

GEN²: A Generative Prediction-Correction Framework for Long-time Emulations of Spatially-Resolved Climate Extremes

Mengze Wang^{1†}, Benedikt Barthel Sorensen^{1†},
Themistoklis P. Sapsis^{1*}

¹*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.

*Corresponding author(s). E-mail(s): sapsis@mit.edu;

†These authors contributed equally to this work.

Abstract

Accurately quantifying the increased risks of climate extremes requires generating large ensembles of climate realization across a wide range of emissions scenarios, which is computationally challenging for conventional Earth System Models. We propose GEN², a generative prediction-correction framework for an efficient and accurate forecast of the extreme event statistics. The prediction step is constructed as a conditional Gaussian emulator, followed by a non-Gaussian machine-learning (ML) correction step. The ML model is trained on pairs of the reference data and the emulated fields nudged towards the reference, to ensure the training is robust to chaos. We first validate the accuracy of our model on historical ERA5 data and then demonstrate the extrapolation capabilities on various future climate change scenarios. When trained on a single realization of one warming scenario, our model accurately predicts the statistics of extreme events in different scenarios, successfully extrapolating beyond the distribution of training data.

Keywords: Climate Science, Machine Learning, Dynamical Systems, Reduced Order Modeling

1 Introduction

It is widely expected that the current rapid rate of climate change will lead to an increase in the frequency of extreme weather events such as tropical storm, heatwaves, and droughts [1–3]. These events can have massive negative impacts on society lost lives and economic costs which have already ballooned from several million dollars in 1980 to 368 billion in 2024 [4–7]. Quantifying the increased risk of these extreme events as a function of various climate change scenarios is the first step in providing policymakers with the information needed to implement both plan for and mitigate their impacts on society. However, despite their increasing frequency, extreme weather events remain rare and thus it is impossible to accurately quantify their frequency or severity purely from observations. Accurate risk assessment requires generating sufficient data in the form of ensembles of long time horizon climate simulations. Furthermore, the trajectory of various driving factors, such as greenhouse gas emissions over the next several decades is far from clear, and depends on various social, political, economic, and technological factors we make no attempt to predict here. Therefore the amount of required data is further multiplied by the need to test a variety of emissions scenarios Earth may plausibly encounter. In summary, the challenge of quantifying extreme weather risks reduces to a need to generate large ensemble of climate realizations both across and within projected emissions scenarios.

The current state-of-the-art for generating such climate realizations relies on Earth System Models (ESMs) which solve the dynamical equations governing the earth’s atmosphere, oceans, and biosphere.[8–14]. Relying on ESMs as a primary predictive tool presents two challenges. First, there is a vast array of physical processes and interactions affecting the atmosphere that we do not yet fully understand, and must therefore be empirically parametrized in our numerical models [15–18]. Second, practical computational cost restricts global simulations to a spatial resolution of approximately 100 km. Such a coarse resolution not only precludes the quantification of weather events evolving on smaller length scales, but also leads to inaccuracy in the larger resolved scales. Many studies have sought to ameliorate these challenges through the introduction of data-driven forcing terms which are intended to parametrize the effects of any unknown physics and/or unresolved scales [19–24]. Despite such innovations, simulating the Earth System over a long time horizon on spatially-resolved grids pushes the frontier capabilities of modern high-performance computing. This has motivated the need for cheaper data-driven reduced-order models to augment or even replace numerical ESMs.

One approach driven by the recent advances in machine learning are auto-regressive weather models, which aim to predict the atmospheric state at a certain time as a function of the state at previous times [25–28]. The accuracy of these models generally compares favorably to that of numerical weather predictions, but at a fraction of the cost. However, numerical instabilities limit their predictions to a few days or potentially weeks. Even if the instability issue can be resolved [29], it is computationally expensive to run machine-learning weather models for hundreds of years to predict the climate. As we are interested in the statistics of events occurring on multi-decade or longer time scales, we do not pursue this approach here. An alternative approach is to construct reduced-complexity models for the climate system, or so-called climate

emulators. These models focused on quantifying the parametric relationships between various inputs to the climate system, such as greenhouse gas emissions, and the climate response. One of the most widely used approaches is known as Linear Pattern Scaling (LPS), where the local climate variables are assumed to be linear functions of the global mean temperature [30–32]. The general LPS approach has recently been expanded in a variety of ways including accounting for physical processes such as emission history [33, 34] and internal variability [35, 36] as well as the incorporation of more sophisticated metrics for modeling spatial correlation [37]. Furthermore, various deep learning based alternatives to LPS have been proposed [38, 39]. However, Lütjens et al. [40] recently compared the performance of various emulators on ClimateBench [41] and found the benefits of deep learning emulators are at best unclear. Additionally, emulation generally predicts time averaged quantities, and only a few recent studies have explored emulating extreme events such as the annual maximum temperature [42] and heat wave duration [42, 43].

Despite the success of LPS and its variants, the inherent nonlinearity and non-Gaussianity of the climate system places an upper bound on their potential to predict the full statistics of extremes, such as joint distributions of different variables. Correcting the biases of these emulators is known as *debiasing*. The most widely used strategy for debiasing coarse-resolution climate models is to augment numerical models with machine-learned parametric forcing terms, which aim to mimic the effects of the unresolved “sub-grid scale” dynamics [19–24]. However, like the fully auto-regressive models mentioned previously, these *intrusive* approaches suffer from instabilities when integrated over long (10+ year) time horizons [21, 44]. An alternative strategy is *non-intrusive* debiasing, which corrects the output of imperfect models in a post-processing manner – thereby bypassing the stability issue. The challenge with non-intrusive debiasing is that learning a map between two arbitrary chaotic trajectories is generally ill-posed, and any such map will not generalize to unseen data during training. Learning a generalizable map requires *paired* training data that are minimally affected by chaotic divergence. This is possible through the framework introduced by Barthel Sorensen et al. [45], which relies on training a correction operator on a surrogate model nudged towards a high fidelity reference. By formulating the supervised learning problem directly between paired trajectories, this strategy facilitates learning the dynamics with very little training data, which in turn enables the extrapolation of statistics when the learned map is applied to much longer trajectories [45, 46] and out-of-sample climate change scenarios [47]. However, these innovations still require expensive ESM simulations to generate the data needed for training and inference. In this work we aim to replace these expensive computations with parsimonious climate emulators.

Our approach, which we refer to as “GEN²” – as it consists of two generative steps applied in succession: (1) A Gaussian emulator that correctly captures the second-order spatio-temporal statistics of the climate, (2) A diffusion-model-based debiasing step trained using the nudging framework introduced by Barthel Sorensen et al. [45]. The emulation step is built on the stochastic model introduced by Wang et al. [48], which we have extended to emulate multiple variables, including wind speed, temperature, and humidity. The emulator is also refined to capture the spatio-temporal

spectra of the variable of interest. The debiasing step is achieved using a conditional diffusion model [49, 50] whose architectural backbone is based on U-Net introduced by Ronneberger et al. [51]. The choice of diffusion model allows for significantly improved debiasing capabilities as compared to the simpler auto-encoder based models used in previous studies, as has been recently demonstrated on tasks including debiasing [52] and down-scaling [53, 54]. Additionally, it is an inherently probabilistic model meaning that a single input can be used to generate a distribution of outputs. We first validate our approach on historical ERA5 data [55], and then demonstrate its extrapolation capabilities on various climate change scenarios for the coming century.

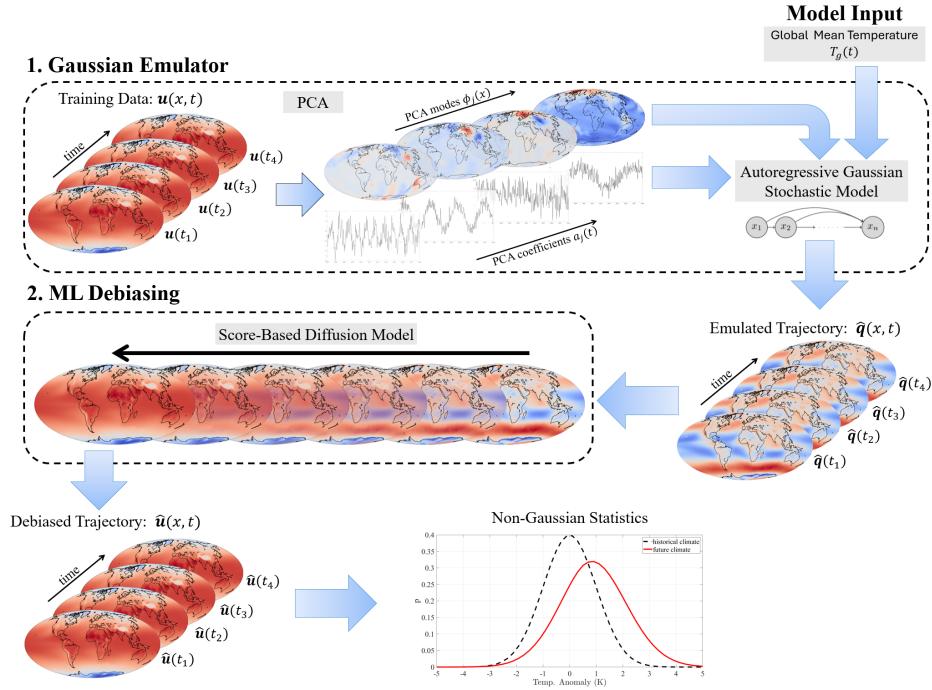


Fig. 1 Schematic diagram of the proposed two-step GEN² emulation framework consisting of an initial Gaussian emulator followed by a ML correction. The model takes as an input a time series of global mean temperature: $T_g(t)$ and outputs climate trajectory defined over the entire globe $\hat{u}(x, t)$. The Gaussian emulator assumes a spatial basis of PCA modes computed from the training data, and models the temporal coefficients as an autoregressive Gaussian process. The ML debiasing step consists of a score-based diffusion model trained to debias long term chaotic trajectories. *The specific data-images shown are chosen purely for illustrative purposes, and the bias of the emulator is exaggerated.

2 Results

Our goal is to accurately quantify the spatial and temporal statistics of low probability, high impact weather events over long time horizons given solely the global mean

temperature. As the latter is known to be proportional to cumulative CO_2 emissions, this allows us to directly quantify the spatial distribution of evolving extreme weather risk in a changing climate. We will use the term “climate” to refer to the statistics of the atmospheric state as quantified by the zonal and meridional wind speed: U, V , temperature: T , and specific humidity: Q . The GEN² framework takes a prescribed global mean temperature trajectory $T_g(t)$ as input—a scalar quantity—and outputs the full field trajectory of prognostic variables defined on a global grid, whose resolution is determined by the specific training data used. As illustrated in figure 1, the model consists of two components: an initial Gaussian emulation step and a diffusion-based ML debiasing step. The Gaussian emulation derives its spatial structure from the training data and assumes temporal dynamics are Gaussian processes conditioned on the global mean temperature. The subsequent debiasing step aims to reconstruct the strongly non-Gaussian tail statistics of extreme weather events. Both components are trained on the same reference dataset, although this is not necessary in practice. For brevity, the following discussion focuses on a subset of variables and statistics, with additional results provided in the Supplemental Information (see Supplementary Notes).

2.1 Historical Validation

We first validate our proposed modeling framework on historical ERA5 reanalysis data [55]. The training dataset includes U, V, T, Q from 1979 to 2018. The temporal sampling frequency is three hours, and the data are projected onto a $1.5^\circ \times 1.5^\circ$ resolution grid, i.e. approximately 100km. Once trained, the GEN² model is used to generate a 1979–2018 trajectory, and the climate of this predicted trajectory is compared against the actual ERA5 data. In figure 2, we show several metrics illustrating the ability of our model to capture the full richness of the ERA5 data. Unless otherwise stated, these metrics are computed using the fluctuation fields, defined as the deviation from the known climatological mean. All the metrics are temporally averaged from 1979 to 2018. More detailed definitions are provided in Supplementary Methods (section 3). Figure 2(a) compares the 40-year standard deviation, 97.5% quantile, skewness, and kurtosis of the zonal wind (U). In all cases, our model accurately predicts both the qualitative and quantitative structure of the statistics – although the skewness is slightly underestimated in the Pacific Ocean. To systematically compare the full statistics, especially the extreme events, in panel 2(b), we plot the probability density function (PDF) of U in log scale at four representative locations - Los Angeles, Boston, Athens, and Hong Kong. The prediction of the conditional Gaussian emulator, without ML correction, is also provided for reference. Interestingly, the conditional Gaussian emulator itself is already capable of capturing the distributions relatively well at locations where the distributions are weakly non-Gaussian. The ML debiasing step maintains or slightly improves the prediction at these locations. At Hong Kong, where the PDF is more strongly skewed, the ML debiasing step significantly corrects the tails of the distributions. More examples illustrating the debiasing power of the ML correction are included in Supplemental Information (Supplementary figure 3, 4 and table 1-4). Beyond these single-point statistics, we also evaluate the spatio-temporal coherence of the predicted fields, by plotting the Wheeler-Kiladis spectrum [56] of U ,

which quantifies the dispersion relationships of equatorial waves. As shown in figure 2(c), our model captures the characteristic frequency-wavenumber correlations corresponding to Kelvin waves observed in the data [56, 57] – a remarkable observation given that our model includes no physics to enforce these dispersion relations.

To further quantify the structure of the predicted fields, we compute the spatial two point correlation coefficients of temperature $\rho(T(\mathbf{x}_0), T(\mathbf{x}))$ and zonal wind $\rho(U(\mathbf{x}_0), U(\mathbf{x}))$, centered at each of the four previously analyzed locations. The results are shown in figure 3 (a,b). Moreoever, the cross-variable correlations at the same location, including zonal-meridonal wind correlation $\rho(U(\mathbf{x}), V(\mathbf{x}))$ as well as temperature-humidity correlation $\rho(T(\mathbf{x}), Q(\mathbf{x}))$, are plotted in figure 3 (c,d). Accurately capturing these correlations is crucial for accurately quantifying the risks of extreme weather events, which often occur due the concurrent incidence of extreme excursions in multiple climate variables, such as droughts being characterized by high temperatures and low humidity [1, 2, 58, 59]. Again, our model captures the structure of the underlying data exceptionally well, see for example the negative correlation between temperature and humidity in India and the southwestern United States – both places known to be susceptible to drought and extreme heat. We also capture the highly nontrivial wind patterns quantified by the correlation of U and V over both Europe and the United States.

2.2 Climate Change Scenarios

Having now demonstrated the capability and flexibility of our approach to reproduce the highly nontrivial structure of the historical climate, we now apply our method to forward looking climate change scenarios. Specifically, we consider the MPI-ESM1-2-LR model outputs of the Coupled Model Intercomparison Project Phase 6 (CMIP6) as our reference data. Such a choice is based on two considerations. First, the MPI model dataset has multiple ensemble members and climate change scenarios available. Second, this model has demonstrated adequate skill in the quantification of climate extremes in a recent benchmarking study of CMIP6 models [60]. We focus on four climate change scenarios, SSP126, SSP245, SSP370, and SSP585, each corresponding to a different level of global emissions, and accordingly a different trajectory of global mean temperature.

The crucial test of any data-driven model is its ability to extrapolate beyond the distribution of the data seen in training. We demonstrate this capability in two ways, first we will show the ability to extrapolate statistics *within* a single climate change scenario, and second the ability to extrapolate to unseen scenarios. We train on 1 realization of the most extreme emission scenario, SSP585, and evaluate our model on 10 realizations of the climate under all 4 warming scenarios – both the SSP585 scenario seen in training and the three other unseen scenarios.

We first demonstrate in figure 4 the ability of our model to extrapolate within scenario. Here we show the probability densities of temperature fluctuations in Hong Kong, Los Angeles, Boston, and Athens in 2090-2099 under the SSP585 warming scenario. The PDF's are computed by Monte Carlo sampling and smoothed by a moving average filter to improve readability. Although only one member is used for training (red circles), the generated 10 realizations (blue triangles) successfully capture the tail

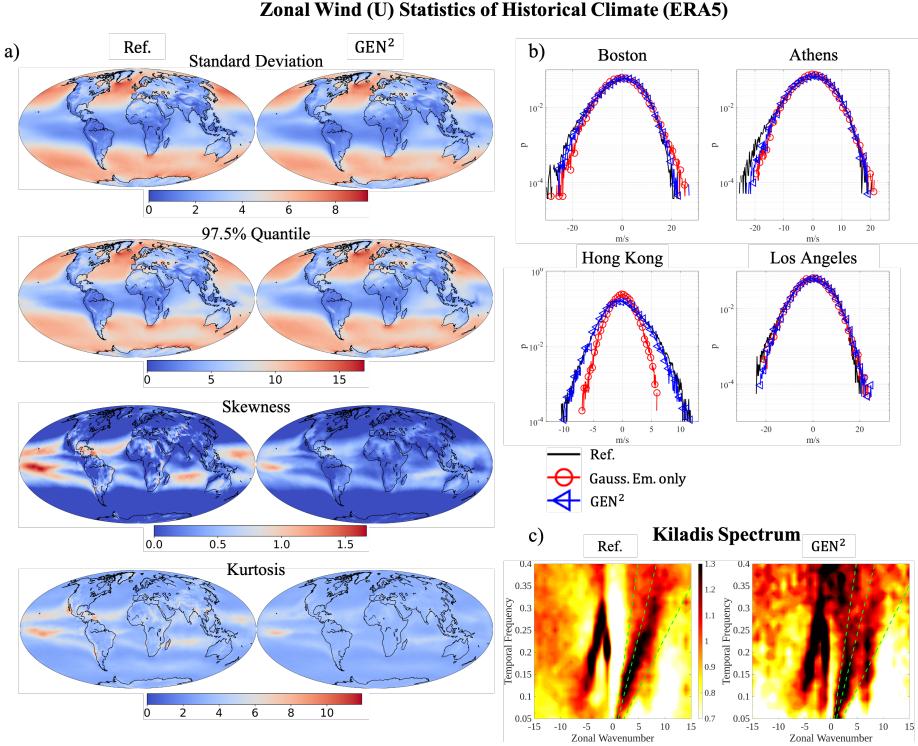


Fig. 2 Zonal wind statistics of historical climate, evaluated using 1979–2018 reference ERA5 data and GEN^2 prediction. (a) Local standard deviation, 97.5% quantile, skewness and kurtosis. (b) Log probability density functions at areas including Boston, Athens, Hong Kong, and Los Angeles. Black: reference; Red: Gaussian emulator only; Blue: GEN^2 . (c) Wheeler-Kiladis space time spectra. Green dashed lines: Kelvin waves with depth $H = 12m, 50m, 150m$ (Phase speed of Kelvin waves is \sqrt{gH}).

statistics of the true 10 members (black squares). If the underlying system is ergodic, this type of extrapolation from 1 to multiple realizations is equivalent to training the model on a short time window and testing on a longer time window, as demonstrated in Barthel Sorensen et al. [45, 46].

We next demonstrate the ability of our approach to extrapolate beyond the scenario seen in training, as shown in figure 5. Figure 5(a) illustrates the evolution of global mean temperature corresponding to the four emissions scenarios studied here. Each curve shows the ensemble average over 10 members. Figure 5(b) compares the 97.5% quantile of temperature predicted by our model to the reference data at the end of the century (2090–2099) for three different scenarios SSP125, 245, 585. Since these are the quantiles of the climatological-mean-subtracted fields, the peak values are observed at the poles (as opposed to the equator), indicating that the impacts of extreme temperature fluctuations will be most pronounced in the polar regions under strong global warming. Figure 5(c) shows the root-mean square error (RMSE) in the predicted quantiles for each scenario as a function of time – that is to say the RMSE of the fields shown in figure 5(b) computed for each decade. Our model is able to successfully and

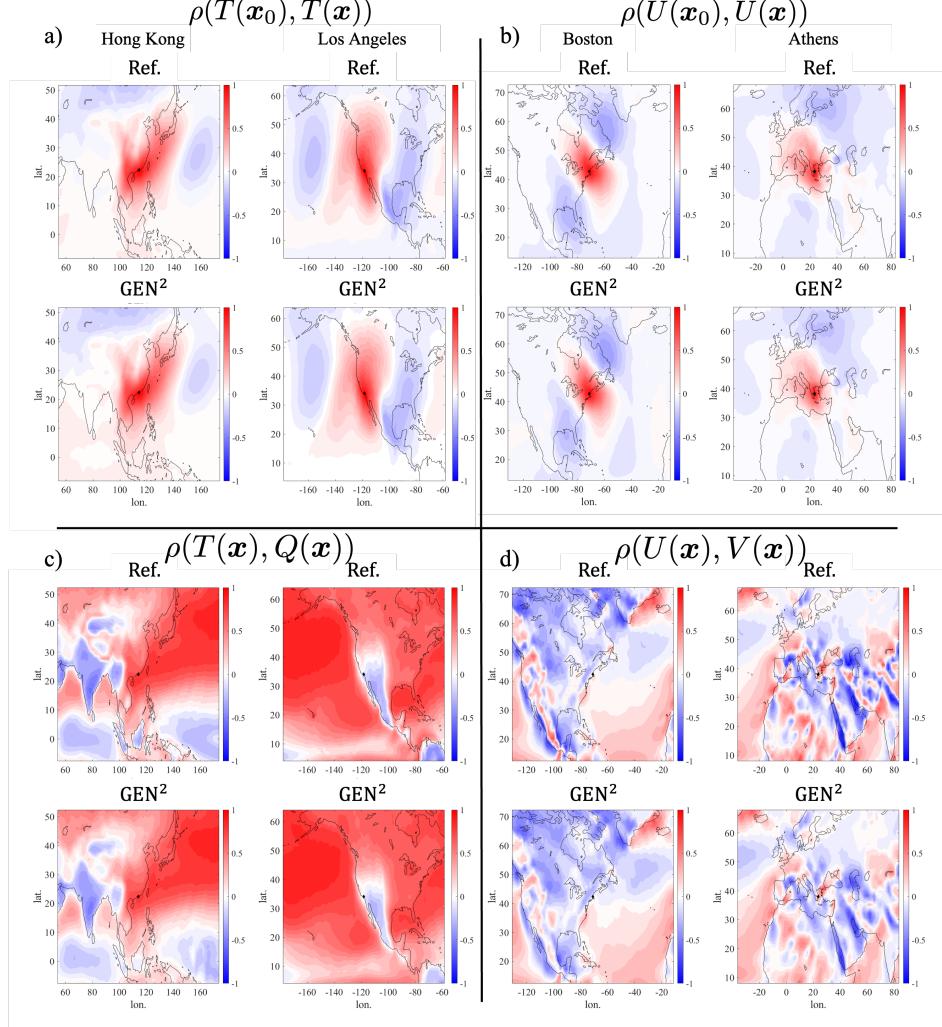


Fig. 3 Regional correlation coefficients. (a,b) Two point correlations of temperature (T) and zonal wind (U), centered at four different cities. (c,d) Local cross-variable correlations between (c) temperature and humidity, and (d) zonal and meridional wind speed. For each sub-figure reference ERA5 data is shown in the top row and the GEN² prediction in the bottom row. All results are evaluated over the 40 year period 1987-2018 and at surface elevation.

consistently predict the quantiles in unseen scenarios, achieving comparably low error ($< 0.5K$) across all scenario despite the fact that only data from one of the scenarios was seen in training. This is a critical ability in any climate emulator, as it means that new scenarios of interest can be reliably investigated without additional numerical simulations or retraining of exiting data-driven models.

To further highlight the ability of our model to replicate warming scenarios not seen in training, we zoom in to the region centered around Boston under the SSP126 scenario, which is the most dissimilar from the SSP585 data seen in training. Figure 6

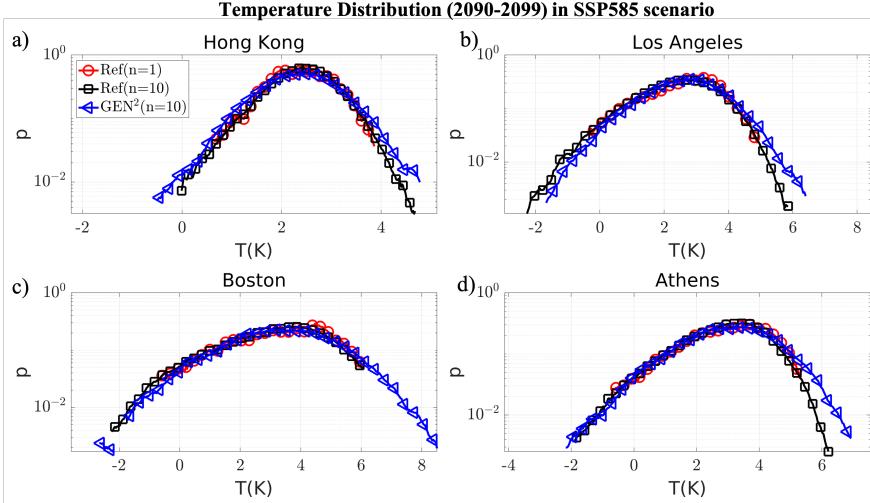


Fig. 4 Demonstration of extrapolation within warming scenario. Probability densities of temperature in (a) Hong Kong , (b) Los Angeles, (c) Boston, and (d) Athens in 2090-2099 under SSP585 warming scenario. The ensemble statistics of 10 realizations of the reference data and GEN² prediction are shown in black squares and blue triangles respectively. The density of the single realization of reference data used to train the GEN² is shown in red circles. Results labeled “Ref.” represent reference simulation data, “GEN²” represent our model predictions.

shows the two point correlations (panel a), PDFs of U and T (panel b), and the joint PDF of T and Q (panel c). All these statistics are computed using the fluctuation fields and averaged from 2090 to 2099. In all cases our model prediction captures the highly non-Gaussian and non-isotropic structure observed in the reference data. Our model again manages to extrapolate the tails in the local PDFs (panel b). The shape of joint PDFs in (c) indicate that fluctuations in humidity are positively correlated with fluctuations in temperature, representing the relative prevalence of dry cold snaps and humid heatwaves – patterns not unfamiliar to residents of New England and accurately predicted by GEN² approach.

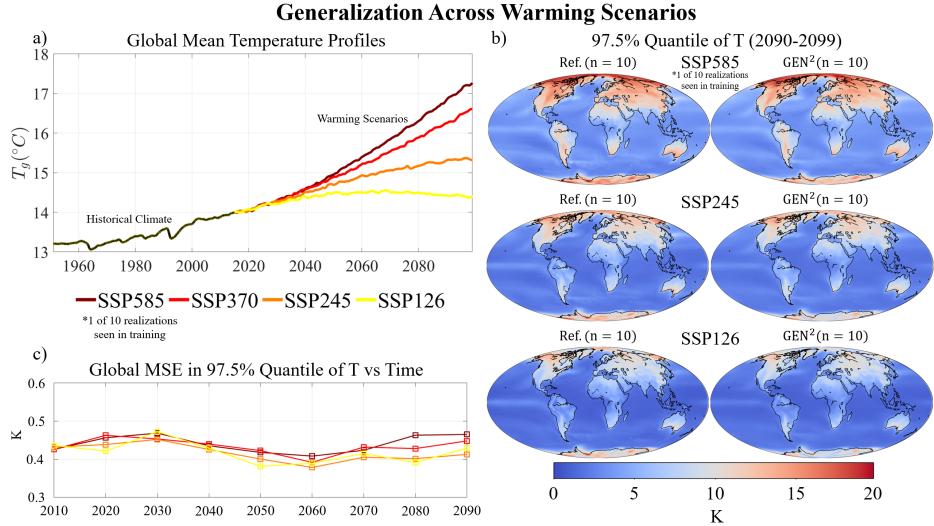


Fig. 5 Demonstration of generalizability across climate change scenarios. (a) Illustration of Global mean temperature profile (which serve as inputs to our model) corresponding to various climate change scenarios - figure shows ensemble average over 10 realizations. Only one of the realization of the SSP585 scenario is used for training. (b) Global field of 97.5% quantiles of 10 realizations of temperature fluctuations for 3 different warming scenarios for the years 2090-2099. Results labeled “Ref.” represent reference simulation data, “GEN²” represent our model predictions. (c) Global root-mean-square error of the same temperature fluctuation quantiles over time. Each data point represents the RMSE of the two quantile fields in (b) computed for each decade.

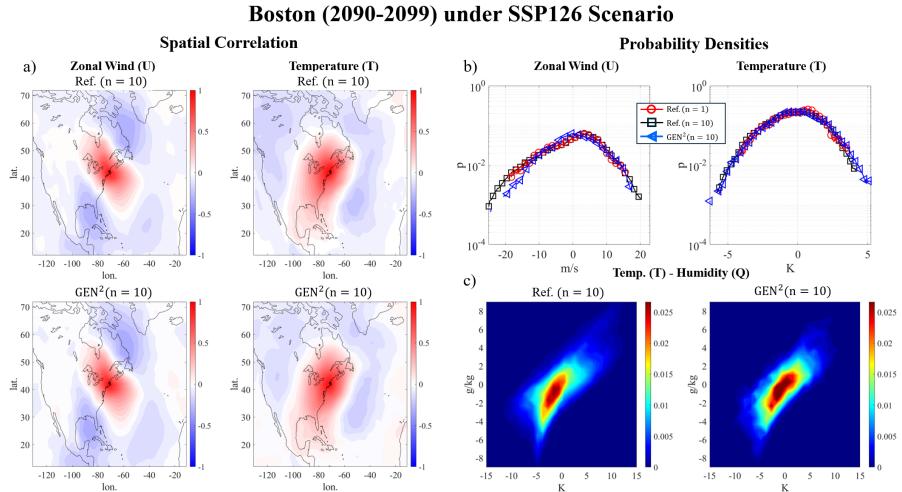


Fig. 6 Demonstration of regional climate statistics. (a) Spatial correlation coefficient, (b) probability density and (c) joint probability density of zonal wind (U) and temperature (T) over Boston. Results labeled “Ref.” represent reference simulation data, “GEN²” represent our model predictions. All statistics are computed from 10 ensemble members for the years 2090-2099.

3 Discussion

The prediction-correction framework presented here provides a more accessible and computationally affordable alternative to Earth System Models for generating large ensembles of climate realizations. Once trained, our GEN² model generates predictions at a cost of approximately 720 simulated years-per-day (YPD), whereas Earth System Models achieve approximately 40 YPD [14]. Compared to traditional data-driven emulators such as pattern scaling, our model is able to capture non-Gaussian statistics as well as significantly improve the estimation of higher-order moments, spatial correlations, and highly non-trivial spatio-temporal features such as the Kiladis spectrum. Compared with other ML-based emulators that may offer sufficient flexibility to capture higher-order features, our approach achieves a lower computational cost of both training and inference, by obtaining a first-pass prediction using the conditional Gaussian emulator and limiting the more costly ML step to a debiasing operation. Another benefit of the initial Gaussian emulation is the ability to incorporate different physics by customizing which statistics are regressed on. For example, while we chose to enforce temporal coherence over several days, with sufficient data one could also choose to enforce inter-annual or season-to-season correlations. This would generally be much more difficult with purely ML based emulators.

The GEN² framework can be generalized along different pathways to further improve its accuracy, and we outline two possibilities below. First, our debiasing operator is based on an image diffusion model, which operates on each snapshot in time independently and thus can not correct biases in temporal dynamics. Debiasing in time would require the use of a video diffusion model, but this approach would significantly increase the training and inference cost. Second, quantifying very localized weather statistics requires additional localized super-resolution (or downscaling) of the predicted fields – a feature not incorporated in our framework. However, many existing techniques [52] could be directly applied to the output of our model, and this remains a topic of ongoing work. In summary, the GEN² framework is capable of reconciling the computational cost constraints with the need for accurate climate extreme emulation. This framework serves as a guide for future developments of climate emulators, enabling policymakers and stakeholders to explore the parameter space of emission trajectories and interventions.

4 Methods

Consider the state variable of the climate system as a function of space and time, $\mathbf{u}(\mathbf{x}, t)$. Its underlying dynamical system is generally chaotic and high-dimensional, which makes it intractable to capture the full dynamics using deterministic data-driven models. Faced with this challenge, we seek a stochastic model that approximates the dynamics as a function of the emission scenario, quantified by the global mean temperature T_g . Specifically, we seek to parameterize the following conditional probability distribution

$$p(\mathbf{u}(\mathbf{x}, t)|T_g(t)) \sim \mathcal{G}_\theta[T_g(t)] \quad (1)$$

through a prediction-correction process

$$\hat{\mathbf{q}}(\mathbf{x}, t) \sim \mathcal{G}_{\theta,1} [T_g(t)] \quad (2)$$

$$\hat{\mathbf{u}}(\mathbf{x}, t) \sim \mathcal{G}_{\theta,2} [\hat{\mathbf{q}}(\mathbf{x}, t)], \quad (3)$$

where $\mathcal{G}_{\theta,1}$ is a conditional Gaussian emulator, linearly driven by T_g , and $\mathcal{G}_{\theta,2}$ is a general nonlinear non-Gaussian stochastic model, parametrized by θ . In general we wish the model (1) to fulfill three main aims. (1) Forward evaluation of the model must be cheap so that a large number of ensembles can be rapidly generated. (2) The model must be stable over arbitrarily long time horizons. (3) The model must be capable of extrapolating beyond the distribution of the data seen in training. A diagram describing the full model is shown in figure 1.

4.1 Step 1: Conditional Gaussian Emulation

The conditional Gaussian emulator is built on the framework introduced by Wang et al. [48]. We extend this approach to emulate multiple variables and capture the correct spatio-temporal correlations. The emulated state, $\hat{\mathbf{q}}(\mathbf{x}, t)$, is constructed as the superposition of the climatological mean $\bar{\mathbf{u}}$ and the PCA modes,

$$\hat{q}_k(\mathbf{x}, t) = \bar{u}_k(\mathbf{x}, t) + \sigma_{g,k} \sum_{i=1}^I \hat{a}_i(t) \phi_k^{(i)}(\mathbf{x}). \quad (4)$$

where the subscript $k = 1, 2, 3, 4$ corresponds to U, V, T, Q components, $\sigma_{g,k}$ is the globally-averaged standard deviation, and $\phi_k^{(i)}$ is the i^{th} PCA mode. The quantities $\bar{\mathbf{u}}$, $\sigma_{g,k}$, and $\phi_k^{(i)}$ are all assumed to be known.

The time series of PCA coefficients are modelled as Gaussian processes conditioned on T_g , which characterizes the climate change. Specifically, the modelled time series,

$$\hat{a}_i(t) = \hat{\mu}_i(T_g) + \hat{\sigma}_i(T_g) \hat{\eta}_i(t), \quad (5)$$

consist of the seasonal mean $\hat{\mu}_i(T_g)$ and daily fluctuations $\hat{\eta}_i(t)$ scaled by the seasonal standard deviation $\hat{\sigma}_i(T_g)$. Here $\hat{\mu}_i$ and $\hat{\sigma}_i$ are assumed piecewise constant in each season and varying linearly with the seasonal average of the global mean temperature T_g (as demonstrated in Supplementary Figure 1). The parameters of the linear models are estimated by performing least square regression on the data in each season respectively. The daily fluctuations $\eta_i(t)$ are modelled as zero-mean multivariate Gaussian processes, whose covariance matrices are assumed constant in each season and estimated from data. Once trained, this emulator takes as input a scalar valued time series of global mean temperature $T_g(t)$ and outputs a prediction of the full spatio-temporal evolution $\hat{\mathbf{q}}(\mathbf{x}, t)$. A more detailed mathematical description of the model is included in the Supplementary Methods.

In order to accurately represent the global climate system while maintaining computational efficiency, we only keep the first 500 PCA modes that account for approximately 80% of the total variance of the training data. As such, the bias of the

emulator arises from two sources: the non-Gaussian statistics of the first 500 modes and the ignored higher-order modes. The latter generally correspond to the small-scale and more extreme events. These biases will be corrected through the deep learning model in the next section.

4.2 Step 2: Nudged Emulation and Non-Gaussian Correction

Due to the chaotic nature of the Earth system and herein the stochasticity of the emulator, the emulated state $\hat{\mathbf{q}}(\mathbf{x}, t)$ will significantly deviate from contemporaneous reference $\mathbf{u}(\mathbf{x}, t)$, making it fundamentally difficult to learn a map $\mathcal{G}_{\theta,2}$ from $\hat{\mathbf{q}}(\mathbf{x}, t)$ to $\mathbf{u}(\mathbf{x}, t)$. To address this challenge, we construct a *nudged* trajectory $\hat{\mathbf{a}}^\nu$ that stays close to the reference trajectory \mathbf{u} while approximately satisfying the equations of the conditional Gaussian emulator. If we re-formulate the emulator (5) as a dynamical system,

$$\frac{d\hat{\mathbf{a}}}{dt} = f(\hat{\mathbf{a}}, t), \quad (6)$$

the nudged PCA time series are defined as,

$$\frac{d\hat{\mathbf{a}}^\nu}{dt} = f(\hat{\mathbf{a}}^\nu, t) - \frac{1}{\tau} (\hat{\mathbf{a}}^\nu - \mathbf{a}). \quad (7)$$

The vector \mathbf{a} denotes the first 500 modes that are included into the emulator. Compared with equation (6), the nudged emulator (7) features an additional feedback term that forces $\hat{\mathbf{a}}^\nu$ towards the true PCA coefficients \mathbf{a} . From a physical perspective, the nudging term in (7) enforces the slow dynamics of $\hat{\mathbf{a}}^\nu$ to follow the reference, while allowing fast and more extreme dynamics to freely evolve [45]. The nudging timescale τ is a user-defined parameter. Generally, τ should be selected to ensure that the nudging term is an order of magnitude than the other terms in the governing equations. In this study, τ is set as six hours for both ERA5 and CMIP6 data, which is also consistent with previous studies [45]. Combining $\hat{\mathbf{a}}^\nu$ with the PCA mode shapes gives us the nudged spatial-temporal fields $\hat{\mathbf{q}}^\nu(\mathbf{x}, t)$, which will be utilized together with the reference data $\mathbf{u}(\mathbf{x}, t)$ to learn $\mathcal{G}_{\theta,2}$.

The debiasing operator $\mathcal{G}_{\theta,2}$ is parameterized as a conditional score-based diffusion model [49, 50]. This probabilistic approach accounts for the potentially non-unique relationship between $\hat{\mathbf{q}}^\nu$ and \mathbf{u} . Once the diffusion model is trained to learn $p(\mathbf{u}|\hat{\mathbf{q}}^\nu)$, it can take any free-running emulation $\hat{\mathbf{q}}(\mathbf{x}, t)$ as an input or conditional information to produce the debiased fields $\hat{\mathbf{u}}(\mathbf{x}, t)$. For example, in CMIP6 data, $\hat{\mathbf{q}}(\mathbf{x}, t)$ could come from a realization of the climate change scenario that is unseen during training, in order to evaluate the capability of our framework to extrapolate beyond the training scenarios. More details about the conditional diffusion model are provided in the Supplemental Methods.

Supplementary information. This article has accompanying supplemental materials.

Acknowledgments. This research was partially supported by the Vannevar Bush grant N000142512059, as well as the project Bringing Computation to the Climate

Challenge (BC3), supported by Schmidt Sciences through the MIT Climate Grand Challenges.

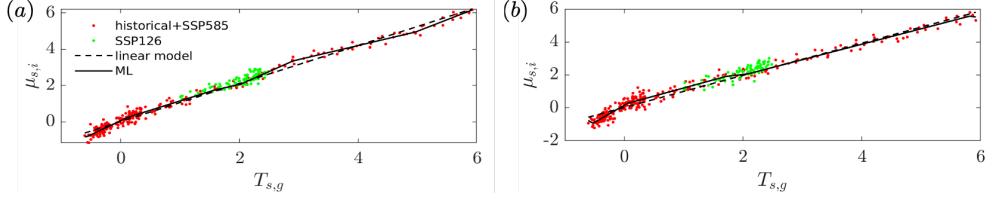


Fig. A1 Jun-Aug mean of (a) the first and (b) second PCA coefficients in each year of CNRM-CM6-1-HR dataset, from 1850 to 2100, plotted versus the global mean temperature. Red dots: true seasonal mean obtained from the historical and SSP5-8.5 scenario. Green dots: SSP1-2.6 scenario. Black dashed line: linear regression; Solid line: machine-learned function.

Appendix A Stochastic Emulator

Here we describe in detail the first part of our climate modeling framework, the linear stochastic emulation. In summary, the emulator takes as input a time series of the global mean temperature $T_g(t)$ and outputs a time series of the local state of the climate $\mathbf{u}(\mathbf{x}, t) = [u_1, u_2, u_3, u_4]^\top = [U(\mathbf{x}, t), V(\mathbf{x}, t), T(\mathbf{x}, t), Q(\mathbf{x}, t)]^\top$ where U, V, T, Q are the zonal and meridional wind speeds, temperature and humidity respectively and the spatial dimensions $\mathbf{x} = (\theta, \varphi)$ are the longitude and latitude, $\theta \in [-\pi/2, \pi, 2]$ and $\varphi \in [0, 2\pi]$. The time step size of t is three hours for ERA5 dataset and one day for the CMIP6 MPI model. Consistent with the formulation of modern climate models – and to reduce the data to a manageable size – our model operates at a fixed altitude, and thus the spatial dimension is 2D. We focus here exclusively on the near-surface climate, but our model could be directly applied to any altitude.

Stated succinctly, our approach consider a principal component analysis (PCA) of the climate data $\mathbf{u}(\mathbf{x}, t) = \sum_j a_j(t) \phi_j(\mathbf{x})$ and attempts to model the temporal coefficients $a_j(t)$ for a given spatial basis $\phi_j(\mathbf{x})$. Our emulator is therefore built on three fundamental assumptions:

1. The PCA basis $\phi_j(\mathbf{x})$ computed from the climate during a sufficiently long time period (e.g. historical and SSP5-8.5 scenario) remains an efficient basis for describing other future climate change scenarios;
2. The seasonal mean and variance of the coefficients $a_j(t)$ vary linearly with global mean temperature.
3. The statistics of daily fluctuations, given the season, are independent of the year and the climate change scenarios.

The construction of the emulator can be divided into two distinct steps: dimensionality reduction and stochastic modeling of PCA time series. The emulator is then nudged towards the observation data to facilitate machine-learning-based debiasing. We now describe each of these steps in detail.

A.1 Dimensionality Reduction

First we describe how the spatial PCA basis $\phi_j(\mathbf{x})$, which provides the structure for our emulator, is computed. Given a dataset consisting of N years, we extract the

climatological mean $\bar{\mathbf{u}}(\mathbf{x}, t)$, defined as phase-average of \mathbf{u} on the same calendar day (e.g. Jan 1st),

$$\bar{\mathbf{u}}(\mathbf{x}, t) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{u}(\mathbf{x}, t + nT), \quad 1 \leq t \leq T. \quad (\text{A1})$$

, where the period T is one year. When the emulator is trained on the daily maximum data from the MPI model, the climatological mean (A1) only quantifies the seasonal cycle. For the three-hourly ERA5 data, $\bar{\mathbf{u}}$ accounts for not only the seasonal variation but also the diurnal cycles. To obtain the scaling of each state variable, we compute the global-and-time-averaged standard deviation,

$$\sigma_{g,k} = \left[\frac{1}{\mathcal{T}S} \int_0^{\mathcal{T}} \int_S (u_k(\mathbf{x}, t) - \bar{u}_k(\mathbf{x}, t))^2 \cos \theta d\theta d\varphi dt \right]^{1/2}. \quad (\text{A2})$$

The notations \mathcal{T} and S are the duration of training window and the Earth's surface, respectively. The data are then centered to have zero climatological mean and scaled by the global standard deviation,

$$u'_k(\mathbf{x}, t) = (u_k(\mathbf{x}, t) - \bar{u}_k(\mathbf{x}, t)) / \sigma_{g,k}. \quad (\text{A3})$$

Now that each component of q'_k has the same order of magnitude, we construct its spatial covariance function,

$$\mathcal{R}_{jk}(\mathbf{x}, \mathbf{x}^*) = \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} u'_j(\mathbf{x}, t) u'_k(\mathbf{x}^*, t) dt, \quad j, k = 1, 2, 3, 4. \quad (\text{A4})$$

The PCA modes are acquired by solving the eigenvalue problem,

$$\int_S \sum_k \mathcal{R}_{jk}(\mathbf{x}, \mathbf{x}^*) \phi_k(\mathbf{x}^*) \cos \theta d\theta d\varphi = \lambda \phi_j(\mathbf{x}), \quad j = 1, 2, 3, 4, \quad (\text{A5})$$

This set of equations has multiple solutions $(\lambda^{(i)}, \phi^{(i)})$, $i = 1, 2, 3, \dots$, which are the PCA eigenvalues and mode shapes, respectively. Without loss of generality we rank the eigenpairs such that the eigenvalues, which represent variance, satisfy $\lambda_1 > \lambda_2 > \dots > \lambda_I$. The temporal PCA coefficients which govern the time dependence of the spatial PCA modes are found by projecting the normalized fluctuation field onto $\phi^{(i)}$,

$$a_i(t) = \int_S \sum_k u'_k(\mathbf{x}, t) \phi_k^{(i)}(\mathbf{x}) \cos \theta d\theta d\varphi. \quad (\text{A6})$$

The state of the climate can then be expressed as superposition of PCA modes,

$$u_k(\mathbf{x}, t) = \bar{u}_k(\mathbf{x}, t) + \sigma_{g,k} \sum_{i=1}^I a_i(t) \phi_k^{(i)}(\mathbf{x}). \quad (\text{A7})$$

When the number of PCA modes I is equal to the number of grid points or the number of snapshots, whichever is smaller, we recover the full field, and any smaller value of I represents a truncation. In this work, we always retain 500 PCA modes, which represent 79.6% of the total variance of 1979-2018 ERA5 data and 78.2% of 1950-2100 MPI data. To reiterate, we assume that the climatological mean $\bar{u}_k(\mathbf{x}, t)$, global standard deviation $\sigma_{g,k}$, and PCA mode shapes $\phi_k^{(i)}$ are unchanged with time or future scenarios. As a result, we focus purely on modeling $a_i(t)$, and the emulated state is written as,

$$\hat{q}_k(\mathbf{x}, t) = \bar{u}_k(\mathbf{x}, t) + \sigma_{g,k} \sum_{i=1}^{500} \hat{a}_i(t) \phi_k^{(i)}(\mathbf{x}). \quad (\text{A8})$$

Here the notations with $\hat{\cdot}$ are emulated quantities.

A.2 Stochastic emulator of PCA time series

A.2.1 Seasonal Decomposition

Our goal is to construct a time series of $\hat{a}_i(t)$ that *statistically* resembles the reference data $a_i(t)$. Although we have removed the climatological mean, the statistics of $a_i(t)$ still exhibit seasonal variation that is important to take into account. Therefore, we divide $a_i(t)$ into four seasons $a_{s,i}(t)$ – of approximately equal length – and model them separately where the additional subscript $s = 1, 2, 3, 4$ represents winter (Dec-Feb), spring (Mar-May), summer (Jun-Aug) and autumn (Sep-Nov). The number of days in each season is 90, 92, 92, and 91 respectively.

A.2.2 Formulation and Estimation of Model Parameters

We postulate a decomposition of the time series of PCA coefficients,

$$\hat{a}_{s,i}(t) = \hat{\mu}_{s,i}(T_{s,g}) + \hat{\sigma}_{s,i}(T_{s,g}) \hat{\eta}_{s,i}(t), \quad (\text{A9})$$

which is a superposition of the seasonal mean $\hat{\mu}_{s,i}$ and fluctuations parameterized through an envelope of the seasonal variance $\hat{\sigma}_{s,i}^2$. The seasonal mean and variance are assumed to be functions of the global mean temperature $T_{s,g}$, defined as the seasonal average of the daily T_g . The time-dependent daily fluctuations in each season are modelled as autoregressive Gaussian processes $\hat{\eta}_{s,i}(t)$. We will now discuss the formulation and computation of each of these terms in detail.

Linear Regression of Seasonal Mean and Variance. For each season s and each mode i , in the n th year, we compute the $T_{s,g}$ as well as the seasonal mean $\mu_{s,i}$ and variance $\sigma_{s,i}^2$ of the PCA coefficients of the reference data $a_{s,i}(t)$. Note that for each s, i , and n , the mean $\mu_{s,i}$ and variance $\sigma_{s,i}^2$ are constants – we generally omit explicit notation of the year n to avoid notational clutter. Grouping these values by season s and mode i allows us to perform a linear regression using $\{\mu_{s,i}(n), T_{s,g}(n)\}$ and

$$\{\sigma_{s,i}^2(n), T_{s,g}(n)\}$$

$$\begin{aligned}\hat{\mu}_{s,i}(T_{s,g}) &= \hat{p}_{s,i,0} + \hat{p}_{s,i,1}T_{s,g} \\ \hat{\sigma}_{s,i}^2(T_{s,g}) &= \hat{q}_{s,i,0} + \hat{q}_{s,i,1}T_{s,g},\end{aligned}\tag{A10}$$

an assumption which is justified by the linear trends which have been observed in data by a number of sources [31, 32] and illustrated in figure A1.

Time Lagged Cross-Mode Covariance. After extracting the linear trends of the seasonal mean and standard deviation in response to the global mean temperature, we remove these trends from the true PCA coefficients, resulting in the residuals $\eta_{s,i} = (a_{s,i} - \hat{\mu}_{s,i}) / \hat{\sigma}_{s,i}$. To accurately capture the spatio-temporal dynamics, our model must reflect not only the contemporaneous correlations between different modes, e.g. $\eta_{s,1}(t)$ and $\eta_{s,2}(t)$, but also their correlations across time. To this end, we define the *time-lagged cross-mode covariance*,

$$\boldsymbol{\Sigma}_s(m) = \frac{1}{T_s} \int_{T_s} \boldsymbol{\eta}_s(t) \boldsymbol{\eta}_s^\top(t + m\Delta t) dt, \quad m = 0, 1, \dots, M,\tag{A11}$$

where $\boldsymbol{\eta}_s = [\eta_{s,1}, \eta_{s,2}, \dots, \eta_{s,M}]^\top$ is the vector of fluctuations of each PCA mode, $M\Delta t$ is maximum time lag considered, and T_s represents the set of time indices corresponding to season s across all training years.

Now we want to model the observed fluctuations $\eta_{s,i}(t)$ as a multivariate Gaussian process $\hat{\eta}_{s,i}(t)$, which has the same covariance matrix $\boldsymbol{\Sigma}_s(m)$ as $\eta_{s,i}(t)$. To further simplify our notation, the subscript s will be omitted. Mathematically, we seek to construct an autoregressive model of order M ,

$$\hat{\boldsymbol{\eta}}(t) = \hat{\boldsymbol{\Psi}}_1 \hat{\boldsymbol{\eta}}(t - \Delta t) + \hat{\boldsymbol{\Psi}}_2 \hat{\boldsymbol{\eta}}(t - 2\Delta t) + \dots + \hat{\boldsymbol{\Psi}}_M \hat{\boldsymbol{\eta}}(t - M\Delta t) + \boldsymbol{\epsilon}(t)\tag{A12}$$

where the noise term is a multivariate Gaussian random vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$. The unknown matrices $\hat{\boldsymbol{\Psi}}_1, \hat{\boldsymbol{\Psi}}_2, \dots, \hat{\boldsymbol{\Psi}}_M, \hat{\mathbf{R}}$ are solved such that the simulated process (A12) satisfy the given covariance matrices with different time lags $\boldsymbol{\Sigma}(0), \boldsymbol{\Sigma}(1), \dots, \boldsymbol{\Sigma}(M\Delta t)$. By multiplying both sides of (A12) by $\hat{\boldsymbol{\eta}}(t - i\Delta t)$ and averaging in time, we can derive a set of equations, the so-called Yule-Walker equations,

$$\begin{bmatrix} \boldsymbol{\Sigma}(0) & \boldsymbol{\Sigma}^\top(1) & \cdots & \boldsymbol{\Sigma}^\top(M-1) \\ \boldsymbol{\Sigma}(1) & \boldsymbol{\Sigma}(0) & \cdots & \boldsymbol{\Sigma}^\top(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}(M-1) & \boldsymbol{\Sigma}(M-2) & \cdots & \boldsymbol{\Sigma}(0) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\Psi}}_1^\top \\ \hat{\boldsymbol{\Psi}}_2^\top \\ \vdots \\ \hat{\boldsymbol{\Psi}}_M^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}(1) \\ \boldsymbol{\Sigma}(2) \\ \vdots \\ \boldsymbol{\Sigma}(M) \end{bmatrix}.\tag{A13}$$

which may be readily solved for the $\hat{\boldsymbol{\Psi}}_j$ [61]. The corresponding noise covariance is then given by

$$\hat{\mathbf{R}} = \boldsymbol{\Sigma}(0) - \sum_{m=1}^M \hat{\boldsymbol{\Psi}}_m \boldsymbol{\Sigma}(m).\tag{A14}$$

After solving equations (A13,A14), the matrices $\hat{\Psi}_1, \hat{\Psi}_2, \dots, \hat{\Psi}_M, \hat{\mathbf{R}}$ are substituted into the autoregressive model (A12) to simulate the daily fluctuations. The complete procedures for running the emulator are summarized in Algorithm 1.

Algorithm 1 Stochastic emulator of global climate.

Input: Temporal evolution of global mean temperature $T_g(t)$

Output: Emulated statistics of climate variables

Step 1: Emulate seasonal mean and variance;

- Compute seasonal global mean temperature $T_{s,g}$ for each season s ;
- For each mode i , predict the seasonal mean $\hat{\mu}_{s,i}(T_{s,g})$ and variance $\hat{\sigma}_{s,i}^2(T_{s,g})$.

Step 2: Generate stochastic daily fluctuations;

- At every time step t , sample a Gaussian random vector $\epsilon \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$;
- Compute the vector autoregressive process $\hat{\eta}(t)$ according to equation (A12).

Step 3: Construct time series of spatial fields;

- Combine seasonal mean and variance with daily fluctuations to obtain $\hat{a}_i(t)$;
- Multiply PCA coefficients by their mode shapes and superpose all the modes;
- Denormalize by σ_g and \bar{u} to compute U, V, T, Q in physical space (A8).

Step 4: Estimate statistics of interest;

- Average the spatial fields of U, V, T, Q over window to calculate statistics;
 - If needed, input $T_g(t)$ from a different ensemble member, repeat steps 1-3 and average over multiple members.
-

A.3 Nudging the Stochastic Emulator

The stochastic emulator introduced previously was designed to capture the second-order statistics of the leading PCA modes. While this emulator has demonstrated effectiveness in representing the conditional Gaussian distribution of certain variables, such as temperature [48], it inherently struggles to reproduce the non-Gaussian characteristics of climate data, including extreme events associated with higher-order PCA modes. A common approach to addressing this limitation involves using machine-learning models to debias the emulator. However, due to the stochastic nature of the emulated spatiotemporal data, instantaneous matches with reference data are not achievable. For instance, an emulated wind speed field on January 1st, 2025, would significantly differ from the corresponding ground truth dataset, whether sourced from ERA5 or CMIP6. Ideally, an infinite ensemble of realizations could be produced by the emulator, enabling selection of instances closest to reference observations for training a debiasing model. However, this method is impractical. A more realistic alternative is the nudging approach [62–65], where the emulator is forced by the deviation from the reference data to produce a time series of fields that approximately maintain the emulator’s statistical characteristics while closely aligning with the observed ground truth. Herein we interpret how to nudge the stochastic emulator. In the following, we detail how to implement nudging within the stochastic emulator framework.

Recall the formulation of the emulator (A8,A9),

$$\hat{a}_{s,i}(t) = \hat{\mu}_{s,i}(T_{s,g}) + \hat{\sigma}_{s,i}(T_{s,g}) \hat{\eta}_{s,i}(t) \quad (\text{A15})$$

$$\hat{q}_k(\mathbf{x}, t) = \bar{u}_k(\mathbf{x}, t) + \sigma_{g,k} \sum_{i=1}^{500} \hat{a}_i(t) \phi_k^{(i)}(\mathbf{x}). \quad (\text{A16})$$

The only stochastic component is the time series of daily fluctuations $\hat{\eta}_{s,i}$. All other parts are deterministic and constructed to align with the reference data. Therefore we focus on nudging $\hat{\eta}_{s,i}$, given the true fluctuations $\eta_{s,i}$. Hereafter we omit any subscripts to simplify the notation.

The nudged emulator, denoted as $\boldsymbol{\nu}$, is designed to follow the dynamics of the free-running emulator $\hat{\boldsymbol{\eta}}$, while driven by the deviation from the reference data,

$$\dot{\boldsymbol{\nu}} = \dot{\hat{\boldsymbol{\eta}}} - \frac{1}{\tau} (\boldsymbol{\nu} - \boldsymbol{\eta}). \quad (\text{A17})$$

The relaxation time scale τ is a constant that is independent from the season or the PCA mode. Equation (A17) has a closed-form solution,

$$\boldsymbol{\nu}(t) = \boldsymbol{\nu}(0)e^{-t/\tau} + \int_0^t e^{-(t-s)/\tau} \left(\dot{\hat{\boldsymbol{\eta}}}(s) + \frac{1}{\tau} \boldsymbol{\eta}(s) \right) ds. \quad (\text{A18})$$

The time derivative term $\dot{\hat{\boldsymbol{\eta}}}$ is approximated using the first-order Euler scheme and computed from the free-running emulator data. Combining the nudged time series of daily fluctuations $\boldsymbol{\nu}$ with seasonal mean and variance, we obtain the complete nudged PCA time series and the spatiotemporal fields,

$$\hat{a}_{s,i}^\nu(t) = \hat{\mu}_{s,i}(T_{s,g}) + \hat{\sigma}_{s,i}(T_{s,g}) \hat{\nu}_{s,i}(t) \quad (\text{A19})$$

$$\hat{q}_k^\nu(\mathbf{x}, t) = \bar{u}_k(\mathbf{x}, t) + \sigma_{g,k} \sum_{i=1}^{500} \hat{a}_i^\nu(t) \phi_k^{(i)}(\mathbf{x}). \quad (\text{A20})$$

The interpretation of nudging and the selection of τ have been thoroughly discussed in [46]. Briefly speaking, the relaxation timescale τ serves to separate the time scales between slow and fast dynamics. The feedback term in equation (A17) drives the slow dynamics of $\boldsymbol{\nu}$ towards the reference trajectory $\boldsymbol{\eta}$ in the state space, while allowing the fast dynamics of $\boldsymbol{\nu}$ to freely evolve. Thus, when pairs of the nudged and reference data are used for training a machine-learning model, we are essentially learning a map that corrects the fast features of the imperfect emulator and improve the performance on extreme events. In our case, the relaxation timescale is set as $\tau = 6\text{hrs}$, consistent with previous work [45]. Minor adjustments of τ , such as to 3 or 12 hours, do not significantly alter the results.

The feedback term in (A17), although driving the nudged emulator towards the reference, introduces artificial dissipation not present in the free-running emulator. Such an effect leads to a distribution of $\boldsymbol{\nu}$ and $\hat{\boldsymbol{q}}^\nu$ that is slightly different from the

free-running emulator. In order for a neural network trained on the nudged dataset to generalize to unseen free-running emulator data, this discrepancy must be remedied. To this end, we rescale the nudged solution $\hat{\mathbf{q}}^\nu$ in each season so that its mean and variance match those of the free-running emulator $\hat{\mathbf{q}}$ at each grid point.

Appendix B Machine Learned Debiasing

B.1 Conditional Score-based Diffusion model

Here we describe the training strategy and network architecture used in the ML correction step of our model. Our model relies on the framework introduced by Barthel Sorensen et al. [45], which aims to learn a deterministic map from the nudged trajectory to the reference trajectory,

$$\mathbf{u} = \mathcal{F}(\hat{\mathbf{q}}^\nu). \quad (\text{B21})$$

In practice, such a mapping is not necessarily deterministic. There could exist multiple reference state \mathbf{u} that are close to the same nudged state $\hat{\mathbf{q}}^\nu$. Therefore, we generalize this framework by learning a conditional probability distribution function,

$$p(\mathbf{u} | \hat{\mathbf{q}}^\nu). \quad (\text{B22})$$

If the mapping is actually deterministic, the conditional probability distribution will collapse to a Dirac delta function $\delta(\mathbf{u} - \mathcal{F}(\hat{\mathbf{q}}^\nu))$. Once the conditional PDF (B22) is learned, we can provide the free-running emulation $\hat{\mathbf{q}}$ as the conditional information to generate debiased estimations of the state variables $\hat{\mathbf{u}}$,

$$\hat{\mathbf{u}}(\mathbf{x}, t) \sim \mathcal{G}_{\theta, 2}[\hat{\mathbf{q}}(\mathbf{x}, t)] \quad (\text{B23})$$

Although learning and sampling high-dimensional PDFs were long considered intractable, these tasks have recently become practical thanks to advances in deep generative models. In this study, we adopt conditional score-based diffusion model [49, 66] that has been demonstrated effective for geophysical datasets [50]. Other frameworks, such as flow matching [67] and stochastic interpolant [68], could likewise address the debiasing problem considered here. The choice of the generative model is beyond the scope of this work and will be investigated in the future.

Our implementation of score-based diffusion model follows that of Bischoff and Deck [50]. To simplify the notation, we will use \mathbf{q} to represent the nudged emulation $\hat{\mathbf{q}}^\nu$. The diffusion model consists of a forward diffusion process, which maps the data distribution to normal distribution, and a reverse denoising process that transforms Gaussian noise to a sample or image of the climate state. Specifically, given an initial condition $\mathbf{u}(t=0) \sim p_{\text{data}}(\mathbf{u}|\mathbf{q})$ drawn from the training data, the forward diffusion process is defined by the stochastic differential equation (SDE),

$$d\mathbf{u} = g(t)d\mathbf{W}, \quad (\text{B24})$$

where the diffusion coefficient $g(t)$ is a non-negative prescribed function and \mathbf{W} is a Wiener process. Note that the diffusion time t is within $[0, 1]$ and should be distinguished from the physical time t . At any time t , the solution to the SDE (B24) is a “noised” image $\mathbf{u}(t)$, which follows a normal distribution conditioned on $\mathbf{u}(0)$,

$$\mathbf{u}(t) \sim \mathcal{N}(\mathbf{u}(0), \sigma^2(t)) = p(\mathbf{u}(t)|\mathbf{u}(0)), \quad (\text{B25})$$

where the variance $\sigma^2(t)$ depends on $g(t)$,

$$\sigma^2(t) = \int_0^t g^2(t') dt'. \quad (\text{B26})$$

The marginal distribution of $\mathbf{u}(t)$ after integrating out $\mathbf{u}(0)$ is defined as $p_t(\mathbf{u}(t)|\mathbf{q})$, which is generally non-Gaussian. The diffusion coefficient $g(t)$ is chosen such that at $t = 1$, the variance $\sigma^2(t = 1)$ has much larger magnitude than the original $\mathbf{u}(t = 0)$. Therefore, $\mathbf{u}(t = 0)$ has lost all memory of initial condition, and

$$p(\mathbf{u}(1)|\mathbf{u}(0)) = p(\mathbf{u}(1)) = \mathcal{N}(\mathbf{0}, \sigma^2(1)). \quad (\text{B27})$$

According to Anderson's theorem [69], the reverse of equation (B24) is also a diffusion process, running backward in time and governed by the following SDE,

$$d\mathbf{u} = -g^2(t)\mathbf{s}(\mathbf{u}, \mathbf{q}, t)dt + g(t)d\bar{\mathbf{W}}, \quad (\text{B28})$$

where $\bar{\mathbf{W}}$ is a reverse-time Wiener process, and $\mathbf{s}(\mathbf{u}, t)$ is the conditional score,

$$\mathbf{s}(\mathbf{u}, \mathbf{q}, t) = \nabla_{\mathbf{u}} \log p_t(\mathbf{u}(t)|\mathbf{q}). \quad (\text{B29})$$

If we have access to the score for all t , we can derive the reverse diffusion process, simulate it from $t = 1$ to $t = 0$, and generate samples that follow the data distribution. To this end, we approximate the score by a neural network, $\mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t)$, which is obtained by minimizing the score-matching loss or Fisher's divergence,

$$\mathcal{L}_{SM}(\theta) := \frac{1}{2} \mathbb{E}_{\substack{t \sim U(0,1) \\ \mathbf{u}(t), \mathbf{q} \sim p_t(\mathbf{u}(t)|\mathbf{q})}} \left[\sigma^2(t) \|\nabla_{\mathbf{u}} \log p_t(\mathbf{u}| \mathbf{q}) - \mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t)\|_2^2 \right], \quad (\text{B30})$$

where $U(0, 1)$ stands for a uniform distribution from 0 to 1. However, the loss function (B30) cannot be directly optimized, since the true conditional score $\nabla_{\mathbf{u}} \log p_t(\mathbf{u}| \mathbf{q})$ is unknown. Taking advantage of the Gaussian property of the forward diffusion process (B25, B27), [49] showed that \mathcal{L}_{SM} is equal to the following loss up to an additive term,

$$\mathcal{L}(\theta) := \frac{1}{2} \mathbb{E}_{\substack{t \sim U(0,1) \\ \mathbf{u}(0), \mathbf{q} \sim p_t(\mathbf{u}_0 | \mathbf{q}) \\ \mathbf{u}(t) \sim p(\mathbf{u}(t) | \mathbf{u}(0))}} \left[\sigma^2(t) \|\nabla_{\mathbf{u}} \log p_t(\mathbf{u}(t) | \mathbf{u}(0)) - \mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t)\|_2^2 \right]. \quad (\text{B31})$$

This expression only involves $\nabla_{\mathbf{u}} \log p_t(\mathbf{u}(t) | \mathbf{u}(0))$ which can be computed analytically from the forward diffusion process.

Once the conditional score is learned through training, it can be substituted into the backward SDE (equation B28) to generate a debiased sample. The backward SDE is simulated using Euler-Maruyama scheme. The complete procedures of training and sampling are summarized in Algorithm 2.

Algorithm 2 Conditional score-based diffusion model.

Training

Input: Reference data and nudged emulation $\{\mathbf{u}_i, \mathbf{q}_i\} \sim p(\mathbf{u}|\mathbf{q})$

repeat

- $\mathbf{u}(0), \mathbf{q} \sim p(\mathbf{u}|\mathbf{q})$
- $t \sim U(0, 1)$
- $\mathbf{u}(t) \sim p(\mathbf{u}(t)|\mathbf{u}(0))$
- Take gradient descent step on $\nabla_\theta \left[\sigma^2(t) \|\nabla_{\mathbf{u}} \log p_t(\mathbf{u}(t)|\mathbf{u}(0)) - \mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t)\|_2^2 \right]$

until converged

Output: Trained neural network $\mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t)$.

End

Sampling

Input: Snapshot of emulated state \mathbf{q}

$\mathbf{u}(1) \sim \mathcal{N}(\mathbf{0}, \sigma^2(1))$

for $t = 1$ **to** 0 **do**

- Evaluate $\mathbf{s}_\theta(\mathbf{u}(t), \mathbf{q}, t)$
- $\mathbf{u}(t - \Delta t) = \mathbf{u}(t) + g^2(t) \mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t) \Delta t - g(t) (\bar{\mathbf{W}}(t) - \bar{\mathbf{W}}(t - \Delta t))$
- $t \leftarrow t - \Delta t$

end for

Output: Debiased snapshot $\hat{\mathbf{u}} = \mathbf{u}(t = 0)$

End

B.2 Network Architecture and Training Parameters

Before feeding the emulation and reference data into the diffusion model, we remove the true climatological mean and scale each variable (U, V, T, Q) by twice its own globally-averaged standard deviation. In other words, we focus on correcting the fluctuation fields provided by the emulator, and each variable is scaled to the same order of magnitude. Regarding the diffusion coefficient $g(t)$ in equation (B24), we adopt the “variance-exploding” schedule,

$$g(t) = \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^t \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)}, \quad (\text{B32})$$

where $\sigma_{\min} = 0.01$ and σ_{\max} is chosen as the maximum 2-norm distance between any two snapshots of \mathbf{u} .

The neural network architecture we adopted is a U-Net [50, 51], denoted as

$$\mathbf{s}_\theta(\mathbf{u}, \mathbf{q}, t) = \mathcal{U}(\mathbf{X}, t; \boldsymbol{\theta}). \quad (\text{B33})$$

The first input X is a tensor of size (N_1, N_2, C_{in}) , where (N_1, N_2) are longitude and latitude dimensions, and C_{in} is the number of channels. In our case, $C_{in} = 8$, including (U, V, T, Q) from the nudged emulation and reference data, respectively. The output of our U-Net is another tensor of size (N_1, N_2, C_{out}) . The number of output channels is $C_{out} = 4$. The U-Net architecture consists of

1. A lifting layer which increases the number of channels from C_{in} to 32;
2. Three downsampling convolutional layers, each of which reduces the spatial dimension and increase the number of channels by a factor of 2;
3. Eight residual blocks [70] to promote continuity in the latent space;
4. Three nearest neighbor up sampling layer and convolution layers which mirror the downsampling operations;
5. A Final projection layer that decreases the number of channels to C_{out} .

Our choice of the optimizer follows [49] and [50]. An Adam optimizer is adopted with a learning rate of $\lambda_0 = 2e - 4$, $\epsilon = 1e - 8$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The gradient norm clipping is employed to a value of 1.0. For both the ERA5 and CMIP6 datasets, we set the batchsize as 8, and train the U-Net for 200 epochs.

Appendix C Data Post-processing and Evaluation Metrics

This section provides detailed definitions and calculation methods for all statistics and metrics presented in the main figures. Given that the climatological mean $\bar{\mathbf{u}}(\mathbf{x}, t)$ (equation A1) is assumed known, our analysis focuses on evaluating the statistics of the fluctuation fields, $\mathbf{u} - \bar{\mathbf{u}}$. This approach enables a clearer comparison between reference statistics and GEN² prediction. In the following subsections, the fluctuations from the climatological mean, $\mathbf{u} - \bar{\mathbf{u}}$, will be simply written as \mathbf{u} for notational convenience.

C.1 Single-point and two-point statistics

Without loss of generality, we use the zonal wind speed $U(\mathbf{x}, t)$ as an example. To evaluate the statistics of U at location \mathbf{x} , we perform a time average (e.g. from 1979 to 2018 for ERA5 dataset). If we have N_t time steps available, the mean and standard deviation are computed as,

$$\mu_U(\mathbf{x}, t_j) = \frac{1}{N_t} \sum_{j=1}^{N_t} U(\mathbf{x}, t_j) \quad \text{and} \quad \sigma_U(\mathbf{x}) = \sqrt{\frac{1}{N_t - 1} \sum_{j=1}^{N_t} (U(\mathbf{x}, t_j) - \mu_U(\mathbf{x}, t_j))^2}. \quad (\text{C34})$$

To obtain the unbiased skewness, we first compute

$$s_U(\mathbf{x}) = \frac{\frac{1}{N_t} \sum_{j=1}^{N_t} (U(\mathbf{x}, t_j) - \mu_U(\mathbf{x}, t_j))^3}{\left(\frac{1}{N_t} \sum_{j=1}^{N_t} (U(\mathbf{x}, t_j) - \mu_U(\mathbf{x}, t_j))^2 \right)^{3/2}}, \quad (\text{C35})$$

which is then substituted into

$$s_U^0(\mathbf{x}) = \frac{\sqrt{N_t(N_t - 1)}}{N_t - 2} s_U(\mathbf{x}). \quad (\text{C36})$$

The calculation of unbiased kurtosis also consists of two steps:

$$k_U(\mathbf{x}) = \frac{\frac{1}{N_t} \sum_{j=1}^{N_t} (U(\mathbf{x}, t_j) - \mu_U(\mathbf{x}, t_j))^4}{\left(\frac{1}{N_t} \sum_{j=1}^{N_t} (U(\mathbf{x}, t_j) - \mu_U(\mathbf{x}, t_j))^2 \right)^2}, \quad (\text{C37})$$

$$k_U^0(\mathbf{x}) = \frac{N_t - 1}{(N_t - 2)(N_t - 3)} ((N_t + 1)k_U(\mathbf{x}) - 3(N_t - 1) + 3). \quad (\text{C38})$$

At the same location \mathbf{x} , the correlation coefficient between two variables U and V are defined as,

$$\rho(U, V) = \frac{\text{cov}(U(\mathbf{x}, t), V(\mathbf{x}, t))}{\text{cov}(U(\mathbf{x}, t), U(\mathbf{x}, t)) \text{cov}(V(\mathbf{x}, t), V(\mathbf{x}, t))}, \quad (\text{C39})$$

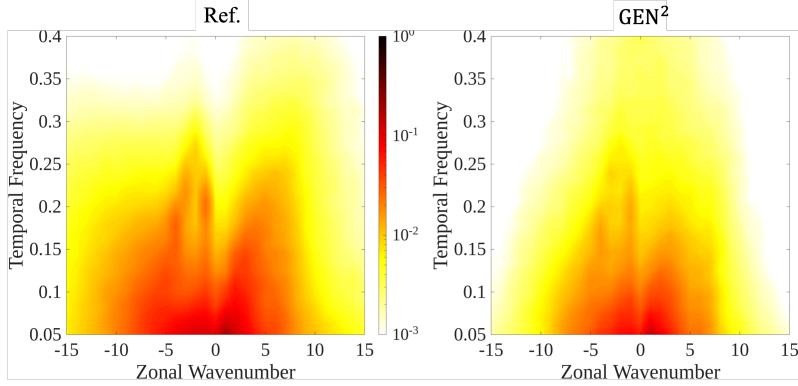


Fig. C2 Raw Wheeler-Kiladis spectrum of zonal wind, computed using 1979-2018 ERA5 reference data and GEN² prediction. These spectra are normalized by the “background power” to obtain the spectra in figure 2(c) of the main manuscript.

where $\text{cov}(U(\mathbf{x}, t), V(\mathbf{x}, t))$ is the time-averaged covariance between U and V .

The two-point correlation coefficient of U is defined as,

$$\rho(U(\mathbf{x}_0), U(\mathbf{x})) = \frac{\text{cov}(U(\mathbf{x}_0, t), U(\mathbf{x}, t))}{\text{cov}(U(\mathbf{x}_0, t), U(\mathbf{x}_0, t)) \text{cov}(U(\mathbf{x}, t), U(\mathbf{x}, t))}. \quad (\text{C40})$$

The anchor point \mathbf{x}_0 is selected as major cities (e.g. Boston, Hong Kong) in figure 3, 6 and in section D.

The global root-mean-square error (RMSE) of an arbitrary statistic \mathcal{Q} (e.g. in figure 5 and table D3,D4) is defined as,

$$\text{RMSE}(\mathcal{Q}) = \left[\frac{1}{S} \int_S \left(\hat{\mathcal{Q}}(\theta, \varphi, t) - \mathcal{Q}(\theta, \varphi, t) \right)^2 \cos \theta d\theta d\varphi \right]^{1/2}, \quad (\text{C41})$$

where $\hat{\mathcal{Q}}$ is the GEN² prediction and \mathcal{Q} is the reference statistics.

C.2 Wheeler-Kiladis spectrum

The Wheeler-Kiladis spectrum is computed following the procedure described in Wheeler and Kiladis [56], Kiladis et al. [57]. Given data as a function of longitude, latitude, and time $[(\theta, \phi, t)]$ the latitude range is first truncated to $\phi \in [-15^\circ, 15^\circ]$. Then for each latitude ϕ_j , and time t_j the Fourier spectrum is computed in the azimuthal direction θ giving rise to a azimuthal wavenumber m . Then for each ϕ_j and m_j the time series of data is split into a series of overlapping segments. Following [56, 57] we set the length of each segment to 96 days and the overlap to 65 days. Each segment is then detrended using a linear fit and Fourier transformed in time - giving a temporal frequency f . After averaging over all latitudes and all temporal segments, the raw Wheeler-Kiladis spectra are shown in figure C2.

Due to the strong “redness” of the spectra, detailed features corresponding to the equatorial waves are obscured. To better identify the ridges of the spectra, we

first apply a 1-2-1 filter ten times to obtain a much smoother “background” spectra. Then the raw spectra in figure C2 are divided by the background [56]. The results, as shown in figure 2(c) of the main manuscript, more clearly show the spectral peaks that correspond to different types of equatorial waves. Note that the spectra in our results should not be directly compared against the plots in [56]. The reason is that their analysis was based on the long-wave radiation data, which are proxy for cloudiness, whereas our analysis focuses on near-surface zonal wind speed.

Statistics	RMSE of U stats			RMSE of V stats		
	Em.	GEN ²	Change	Em.	GEN ²	Change
Std	0.43	0.15	-65%	0.47	0.13	-73%
97.5% quantile	1.14	0.46	-60%	0.99	0.33	-66%
Skewness	0.41	0.19	-53%	0.27	0.14	-48%
Kurtosis	0.84	0.57	-33%	0.68	0.46	-33%

Table D1 RMSE of single-point statistics of U , V , defined as equation (C41). Columns labeled as “Em.” are the RMSE of the prediction of conditional Gaussian emulator, and “GEN²” is the full-model prediction. “Change” columns are the relative error change from conditional Gaussian emulator to GEN², more precisely, $(\text{RMSE}(\text{GEN}^2) - \text{RMSE}(\text{Em.})) / \text{RMSE}(\text{Em.})$.

Statistics	RMSE of T stats			RMSE of Q stats		
	Em.	GEN ²	Change	Em.	GEN ²	Change
Std	0.23	0.10	-56%	0.19	0.05	-72%
97.5% quantile	0.67	0.35	-48%	0.46	0.17	-63%
Skewness	0.34	0.20	-42%	0.50	0.27	-47%
Kurtosis	0.89	0.68	-24%	1.78	1.38	-23%

Table D2 Same as table D1, but for temperature T and Q .

Appendix D Additional Results: Bias Reduction via ML Correction

To illustrate the debiasing capabilities of the ML correction step, we show in figure D3 the bias in the 97.5% quantile, predicted with or without ML correction. As explained at the beginning of Appendix C, the statistics are evaluated for the fluctuation fields. Before applying ML correction (middle column), the error of the conditional Gaussian emulator is already moderately accurate. For example, the error of T at most locations is within $3K$, and the highest error is within $3K$. The ML model (right column) consistently reduces the error of all the state variables at almost all the locations. A more quantitative comparison is provided in table D1,D2. The bias in standard deviation, quantile, skewness, and kurtosis are all significantly reduced by ML correction.

The bias reduction in two-point correlations are shown in figure D4. We select Lagos and Tehran for visualization, because the bias of the conditional Gaussian emulator is more pronounced at these two locations. Top panels in figure D4 are the bias of the conditional Gaussian emulator, without ML correction. Bottom panels are the bias of GEN² prediction.

Table D3 and D4 summarize the error of two-point correlations at more locations, selected from different regions and climate over the world. At most locations, the conditional Gaussian emulator already achieves an accurate prediction, with RMSE lower than 0.03. After applying ML correction, the errors are significantly reduced at all the locations considered. These results demonstrate the robustness of the ML correction.

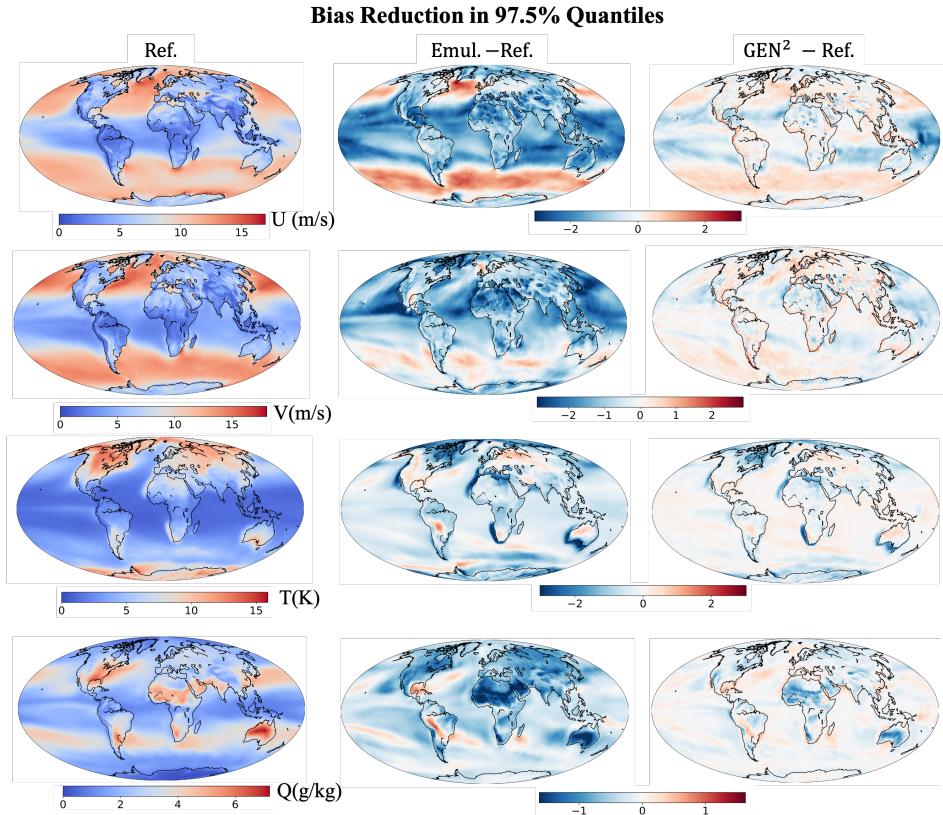


Fig. D3 Left column: 97.5% quantile of zonal wind, meridional wind, temperature, computed from reference data. Middle column: bias of conditional Gaussian emulation. Right column: bias of “GEN²” approach.

Bias Reduction in 2-Point Correlation From ML Correction

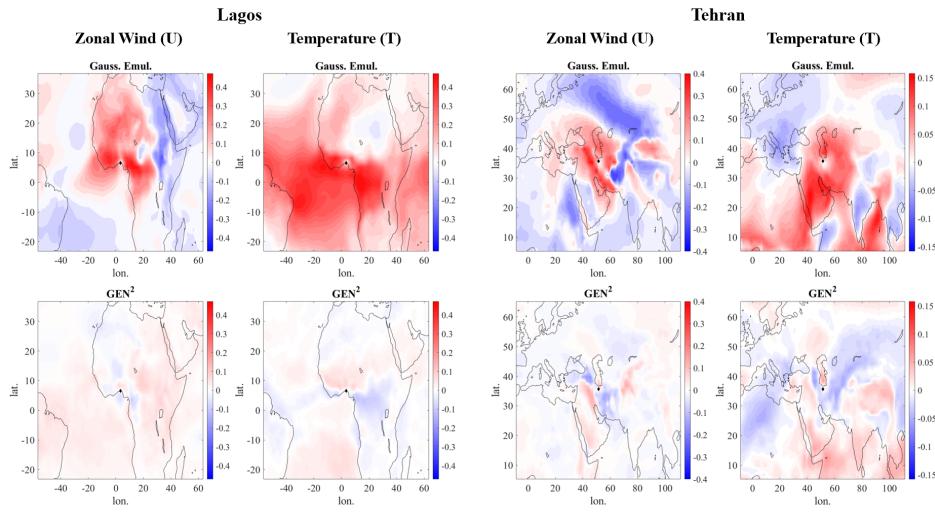


Fig. D4 Bias in two-point correlations of zonal wind and temperature, centered at Lagos and Tehran. Contour plots represent bias relative to reference ERA5 data. Panels labeled “Gauss. Emul.” correspond to predictions of the conditional Gaussian emulator *only* (no ML correction) and “GEN²” represents the full model prediction (with ML correction).

City	Anchor point \mathbf{x}_0		RMSE of $\rho(U(\mathbf{x}_0), U(\mathbf{x}))$			RMSE of $\rho(V(\mathbf{x}_0), V(\mathbf{x}))$		
	Em.	Lat(N)	GEN ²	Change	Em.	GEN ²	Change	
Boston	-71.1	42.4	0.016	0.009	-43%	0.018	0.008	-53%
Los Angeles	-118.2	34.1	0.028	0.011	-62%	0.026	0.009	-67%
Chicago	-87.6	41.9	0.020	0.008	-59%	0.017	0.008	-54%
Houston	-95.4	29.8	0.020	0.009	-57%	0.017	0.009	-48%
Kansas City	-94.6	39.1	0.022	0.008	-64%	0.017	0.008	-51%
London	-0.1	51.5	0.018	0.009	-50%	0.019	0.008	-60%
Anchorage	-149.9	61.2	0.019	0.012	-36%	0.021	0.009	-58%
Paris	2.4	48.9	0.019	0.009	-53%	0.019	0.008	-57%
Athens	23.7	38.0	0.024	0.010	-57%	0.022	0.009	-60%
Moscow	37.6	55.8	0.024	0.010	-58%	0.022	0.008	-66%
Stockholm	18.1	59.3	0.022	0.010	-57%	0.020	0.008	-61%
Tokyo	139.7	35.7	0.017	0.008	-51%	0.017	0.009	-48%
Hong Kong	114.2	22.3	0.026	0.009	-64%	0.025	0.009	-65%
New Delhi	77.1	28.6	0.028	0.010	-64%	0.028	0.010	-65%
Tehran	51.4	35.7	0.030	0.011	-63%	0.034	0.010	-70%
Astana	71.5	51.2	0.022	0.009	-60%	0.022	0.009	-60%
Cairo	31.2	30.0	0.029	0.009	-70%	0.027	0.008	-69%
Cape Town	18.4	-33.9	0.018	0.009	-50%	0.022	0.009	-60%
Lagos	3.4	6.5	0.043	0.015	-65%	0.040	0.009	-78%
Kisangani	25.2	0.1	0.047	0.015	-69%	0.036	0.009	-76%
Mombasa	39.7	-4.0	0.041	0.018	-55%	0.038	0.010	-74%
Sydney	151.2	-33.9	0.020	0.009	-56%	0.020	0.008	-58%
Brasília	-47.9	-15.8	0.029	0.012	-58%	0.044	0.010	-78%
Bogota	-74.1	4.7	0.054	0.023	-57%	0.034	0.017	-49%
Buenos Aires	-58.4	-34.6	0.020	0.009	-55%	0.019	0.009	-54%

Table D3 RMSE of two-point correlation of U, V . Columns labeled as “Em.” are the RMSE of the prediction of conditional Gaussian emulator, and “GEN²” is the full-model prediction. “Change” columns are the relative error change from conditional Gassuain emulator to GEN², more precisely, $(\text{RMSE}(\text{GEN}^2) - \text{RMSE}(\text{Em.})) / \text{RMSE}(\text{Em.})$.

City	Anchor point x_0		RMSE of $\rho(T(x_0), T(x))$			RMSE of $\rho(Q(x_0), Q(x))$		
	Em.	Lat(N)	GEN ²	Change	Em.	GEN ²	Change	
Boston	-71.1	42.4	0.017	0.012	-30%	0.016	0.009	-43%
Los Angeles	-118.2	34.1	0.021	0.011	-48%	0.022	0.010	-54%
Chicago	-87.6	41.9	0.014	0.010	-29%	0.017	0.009	-46%
Houston	-95.4	29.8	0.019	0.014	-28%	0.015	0.008	-47%
Kansas City	-94.6	39.1	0.014	0.010	-30%	0.016	0.009	-45%
London	-0.1	51.5	0.023	0.010	-55%	0.022	0.011	-49%
Anchorage	-149.9	61.2	0.026	0.013	-48%	0.026	0.014	-48%
Paris	2.4	48.9	0.023	0.010	-54%	0.023	0.012	-47%
Athens	23.7	38.0	0.018	0.012	-35%	0.023	0.009	-60%
Moscow	37.6	55.8	0.017	0.011	-36%	0.021	0.010	-51%
Stockholm	18.1	59.3	0.021	0.010	-53%	0.021	0.011	-49%
Tokyo	139.7	35.7	0.019	0.011	-40%	0.015	0.009	-39%
Hong Kong	114.2	22.3	0.023	0.014	-41%	0.020	0.011	-46%
New Delhi	77.1	28.6	0.031	0.018	-43%	0.022	0.011	-47%
Tehran	51.4	35.7	0.030	0.018	-42%	0.033	0.010	-68%
Astana	71.5	51.2	0.017	0.011	-32%	0.026	0.011	-59%
Cairo	31.2	30.0	0.031	0.014	-55%	0.027	0.008	-69%
Cape Town	18.4	-33.9	0.022	0.012	-45%	0.022	0.010	-55%
Lagos	3.4	6.5	0.094	0.018	-81%	0.034	0.013	-63%
Kisangani	25.2	0.1	0.072	0.015	-79%	0.046	0.011	-77%
Mombasa	39.7	-4.0	0.100	0.023	-77%	0.059	0.015	-75%
Sydney	151.2	-33.9	0.025	0.012	-52%	0.021	0.010	-53%
Brasilia	-47.9	-15.8	0.042	0.013	-68%	0.030	0.012	-62%
Bogota	-74.1	4.7	0.092	0.026	-71%	0.054	0.025	-53%
Buenos Aires	-58.4	-34.6	0.019	0.012	-38%	0.017	0.010	-42%

Table D4 Same as table D3, but for temperature T and humidity Q .

References

- [1] Raymond, C., Horton, R.M., Zscheischler, J., Martius, O., AghaKouchak, A., Balch, J., Bowen, S.G., Camargo, S.J., Hess, J., Kornhuber, K., Oppenheimer, M., Ruane, A.C., Wahl, T., White, K.: Understanding and managing connected extreme events. *Nature Climate Change* **10**(7), 611–621 (2020) <https://doi.org/10.1038/s41558-020-0790-4>
- [2] Robinson, A., Lehmann, J., Barriopedro, D., Rahmstorf, S., Coumou, D.: Increasing heat and rainfall extremes now far outside the historical climate. *npj Climate and Atmospheric Science* **4**(1), 1–4 (2021) <https://doi.org/10.1038/s41612-021-00202-w>
- [3] Fischer, E.M., Sippel, S., Knutti, R.: Increasing probability of record-shattering climate extremes. *Nature Climate Change* **11**(8), 689–695 (2021) <https://doi.org/10.1038/s41558-021-01092-9>
- [4] Allen, S., Barros, V., (Canada, I., (UK, D., Cardona, O., Cutter, S., Dube, O.P., Ebi, K., (USA, C., Handmer, J., (Australia, P., Lavell, A., (USA, K., Mastrandrea, M., Mcbean, G., Mechler, R., (UK, T., Nicholls, N., (Norway, K., (USA, T.: 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. Technical report (November 2012). <https://doi.org/10.13140/2.1.3117.9529>
- [5] Houser, T., Hsiang, S., Kopp, R., Larsen, K., Delgado, M., Jina, A., Mastrandrea, M., Mohan, S., Muir-Wood, R., Rasmussen, D.J., Rising, J., Wilson, P., Fisher-Vanden, K., Greenstone, M., Heal, G., Oppenheimer, M., Stern, N., Ward, B., Bloomberg, M.R., Paulson, H.M., Steyer, T.F.: Economic Risks of Climate Change: An American Prospectus. Columbia University Press, ??? (2015). <https://doi.org/10.7312/hous17456> . <https://www.jstor.org/stable/10.7312/hous17456>
- [6] AON: 2025 Climate & catastrophe insight (2025)
- [7] Fiedler, T., Pitman, A.J., Mackenzie, K., Wood, N., Jakob, C., Perkins-Kirkpatrick, S.E.: Business risk and the emergence of climate analytics. *Nature Climate Change* **11**(2), 87–94 (2021) <https://doi.org/10.1038/s41558-020-00984-6>
- [8] Smagorinsky, J.: General circulation experiments with the primitive equations: I. the basic experiment. *Monthly Weather Review* **91**(3), 99–164 (1963) [https://doi.org/10.1175/1520-0493\(1963\)091<0099:GCEWTP>2.3.CO;2](https://doi.org/10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2)
- [9] Smagorinsky, J., Manabe, S., Holloway, J.L.: NUMERICAL RESULTS FROM A NINE-LEVEL GENERAL CIRCULATION MODEL OF THE ATMOSPHERE. *Monthly Weather Review* **93**(12), 727–768 (1965) [https://doi.org/10.1175/1520-0493\(1965\)093<0727:NRFANL>2.3.CO;2](https://doi.org/10.1175/1520-0493(1965)093<0727:NRFANL>2.3.CO;2)

- [10] Manabe, S., Smagorinsky, J., Strickler, R.F.: SIMULATED CLIMATOLOGY OF A GENERAL CIRCULATION MODEL WITH A HYDROLOGIC CYCLE. *Monthly Weather Review* **93**(12), 769–798 (1965) [https://doi.org/10.1175/1520-0493\(1965\)093<0769:SCOAGC>2.3.CO;2](https://doi.org/10.1175/1520-0493(1965)093<0769:SCOAGC>2.3.CO;2)
- [11] Mintz, Y.: Very Long-Term Global Integration of the Primitive Equations of Atmospheric Motion: An Experiment in Climate Simulation. In: Billings, D.E., Broecker, W.S., Bryson, R.A., Cox, A., Damon, P.E., Donn, W.L., Eriksson, E., Ewing, M., Fletcher, J.O., Hamilton, W., Jerzykiewicz, M., Kutzbach, J.E., Lorenz, E.N., Mintz, Y., Mitchell, J.M., Saltzman, B., Serkowski, K., Shen, W.C., Suess, H.E., Tanner, W.F., Weyl, P.K., Worthington, L.V., Mitchell, J.M. (eds.) Causes of Climatic Change: A Collection of Papers Derived from the INQUA—NCAR Symposium on Causes of Climatic Change, August 30–31, 1965, Boulder, Colorado. Meteorological Monographs, pp. 20–36. American Meteorological Society, Boston, MA (1968). https://doi.org/10.1007/978-1-935704-38-6_3
- [12] Taylor, M.A., Cyr, A.S., Fournier, A.: A non-oscillatory advection operator for the compatible spectral element method. In: International Conference on Computational Science, pp. 273–282 (2009). Springer
- [13] Dennis, J.M., Edwards, J., Evans, K.J., Guba, O., Lauritzen, P.H., Mirin, A.A., St-Cyr, A., Taylor, M.A., Worley, P.H.: Cam-se: A scalable spectral element dynamical core for the community atmosphere model. *The International Journal of High Performance Computing Applications* **26**(1), 74–89 (2012)
- [14] Golaz, J.-C., Van Roekel, L.P., Zheng, X., Roberts, A.F., Wolfe, J.D., Lin, W., Bradley, A.M., Tang, Q., Maltrud, M.E., Forsyth, R.M., et al.: The doe e3sm model version 2: overview of the physical model and initial model evaluation. *Journal of Advances in Modeling Earth Systems* **14**(12) (2022)
- [15] Stensrud, D.J.: Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge University Press, Cambridge (2007). <https://doi.org/10.1017/CBO9780511812590>. <https://www.cambridge.org/core/books/parameterizationschemes/C7C8EC8901957314433BE7C8BC36F16D>
- [16] Holloway, C.E., Neelin, J.D.: Moisture Vertical Structure, Column Water Vapor, and Tropical Deep Convection. *Journal of the Atmospheric Sciences* **66**(6), 1665–1683 (2009) <https://doi.org/10.1175/2008JAS2806.1>
- [17] Friend, A.D., Lucht, W., Rademacher, T.T., Keribin, R., Betts, R., Cadule, P., Ciais, P., Clark, D.B., Dankers, R., Falloon, P.D., Ito, A., Kahana, R., Kleidon, A., Lomas, M.R., Nishina, K., Ostberg, S., Pavlick, R., Peylin, P., Schaphoff, S., Vuichard, N., Warszawski, L., Wiltshire, A., Woodward, F.I.: Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO₂. *Proceedings of the National Academy of Sciences* **111**(9),

- 3280–3285 (2014) <https://doi.org/10.1073/pnas.1222477110>
- [18] Bloom, A.A., Exbrayat, J.-F., Velde, I.R., Feng, L., Williams, M.: The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *Proceedings of the National Academy of Sciences* **113**(5), 1285–1290 (2016) <https://doi.org/10.1073/pnas.1515160113>
- [19] Rasp, S., Pritchard, M.S., Gentine, P.: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences* **115**(39), 9684–9689 (2018) <https://doi.org/10.1073/pnas.1810286115>
- [20] Brenowitz, N.D., Bretherton, C.S.: Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining. *Journal of Advances in Modeling Earth Systems* **11**(8), 2728–2744 (2019) <https://doi.org/10.1029/2019MS001711>
- [21] Yuval, J., O’Gorman, P.A., Hill, C.N.: Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision. *Geophysical Research Letters* **48**(6), 2020–091363 (2021) <https://doi.org/10.1029/2020GL091363>
- [22] Watt-Meyer, O., Brenowitz, N.D., Clark, S.K., Henn, B., Kwa, A., McGibbon, J., Perkins, W.A., Bretherton, C.S.: Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters* **48**(15), 2021–092555 (2021) <https://doi.org/10.1029/2021GL092555>
- [23] Bretherton, C.S., Henn, B., Kwa, A., Brenowitz, N.D., Watt-Meyer, O., McGibbon, J., Perkins, W.A., Clark, S.K., Harris, L.: Correcting Coarse-Grid Weather and Climate Models by Machine Learning From Global Storm-Resolving Simulations. *Journal of Advances in Modeling Earth Systems* **14**(2), 2021–002794 (2022) <https://doi.org/10.1029/2021MS002794>
- [24] Clark, S.K., Brenowitz, N.D., Henn, B., Kwa, A., McGibbon, J., Perkins, W.A., Watt-Meyer, O., Bretherton, C.S., Harris, L.M.: Correcting a 200 km Resolution Climate Model in Multiple Climates by Machine Learning From 25 km Resolution Simulations. *Journal of Advances in Modeling Earth Systems* **14**(9), 2022–003219 (2022) <https://doi.org/10.1029/2022MS003219>
- [25] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Marandani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *arXiv* (2022). <https://doi.org/10.48550/arXiv.2202.11214> . <http://arxiv.org/abs/2202.11214>
- [26] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., *et al.*: Learning skillful medium-range global weather forecasting. *Science* **382**(6677), 1416–1421 (2023)

- [27] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Vaughan, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J.A., Dong, H., Gupta, J.K., Thambiratnam, K., Archibald, A.T., Wu, C.-C., Heider, E., Welling, M., Turner, R.E., Perdikaris, P.: A Foundation Model for the Earth System (2024). <https://arxiv.org/abs/2405.13063>
- [28] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**(7970), 533–538 (2023)
- [29] Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S.K., Henn, B., Duncan, J., Brenowitz, N.D., Kashinath, K., Pritchard, M.S., Bonev, B., et al.: Ace: A fast, skillful learned global atmospheric model for climate prediction. arXiv preprint arXiv:2310.02074 (2023)
- [30] Mitchell, T.D.: Pattern scaling: an examination of the accuracy of the technique for describing future climates. *Climatic change* **60**(3), 217–242 (2003)
- [31] Tebaldi, C., Arblaster, J.M.: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change* **122**(3), 459–471 (2014) <https://doi.org/10.1007/s10584-013-1032-9>
- [32] Osborn, T.J., Wallace, C.J., Lowe, J.A., Bernie, D.: Performance of Pattern-Scaled Climate Projections under High-End Warming. Part I: Surface Air Temperature over Land (2018) <https://doi.org/10.1175/JCLI-D-17-0780.1>
- [33] Castruccio, S., McInerney, D.J., Stein, M.L., Crouch, F.L., Jacob, R.L., Moyer, E.J.: Statistical emulation of climate model projections based on precomputed gcm runs. *Journal of Climate* **27**(5), 1829–1844 (2014)
- [34] Freese, L.M., Fiore, A.M., Selin, N.E.: Spatially resolved temperature response functions to co2 emissions. *Authorea Preprints* (2024)
- [35] Beusch, L., Gudmundsson, L., Seneviratne, S.I.: 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)
- [36] Link, R., Snyder, A., Lynch, C., Hartin, C., Kravitz, B., Bond-Lamberty, B.: Fldgen v1. 0: an emulator with internal variability and space–time correlation for earth system models. *Geoscientific Model Development* **12**(4), 1477–1489 (2019)
- [37] Alexeeff, S.E., Nychka, D., Sain, S.R., Tebaldi, C.: Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments. *Climatic Change* **146**, 319–333 (2018)
- [38] Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls,

- G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., Roesch, C.: ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems* **14**(10), 2021–002954 (2022) <https://doi.org/10.1029/2021MS002954>
- [39] Kaltenborn, J., Lange, C., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., Rolnick, D.: Climateset: A large-scale climate model dataset for machine learning. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*, vol. 36, pp. 21757–21792. Curran Associates, Inc., ??? (2023)
- [40] Lütjens, B., Ferrari, R., Watson-Parris, D., Selin, N.: The impact of internal variability on benchmarking deep learning climate emulators. arXiv (2024). <https://doi.org/10.48550/arXiv.2408.05288> . <http://arxiv.org/abs/2408.05288>
- [41] Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., et al.: Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems* **14**(10), 2021–002954 (2022)
- [42] Quilcaille, Y., Gudmundsson, L., Beusch, L., Hauser, M., Seneviratne, S.I.: Showcasing mesmer-x: Spatially resolved emulation of annual maximum temperatures of earth system models. *Geophysical Research Letters* **49**(17), 2022–099012 (2022)
- [43] Tebaldi, C., Armbruster, A., Engler, H., Link, R.: Emulating climate extreme indices. *Environmental Research Letters* **15**(7), 074006 (2020)
- [44] Wikner, A., Harvey, J., Girvan, M., Hunt, B.R., Pomerance, A., Antonsen, T., Ott, E.: Stabilizing Machine Learning Prediction of Dynamics: Noise and Noise-inspired Regularization. arXiv (2022). <https://doi.org/10.48550/arXiv.2211.05262> . <http://arxiv.org/abs/2211.05262>
- [45] Barthel Sorensen, B., Charalampopoulos, A., Zhang, S., Harrop, B.E., Leung, L.R., Sapsis, T.P.: 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), 2023–004122 (2024) <https://doi.org/10.1029/2023MS004122>
- [46] Barthel Sorensen, B., Zepeda-Núñez, L., Lopez-Gomez, I., Wan, Z.Y., Carver, R., Sha, F., Sapsis, T.: A probabilistic framework for learning non-intrusive corrections to long-time climate simulations from short-time training data. arXiv (2024). <https://doi.org/10.48550/arXiv.2408.02688> . <http://arxiv.org/abs/2408.02688>
- [47] Zhang, S., Harrop, B., Leung, L.R., Charalampopoulos, A.-T., Barthel Sorensen,

- B., Xu, W., Sapsis, T.: A Machine Learning Bias Correction on Large-Scale Environment of High-Impact Weather Systems in E3SM Atmosphere Model. *Journal of Advances in Modeling Earth Systems* **16**(8), 2023–004138 (2024) <https://doi.org/10.1029/2023MS004138>
- [48] Wang, M., Souza, A., Ferrari, R., Sapsis, T.: Spatially-resolved emulation of climate extremes via machine learning stochastic models (2023)
- [49] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. arXiv (2021). <https://doi.org/10.48550/arXiv.2011.13456> . <http://arxiv.org/abs/2011.13456>
- [50] Bischoff, T., Deck, K.: Unpaired Downscaling of Fluid Flows with Diffusion Bridges. arXiv (2023). <https://doi.org/10.48550/arXiv.2305.01822> . <http://arxiv.org/abs/2305.01822>
- [51] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv (2015). <https://doi.org/10.48550/arXiv.1505.04597> . <http://arxiv.org/abs/1505.04597>
- [52] Wan, Z.Y., Baptista, R., Chen, Y.-f., Anderson, J., Boral, A., Sha, F., Zepeda-Núñez, L.: Debias Coarsely, Sample Conditionally: Statistical Downscaling through Optimal Transport and Probabilistic Diffusion Models. arXiv (2023). <https://doi.org/10.48550/arXiv.2305.15618> . <http://arxiv.org/abs/2305.15618>
- [53] Lopez-Gomez, I., Wan, Z.Y., Zepeda-Núñez, L., Schneider, T., Anderson, J., Sha, F.: Dynamical-generative downscaling of climate model ensembles. arXiv (2024). <https://doi.org/10.48550/arXiv.2410.01776> . <http://arxiv.org/abs/2410.01776>
- [54] Wan, Z.Y., Lopez-Gomez, I., Carver, R., Schneider, T., Anderson, J., Sha, F., Zepeda-Núñez, L.: Statistical Downscaling via High-Dimensional Distribution Matching with Generative Models. arXiv (2024). <https://doi.org/10.48550/arXiv.2412.08079> . <http://arxiv.org/abs/2412.08079>
- [55] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., *et al.*: The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**(730), 1999–2049 (2020)
- [56] Wheeler, M., Kiladis, G.N.: Convectively Coupled Equatorial Waves: Analysis of Clouds and Temperature in the Wavenumber–Frequency Domain (1999)
- [57] Kiladis, G.N., Wheeler, M.C., Haertel, P.T., Straub, K.H., Roundy, P.E.: Convectively coupled equatorial waves. *Reviews of Geophysics* **47**(2) (2009) <https://doi.org/10.1029/2008RG000266>

- [58] Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., Yiou, P., Zscheischler, J.: Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications* **14**(1), 2145 (2023) <https://doi.org/10.1038/s41467-023-37847-5>
- [59] Zscheischler, J., Westra, S., Hurk, B.J.J.M., Seneviratne, S.I., Ward, P.J., Pitman, A., AghaKouchak, A., Bresch, D.N., Leonard, M., Wahl, T., Zhang, X.: Future climate risk from compound events. *Nature Climate Change* **8**(6), 469–477 (2018) <https://doi.org/10.1038/s41558-018-0156-3>
- [60] Wehner, M., Gleckler, P., Lee, J.: 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 (2020)
- [61] Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control, Third edition edn. Prentice Hall, Englewood Cliffs, N.J. (1994). <http://www.gbv.de/dms/bowker/toc/9780130607744.pdf>
- [62] Storch, H.v., Langenberg, H., Feser, F.: A Spectral Nudging Technique for Dynamical Downscaling Purposes. *Monthly Weather Review* **128**(10), 3664–3673 (2000) [https://doi.org/10.1175/1520-0493\(2000\)128<3664:ASNTFD>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<3664:ASNTFD>2.0.CO;2)
- [63] Miguez-Macho, G., Stenchikov, G.L., Robock, A.: Regional Climate Simulations over North America: Interaction of Local Processes with Improved Large-Scale Flow. *Journal of Climate* **18**(8), 1227–1246 (2005) <https://doi.org/10.1175/JCLI3369.1>
- [64] Sun, J., Zhang, K., Wan, H., Ma, P.-L., Tang, Q., ZHANG, S.: Impact of Nudging Strategy on the Climate Representativeness and Hindcast Skill of Constrained EAMv1 Simulations. *Journal of Advances in Modeling Earth Systems* **11** (2019) <https://doi.org/10.1029/2019MS001831>
- [65] Huang, Z., Zhong, L., Ma, Y., Fu, Y.: Development and evaluation of spectral nudging strategy for the simulation of summer precipitation over the Tibetan Plateau using WRF (v4.0). *Geoscientific Model Development* **14**(5), 2827–2841 (2021) <https://doi.org/10.5194/gmd-14-2827-2021>
- [66] Batzolis, G., Stanczuk, J., Schönlieb, C.-B., Etmann, C.: Conditional image generation with score-based diffusion models. arXiv preprint arXiv:2111.13606 (2021)
- [67] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
- [68] Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)

- [69] Anderson, B.D.: Reverse-time diffusion equation models. *Stochastic Processes and their Applications* **12**(3), 313–326 (1982)
- [70] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv (2015). <https://doi.org/10.48550/arXiv.1512.03385> . <http://arxiv.org/abs/1512.03385>