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

Tong Li[1,2], Xuebin Zhang[3], Zhihong Jiang[1*]

[1]Joint International Research Laboratory of Climate and Environment Change, Key Laboratory of Meteorological Disaster of Ministry of Education (KLME), Nanjing University of Information Science and Technology, Nanjing, China, [2]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disaster, Nanjing University of Information Science and Technology, Nanjing, China, [3]Climate Research Division, Environment and Climate Change Canada, Toronto, ON, Canada

Corresponding author: Zhihong Jiang (zhjiang@nuist.edu.cn)

Key Points:

- The improvement of weighted projection is tied to the level the model's performance metric represents climate change signal.
- Regional temperature trend as a metric yields the most added value for model weighting, but the climatology is not a useful metric.
- Weighted projection of summer warming over China for the mid-century is similar to unweighted, with 18% smaller uncertainty range.

Abstract

Weighting models according to their performance has been used in constructing multi-model regional climate change scenarios. But the added value of model weighting is not always examined. Here we apply an imperfect model framework to examine the added value of model weighting in projecting summer temperature changes over China. Members of large ensemble initial condition simulations by three climate models of different climate sensitivities under the historical forcing and future scenarios are used as pseudo-observations for the past and the future. Performance of the models participating in the 6^{th} phase of the coupled model intercomparison project (CMIP6) in simulating past climate are evaluated against the pseudo-observations based on climatology, trends in global, regional and local temperatures. The performance along with model's independence are used to determine the model weights for future projection. The weighted projections are then compared with the pseudo-observations for the future. We find that regional trend as a metric of model performance yields the best skill for future projection while past climatology as performance metric does not improve future projection. Trend at the grid-box scale is also not a good performance indicator as small scale trend is highly uncertain. Projected summer warming based on model weighting is similar to that of unweighted

projection, at 2.3°C increase relative to 1995-2014 by the middle of the 21st century under SSP8.5 scenario, but the 5th-95th uncertainty range of the weighted projection is 18% smaller with the reduction mainly in the upper bound, with the largest reduction in the northern Tibetan Plateau.

Plain Language Summary

Proper synthesis of climate models' simulations is essential for a reliable future projection. Model synthesis requires evaluating model's performance. But model's performance will be different when different metrics are used, and past performance may not represent future performance in producing climate projections. Here we use large-ensemble historical and future projection simulations by three models as pseudo "true", for both the past and the future, to test the usefulness of model evaluation for future projection. Our test focuses on summer temperature projections over China. We evaluated the models' performance on different aspects including historical climatology, historical trends in global, regional and local temperatures. We find that projections based on the performance of reproducing past regional trend are generally more accurate; while that based on producing historical climatology or grid box scale trend has no or only small effect. Compared with the projection that does not consider model's performance, the weighted projection for the middle of the 21st century summer temperature, based on model's performance in reproducing the observed regional trend, is similar but with a reduction of uncertainty range by about 18%. The largest reduction at more than 0.4°C is observed in the northern Tibetan Plateau and parts of Northeast China.

1. Introduction

The Intergovernmental Panel on Climate Change in its 6th 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(). 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 CMIP6 driven by various emissions scenarios have provided a range of plausible future climate projections Eyring et al., 2016IPCC, 2021(;). Their proper synthesis can provide a coherent projection.

Traditionally, a ''democratic'' approach, i.e., simulations by each model being given with equal weight, has been used to synthesis multi-model projections. Projections by multi-models synthesized with this approach are more robust than those based on simulations by a single model Eyring et al., 2019Knutti, 2010Tebaldi & 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 Chen et al., 2011Giorgi & Mearns, 2003Li et al., 2021(; ;). Existing multi-model ensembles such as those produced through the World Climate Research Program's Coupled Model Intercomparison Project are ensembles of opportunity and are not designed to explore plausible model structures and epistemic uncertainty Knutti, 2010Benjamin M Sanderson et al., 2015Shiogama et al., 2022(; ;). Some models share components, making them not completely independent. This aspect needs to be considered when synthesizing multi-model 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 has been widely used to project future 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(), 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; (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 very reasonable, the validity of the second assumption is not obvious. It's not always possible to test the validity of the second assumption because the future is not known. As observations for the future do not exist, 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., 2019Bishop & Abramowitz, 2012Tebaldi & Knutti, 2007(; ;). But it is unclear if such performance measure is still valid for out of sample situation, i.e., in projecting future climates Knutti et al., 2010Tebaldi & Knutti, 2007(;). Different metrics have been used to evaluate model, resulting in difference in the level of model performance and thus different weighting schemes for the same set of models. For example, two dominate metrics, past climatology and past trend of a variable over a region are both used in model evaluation 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 processes such as east Asia summer

monsoon or feedbacks unique to the region.

The so-called "imperfect model test" or "model-as-truth test" provides an approach for estimating the skill of a future projection Abramowitz et al., 2019Eyring et al., 2019(;). This approach uses a particular model as pseudo "true" real-world and calibrates the remaining ensemble to the "truth" over historical time and then produce projection for the future represented by that model. As future world can be simulated by the same model, the true future becomes knowable and as a result, the performance of the projection can be compared with this known "truth". This imperfect model test has previously been applied to each member of each model in turn, then the results across all cases were synthesized Brunner et al., 2020Herger et al., 2019(;). However, as models may only have one or a few realizations, the limited sample size makes it difficult to separate internal variability and structural differences among the models, making it hard to interpret evaluation results Frankcombe et al., 2018Suarez-Gutierrez et al., 2021(;). In this regard, large-ensemble initial condition simulations have a unique advantage by providing multiple pseudo-observations Deser et al., 2020Milinski et al., 2020(;).

To explore the effect of the use of different metrics on projection skill and identify a more suitable spatial scale on which the model performance should be evaluated for the purpose of future projection, here we conduct imperfect model tests with model performance being evaluated by two metrics, climatology and long-term trends and on various spatial scales. To demonstrate the utility of our approach, we will focus on summer mean temperature over China. This is because various aspects of summer heatwaves are clearly and directly connected to summer mean temperature with higher mean temperature corresponding to longer, more frequent, and severer 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. Followed in section 3 are the main results of the skill assessment and future projections. Finally, general conclusions and discussion are provided in section 4.

2. Data and Methods

2.1 Data Used

CMIP6 simulations. We make use of 204 simulations conducted with 25 models participated in CMIP6. Table S1 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 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; 2. A large range of climate sensitivity of the models, with climate sensitivity lies in the upper, the middle, and the bottom of available CMIP6 models (Figure

S1-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. (2013)) are used. Historical data 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., 2020Tokarska et al., 2020(;). We focus on projected changes in the mid-21st century 2041-2060 relative to 1995-2014 baseline. Model data come with different spatial resolution, they are interpolated onto a common $2.5\circ \times 2.5\circ$ grids using bilinear interpolation.

Observational data. To produce future projections over China, we use a gridded temperature dataset CN05.1 Wu & Gao, 2013() to evaluate the model's performance after we have identified the most relevant model weighting scheme. The monthly gridded dataset covers 1961-2015, with a spatial resolution of $0.25\circ \times 0.25\circ$. We use the data from the same period 1971-2014, and re-grid it to $2.5\circ \times 2.5\circ$ resolution before used for model evaluation.

2.2 Imperfect Model Test Framework

To explore how skillful a weighting scheme established based on historical data for future projection, we conducted a series of model-as-truth tests under the "Imperfect model test" framework. This process involves two steps: estimating model weight based on historical simulation and evaluate the skill of the weight-based projection by comparing with model simulated future climate. These steps are detailed below.

2.2.1 Metrics for estimating distance

The similarity between observational record and a model simulation, or between simulations by two models, is measured by a distance measure based on suitably constructed metrics. The metrics under consideration include the following: 1) spatial distribution of temperature climatology (on $2.5°\times2.5°$ grid, referred as climatology metric below), 2) trend in regional mean temperature (referred as trend metric), 3) spatial distribution of trend on $2.5\times2.5°$ grid (referred as trend pattern metric), and 4) the combination of the climatology and trend metrics that is referred as composite metric. For the latter, the distance between two models or between a model and the observation is the average of the relevant climatology distance and trend distance. Trends are estimated based on least square fit to the area-weighted regional mean temperature series. We also used the non-parametric Sen's slope estimator for trends, the results are essentially the same as that of the best linear fit.

2.2.2 Weighting method ClimWIP

We follow the ClimWIP approach for determining model's weight. This method was proposed by Knutti et al. (2017) based on Benjamin M Sanderson et al. (2015), and has been used by many researchers Amos et al., 2020Liang et al., 2020Merrifield et al., 2020(; ;). The basic idea is that models agree more

poorly with observations and that largely duplicate existing models get less weight Knutti et al., 2017(). The weight w_i for the model i is given according to the following equation:

$$w_i = \frac{e^{-(D_i/\sigma_D)^2}}{\left(1+\sum_{j\neq i}^{M} e^{-(S_{\mathrm{ij}}/\sigma_S)^2}\right)} \quad (1)$$

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. When climatology and trend pattern metrics are used, the distance is the root mean square difference of climatological and trend values for all grids within the spatial domain. When the (regional) trend metric is used, the distance is the absolute difference of the trends. For both D_i and S_{ij}, the raw distances are normalized separately by dividing the raw distance by their respective medium values. $_d$ and $_s$ are shape parameters, corresponding to the strength of performance of individual models and independence among models. A larger $_d$ leads to more equally weighting of models and a larger $_s$ means models are treated to be more dependent. The procedure to determinate these two parameters follows Brunner et al (2020a), Lorenz et al (2018) and Knutti et al (2017). M is the total number of models, here is 24. We used the model ensemble mean to compute the weights rather than use individual model runs. This has the advantage of reducing the influence of internal variability, in particular when trend metric is used.

2.2.3 Spatial scale for model evaluation

When model's performance is evaluated on different spatial scales, the results can be different. As we will show later, the use of model climatology and trend pattern as performance metrics do not improve projection skills, the aspect related to spatial scale will be examined only when trend metric is used. We will consider four different spatial scales: a) trends in global summer mean temperature series (referred as global trend metric), b) trends in China-wide summer mean temperature series (referred as regional trend metric, the same as the trend metric mentioned above), c) trends in sub-regions of China where we divide China into East and West China separated by 105°E (ref. Li et al. (2021)), referred as sub-regional trend metric. d) trends in summer mean temperature of individual grids as grid trend metric. For the projections based on trend over a large region in the cases of the a) b) and c) listed above, each model is assigned only one weight, the weight is then applied to all grids within the domain. In the case of d), each grid box has its own set of weights.

2.2.4 Sampling procedure to generate projections

To compare multi-model weighted projection with the "known" future projection distribution provided by the large ensembles, we use a sampling method to generate a multiple model projection that reflects the model weights determined by the model evaluation scheme. Statistics about future projections are determined from the sampled data with all samples treated equally. For each ensemble member that is used as pseudo-observation, we produce 5000 samples of

future projection from the CMIP6 simulations, with the number of samples from an individual model equals to 5000 times the weight of the model (rounded to the nearest integer). The samples from individual models are randomly drawn from the available runs of the model with replacement. For a pseudo-observation with k ensemble members, we generate k sets of 5000 samples, based on which we produce probability distribution of the future projection and then compare it with those of the relevant large ensemble simulation. When producing future projection for summer mean temperature averaged over China, we sample the national mean values. When producing future projection for grid boxes, we sample the 2-dimentional spatial map of projected changes by individual model runs to maintain spatial structure of temperature changes.

2.2.5 Measures for skills assessment

The skill of the multi-model weighted projection against the unweighted projection is assessed on three aspects: a) bias, b) difference in width of the distribution, and c) the similarity between probability distributions. We examine if 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.

Bias. This compares the absolute bias between the median values of multi-model ensemble projection against the "known" projection by the large ensembles. The skill score is defined as following Eq. (2):

$$Bias\ skill\ score = \left|\text{Bias}_{\text{unweighted}}\right| - \left|\text{Bias}_{\text{weighted}}\right| \quad (2)$$

Width. The width between the 5$^{\text{th}}$ percentile and the 95$^{\text{th}}$ percentile is derived as representation of uncertainty range. The width skill score is defined as:

$$Width\ skill\ score = \text{Width}_{\text{unweighted}} - \text{Width}_{\text{weighted}} \quad (3)$$

Similarity between probability distributions. This measures how similar two probability density functions (PDFs) are. Perkins et al. (2007) proposed the use of the area that two PDFs overlap. A larger area of overlap means better agreement between the two PDFs. A perfect match of the PDF would give the value 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

$$Sscore = \sum_1^n minimum(Z_s,\ Z_o) \quad (4)$$

where n is the number of bins for which we use 50. This statistic has 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 bin size n will of course influence 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:

$$Sscore\ skill\ score = \mathrm{Sscore}_{\mathrm{weighted}} - \mathrm{Sscore}_{\mathrm{unweighted}} \quad (5)$$

3. Results

3.1 Skills of the Regional Climatology and Trend Metric

The left panel of Figure 1 presents the projected summer mean temperature over China by middle of the 21st century with simulations of the large ensembles by the three models as targets. Multi-model ensemble projections are constructed by using equal weighting (i.e., unweighted) for individual models or optimal weighting according to the four different metrics. As expected, different metrics consider different aspects, resulting in different projections. Overall, the regional trend metric boosts the consistency between ensemble projections and their targets, but that based on climatology metric or trend pattern metric offers little improvement.

The climatology metric has been widely used, as a default metric, to evaluate model's performance. Sometimes it has also been used to weight models when producing multi-model projection. Our tests show weighted projection based on performance in producing historical climatology does not improve model projection. The weighting does not reduce bias in the projected median change, there is also no clear evidence for it to reduce uncertainty range of the projection. While uncertainty range was reduced when EC-Earth3 and MIROC6 simulations were used as targets, the uncertainty range became slightly larger when CanESM5 simulations were the targets. This is not surprising as present-day climate conditions are not linked to the magnitude of warming Herger et al., 2018Knutti et al., 2010Benjamin M. Sanderson et al., 2017(; ;).

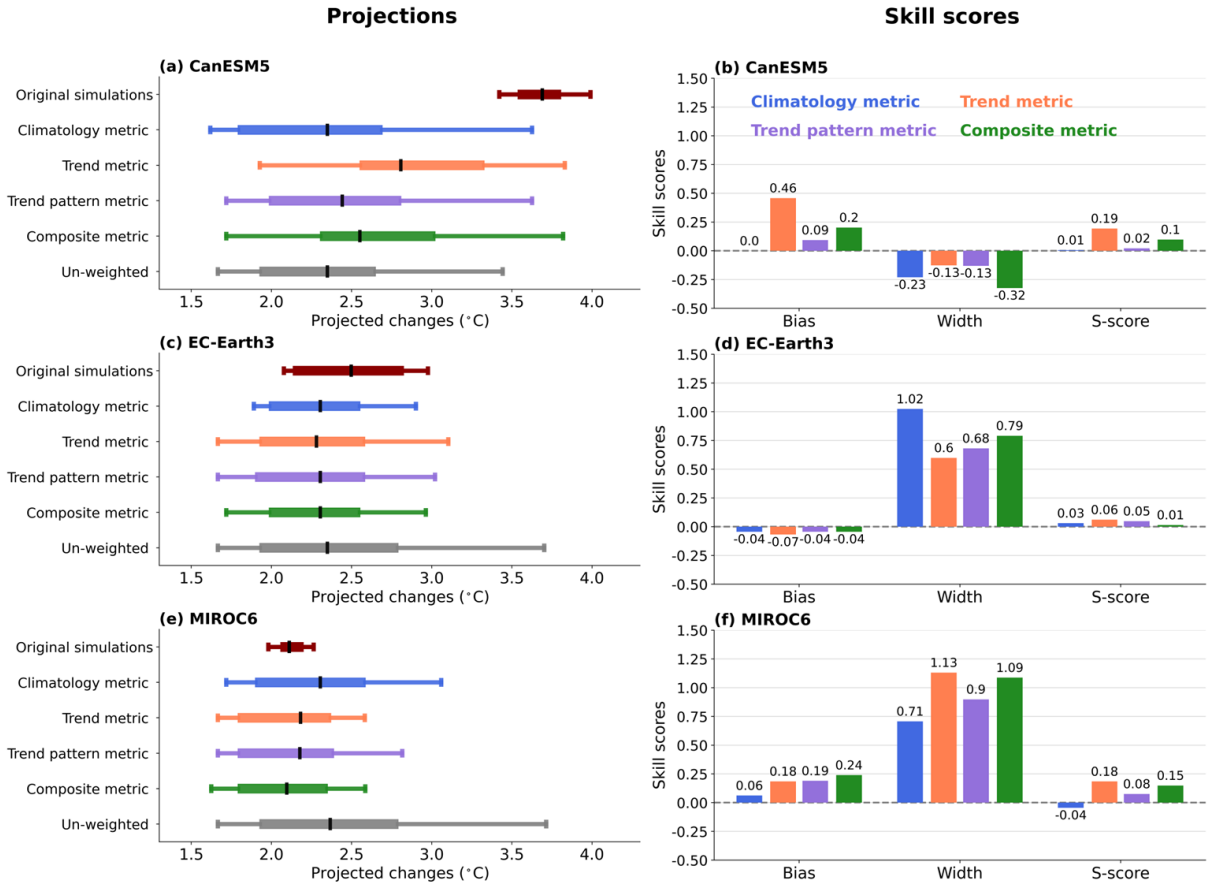


Figure 1. Multi-model projections and their target projections for changes in China's summer mean temperature during 2041-2060. The left panel shows the median (black ticks), the 25th-75th percentiles (boxes), and the 5th-95th percentiles (whiskers) with targets produced by three different models. Unit is (°C). The right panel shows the skill scores of weighted projections relative to the unweighted projection.

Weighted projection based on the performance of reproducing past trend are generally more accurate, with better agreement in the magnitude of projected changes and smaller uncertainty range, and higher skill scores in Bias and S-score. The improvement over unweighted projection is especially clear when the high sensitivity model CanESM5 and low sensitivity model MIROC6 were targets. When simulations of CanESM5 are targets, the unweighted CMIP6 ensemble projection could not reproduce the large magnitude warming simulated by CanESM5, with a cold bias of 1.29°C. The weighted ensemble reduces the bias to about 0.8°C, leading to positive skill scores of Bias and S-score of 0.46 and 0.18, respectively. In the case of MIROC6 as target, the weighted projection shifts the value downwards when compared to the unweighted projection, with

median value closer to the target median value and a considerable reduction in uncertainty range. These indicate that the trend matric has the effect of reducing bias and uncertainty.

In contrast, when spatial distribution of trend is used for model evaluation and weighting, it does not offer any improvement to the projection: it does not reduce the bias nor uncertainty in the projection. This may seem to be counterintuitive as one would expect local trends to be linked to model's sensitivities even though local trends would be noisier. But apparently uncertainty in the local trends due to natural variability renders the spatial distribution of trends usefulness for model evaluation.

The composite metric has been used to provide a comprehensive evaluation of models and also to avoid overconfidence in model weighting Lorenz et al., 2018Merrifield et al., 2020(;). As half of the metric is irrelevant to future warming and half of the metric is directly relevant to warming, it's effect in improving projection is also in-between the effects of its two constituents. Clearly the inclusion of climatology in model weighting reduced the usefulness of trend metric.

It is worth noting the case when the simulations by EC-Earth3 model are the targets as climate sensitivity of the model is in the middle of those of the available CMIP6 models. There is little room for improvement over the unweighted projection about bias. But weighted projections do reduce the projection uncertainty by a large margin, indicating a clear added value of model weighting. While the use of climatology metric resulted in a larger reduction in projection uncertainty, it is unclear if this is an indication of better performance as the S-score corresponding to this metric is lower than that when trend metric is used for weighting.

As many applications require local scale projection, we now present various skill measures at the grid box scale. Figures 2-4 show the skill scores computed at the grid box level when simulations by the three models are used as targets. Overall, these skill scores resemble those computed for the national mean temperatures. When the simulations by CanESM5 and MIROC6 are the targets (Figures 2 and 4), weighting the 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 little effect across the whole region when the climatology metric was used to weight the models, indicating again that better performance in simulating present-day climatology does not guarantee better future projections. The use of trend pattern metric for model weighting offers some improvement regarding bias and S-score but the improvement is very small. When the simulations by EC-Earth3 are the targets (Figure 3), weighting the models based on any metric does not affect the bias or the S-score, but the width of uncertainty range is greatly reduced, indicating again model weighting adds value.

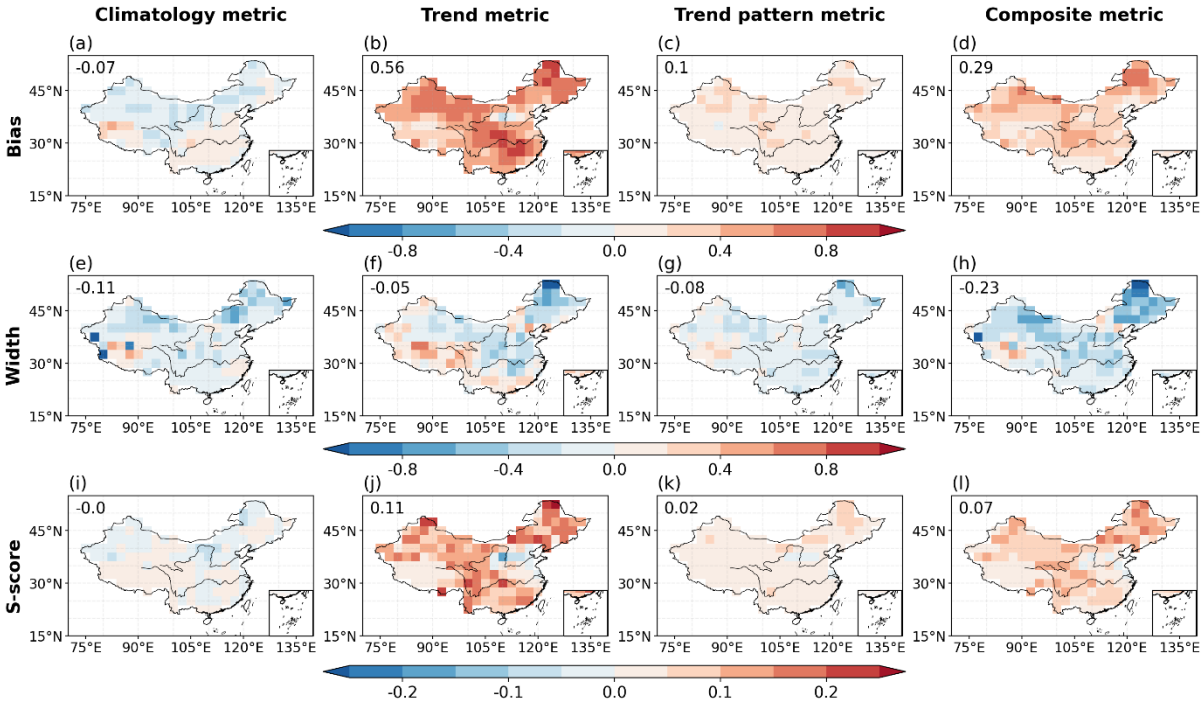


Figure 2. Spatial patterns of the skill scores corresponding to different model performance metrics. The skill scores include Bias (a-d), Width of uncertainty (e-h) and S-score (i-l) of weighted projections relative to unweighted projection when the simulations by CanESM5 are targets. While the skill scores are computed for individual grids separately, the weighs for every grid for the same model are the same. The numbers in the top-left corners inside each panel shows the median value of the skill scores within the spatial domain.

The skill scores are not uniform over the space. For instance, when performance in reproducing regional trend is used to weight models and simulations of CanESM5 are targets, notably better skill scores can be seen in the northeastern region and the Yangtze River Basin while the scores in the lower reach of the Yellow River Basin can be close to zero or even negative (Figure 2j). When simulations by MICRO6 are targets, negative skill scores can be seen in the Tibetan Plateau area and parts of Northwest China (Figures 4 j, l). But these grid-box skill scores should be interpreted in the context that projection on local scale is inherently more uncertain.

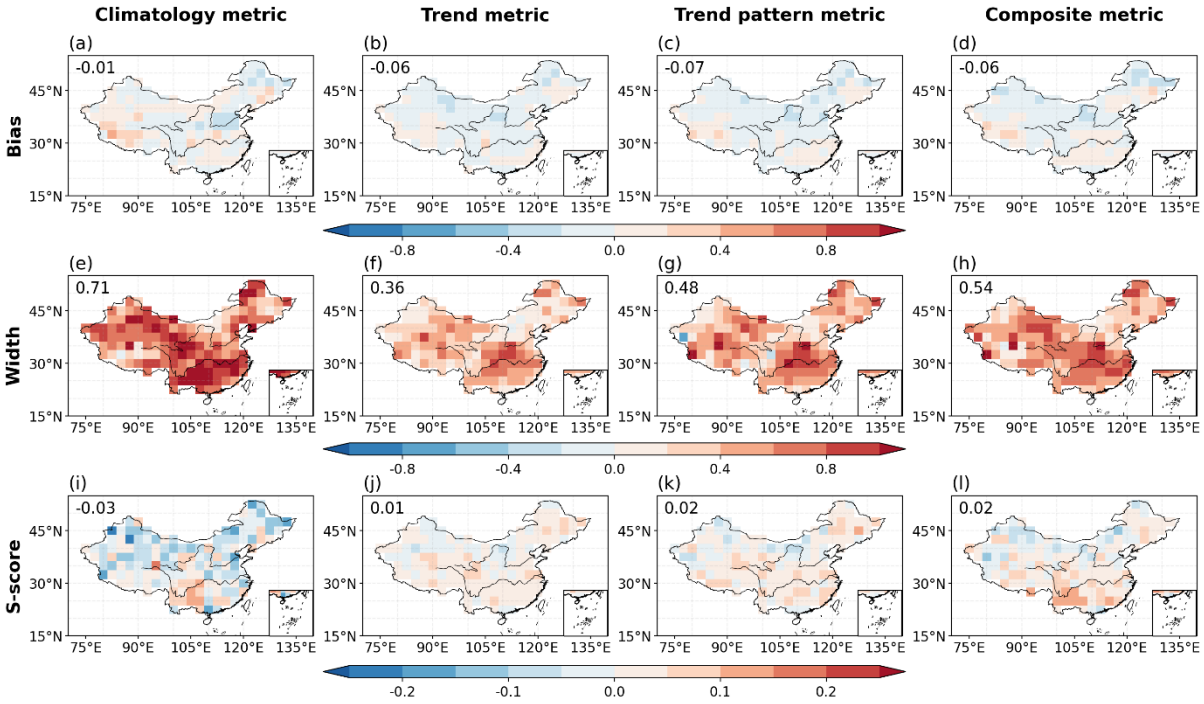


Figure 3. The same as Figure 2, but for the simulations by EC-Earth3 as targets for the projection.

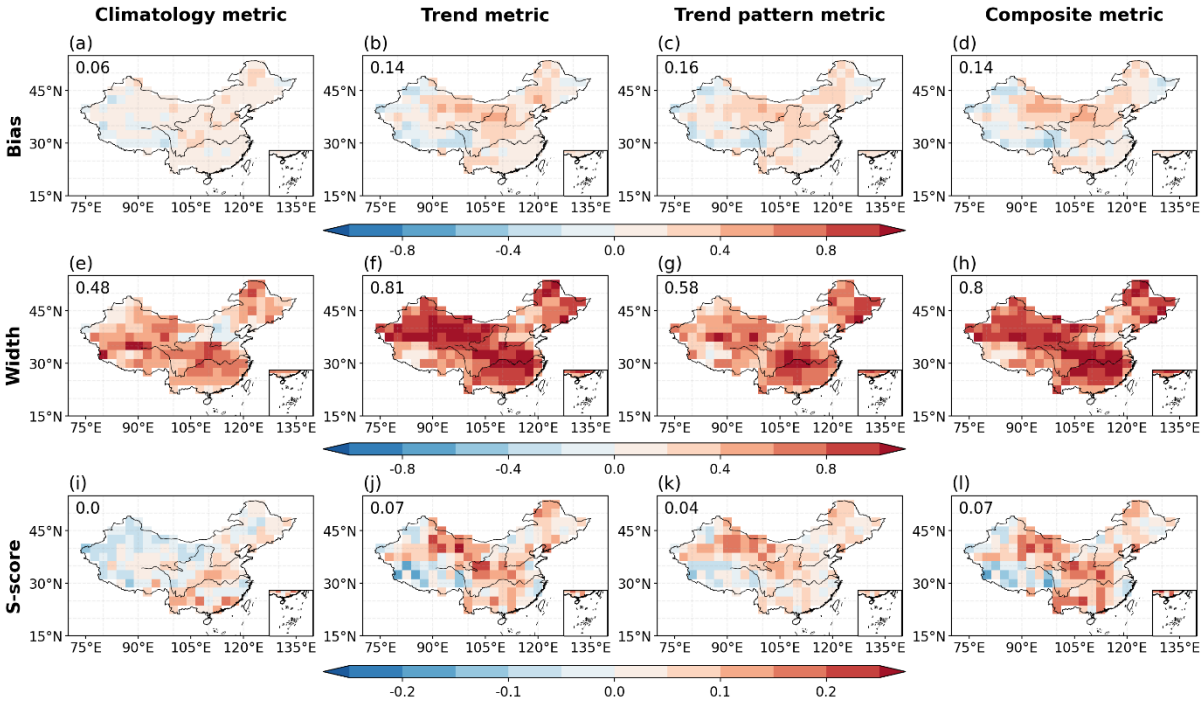


Figure 4. The same as Figure 2, but for the simulations by MIROC6 as targets for the projection.

3.2 Proper Spatial Scale of the Trend Metric

Having identified that the trend metric to be the most effective for model weighting, we now look at connection between the spatial scales of the trend and corresponding skill scores of the weighted projection. Figure 5 displays the results when simulations by CanESM5 are the targets and when temperature trends over the globe, over China, over West and East China or at the grid box scale are used as model's performance metric. In general, trends on different spatial scales as metrics for model weighting do improve projection but their corresponding skill scores can be quite different. When the global mean temperature trend is the performance metric, the skill scores show large spatial differences with large positive scores in West China and smaller or negative scores in East China (first column). When trend in the mean temperature over China is the performance metric, there are uniform improvement in the projection over the whole country, especially in terms of reduction in bias and in matching the probability distribution though improvement in the uncertainty range is very small (second column). Results for the regional trends in the West and the East China mean temperature as performance metric are similar to those of China mean temperature trends, though overall skill scores are slightly smaller (third column). When temperature trends on grid box scale are used as performance measure, the skill scores indicate overall improvement for the weighted

projection when compared to unweighted projection, but the improvement is very minimum (last column).

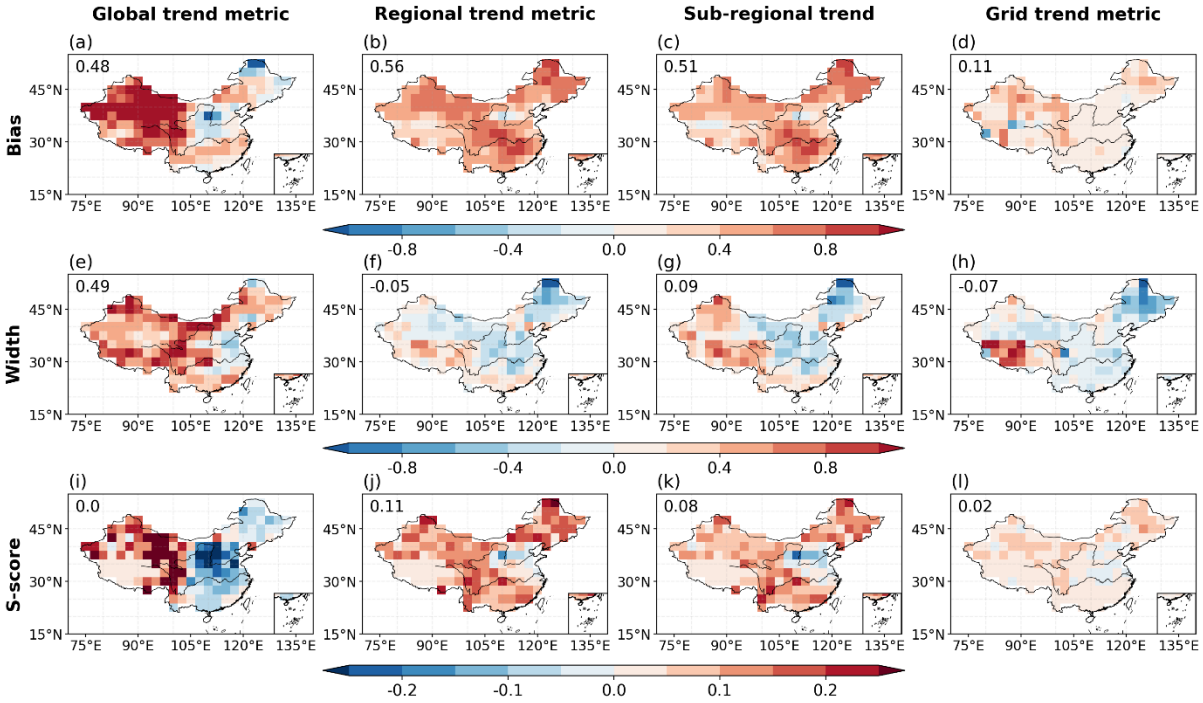
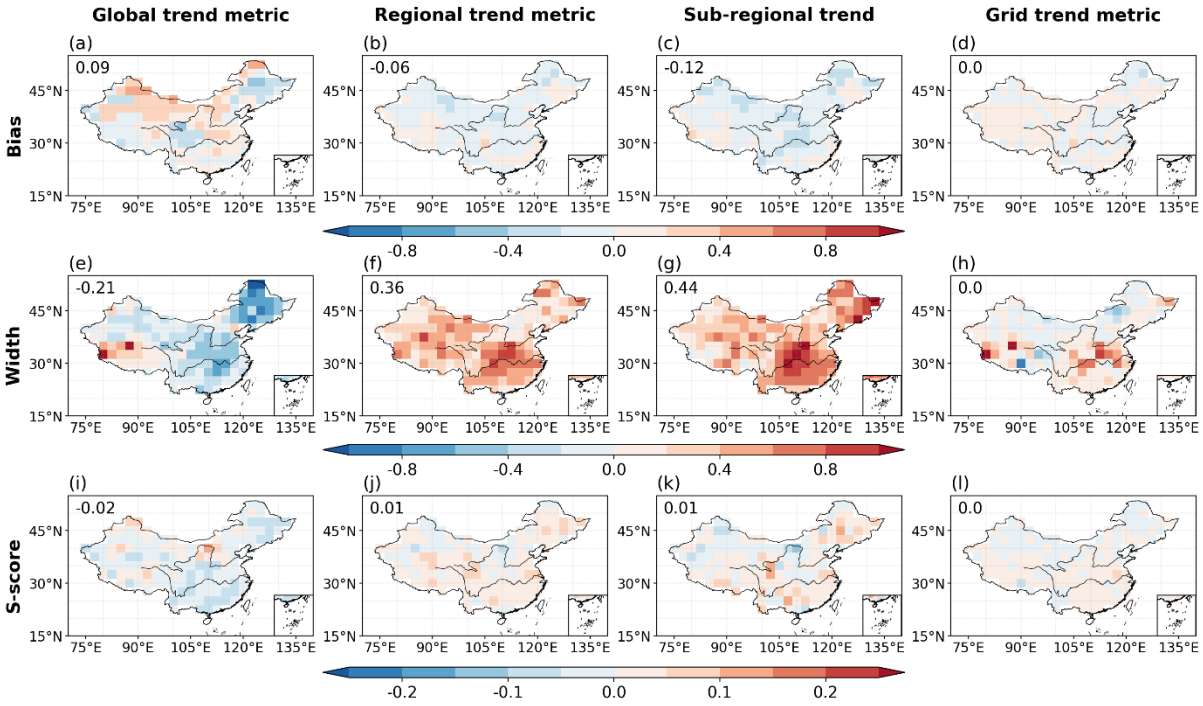


Figure 5. The same as Figure 2 but for trends on different spatial scales as metrics and when the simulations by CanESM5 are the targets. The weights for every grid depend on the metrics being used. The skill scores are computed for individual grids separately.

Figures 6 shows the results when simulations by EC-Earth3 are the targets. As projections by the EC-Earth3 are already in the middle of the projections by the available CMIP6 models, there is not much improvement about bias and S-scores as expected. But the use of regional trends as performance metric for model weighting does reduce uncertainty in the projection. The use of global temperature trend as metric for model weighting increases projection uncertainty. Figure 7 presents the results when simulations by MICRO6 are the targets. The use of global mean temperature trend as performance metric for model weighting has little effect on bias though it greatly reduces the projection uncertainty. The use of regional trend as metric improves projection by reducing bias, uncertainty and improving distributional match across the country, relatively uniformly. The use of grid-box trend as metric again results in little improvement.



Figu

6. The same as Figure 5, but for the simulations by EC-Earth3 as targets for the projection.

Overall, the use of grid-box scale historical trend as performance metric for model weighting offers some improvement than unweighted objection, but the improvement is small. When models are weighted according to their performance in simulating historical global mean temperature trend, the projection for regional mean temperature over China is improved. But the effect on projection on grid box scale can be quite uneven over the space and may not be robust, there can be large improvement in some regions and there can also be poorer projection in other regions depending on the target model simulations. When the models are weighted based on their performance in simulating regional trend over China or large sub-regions 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. Local trend is highly uncertain as a result of natural internal variability, its usefulness as model's performance measure is limited. While global temperature trend is a good indicator of model's climate sensitivity, performance in simulating it by individual models may not reflect well regional and local processes and feedback well.

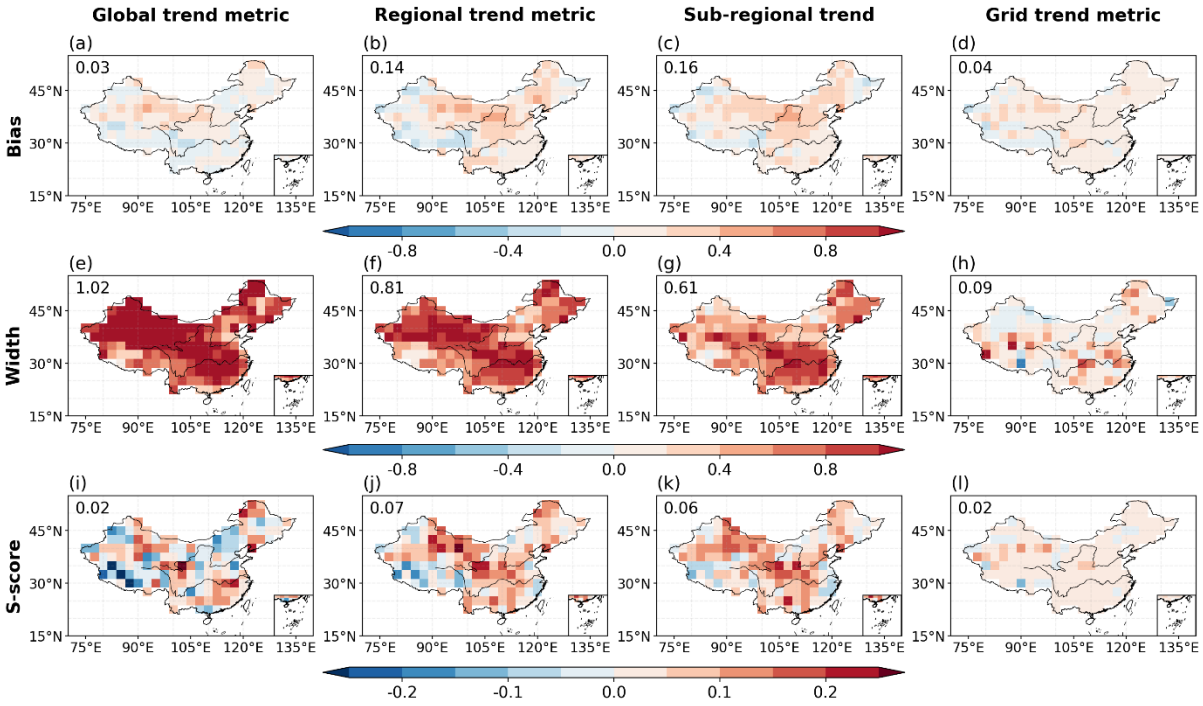


Figure 7. The same as Figure 5, but for the simulations by MIROC6 as targets for the projection.

3.3 Future Projection of China's Summer Temperature

As the regional mean temperature trend is the most effective performance metric for model weighting, we use it to estimate model weighting when constructing summer temperature projection for the middle of the 21st century (2041-2600) over China. Figure 8 shows the projection of national mean temperature. The weighted multi-model ensemble projects a median increase of 2.3°C over the 1995-2014 base period, with the 5-95th percentile range of 1.67°C to 3.4°C. When compared with the unweighted multi-model projection, the weighted projection reduces the uncertainty range by 0.32°C (18%), with the most reduction in the upper bound. This result is consistent with previous studies that used other methods to treat the 'hot tail' CMIP6 models problem IPCC, 2021Nijsse et al., 2020Ribes et al., 2021Tokarska et al., 2020Zeke et al., 2022(; ; ; ;).

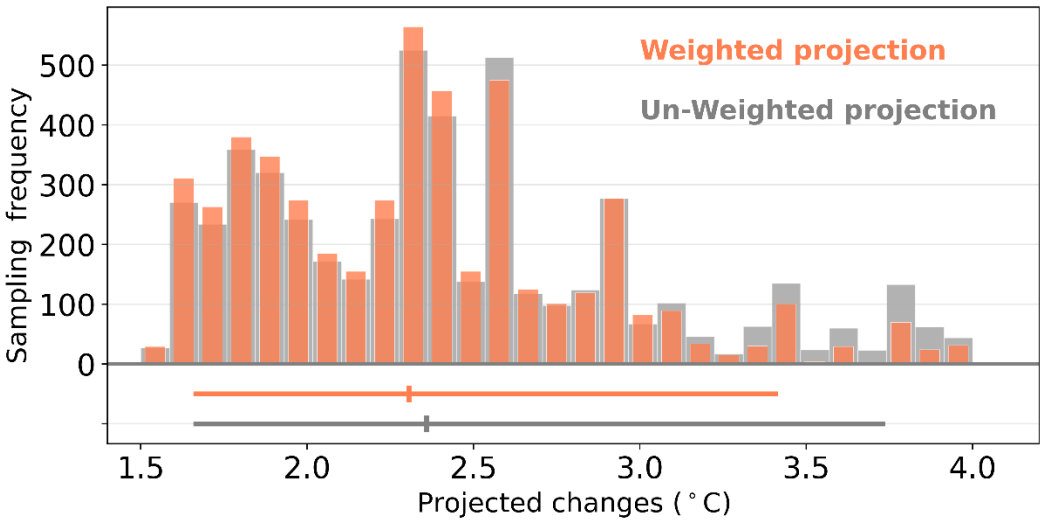


Figure 8. Histogram for projected changes in summer temperature over China for the middle of the 21st century (2041-2600) relative to1995-2014 base period. The histogram shadings show the sampling frequency distribution. The lines at the bottom mark the 5-95% ranges, with the median values marked by the vertical ticks.

Figure 9 displays the median, 5th and 95th percentiles of the projected changes on grid box scale. Warming is widespread the entire region and for all percentiles, with spatial median value of 2.36°C for the median projection and 1.53°C and 3.59°C for the 5th and 95th percentiles, respectively. Larger warming occurs in the northern high-latitude regions, especially in the Northwest China, with median increase as large as 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 0.06°C (Figure 9b). The difference in the 5th percentile projection is even smaller, values are mostly within 0.1°C across the whole region. In contrast, the 95th percentile is much reduced in the weighted projection, with largest reduction at more than 0.4°C in the northern Tibetan Plateau and parts of Northeast China. Due to the reduction in the 95th percentile, the 5th-95th uncertainty range is also reduced by at least 0.2°C.

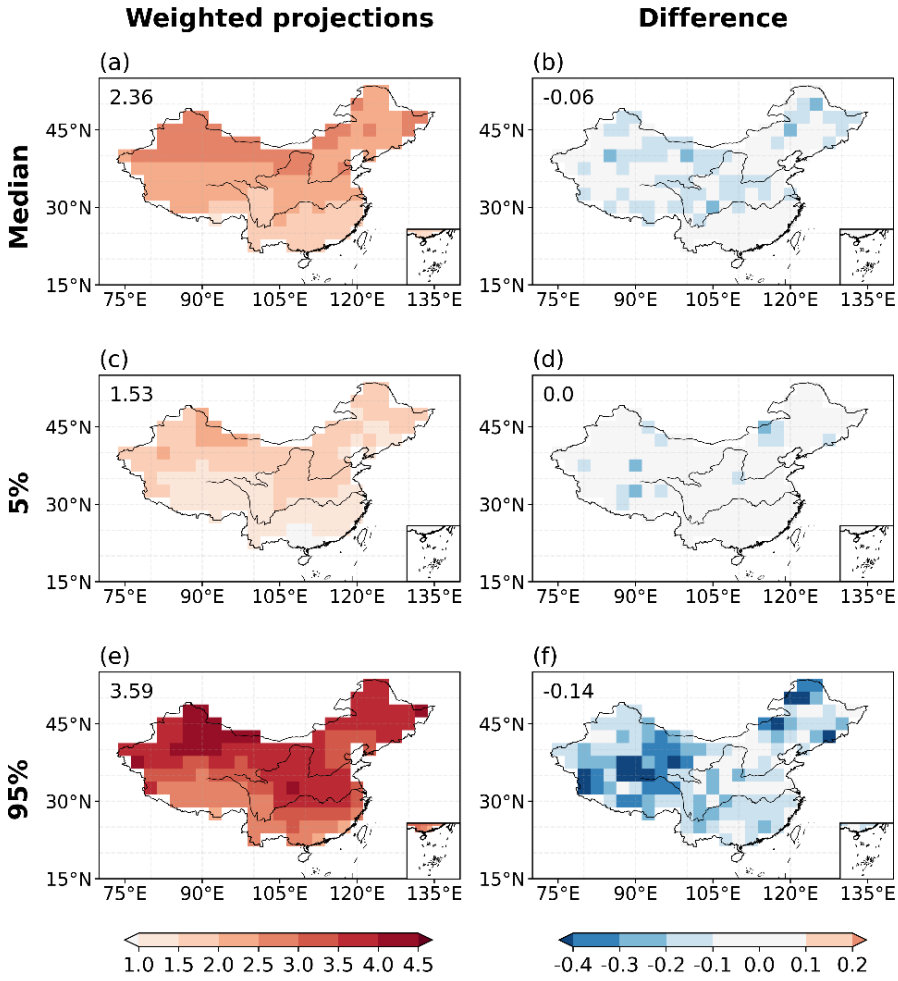


Figure 9. Weighted projection and its difference from unweighted projection for summer mean temperature changes for the middle of the 21st century. The median, the 5th and the 95th percentiles of weighted projection (left panel) and the difference (right panel) are shown. The numbers in the top-left corner inside the panels show the median value of the skill scores within the domain. Unit is (°C).

4. Conclusion 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, trends in mean temperatures on various spatial scales including the global, the regional, and grid-box scales as the bases for model weighting.

We estimated model weighting skills using large ensemble simulations by three climate models of different climate sensitivities.

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 see clearly that when trends in the mean temperature over China or sub-regions of China are used to evaluate climate models' performance for model weighting, the bias and uncertainty in the weighted projections are greatly reduced. We also see that not all model evaluations are created equal. When the model's performance is evaluated based on a popular metric, observed climatology, the weighted projection does not improve upon the unweighted projection. Clearly, the model evaluation needs to fit for the particular purpose. To the first order, changes in temperature are mostly the results of thermodynamic effect of the global warming. It thus makes sense for historical temperature trend to be a relevant performance metric. In the perfect model tests, we saw models of higher sensitivity being given higher weights when the target of the future projection is simulated by a high sensitivity model CanESM5. The reverse is also true when the target of the future projection is simulated by a lower sensitivity model. This also explains why model weighting based on historical climatology did not offer any improvement as there is not a clear link between model's sensitivity and climatology. Therefore, we emphasize that metric for model evaluation must be fit-for-purpose, being relevant to the projected future change for the variable of interest.

Spatial scale on which model's performance is evaluated also plays a role. Trend in regional mean temperature over China or sub-regions of China seem to perform the best. Trend in global mean temperature improved projection for regional mean temperature over China but its performance for projection on grid-box scale is mixed, suggesting that some regionally or locally important processes and feedbacks may not be well represented in the global mean temperature trend. Grid box scale trends offer little improvement in the projection, suggesting that the noisy nature of trends on such fine spatial scale does not provide useful information for selecting better performing models. This is a strong indication that evaluating a model's performance at gid box scale is not a useful exercise.

For the model weighting to be effective, the metric for evaluating the model's performance must meet two conditions. 1) The observed metric must be related to climate response to external forcing, with signal separatable 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 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 heat waves 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 because of connection between atmospheric moisture and temperature.

Weighting the CMIP6 models based on their performance in simulating the observed summer temperature trend in China, we project summer temperature in China will increase by about 2.3°C with the 5-95th percentiles range of 1.67°C to 3.40°C, by the middle of the 21st century (2041-2060). Compared with unweighted projection, the median and the 5th percentile change little, but the 95th percentile is reduced by 0.32°C. This is in line with some studies that suggest climate sensitivities in some CMIP6 models to be too high Sherwood et al., 2020Zeke et al., 2022(;). The weighted projection has smaller uncertainty range compared with that of unweighted projection, with a reduction of 18%. Larger reduction in the uncertainty is observed in the northern Tibetan Plateau area and parts of Northeastern China, 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, the confidence about this 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 Key Research and Development Program of China (Grant 2017YFA0603804), the National Natural Science Foundation of China (42275184), the Postgraduate Research and Practice Innovation Program of Government of Jiangsu Province (KYCX21_0940), and the Visiting Fellowship from China Scholarship Council (NO. 202209040009).

Data Availability Statement

The CMIP6 model data that support the findings of this study are openly available at the following URL/DOI: https://esgf-node.llnl.gov/search/cmip6/. The high-quality in situ dataset (CN05.1) is available through Wu and Gao (2013).

Reference

https://doi.org/10.5194/esd-10-91-2019

https://doi.org/10.5194/acp-20-9961-2020

https://doi.org/10.1007/s00382-012-1610-y

https://doi.org/10.1088/1748-9326/ab492f

https://doi.org/10.5194/esd-11-995-2020

https://doi.org/10.1175/2011jcli4102.1

https://doi.org/10.1038/s41558-020-0731-2

https://doi.org/10.5194/gmd-9-1937-2016

https://doi.org/10.1038/s41558-018-0355-y

https://doi.org/10.1175/jcli-d-17-0662.1

https://doi.org/10.1029/2003gl017130

https://doi.org/10.5194/esd-9-135-2018

https://doi.org/10.1007/s00382-019-04690-8

https://doi.org/10.1007/s10584-010-9800-2

https://doi.org/10.1175/2009jcli3361.1

https://doi.org/10.1002/2016gl072012

https://doi.org/10.1007/s13351-021-0067-5

https://doi.org/10.1029/2019gl086757

https://doi.org/10.1029/2017jd027992

https://doi.org/10.5194/esd-11-807-2020

https://doi.org/10.5194/esd-11-885-2020

https://doi.org/10.5194/esd-11-737-2020

https://doi.org/10.1007/s10584-013-0905-2

https://doi.org/10.1126/sciadv.abc0671

https://doi.org/10.1175/jcli-d-14-00362.1

https://doi.org/10.5194/gmd-10-2379-2017

https://doi.org/10.1029/2019RG000678

https://doi.org/10.1038/s41586-021-04310-8

https://doi.org/10.1007/s00382-021-05821-w

https://doi.org/10.1038/nclimate2410

https://doi.org/10.1098/rsta.2007.2076

https://doi.org/10.1126/sciadv.aaz9549

https://doi.org/10.6038/cjg20130406

https://doi.org/https://doi.org/10.1038/d41586-022-01192-2

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., . . . Schmidt, G. A. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics, 10*(1), 91-105. Amos, M., Young, P. J., Hosking, J. S., Lamarque, J.-F., Abraham, N. L., Akiyoshi, H., . . . Yamashita, Y. (2020). Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence. *Atmospheric Chemistry and Physics, 20*(16), 9961-9977. Bishop, C. H., & Abramowitz, G. (2012). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics, 41*(3-4), 885-900. Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters, 14*(12). Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics, 11*(4), 995-1012. Chen, W., Jiang, Z., & Li, L. (2011). Probabilistic Projections of Climate Change over China under the SRES A1B Scenario Using 28 AOGCMs. *Journal of Climate, 24*(17), 4741-4756. Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., . . . Ting, M. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change, 10*(4), 277-286. Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development, 9*(5), 1937-1958. Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., . . . Williamson, M. S. (2019). Taking climate model evaluation to the next level. *Nature Climate Change, 9*(2), 102-110. Frankcombe, L. M., England, M. H., Kajtar, J. B., Mann, M. E., & Steinman, B. A. (2018). On the Choice of Ensemble Mean for Estimating the Forced Signal in the Presence of Internal Variability. *Journal of Climate, 31*(14), 5681-5693. Giorgi, F., & Mearns, L. O. (2003). Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophysical Research Letters, 30*(12). Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., & Sanderson, B. M. (2018). Selecting a climate model subset to optimise key ensemble properties. *Earth System Dynamics, 9*(1), 135-151. Herger, N., Abramowitz, G., Sherwood, S., Knutti, R., Angélil, O., & Sisson, S. A. (2019). Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change. *Climate Dynamics, 53*(3-4), 1581-1596. IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* C. U. Press.IPCC. (2022). *Summary for Policymakers. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* C. U. Press.Knutti, R. (2010). The end of model democracy? *Climatic Change, 102*(3-4), 395-404. Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Chal-

lenges in Combining Projections from Multiple Climate Models. *Journal of Climate, 23*(10), 2739-2758. Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*. Li, T., Jiang, Z., Zhao, L., & Li, L. (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. *Journal of Meteorological Research, 35*(1), 184-197. Liang, Y., Gillett, N. P., & Monahan, A. H. (2020). Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. *Geophysical Research Letters, 47*(12). Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. *Journal of Geophysical Research: Atmospheres, 123*(9), 4509-4526. Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., & Knutti, R. (2020). An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics, 11*(3), 807-834. Milinski, S., Maher, N., & Olonscheck, D. (2020). How large does a large ensemble need to be? *Earth System Dynamics, 11*(4), 885-901. Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020). Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth System Dynamics, 11*(3), 737-750. O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., . . . van Vuuren, D. P. (2013). A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Climatic Change, 122*(3), 387-400. Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Sci Adv, 7*(4). Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *Journal of Climate, 28*(13), 5171-5194. Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development, 10*(6), 2379-2395. Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., . . . Zelinka, M. D. (2020). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Rev Geophys, 58*(4), e2019RG000678. Shiogama, H., Watanabe, M., Kim, H., & Hirota, N. (2022). Emergent constraints on future precipitation changes. *Nature, 602*(7898), 612-616. Suarez-Gutierrez, L., Milinski, S., & Maher, N. (2021). Exploiting large ensembles for a better yet simpler climate model evaluation. *Climate Dynamics, 57*(9-10), 2557-2580. Sun, Y., Zhang, X., Zwiers, F. W., Song, L., Wan, H., Hu, T., . . . Ren, G. (2014). Rapid increase in the risk of extreme summer heat in Eastern China. *Nature Climate Change, 4*(12), 1082-1085. Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans A Math Phys Eng Sci, 365*(1857), 2053-2075. Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Sci Adv, 6*(12), eaaz9549. Wu, J., & Gao, X. J. (2013). A gridded daily observation dataset

over China region and comparison with the other datasets (in Chinese). *Chinese Journal of Geophysics Chinese Edition, 56*(4), 1102-1111. Zeke, H., Kate, M., Gavin, A. S., John, W. N.-G., & Mark, Z. (2022). Climate simulations: recognize the 'hot model' problem. *Nature*.