Stochastic Emulators of Spatially Resolved Extreme Temperatures of Earth System Models

Mengze Wang[1], Andre N. Souza[2], Raffaele Ferrari[2], Themistoklis P. Sapsis[1]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Key Points:

- Stochastic emulators are developed to estimate the probability distribution of local daily maximum temperature under climate change.
- Coefficients of Empirical Orthogonal Functions are modelled as functions of the global mean temperature, superposed with Gaussian processes.
- Our approach can accurately emulate the quantile anomaly of daily maximum temperature in future scenarios.

Corresponding author: Themistoklis P. Sapsis, `sapsis@mit.edu`

Abstract

Prediction of extreme events under climate change is challenging but essential for risk management of natural disasters. Although earth system models (ESMs) are arguably our best tool to predict climate extremes, their high computational cost restricts the application to project only a few future scenarios. Emulators, or reduced-complexity models, serve as a complement to ESMs that achieve a fast prediction of the local response to various climate change scenarios. Here we propose a data-driven framework to emulate the full statistics of spatially resolved climate extremes. The variable of interest is the near-surface daily maximum temperature. The spatial patterns of temperature variations are assumed to be independent of time and extracted using Empirical Orthogonal Functions (EOFs). The time dependence is encoded through the coefficients of leading EOFs which are decomposed into long-term seasonal variations and daily fluctuations. The former are assumed to be functions of the global mean temperature, while the latter are modelled as Gaussian stochastic processes with temporal correlation conditioned on the season. The emulator is trained and tested using the simulation data in CMIP6. By generating multiple realizations, the emulator shows significant performance in predicting the temporal evolution of the probability distribution of local daily maximum temperature. Furthermore, the uncertainty of the emulated statistics is quantified to account for the internal variability. The emulation accuracy in testing scenarios remains consistent with the training datasets. The performance of the emulator suggests that the proposed framework can be generalized to other climate extremes and more complicated scenarios of climate change.

Plain Language Summary

Extreme events in the global climate system, such as heat waves and hurricanes, cause incalculable losses every year. Conventional climate models, called Earth System Models (ESMs), are our best tools to predict how climate change may affect the occurrence rate of extreme events in the future. However, these models are relatively slow and expensive to run. We present a framework to design emulators, or reduced-complexity models, to efficiently predict the complete statistics of climate extremes on spatially-resolved grids. Once trained using a few simulations generated from ESMs, the emulator can be used to predict climate change scenarios that were not included in the training data. Our approach is demonstrated for near-surface daily maximum temperature data. The mean, variance, and extreme values of the temperature generated by the emulator are very similar to the statistics generated by ESMs. Furthermore, the emulator provides a speedy quantification of the uncertainty of the predicted statistics. The performance of the emulator suggests that our framework can be generalized to other types of extreme events in the climate system.

1 Introduction

Unprecedented climate extremes, associated with anthropogenic global warming, have been observed worldwide, such as the Russian heatwaves in 2010 and the record-breaking Atlantic hurricane season in 2020 (Meehl & Tebaldi, 2004; Barriopedro et al., 2011; Reed et al., 2022). The annual losses from such weather- and climate-related disasters have surged dramatically, escalating from several billion dollars in 1980 to 200 billion in 2020 (Allen et al., 2012; AON, 2020), not to mention the incalculable loss of lives. Effectively managing the risks of extreme events and minimizing their associated damages necessitates accurate quantification of their likelihood in a rapidly changing global climate. Despite the increased frequency of extreme weather events, their probability at a given time and location is still very low, and thus quantifying their risks requires large ensembles of numerical simulations for very long time horizons. The need for ensembles of simulations amplifies the already high computational cost associated with running fullscale Earth System Models (ESMs) and restricts their application to a limited number of climate change scenarios. In contrast, emulators, or reduced-complexity models, provide a more efficient evaluation of the statistics of extreme events in response to more diverse scenarios. In the present work, we develop a multivariate Gaussian stochastic emulator that estimates the probability distribution of local daily maximum temperature on spatially-resolved grids.

Climate emulators can be broadly categorized by the spatial resolution of their projections. The first type of emulators, also known as simple climate models (SCMs), focus on modelling how global or regional mean fields are influenced by the concentrations of greenhouse gases, emissions of aerosols, and natural effective radiative forcing variations (Meinshausen et al., 2011; Seneviratne et al., 2016). A majority of these emulators have been systematically compared in the Reduced Complexity Model Intercomparison Project (Z. R. Nicholls et al., 2020; Z. Nicholls et al., 2021), by evaluating their prediction accuracy of the global mean temperature. Based on this type of emulators, interactive models have been developed for policymakers and stakeholders to actively examine the impact of energy, economic and public policies on climate change (Kapmeier et al., 2021; Rooney-Varga et al., 2021).

The second type of emulators specialize in predicting the response of local variables to climate change. The most widely used method for this type of emulator is pattern scaling, where the climate variables at different locations are assumed as independent linear functions of the global mean temperature (Mitchell, 2003). Therefore, the global mean temperature predicted by the first-type emulators can be used as an input for pattern scaling, facilitating localized climate predictions in response to a variety of emission scenarios. Over time, the framework of pattern scaling has evolved to encompass a broader range of techniques. These advances include the adoption of response functions to account for past trajectories of CO_2 (Castruccio et al., 2014; Freese et al., 2024), the use of Matern covariance functions for modeling spatial correlation (Alexeeff et al., 2018), and the incorporation of internal variability through autoregressive processes or the spectrum of principal components analysis (Beusch et al., 2020; Link et al., 2019). As modern machine learning methods emerge, researchers have explored diverse architectures to enhance the accuracy of local climate emulation, utilizing inputs ranging from globally-averaged emissions to spatial distribution of aerosols. Most of these machine learning models have been evaluated on the benchmark datasets, with ClimateBench (Watson-Parris et al., 2022) and ClimateSet (Kaltenborn et al., 2023) being the most commonly used ones. Compared with pattern scaling, neural networks can provide a more accurate emulation of certain variables, such as the global precipitation, when trained on sufficiently large ensembles of simulations (Lütjens et al., 2024) albeit with a compromise in the model complexity.

Both classes of emulators have been typically used to predict time-averaged quantities. Only a few recent studies have explored emulating the statistics of climate extremes, such as the annual maximum temperature and the duration of hot waves within a year (Tebaldi et al., 2020; Quilcaille et al., 2022). Furthermore, no prior work has been reported on the emulation of probability distribution of local climate variables, which constitutes the primary objective of our research. We introduce a stochastic model to emulate the statistics of climate extremes, utilizing temperature-related extreme events as a prototypical application. We first extract the empirical orthogonal functions (EOF) (Lorenz, 1956; Hannachi et al., 2007) of the spatial patterns of near-surface daily maximum temperature (TMX) fields to reduce the dimensionality of the system while maintaining a high spatial resolution. Driven by the observed nearly-Gaussian character of the EOF statistics (conditioned over season and year), we model the temporal evolution of the EOF coefficients as Gaussian stochastic processes (Mohamad & Sapsis, 2015; Arbabi & Sapsis, 2022), characterized by long-term trends, seasonal variations, and colored noise. The mean, variance and covariance of the EOF coefficients are parameterized using the

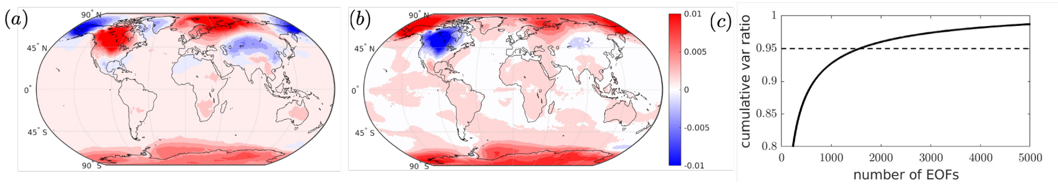

Figure 1. (a,b) The first and second spatial EOFs of daily maximum temperature, computed using CNRM-CM6-1-HR simulation data. (c) Cumulative variance ratio represented by leading EOFs.

global mean temperature and season, thus generalizing our emulator to more diverse climate change scenarios. A similar framework has been applied to emulate monthly-averaged temperature and humidity (Geogdzhayev et al., 2024). Our work will focus on daily maximum temperature and its full statistics.

The content of this paper is organized as follows. In §2 we introduce the simulation data used for training and testing the emulator. The mathematical framework of the emulator is described in §3, including the dimensionality reduction method in §3.1 and stochastic modeling of time series in §3.2. The emulation results are presented in §4, followed by a summary of the main conclusions and discussion in 5.

2 Data

Among all the ESMs in Coupled Model Intercomparison Project Phase 6 (CMIP6), we adopted the CNRM-CM6-1-HR and MPI-ESM1-2-LR model outputs as our reference dataset. Both models achieved reasonable skill scores on simulating the statistics of climate extremes according to a recent evaluation of the performance of CMIP6 models (Wehner et al., 2020). The CNRM-CM6-1-HR model provides the highest spatial resolution (nominal resolution 50km) among CMIP6 models, which best fits our needs to develop a spatially-resolved emulator. However, this model only has one realization available, which is insufficient to assess the influence of climate internal variability on the emulator. The MPI-ESM1-2-LR data feature a large ensemble of realizations, although the spatial resolution (250km) is problematic for studying climate extremes. Therefore, the majority of our results will focus on emulation of CNRM-CM6-1-HR data, while the large ensemble data of MPI-ESM1-2-LR will be utilized to investigate the impact of internal variability and ensemble size on the performance of the emulator.

Two variables are collected from the CNRM-CM6-1-HR and MPI-ESM1-2-LR model outputs: (i) Near-surface daily mean temperature (the *tas* variable in CMIP6), used to compute the global mean temperature; (ii) Near-surface daily maximum temperature (the *tasmax* variable in CMIP6). Here "near surface" refers to two-meter height according to the CMIP6 convention. The CMIP6 simulations cover a historical period from 1850 to 2014, followed by a set of future scenarios until 2100. The CNRM-CM6-1-HR model offers only one realization for both the historical period and each future scenario, whereas the MPI-ESM1-2-LR model provides 50 realizations. To train the emulator, we utilize the simulation data within the historical period and the SSP5-8.5 future scenario for each ESM. The SSP1-2.6 future scenario is utilized for testing purposes.

3 Methods

3.1 Data pre-processing: dimensionality reduction

Since we focus on the near-surface temperature, the spatial location $\boldsymbol{x}$ is described by the latitude and longitude coordinates, $\boldsymbol{x} = (\theta, \varphi)$, where $\theta \in [-\pi/2, \pi/2]$ and $\varphi \in [0, 2\pi)$. The time step size is one day, and the number of days since 01/01/1850 0:00 is represented as t. The daily maximum temperature (TMX) at location $\boldsymbol{x}$ and time t for the ensemble member ω is denoted as $q(\boldsymbol{x}, t, \omega)$. The climatological mean $\bar{q}(\boldsymbol{x}, t)$ is extracted by phase-averaging TMX for the same calendar day and location across the historical period, 1850-2014, and over the entire ensemble. In other words, at an arbitrary time t, $\bar{q}(\boldsymbol{x}, t) = \bar{q}(\boldsymbol{x}, \mathrm{mod}(t, 365))$. The fluctuations of TMX are decomposed as superposition of Empirical Orthogonal Functions (EOFs), $\phi_i(\boldsymbol{x})$,

$$q'(\boldsymbol{x}, t, \omega) \coloneqq q(\boldsymbol{x}, t, \omega) - \bar{q}(\boldsymbol{x}, t) = \sum_i a_i(t, \omega)\phi_i(\boldsymbol{x}). \tag{1}$$

In order to compute the EOFs, we construct the spatial covariance function $\mathcal{R}(\boldsymbol{x}, \boldsymbol{x}^*)$ that quantifies the covariance between fluctuating TMX at two arbitrary locations $\boldsymbol{x}$ and $\boldsymbol{x}^*$,

$$\mathcal{R}(\boldsymbol{x}, \boldsymbol{x}^*) = \langle q'(\boldsymbol{x}, t, \omega) q'(\boldsymbol{x}^*, t, \omega)\rangle_{t\omega}\,. \tag{2}$$

The notation $\langle\cdot\rangle_{t\omega}$ represents averaging over time and the ensemble. The EOFs are defined as the eigenfunctions of $\mathcal{R}(\boldsymbol{x}, \boldsymbol{x}^*)$, taking into account the curvature of the Earth's surface S,

$$\int_S \mathcal{R}(\boldsymbol{x}, \boldsymbol{x}^*)\phi_i(\boldsymbol{x^*})\cos\theta^* d\theta^* d\varphi^* = \lambda_i \phi_i(\boldsymbol{x}). \tag{3}$$

The coefficient of each EOF at time t is obtained by projecting $q'(\boldsymbol{x}, t, \omega)$ onto $\phi_i(\boldsymbol{x})$,

$$a_i(t, \omega) = \int_S q'(\boldsymbol{x}, t, \omega)\phi_i(\boldsymbol{x})\cos\theta d\theta d\varphi. \tag{4}$$

Similar to the climatological mean, the EOFs are also computed from the historical data. However, we only utilize the TMX snapshots on every five days, rather than daily data, because TMX on adjacent days are highly correlated. Our choice of five-day interval is based on the observation that on this timescale the autocorrelation coefficient of TMX at most locations decreases to approximately 0.5 (Kalvová & Nemešová, 1998), striking a reasonable balance between data independence and comprehensive representation of temperature variability.

For CNRM-CM6-1-HR data, since only one realization is available, the number of snapshots (1.2×10^4) is much smaller than the number of grids (2.6×10^5). As such, it is unnecessary to store the large covariance matrix (2), and the method of snapshots is adopted to solve the eigenvalue problem (3) more efficiently. Specifically, we compute the temporal covariance matrix of q', whose size is the square of the number of snapshots. The eigen-decomposition of the temporal covariance matrix is then performed to get its eigenvalues and eigenfunctions, which can be linearly transformed to get the eigenpairs $(\lambda_i, \phi_i(\boldsymbol{x}))$ of the spatial covariance $\mathcal{R}(\boldsymbol{x}, \boldsymbol{x}^*)$. More details can be found in Sirovich (1987) and Taira et al. (2020). For MPI-ESM1-2-LR data, the number of grids (1.8×10^4) is comparable or smaller than the total number of snapshots ($1.2 \times 10^4\times$ the number of realizations adopted), and we directly solve equation (3) to obtain the eigenfunctions of the spatial covariance.

The first two EOFs of the CNRM-CM6-1-HR data are visualized in figure 1(*a*,*b*). They account for 2.9% and 2.7% of the total variance, respectively. Both EOFs are reminiscent of the Arctic Oscillation/Northern Hemisphere Annular Mode (Thompson & Wallace, 1998) and the Southern Hemisphere Annular Mode (Fogt & Marshall, 2020). Unlike previous studies that focused on the first few EOFs to extract the physically significant modes (Wallace & Gutzler, 1981; Amaya, 2019), our objective is to reconstruct the

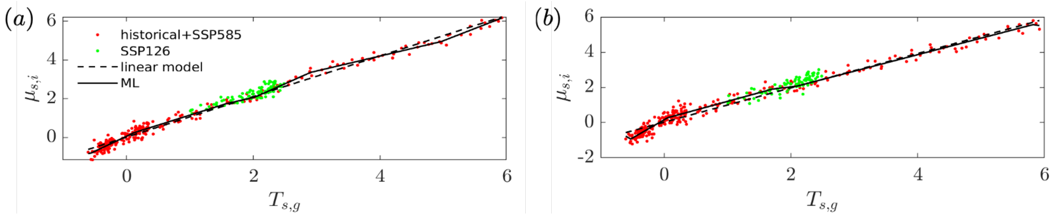

Figure 2. Jun-Aug mean of (a) the first and (b) second EOF coefficients in each year of CNRM-CM6-1-HR dataset, from 1850 to 2100, plotted versus the global mean temperature. Red dots: true seasonal mean obtained from the historical and SSP5-8.5 scenario. Green dots: SSP1-2.6 scenario. Black dashed line: linear regression; Solid line: machine-learned function.

full probability distribution of local TMX with sufficient accuracy and efficiency. Therefore, we retain the first 2,000 EOFs for the CNRM-CM6-1-HR model, which altogether represent approximately 95% of the total variance (figure 1c) of the respective datasets.

3.2 Multivariate Gaussian stochastic emulator of EOF time series

Assuming the climatological mean and EOFs remain invariant with respect to time and future scenarios, our stochastic emulator of the daily maximum temperature is formulated as,

$$\hat{q}(\boldsymbol{x}, t, \hat{\omega}) = \bar{q}(\boldsymbol{x}, t) + \sum_{i=1}^{I} \hat{a}_i(t, \hat{\omega})\phi_i(\boldsymbol{x}). \tag{5}$$

A notable difference between equation (5) and the decomposition of true TMX fluctuations (1) is the EOF coefficient, where a is the true coefficient obtained from projection (4) and $\hat{a}$ is estimated from the emulator. The emulation index $\hat{\omega}$ is also different from the ensemble member ω, since the emulator can be used to generate more realizations than the training data.

The time series of $\hat{a}$ in season s and for a given global mean temperature, is modelled as superposition of long-term trends and Gaussian-distributed daily fluctuations that encode temporal correlation:

$$\hat{a}_{s,i}(t, \hat{\omega}) = \hat{\mu}_{s,i}\left(T_{s,g}\right) + \hat{\sigma}_{s,i}\left(T_{s,g}\right) \sum_{j=1}^{I} \hat{l}_{s,ij}\hat{\eta}_{s,j}(t, \hat{\omega}), \quad i = 1, 2, \ldots, I. \tag{6}$$

The subscript $s = 1, 2, 3, 4$ corresponds to Northern Hemisphere spring (Mar-May), summer (Jun-Aug), autumn (Sep-Nov), and winter (Dec-Feb) respectively. The seasonal mean $\hat{\mu}_{s,i}$ and variance $\hat{\sigma}^2_{s,i}$ are parameterized as a function of the seasonally-averaged global mean temperature, $T_{s,g}$. The correlation between the ith and jth EOFs in season s is assumed constant and accounted for by $\hat{l}_{s,ij}$. The daily fluctuations of the EOF coefficients are modelled as superposition of Gaussian autoregressive processes $\hat{\eta}_{s,j}(t, \hat{\omega})$. Here $\hat{\eta}_{s,j}$ and $\hat{\eta}_{s,k}$ are uncorrelated when $j \neq k$, and the time series of $\hat{\eta}_{s,j}$ are emulated usiing the autocorrelation computed from training data. Specifically, consider a time window in season s of the y-th year, denoted as $t \in [t_{ys}, t_{ys}+N_s]$. The starting time, t_{ys}, corresponds to the first day of each season: Mar 1st, Jun 1st, Sep 1st, and Dec 1st, for $s = \{1, 2, 3, 4\}$. The duration of each time window, N_s, is given by $N_s = \{92, 92, 91, 90\}$ days respectively. Within $t \in [t_{ys}, t_{ys} + N_s]$, the emulated daily fluctuations $\hat{\eta}_{s,j}(t, \hat{\omega})$ satisfy

$$\hat{\eta}_{s,j}(t, \hat{\omega}) = \sum_{n=1}^{t-t_{ys}} c_{s,j}(n)\hat{\eta}_{s,j}(t-n, \hat{\omega}) + g_{s,j}(n)\epsilon_{s,j}(n), \quad \epsilon_{s,j}(n) \sim \mathcal{N}(0, 1), \quad t \in [t_{ys}, t_{ys}+N_s]. \tag{7}$$

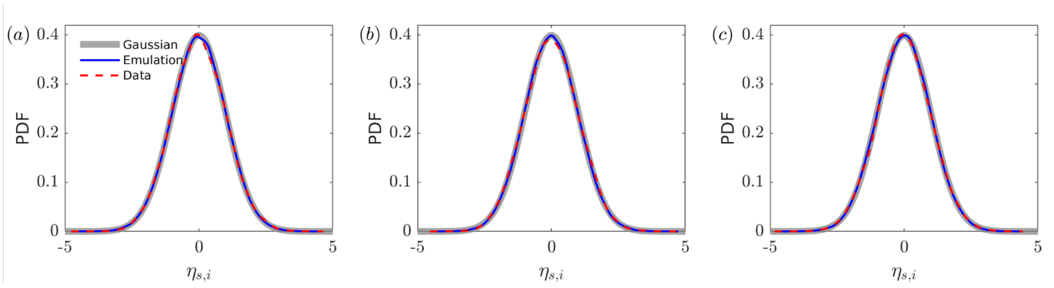

Figure 3. Probability density function (PDF) of the 1st, 2nd, and 500th component of the Jun-Aug $\boldsymbol{\eta}_s$: (*a*) $\eta_{2,1}$,(*b*) $\eta_{2,2}$, (*c*)$\eta_{2,500}$. Red dashed lines: PDF computed using CNRM-CM6-1-HR historical and SSP5-8.5 future scenario data, from 1850 to 2100; gray lines: Gaussian fit of $\eta_{2,i}$ data; blue lines: PDF of 10 emulations of 1850-2100 $\hat{\eta}_{2,i}$

The parameters $c_{s,j}(n)$ and $g_{s,j}(n)$ are independent of the year and will be estimated from the training data, while the standard normal random number $\epsilon_{s,j}(n)$ varies with the year and the emulation. The emulator (6) can also be written more compactly in vector form,

$$\hat{\mathbf{a}}_s(t,\hat{\omega}) = \hat{\boldsymbol{\mu}}_s\left(T_{s,g}\right) + \hat{\mathbf{D}}_s\left(T_{s,g}\right)\hat{\mathbf{L}}_s\hat{\boldsymbol{\eta}}_s(t,\hat{\omega}), \tag{8}$$

where $\hat{\mathbf{a}}_s$, $\hat{\boldsymbol{\mu}}_s$, and $\hat{\boldsymbol{\eta}}_s$ are $I\times 1$ column vectors. The notation $\hat{\mathbf{D}}_s$ is a diagonal matrix, and each element on the diagonal is $\hat{\sigma}_{s,i}$. The matrix $\hat{\mathbf{L}}_s$ is lower triangular, where each entry corresponds to $\hat{l}_{s,ij}$.

It is important to emphasize here that the formulated emulator is conditionally Gaussian, i.e. for a fixed season and global mean temperature, the daily fluctuations are, by design, normally distributed. While this does not necessarily imply that long term statistics will have a Gaussian character, since we also have the variation of the global mean temperature, it does not allow for the possibility of daily temperature extremes that have (for a given season and global mean temperature) a non-Gaussian distribution, e.g. follow heavy tails. For the present context, direct comparisons suggest that this is a acceptable assumption. However, for other variables this aspect may introduce limitations. We plan to extend the framework to address these potential limitations in future work.

The unknown parameters (which are functions of $T_{s,g}$) in the emulator (6,7) are estimated using the true EOF coefficients $a_i(t,\omega)$ (4) and the global mean temperature $T_{s,g}$ from 1850 to 2100 (historical and SSP5-8.5 scenario). Given $a_i(t,\omega)$ data, we first compute the actual seasonal mean $\mu_{s,i}$ and standard deviation $\sigma_{s,i}$ in each year, averaged over the entire ensemble. Two examples of the Jun-Aug mean $\mu_{s,i}$ versus the corresponding $T_{s,g}$ are shown in figure 2 (red dots). These relationships are mostly linear and independent of the future scenario (SSP1-2.6 shown in green dots), which motivate us to regress $\hat{\mu}_{s,i}$ as a linear function of $T_{s,g}$ (black dashed lines). Similar linear relationships are also observed for the variance $\sigma^2_{s,i}$ and also for higher-ranked EOFs. Nonlinear functions are also attempted using fully-connected neural networks. For each $\hat{\mu}_{s,i}$ or $\hat{\sigma}^2_{s,i}$, the neural network is designed with two hidden layers, each containing three neurons, utilizing the ReLU activation function. The learned nonlinear functions are shown as black solid lines in figure 2, which provide slightly better agreement with the training data. A more systematic comparison of the emulation results using linear and nonlinear functions will be provided in §4.1. We also explored alternative network architectures with varying numbers of layers and neurons, as well as different activation functions, including Sigmoid and Tanh. However, these modifications did not yield significant improvements and the associated results are not shown.

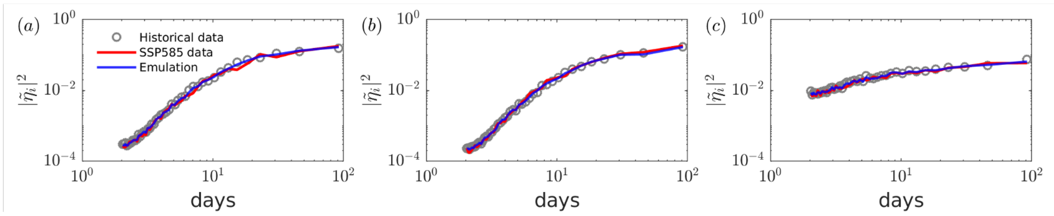

Figure 4. Spectra of the 1st, 2nd, and 500th component of the Jun-Aug $\boldsymbol{\eta}_s$: (a) $\eta_{2,1}$,(b) $\eta_{2,2}$, $(c)\eta_{2,500}$. Gray circles: spectra averaged using CNRM-CM6-1-HR historical (1850-2014) data; red lines: CNRM-CM6-1-HR SSP5-8.5 (2015-2100) data; blue lines: 10 emulations of 2015-2100 spectra.

After extracting the variation of the seasonal mean and standard deviation in response to the global mean temperature, $\hat{\mu}_{s,i}(T_{s,g})$ or $\hat{\sigma}^2_{s,i}(T_{s,g})$, we remove these trends from the true EOF coefficients, resulting in the residuals $(a_{s,i} - \hat{\mu}_{s,i})/\hat{\sigma}_{s,i}$. We then evaluate their cross-correlations,

$$\hat{\mathbf{\Sigma}}_s = \left\langle \hat{\mathbf{D}}_s^{-1} \left(\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s\right) \left(\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s\right)^\top \hat{\mathbf{D}}_s^{-\top} \right\rangle_{t\omega}, \quad \hat{\mathbf{\Sigma}}_s = \hat{\mathbf{L}}_s \hat{\mathbf{L}}_s^\top, \tag{9}$$

The time average is performed from 1850 to 2100 for each season respectively. While the actual cross correlations fluctuate over time, they remain statistically stationary for most EOFs, justifying the choice of a constant matrix model. Generalization of (9) to time-dependent correlations requires large-ensemble data and will be discussed in §4.2. The last equality in (9) is a Cholesky decomposition of $\hat{\mathbf{\Sigma}}_s$. Multiplying the residuals by $\hat{\mathbf{L}}_s^{-1}$ produces uncorrelated time series,

$$\boldsymbol{\eta}_s(t,\omega) = \hat{\mathbf{L}}_s^{-1} \hat{\mathbf{D}}_s^{-1} \left(\mathbf{a}_s(t,\omega) - \hat{\boldsymbol{\mu}}_s\right), \tag{10}$$

which satisfies

$$\left\langle \boldsymbol{\eta}_s(t,\omega) \boldsymbol{\eta}_s(t,\omega)^\top \right\rangle_{t\omega} = \mathbf{I}. \tag{11}$$

Here $\mathbf{I}$ is an identity matrix with a size equal to the number of adopted EOFs . In other words, each entry of $\boldsymbol{\eta}_s(t,\omega)$ has unit variance, and different entries are uncorrelated.

To justify our assumption that $\eta_{s,j}(t,\omega)$ in season s can be modelled as Gaussian processes (equation 7) with the same autocorrelations across different years, we evaluate the statistics $\eta_{s,j}(t,\omega)$ in figure 3,4. The probability density functions of the 1st, 2nd, and 500th component of Jun-Aug $\eta_{s,j}$ are computed using historical and SSP5-8.5 scenario data, from 1850 to 2100. The profiles are plotted by red dashed lines in figure 3, which almost overlap with the fitted Gaussian distributions (gray lines). While not shown here, the other components of $\eta_{s,j}(t,\omega)$ also exhibit approximately Gaussian distributions. To examine the time dependence of the second-order statistics of each component of $\boldsymbol{\eta}_s$, we compute the Fourier spectra of $\boldsymbol{\eta}_s$ in Jun-Aug of each year and average them over two distinct time windows, 1850-2014 and 2015-2100 of SSP5-8.5 scenario. As visualized in figure 4, the spectra of three components of $\boldsymbol{\eta}_s$ remain approximately unchanged over time. Therefore, the statistics averaged over the entire 1850-2100 period are used to generate the the surrogate Gaussian processes $\hat{\eta}_{s,j}$ that represent stochastic realizations of daily fluctuations. Simulation of the Gaussian processes is based on the exact time-domain method which utilizes the autocorrelation of $\boldsymbol{\eta}_s$. This approach has been demonstrated more robust against uncertainty of statistics than the frequency-domain method (Percival, 1993). The PDFs of the simulated $\hat{\boldsymbol{\eta}}_s$ in figure 3(blue lines) indeed follow Gaussian distribution, and the Fourier spectra of the simulated processes align with the true spectra, as illustrated in figure 4.

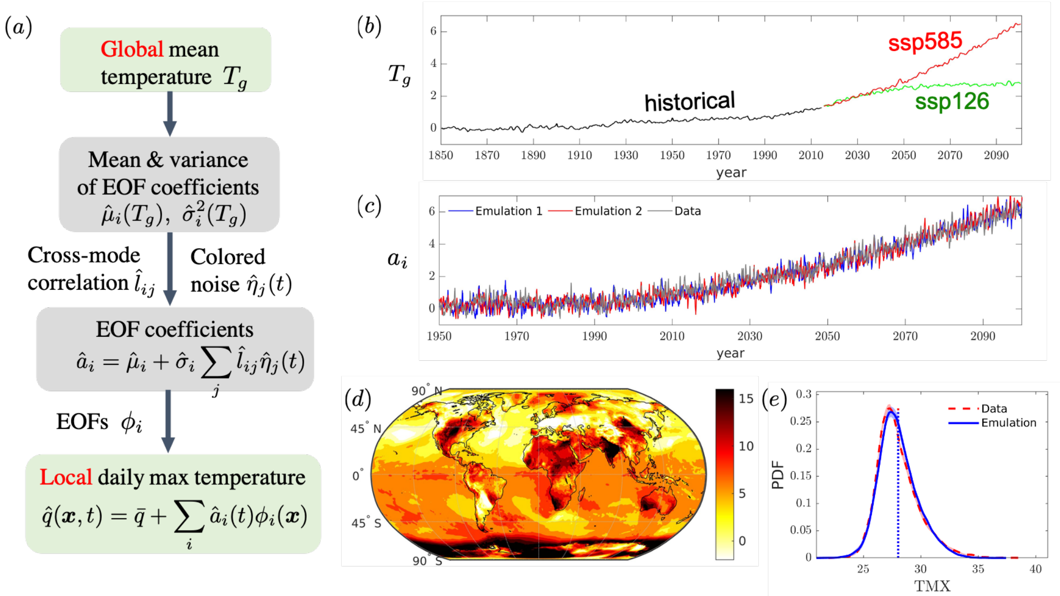

Figure 5. (*a*) Flow chart showing the structure of the emulator. Given the global mean temperature T_g, the emulator predicts the local daily maximum temperature on spatially-resolved grids. (*b*) One-year moving average of the global mean temperature, shown for different scenarios. (*c*) Example time series of the true and emulated EOF coefficients. (*d*) Sample outputs from the emulator: reconstruction of the TMX field. (*e*) An example of the probability density function of local TMX, averaged in Jun-Aug over a ten-year window. The vertical lines mark the mean values.

The steps of the emulation are summarized schematically in figure 5*a*. Starting from the temporal evolution of the global mean temperature (panel *b*), the seasonal mean and variance of the EOF coefficients are estimated from the learned relationships $\hat{\mu}_{s,i}(T_{s,g})$, $\hat{\sigma}^2_{s,i}(T_{s,g})$. The daily fluctuations are constructed as the stochastic autoregressive processes $\hat{\eta}_{s,j}(t,\omega)$, which are scaled by $\hat{l}_{s,ij}$ and superposed to account for the cross correlation between different EOFs. Combining the scaled daily fluctuations with long-term trends, we obtain the emulated time series of the EOF coefficients, exhibiting the same first and second order statistics as the true time series (panel *c*). Given the time series and shape of EOFs, the final output of the emulator is the temporal evolution of gridded local TMX. A sample snapshot of TMX is visualized in panel *d*. To acquire converged probability distribution of local TMX, especially for the tails that represent extreme events, the statistics are computed by averaging over a decadal window in time and a $1^\circ \times 1^\circ$ region in space. Panel *e* shows a sample comparison between the emulated and true probability density function (PDF). The blue region marks the uncertainty of the distribution, estimated by performing multiple emulations. We note the non-Gaussian character of the target and approximated PDF, which is the result of considering the statistics over a time window that the global average temperature changes.

4 Emulation results

4.1 Emulation of CNRM-CM6-1-HR dataset

The performance of the emulator is firstly evaluated in detail for Jun-Aug, when TMX is the most extreme in Northern Hemisphere. Results in other seasons will be briefly discussed at the end of this section. To differentiate between the emulator that adopts

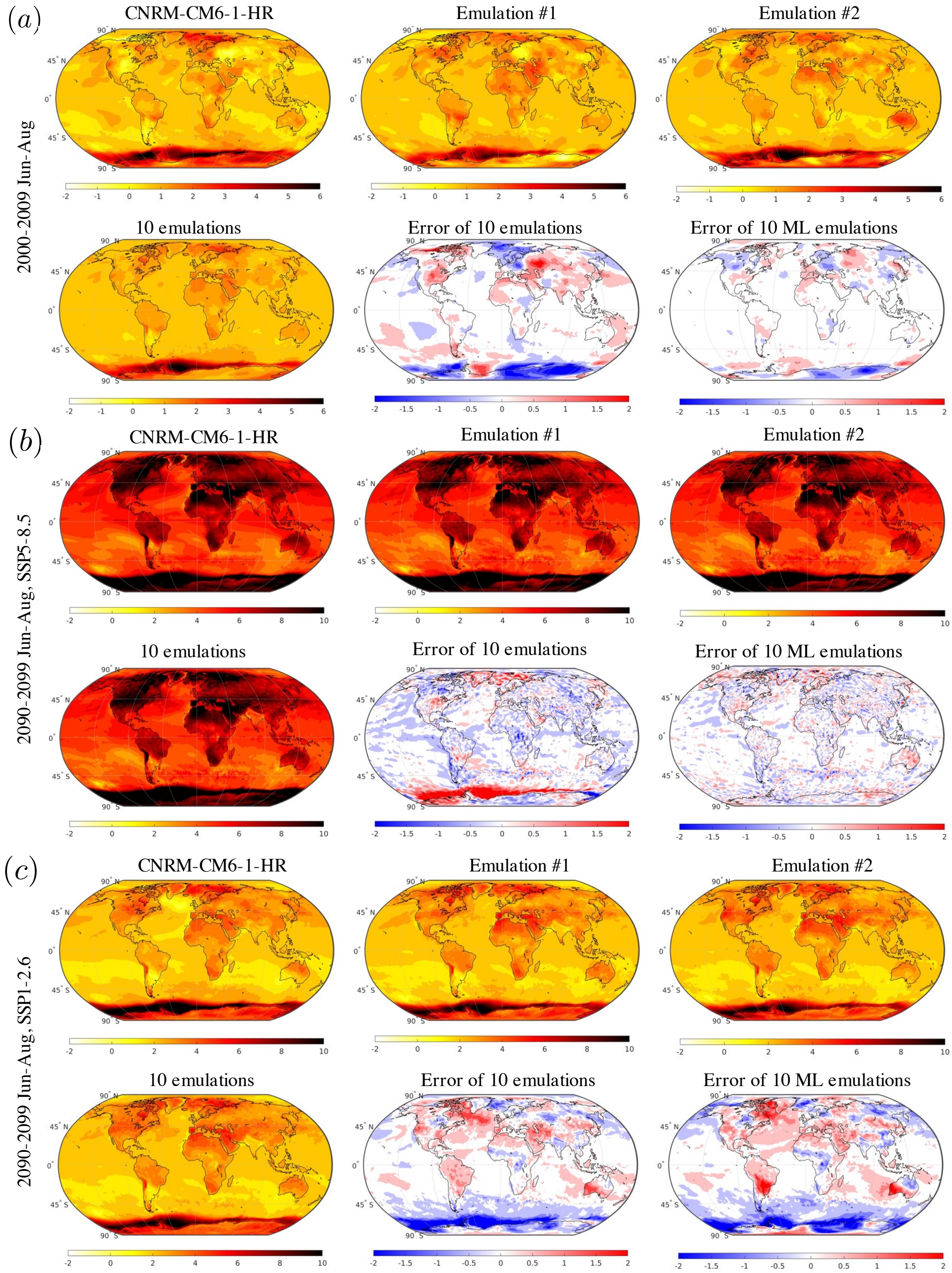

Figure 6. Mean anomaly of Jun-Aug daily maximum temperature, averaged over (*a*) 2000-2009, (*b*) 2090-2099 of the SSP5-8.5 scenario, and (*c*) 2090-2099 of the SSP1-2.6 scenario. Each subfigure shows the true mean from CNRM-CM6-1-HR ESM, two sample emulations, average of 10 emulations, error of 10 emulations, and the error of 10 ML emulations. Reference: 1850-1900 Jun-Aug mean TMX.

linear and nonlinear model for the long-term trends, the former is referred to as "emulation" and the latter is denoted as "machine-learning (ML) emulation". Figure 6 shows the mean of local TMX across three decadal periods: 2000-2009 within the historical period, 2090-2099 of the SSP5-8.5 scenario, and 2090-2099 of the SSP1-2.6 scenario. The reference mean, computed from the CNRM-CM6-1-HR data, is compared against two sample emulations, the average of ten emulations, and ML emulations. The emulator accurately captures the evolution of local TMX under both high and low warming scenarios. Significant anomalies in regions such as the Arctic, western coast of South America, North Africa, West Asia and Southern Ocean are well reproduced. Errors are within 1°C at most locations, with the highest errors reaching 2°C. Using ML model for the seasonal mean and variance appreciably improves the emulation accuracy. Despite training on historical and SSP5-8.5 data only, the emulator performance on the unseen SSP1-2.6 scenario demonstrates its potential for application across various climate change pathways.

The errors of the emulated mean in figure 6(*a*-*c*) arise from different contributions. In figure 6*a*, the discrepancy between the emulations and the true mean mainly originates from the modeling assumption that the seasonal mean is fully determined by the global mean temperature, $\hat{\mu}_{s,i}(T_{s,g})$. As discussed in §3.2 (c.f. figure 2), a single global mean temperature $T_{s,g}$ can correspond to multiple values of the mean EOF coefficients $\mu_{s,i}$, due to the internal variability of the climate system and the neglected influence of the past global mean temperature or emission history. The internal variability of the CNRM-CM6-1-HR simulation is difficult to quantify, since only one realization is available. However, the variability captured by the emulator can be readily assessed by performing multiple emulations. Comparing the pattern of errors with the two emulations in figure 6*a*, we observe that most high-error regions also exhibit high variability, such as Europe and the Southern Ocean. In addition, the error magnitude aligns with the variability, indicating that the error can be further reduced if more realizations of the ESM are available for training the emulator and computing the local statistics. In figure 6*b*, smaller-scale fluctuation of the errors become more apparent, which stems from the changing shape of the leading EOFs under different warming conditions. Recall that the EOFs were computed only using the historical data. The leading historical EOFs adopted in the emulator may represent a lower variance in the SSP5-8.5 scenario, which results in higher emulation errors contributed by truncating EOFs. This issue can be mitigated by including SSP5-8.5 data into the calculation of EOFs, though similar errors might recur when the emulator is applied to unseen scenarios. The error in SSP1-2.6 scenario (figure 6*c*) is slightly higher than SSP5-8.5, due to the trained model of long-term trends not being optimal for SSP1-2.6. The error of ML emulations are even higher than linear emulations for SSP1-2.6, such as in South America, which indicates that the superior performance of ML emulator in SSP5-8.5 is likely due to overfitting. Nevertheless, the sensitivity of the seasonal mean to warming condition is modest, and the emulation error remains the same order of magnitude across different scenarios.

The standard deviation of local TMX is presented in figure 7. In historical periods, such as 2000-2009 shown in figure 7*a*, the standard deviation is reconstructed accurately for most locations. The error from ten emulations is almost identical to the ML emulations, suggesting a predominantly linear relationship between the variance of most EOF coefficients and the global mean temperature, $\hat{\sigma}_{s,i}(T_{s,g})$. From 2000-2009 to 2090-2099 in SSP5-8.5 scenario (panel *b*), the standard deviation slightly increases in most regions, such as North America, North Africa and West Asia. In contrast, the standard deviation in Greenland and Southern Ocean shows a significant reduction, likely due to diminished ice coverage (Räisänen, 2002; Gao et al., 2015). These trends are consistent with the observational data (Huntingford et al., 2013) and ESM simulations using other models (Olonscheck & Notz, 2017). The performance of the emulator is the least satisfactory in regions associated with the most significant trends. For example, the enhanced variance in North Africa is not captured, and the decreasing trend in the Southern Ocean

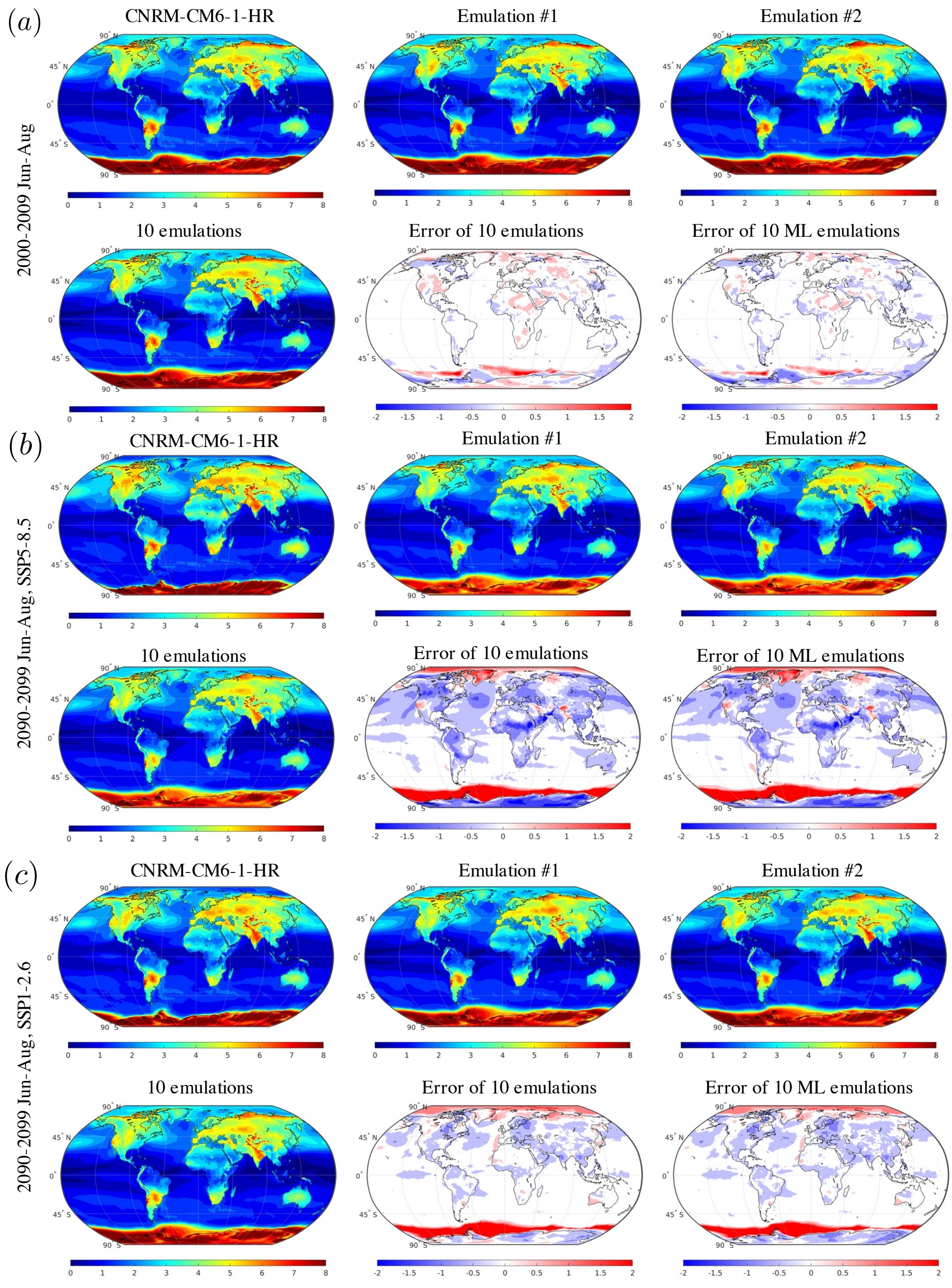

Figure 7. Standard deviation of Jun-Aug daily maximum temperature, averaged over (*a*) 2000-2009, (*b*) 2090-2099 of the SSP5-8.5 scenario, and (*c*) 2090-2099 of the SSP1-2.6 scenario. Each subfigure shows the true mean from CNRM-CM6-1-HR ESM, two sample emulations, average of 10 emulations, error of 10 emulations, and the error of 10 ML emulations.

is only partially reproduced. These limitations can be alleviated by relaxing the assumption of the emulator that cross-EOF correlations $\hat{\mathbf{L}}_s$ are constant, which is explored in §4.2. Nonetheless, the underlying climate dynamics, such as the removal of polar amplification due to the loss of ice coverage, is non-linear and non-local, requiring more judicious treatment in the construction of emulators. In the SSP1-2.6 scenario (figure 7*c*), changes of standard deviation progress more slowly, and the corresponding emulation errors are less severe than in the SSP5-8.5 scenario.

We visualize in figure 8 the 97.5% quantile as an example of extreme temperature. It is important to note that the baseline temperature for anomalies in figure 8 differs from that in figure 6; here, it is based on the 1850-1900 97.5% quantile rather than the 1850-1900 average. Within 2000-2009, the emulated quantile (figure 8*a*) is less accurate than the mean (c.f. figure 6*a*), which is anticipated due to the compounded error from the emulated standard deviation affecting the quantile estimation. Moreover, the predicted quantile exhibits greater uncertainty across different emulations, further contaminating the accuracy of averaged emulations. In SSP5-8.5 2090-2099 (figure 8*b*), the increase of quantile is similar to the mean (figure 6*b*) at most locations. An interesting trend can be observed in South Asia: the quantile grows more significantly than the mean in India but slightly decreases in Ganges Delta. Since the standard deviation in South Asia remains approximately unaffected by the global warming, the change of extreme temperature predominantly indicates heavier or thinner tails of the probability distribution. These trends are successfully identified by the emulator. The highest error of the emulated quantile occurs in Greenland and the Southern Ocean, due to the overestimated standard deviation as discussed in figure 7. Other error patterns primarily originate from the internal variability, as explored by analyzing the temporal evolution of the emulated quantile from 2010 to SSP5-8.5 2100 (Appendix Appendix A). When applied to the testing data under the SSP1-2.6 scenario (figure 8*c*), the emulator effectively captures the warming patterns of extreme temperatures with accuracy comparable to the training data in figures 8(*a*,*b*). Using the ML model for long-term trends does not improve the quantiles of TMX in SSP1-2.6 scenario.

The probability density functions of local TMX are plotted in figure 9 at three $1^\circ\times 1^\circ$ small regions that include major cities: Boston, situated in proximity to the Atlantic Ocean; Tehran, featured by the semi-arid climate with hot dry summers; Shanghai, characterized by the subtropical maritime monsoon climate. All these locations exhibit a significant increase of the extreme temperature in SSP5-8.5 scenario (c.f. figure 8). Overall the emulated PDFs closely match their true profiles, although the deviations in the SSP5-8.5 scenario are more appreciable. Since the size of samples (3,680) to estimate the true PDF might be insufficient, we quantify the uncertainty by bootstrap resampling, as marked by red shaded regions in figure 9. The uncertainty of emulated PDFs are quantified using one standard deviation of ten emulations, as shown by blue shaded areas. Taking the uncertainty of PDFs into consideration, the mismatch between emulated and true profiles are less severe. Note that the non-Gaussian shape of the PDF at Tehran (middle row in figure 9) is accurately replicated by the emulator, due to the effect of mixing instantaneous Gaussian TMX with different mean and variance, as discussed at the end of §3.2. The accurate emulation of the PDFs demonstrate the capacity of the emulator to predict any statistics of theoretical and practical interest, including skewness, kurtosis, and climate extreme indices.

The performance of the emulator in different seasons is examined by the root-mean-square error (RMSE) of the statistics and summarized in figure 10. Given a statistic of the reference daily maximum temperature $\mathcal{Q}$ and its estimation $\hat{\mathcal{Q}}$, the RMSE is defined as,

$$\mathrm{RMSE} = \left(\frac{1}{S}\int_S \left(\hat{\mathcal{Q}} - \mathcal{Q}\right)^2 \cos\theta d\theta d\varphi\right)^{1/2}. \tag{12}$$

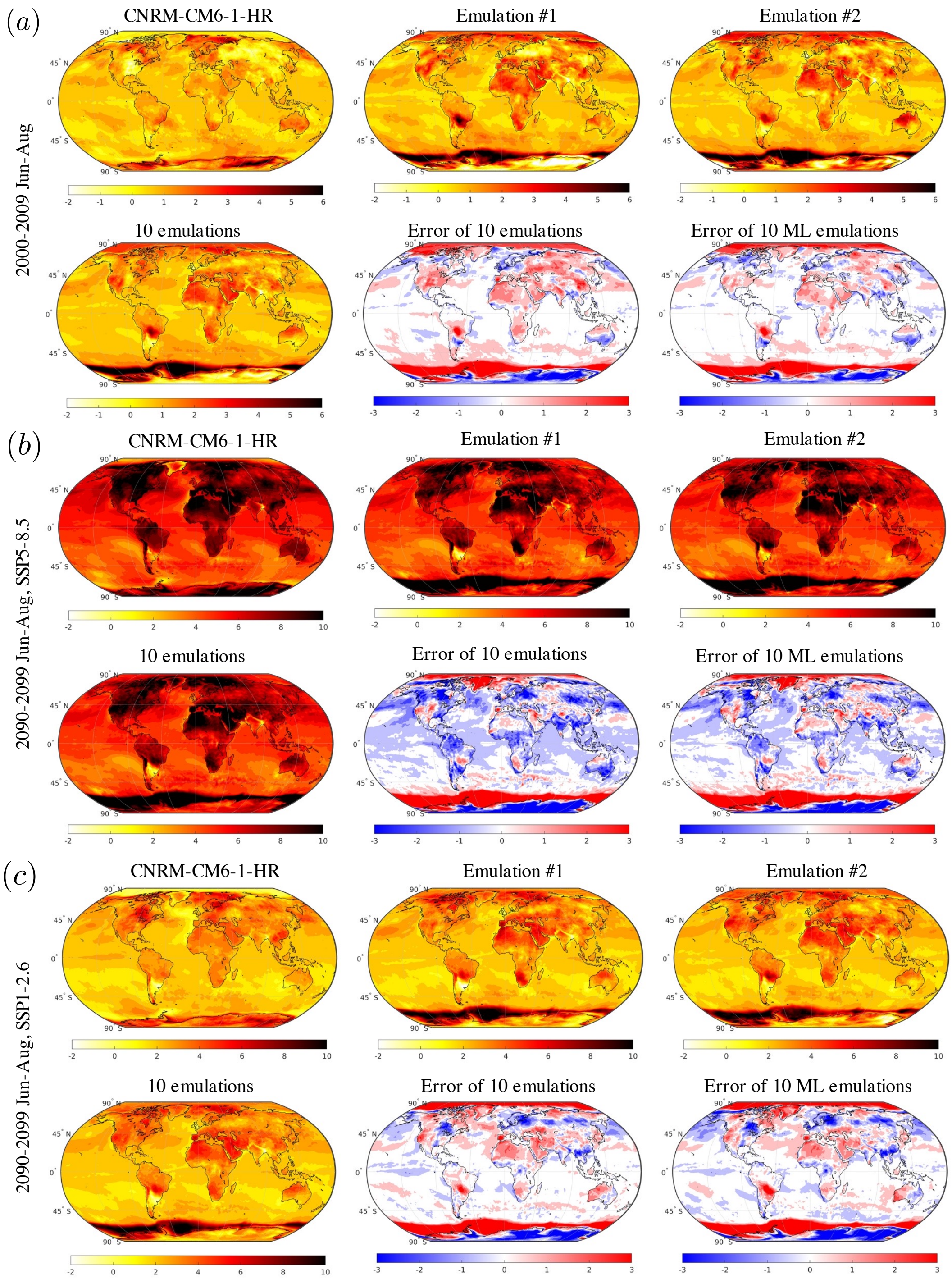

Figure 8. Extreme anomaly of Jun-Aug daily maximum temperature, quantified by the 97.5% quantile of local TMX distribution. The quantiles are evaluated using data from (*a*) 2000-2009, (*b*) 2090-2099 of the SSP5-8.5 scenario, and (*c*) 2090-2099 of the SSP1-2.6 scenario. Each subfigure shows the true mean from CNRM-CM6-1-HR ESM, two sample emulations, average of 10 emulations, error of 10 emulations, and the error of 10 ML emulations. Reference: 1850-1900 Jun-Aug 97.5% quantile of TMX.

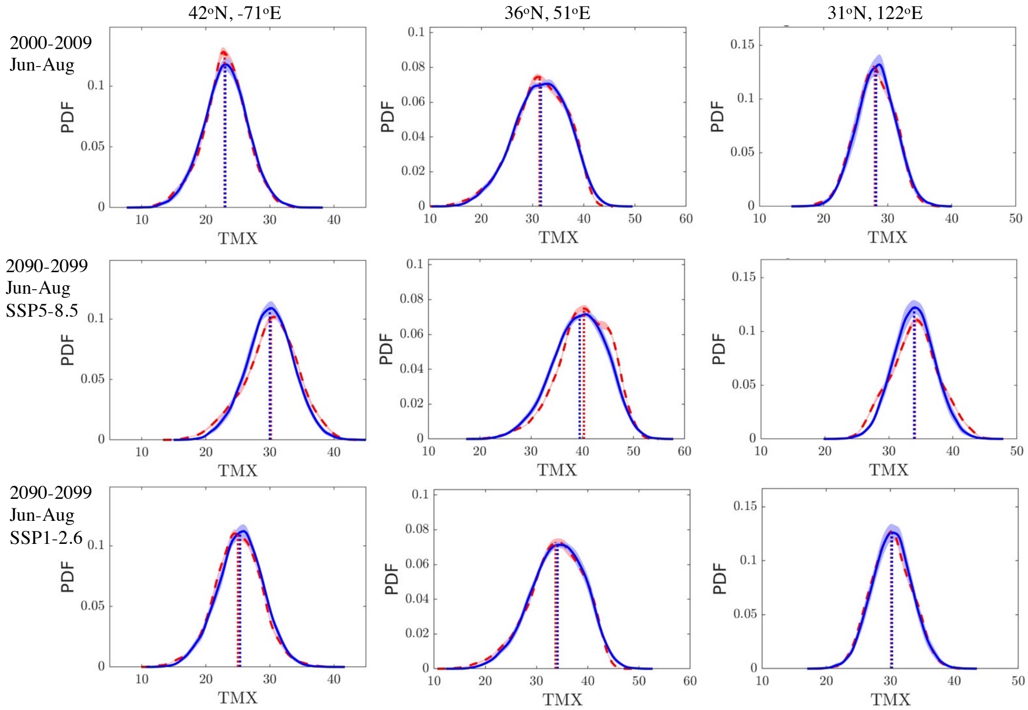

Figure 9. Probability density function (PDF) of local daily maximum temperature, averaged over three $1^\circ \times 1^\circ$ regions that include major cities. Left to right columns: Boston ($42^\circ N$,$-71^\circ E$), Tehran ($36^\circ N$,$51^\circ E$) and Shanghai ($31^\circ N$,$122^\circ E$). Red dashed line: CNRM-CM6-1-HR simulation data; red shaded region: uncertainty of the true PDF computed by bootstrapping; solid line: 10 emulations; blue shaded region: uncertainty of PDf quantified by one standard deviation of 10 emulations. The PDF are evaluated in decadal windows: (top row) historical, 2000-2009; (middle row) 2090-2099, SSP5-8.5 scenario; (bottom row) 2090-2099, SSP1-2.6 scenario. TMX are shown using degree Celsius.

The error in mean TMX remains relatively consistent across seasons and future scenarios. Similarly, the standard deviation error is nearly stationary and independent of seasons over historical periods. However, in SSP5-8.5 future scenario, seasonal variation becomes more pronounced, with the error in Sep-Nov at the end of the century almost doubling that of Dec-Feb. The end period of SSP5-8.5 scenario is the most difficult to predict, because of the reduced representation accuracy of leading EOFs trained from historical data. Additionally, the availability of only a single realization limits the emulator's ability to accurately estimate the most extreme warming conditions. The more pronounced error in Sep-Nov is due to the more significant influence of global warming on Sep-Nov statistics of TMX. Specifically, the Sep-Nov standard deviation of TMX is decreasing not only in the Southern Ocean, but also in the Arctic, which are not accurately captured by the emulator (see Appendix B for global distribution of standard deviations). The SSP1-2.6 future scenario exhibits similar seasonal error variations, albeit with generally lower magnitudes compared to SSP5-8.5. Regarding the 97.5% quantiles, their RMSE patterns align closely with those observed for the standard deviation, reflecting the same underlying climate dynamics. Despite these seasonal variations, the overall error magnitude remains relatively consistent across all four seasons throughout the emualated time and scenarios, which justifies the application of the emulator across the entire annual cycle.

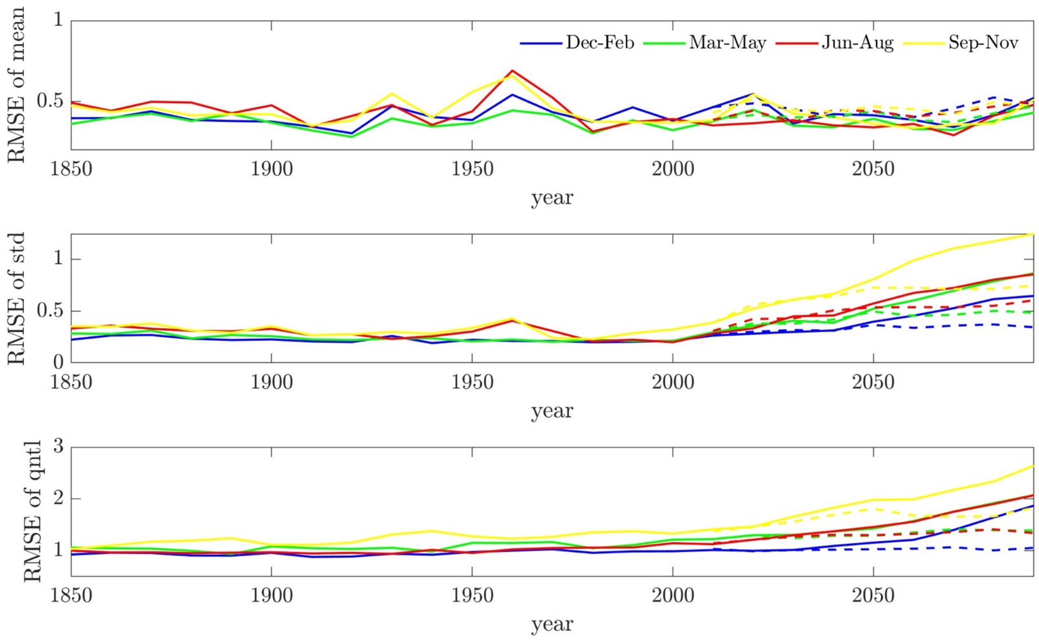

Figure 10. Root-mean-square error of the mean, standard deviation, and 97.5% quantile of TMX in different seasons. Solid lines: historical and SSP5-8.5 future scenario; dashed lines: SSP1-2.6 future scenario. Blue, green, red, yellow: errors averaged in Dec-Feb, Mar-May, Jun-Aug, Sep-Nov.

4.2 Emulation of MPI-ESM1-2-LR large-ensemble dataset

When a large ensemble of realizations are available, the assumption of constant cross-mode covariance in the emulator (equation 9) can be relaxed. Specifically, we generalize the emulator of EOF time series (equation 6) by modeling $\hat{l}_{s,ij}$ as a function of the global mean temperature,

$$\hat{a}_{s,i}(t,\hat{\omega}) = \hat{\mu}_{s,i}\left(T_{s,g}\right) + \sum_{j=1}^{I} \hat{l}_{s,ij}\left(T_{s,g}\right)\hat{\eta}_{s,j}(t,\hat{\omega}), \quad i = 1,2,\ldots,I. \tag{13}$$

In order to estimate the relation between $\hat{l}_{s,ij}$ and $T_{s,g}$, we follow similar procedures as §3.2. Given the true EOF time series $\mathbf{a}(t)$, we remove the linear trends of seasonal mean $\hat{\boldsymbol{\mu}}_s(T_{s,g})$, compute the covariance of $\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s$ in each year, and perform Cholesky decomposition of the covariance matrix,

$$\bar{\boldsymbol{\Sigma}}_s(t) = \left\langle \left(\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s\right)\left(\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s\right)^\top \right\rangle_{s\omega}, \quad \bar{\boldsymbol{\Sigma}}_s(t) = \bar{\mathbf{L}}_s(t)\bar{\mathbf{L}}_s^\top(t), \tag{14}$$

where $\langle\cdot\rangle_{s\omega}$ denotes an average over the ensemble and season s in each year. An intuitive but risky idea is modelling each entry of $\bar{\boldsymbol{\Sigma}}_s(t)$ as a linear function of the global mean temperature. Such a strategy cannot guarantee the positive definite property of the estimated covariance matrix. This limitation can be overcome by modelling $\bar{\mathbf{L}}_s(t)$ as linear functions of $T_{s,g}$,

$$\hat{\mathbf{L}}_s(T_{s,g}) = \hat{\mathbf{P}}_{s,0} + T_{s,g}\hat{\mathbf{P}}_{s,1}. \tag{15}$$

Since $\hat{\mathbf{L}}_s(T_{s,g})$ is lower triangular, $\hat{\mathbf{P}}_{s,0}$ and $\hat{\mathbf{P}}_{s,1}$ inherit this property, and each of their non-zero entries is computed by the method of least squares. Multiplying $\mathbf{a}_s - \hat{\boldsymbol{\mu}}_s$ by $\hat{\mathbf{L}}_s^{-1}(T_{s,g})$, we can extract the time series that are approximately uncorrelated in each season of each year,

$$\boldsymbol{\eta}_s(t,\omega) = \hat{\mathbf{L}}_s^{-1}(T_{s,g})\left(\mathbf{a}_s(t,\omega) - \hat{\boldsymbol{\mu}}_s\right). \tag{16}$$

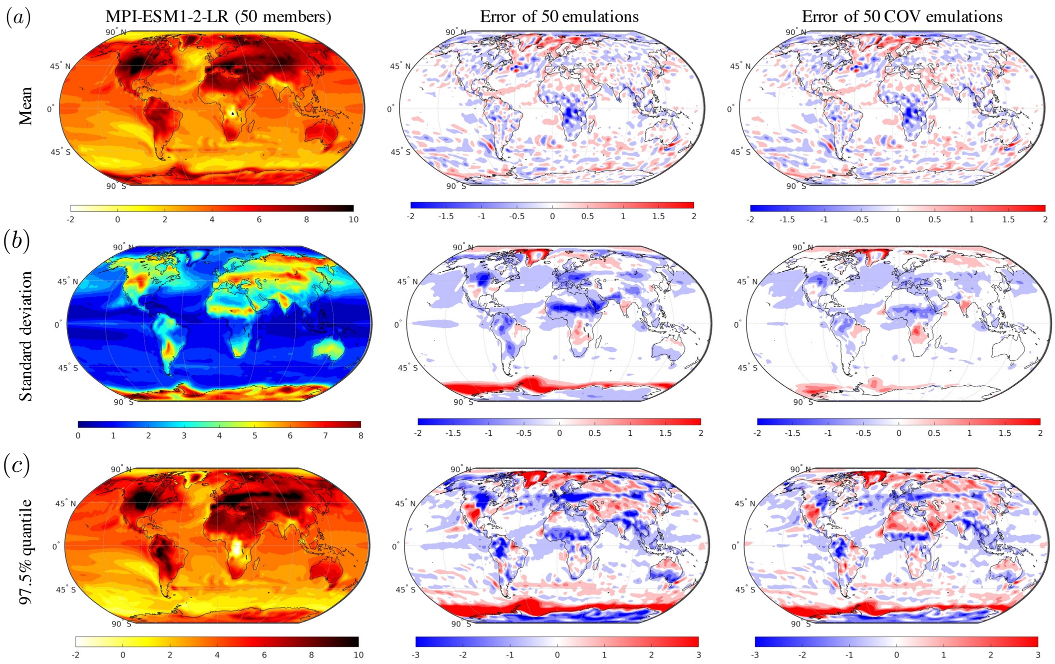

Figure 11. Statistics of Jun-Aug daily maximum temperature of MPI-ESM1-2-LR dataset and the emulations. All the statistics are evaluated in 2090-2099 of the SSP5-8.5 scenario. (a) Mean anomaly from 1850-1900; (b) Standard deviation; (c) Anomaly of 97.5% quantile of local TMX distribution from the 1850-1900 value.

The autocorrelation of each component of $\boldsymbol{\eta}_s$ will be used to generate Gaussian processes. The remaining procedures for constructing the emulator are the same as in §3.2 and therefore not repeated here for conciseness.

Although the generalization introduced in (13-15) has the potential to improve the performance of the emulator, it is only applicable when the data are sufficient to obtain converged time-dependent covariance matrices. A minimum requirement for the amount of data is that the number of samples for computing the covariance matrix (14) must exceed the number of EOFs, or equivalently the size of $\bar{\boldsymbol{\Sigma}}_s(t)$. This requirement is not satisfied by the CNRM-CM6-1-HR dataset. For example, in Northern Hemisphere summer of every year we have 92 samples to compute $\bar{\boldsymbol{\Sigma}}_s(t)$, but the number of EOFs used in the emulator is 2,000. As a result, the computed covariance matrix is not even full rank, consisting of spurious correlations that contaminate the dependence on time or global mean temperature.

To distinguish from the emulator introduced in §3.2, all the results generated using (13-15) will be termed as COV emulations. Both types of emulators are applied to the MPI-ESM1-2-LR dataset to compare their performance. Different from the CNRM-CM6-1-HR dataset that requires 2,000 EOFs to represent 95% of the total variance, only 1,000 EOFs are sufficient to model the MPI-ESM1-2-LR dataset due to lower spatial resolutions. All the 50 realizations of the historical and SSP5-8.5 scenarios are used to compute the EOFs and train the stochastic emulators of the EOF time series.

Since the error of emulated statistics were highest in SSP585 2090-2099 for the CNRM-CM6-1-HR dataset, we focus on this time window to compare the performance of the emulators. The results are visualized in figure 11. Overall the warming trend predicted by MPI-ESM1-2-LR model is less pronounced than the CNRM-CM6-1-HR model, which

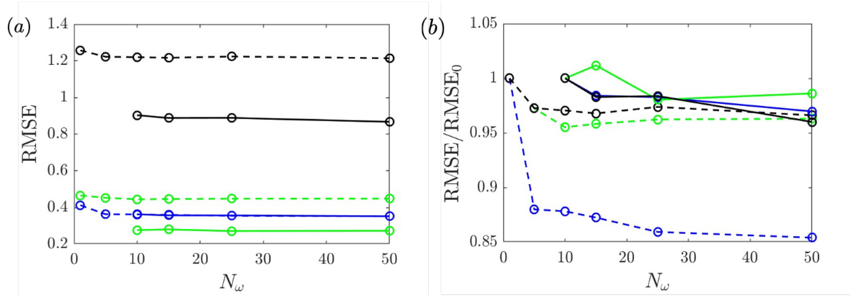

Figure 12. (a) Absolute and (b) Relative root-mean-square error of emulated statistics versus the number of realizations used for training the emulator. The statistics are evaluated in SSP5-8.5 2090-2099 Jun-Aug. Dashed lines: error of 50 emulations; Solid lines: error of 50 COV emulations. Blue, green, black: error of the mean, standard deviation, and 97.5% quantile. The relative errors in (b) are normalized by the values associated with the smallest N_ω.

is consistent with previous studies on equilibrium climate sensitivity of ESMs (Tokarska et al., 2020). In figure 11a, the error of the mean anomaly of both emulators are almost identical, which is expected since the same linear model is adopted for the seasonal mean of EOF coefficients. The error of local standard deviation, as shown in panel b, is significantly reduced by modeling the variations of covariance matrix. For example, the highest errors in North Africa and the Southern Ocean are decreased by approximately $2°C$, which confirms the speculation in §4.1 that these errors are mostly associated with time-dependent cross-mode correlations. As a result of more accurate estimation of local variance in COV emulations, the quantiles in panel c are also reproduced with lower errors.

To assess the influence of ensemble size of the training data on both emulators, we calculated the root-mean-square error (RMSE) of the emulated statistics. The results are reported in figure 12 for the mean, standard deviation, and 97.5% quantile in SSP5-8.5 2090-2099 Jun-Aug. When N_ω realizations are available for training the emulator, the true statistics $\mathcal{Q}$ are also evaluated using the same N_ω realizations, while the emulators are always performed 50 times to generate converged statistics, $\hat{\mathcal{Q}}$. In panel a, compared with the constant-covariance emulator (dashed lines), the COV emulator (solid lines) achieves approximately 40% error reduction in the standard deviation and 30% in the quantile. However, the COV emulator requires at least ten realizations to ensure the positive definiteness of the covariance matrices. To highlight the dependence of emulation error on the ensemble size N_ω, the RMSE is normalized by the value associated with the smallest N_ω attempted. The results are shown in figure 12b. For the constant-covariance emulator (dashed lines), as the size of ensemble is increased from one to ten, the RMSE of mean, standard deviation and quantile are respectively decreased by 12%, 4.5% and 3.0%. These error reductions suggest that the emulation accuracy is generally improved when the impact of climate internal variability is alleviated in the training data. Such a trend is also consistent with conclusions of previous studies (Tebaldi et al., 2021) that approximately ten realizations are required to capture the ensemble variance accurately. As the ensemble size reaches 50, further error reduction becomes negligible for the standard deviation (green dashed) and quantile (black dashed), suggesting diminishing returns from larger training datasets. In contrast, the COV emulator shows continued improvement, with a reduction in error of 1.4% for the standard deviation and 4.0% for the quantile, since larger-ensemble data can still help improve the emulated covariance matrices. Despite these gains, the COV emulator constructed with ten ensemble members already provides an accurate estimation of the statistics of extreme temperature. These results indicate that as long as the amount of training data are sufficient to construct the COV emulator, the performance of the emulator is robust against the ensemble size of realizations.

5 Conclusions and Discussion

We have developed a framework of a spatially resolved stochastic emulator that estimates the full statistics of climate extremes. The emulator was trained and tested using the daily maximum temperature data from CNRM-CM6-1-HR and MPI-ESM1-2-LR Earth system simulations in CMIP6. To reduce the dimensionality of the global climate system and achieve speedy emulations, we extract empirical orthogonal functions of daily maximum temperature data and assume their shapes remain unchanged across different climate change scenarios. The time series of EOF coefficients are decomposed as the combination of long term trends of seasonal statistics and conditionally Gaussian daily fluctuations. The former, including seasonal mean and variance, are approximated as linear or machine-learned functions of the global mean temperature, while the daily fluctuations are modeled as Gaussian autoregressive processes that are scaled by the cross correlations of different EOFs.While the statistics of the emulator, conditioned on season and global mean temperature, are assumed to be Gaussian, the long term statistics of the model do not produce normal distribution due to variation of the global mean temperature. However, the possibility of heavy tailed daily temperature fluctuations is not covered and is left for future work.

The performance of the emulator is evaluated on the CNRM-CM6-1-HR dataset due to its high spatial resolution. Trained on historical and SSP5-8.5 scenario, the emulated time series accurately reproduce the evolution of the seasonal mean and the Fourier spectra of daily fluctuations. After generating the spatiotemporal evolution of the instantaneous daily maximum temperature, the emulator's performance is systematically evaluated on the ten-year Jun-Aug statistics, including the mean, standard deviation, quantile, and the full probability density function. Remarkably, the emulator reproduces the quantile anomaly in response to climate change and effectively captures the non-Gaussian profiles of the local PDF. When tested on the SSP1-2.6 scenario that is not included in the training data, the full statistics are also accurately predicted, which demonstrates the potential of the emulator to be applied to various climate change scenarios. While using neural networks to represent the impact of global warming improves the emulator's performance on the training SSP5-8.5 scenario compared to linear functions, this improvement does not extend to the SSP1-2.6 scenario used for validation.

Based on MPI-ESM1-2-LR large-ensemble datasets, we further developed the emulator by modelling the variation of the cross-mode covariance as linear functions of the global mean temperature. Such a refinement helps reduce the root-mean-square error of emulated local statistics by 50%. By progressively increasing the number of ensemble members in the training data, we assessed the impact of climate internal variability on performance of both emulators. Overall the RMSE of statistics decrease with larger ensemble. When more than ten members are included, the accuracy of the constant-covariance emulator approximately saturates, but COV emulator shows continued improvement. As long as there are sufficient training data to construct the COV emulator, its performance remains relatively stable regardless of the ensemble size of realizations.

There are numerous pathways for generalizing the emulator to further improve its accuracy, and we outline a few possibilities below. First, the time-lagged covariance between different EOFs can be included into the emulator to achieve a better estimation of the full probability distribution of local temperature (Wan et al., 2021). Second, instead of using the global mean temperature as the driver, the emulator can be parameterized using the emission history of greenhouse gases, the equivalent radiative forcing, or aerosol concentrations (Castruccio et al., 2014; Freese et al., 2024). Such an extension will take into account the memory effect and facilitate the application of the emulator into scenarios where the evolution of global mean temperature is non-monotonic. Third, the Empirical Orthogonal Functions can be replaced by more state-of-the-art deep learning methods, such as Autoencoders, to nonlinearly reduce the dimensionality of the climate system (Kramer, 1991). Lastly, a recently proposed non-intrusive machine-learning framework shows promise for further improving the emulator's accuracy (Barthel Sorensen et al., 2024). This approach focuses on learning a debiasing operator that takes the emulated time series of temperature fields as input and corrects them to better match the reference data from ESMs. Once trained on a few scenarios, this debiasing operator can be applied to correct the emulations in other unseen climate change scenarios. Despite these potential enhancements, the emulator successful estimation of extreme temperature statistics is promising and suggests its applicability to other variables, such as humidity, precipitation, and wind speed, which will better assist with risk management of climate extremes.

Appendix A Temporal evolution of emulated quantile in SSP5-8.5 scenario

In this appendix, we provide more details about the temporal evolution of statistics of extreme temperature in SSP5-8.5 scenario. Similar to figure 8, we evaluate the 97.5% quantile of the local TMX using ten-year Jun-Aug data. The anomaly of quantiles against 1850-1900 reference are visualized in figure A1 and A2 from 2010 to 2089. Overall the regions with the most rapid increase of extreme temperature are correctly identified by the emulator. Two categories of error patterns can be observed. The first type is relatively independent of time, such as the overestimated quantile in Greenland. The second type is more stochastic, sometime even changing signs across different time windows, such as the North America and southern Africa. These error patterns are probably associated with the internal variability of the global climate system and require more realizations of the Earth system simulations to converge.

Appendix B Emulated statistics in other seasons

This appendix presents the statistics of TMX across different seasons and their corresponding emulation errors. The local standard deviation in 2090-2099 of the SSP5-8.5 scenario is shown in figure B1. In Dec-Feb, the error reaches its maximum in the Arctic, contrasting with the Jun-Aug pattern where the error peaks in the Southern Ocean (c.f. figure B1). This seasonal difference is likely associated with the sea ice coverage. During Dec-Feb, Antarctic sea ice consistently retreats almost to the coastline in both historical and global warming scenarios. Therefore, the standard deviation of TMX in this season is less affected by warming conditions compared to Jun-Aug. Mar-May and Sep-Nov present a more complex picture. During these transitional seasons, sea ice coverage in both polar regions is highly sensitive to climate change. The emulator struggles to capture the associated trends in standard deviations, resulting in high errors in these areas. The error patterns of 97.5% quantile are analogous to the standard deviation, as shown in figure B2.

Open Research

All code to reproduce this work is available at `https://github.com/mzwang2012/sEM_TMX.git`. The raw data from CMIP6 were retrieved through the Earth System Grid Federation interface `https://aims2.llnl.gov/search/cmip6/`.

Acknowledgments

This research was part of the Bringing Computation to the Climate Challenge (BC3) project and supported by Schmidt Sciences through the MIT Climate Grand Challenges.

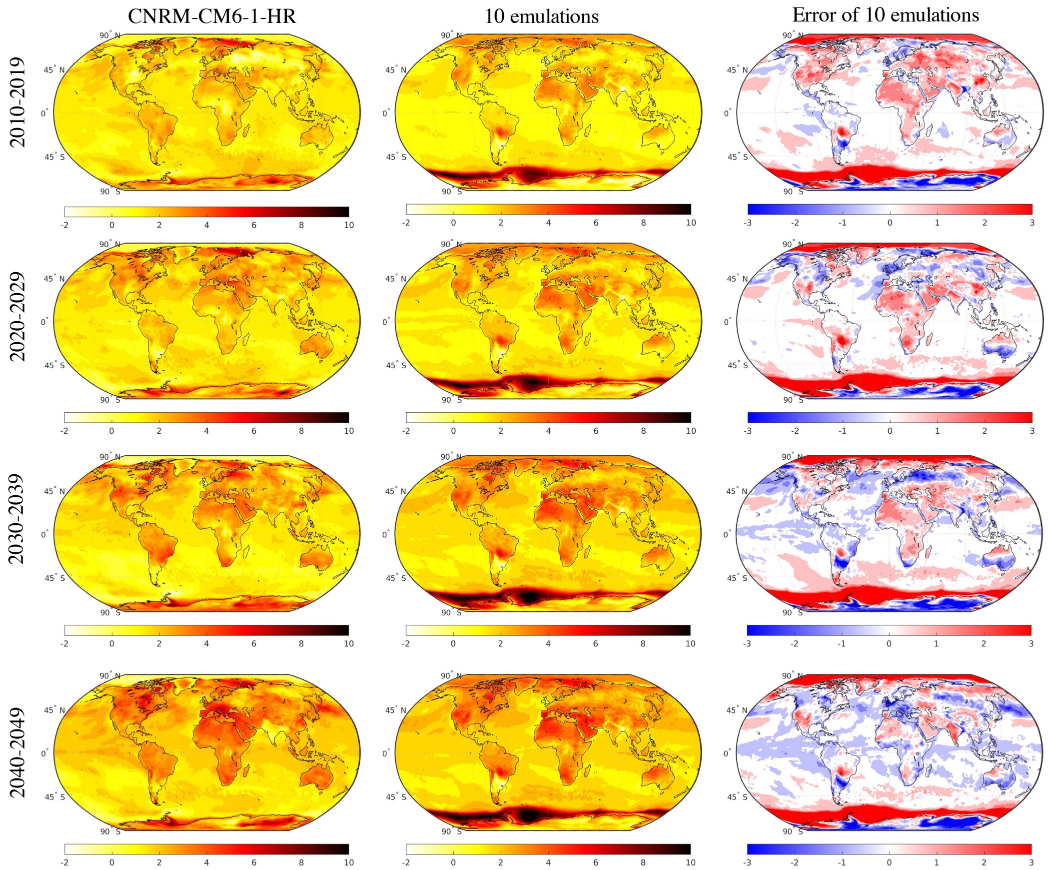

Figure A1. Extreme anomaly of ten-year Jun-Aug daily maximum temperature, quantified by the 97.5% quantile of local TMX distribution. The quantiles are evaluated for SSP5-8.5 scenario within 2010-2019, 2020-2029, 2030-2039, 2040-2049, respectively. Reference: 1850-1900 Jun-Aug 97.5% quantile of TMX.

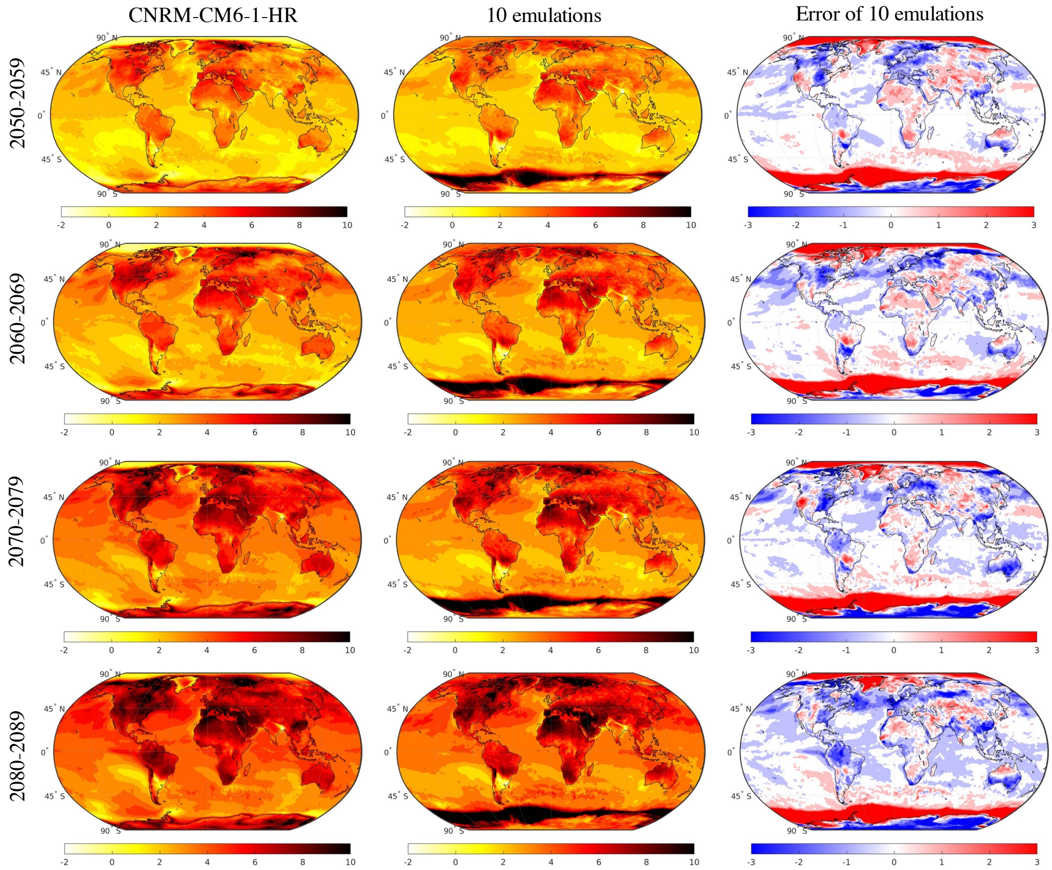

Figure A2. Same as figure A1, but shown for 2050-2059, 2060-2069, 2070-2079, 2080-2089, respectively.

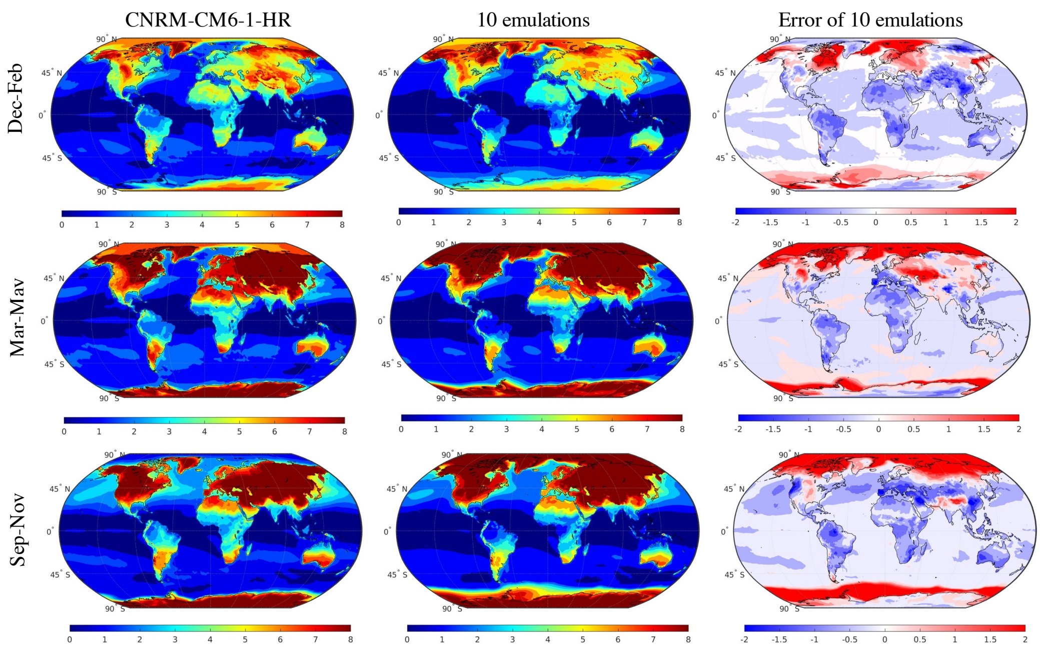

Figure B1. Standard deviation of ten-year seasonal daily maximum temperature, evaluated for Dec-Feb, Mar-May, and Sep-Nov in 2090-2099 of the SSP5-8.5 future scenario.

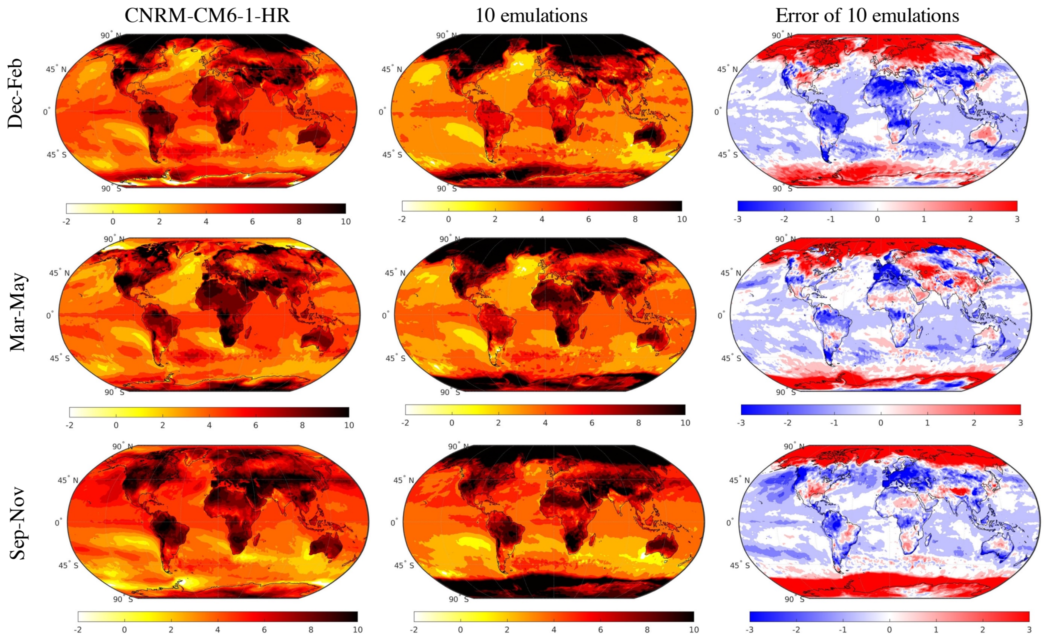

Figure B2. Extreme anomaly of ten-year seasonal daily maximum temperature, quantified by the 97.5% quantile of local TMX distribution. The quantiles are evaluated for Dec-Feb, Mar-May, and Sep-Nov in 2090-2099 of the SSP5-8.5 future scenario.. Reference: 1850-1900 97.5% quantile of TMX of each season.

References

Alexeeff, S. E., Nychka, D., Sain, S. R., & Tebaldi, C. (2018). Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments. *Climatic Change*, *146*, 319–333.

Allen, S., Barros, V., (Canada, I., (UK, D., Cardona, O., Cutter, S., . . . (USA, T. (2012, nov). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change.* doi: 10.13140/2.1.3117.9529

Amaya, D. J. (2019). The pacific meridional mode and enso: A review. *Current Climate Change Reports*, *5*(4), 296–307.

AON. (2020). *Weather, climate & catastrophe insight, 2020 annual report.*

Arbabi, H., & Sapsis, T. (2022). Generative stochastic modeling of strongly nonlinear flows with non-gaussian statistics. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(2), 555–583.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., & García-Herrera, R. (2011). The hot summer of 2010: Redrawing the temperature record map of europe. *Science*, *332*(6026), 220-224.

Barthel Sorensen, B., Charalampopoulos, A., Zhang, S., Harrop, B., Leung, L., & Sapsis, T. P. (2024). A non-intrusive machine learning framework for debiasing long-time coarse resolution climate simulations and quantifying rare events statistics. *Journal of Advances in Modeling Earth Systems*, *16*(3), e2023MS004122.

Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020). Emulating earth system model temperatures with mesmer: from global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*, *11*(1), 139–159.

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., & Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed gcm runs. *Journal of Climate*, *27*(5), 1829–1844.

Fogt, R. L., & Marshall, G. J. (2020). The southern annular mode: variability, trends, and climate impacts across the southern hemisphere. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(4), e652.

Freese, L. M., Fiore, A. M., & Selin, N. E. (2024). Spatially resolved temperature response functions to co2 emissions. *Authorea Preprints*.

Gao, Y., Leung, L. R., Lu, J., & Masato, G. (2015). Persistent cold air outbreaks over north america in a warming climate. *Environmental Research Letters*, *10*(4), 044001.

Geogdzhayev, G., Souza, A., Ferrari, R., & Flierl, G. R. (2024). *A statistical emulator design for averaged climate fields.* (Personal Communications)

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(9), 1119–1152.

Huntingford, C., Jones, P. D., Livina, V. N., Lenton, T. M., & Cox, P. M. (2013). No increase in global temperature variability despite changing regional patterns. *Nature*, *500*(7462), 327–330.

Kaltenborn, J., Lange, C., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., ... Rolnick, D. (2023). Climateset: A large-scale climate model dataset for machine learning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 21757–21792). Curran Associates, Inc.

Kalvová, J., & Nemešová, I. (1998). Estimating autocorrelations of daily extreme temperatures in observed and simulated climates. *Theoretical and applied climatology*, *59*, 151–164.

Kapmeier, F., Greenspan, A., Jones, A., & Sterman, J. (2021). Science-based analysis for climate action: how hsbc bank uses the en-roads climate policy simulation. *System dynamics review: the journal of the System Dynamics Society*, *37*(4), 333–352.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, *37*(2), 233–243.

Link, R., Snyder, A., Lynch, C., Hartin, C., Kravitz, B., & Bond-Lamberty, B. (2019). Fldgen v1. 0: an emulator with internal variability and space–time correlation for earth system models. *Geoscientific Model Development*, *12*(4), 1477–1489.

Lorenz, E. N. (1956). *Empirical orthogonal functions and statistical weather prediction* (Vol. 1). Massachusetts Institute of Technology, Department of Meteorology Cambridge.

Lütjens, B., Ferrari, R., Watson-Parris, D., & Selin, N. (2024). The impact of internal variability on benchmarking deep learning climate emulators. *arXiv preprint arXiv:2408.05288*.

Meehl, G. A., & Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, *305*(5686), 994–997.

Meinshausen, M., Raper, S. C., & Wigley, T. M. (2011). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, magicc6–part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, *11*(4), 1417–1456.

Mitchell, T. D. (2003). Pattern scaling: an examination of the accuracy of the technique for describing future climates. *Climatic change*, *60*(3), 217–242.

Mohamad, M. A., & Sapsis, T. P. (2015). Probabilistic description of extreme events in intermittently unstable dynamical systems excited by correlated stochastic processes. *SIAM/ASA Journal on Uncertainty Quantification*, *3*(1), 709–736.

Nicholls, Z., Meinshausen, M., Lewis, J., Corradi, M. R., Dorheim, K., Gasser, T., ... others (2021). Reduced complexity model intercomparison project phase 2: Synthesizing earth system knowledge for probabilistic climate projections. *Earth's Future*, *9*(6), e2020EF001900.

Nicholls, Z. R., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., ... others (2020). Reduced complexity model intercomparison project phase 1: Protocol, results and initial observations. *Geoscientific Model Developments*.

Olonscheck, D., & Notz, D. (2017). Consistently estimating internal climate variability from climate model simulations. *Journal of Climate*, *30*(23), 9555–9573.

Percival, D. B. (1993). Simulating gaussian random processes with specified spectra. *Computing Science and Statistics*, 534–534.

Quilcaille, Y., Gudmundsson, L., Beusch, L., Hauser, M., & Seneviratne, S. I. (2022). Showcasing mesmer-x: Spatially resolved emulation of annual maximum temperatures of earth system models. *Geophysical Research Letters*, *49*(17), e2022GL099012.

Räisänen, J. (2002). Co2-induced changes in interannual temperature and precipitation variability in 19 cmip2 experiments. *Journal of Climate*, *15*(17), 2395–2411.

Reed, K. A., Wehner, M. F., & Zarzycki, C. M. (2022). Attribution of 2020 hurricane season extreme rainfall to human-induced climate change. *Nature communications*, *13*(1), 1905.

Rooney-Varga, J. N., Hensel, M., McCarthy, C., McNeal, K., Norfles, N., Rath, K., ... Sterman, J. D. (2021). Building consensus for ambitious climate action through the world climate simulation. *Earth's Future*, *9*(12), e2021EF002283.

Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., & Wilby, R. L. (2016). Allowable co2 emissions based on regional and impact-related climate targets. *Nature*, *529*(7587), 477–483.

Sirovich, L. (1987). Turbulence and the dynamics of coherent structures. i. coherent structures. *Quarterly of applied mathematics*, *45*(3), 561–571.

Taira, K., Hemati, M. S., Brunton, S. L., Sun, Y., Duraisamy, K., Bagheri, S., ... Yeh, C.-A. (2020). Modal analysis of fluid flows: Applications and outlook. *AIAA journal*, *58*(3), 998–1022.

Tebaldi, C., Armbruster, A., Engler, H., & Link, R. (2020). Emulating climate extreme indices. *Environmental Research Letters*, *15*(7), 074006.

Tebaldi, C., Dorheim, K., Wehner, M., & Leung, R. (2021). Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates. *Earth System Dynamics*, *12*(4), 1427–1501.

Thompson, D. W., & Wallace, J. M. (1998). The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, *25*(9), 1297–1300.

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. *Science advances*, *6*(12), eaaz9549.

Wallace, J. M., & Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly weather review*, *109*(4), 784–812.

Wan, Z. Y., Dodov, B., Lessig, C., Dijkstra, H., & Sapsis, T. P. (2021). A data-driven framework for the stochastic reconstruction of small-scale features with application to climate data sets. *Journal of Computational Physics*, *442*, 110484.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., ... others (2022). Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002954.

Wehner, M., Gleckler, P., & Lee, J. (2020). Characterization of long period return values of extreme daily temperature and precipitation in the cmip6 models: Part 1, model evaluation. *Weather and Climate Extremes*, *30*, 100283.