
Spatially-resolved emulation of climate extremes via machine learning stochastic models

Mengze Wang

Massachusetts Institute of Technology
Boston, MA 02139
mzwang@mit.edu

Andre Souza

Massachusetts Institute of Technology
Boston, MA 02139
sandre@mit.edu

Raffaele Ferrari

Massachusetts Institute of Technology
Boston, MA 02139
rferrari@mit.edu

Themistoklis Sapsis

Massachusetts Institute of Technology
Boston, MA 02139
sapsis@mit.edu

Abstract

Emulators, or reduced-complexity models, serve as an ideal complement to earth system models (ESM) by providing the climate information under various scenarios at much lower computational costs. We develop an emulator of climate extremes that produce the temporal evolution of probability distributions of local variables on a spatially resolved grid. The representative modes of climate change are identified using principal component analysis (PCA), and the PCA time series are approximated using stochastic models. When applied to ERA5 data, the model accurately reproduces the quantiles of local daily maximum temperature and effectively captures the non-Gaussian statistics. We also discuss potential generalization of our emulator to different climate change scenarios.

1 Introduction

Prediction of extreme events under climate change is challenging but essential for local communities, businesses, and governments to manage risks [1, 2, 3]. For example, an increasing number of heat wave events have been reported during the Northern Hemisphere summer, which has led to severe consequences for local economies [4]. Due to the chaotic property of the climate system and the unlikely nature of the extreme events, obtaining the statistics of extreme events from ensembles of Earth system models (ESM) is prohibitively expensive [5]. There is a growing need for emulators, or reduced-complexity models, that produce an accurate and efficient estimation of the statistics of extreme events in response to different policy scenarios.

Existing emulators that assess climate change can be categorized by the type of their output. A majority of emulators focused on predicting the statistics of the global-mean or regional-mean quantities [6, 7], including temperature or precipitation anomaly. The second category of emulators specialize in generating spatially-resolved local statistics, as well as the response to climate change. These emulators are generally based on pattern scaling, where the local responses are assumed as a linear function of global mean temperature [8]. The spatial correlation between grid points can be modelled using a Matern covariance function [9], and internal variability has been represented using autoregressive process or spectrum of principal components [10, 11]. However, the focus of most spatially-resolved emulators remains on mean quantities. Only a few recent studies have explored emulating the evolution of climate extreme indices, such as annual maximum temperature [12] or the duration of hot waves within a year [13]. To the authors' best knowledge, no prior work has reported

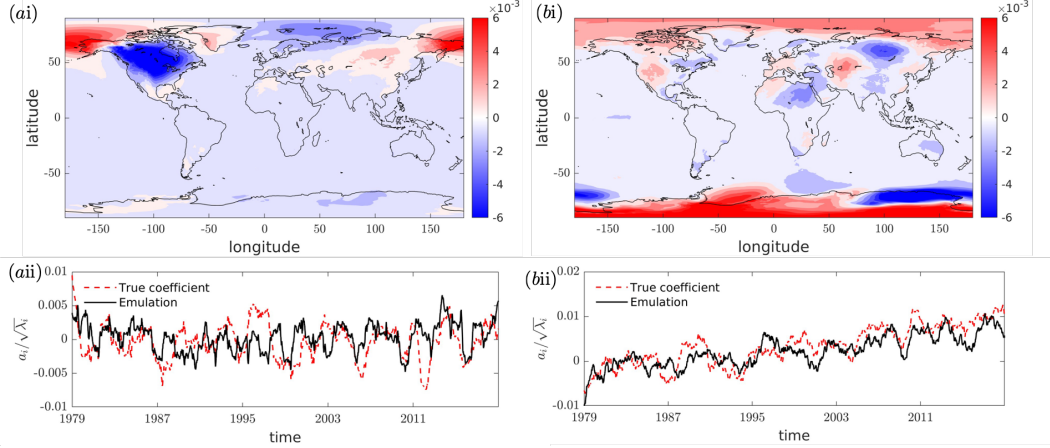


Figure 1: (i) Visualization of PCA modes extracted from 1979-1998 TMX data: (a) The most energetic mode; (b) The fastest growing mode. (ii) One-year moving averaged PCA time series from 1979 to 2018, normalized by the square root of the corresponding eigenvalues.

on the emulation of probability distribution of local climate variables, which constitutes the primary objective of our research.

We introduce a machine learning stochastic model to emulate the statistics of climate extremes, utilizing temperature-related extreme events as a prototypical application. We first perform a principal component analysis (PCA) of the global temperature fields to reduce the dimensionality of the system, while maintaining a high spatial resolution. The leading modes that represent climate change are identified. The temporal evolution of the PCA coefficients are modelled as stochastic processes, characterized by long-term trends, seasonal variations, and colored noise, thus making it possible to capture non-Gaussian statistics of local quantities. Our model can be further parameterized by the global mean temperature and generalized across diverse climate change scenarios.

2 Data and Methods

In this study we consider near-surface daily maximum temperature (TMX) data on a spatially resolved grid ($0.25^\circ \times 0.25^\circ$), obtained from ERA5 reanalysis [14]. The 1979-1998 data will be used for training, and 1999-2018 data for testing. At location \mathbf{x} and time t , the TMX is denoted as $q(\mathbf{x}, t)$. The climatological mean $\bar{q}(\mathbf{x}, t)$ is extracted by averaging TMX on the same day and location over a multi-year period. In other words, $\bar{q}(\mathbf{x}, t)$ has an annual periodicity $\bar{q}(\mathbf{x}, t + T) = \bar{q}(\mathbf{x}, t)$, where T corresponds to one year. The fluctuating fields are then decomposed as superposition of PCA modes $\phi_j(\mathbf{x})$,

$$q(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t) + \sum_j a_j(t)\phi_j(\mathbf{x}). \quad (1)$$

Two representative PCA modes are visualized in figure 1. Panel (a) shows the most energetic mode that is reminiscent of Arctic Oscillation/Northern Hemisphere Annular Mode [15]. The projection of instantaneous fields onto this particular mode (red dashed line in aii) is close to statistically stationary. Panel (b) displays the mode that experiences the most rapid linear growth in response to climate change; notably, the shape of this mode suggests increased heating in the polar regions.

Assuming the climatological mean and PCA modes in (1) are known, we model the time series of PCA coefficients as long term trends superposed with colored noise,

$$\mathbf{a}(t) = \boldsymbol{\mu}(t) + \mathbf{C}(t)\mathbf{w}(t). \quad (2)$$

The vector $\mathbf{a} = [a_1, a_2, \dots, a_J]$ consists of the coefficients of the first J PCA modes. The long-term trends $\boldsymbol{\mu}(t)$ are approximated as linear functions, $\mu_j(t) = p_{0j} + p_{1j}t$. The parameters p_{0j}, p_{1j} are estimated from a linear regression of the true PCA time series. Adaptation of $\boldsymbol{\mu}(t)$ to higher-order polynomials can be easily achieved using sparse regression. The piecewise-constant matrix \mathbf{C} accounts for cross-mode correlations, and the specific value of \mathbf{C} depends on seasons, which is

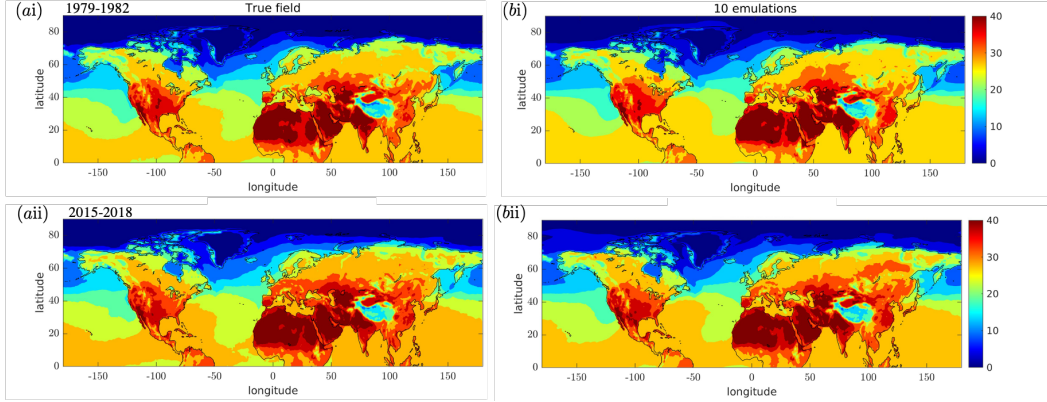


Figure 2: (a) 97.5% quantiles of local TMX distributions, computed using (i) 1979-1982 summer (June, July, Aug) and (ii) 2015-2018 summer data; (b) quantiles for the same 4-year window, computed from 10 emulations.

computed from a Cholesky decomposition of the seasonal-averaged correlation matrix of the residual ($\mathbf{a} - \boldsymbol{\mu}$). After removing the long-term trend and seasonal dependence, we store the spectra of the remaining part $\mathbf{C}^{-1}(\mathbf{a} - \boldsymbol{\mu})$, and the colored noise $\mathbf{w} = [w_1, w_2, \dots, w_J]$ is simulated from the stored spectra. For each realization of the stochastic process $\mathbf{w}(t)$, we substitute it into (2,1) to generate the estimated TMX fields, which will be referred as an emulation. More details of the models are provided in appendix A.

3 Results

The performance of our model is first evaluated in the context of matching PCA time series, as shown in figure 1(ii). Although only 1979-1998 data are available for training, the model output (black lines) correctly captures the long-term trend and variance of the true PCA time series (red dashed lines) even beyond the training window. Since the statistical properties of the PCA time series are reproduced accurately, we keep the first 500 modes and combine them with the mode shapes to emulate local statistics of the climate system. In order to acquire converged statistics from TMX snapshots, samples are collected from a localized neighborhood around each geographic coordinate (θ, φ) , specifically within the range $[\theta - 0.25^\circ, \theta + 0.25^\circ] \times [\varphi - 0.25^\circ, \varphi + 0.25^\circ]$. These samples are gathered during the summer months (June, July, August) in the Northern Hemisphere over a four-year window. Overall approximately 3,300 samples are used to compute any statistics of local TMX. The local mean and standard deviation are emulated with a reasonable accuracy, see appendix B. Below we focus on emulation of climate extremes.

The extreme temperatures corresponding to the 97.5% quantile are shown in figure 2. Similar indices of climate extreme events have been reported in previous emulators [13, 10, 12], but never on spatially resolved grids. Compared with the true fields in 1979-1982 (panel ai), the emulator (panel bi) successfully reproduces the spatial pattern of the quantiles, especially the hot regions such as mid-west America, north Africa, and west Asia. In 2015-2018 (panel ii), the warming trends in Europe, east Asia and mid-south US are predicted by the emulator with reasonable accuracy. The mismatch between the true quantiles and our model could arise from three sources: (i) interannual trends such as El Niño-Southern Oscillation (ENSO) which are not modelled; (ii) internal variability of climate extremes; (iii) higher-order statistics of the true PCA time series that are not included in the model. Upon analyzing the statistics in multiple four-year windows (appendix B), we ascertain that regions characterized by the highest estimation errors predominantly align with areas exhibiting substantial internal variability. Consequently, with the current dataset, it is difficult to differentiate model-induced errors from the inherent variability of the climate.

The probability density functions of local TMX are plotted at three geographically distinct locations: Tehran, featured by arid continental climate; Houston and Hong Kong, which are situated in proximity to the Atlantic and Pacific ocean, respectively. Our analysis reaffirms that the model successfully captures the increasing trend in local mean TMX. While the actual PDFs exhibit non-Gaussian

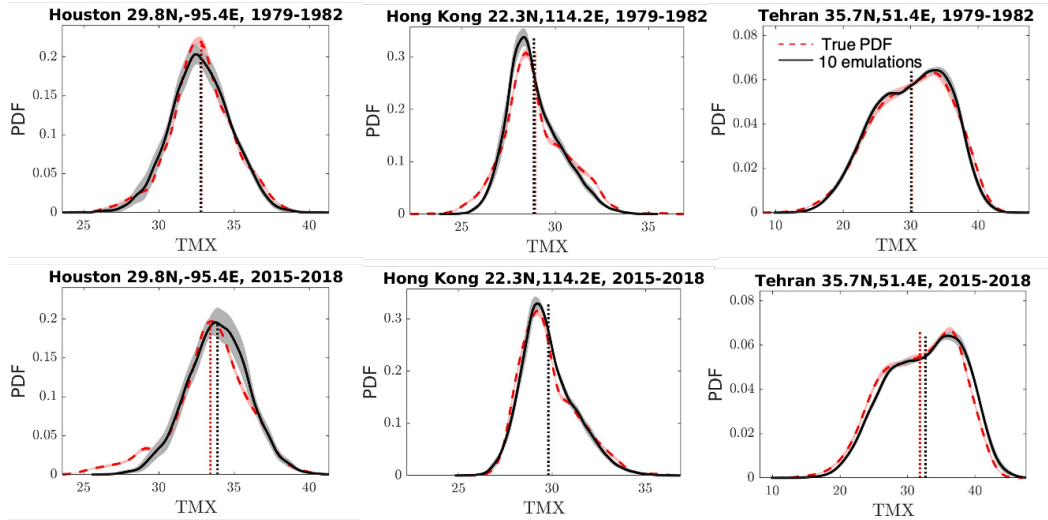


Figure 3: Probability density functions (PDF) of local TMX at selected cities. Red shaded region: uncertainties of the true PDF, computed by bootstrapping; Gray shaded region: one-standard-deviation uncertainty, computed from 10 emulations.

behavior (red dashed lines), their shapes are accurately replicated by the emulator (black lines) with small uncertainties. Emulating the tails of these distributions are more difficult, due to more pronounced internal variations and heightened data requirements for convergence. Nonetheless, our emulator demonstrates the capacity to approximate the tails with an acceptable level of accuracy.

4 Conclusions and future work

Utilizing ERA5 reanalysis data of daily maximum temperature, we attempted to develop a spatially-resolved emulator of climate extremes. Dimensionality reduction of the climate system was accomplished through principal component analysis, and the corresponding time series of PCA coefficients were modelled as stochastic processes featured by varying mean and seasonal correlations. We demonstrated the capacity of our emulator to predict spatially-resolved extreme local temperatures and the probability distribution of TMX, which has never been achieved in the existing literature on emulators for climate extremes.

Generalization of our stochastic model using more cutting-edge deep learning methods should be straightforward. For instance, the PCA can be replaced by autoencoders to perform nonlinear dimensionality reduction. The PCA time series may also be better represented using deep generative models such as diffusion model. In addition, our emulator can be further parameterized as functions of global mean temperature, thereby extending its applicability across a range of climate change scenarios. This avenue for generalization is currently the subject of ongoing research.

Acknowledgements

This research was supported by MIT Climate Grand Challenges.

References

- [1] Hans O Pörtner, Debra C Roberts, Helen Adams, Carolina Adler, Paulina Aldunce, Elham Ali, Rawshan Ara Begum, Richard Betts, Rachel Bezner Kerr, Robbert Biesbroek, et al. Climate change 2022: impacts, adaptation and vulnerability. Technical report, IPCC, 2022.
- [2] Tamma Carleton, Amir Jina, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Robert E Kopp, Kelly E McCusker, Ishan Nath, et al. Valuing the

- global mortality consequences of climate change accounting for adaptation costs and benefits. *The Quarterly Journal of Economics*, 137(4):2037–2105, 2022.
- [3] Elisabeth Vogel, Markus G Donat, Lisa V Alexander, Malte Meinshausen, Deepak K Ray, David Karoly, Nicolai Meinshausen, and Katja Frieler. The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, 14(5):054010, may 2019.
- [4] Ruonan Zhang, Chenghu Sun, Jieshun Zhu, Renhe Zhang, and Weijing Li. Increased european heat waves in recent decades in response to shrinking arctic sea ice and eurasian snow cover. *NPJ Climate and Atmospheric Science*, 3(1):7, 2020.
- [5] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.
- [6] Sonia I Seneviratne, Markus G Donat, Andy J Pitman, Reto Knutti, and Robert L Wilby. Allowable co2 emissions based on regional and impact-related climate targets. *Nature*, 529(7587):477–483, 2016.
- [7] Zebedee RJ Nicholls, Malte Meinshausen, Jared Lewis, Robert Gieseke, Dietmar Dommenges, Kalyn Dorheim, Chen-Shuo Fan, Jan S Fuglestedt, Thomas Gasser, Ulrich Golüke, et al. Reduced complexity model intercomparison project phase 1: Protocol, results and initial observations. *Geoscientific Model Developments*, 2020.
- [8] Timothy D Mitchell. Pattern scaling: an examination of the accuracy of the technique for describing future climates. *Climatic change*, 60(3):217–242, 2003.
- [9] Stacey E Alexeeff, Doug Nychka, Stephan R Sain, and Claudia Tebaldi. Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments. *Climatic Change*, 146:319–333, 2018.
- [10] Lea Beusch, Lukas Gudmundsson, and Sonia I Seneviratne. 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, 2020.
- [11] Robert Link, Abigail Snyder, Cary Lynch, Corinne Hartin, Ben Kravitz, and Ben Bond-Lamberty. Fldgen v1. 0: an emulator with internal variability and space–time correlation for earth system models. *Geoscientific Model Development*, 12(4):1477–1489, 2019.
- [12] Yann Quilcaille, Lukas Gudmundsson, Lea Beusch, Mathias Hauser, and Sonia I Seneviratne. Showcasing mesmer-x: Spatially resolved emulation of annual maximum temperatures of earth system models. *Geophysical Research Letters*, 49(17):e2022GL099012, 2022.
- [13] C Tebaldi, A Armbruster, HP Engler, and R Link. Emulating climate extreme indices. *Environmental Research Letters*, 15(7):074006, 2020.
- [14] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
- [15] David WJ Thompson and John M Wallace. The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, 25(9):1297–1300, 1998.
- [16] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [17] Donald B Percival. Simulating gaussian random processes with specified spectra. *Computing Science and Statistics*, pages 534–534, 1993.

Appendix A Detailed model formulation

A.1 Principal component analysis of data on spherical coordinate

Since the TMX data are collected from a approximately spherical surface, we adopt spherical integral to measure the distance between samples, instead of vector inner product as in conventional PCA formulation [16]. To simplify our discussion, the TMX snapshots or samples are denoted as $\{\mathbf{q}_n\}_{n=1}^N$, where \mathbf{q}_n is a column vector containing TMX data from different spatial locations. After removing the climatological mean from the dataset, $\mathbf{q}'_n = \mathbf{q}_n - \bar{\mathbf{q}}$, we search for an orthonormal basis $\Phi = [\phi_1, \phi_2, \dots, \phi_J]$ and a reduced-order representation of the fluctuation fields,

$$\hat{\mathbf{q}}_n = \Phi \mathbf{a}_n, \quad (3)$$

which minimize a cost function

$$J(\Phi, \mathbf{A}) = \frac{1}{N} \sum_{n=1}^N \|\mathbf{q}'_n - \hat{\mathbf{q}}_n\|_S^2. \quad (4)$$

The matrix \mathbf{A} consists of column vectors \mathbf{a}_n . The norm in (4) is defined as integral over the earth surface S ,

$$\|\mathbf{q}\|_S^2 = \int_S q^2(\theta, \varphi) \sin \theta d\theta d\varphi \approx \mathbf{q}^\top \mathbf{S} \mathbf{q}, \quad (5)$$

where (θ, φ) are latitude and longitude coordinates. The approximate equality in (5) refers to discretization, and \mathbf{S} is a diagonal matrix holding the integral weights at different locations. Introducing a matrix that stores all the fluctuation snapshots $\{\mathbf{q}'_n\}_{n=1}^N$,

$$\mathbf{Q} = [\mathbf{q}'_1, \mathbf{q}'_2, \dots, \mathbf{q}'_N], \quad (6)$$

the optimal orthonormal basis Φ that minimizes (4) is the eigenvector of matrix $\mathbf{Q}\mathbf{Q}^\top \mathbf{S}$. In other words, Φ solves the eigenvalue problem,

$$\mathbf{Q}\mathbf{Q}^\top \mathbf{S} \phi_j = \lambda_j \phi_j. \quad (7)$$

The leading PCA modes are visualized in figure 1 and used in (1) to reduce the dimensionality of the model.

A.2 Estimation of model parameters

To determine the parameters in our stochastic model,

$$\mathbf{a}(t) = \boldsymbol{\mu}(t) + \mathbf{C}(t)\mathbf{w}(t), \quad (8)$$

we project the 1979-1998 TMX fields onto the first 500 PCA modes to obtain the true coefficients $\mathbf{a}(t)$. We perform a linear regression for each component $a_j(t)$ respectively to extract the long-term trend $\mu_j(t) = p_{0j} + p_{1j}t$. Although the coefficients among different modes are uncorrelated when averaged over the full years from 1979 to 1998, there exists strong seasonal cross-mode correlations, which are modelled by the $\mathbf{C}(t)$ matrix. In each season s ($s = 1, 2, 3, 4$ corresponding to spring, summer, autumn, winter), the matrix $\mathbf{C}(t) = \mathbf{C}_s$ is assumed constant. Therefore \mathbf{C}_s can be computed by a Cholesky decomposition of the seasonal-averaged correlation matrix,

$$\boldsymbol{\Sigma}_s = \mathbb{E}_s \left[(\mathbf{a} - \boldsymbol{\mu})(\mathbf{a} - \boldsymbol{\mu})^\top \right], \quad \boldsymbol{\Sigma}_s = \mathbf{C}_s \mathbf{C}_s^\top. \quad (9)$$

The notation \mathbb{E}_s represents taking the expectation in season s . Finally, the remaining part of the time series in season s is,

$$\mathbf{w}_s(t) = \mathbf{C}_s^{-1} (\mathbf{a} - \boldsymbol{\mu}). \quad (10)$$

The spectrum of each component of $\mathbf{w}_s(t)$ is computed and stored, which will be used to generate surrogate stochastic processes $\hat{\mathbf{w}}_s(t)$ [17], assuming different components of $\hat{\mathbf{w}}_s(t)$ are independent. In summary, the emulator for the PCA time series (8) can be written as

$$\hat{\mathbf{a}}(t) = \boldsymbol{\mu}(t) + \mathbf{C}(t)\hat{\mathbf{w}}(t). \quad (11)$$

Given a realization of the stochastic process $\hat{\mathbf{w}}(t)$, we combine the corresponding $\hat{\mathbf{a}}(t)$ with PCA modes $\phi_j(\mathbf{x})$ to achieve an emulation of instantaneous TMX fields.

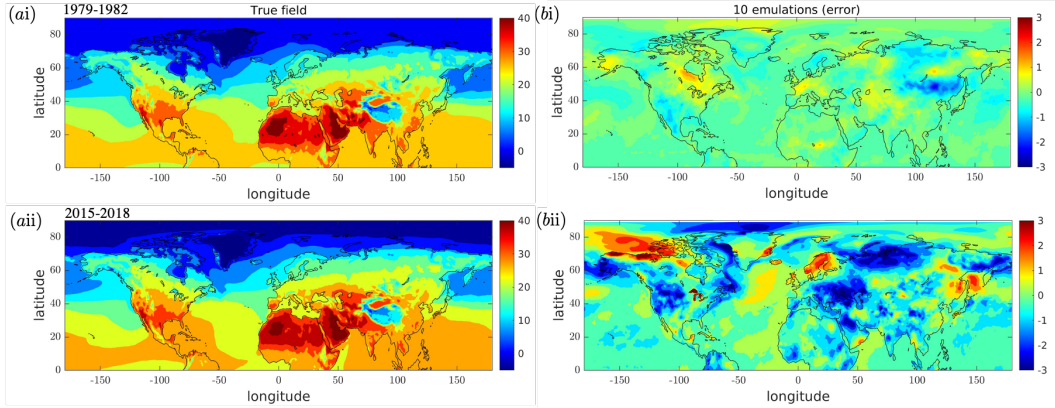


Figure 4: (a) Mean of local TMX distributions, computed using (i) 1979-1982 summer (June, July, Aug) and (ii) 2015-2018 summer data; (b) Error of emulated local mean TMX, computed by subtracting the true mean from emulations.

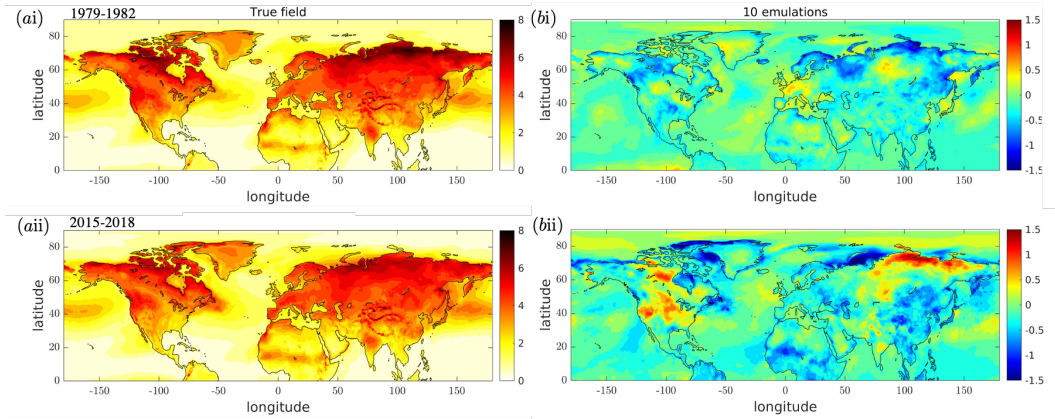


Figure 5: (a) Standard deviation of local TMX, computed using (i) 1979-1982 summer (June, July, Aug) and (ii) 2015-2018 summer data; (b) Error of emulated standard deviation, computed by subtracting the true std from emulations.

Appendix B Additional results

The mean and standard deviation of local TMX are plotted in figure 4,5 to evaluate the performance of our emulator. Note that the regions with highest errors vary across different four-year windows due to climate internal variability, which is more evident in quantile plots from 1979 to 2018 (figure 6,7).

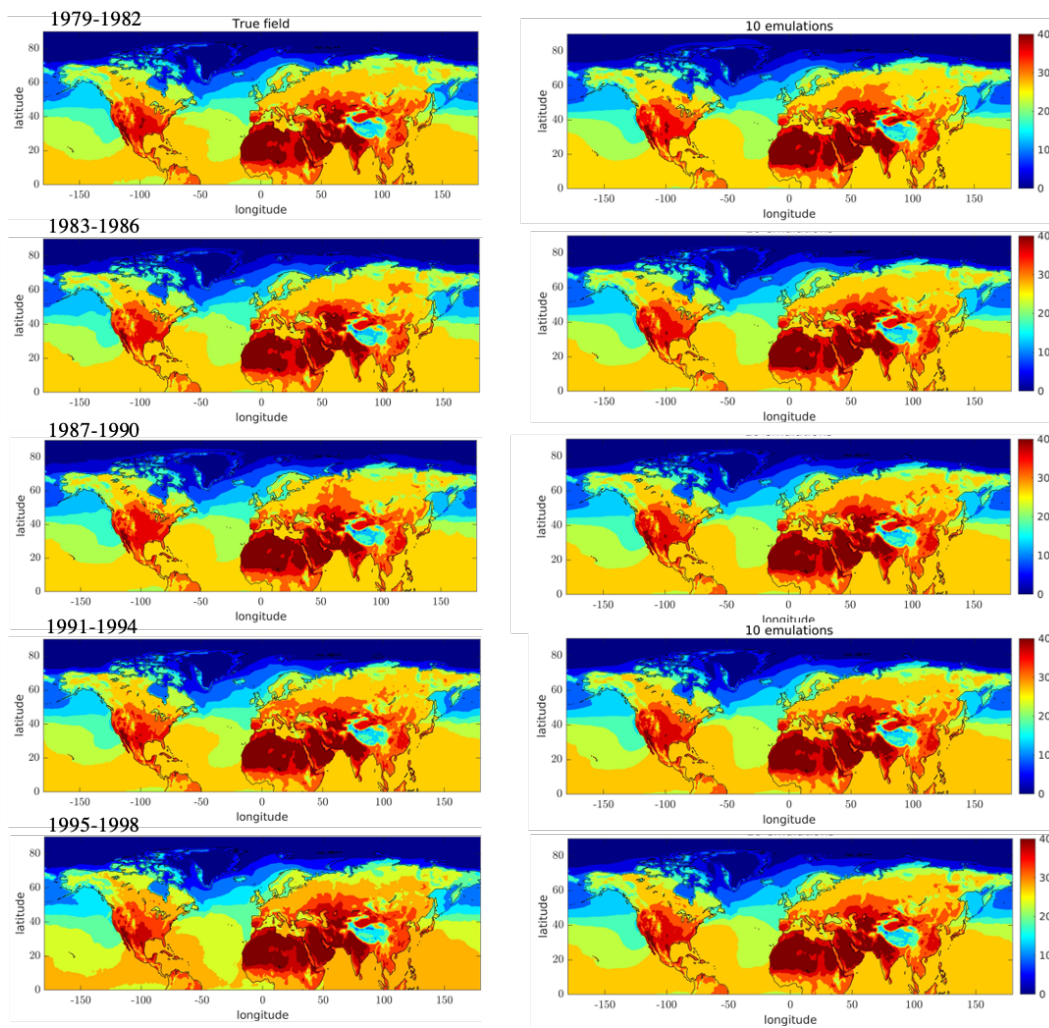


Figure 6: 97.5% quantiles of local TMX distributions, computed using summer data from different four-year windows. Left: true quantiles; right: quantiles computed from 10 emulations.

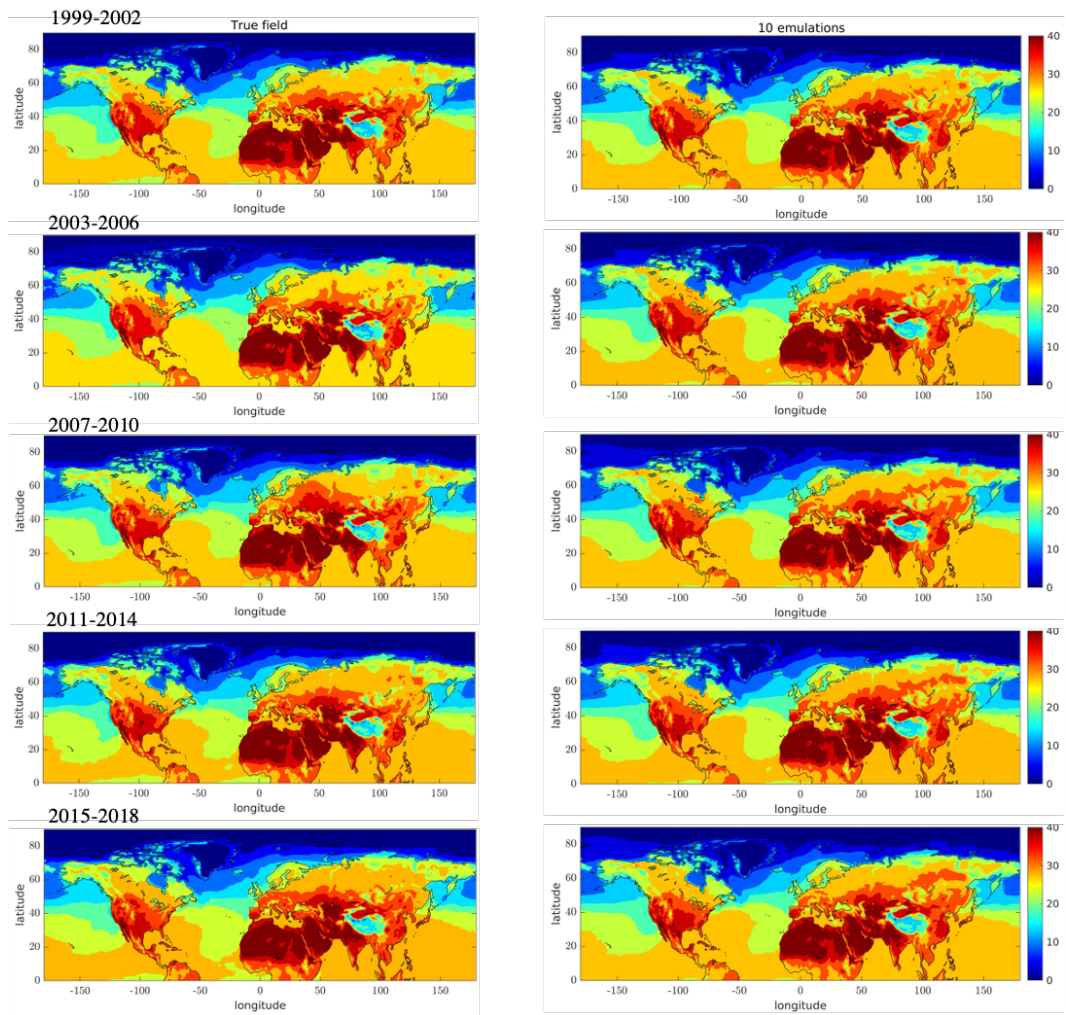


Figure 7: See caption of the previous figure. Here the results are shown from 1999 to 2018.