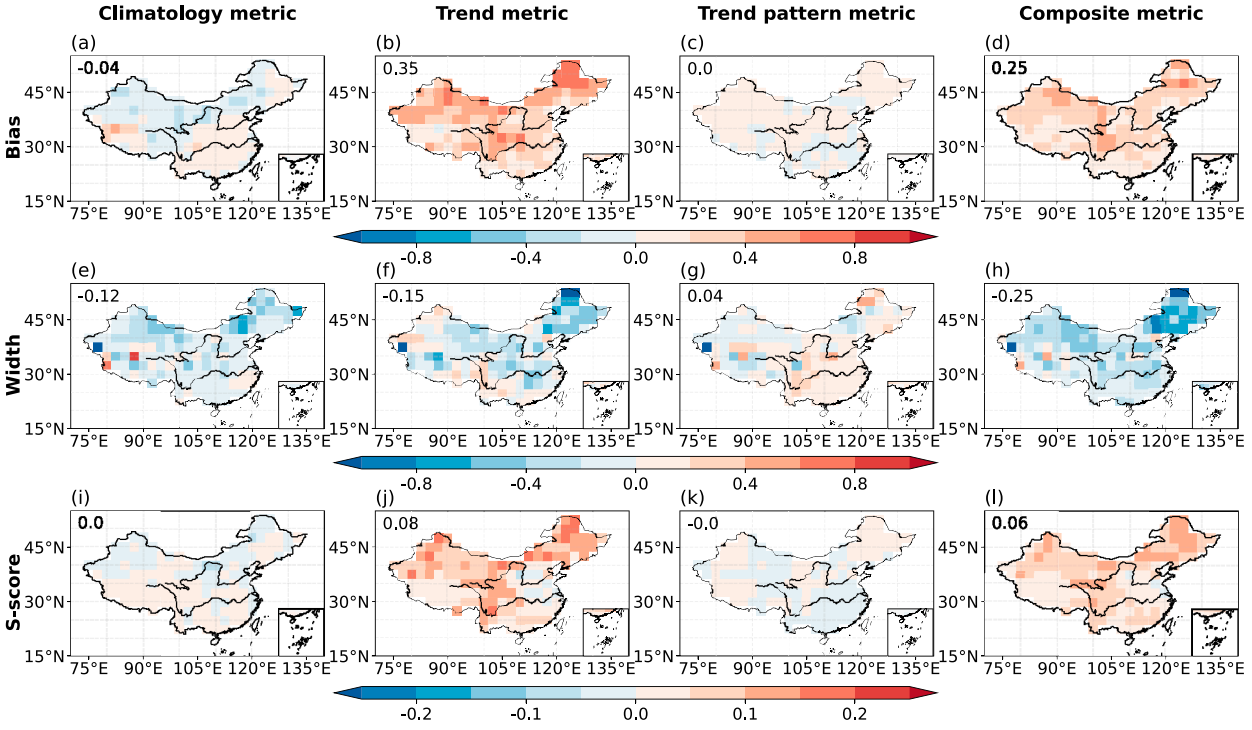

FIG. 3. Spatial patterns of the skill scores corresponding to different model performance metrics. The skill scores include (a)–(d) bias, (e)–(h) width of uncertainty, and (i)–(l) S-score of weighted projections relative to unweighted projection when the simulations by CanESM5 are pseudo-observations. While the skill scores are computed for individual grids separately, the weights for every grid for the same model are the same. The numbers in the top-left corners in each panel show the median value of the skill scores within the spatial domain.

projection, but their corresponding skill scores can be quite different. When the global mean temperature trend is the performance metric, the S-score skill scores show large spatial differences with large positive scores in West China and smaller or negative scores in Southeast China (Fig. 6i). When the trend in the mean temperature over China is the performance metric, improvement in the projection is more uniform across the country, especially in terms of reduction in bias and in matching the probability distribution although improvement in the uncertainty range is small (second column in Fig. 6). Results for the regional trends in the West and East China mean temperature as a performance metric are similar to those of China mean temperature trends, although skill scores are slightly smaller overall (third column in Fig. 6). When the temperature trends on grid box scale are used as a performance measure, the skill scores indicate overall improvement for the weighted projection when compared to unweighted projection, but the improvement is very minimal (last column in Fig. 6).

Figure 7 shows the results when simulations by EC-Earth3 are the pseudo-observations. As the projections by the EC-Earth3 are already in the middle of the projections by the available CMIP6 models, there is not much room for improvement regarding bias and S-scores. But the use of regional trends as a performance metric for model weighting does reduce uncertainty in the projection. On the contrary, the use of global temperature trend as a metric for model weighting increases projection uncertainty. Figure 8 presents the results when simulations by MICRO6 are the pseudo-observations. The use of global, regional, and subregional trends as metrics improve the projection by reducing bias, uncertainty, and improving distributional match across the country, relatively uniformly. The use of grid box trend as a metric again results in little improvement.

Overall, the results suggest that the level of uncertainty in the trend estimate and representativeness of trend for the region of interest play an important role in the performance of model weighting. When grid box scale historical trend is used as a performance metric it offers some improvement than unweighted projection in China region, but the improvement is small. This is because local trend is highly affected by natural internal variability and thus its estimate is also highly uncertain. When models are weighted according to their performance in simulating the historical global mean temperature trend, the effect on projection for different regions can be quite different, indicating that the global-scale temperature trend may not reflect important regional process or feedback well. When the models are weighted based on their performance in simulating regional trend over China or large subregions of the country, future projections on both national and grid box scales are improved, and the improvement is generally consistent regardless the targets and across the space.

c. Future projection of China's summer temperature

As the regional mean temperature trend is generally the most effective performance metric, multimodel projection for

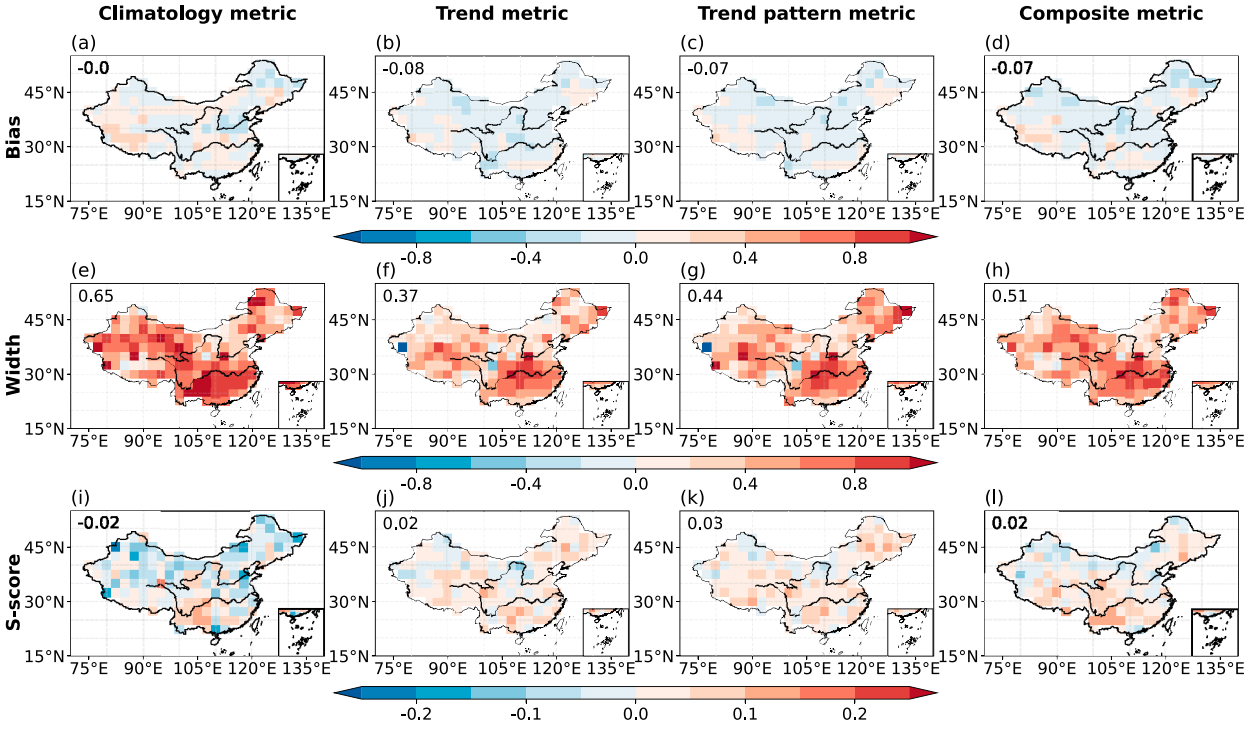

FIG. 4. As in Fig. 3, but for the simulations by EC-Earth3 as pseudo-observations for the projection.

summer temperature over China is produced based on this metric. Figure 9 shows the time series of projected national mean summer temperature under the SSP5–8.5 scenario. From the figure, it appears that the weighted multimodel projection shows a weaker warming when compared with the unweighted projection, especially in the upper-95th-percentile bound. There is also some reduction in the lower-5th-percentile bound of the weighted projection, but it is lower than the upper

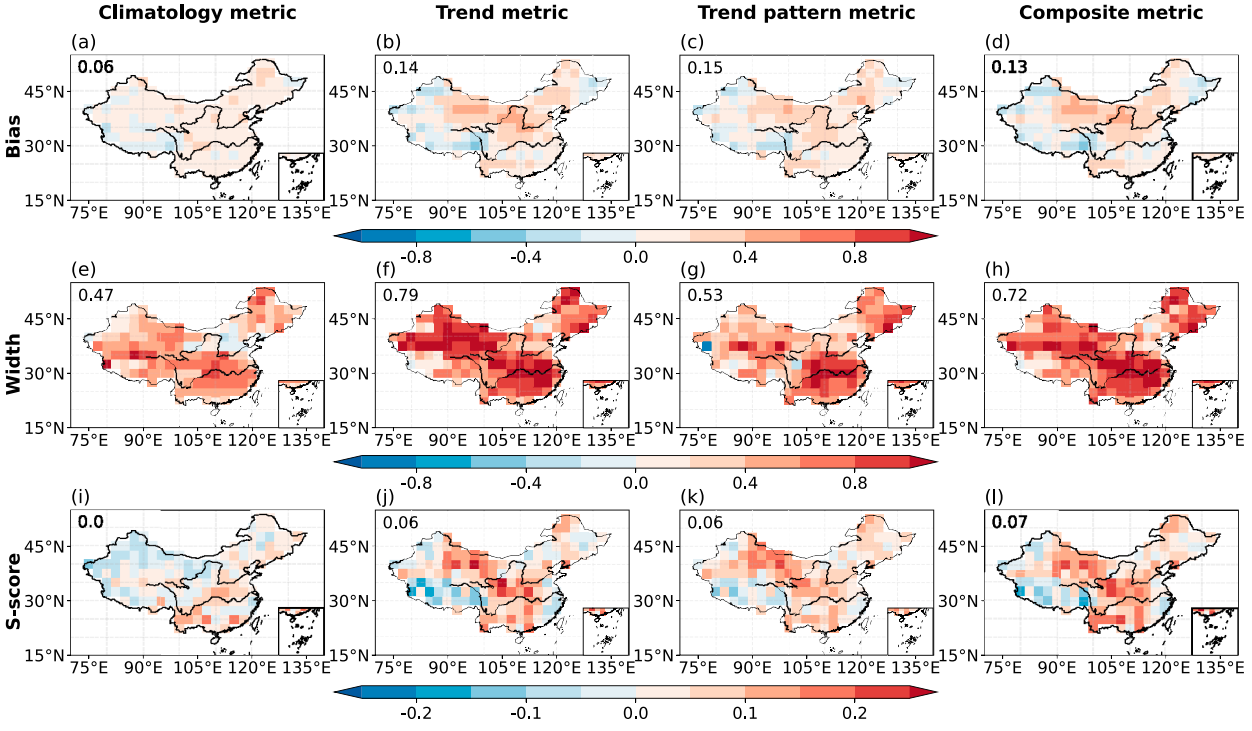

FIG. 5. As in Fig. 3, but for the simulations by MIROC6 as pseudo-observations for the projection.

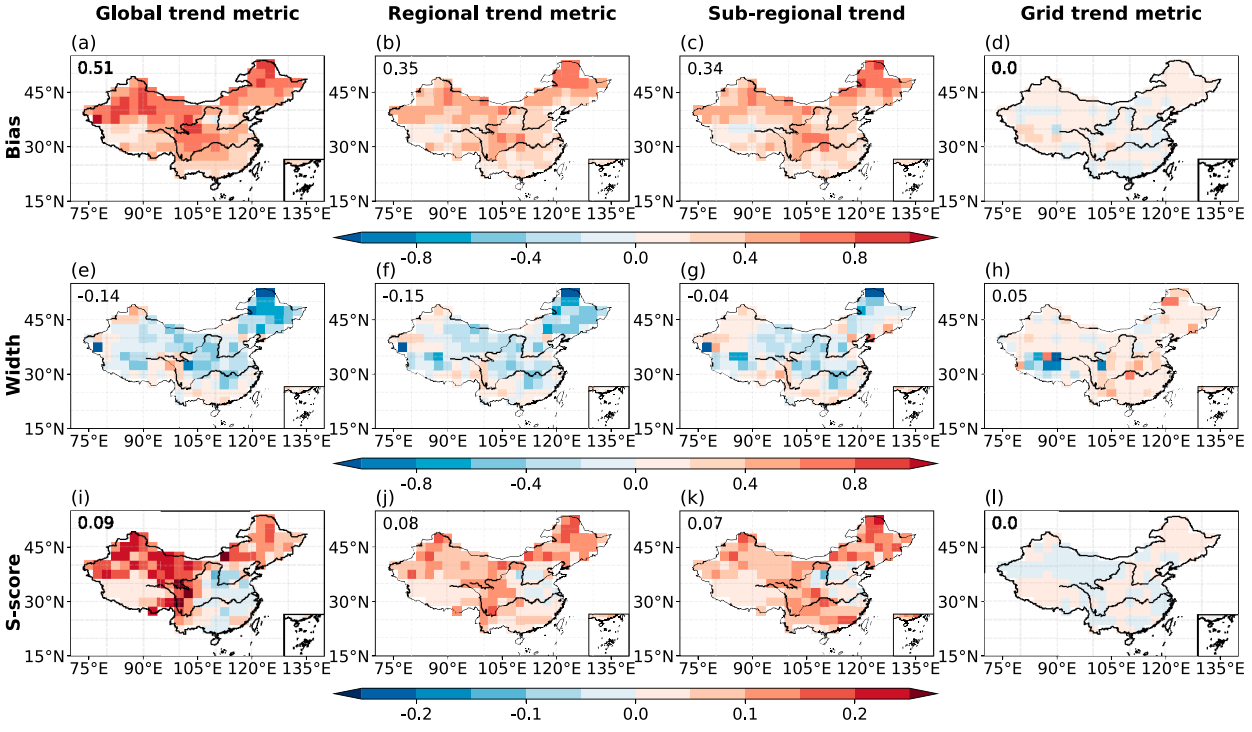

FIG. 6. As in Fig. 3 but for trends on different spatial scales as metrics and when the simulations by CanESM5 are the pseudo-observations. The weights for every grid depend on the metrics being used. The skill scores are computed for individual grids separately.

bound, resulting in a narrower uncertainty range of the weighted projection. This result is consistent with previous studies that used other methods to address the "hot tail" CMIP6 model problem, as the contribution of models that have relatively higher sensitivity is reduced (IPCC 2021; Nijsse et al. 2020; Ribes et al. 2021; Tokarska et al. 2020; Hu et al. 2022; Hausfather et al. 2022).

Focusing on the warming in specific time periods, such as the midcentury and end of century, the above comparison between weighted and unweighted projections becomes more apparent (Fig. 10). In the mid-twenty-first century, the weighted multimodel ensemble projects a median increase of 2.3°C over the 1995–2014 base period, with a 5th–95th-percentile range of 1.67°–2.9°C. When compared with the unweighted multimodel projection, the weighted projection shows a slightly lower median value by about 0.04°C, and a 0.77°C (38%) reduction in the uncertainty range, particularly as a reduction in the upper bound. In the end of the century, the differences between weighted and unweighted projections are more pronounced, with the weighted ensemble projecting a warming of 5.27°C, while the unweighted ensemble projects 5.44°C. The uncertainty range of the weighted ensemble is 3.41°–7.22°C, which is lower than the unweighted range by 0.87°C (23%).

Figure 11 displays the median and 5th and 95th percentiles of the weighted projection and its difference from the unweighted projection on grid box scale in the mid-twenty-first century. Warming is widespread over the entire region and for all percentiles, with the spatial median value of 2.31°C for the median projection and 1.52° and 3.2°C for the 5th and 95th percentiles, respectively. Larger warming occurs in the northern high-latitude regions, especially in Northwest China, with the median increase of up to 2.75°C and the 95th-percentile warming more than 4.0°C. The magnitude of warming decreases from northwest to southeast, with a median warming as small as 1.75°C in Southeast China.

Compared with the unweighted projection, there is little difference in the median value of the projected change, though the weighted projection tends to be slightly cooler with a spatial median value of about 0.12°C (Fig. 11b). The difference in the 5th-percentile projection is even smaller, with values mostly within 0.2°C across the whole region. In contrast, the 95th percentile is much reduced in the weighted projection, with the largest reduction of more than 0.8°C in the Southeast China, parts of the Tibetan Plateau region, and Northeast China, which are also the region with the largest uncertainty (Fig. S4d). Due to the reduction in the 95th percentile, the 5th–95th-percentile uncertainty range is also reduced by at least 0.4°C.

4. Conclusions and discussion

In this study, we examined the skills of model weighting based on various model performance metrics in producing summer temperature projections over China. We considered models' performance in reproducing the observed historical climatology and trends on various spatial scales, including the global, the regional, and grid-box scales as bases for model weighting. We estimated model weighting skills using large-ensemble simulations by three climate models of different climate sensitivities.

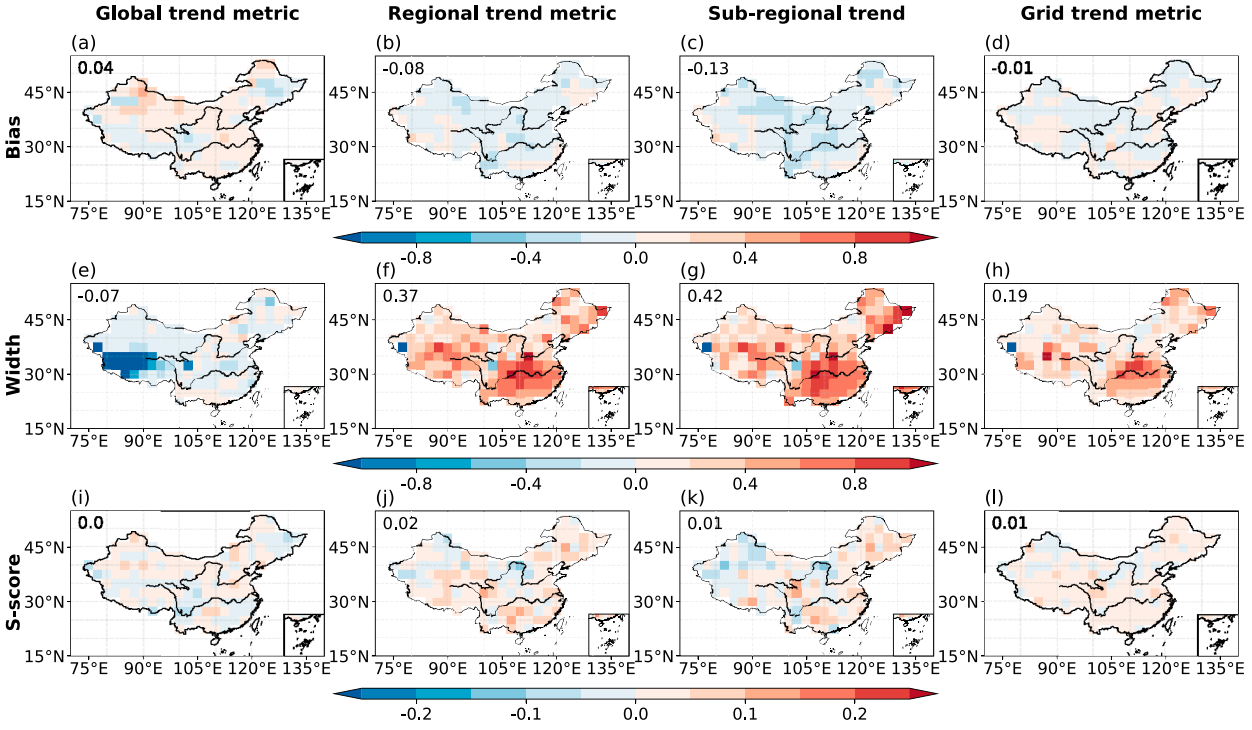

FIG. 7. As in Fig. 6, but for the simulations by EC-Earth3 as pseudo-observations for the projection.

Our results clearly demonstrate that model weighting has added values over unweighted (or equal weighting) if a proper metric is used to evaluate the model's performance. We indicate clearly that using trends in the mean temperature over China or subregions of China for model weighting yields better results, leading to more consistent projections with the pseudo-observations. Changes in temperature are mostly the results of thermodynamic effect of global warming, it thus

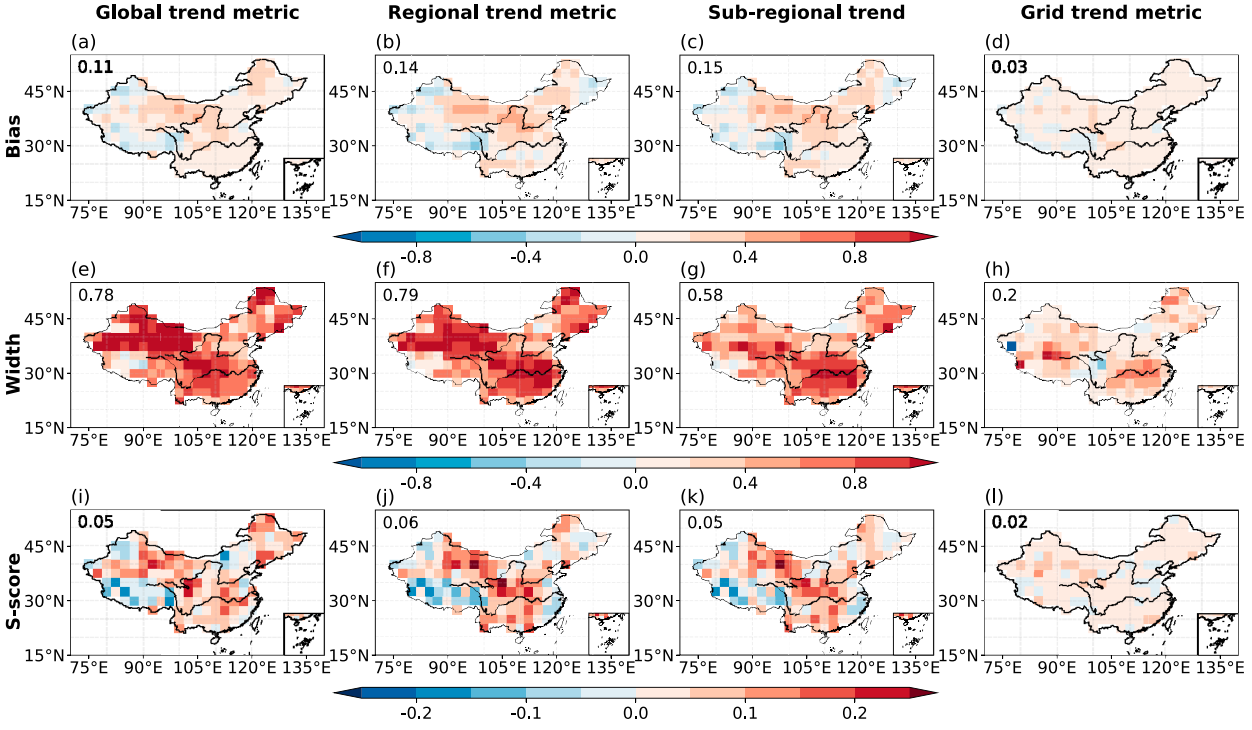

FIG. 8. As in Fig. 6, but for the simulations by MIROC6 as pseudo-observations for the projection.

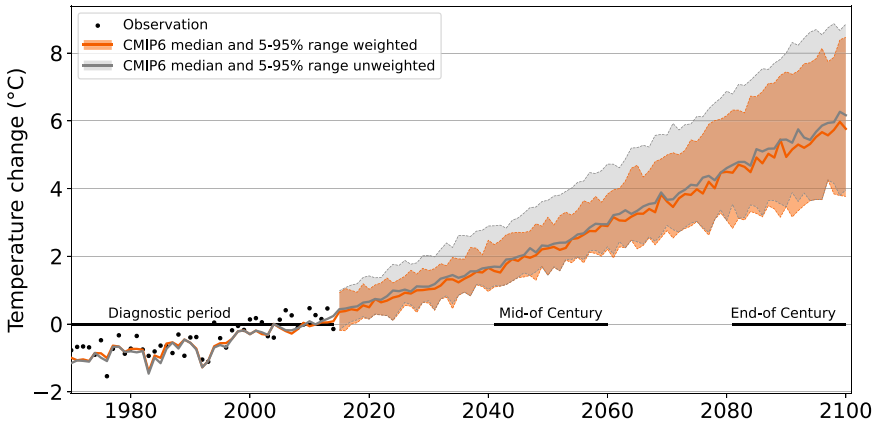

FIG. 9. Time series of summer temperature change (°C) in China region under the SSP5–8.5 scenario (relative to 1995–2014). The solid lines are the median value from weighted (orange) and unweighted (gray) ensembles; the shading represent the 5%–95% ranges.

makes sense for historical temperature trend on relatively large scale to be a relevant performance metric that reflect climate change signal. A large body of literature points that the observed long-term warming trend in China region can be robustly attributed to anthropogenic forcing and thus reflects the climate change signal, which means that the historical trend can well represent the forced response for both the model world and real world (Li et al. 2020; Sun et al. 2021).

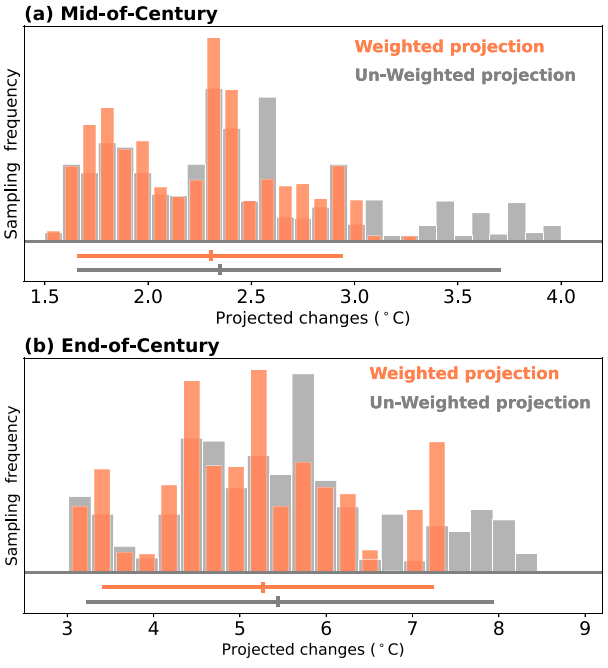

FIG. 10. Histogram for projected changes in summer temperature over China for (a) the middle of the twenty-first century (2041–60) and (b) the end of the twenty-first century (2081–2100) relative to the 1995–2014 base period. The histogram shading shows the sampling frequency distribution. The lines at the bottom mark the 5%–95% ranges, with the median values marked by the vertical ticks.

When the model's performance is evaluated based on a popular metric climatology, there is no solid evidence of improvement in the prediction bias as there is not a clear and strong relationship between models' climate sensitivity and historical climatology. Clearly, the model evaluation needs to fit for the particular purpose and should have a clear physical basis.

Spatial scale on which model's performance is evaluated also plays a role. The trend in regional mean temperature over China or subregions of China seems to perform better. The projection on gridbox scale exhibits regional difference when using the trend in global mean temperature, suggesting that some regionally important processes and feedbacks may not be well represented in the large-scale trend metric. Grid box scale trends offer little improvement in the projection over China region, suggesting that the noisy nature of trends on such fine spatial scale does not provide useful information for selecting better-performing models.

For the model weighting to be effective, the metric for evaluating the model's performance must meet two conditions. 1) The metric should well represent models' forced response, with signal separable from internal variability. This way, models' behavior is evaluated against climate response rather than noise. 2) The metric must be relatable to future changes of the variable of interest. As we have demonstrated that the historical trend in summer mean temperature over China is effective as a metric for model weighting for the purpose of projecting summer mean temperature in the future, it is possible to use this metric to produce weighted projection for different aspects of heatwaves, as the frequency, the magnitude, and the duration of heatwaves are closely related to summer mean temperature (Sun et al. 2014). It may also be feasible to weight the model based on this metric to project future changes in extreme precipitation of short duration due to the connection between atmospheric moisture and temperature.

By weighting the CMIP6 models based on their performance in simulating the observed summer temperature trend in China, we project that summer temperature in China will increase by about 2.3°C with the 5th–95th-percentiles range of 1.67°–2.9°C, by the middle of the twenty-first century (2041–60).

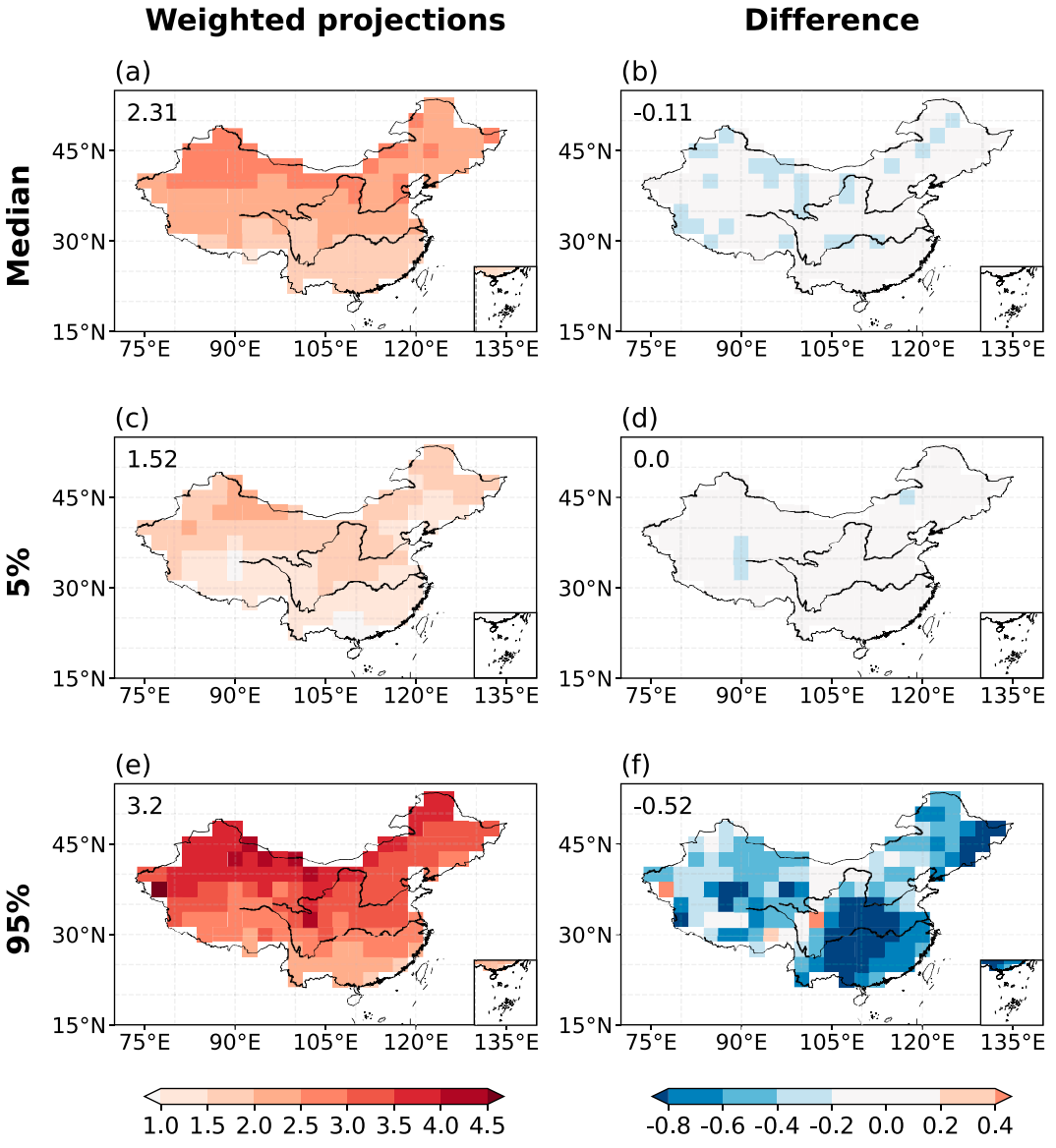

FIG. 11. Weighted projection and its difference from unweighted projection for summer mean temperature changes (°C) for the middle of the twenty-first century. The median and the 5th and the 95th percentiles of (left) weighted projection and (right) the difference are shown. The numbers in the top-left corner in each panel show (left) the median value of the warming and (right) the difference within the domain.

Compared with the unweighted projection, the median and the 5th percentile change little, but the 95th percentile is reduced by 0.77°C. This is in line with some studies that suggest climate sensitivities in some CMIP6 models to be too high (Sherwood et al. 2020; Hausfather et al. 2022). The weighted projection has a smaller uncertainty range compared with that of the unweighted projection, with a reduction of 38%. A larger reduction in the uncertainty is observed in the Southeast China and parts of Northeast China region, with a magnitude as large as 0.4°C. As the model weighting scheme has shown to be effective in a set of imperfect model tests, as well as we have considered the influence of internal variability on regional observed trend and used all available runs of models to generate the future projection, the confidence about the reduction in uncertainty is high.

Acknowledgments. We acknowledge Lukas Brunner and Ruth Lorenz for publishing their weighting code. This research was supported by the National Natural Science Foundation of China (Grant 42275184) and the National Key Research and Development Program of China (Grant 2017YFA0603804), and the Postgraduate Research and Practice Innovation Program of Government of Jiangsu Province (KYCX21_0940).

Data availability statement. The CMIP6 model data that support the findings of this study are openly available at the

following URL: https://esgf-node.llnl.gov/search/cmip6/. The high-quality in situ dataset (CN05.1) is available through Wu and Gao (2013).

APPENDIX

The Shape Parameters in the ClimWIP Method

Shape parameters σ_d and σ_s play crucial roles in capturing the strength of model performance and independence; σ_d regulates the degree of model performance on the weights, while σ_s regulates how model similarities against each other are balanced. A larger value of σ_d leads to a more uniform weighting of models, while a small value of σ_d leads to a more aggressive weighting. The σ_s determines the typical distance between two models considered to be similar. Increasing σ_s allows for two models to be farther apart and still be considered almost identical. More details and discussions of the method can be found in Knutti et al. (2017), Lorenz et al. (2018), and Brunner et al. (2020).

To estimate the two shape parameters, we utilize the same approach as Knutti et al. (2017) and Lorenz et al. (2018); it is also an imperfect model test setup but using all available models as "pseudo-observations." For every model as "truth," we calculate the weighted projection for $\sigma_d \times \sigma_s$ combinations (More exactly, we varied the σ_d and σ_s within the range of 0.2–0.8, with the step of 0.04, like what is shown in Fig. S5). After repeating this process M times (each model having been selected to represent the "truth" once), we choose the parameter combinations so that the model as truth lies within the 5th–95th-percentile range at least 80% of the time. This inside ratio 80% threshold was commonly chosen to avoid overconfidence in the weighting and ensure a reasonably broad range. Since the effect of σ_d on the ratio and the final result is larger than the σ_s value (Knutti et al. 2017; Brunner et al. 2020), we prioritize selecting a minimum value for σ_d. In this work, we use three large ensembles as pseudo-observations, for each of them and each of the metrics, we conduct the process of selecting the best shape parameters. For the case that climatology metric is used as performance metric, the results of inside ratio for varying parameters σ_d and σ_s are shown in Fig. S5. The optimized shape parameters for all cases are summarized in Table S2.

Generally, when the trend metric is used, larger values of σ_d are selected compared to when other metrics are used. As for σ_s, it determines the typical distance between two models considered to be similar. One standard to check its suitability is by comparing it with the intermember distances and intermodel distances to see if it is somewhere between these two. In the case of the climatology metric, the value of σ_s is selected as 0.44 for all three large ensembles cases. In the context of the ensemble used, the median of the generalized distance between pair models is about 1.04, while the median of the generalized distance between pair members of CanESM5, EC-Earth3, and MIROC6 are 0.21, 0.12, and 0.07, respectively. Therefore, the selected σ_s value is suitable as a typical distance to separate the intermodel and intermember distances.

REFERENCES

Abramowitz, G., and Coauthors, 2019: ESD reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.*, **10**, 91–105, https://doi.org/10.5194/esd-10-91-2019.

Amos, M., and Coauthors, 2020: Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence. *Atmos. Chem. Phys.*, **20**, 9961–9977, https://doi.org/10.5194/acp-20-9961-2020.

Bishop, C. H., and G. Abramowitz, 2012: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.*, **41**, 885–900, https://doi.org/10.1007/s00382-012-1610-y.

Boé, J., 2018: Interdependency in multimodel climate projections: Component replication and result similarity. *Geophys. Res. Lett.*, **45**, 2771–2779, https://doi.org/10.1002/2017GL076829.

Brunner, L., R. Lorenz, M. Zumwald, and R. Knutti, 2019: Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.*, **14**, 124010, https://doi.org/10.1088/1748-9326/ab492f.

——, A. G. Pendergrass, F. Lehner, A. L. Merrifield, R. Lorenz, and R. Knutti, 2020: Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth Syst. Dyn.*, **11**, 995–1012, https://doi.org/10.5194/esd-11-995-2020.

Chen, W., Z. Jiang, and L. Li, 2011: Probabilistic projections of climate change over China under the SRES A1B scenario using 28 AOGCMs. *J. Climate*, **24**, 4741–4756, https://doi.org/10.1175/2011JCLI4102.1.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

——, and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat. Climate Change*, **9**, 102–110, https://doi.org/10.1038/s41558-018-0355-y.

Frankcombe, L. M., M. H. England, J. B. Kajtar, M. E. Mann, and B. A. Steinman, 2018: On the choice of ensemble mean for estimating the forced signal in the presence of internal variability. *J. Climate*, **31**, 5681–5693, https://doi.org/10.1175/JCLI-D-17-0662.1.

Giorgi, F., and L. O. Mearns, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.*, **30**, 1629, https://doi.org/10.1029/2003GL017130.

Hall, A., P. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nat. Climate Change*, **9**, 269–278, https://doi.org/10.1038/s41558-019-0436-6.

Hausfather, Z., K. Marvel, G. A. Schmidt, J. W. Nielsen-Gammon, and M. Zelinka, 2022: Climate simulations: Recognize the 'hot model' problem. *Nature*, **605**, 26–29, https://doi.org/10.1038/d41586-022-01192-2.

Herger, N., G. Abramowitz, R. Knutti, O. Angélil, K. Lehmann, and B. M. Sanderson, 2018: Selecting a climate model subset

to optimise key ensemble properties. *Earth Syst. Dyn.*, **9**, 135–151, https://doi.org/10.5194/esd-9-135-2018.

——, ——, S. Sherwood, R. Knutti, O. Angélil, and S. A. Sisson, 2019: Ensemble optimisation, multiple constraints and overconfidence: A case study with future Australian precipitation change. *Climate Dyn.*, **53**, 1581–1596, https://doi.org/10.1007/s00382-019-04690-8.

Hu, D., D. Jiang, Z. Tian, and X. Lang, 2022: How skillful was the projected temperature over China during 2002–2018? *Sci. Bull.*, **67**, 1077–1085, https://doi.org/10.1016/j.scib.2022.02.004.

IPCC, 2021: *Climatic Change 2021: The Physical Science Basis*. Cambridge University Press, 2391 pp.

——, 2022: Summary for policymakers. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*, H.-O. Pörtner et al., Eds., Cambridge University Press, 3–33, https://doi.org/10.1017/9781009325844.001.

Knutti, R., 2010: The end of model democracy? *Climatic Change*, **102**, 395–404, https://doi.org/10.1007/s10584-010-9800-2.

——, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1.

——, D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, https://doi.org/10.1002/grl.50256.

——, J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.*, **44**, 1909–1918, https://doi.org/10.1002/2016GL072012.

Li, C., Y. Sun, F. Zwiers, D. Wang, X. Zhang, G. Chen, and H. Wu, 2020: Rapid warming in summer wet bulb globe temperature in China with human-induced climate change. *J. Climate*, **33**, 5697–5711, https://doi.org/10.1175/JCLI-D-19-0492.1.

Li, T., Z. Jiang, L. Zhao, and L. Li, 2021: Multi-model ensemble projection of precipitation changes over China under global warming of 1.5 and 2°C with consideration of model performance and independence. *J. Meteor. Res.*, **35**, 184–197, https://doi.org/10.1007/s13351-021-0067-5.

Liang, Y., N. P. Gillett, and A. H. Monahan, 2020: Climate model projections of 21st century global warming constrained using the observed warming trend. *Geophys. Res. Lett.*, **47**, e2019GL086757, https://doi.orgs/10.1029/2019GL086757.

Lorenz, R., N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti, 2018: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.*, **123**, 4509–4526, https://doi.org/10.1029/2017JD027992.

Merrifield, A. L., L. Brunner, R. Lorenz, I. Medhaug, and R. Knutti, 2020: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth Syst. Dyn.*, **11**, 807–834, https://doi.org/10.5194/esd-11-807-2020.

Milinski, S., N. Maher, and D. Olonscheck, 2020: How large does a large ensemble need to be? *Earth Syst. Dyn.*, **11**, 885–901, https://doi.org/10.5194/esd-11-885-2020.

Nijsse, F. J. M. M., P. M. Cox, and M. S. Williamson, 2020: Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth Syst. Dyn.*, **11**, 737–750, https://doi.org/10.5194/esd-11-737-2020.

O'Neill, B. C., E. Kriegler, K. Riahi, K. L. Ebi, S. Hallegatte, T. R. Carter, R. Mathur, and D. P. van Vuuren, 2014: A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Climatic Change*, **122**, 387–400, https://doi.org/10.1007/s10584-013-0905-2.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Climate*, **20**, 4356–4376, https://doi.org/10.1175/jcli4253.1.

Ribes, A., S. Qasmi, and N. P. Gillett, 2021: Making climate projections conditional on historical observations. *Sci. Adv.*, **7**, eabc0671, https://doi.org/10.1126/sciadv.abc0671.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1.

——, M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geosci. Model Dev.*, **10**, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017.

Sherwood, S. C., and Coauthors, 2020: An assessment of Earth's climate sensitivity using multiple lines of evidence. *Rev. Geophys.*, **58**, e2019RG000678, https://doi.org/10.1029/2019RG000678.

Shiogama, H., M. Watanabe, H. Kim, and N. Hirota, 2022: Emergent constraints on future precipitation changes. *Nature*, **602**, 612–616, https://doi.org/10.1038/s41586-021-04310-8.

Suarez-Gutierrez, L., S. Milinski, and N. Maher, 2021: Exploiting large ensembles for a better yet simpler climate model evaluation. *Climate Dyn.*, **57**, 2557–2580, https://doi.org/10.1007/s00382-021-05821-w.

Sun, Y., X. Zhang, F. W. Zwiers, L. Song, H. Wan, T. Hu, H. Yin, and G. Ren, 2014: Rapid increase in the risk of extreme summer heat in eastern China. *Nat. Climate Change*, **4**, 1082–1085, https://doi.org/10.1038/nclimate2410.

——, ——, Y. Ding, D. Chen, D. Qin, and P. Zhai, 2021: Understanding human influence on climate change in China. *Natl. Sci. Rev.*, **9**, nwab113, https://doi.org/10.1093/nsr/nwab113.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **A365**, 2053–2075, https://doi.org/10.1098/rsta.2007.2076.

Tokarska, K. B., M. B. Stolpe, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti, 2020: Past warming trend constrains future warming in CMIP6 models. *Sci. Adv.*, **6**, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549.

von Storch, H., 1995: Misuses of statistical analysis in climate research. *Analysis of Climate Variability: Applications of Statistical Techniques*. Springer-Verlag Berlin, 11–26.

——, and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Wu, J., and X.-J. Gao, 2013: A gridded daily observation dataset over China region and comparison with the other datasets (in Chinese). *Chin. J. Geophys.*, **56**, 1102–1111, https://doi.org/10.6038/cjg20130406.

Zhang, X., L. A. Vincent, W. D. Hogg, and A. Niitsoo, 2000: Temperature and precipitation trends in Canada during the 20th century. *Atmos.–Ocean*, **38**, 395–429, https://doi.org/10.1080/07055900.2000.9649654.