Machine learning to optimize climate projection over China with multi-model ensemble simulations

Tong Li, Zhihong Jiang, Hervé Le Treut, Laurent Li, Lilong Zhao, Lingling Ge

Machine learning to optimize climate projection over China with multi-model ensemble simulations

Tong Li [1], Zhihong Jiang [2*], Hervé Le Treut[3], Laurent Li[3], Lilong Zhao[1] and Lingling Ge[4]

1 Joint International Research Laboratory of Climate and Environment Change, Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disaster, Nanjing University of Information Science and Technology, Nanjing 210044, China;

2 Key Laboratory of Meteorological Disaster of Ministry of Education, Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disaster, Nanjing University of Information Science and Technology, Nanjing 210044, China;

3 Laboratoire de Météorologie Dynamique, IPSL, CNRS, Sorbonne Université, Ecole Normale Supérieure, Ecole Polytechnique, Paris 75005, France;

4 Jiangsu Key Laboratory of Big Data Analysis Technology, School of Information and Control,

Nanjing University of Information Science and Technology, Nanjing 210044, China.

*Corresponding author. E-mail address: zhjiang@nuist.edu.cn (Z. Jiang). Tel: +86-025-58731135.

Abstract

The multi-model ensemble approach is generally considered as the best way to explore the advantage and to avoid the weakness of each individual model, and ultimately to achieve the best climate projection. But the design of an optimal strategy and its practical implementation still constitutes a challenge. Here we use the Random Forest (RF) algorithm (from the category of Machine Learning) to explore the information offered by the multi-model ensemble simulations within the Coupled Model Intercomparison Project Phase 6. Our objective is to achieve a more reliable climate projection (mean climate and extremes) over China. RF is furthermore compared to two other ensemble-processing strategies of different nature, one is the basic arithmetic mean (AM), and another is the linear regression (LR) across the ensemble members. Our results indicate that RF effectively enhances the capability in capturing spatial climate characteristics. Regions with complex topography, such as the Tibetan Plateau and its periphery, show the most significant improvements. RF projects less future warming but enhanced wet conditions across China. It also produces larger spatial variability and more small-scale features. The most obvious increase of precipitation is in the northern part and the periphery of the Tibetan Plateau. The projected changes in RF for strong precipitation are almost twice higher than in AM, while in the northwestern area, weaker increases of precipitation are projected by RF, which indicates larger spatial inhomogeneity of its projection.

Keywords: Ensemble-processing strategy, Climate extremes, Observational constraint, Random Forest, China regional climate.

1. Introduction

Global warming has altered the mean and extreme climate in many regions of the world, and this warming trend will undoubtedly continue (Hulme 2016). Global Climate Models (GCMs) play a crucial role in generating future projections to examine the potential impacts of climate change. The ability of GCMs to reproduce observed features of the past and current climate increases our confidence to correctly make future projections (Palmer *et al* 2005; Semenov and Stratonovitch 2010). Climate projection is inevitably accompanied by uncertainties, with available physically-based models being imperfect (Knutti *et al* 2013; Hidalgo and Alfaro 2015). The multi-model ensemble approach is useful to explore the advantage and to avoid the weakness of individual models, and ultimately to achieve the best climate projection. But the design of an optimal ensemble-processing strategy and its practical implementation still constitute a challenge (Knutti *et al* 2010; Knutti *et al* 2013). The arithmetic mean (hereafter called AM) is the simplest and mostly-used method to deal with a multi-model ensemble (Knutti *et al* 2010; Sanderson *et al* 2015). Subsequently, more complex statistical methods such as the Bayesian methods (Robertson *et al* 2004; Tan *et al* 2016) or weighted averages, which consider the simulation skills and model inter-dependence, have been developed (Xu *et al* 2010; Jiang *et al* 2015; Knutti *et al* 2017; Brunner *et al* 2020). These methods allow tuning particular parameters or weights and constraining uncertainties with historical observations. Most of these strategies or methods, however, rely on the concept of linear regression based on some specific relationships or indices, potentially neglecting useful information.

With observations as a target or a constraint, machine learning (ML) is a useful tool to extract more information from multi-model data. Significant advancements have been reported with application of heuristic machine learning for uses in weather forecast, climate prediction, and reconstruction of missing climate information (Ham *et al* 2019; Reichstein *et al* 2019; Kadow *et al* 2020). ML has considerable advantages in solving non-linear, high-dimensional, and hierarchical problems to retrieve implicit patterns in complex relationships (Alizamir *et al* 2018; Guo *et al* 2019; Li *et al* 2020). With such general properties, ML can better extract important dynamical and physical processes within climate models and fully explore useful information (Wang *et al* 2018; Reichstein *et al* 2019). This would lead to a hybrid approach for future climate projection, which combines the strengths of physical modelling and mathematical algorithms of machine learning (Reichstein *et al* 2019; Watson-Parris 2020).

Under the framework of the Coupled Model Intercomparison Project (CMIP), a large number of climate simulations have been performed and released publicly. CMIP is an unprecedent effort and has entered its sixth phase (CMIP6) (Eyring *et al* 2016), with more models and a larger ensemble of simulations compared to its predecessor (CMIP5) (Liang *et al* 2020; Zhu, H *et al* 2020). It offers exciting new opportunities for expanding our knowledge of the Earth system through the exploration of big data with advanced ML concepts and algorithms. The present study uses the Random Forest (RF), a powerful ML algorithm that is based on the decision tree and able to extract non-linear relations and behaviors (Breiman *et al* 1984; Breiman 2001). For the purpose of demonstration, RF is contrasted to the arithmetic mean (AM), the simplest ensemble-processing strategy, as well as the basic linear regression (LR) applied to the ensemble members. We want to check whether RF can effectively enhance our skill to mimic observed properties and to make reliable future climate projections. This work is a part of our general efforts of climate change mitigation and adaptation in China. It focuses on the recommended targets of 1.5°C, and 2°C global warming levels, following the Paris Agreement (UNFCCC 2015). The geographic domain of our investigation is mainland China

where a reliable dataset of observed climate is available.

The rest of the paper is organized as follows. Section 2 describes the data, methodology, and the three algorithms involved in our study, together with the skill metrics for evaluation. Followed in Section 3 are the main results of the methodological assessment in present-day and future climate projection. Finally, conclusions and a few discussions are provided in Section 4.

2. Data and Methods

2.1 Study area and data used

This work focuses on mainland China, a territory highly susceptible to climate change due to its complex topography and strongly-pronounced monsoonal characteristics (Fu *et al* 2008; Piao *et al* 2010). A high-quality in-situ dataset (CN05.1), including conventional surface climatic variables, is employed for the calibration of all our approaches to develop a reliable multi-model ensemble-processing strategy. The daily gridded dataset covers 1961–2014, with a spatial resolution of 0.25° × 0.25° over whole China. Wu and Gao (2013) provide detailed information about this dataset.

On the other hand, 24 CMIP6 models' historical simulations and future projections from Shared Socioeconomic Pathway (SSP5-8.5) are used to construct the multi-model ensemble and to generate 1.5, 2°C and 3°C warming projection. These models were selected on the sole criterion of data availability for our purpose of determining warming targets at 1.5, 2°C and 3°C. All CMIP6 data were retrieved through the data portals of the Earth System Grid Federation, which can be obtained from https://esgf-node.llnl.gov/search/cmip6/. Some essential characteristics of the used models are listed in Table S1. Only their first realization (r1i1p1f1) was used in this work.

2.2 Methods

2.2.1 Climate indices

The present study employed six quantitative indices, including mean temperature (TAS), annual maximum (hottest daytime) temperature (TXx), annual minimum (coldest nighttime) temperature (TNn), total precipitation in wet days (PRCPTOT), annual maximum consecutive 5-day precipitation amount (RX5DAY) and annual total precipitation for events exceeding the 95th percentile (R95P, an indication of strong precipitation). These indices are useful in capturing climate change information and have been widely used to identify and monitor extreme climate (Zhang *et al* 2011; Zhu, H *et al* 2020). They are derived from daily precipitation and temperature CMIP6 datasets following the recommendation by the Expert Team on Climate Change Detection and Indices (ETCCDI) (http://etccdi.pacificclimate.org/). Indices from different models and observation were first calculated at their original grid and then interpolated, using bilinear interpolation, onto a common 1° × 1° grid comprising 928 geographic locations across China. The three ensemble-processing strategies, AM, LR and RF, were then practiced on this common grid to ensure fairness and to facilitate their inter-comparison.

The study adopted the criteria used by Shi *et al* (2018) in defining the calendar year for models to reach 1.5°C and 2°C global warming thresholds. A time window of 21 years, including the ten years before and after the nominative year, is used to deduce the climate

statistics. A similar approach has been utilized in a few recent studies (e.g., Sun *et al* (2019); Guo *et al* (2020)).

2.2.2 Strategies in processing multi-model ensemble

Figure 1 shows an overall flow chart of our designed processing. Historical simulations, together with observation, are divided into the training period 1961–1994 (34 years) and the testing period 1995–2014 (20 years). The testing period also serves as the historical reference for future warming projection. Our procedure is separately applied to each of the six climate indices with the general goal to explore, as much as possible, the properties of observation. The basic principle is to minimize the loss function (here the Mean Squared Error) representing the deviation between the multi-model ensemble output and the observation. Once the training procedure is accomplished, the optimized multi-model ensemble-processing scheme can then be used to produce results for the testing period. Finally, future projections under the 1.5°C, 2°C and 3°C global warming were conducted.

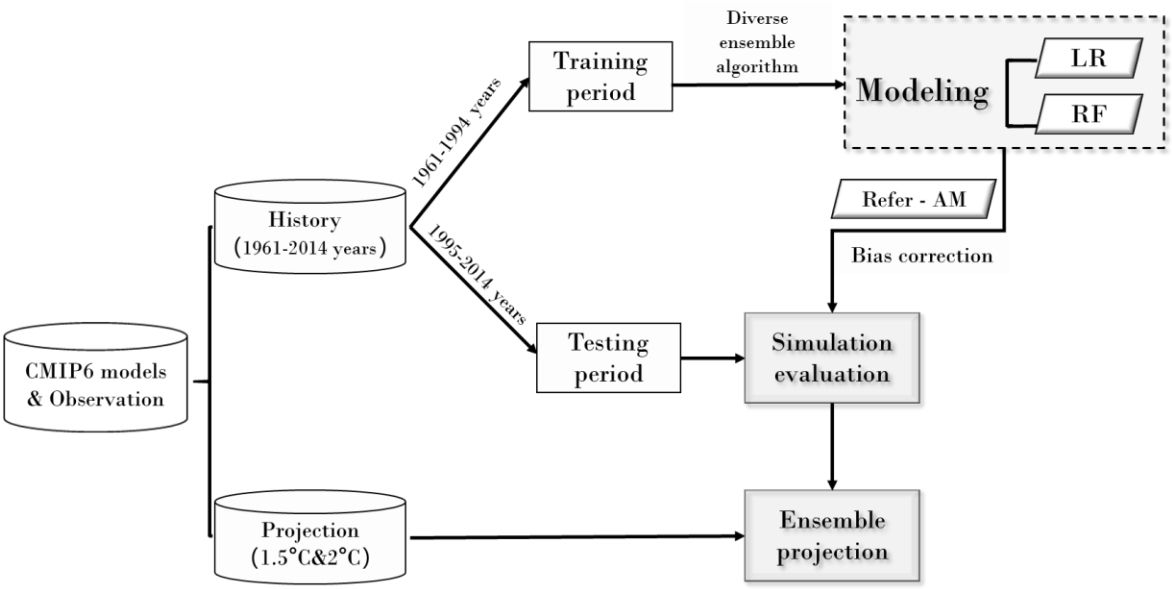

Figure 1. Schematic showing the design and operating process to deal with multi-model ensemble simulations. Historical simulations, together with observation, are used to train different multi-model ensemble-processing strategies, and to assess their performance. The validated strategy is then used to make projections of future climate.

The arithmetic mean is the simplest and widely-used ensemble-processing strategy. There is no parameter to optimize and it is incapable of learning from training data, which would constitute a biased reference to fairly evaluate other ensemble-processing strategies. To ensure a fair comparison with LR or RF, a linear scaling is used in AM to remove biases of climate models with their domain-mean deviation from observation (Lenderink *et al* 2007; Teutschbein and Seibert 2012). The temperature (T) is corrected with an additive term on original value and precipitation (P) with a multiplier.

$$T_{cor} = T_{ori} + \mu\left(T_{obs}\right) - \mu\left(T_{ori}\right) \quad (1)$$

$$P_{cor} = P_{ori} \times \frac{\mu(P_{obs})}{\mu(P_{ori})} \quad (2)$$

where subscripts denote corrected (*cor*), raw (*ori*), and observed (*obs*) values, and μ represents averaging over the domain.

The result of AM after bias correction is included here as a comparison baseline. It is worth mentioning that this linear scaling bias correction has no impact on the projections. This is due to the fact that methods measuring future changes are absolute change for temperature and relative change for precipitation.

LR is a basic linear algorithm, suitable for resolving regression problems across multiple models or members in an ensemble. It fits a linear model to minimize the sum of squared errors. Its general form can be written as: $Y = a_0 + A \bullet X$, where $X(i, k)$ is the input spatial field (i = 1, …, 928) from the 24 models (k=1, …, 24) and $Y(i)$ is the output spatial field (i = 1, …, 928). The regression coefficients a_0 and A_k, (k=1, …, 24) were fitted with data in the training period comprising 34 years from 1961 to 1994.

A linear model is not always inferior to nonlinear models, depending on the nature of the problem to resolve (Choubin *et al* 2016; Xu *et al* 2020). The practical realization of LR used in this paper was done through the function "LinearRegression" in the module "sklearn.linear_model" in python 3.8 (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression).

Random Forest solves regression problems by growing an ensemble of decision trees based on binary recursive partitioning (Breiman *et al* 1984; Breiman 2001). Although each individual regression (done at the level of leaves or terminal nodes) is still linear, but it is operated in a reduced range among the total samples. This is why RF can solve non-linear problems and reveal complex behaviors hidden in the data samples. Its randomness manifests in two particular points. Firstly, the samples used to construct each decision tree of the forest is a random subset of the total samples. They are generally drawn with replacement under the strategy of bootstrapping. Secondly, for each partitioning node, only a randomly-formed subset of features is used to split samples into binary branches. The size of this subset is generally around the square root of the number of total features. Under such conditions, RF is quite time consuming for its operation, but it has an excellent performance, with large tolerance to imperfections of samples, and good capacity to avoid overfitting. In our work, the function "RandomForestRegressor" from the python package "sklearn.ensemble" (Pedregosa *et al.* 2012) was used (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor). For the training procedure, we have data covering 34 years, from 1961 to 1994, and 928 spatial points. The total number of samples into our RF training is thus 34×928=31552. Each of the 24 climate models is treated as a feature in our RF implementation. After RF is trained, it is used in the testing period from 1995 to 2014 to validate its performance. Similarly, it is used to make the future projection under the specific warming thresholds.

The "Bayesian Optimization" was used to find the best hyperparameters implemented in the RF algorithm (Shahriari *et al* 2016) (http://rmcantin.github.io/bayesopt/html/boptttheory.html). It has a higher efficiency than other

methods, such as “Grid Search” or “Randomized Search”. We thus optimized 4 important parameters of the Random Forest algorithm, the number of trees in the forest (n_estimators), the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the number of features to consider when looking for the best split (max_features). Within 30 iterations, the Bayesian optimization process generally converges to optimal parameters for a specific climate index. The optimal parameters for the six indices are shown in Table S2.

As other statistical tools, machine learning methods don’t inspire confidence if we can’t ensure an appropriate interpretation on their derived features, patterns, and rules. In the RF model which consists of establishing a set of decision trees with internal nodes and leaves, the importance of input features or variables (climate models in our case) can be measured by the variance reduction attributed to each feature (total variance before the splitting node minus the sum of the same variance in the two split groups). In our case of multiple decision trees, the final measure of importance is the sum from all trees in the forest. It is furthermore normalized among all features or variables to ensure that the total sum is unity. This “relative importance” can help understanding the importance of each climate model in the ensemble-processing strategy. Relevant analysis and results are shown in Supplementary Materials Text. S1 and Figure S1.

2.2.3 Skill evaluation metrics

Taylor diagram (Taylor 2001) and skill score are standard tools providing a concise statistical summary of spatial characteristics between the simulation and observation. The Taylor diagram can show three aspects of statistical information: pattern correlation coefficient, a ratio of the centered standard deviations, and root mean square error, any two of them being independent (Li *et al* 2021). A good simulation would be that both the pattern correlation coefficient and the ratio of standard deviations are close to 1, and the root mean square error is close to 0 (Taylor 2001; Jiang *et al* 2015).

Taylor skill score (TSS), calculated as in Eq. (3), is a numerical summary of the Taylor diagram to express a synthetic measure.

$$\mathrm{TSS} = \frac{4\left(1+R_m\right)^2}{\left(\frac{\sigma_m}{\sigma_o}+\frac{\sigma_o}{\sigma_m}\right)^2\left(1+R_0\right)^2} \tag{3}$$

where R_m is the spatial correlation coefficient of climatological mean between simulation and observation, R_o is the maximum correlation coefficient attainable set here to 0.999, σ_m and σ_o are the standard deviations of the simulated and observed spatial patterns in climatological means, respectively. The closer the value of TSS is to 1, the better the agreement between the simulation and observation. This skill score has been generally used in many previous researches (Wang *et al* 2018; Zhu, H *et al* 2020; Ngoma *et al* 2021).

3. Results

3.1 Performance evaluation

To assess the ability of our three schemes dealing with the multi-model ensemble simulations, the spatial patterns and corresponding distribution boxplots for biases of all indices against observations across China during the validation period are examined (figure 2 and figure 3). Darker colors and far-from-zero bars represent higher deviations from observation. To facilitate visual inspection and interpretation of differential fields, the climatology from observation in the validation period from 1995 to 2014 is exhibited in figure S2. A general feature that can be observed in figures 2 and 3 is that the three schemes exhibit similar patterns of spatial bias distribution, and AM (with a bias correction included) shows the largest biases. Compared with AM, biases from LR and RF are reduced across almost the whole domain.

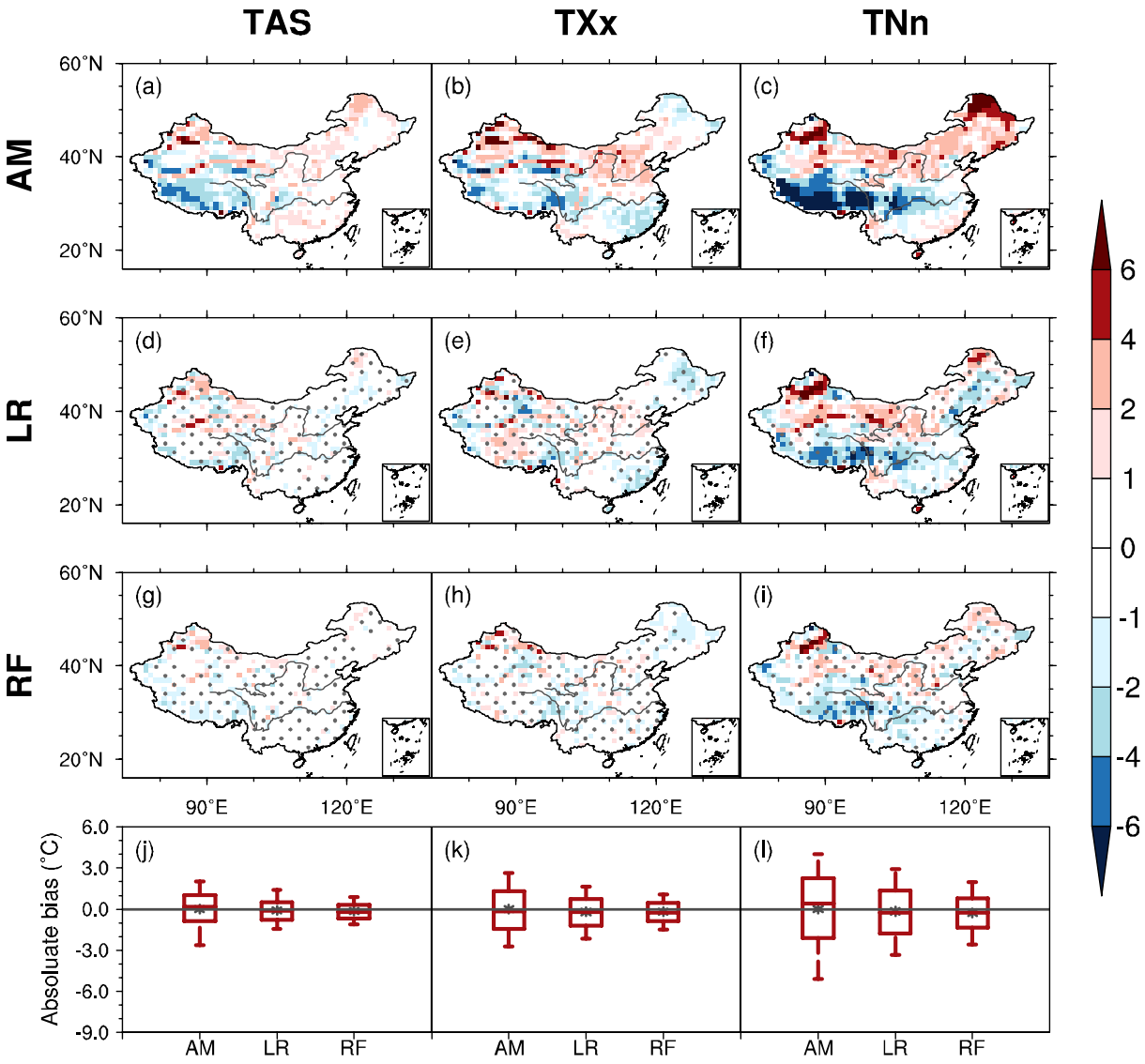

Figure 2. Spatial distributions (a–i) and corresponding boxplots (j–l) of the absolute biases from AM (a–c), LR (d–f), and RF (g–i) algorithms for mean and extreme temperature indices in the validation period (unit: °C). From left to right are TAS (column 1), TXx (column 2), TNn (column 3), respectively. Areas with significant amelioration based on AM above the 0.95 confidence level are marked with gray dots in the LR and RF panels, according to Student's t-test. The upper and lower limits of box are the first and third quartile, the horizontal line

and the asterisk in the box are the mean and median values, respectively, and the whiskers show the 10th and 90th percentile values.

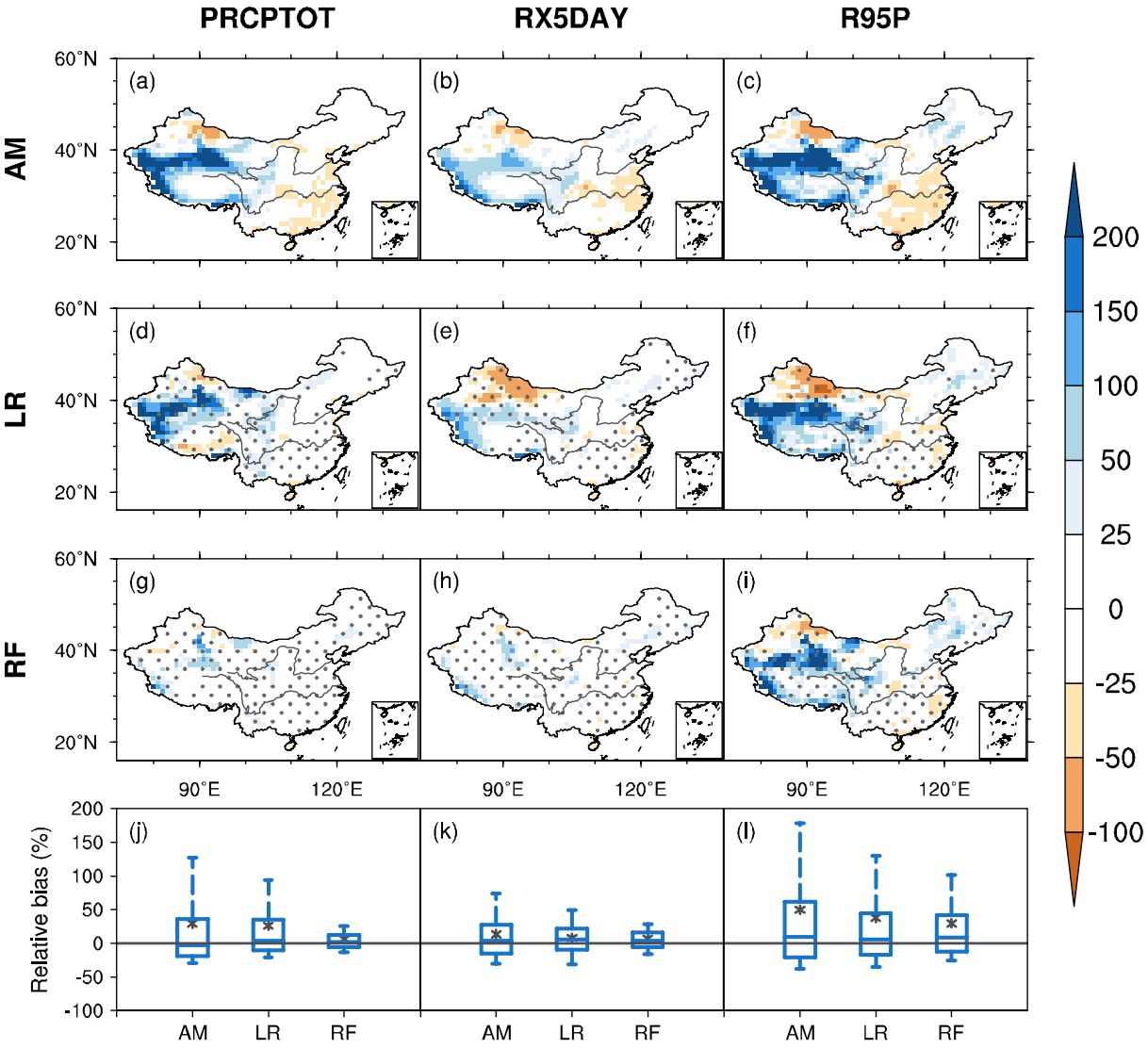

Figure 3. Same as figure 2, but the relative biases for mean and extreme precipitation indices (unit: %). From left to right are PRCPTOT (column 1), RX5DAY (column 2), R95P (column 3), respectively. Warm and cold colors indicate dry and wet biases respectively.

Cold biases ($\Delta T < -6°C$) from AM are mainly concentrated in the Tibetan Plateau and the middle and upper reaches of the Yangtze River for all temperature indices. They are significantly reduced in LR and RF (figure 2(d)–(i) vs 2(a)–(c)). The amelioration of RF is especially remarkable, only some scattered areas exist with bias exceeding 2°C. Higher cold biases (with focus on the 10th percentile biases in boxplots) depict a decrease from AM to RF, with values -2.62°C to -1.12°C for TAS, -2.73°C to -1.46°C for TXx, and -5.07°C to -2.58°C for TNn.

Similar characteristic holds true for precipitation indices (figure 3(d)–(i) vs 3(a)–(c)). Areas with large biases in AM are reduced in LR and RF, especially in the Tibetan Plateau and its periphery where there are the largest wet biases. Higher wet biases (with focus on the 90th percentile in the boxplots) are reduced from 127% in AM to 25% in RF for PRCPTOT. Similarly,

RX5DAY shows a reduction from 74% to 28%, and R95P from 178% to 101%. RF is the best performing, and the biases are lower than 50% over almost the whole territory of China for PRCPTOT and RX5DAY. Higher wet biases for R95P exist in the Tarim Basin and the Qilian Mountains with complex topography.

Taylor diagram and Taylor skill score are presented in figure 4 to show a concise statistical analysis of the three ensemble-processing strategies in the evaluation period. There is a general weak performance with AM (gray markers). The correlation coefficients of all temperature indices deduced from AM are between 0.94–0.97, the standardized deviations vary between 0.96–1.05, and the Taylor skill scores are lower than 0.97. LR and RF schemes show an extra improvement, compared to AM. The best-performing RF gives correlation coefficient, and Taylor skill scores all superior to 0.98–0.99. Precipitation indices from AM show an unsatisfactory performance, with all spatial correlation coefficients less than 0.88, standardized deviations between 0.64–0.80, and the lowest value of Taylor skill scores reaching only 0.71. RF has the best efficiency, with precipitation indices comparable to temperature indices. In terms of Taylor skill scores, RF improves them from 0.82 in AM to 0.98 for PRCPTOT, from 0.79 to 0.95 for RX5DAY and from 0.71 to 0.89 for R95P.

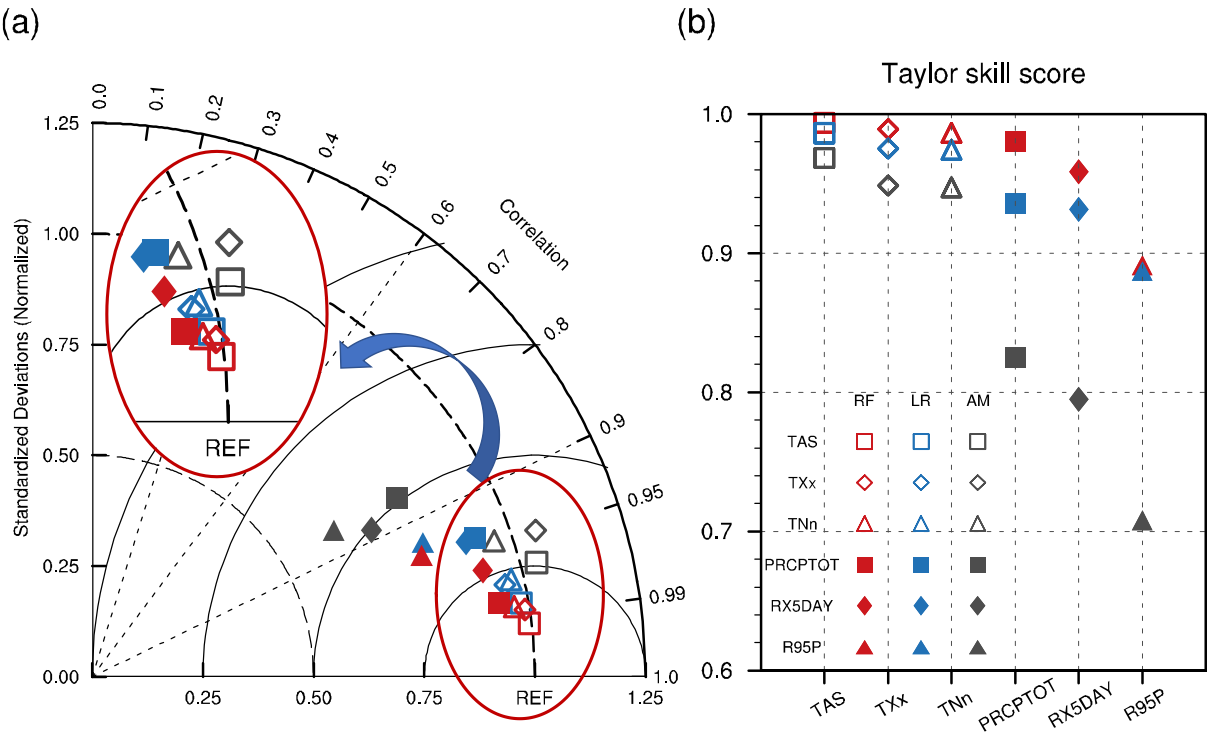

Figure 4. Taylor diagram (a) and Taylor skill score scatter plot (b) showing the mean and extreme temperature and precipitation indices under the three ensemble-processing schemes (represented with colors) during the validation period. Different symbols represent different indices, with hollow symbols for temperature indices and solid symbols for precipitation indices.

Overall, the results provide clear evidence that LR and RF schemes effectively enhance the capability of reproducing the spatial climate characteristics in China, especially in western China where, with complex topography, most significant biases manifest in AM. RF has the best performance, with Taylor skill scores of all temperature indices at the level of 0.98–0.99, and remarkably improves the skill scores of precipitation indices to a level higher than 0.89. Temperature indices generally have a better performance than precipitation, but the

improvement for precipitation indices is more significant and substantial.

Beyond the mean state, it is also interesting to check how well our ensemble-processing schemes can produce their interannual variability. We now assess the temporal standard deviation during the validation period from 1995 to 2014, with the interannual standard deviation from observation shown in figure S3. The result of evaluation is shown in figures S4, and expressed as a ratio of standard deviations between the simulation and observation. Only the mean states of TAS and PRCPTOT are exhibited as illustration. This ratio is generally smaller than 1.0, reflecting the fact that the ensemble-processing strategies present a reduced interannual variability. Such a result is expected, since any ensemble-processing strategy, due to its nature of mixing different simulations, reduces the interannual variability. In the case of AM, if all members in the ensemble are sequentially independent and possess an identical standard deviation, then the ensemble average from N members would reduce the standard deviation by a factor of $1/\sqrt{N}$. In our configuration of 24 models, this factor is about 0.20. The actual ratio for the mean temperature indices is larger than this expected value, but its counterpart for precipitation indices is smaller (all indices are not shown). We believe that this behavior is due to the fact that temperature indices have a consistent warm trend among models, but precipitation indices do not. Let us now inspect the cases of RF and LR, since a regression relationship is used to combine the 24 models (or a subset), the reduction of interannual variability is less pronounced. It is necessary to point out that when the regression is ill-fitted (with large negative coefficients for certain members, for example), the interannual variability can even be augmented.

3.2 Projection of future climate

Given its good performance in dealing with multi-model ensemble simulations, RF is now used for the regional projection of future climate for 1.5°C, 2°C and 3°C global warming targets (relative to preindustrial), under the SSP5-8.5 emission scenario. The widely-used AM scheme is also shown as a baseline and reference. As a conventional practice, the target warming levels are relative to pre-industrial (1861–1900), while the projected changes are relative to 1995–2014. For the sake of conciseness, only temperature and precipitation indices under 1.5°C and 2°C warming targets are given in the main text, the results under 3°C warming target being placed in Supplementary Materials.

3.2.1 Temperature indices

The land fraction from whole China territory with projected changes exceeding the abscissa's values is plotted in figure 5(a)–(c) in the form similar to a curve of the cumulative probability distribution function. Results are shown for both RF and AM, for all temperature indices, and for the 1.5°C and 2°C warming targets, respectively. figure 5(d)–(o) show their corresponding spatial pattern of changes, while the difference between RF and AM under the 2°C warming level is shown in figure 5 (p)–(r).

Let us firstly examine the median value which is an emblematic figure since it separates the entire territory across China into two equal halves. Changes of mean and extreme temperature projected by RF are lower than those by AM (figure 5). Under the 2°C warming target, but relative to nowadays, RF shows a median change of TAS, TXx, and TNn at 1.35°C, 1.37°C and 1.64°C, which are lower than the counterpart in AM, by about 0.23°C, 0.31°C and 0.15°C. Recent studies based on CMIP6 models show higher transient climate response and

equilibrium climate sensitivity than what shown by previous versions of these models in CMIP5. Consequently, the projection of future climate in CMIP6 is also stronger than in CMIP5 (Gettelman *et al* 2019; Nijsse *et al* 2020; Zelinka *et al* 2020). However, with some observational constraints, the projected warming is reduced compared with non-constrained projection (Brunner *et al* 2020; Liang *et al* 2020; Tokarska *et al* 2020). Our results of multi-model ensemble projection seem to agree with this conclusion. Observation plays important role in the machine learning RF ensemble-processing scheme, similar to a role of observation-based constraining, which lowers the projected warming compared to unconstrained AM.

Obvious differences are detected between the land fraction curves of RF and AM. Extreme changes have higher probability of occurrence in more areas in RF as its curves have longer tails. Under the 2°C global warming target, AM does not project warmer mean temperature exceeding 2.6°C, but RF suggests a likelihood of 9% over China with such a warming level. In terms of geographic distribution of TAS, larger spatial variability is detected in RF, as the spatial standard deviations are almost twice larger than that in AM (figure 5(j) vs. 5(m)). Large magnitudes of warming projected by RF are found in the western part of Northeastern China and the north part of Northwestern China under 1.5°C warming. Under the 2°C warming target, the warming in these areas would further expand and strengthen, the northern and eastern periphery of the Tibetan Plateau also shows significant warming, exceeding 2.5°C. Meanwhile, AM projects a smoother distribution, with warming uniformly enhanced (exceeding 2.5°C warming) in the area north of 45°N and part of the Tibetan Plateau. From the difference between RF and AM (figure 5(p)), it is clear that, except a few regions with drastic increases in RF, the warming projected from RF is generally lower about 0.25°C–1°C than that of AM in almost the whole country, especially in the Tibetan Plateau, where the difference is significant under the 0.95 confidence level.

Regarding the spatial pattern, a similar behavior holds for extreme temperature TXx. Significantly enhanced warming over Northeastern China, the Tianshan Mountains, as well as the Loess Plateau is projected in RF, with a magnitude of 2.5°C above current world under the 2°C warming target. Areas with larger increases from AM are evenly distributed in Northeastern China and the entire Northwest region. The change of TXx in different regions has large distinction in the projection of RF, the spatial standard deviations are more than three times larger than that in AM under both 1.5°C and 2°C warming targets (figure 5(e) vs. 5(h) and figure 5(k) vs. 5(n)).

For the minimum temperature TNn, sensitive areas from RF are distributed in Northeastern China, in the Yellow River Basin and in sparse areas in Northwestern China, while the warming projected by AM is more widely distributed in the southeast, extending to the south of the Yangtze River Basin.

These results show broad similarities with those from GCMs (Shi *et al* 2018; Sui *et al* 2018; Yang *et al* 2018), i.e., Northwestern China, Northeastern China, and the Tibetan Plateau are particularly sensitive to global warming. Compared to AM, RF shows more detailed information and larger inhomogeneity, and it exhibits a closer correlation with topography. More pronounced hotspots can be observed in RF.

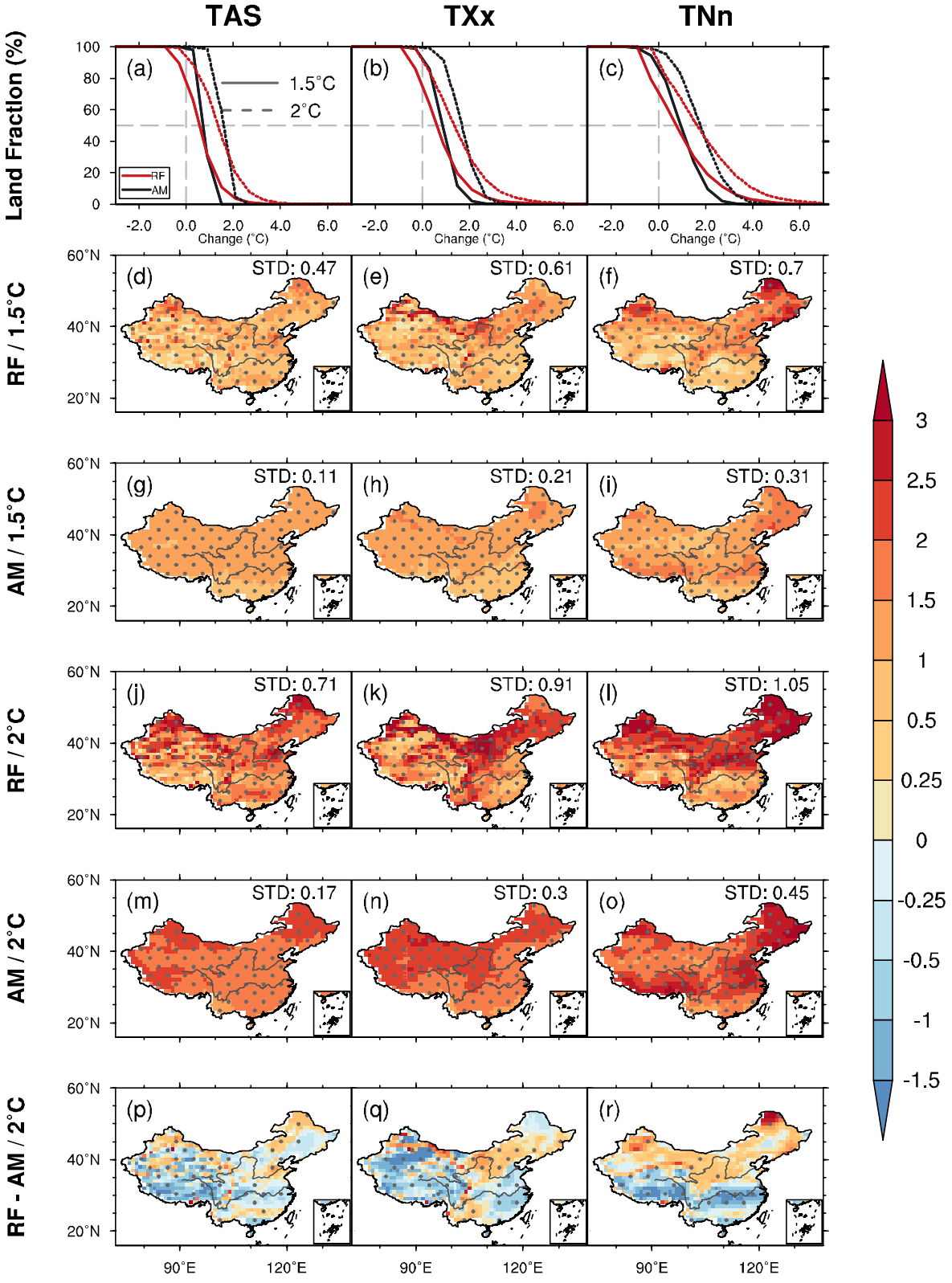

Figure 5. The land fractions (a–c) and corresponding spatial distributions (d–r) of the changes projected from RF (colored lines and panels d–f, j–l) and AM (black lines and panels g–i, m–o) for TAS (column 1), TXx (column 2), TNn (column 3) at the 1.5°C and 2°C global warming relative to the reference period. The solid lines in land fraction plots and rows 2, 3 (panels d–i) are at 1.5°C global warming target; dash lines and rows 4, 5 (panels j–o) are at 2°C. Panels (p–r) are the spatial distributions of distinctions between RF and AM at the 2°C global warming. The spatial standard deviations (STD) over the country are given on the top of panels d–o. Areas with significant

changes above the 0.95 confidence level with reference period are marked with gray dots in panels (d–o), and areas with significant differences above the 0.95 confidence level between RF and AM are marked with gray dots in panels (p–r), according to Student's t-test (unit: °C).

3.2.2 Precipitation indices

Mean and extreme precipitation projections are presented in figure 6, both RF and AM project increased precipitation over most of China in response to global warming. For the median value across China that separates the whole territory into two equal halves, RF shows an increase of 3%, 4%, and 19 % for PRCPTOT, RX5DAY, and R95P under the 2°C global warming, which is almost the same as the counterpart in AM (only about 0.6%, 0.1% and 1.5% higher). For the change of total precipitation (PRCPTOT) under the 1.5°C and 2°C global warming targets, the fraction of lands where increase exceeds 40% is projected to be almost non-existent over China in AM, while that fraction in RF is above 6%. That means a higher risk for intense precipitation in RF projection.

In terms of geographic distribution, RF and AM show good consistency, but there are substantial differences of magnitude (figure 6(d)–(r)). Small-scale features in RF are more significant, and the amplitude of increase is also higher. Large-increase areas of total precipitation (more than 30%) in RF are concentrated in the region of the Tsaidam Basin and Qilian Mountains. For the case of AM, enhanced precipitation (10 to 20%) is more evenly distributed in the whole western area, extends from the Tibetan Plateau, northeastward, stretching to the Loess Plateau and its northern area. In the northwestern area, significant lower precipitation change is projected by RF compared with AM. In other areas, the changes of total precipitation projected by RF are generally more notable than that in AM.

Changes of RX5DAY show a close resemblance to total precipitation in terms of intensity and main geographic patterns. But precise areas of remarkable increase have some differences, especially in RF. Significant enhancement is found in most part of the Tibetan Plateau and patchy areas in Northeastern China, where the magnitudes exceed 20% under the 2°C global warming. Contrasted with RF, AM suggests smaller increases of RX5DAY, but with a more homogeneous geographic distribution. Almost all the territory would see an increase within the 15% threshold. As shown in figure 5(q), significant differences between AM and RF are found in the northwestern region.

Changes of R95P exceeding 50% in RF concentrate in the Tibetan Plateau and the Yellow River Basin, where the changes are almost twice higher than in AM (higher about 20%–30%). Meanwhile, in the southeastern and northwestern regions, the projected increase in strong precipitation from RF are not noticeable, which is lower than that projected by AM.

Further comparison of our RF projections with previous studies using high-resolution regional climate models (RCMs) shows some similarities, especially in complex-terrain areas. Zhu, X *et al* (2020), using WRF v3.7.1, showed that, for total precipitation and extreme events, the Tibetan Plateau and regions outside China's northwestern boundaries are particularly sensitive to climate change, conclusion very consistent with our results. Similar patterns from our RF projection for RX5DAY were also present in Li, H *et al* (2018) using five RCMs involved in the CORDEX-East Asia project. Our results are also comparable to Li, D *et al* (2018) using FROALS as a dynamical downscaling model, together with a statistical downscaling tool. It is worthy of note that our projected R95P pattern in RF is very close to what found with WRF v3.5.1 when it was applied to China (Bao *et al* 2015).

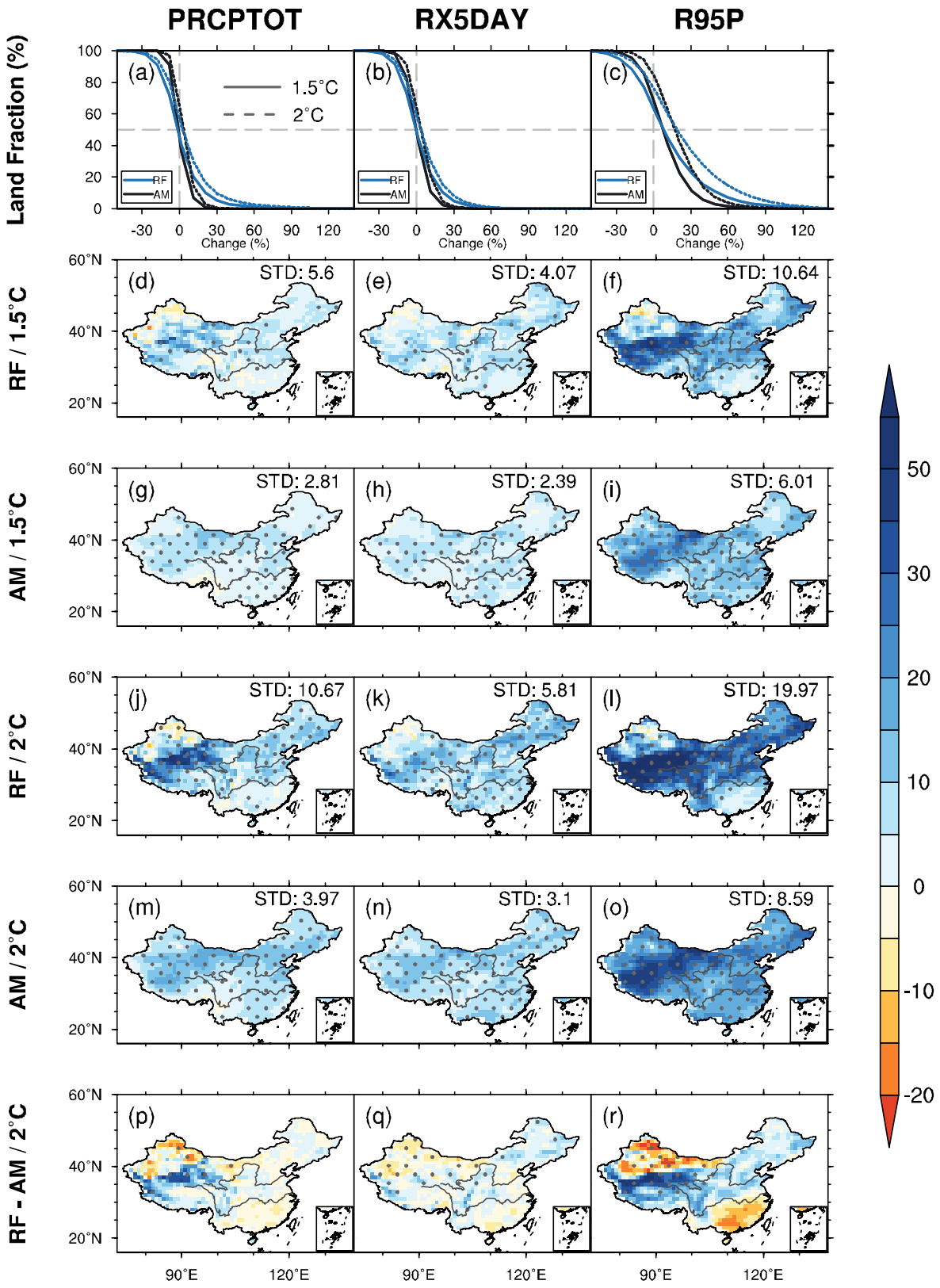

Figure 6. Same as figure 5, but the relative changes of precipitation indices PRCPTOT, R95P, and RX5DAY (unit: %).

The machine learning RF algorithm uses the concept of multi-regression decision trees. It can efficiently solve nonlinear regression problems and achieve good matching to observation,

in both temporal and spatial domains, as well demonstrated in Crawford *et al* (2019) and Pang *et al* (2017). Our results shown here are consistent with its intrinsic properties and with our expectation on it.

4. Conclusion and Discussion

In this work, three different ensemble-processing strategies, AM (arithmetic mean), LR (linear regression), and RF (Random Forest, machine learning decision tree algorithm), are used to explore information offered by the multi-model ensemble climate simulations of CMIP6. The main idea was to find the best way of processing the ensemble simulations to mimic observational climatic properties and to give a more reliable projection of future climate. AM is the simplest and most intuitive strategy. LR advocates the vision of a linear-regression approach to establish the relationship between simulations and observations, but it cannot necessarily represent any physical rules governing the climate system. RF is one of the most advanced machine-learning algorithms. It can extract non-linear and complex relations among climate models, instead of making a simple evaluation of models' apparent performance as in other ensemble-processing strategies. This leads to a hybrid approach that we advocate for climate change issues, which combines physical modelling and machine learning strengths, thus giving confidence in retrieving more valuable information.

The performance of the three schemes was assessed in the validation period (20 years, from 1995 to 2014). Compared with AM, LR and RF effectively enhance the capability of capturing spatial climate characteristics over China. Improvement in areas with complex terrain is the most significant such as in the periphery of the Tibetan Plateau. RF performs well, with the Taylor skill score of temperature indices being of 0.98–0.99, and that of precipitation indices higher than 0.89. It was also revealed that the internal variability, such as the interannual-scale standard deviation, can not be correctly reproduced by any of our ensemble-processing strategies which were designed, after all, to calculate the mean state of our expectation.

After an inter-comparison of performance, RF was selected as the optimal scheme and used to investigate climate changes in the 1.5°C, 2°C and 3°C warmer worlds under the SSP5-8.5 emission scenario. Compared with AM, RF shows less warming and enhanced wet conditions at the national scale of China. In terms of median changes across China, mean temperature (TAS), annual maximum (hottest daytime) temperature (TXx), and annual minimum (coldest nighttime) temperature (TNn) show 1.35°C, 1.37°C and 1.64°C warming relative to 1995–2014 period, respectively, under the 2°C global warming level, when RF is used. They are lower than their counterpart in AM, especially for TXx, lower about 0.31°C. The median changes of total precipitation in wet days (PRCPTOT), annual maximum consecutive 5-day precipitation amount (RX5DAY), and annual total precipitation for events exceeding the 95th percentile (R95P) projected in RF are 3%, 4%, and 19%, respectively, similar with the counterpart in AM.

Regarding the geographic distribution, RF would see larger warming in Northeastern China and the northern part of Northwestern China. Tianshan Mountain, Loess Plateau area for TXx, and the Yellow River Basin for TNn are also regions of hotspots. Meanwhile except the regions with intensified warming, the warming projected from RF is generally lower than that of AM. That indicates a larger spatial variability and more pronounced local-scale characteristics of RF. For the projection of TXx, the spatial standard deviation can be three times larger compared with that in AM.

RF also projects more intense precipitation in most part of China. For example, in the region of the Tsaidam Basin and the Qilian Mountains, the projected changes in RF for the strong precipitation (partly exceeding 50% under the 2°C warming) are almost twice higher than in AM. Meanwhile in the northwestern area, for all precipitation indices, weaker increases of precipitation compared with AM are projected by RF. AM shows however much more homogeneous features.

It is interesting to point out that the geographic structure of climate projection in RF shows a resemblance to that from dynamical downscaling with high-resolution models or from statistical downscaling (Li, D *et al* 2018; Zhu, X *et al* 2020). This indicates that the machine-learning algorithm RF could capture detailed information at local scale, certainly due to its ability to behave as do those dynamic models with higher spatial resolution. This is quite reasonable since the high-resolution observation seems to play its role in constraining the ensemble-processing strategy RF which is able to manipulate complex nonlinear processes across multiple models. We believe that using advanced machine-learning techniques can provide a new perspective to retrieve more information from large amounts of data and make more reliable climate projections.

Acknowledgments

We would like to acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP6. We thank the climate modeling groups for producing simulations and making their model outputs available. This work is supported by the National Key Research and Development Program of China (Grant 2017YFA0603804, Grant 2018YFC1507704).

References

Alizamir, M., M. Azhdary Moghadam, A. Hashemi Monfared, and A. Shamsipour 2018 Statistical downscaling of global climate model outputs to monthly precipitation via extreme learning machine: A case study. *Environmental Progress & Sustainable Energy*. **37** 1853-1862

Bao, J., J. Feng, and Y. Wang 2015 Dynamical downscaling simulation and future projection of precipitation over China. *Journal of Geophysical Research: Atmospheres*. **120** 8227-8243

Breiman, L. 2001 Random forests. *Machine Learning*. **45** 5-32

Breiman, L., J. Friedman, R. Olshen, and C. J. Stone, 1984: Classification and regression trees. Chapman and Hall/CRC, 368 p.

Brunner, L., 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 System Dynamics*. **11** 995-1012

Choubin, B., S. Khalighisigaroodi, A. Malekian, and O. Kisi 2016 Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. *Hydrological Sciences Journal-journal Des Sciences Hydrologiques*. **61** 1001-1009

Crawford, J., K. Venkataraman, and J. Booth 2019 Developing climate model ensembles: A comparative case study. *J. Hydrol.* **568** 160-173

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

Fu, C., Z. Jiang, Z. Guan, J. He, and Z. Xu, 2008: *Impacts of climate change on water resources and agriculture in China.* Springer Berlin Heidelberg.

Gettelman, A., C. Hannay, J. T. Bacmeister, R. Neale, A. G. Pendergrass, G. Danabasoglu, J. Lamarque, J. T. Fasullo, D. A. Bailey, and D. M. Lawrence 2019 High climate sensitivity in the community earth system model version 2 (cesm2). *Geophys. Res. Lett.* **46** 8329-8337

Guo, L., Z. Jiang, M. Ding, W. Chen, and L. Li 2019 Downscaling and projection of summer rainfall in eastern China using a nonhomogeneous hidden Markov model. *Int. J. Climatol.* **39** 1319-1330

Guo, L., Z. Jiang, D. Chen, H. Le Treut, and L. Li 2020 Projected precipitation changes over China for global warming levels at 1.5 °C and 2 °C in an ensemble of regional climate simulations: Impact of bias correction methods. *Climatic Change*. **162** 623-643

Ham, Y. G., J. H. Kim, and J. J. Luo 2019 Deep learning for multi-year enso forecasts. *Nature*. **573** 568-572

Hidalgo, H. G., and E. J. Alfaro 2015 Skill of CMIP5 climate models in reproducing 20th century basic climate features in central America. *Int. J. Climatol.* **35** 3397-3421

Hulme, M. 2016 1.5 °C and climate research after the Paris agreement. *Nature Climate Change*. **6** 222-224

Jiang, Z. H., W. Li, J. J. Xu, and L. Li 2015 Extreme precipitation indices over China in CMIP5 models. Part I: Model evaluation. *J. Climate*. **28** 8603-8619

Kadow, C., D. M. Hall, and U. Ulbrich 2020 Artificial intelligence reconstructs missing climate information. *Nature Geoscience*. **13** 408-413

Knutti, R., D. Masson, and A. Gettelman 2013 Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.* **40** 1194-1199

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl 2010 Challenges in combining projections from multiple climate models. *J. Climate*. **23** 2739-2758

Knutti, R., 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

Lenderink, G., A. Buishand, and W. van Deursen 2007 Estimates of future discharges of the river rhine using two scenario methodologies: Direct versus delta approach. *Hydrology and Earth System Sciences*. **11** 1145-1159

Li, D., L. Zou, and T. Zhou 2018 Extreme climate event changes in China in the 1.5 and 2 °C warmer climates: Results from statistical and dynamical downscaling. *Journal of Geophysical Research: Atmospheres*. **123** 10215-10230

Li, H., H. Chen, H. Wang, and E. Yu 2018 Future precipitation changes over China under 1.5 degrees C and 2.0 degrees C global warming targets by using cordex regional climate models. *Sci Total Environ*. **640-641** 543-554

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. Meteorol. Res.* **35** 184-197

Li, X., Z. Li, W. Huang, and P. Zhou 2020 Performance of statistical and machine learning ensembles for daily temperature downscaling. *Theor. Appl. Climatol.* **140** 571-588

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

Ngoma, H., W. Wen, B. Ayugi, H. Babaousmail, R. Karim, and V. Ongoma 2021 Evaluation of precipitation simulations in CMIP6 models over uganda. *Int. J. Climatol.* 1-26

Nijsse, F. J. M. M., P. M. Cox, and M. S. Williamson 2020 An emergent constraint on transient climate response from simulated historical warming in CMIP6 models. *Earth System Dynamics*. **2020** 1-14

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblasreyes, T. Jung, and M. Leutbecher 2005 Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.* **33** 163-193

Pang, B., J. Yue, G. Zhao, and Z. Xu 2017 Statistical downscaling of temperature with the random forest model. *Advances in Meteorology*. **2017** 1-11

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2012 Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*. **12** 2825-2830

Piao, S., P. Ciais, Y. Huang, Z. Shen, S. Peng, J. Li, L. Zhou, H. Liu, Y. Ma, and Y. Ding 2010 The impacts of climate change on water resources and agriculture in China. *Nature*. **467** 43-51

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat 2019 Deep learning and process understanding for data-driven earth system science. *Nature*. **566** 195-204

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard 2004 Improved combination of multiple atmospheric gcm ensembles for seasonal prediction. *Mon. Wea. Rev.* **132** 2732-2744

Sanderson, B. M., R. Knutti, and P. Caldwell 2015 Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*. **28** 5150-5170

Semenov, M. A., and P. Stratonovitch 2010 Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Climate Research*. **41** 1-14

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas 2016 Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*. **104** 148-175

Shi, C., Z. H. Jiang, W. L. Chen, and L. Li 2018 Changes in temperature extremes over China under 1.5 °C and 2 °C global warming targets. *Advances in Climate Change Research*. **9** 120-129

Sui, Y., X. Lang, and D. Jiang 2018 Projected signals in climate extremes over China associated with a 2 °C global warming under two RCP scenarios. *Int. J. Climatol.* **38** e678-e697

Sun, C., Z. Jiang, W. Li, Q. Hou, and L. Li 2019 Changes in extreme temperature over China when global warming stabilized at 1.5 °C and 2.0 °C. *Sci Rep*. **9** 14982

Tan, J., Z. Jiang, and T. Ma 2016 Projections of future surface temperature change and uncertainty over China based on bayesian model averaging. *Acta Meteor. Sinica*. **74** 583-597

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*. **106** 7183-7192

Teutschbein, C., and J. Seibert 2012 Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.* **456-457** 12-29

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. *Science Advances*. **6** eaaz9549

UNFCCC: United nations framework convention on climate change (2015). Decision 1/cp.21. The Paris agreement. [Available online at http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.]

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

Watson-Parris, D. 2020 Machine learning for weather and climate are worlds apart. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **379** 20200098

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

Xu, L., N. Chen, X. Zhang, and Z. Chen 2020 A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Climate Dyn.* **54** 3355-3374

Xu, Y., X. Gao, and F. Giorgi 2010 Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections. *Climate Research*. **41** 61-81

Yang, Y., J. Tang, S. Wang, and G. Liu 2018 Differential impacts of 1.5 and 2 °C warming on extreme events over China using statistically downscaled and bias-corrected CESM Low-Warming experiment. *Geophys. Res. Lett.* **45** 9852-9860

Zelinka, M. D., T. A. Myers, D. T. Mccoy, S. Pochedley, P. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor 2020 Causes of higher climate sensitivity in CMIP6 models. *Geophys. Res. Lett.* **47**

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers 2011 Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews Climate Change*. **2** 851-870

Zhu, H., Z. Jiang, J. Li, W. Li, C. Sun, and L. Li 2020 Does CMIP6 inspire more confidence in simulating climate extremes over China? *Adv. Atmos. Sci.* **37** 1119−1132

Zhu, X., Z. Wei, W. Dong, Z. Ji, X. Wen, Z. Zheng, D. Yan, and D. Chen 2020 Dynamical downscaling simulation and projection for mean and extreme temperature and precipitation over central asia. *Climate Dyn.* **54** 3279-3306

Supplementary Material for

Machine learning to optimize climate projection over China with multi-model ensemble simulations

Contents of this file

Table. S1. Model number, model name, modeling center and country, and atmospheric resolution of 24 CMIP6 global climate models (Expansions of acronyms are available at http://www.ametsoc.org/PubsAcronymList)

Model Number	Model Name	Modeling Center/ Country	Resolution (lat×lon)
1	ACCESS-CM2	Commonwealth Scientific and Industrial Research Organisation /Australia	1.25°×1.875°
2	ACCESS-ESM1-5		1.25°×1.875°
3	BCC-CSM2-MR	Beijing Climate Center China Meteorological Administration /China	1.125°×1.125°
4	CanESM5	Canadian Centre for Climate Modelling and Analysis /Canada	2.8°×2.8°
5	CNRM-CM6-1	Centre National de Recherches Météorologiques– Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique /France	1.4°×1.4°
6	CNRM-ESM2-1		1.4°×1.4°
7	EC-Earth3	EC-EARTH consortium	0.7°×0.7°
8	EC-Earth3-Veg		0.7°×0.7°
9	FGOALS-g3	Chinese Academy of Sciences /China	2.25°×2°
10	GFDL-CM4	NOAA Geophysical Fluid Dynamics Laboratory /USA	1°×1.25°
11	GFDL-ESM4		1°×1.25°
12	HadGEM3-GC31-LL	Met Office Hadley Centre /UK	1.25°×1.875°
13	INM-CM4-8	Institute for Numerical Mathematics, Russian Academy of Science /Russia	1.5°×2°
14	INM-CM5-0		1.5°×2°
15	IPSL-CM6A-LR	Institut Pierre-Simon Laplace /France	1.26°×2.5°
16	MIROC6	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, The University of Tokyo, National Institute for Environmental Studies, and RIKEN Center for Computational Science /Japan	1.4°×1.4°
17	MIROC-ES2L		2.8°×2.8°
18	MPI-ESM-1-2-HR	Max Planck Institute for Meteorology /Germany	0.9375°×0.9375°
19	MPI-ESM-1-2-LR		1.875°×1.875°
20	MRI-ESM2-0	Meteorological Research Institute /Japan	1.125°×1.125°
21	NESM3	Nanjing University of Information Science and Technology /China	1.875°×1.875°
22	NorESM2-LM	Norwegian Climate Centre /Norway	1.875°×2.5°
23	NorESM2-MM		0.9375°×1.25°
24	UKESM1-0-LL	Met Office Hadley Centre /UK	1.25°×1.875°

Table. S2. Optimal parameters for RF models in reproducing the six observed climate indices. The four tunable hyperparameters used in sklearn.ensemble.RandomForestRegressor are: "n_estimators", the number of trees in the forest; "max_depth", the maximum depth of the tree; "min_samples_split", the minimum number of samples required to split an internal node; and "max_features", the number of features (expressed as a fraction of the number of total features) to consider when looking for the best split.

Climate Index	n_estimators	max_depth	min_samples_split	max_features
TAS	697	31	5	0.52
TXx	909	18	5	0.32
TNn	1697	29	9	0.25
PRCPTOT	1470	20	5	0.27
RX5DAY	1528	45	13	0.16
R95P	1628	30	9	0.12

Text. S1. Analysis of the "relative importance" in RF algorithm

In this study, our goal is to use multi-model climate simulations (from 24 CMIP6 models) as input to make the best combination possible. Figure S1 shows the "relative importance" of each model when RF is applied to each of the indices. This helps to further understand which of the CMIP6 models are more suitable for climate simulations in China. It can be seen that the importance for different indices is variable, and there are only some similarities among the three precipitation indices. TAS shows larger disparity with obviously big and small contributions. EC-Earth3-Veg reaches 0.35, while some other models manifest almost no contribution. In the contrary, R95P shows all models' importance between 0.01 and 0.08, indicating a quite uniform distribution. Considering the nature and properties of the "relative importance" as presented in the main text, we need to note that this measure does reveal the performance of climate models, but it is also very dependent on hyper-parameters and precise implementation of the RF algorithm.

Nevertheless, what shown in Figure S1, if interpreted as a kind of performance of climate models, is roughly consistent with results from previous studies that were recently reported (Dong and Dong 2021; Yang *et al* 2021). The two models from EC-EARTH (EC-Earth3-Veg and EC-Earth3) are generally well ranked for all indices except TNn, especially for their mean climate state (TAS and PRCPTOT). HadGEM3-GC31-LL has a good performance for all the precipitation indices but a mediocre one for the temperature indices. INM-CM4-8 and INM-CM5-0, with a strong relationship and from a same institution, have similar relative importances for all climate indices.

Text. S2. Projection under the 3°C warming target

Under the 3°C global warming threshold, projected temperature and precipitation indices (figure S5) show consistent behaviors with what shown in the main text for the case of 1.5°C and 2°C warming targets, but there are more significant changes relative to nowadays.

Corresponding to the global warming threshold of 3°C, the mean temperature TAS over China projected by RF increases by 2.91°C. TXx and TNn increase about 2.95°C and 3.60°C, respectively. These temperature changes are all greater than their global averages. In terms of geographic distribution, stronger changes are generally observed in high latitudes than in mid-low latitudes.

For precipitation indices, PRCPTOT, RX5DAY and R95P all tend to increase in RF and AM. RF shows more detailed local features in terms of geographic distribution, consistent with the 1.5 and 2°C global warming. In terms of intensity of changes projected from RF, PRCPTOT increases by 18%, RX5DAY by 17%, and R95P by 47%. For all precipitation indices, the greatest projected changes appear in the north part of the Tibetan Plateau, with extension to northeast regions.

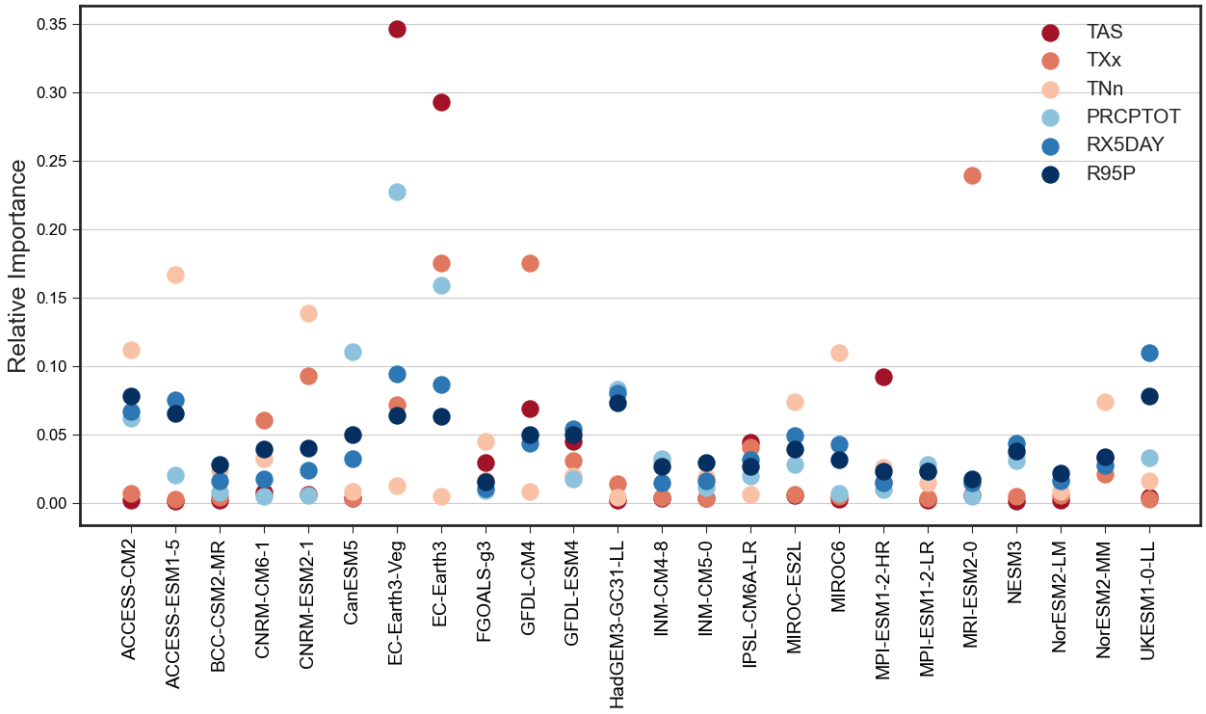

Figure S1. Relative importance of input features (24 CMIP6 models) deduced with the RF ensemble-processing strategy for the mean and extreme temperature and precipitation indices.

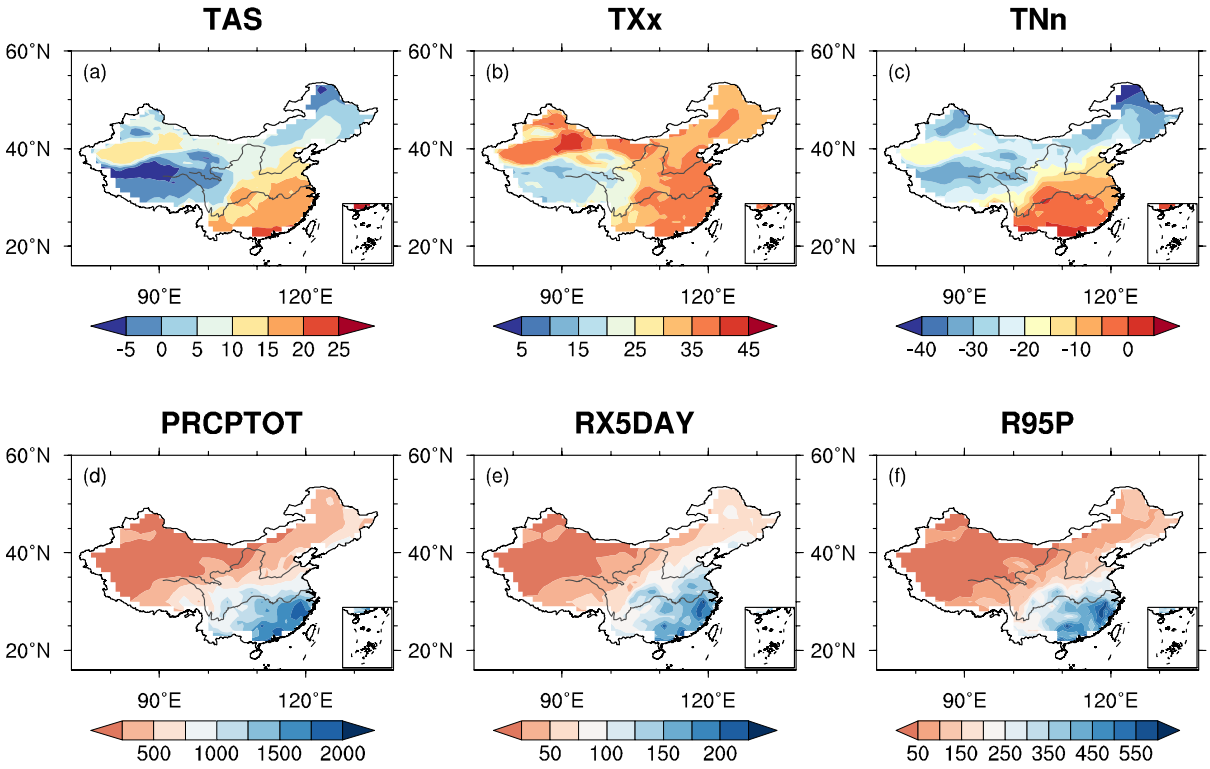

Figure S2. Spatial distributions of observed climatology for the mean and extreme temperature and precipitation indices in the validation period from 1995 to 2014 (unit: °C for temperature indices in panels a to c, and mm for precipitation indices in panels d to f).

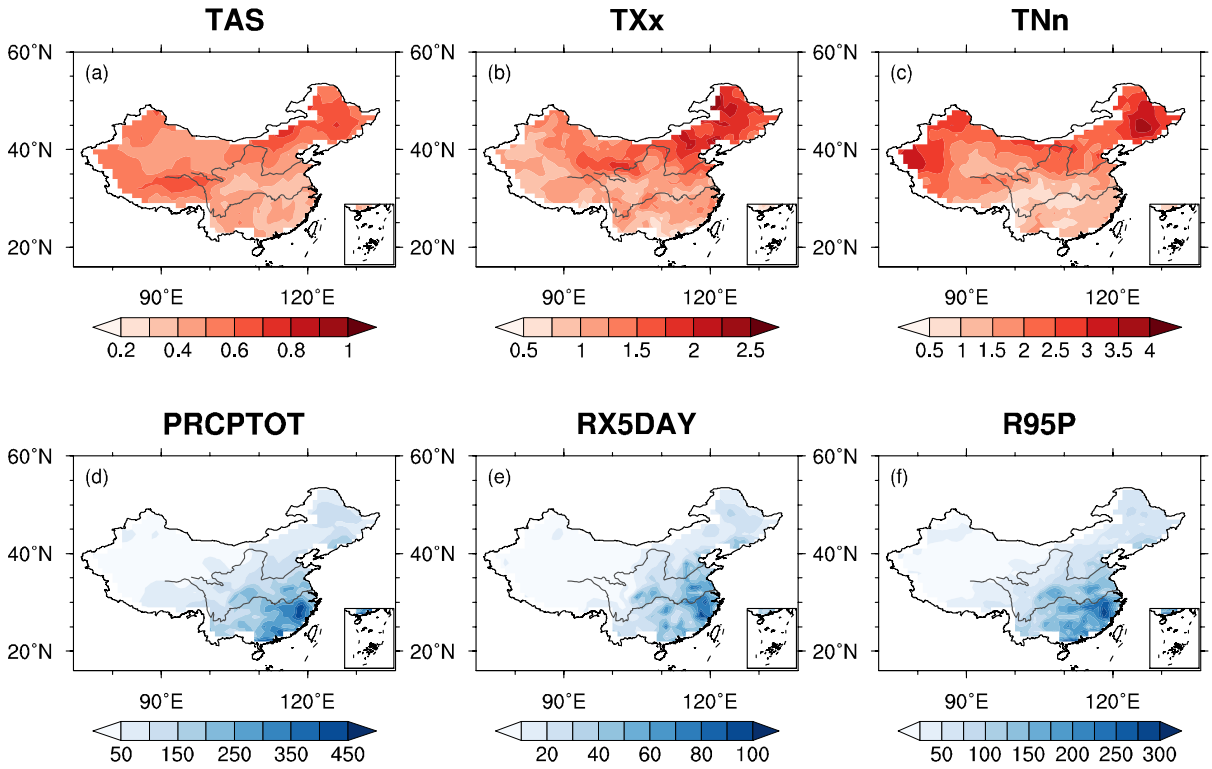

Figure S3. Similar as figure S2, but for the spatial distributions of observed interannual standard deviation (unit: °C for temperature indices in panels a to c, and mm for precipitation indices in panels d to f).

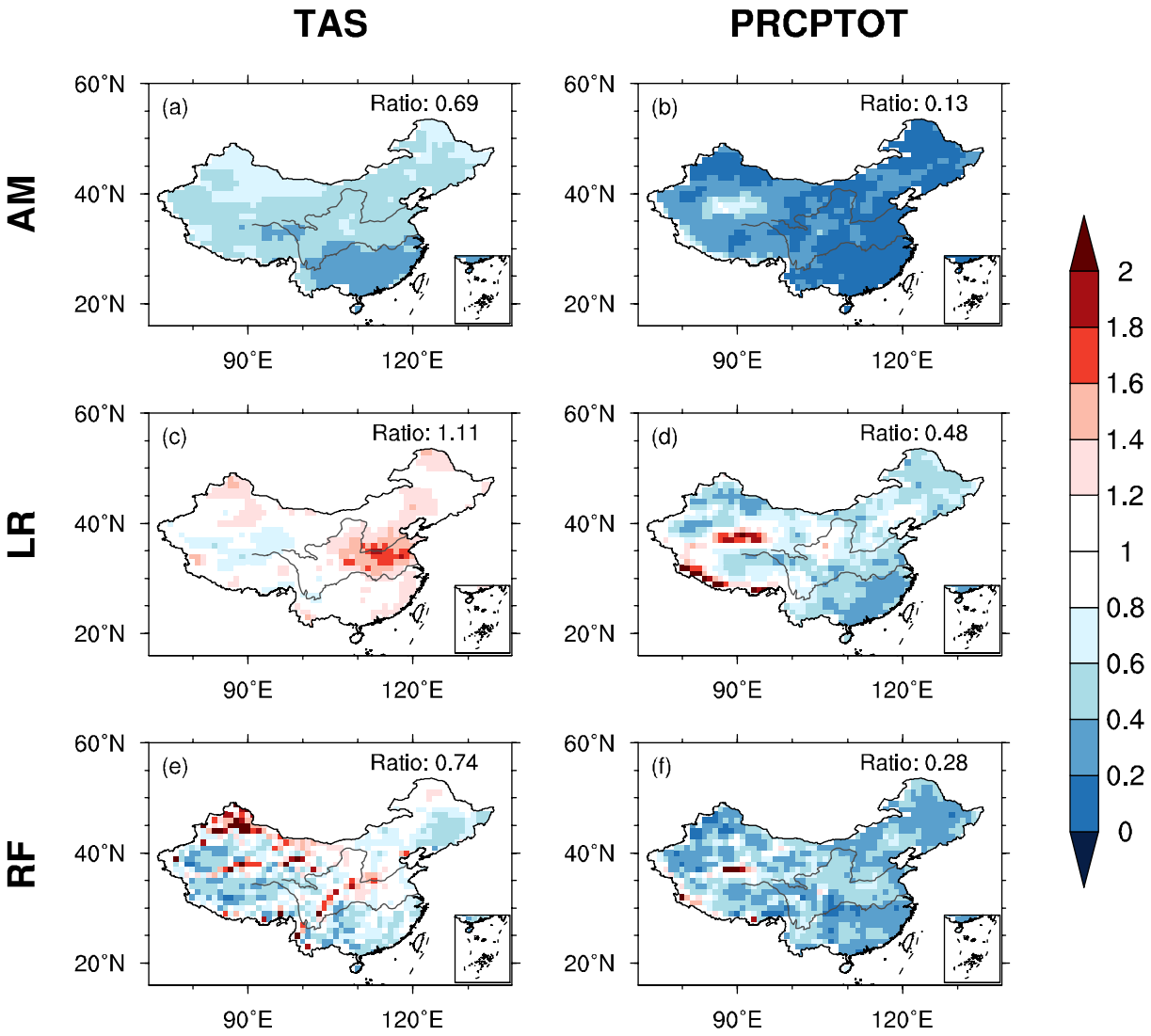

Figure S4. Spatial distributions of the ratio between simulated and observed temporal standard deviations in the validation period for AM (upper panels), LR (middle panels), and RF (lower panels). Left panels are for mean temperature TAS and right panels total precipitation PRCPTOT. The areal average in the domain is given on the top-right corner of each panel.

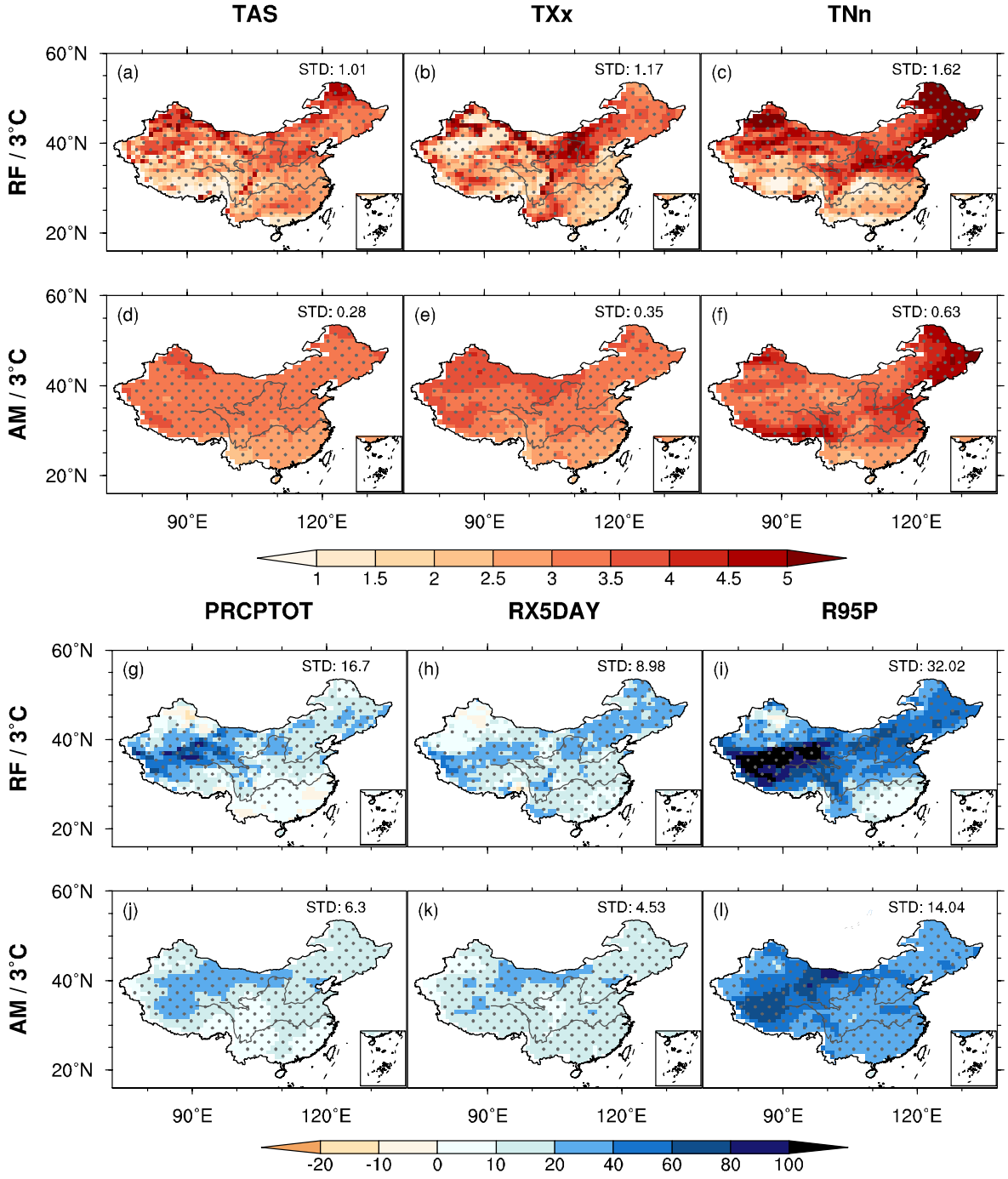

Figure S5. Spatial distribution of changes of 6 climate indices (STD is the spatial standard deviation) obtained from RF and AM under the 3°C warming target. Similar to the cases of 1.5°C and 2°C presented in the main text. Units are °C for the temperature indices, and % for the precipitation indices.

Reference

Dong, T., and W. Dong 2021 Evaluation of extreme precipitation over asia in CMIP6 models. *Climate Dyn.* **10** 1-19

Yang, X., B. Zhou, Y. Xu, and Z. Han 2021 CMIP6 evaluation and projection of temperature and precipitation over China. *Adv. Atmos. Sci.* **38** 817-830