

Score-based generative emulation of impact-relevant Earth system model outputs

Shahine Bouabid, Andre Nogueira Souza, Raffaele Ferrari

Department of Earth, Atmospheric, and Planetary Sciences

Massachusetts Institute of Technology, Cambridge, MA, United States

Abstract

Policy targets evolve faster than the Couple Model Intercomparison Project cycles, complicating adaptation and mitigation planning that must often contend with outdated projections. Climate model output emulators address this gap by offering inexpensive surrogates that can rapidly explore alternative futures while staying close to Earth System Model (ESM) behavior. We focus on emulators designed to provide inputs to impact models. Using monthly ESM fields of near-surface temperature, precipitation, relative humidity, and wind speed, we show that deep generative models have the potential to model jointly the distribution of variables relevant for impacts. The specific model we propose uses score-based diffusion on a spherical mesh and runs on a single mid-range graphical processing unit. We introduce a thorough suite of diagnostics to compare emulator outputs with their parent ESMs, including their probability densities, cross-variable correlations, time of emergence, or tail behavior. We evaluate performance across three distinct ESMs in both pre-industrial and forced regimes. The results show that the emulator produces distributions that closely match the ESM outputs and captures key forced responses. They also reveal important failure cases, notably for variables with a strong regime shift in the seasonal cycle. Although not a perfect match to the ESM, the inaccuracies of the emulator are small relative to the scale of internal variability in ESM projections. We therefore argue that it shows potential to be useful in supporting impact assessment. We discuss priorities for future development toward daily resolution, finer spatial scales, and bias-aware training.

1 Introduction

Planning adaptation measures and evaluating mitigation choices depends on assessing the projected impacts of climate change under different future scenarios. The computational cost of Earth System Models (ESMs) has prevented them from keeping pace with updates to these scenarios. For example, although the set of scenarios for the Phase 7 Coupled Model Intercomparison Project (CMIP) is expected to be released this year (van Vuuren et al., 2025), the corresponding simulations will not be available for another few years. Unless the handful of off-the-shelf scenarios already run by ESMs are deemed sufficient to cover future alternative pathways of emissions, fast ESM surrogates, called *climate model output emulators* can be extremely valuable (Jones et al., 2024; Tebaldi et al., 2025). Typically, these emulators do not attempt to reproduce the dynamics of the climate system. Rather, for a subset of variables deemed relevant in applications like impact modeling, they aim to reproduce statistics of the *distribution of climate model outputs*. The goal is to learn statistical features that best characterize the distribution produced by an ensemble of simulations from the reference ESM. This is a crucial nuance, since under this rationale, an emulator only needs to provide a computationally efficient generator that retains the same statistical behavior as the original ESM.

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) (Warszawski et al., 2014) is a concerted effort to collect, harmonize, and distribute data relevant for impact assessment across sectors and future scenarios. It presents a selection of 11 atmospheric climate model output variables that are most consequential for sectors such as agriculture, energy demand, or forestry. This selection defines a priority list for any emulator intended to support physical-risk assessment. The past decade has seen a profusion

of emulator designs which, collectively, have demonstrated skillful reproduction of these variables at the regional scale, with orders-of-magnitude lower computational cost than running new ESM experiments. They include emulators that target selected low-order statistics of climate model output variables such as traditional pattern scaling (e.g. Santer et al., 1990; Tebaldi and Arblaster, 2014; Herger et al., 2015) and impulse-response models (e.g. Lucarini et al., 2017; Freese et al., 2024; Womack et al., 2025a; Winkler and Sierra, 2025; Sandstad et al., 2025), but also emulators targeting the full probability distribution (e.g. Beusch et al., 2020; Nath et al., 2022; Geogdzhayev et al., 2024; Mathison et al., 2025). In theory, this last category supersedes emulating selected statistics, since any statistic can be derived from the distribution.

A survey of the existing “*anthropogenic forcing* → *climate*” emulators shows that most focus on just one or two variables at a time (e.g. Snyder et al., 2019; Nath et al., 2024; Schöngart et al., 2024; Tebaldi et al., 2025), often making parametric assumptions for the variables distributions based on generalized linear models or Gaussian processes (e.g. Castruccio et al., 2014; Link et al., 2019; Goodwin et al., 2020; Quilcaillie et al., 2022; Bouabid et al., 2024). This has proven to be sufficient, especially when the variables of interest are sufficiently aggregated (spatially or temporally) to be well captured by parametric forms; for example, a lognormal distribution for precipitation or a Gaussian one for temperature. However, these approaches become prone to misspecification and are increasingly cumbersome to implement when the objective is to emulate dozens of variables jointly at high resolution, which is needed to assess high-impact climate damages that are often associated with compound risks due to co-occurring events (Zscheischler et al., 2020; Mathison et al., 2023). Capturing joint dependencies requires emulating not only individual variables, but also spatio-temporal and cross-variable correlations, and the number of correlations grows quadratically with the number of variables and spatial resolution. Alternative “stitching” strategies, which connect time slices of existing ESM runs, have been proposed and can provide most atmospheric variables jointly (Tebaldi et al., 2022; Byers et al., 2025). However, they rely on having large volumes of ESM output archives at hand, and cannot generate new realizations.

Deep generative models offer an appealing alternative: they learn the full joint distribution directly, are well suited to high-dimensional structured data, and provide a compressed representation of ESM outputs. Recent studies have already demonstrated their ability to emulate complete atmospheric states on multiple pressure levels when forced by a prescribed sea surface temperature (Watt-Meyer et al., 2023; Brenowitz et al., 2025). While the design of these emulators is primarily focused on accelerating atmospheric simulation for scientific discovery, their success motivates extending them to the impact-focused setting considered here.

In this work, we propose a score-based diffusion emulator that learns the high-dimensional joint probability distribution of monthly climate model fields conditioned on the global mean surface temperature (GMST) anomaly. The emulator produces data on the ESM’s native grid while carrying out its computations on an equal-area spherical HEALPix mesh similar to the setup of Brenowitz et al. (2025). A pattern scaling step maps GMST onto regional mean temperature anomalies, providing spatial structure as conditioning information to the emulator. The proposed model is purposefully designed to be lightweight and run efficiently on a single mid-range graphical processing unit (GPU). We evaluate the model on three distinct ESMs and for a selection of four surface variables relevant for impact assessment: surface temperature, precipitation, relative humidity, and wind speed. Although we restrict ourselves to these four fields, the emulator can accommodate additional variables without modifications to its design. We find that the emulator learns to sample from the joint distribution of these variables at the climate model’s native spatial resolution. It reproduces internal variability and response to forcing. While many impact studies require daily or finer resolution, the current emulator is already useful for applications relying on long-term averages, and provides a foundation for higher-resolution approaches as well as for pairing with downscaling techniques.

The paper is organized as follows. Section 2 describes the data, the emulator design, and introduces evaluation metrics. Section 3 presents the results for the emulation of unforced simulations, and Section 4 evaluates the forced response of the emulator. Section 5 discusses limitations and outlooks and Section 6 concludes.

2 Data and Methods

Section 2.1 describes the CMIP6 simulations used in training and evaluation. Section 2.2 sets the notation and formal objective of emulation. Sections 2.3, 2.4, and 2.5 outline, respectively, the use of pattern scaling as a conditioning field, the score-based diffusion framework, and the neural network architecture. Finally, Section 2.6 introduces the metrics used to assess emulator performance.

2.1 Data

We select variables informed by climate forcing priorities from the ISIMIP protocol, starting with near-surface air temperature, precipitation, near-surface relative humidity, and near-surface wind speed. Other variables (e.g., specific humidity, daily maximum and minimum temperature, downwelling longwave radiation) can be incorporated without changes to the emulator design, but we choose in this work to focus on the four listed above for clarity of presentation. We retain the native spatial resolution of the data ($\sim 1\text{-}2^\circ$) and use monthly averages to keep data volumes manageable, although the same approach can extend to daily data. Each variable is pre-processed to be an anomaly relative to the climatology of its model’s pre-industrial control run (piControl); unless stated otherwise, every variable in this work represents a positive or negative departure from that baseline.

Our focus is on climate model runs over scenarios from the CMIP6 ScenarioMIP protocol (O’Neill et al., 2016) since they constitute the most widely used projections for impact assessment. The associated Shared Socioeconomic Pathways (SSPs) (Riahi et al., 2017) shown in Figure 1, specify different 2014-2100 trajectories for long- and short-lived emissions. Each pathway produces a distinct global mean surface temperature (GMST) trajectory, ranging from Paris-aligned to strongly overshooting. To train the emulator, we use ESM outputs from SSP1-2.6 and SSP5-8.5, and complement these with historical and piControl experiments from the core CMIP6 protocol (Eyring et al., 2016). This constitutes a diverse training set of different emulation regimes, including unforced outputs, strong curbing of emissions, and an increase in emissions. When presenting experiments, we will explicitly mention the scenarios used to evaluate the emulator, as they can vary for different diagnostics.

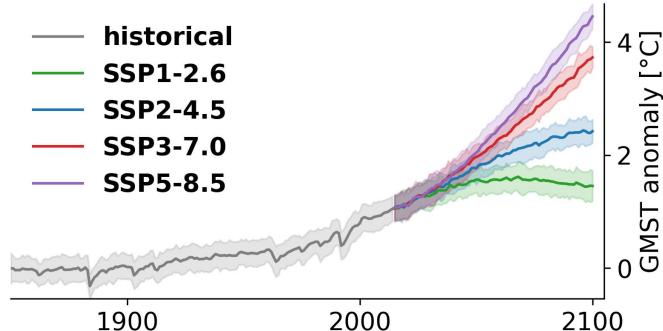


Figure 1: Time series of annual GMST anomalies from the MPI-ESM2-1-LR ScenarioMIP simulations (Schupfner et al., 2021). Anomalies are relative to the climatology of the model’s piControl run. The solid line shows the ensemble mean across 50 members, and shading indicates two standard deviations.

A large ensemble is necessary for the generative model to learn to disentangle the model’s internal variability from the anthropogenic signal (Lütjens et al., 2024). We use large-ensemble simulations from three CMIP6 models — MPI-ESM1-2-LR, MIROC6, and ACCESS-ESM1-5 — each with over 40 members covering historical and SSP scenarios (Wieners et al., 2019; Schupfner et al., 2021; Tatebe and Watanabe, 2018; Shiogama et al., 2023; Ziehn et al., 2019a,b). Each emulator is trained on data from a single ESM; data from multiple models supports inter-model comparison, not cross-model training. Further details on the CMIP6 simulations used are provided in the supplementary.

We draw the reader’s attention to the fact that in the main text, we chose to primarily present results from emulating the MPI-ESM1-2-LR model. Since many of the same conclusions about the emulator’s skill hold across all three ESMs, there is limited value in reproducing every diagnostic in full. We therefore focus on one model to highlight the methodological insights that are largely invariant to the choice of ESM. However, complete figures and diagnostics for all three models are provided in the supplementary material, and inter-model differences are discussed in the main text where relevant.

2.2 Problem setup

Let p_{ESM} denote the ensemble distribution produced by realizations of a reference ESM given a forcing scenario, that is, we consider that individual runs from the ESM correspond to samples drawn from this distribution. We denote x_t a monthly averaged climate model output anomaly at time t , where t indexes a year and a calendar month. It can be thought of as a state, relative to the pre-industrial control climatology, that the climate model could realize at time t , given a forcing trajectory,

$$x_t \sim p_{ESM}(x \mid \text{Forcing}_{\leq t}). \quad (1)$$

In technical terms, p_{ESM} represents the distribution of the forward model induced by the ESM for slightly perturbed initial conditions, given a prescribed forcing. Since our focus is on the development of emulators useful for impact assessment, for practical purposes, x_t can be thought of as multiple concatenated spatial maps of surface variables selected by their relevance for risk assessment.

Because regional climate outcomes tend to display a significant correlation with the GMST anomaly ΔT_t , numerous existing climate model output emulators have adopted it as their main driver (Nath et al., 2022; Quilcaillie et al., 2023; Schöngart et al., 2024; Nath et al., 2024; Geogdzhayev et al., 2024; Wang et al., 2024). This dependency is convenient because the GMST response to changes in emissions can be easily derived from simple climate models — which offer a reliable lower-order approximation of the GMST response to forcing from long- and short-lived forcers (Meinshausen et al., 2011; Gasser et al., 2017; Leach et al., 2021) — or directly from integrated assessment models. This makes emulators ready for the full *emissions → regional climate* pipeline. Formally, the underlying assumption in this shared design for many emulators is that

$$p_{ESM}(x \mid \text{Forcing}_{\leq t}) \approx p_{ESM}(x \mid \Delta T_t). \quad (2)$$

This is, of course, an oversimplifying assumption, which dismisses the effect of forcers’ regionality, the different types of forcing agent, or the climate system memory. Nonetheless, evidence from emulator development literature suggests that replacing the forcing with ΔT_t already captures significant projected variability under plausible future scenarios. This is also expected to align with the requirements of the upcoming FASTMIP emulator intercomparison exercise (Seneviratne et al., 2024). We therefore proceed with this assumption in the remainder of the manuscript and revisit its implications in the discussion. In this context, our objective can be framed as finding a surrogate conditional distribution $q(x \mid \Delta T_t)$, which we can easily sample from, and such that

$$d(q, p_{ESM}) \leq \epsilon, \quad (3)$$

where d is a measure of the discrepancy between probability distributions and $\epsilon > 0$ is a selected threshold. The preferred notions of distance will depend on which statistical features are assessed (e.g., mean, variance, tails), and we discuss in Section 2.6 different choices for d that we use to evaluate the emulator.

2.3 Pattern scaling conditioning

We introduce spatial information in the conditioning signal based on a linear relationship between the global and local mean surface temperature anomaly, known as *pattern scaling*. Pattern scaling is an emulation technique that assumes the forced regional climate response scales linearly with an indicator of the global

climate response (Santer et al., 1990). For surface temperature, such a fixed warming pattern consistently emerges across CMIP6 simulations (Tebaldi and Arblaster, 2014; Osborn et al., 2018) and is supported by energy balance arguments for medium-to-high emission scenarios (Giani et al., 2024).

While other surface variables tend to conform less to this linear approximation (Kravitz et al., 2017), it is recognized that pattern scaling captures the slow large-scale regional changes simulated by ESMs, which matter on a centennial time scale (Tebaldi and Arblaster, 2014; Lee et al., 2021). Therein, owing to its simplicity and competitiveness in benchmarks (Watson-Parris et al., 2022; Lütjens et al., 2024), pattern scaling remains today a recurrent component in the design of emulators aiming to support impact assessment questions (Osborn et al., 2016; Link et al., 2019; Beusch et al., 2020; Tebaldi et al., 2020; Nath et al., 2022; Quilcaille et al., 2022, 2023; Schöngart et al., 2024; Mathison et al., 2025).

Pattern scaling does not represent internal variability and, therefore, cannot serve as an emulator of the climate model output distributions. However, it offers an inexpensive way to augment the conditioning signal with spatial structure. We propose to leverage this and condition our emulator on the projected pattern scaling for regional temperatures rather than ΔT_t . Using the training data, we fit a separate pattern scaling model for each month $P_t^{\text{month}} = \beta_1^{\text{month}} \Delta T_t + \beta_0^{\text{month}}$, where $\beta_1^{\text{month}}, \beta_0^{\text{month}}$ are fixed spatial patterns of warming for a given month. At each time step t , the scalar ΔT_t is thus mapped to a spatially resolved projection P_t^{month} which is passed to the emulator, providing it with structured information about the forced regional response in surface temperature.

In the rest of the paper, we continue to write the conditioning variable as ΔT_t , interpreting pattern scaling as a design step in constructing q . This emphasizes that GMST remains the fundamental driver: to emulate a new scenario, one provides the emulator with the associated GMST trajectory, which is then mapped through a pattern scaling model fitted to the training data.

2.4 Score-based diffusion emulator

We model q using a score-based diffusion framework (Ho et al., 2020; Song et al., 2021), motivated by its suitability for high-dimensional structured data and its demonstrated effectiveness in climate data generation (Bassetti et al., 2024; Mardani et al., 2025; Hess et al., 2025; Brenowitz et al., 2025). However, we emphasize that alternative approaches to learn a surrogate generative process to the climate model — such as variational inference, particle filters, or flow-based generative models — could also be considered and may be better suited to certain applications.

Score-based diffusion models rely on gradually transforming samples from a simple Gaussian distribution into samples from a complex distribution of interest, in our case p_{ESM} . Learning this transformation involves two steps: first, p_{ESM} is gradually mollified until it becomes indistinguishable from a Gaussian distribution; second, a neural network is trained to reverse this process. In practice, we carry out the first step by adding Gaussian noise to climate model outputs $\tilde{x}_t = x_t + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with variance σ^2 . This can be viewed as sampling from a “smoothed” version of the climate model output distribution defined by

$$p_{\text{ESM}}^\sigma = p_{\text{ESM}} * \mathcal{N}(0, \sigma^2), \quad (4)$$

where $*$ denotes convolution. By repeating the convolution with increasing variance levels, the distribution progressively becomes smoother, ultimately becoming indistinguishable from a Gaussian distribution at a sufficiently large σ_{\max} . Then, we train a neural network f_θ to reverse this transformation. We use least-squares regression, predicting the original climate model output from its transformed Gaussian versions at different variance levels σ , by minimizing the loss

$$\mathbb{E} \|x_t - f_\theta(\tilde{x}_t, \sigma, \Delta T_t)\|^2. \quad (5)$$

Such direct regression will generally fail to recover the original climate model output, especially when the injected Gaussian noise is large. However, the key to the success of score-based generative modeling is that the neural network f_θ does not attempt to reverse the transformation outright. Rather, it identifies a vector that nudges transformed samples toward regions of higher density in the climate model output

distribution. This vector is known as the *score function* of the smoothed climate model output distribution, and is defined as the gradient of its log-density $\nabla \log p_{ESM}^\sigma$. Tweedie's formula (Efron, 2011) provides a useful explanation: regressing away Gaussian noise is mathematically equivalent to moving the transformed sample in the direction of the score,

$$\mathbb{E}[x_t | \tilde{x}_t, \sigma, \Delta T_t] = \tilde{x}_t + \sigma^2 \nabla \log p_{ESM}^\sigma(\tilde{x}_t | \Delta T_t). \quad (6)$$

By the property of least-square regression, the trained neural network f_θ constitutes an estimator of the conditional expectation in Eq. 6, and allows for estimating the score function:

$$\nabla \log q^\sigma(\tilde{x}_t | \Delta T_t) = \frac{f_\theta(\tilde{x}_t, \sigma, \Delta T_t) - \tilde{x}_t}{\sigma^2}, \quad (7)$$

where q^σ denotes our learned approximation of the smoothed distribution. This score estimate provides a direction in which we can iteratively displace the noised samples to reverse the transformation process.

This allows us to emulate the climate model output distribution by starting from a new Gaussian sample $\tilde{x} \sim \mathcal{N}(0, \sigma_{\max}^2)$, and evolving it in the direction of the score to produce a sample from the estimated climate model output distribution. Formally, the emulated climate model output distribution q is then defined as q^σ for $\sigma \rightarrow 0$. However, what this really offers is a procedure to draw samples from q , as we never explicitly access the distribution. The choice of schedule for noise levels and the integration scheme used to follow the direction of the score are important design parameters of this sampling algorithm. We detail our technical choices in the supplementary.

2.5 HEALPix UNet architecture

However compelling the score-based diffusion framework can be, diffusion models in reality owe their effectiveness to advances in neural network architectures for structured data, and the computational solutions to support them. For spatially structured data on a square, the undisputed choice of architecture for f_θ is the UNet (Ronneberger et al., 2015). Its effectiveness can be understood through its multi-resolution structure, which implicitly performs transformations analogous to processing signals in a hierarchical wavelet basis, separating information at different spatial scales (Williams et al., 2023). This, however, does not support data on a sphere.

To respect Earth's geometry, we implement a variation of the classical UNet architecture that operates on an equal-area HEALPix mesh (Gorski et al., 2005). This mesh first splits the domain into 12 base diamonds that are exactly equal in area. Each refinement level subdivides every diamond into 4 self-similar diamonds, so the equal-area property is preserved. This avoids the pole-to-equator cell-size distortion and longitude-wrapping discontinuities that arise in equiangular grids, while preserving a hierarchical neighborhood structure ideal for multi-resolution processing. This choice of mesh was also made by Karlbauer et al. (2024) and Brenowitz et al. (2025), and therefore our architectures bear strong similarities.

We expect that impact assessment workflows will generally assume climate model outputs on a regular lat-lon grid, and therefore keep the ESM's native gridding for inputs and outputs. A pair of lightweight bipartite graph neural networks maps fields from the equiangular grid to HEALPix for processing and back again, allowing the model to satisfy spherical fidelity internally while being compatible with climate model output data on the ESM's native grid. Details on specific design choices are provided in the supplementary.

To ensure the emulator is accessible, we designed the architecture to be lightweight. By focusing on a small set of variables relevant for impact assessment, the network is kept compact, with approximately 10 million parameters. This is an order of magnitude smaller than cBottle (Brenowitz et al., 2025), which naturally requires more capacity to emulate entire atmospheric states, comprising 44 variables. The total file size of our emulator is 50 MB and can be downloaded from our repository. This lightweight design allows the emulator to run on a single mid-range GPU; generating one sample takes about 1 second on a T4 (mid-range) and about 0.13 seconds on an H100 (state-of-the-art high-end). Because samples are generated through independent processes, this emulation can be parallelized to accelerate large ensemble generation.

2.6 Distances between distributions

Our goal is for the emulator distribution q to closely approximate the ESM output distribution p_{ESM} under a chosen notion of distance $d(q, p_{ESM})$. We describe below the notions of distances that we use to diagnose different aspects of the discrepancy between the distributions.

Distance in individual statistics Throughout this work, we often define d in terms of error in individual statistics — such as means, variances, correlations, quantiles — as it provides important insights on how two distributions can differ. Biases in moments show errors in the distribution’s location, spread, or skewness; biases in correlations reveal mismatches in dependence structures; and biases in quantiles, often more robust to outliers than moments, are particularly useful to assess differences in the distribution tails. In practice, comparing the actual statistics directly can often make interpretations more natural than comparing errors. Indeed, visual inspection can often reveal sufficient information to judge whether emulator and ESM statistics align, and we therefore rely on it for multiple diagnoses throughout the following sections.

When reporting explicit distances is useful, we privilege defining d in terms of error relative to a reference quantity. This makes the reported values easier to interpret and compare across variables. We use two strategies: relative errors and error-to-noise ratios. Relative errors express the error as a percentage of deviation from the ESM statistic, where an acceptable threshold is chosen by judgment (e.g. 10%). The error-to-noise ratio is defined by dividing the error by the ESM’s internal variability. This is particularly relevant for evaluating errors in means: if the bias is smaller than the natural variability, it is unlikely to be detectable in individual realizations. This naturally sets an acceptable threshold of $d(q, p_{ESM}) \leq 1$, with larger values indicating a meaningful bias.

Earth Mover Distance Beyond individual statistics, we argue that a metric to compare distributions in their entirety should be given a special place. Many useful measures of discrepancy between probability distributions exist, such as the Kullback-Leibler divergence, the total variation distance, or the maximum mean discrepancy; however, measuring discrepancies between full joint probability distributions is computationally prohibitive, and we instead focus on one-dimensional probability distributions. We propose here to use the Earth Mover Distance (EMD), also known as the 1-Wasserstein distance, because we find it to have a more intuitive interpretation when dealing with physical quantities.

Imagine distributions as binned histograms of mass as a function of horizontal location. Then the work to displace a unit of mass against friction is proportional to its weight times the distance. The EMD between p_{ESM} and q corresponds to the minimal amount of work needed to rearrange the masses so that they exactly overlay. For a single grid cell gc and variable var , the distributions are one-dimensional, and the optimal rearrangement simply pairs quantiles together (Santambrogio, 2015). Thus, if F_{ESM} and F_q are the cumulative distribution functions for the ESM and the emulator, the work needed to rearrange quantiles is given by

$$\text{EMD}(gc, \text{var}) = \int |x_{gc, \text{var}} - F_{ESM}^{-1} \circ F_q(x_{gc, \text{var}})| q(x_{gc, \text{var}}) dx_{gc, \text{var}}. \quad (8)$$

The term $|x - F_{ESM}^{-1} \circ F_q(x)|$ represents the distance a unit of mass from q must be shifted when p_{ESM} is held fixed, and $q(x) dx$ gives the amount of mass being displaced. From this expression, we see that the EMD is equally sensitive to small shifts of large bulks of probability or large shifts of small probability mass, since they represent the same amount of work.

Because probability masses are dimensionless and sum to one, Eq. 8 also implies that the EMD inherits the physical units and scale of the underlying variable, which makes it more easily interpretable. For example, if $p(T)$ is a distribution of temperatures, then the EMD to its shifted version $p(T + \delta T)$ exactly equals the shift in temperatures $|\delta T|$. Inspired by this property, we propose to assess the significance of the

misalignment between q and p_{ESM} by comparing their EMD to the typical internal variability at each grid cell. We introduce an EMD-to-noise ratio

$$\text{EMD-to-noise}(\text{gc}, \text{var}) = \frac{\text{EMD}(\text{gc}, \text{var})}{\sigma_{\text{gc}, \text{var}}}, \quad (9)$$

where the noise $\sigma_{\text{gc}, \text{var}}$ is taken as the internal variability of the ESM. The ratio quantifies *how many “standard deviation shifts equivalents” are required to realign the distributions*. Figure 2 compares emulated distributions with their corresponding ESM outputs, together with the associated EMD-to-noise ratios. The apparent degree of misalignment between the distributions increases rapidly with the ratios. As a practical heuristic, we regard EMD-to-noise ratios below 0.5 as indicative that the emulator reproduces the ESM distribution within internal variability, whereas larger values are representative of a significant mismatch in distribution.

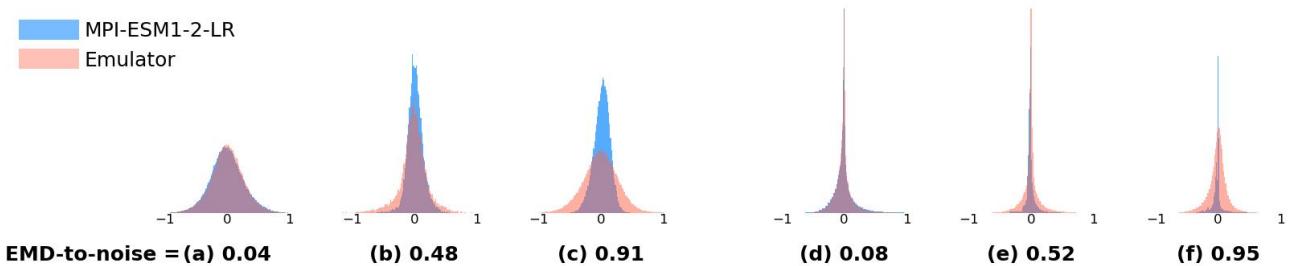


Figure 2: Distributions from the MPI-ESM1-2-LR piControl run and the emulator with $\Delta T_t = 0$, with their EMD-to-noise ratios. Distributions are normalized and shown on a common x-axis. Panels: (a) November wind speed, Mediterranean; (b) October relative humidity, western Africa; (c) July temperature, Arctic; (d) February precipitation, tropical Pacific; (e) July precipitation, Mediterranean; (f) December precipitation, central Africa.

3 Emulation of unforced simulations

We first evaluate if the emulator can represent the monthly internal variability of the climate model under stationary conditions without external forcing. Formally, this amounts to assessing whether in the absence of any forcing we have

$$q(x | \Delta T_t = 0) \approx p_{ESM}(x | \text{Forcing}_{\leq t} = 0). \quad (10)$$

To evaluate this, we train for each ESM an emulator. We then generate a large ensemble of emulated monthly samples in the absence of forcing and compare its statistics to the statistics of the climate model’s piControl run. For a given month, we generate an ensemble of 1000 samples by first drawing $\Delta T \sim \mathcal{N}(0, \sigma_{PI}^2)$, where σ_{PI} is the estimated GMST standard deviation from the piControl run, and then sampling $x \sim q(x | \Delta T)$.

3.1 Grid-cell distributions of variability

We compare moment estimates using monthly fields from all calendar months altogether. Figure 3 shows that for all the emulated variables, the first three moments (mean, variance, skewness) of their emulated variability distribution provide a good match to the moments of the original climate model piControl run. In particular, the standard deviations align remarkably well. The emulator introduces some bias in the mean for all emulated variables. However, it generally remains minimal in magnitude relative to the standard deviation; that is, in individual realizations, this bias in the mean is not detectable and buried under variability. Skewness tends to be underestimated in some regions, but its spatial pattern and overall magnitude remain consistent with the ESM.

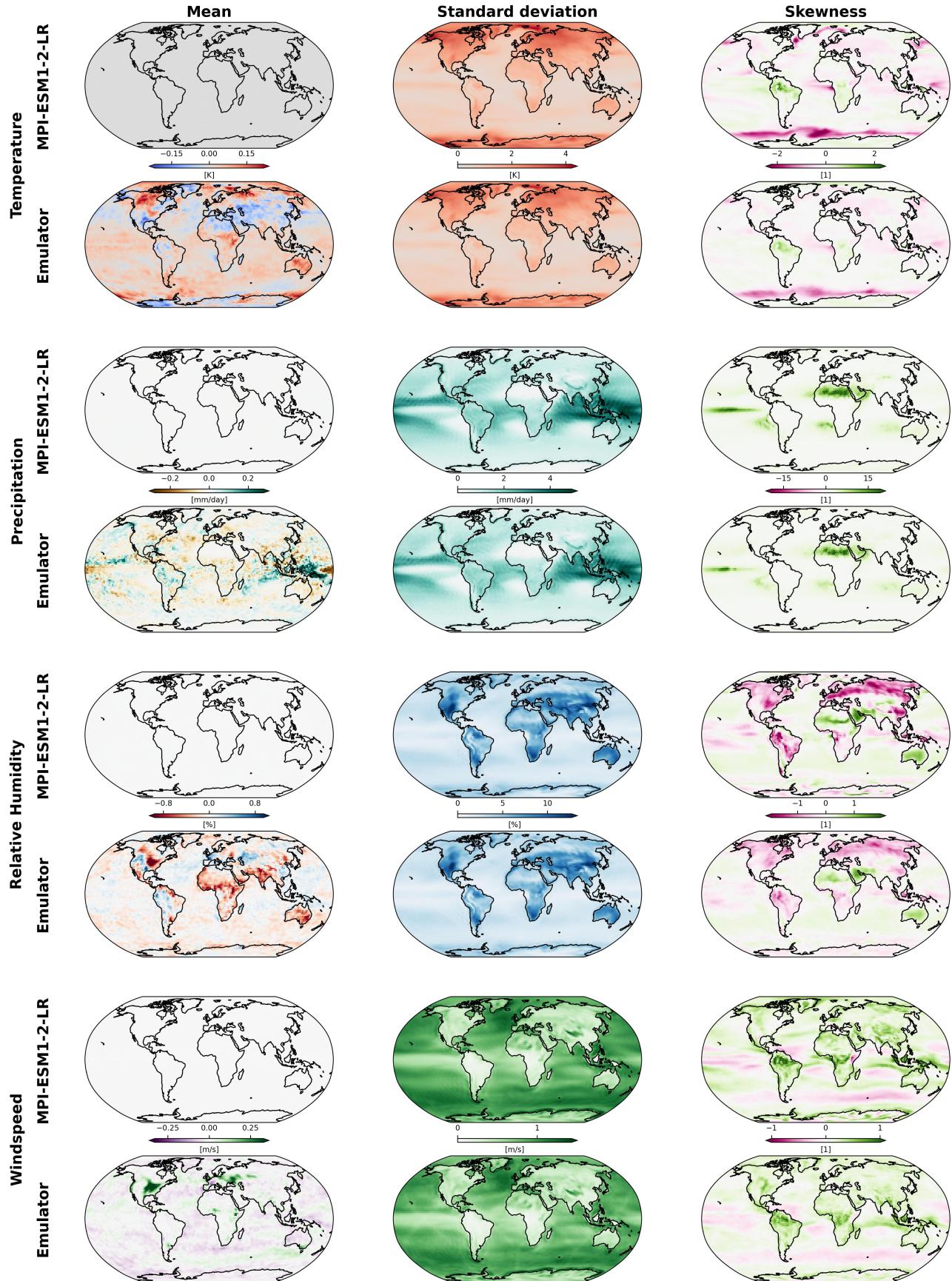


Figure 3: Comparison of moments from the MPI-ESM1-2-LR piControl run and the emulated distribution without forcing. Moments are estimated using anomaly data from all calendar months. Corresponding results for MIROC6 and ACCESS-ESM1-5 are provided in the supplementary.

Some grid cells in the central US, as well as parts of Ukraine and western Russia, show large biases in wind speed and relative humidity compared to the magnitude of variability in these regions. Similar bias patterns appear in the emulations of the two other ESMs we considered. Figure 4 zooms in on this bias, focusing on the central US. It shows that the bias pattern matches the regional anomalies in wind speed and relative humidity simulated by the ESM during the early historical period (1850–1900). At low GMST anomaly levels, this period accounts for roughly half of the training data. Because GMST anomalies in the early historical and piControl runs overlap substantially, this suggests that the generative model is overfitting to spatial anomaly patterns from the historical simulations. Such overfitting is a known tendency of diffusion models, especially for large modes of variability that are denoised at high noise levels (Brenowitz et al., 2025). In addition, since ESM runs over the historical period are generally tuned to match observational records more closely, it is likely that this anomaly appears across many historical realizations and therefore becomes over-represented in the training data at these GMST levels.

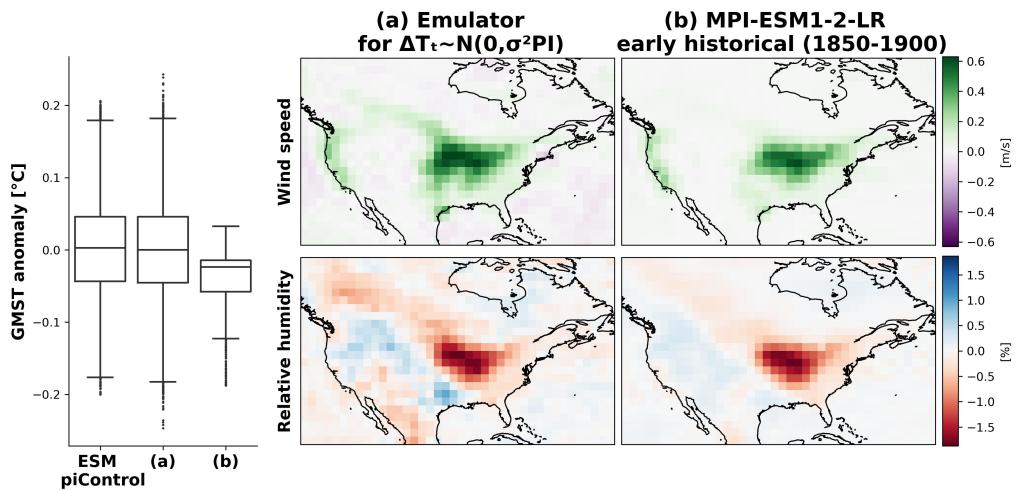


Figure 4: Left: Boxplots of GMST anomalies for the MPI-ESM1-2-LR piControl run, (a) samples drawn from $\mathcal{N}(0, \sigma_{PI}^2)$ used to generate the emulated ensemble, and (b) the MPI-ESM1-2-LR early historical period (1850–1900). Right: Maps of mean wind speed (top) and relative humidity (bottom) anomalies from the MPI-ESM1-2-LR piControl run, the emulated ensemble, and the MPI-ESM1-2-LR early historical period.

Because the emulator offers a sampling procedure, any higher-order moment can be estimated and, in contrast to parametric distributions, the score-based generative approach provides large flexibility on the shapes the distributions can take. This is why, for example, we are able to obtain satisfactory estimates of skewness jointly for both relative humidity and precipitation, even though they strongly differ in the shapes of their distributions.

Moving beyond individual moments of the emulated monthly variability, we now turn to the tougher test of comparing the entire emulated distribution to the climate model distribution. For each season, variable, and grid cell, we compute the EMD-to-noise ratio between the emulated distribution and the ESM piControl distribution and report it in Figure 5. Overall, we find that shifts much smaller than the regional magnitude of variability would produce the same work as that required to match the distributions at each grid cell. This means that, while discrepancies between the climate model output and the emulated distributions exist, they show in most regions a good agreement relative to the scale of internal variability in the climate model.

This is not true everywhere, and the maps identify regions where the approach fails, with an anomalously high EMD-to-noise ratio. The overfitting biases in wind speed and relative humidity in the central US and eastern Europe are clearly visible. For the MIROC6 emulator (see supplementary), several sharply delimited regions show high EMD-to-noise ratios in wind speed, which we attribute to a possible artifact in the model’s wind speed computation.

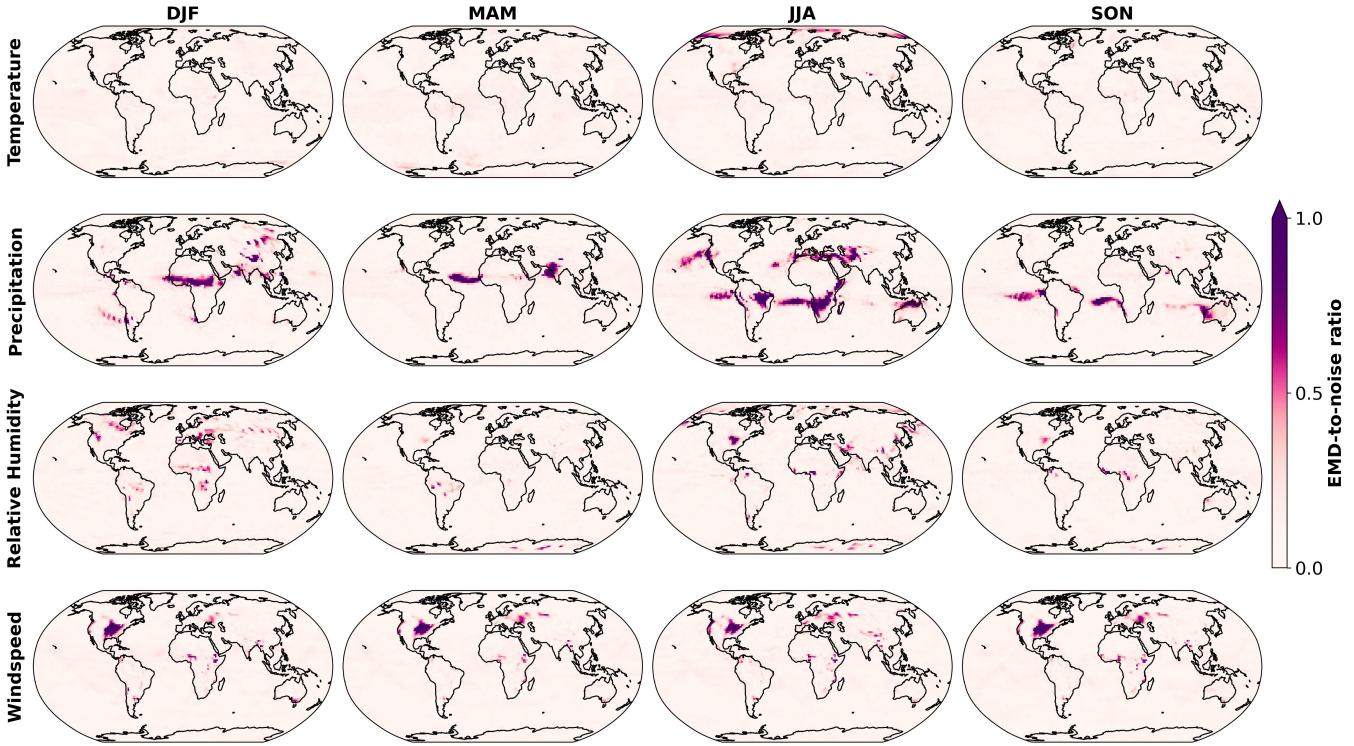


Figure 5: Spatial maps of grid-cell EMD-to-noise ratios between the MPI-ESM1-2-LR piControl run and the emulated large ensemble without external forcing. The reference standard deviation at each grid-cell $\sigma_{\text{gc},\text{var}}$ is floored to 0.1 to prevent spurious inflation of the ratio.

For the three ESMs considered, we find that over a small number of grid cells, in particular over land, the emulator tends to generate “smoothed” distributions that exaggerate internal variability. Examples include the Arctic summer surface temperature, which corresponds to the overestimated monthly variability shown in panel (c) of Figure 2, and spring precipitation over India shown in the supplementary. This phenomenon is most pronounced for the relative humidity if ACCESS-ESM1-5, where the emulator struggles with narrow distributions and smooths them excessively (see supplementary). While the black-box nature of neural networks makes it difficult to identify the cause of these discrepancies, literature on diffusion models shows they learn smooth score functions that closely resemble the scores of Gaussian mixture models (Wang and Vastola, 2024; Aithal et al., 2024). This provides a possible explanation for why distributions are smoothed out. Empirically, we find that increasing the number of parameters of the neural network used to estimate the score function largely mitigates this problem.

A more persistent challenge arises for precipitation across ESMs, where large EMD-to-noise ratios appear across seasons despite attempts at refining the diffusion model. We find that these discrepancies consistently occur in regions marked by some months with very low rainfall followed by months receiving precipitation, e.g., regions that experience seasonal migration of the Intertropical Convergence Zone. Figure 6 shows the 95th percentile of monthly precipitation in the MPI-ESM1-2-LR piControl, taken as an indicator of whether a given month belongs to a dry or precipitating regime. Taking the max–min ratio of this indicator across months identifies regions with pronounced seasonal contrasts, which align closely with those showing large EMD-to-noise ratios. This suggests that the diffusion model struggles to capture important seasonal distributional shifts in precipitation. The overestimation of variability in Arctic summer temperatures discussed earlier may reflect the same difficulty in representing important seasonal changes. A possible reason is that pattern scaling alone may not provide sufficient seasonal control, which can make it challenging for the diffusion model to generate multimodal distributions from a unimodal Gaussian (Xu et al., 2024). This may be even harder if the modes are not equally represented (Sehwag et al., 2022; Qin et al., 2023).

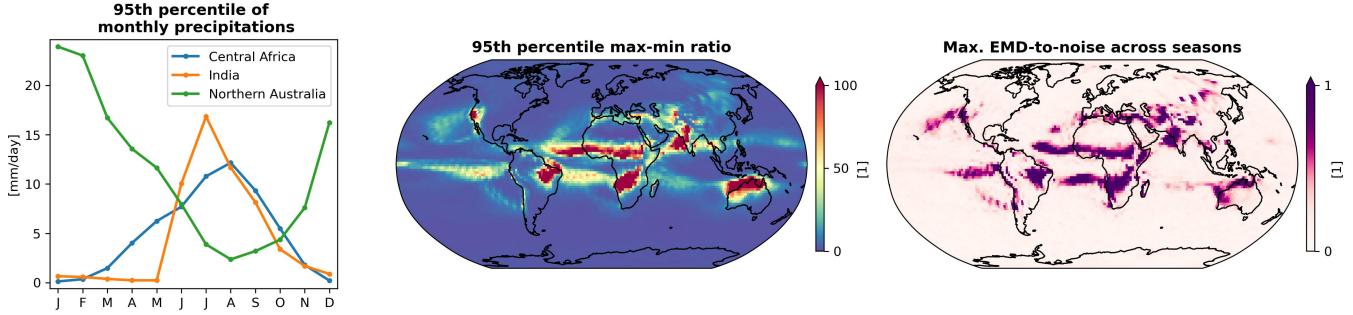


Figure 6: (Left) 95th percentile of monthly precipitation in the MPI-ESM1-2-LR piControl run for three regions where the emulator fails to match the ESM distribution; (Center) Spatial map of the max–min ratio of the monthly 95th percentile at each grid cell, highlighting regions with contrasts between months in dry or precipitating regimes; (Right) Max EMD-to-noise ratio across seasons.

3.2 Cross-correlation across variables and grid-cells

An important argument in favor of deep generative models is their promise to generate samples of multiple variables at every grid cell jointly. Unlike having separate emulators for each variable, such a joint model should preserve cross-correlations across variables and space, as in the ESM output data. The emulator implicitly learns to represent correlations between all grid cells and variables simultaneously. For a climate field with 2° horizontal resolution and four variables, this amounts to more than 2.5 billion correlation terms. Examining every individual correlation is overwhelming; therefore, we focus on two diagnostics: cross-variable correlations at the grid-cell level, which speak to compound consistency, and spatial correlations within each variable, which highlight the scales of variability captured.

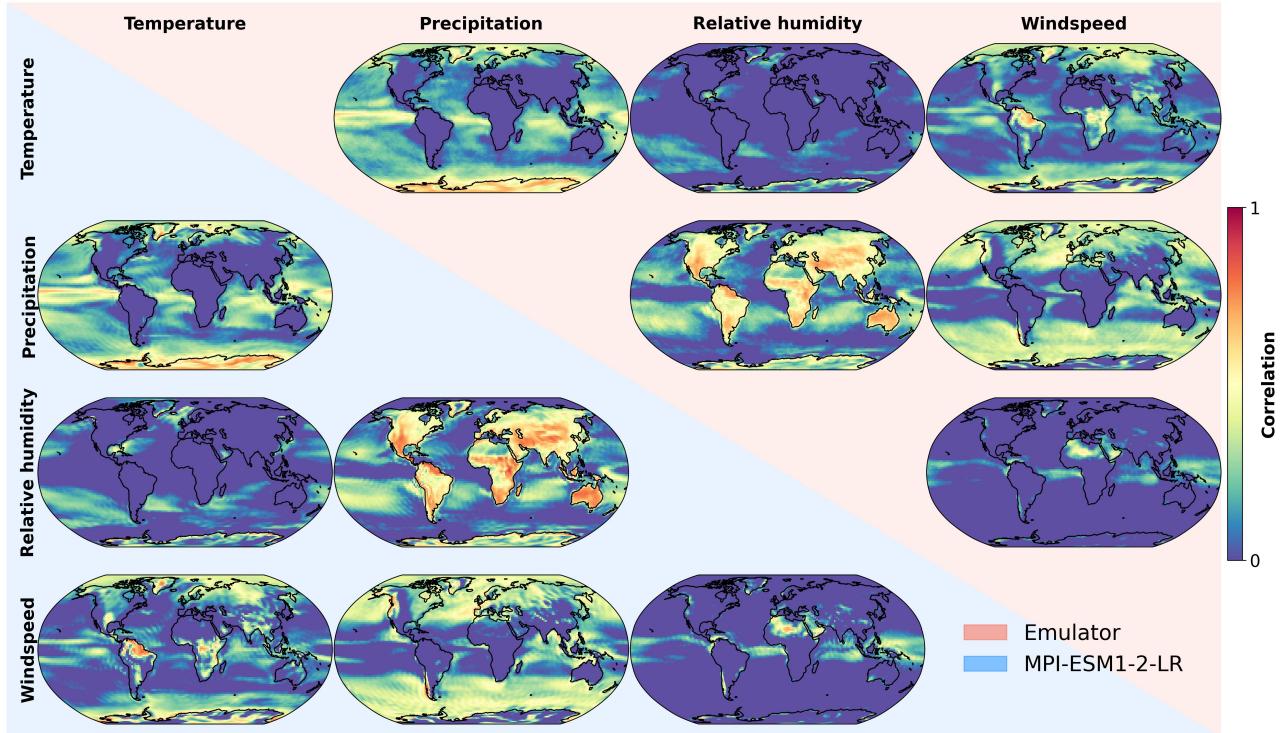


Figure 7: Spatial maps of grid-cell cross-correlations for all months between each pair of the four variables, shown for the MPI-ESM1-2-LR piControl run (bottom left) and for the emulated large ensemble (top right). Corresponding results for MIROC6 and ACCESS-ESM1-5 are provided in the supplementary.

We first examine the cross-variable correlations at individual grid-cells. An effective emulator should generate realizations that satisfy statistical and physical consistency across variables, and therefore open the door to studying projected compound events under alternative emission pathways. Figure 7 compares the grid-cell cross-correlation between the variables obtained from the proposed generative emulator, with those estimated from the MPI-ESM1-2-LR piControl run. It shows the variables' cross-correlations from emulator matches those from the reference ESM it aims to emulate. Because the current emulator generates monthly snapshots and does not incorporate temporal dependencies, its utility for temporal compounding event studies is limited (daily time series are typically required). However, these results suggest that generative models can reproduce statistical consistency across variables, and therefore have potential in this area. With additional compute to condition the emulator on past states and generate daily samples, the emulator can be extended to fit this purpose, following for example the work of Bassetti et al. (2024).

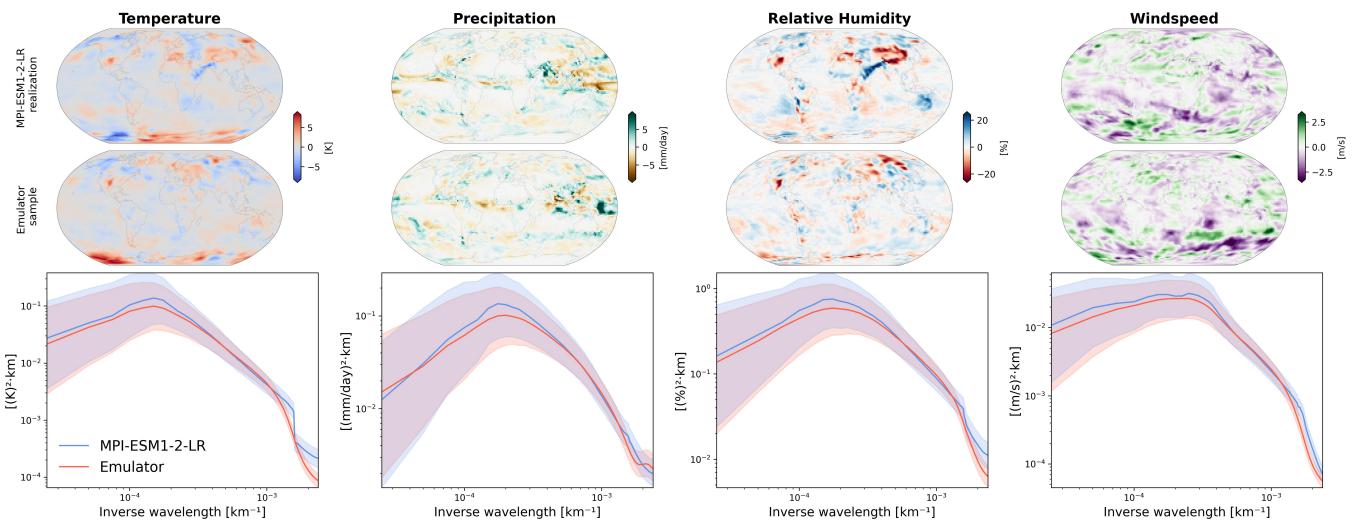


Figure 8: Top: Examples of individual realizations from the MPI-ESM2-1-LR and emulator for the same randomly selected month (July) illustrating the spatial structure of realizations. Bottom: global zonally averaged spherical power spectra from an ensemble of individual realizations for each variable. Spectral densities are computed for all available data in the MPI-ESM2-1-LR piControl run and the emulated large ensemble without forcing. Full lines indicate the median power density and shaded area the 5-95% interquartile range.

We next examine spatial correlations within each variable by studying the spatial structure of individual emulator realizations. For each variable, we compute the global, zonally averaged spherical power spectra of individual realizations and form an ensemble of power spectral densities. Figure 8 shows an example of a spatial sample for each variable along with the ensemble power spectral density. Large-scale spatial structures are generally realistically reconstructed, but as we get closer to the ESM native grid resolution, this turns out to be more challenging. This indicates that the emulator should be able to reproduce synoptic-scale structures such as heat domes or monsoons, but is less reliable at replicating ESM output features confined to only a few grid cells.

At scales close to the numerical meshgrid, ESM projections themselves are influenced by parameterization biases, and impact models typically do not treat them as perfectly accurate. They rely on bias correction, downscaling, and multi-model ensembles to correct inputs or explicitly incorporate uncertainty, thereby ensuring that sensitivities to climate model errors do not mislead conclusions (Falloon et al., 2014; Maraun, 2016; Lange, 2019). In this context, an emulator that exhibits fine-scale inaccuracies that remain within the tolerated envelope of input uncertainty should be just as useful as the original ESM for impact studies.

4 Emulation of forced simulations

We now evaluate if the emulator can reproduce the climate model output distribution in a climate change scenario, when an anthropogenic greenhouse gas forcing is applied. The goal is to assess how well it performs when $\Delta T_t \neq 0$.

Because emulators are intended to explore alternative emission pathways not run by climate models, we evaluate them on scenarios that were excluded from the training set for this purpose. We choose SSP2-4.5, an intermediate scenario with emission peaking by mid-century, and SSP3-7.0, a high-emission scenario with fragmented climate policies because of regional rivalry. These two scenarios lie between the strong mitigation and very high forcing trajectories used for training, and therefore test the emulator's skill at projecting within the range of ScenarioMIP forcings. The emulator is an interpolation algorithm and is not expected to have skill in extrapolating beyond that range.

For each prediction scenario, we generate a large ensemble of emulated climates conditioned on the GMST time series. At a given time $t = (\text{year}, \text{month})$, the large ensemble is generated by computing the GMST anomaly ΔT_t from the reference climate model output, and then drawing $x_t \sim q(x | \Delta T_t)$ 50 times.

4.1 Forced trends

We first evaluate whether the emulator is able to reproduce the forced climate change trends simulated by the ESM. To assess this, we compute mean annual anomalies for each variable over the SSP3-7.0 scenario using ESM output data and an emulated large ensemble. We choose SSP3-7.0 because it has a stronger forcing, so the climate change signal is expected to emerge more clearly.

Temperature and relative humidity trends We study the emulated trends averaged for four characteristic regions with different warming rates following Giani et al. (2024): land, tropical ocean (ocean between 10°S-10°N), Southern Ocean (ocean poleward of 55°S), and Arctic region (poleward of 80°N). Land warms faster than the oceans due to a lower heat capacity and limited evaporative cooling, which also results in a drop in relative humidity (Byrne and O’Gorman, 2016, 2018). The Arctic has the fastest warming rate because of positive feedbacks, and displays high variability (Previdi et al., 2021). The greater heat capacity and unlimited moisture supply in the oceans should lead to a slower warming and smaller changes in relative humidity per degree Kelvin (Byrne and O’Gorman, 2018). The Southern Ocean warming is particularly muted due to strong wind-driven upwelling of cold waters and melting of ice shelf (Armour et al., 2016; Dong et al., 2022).

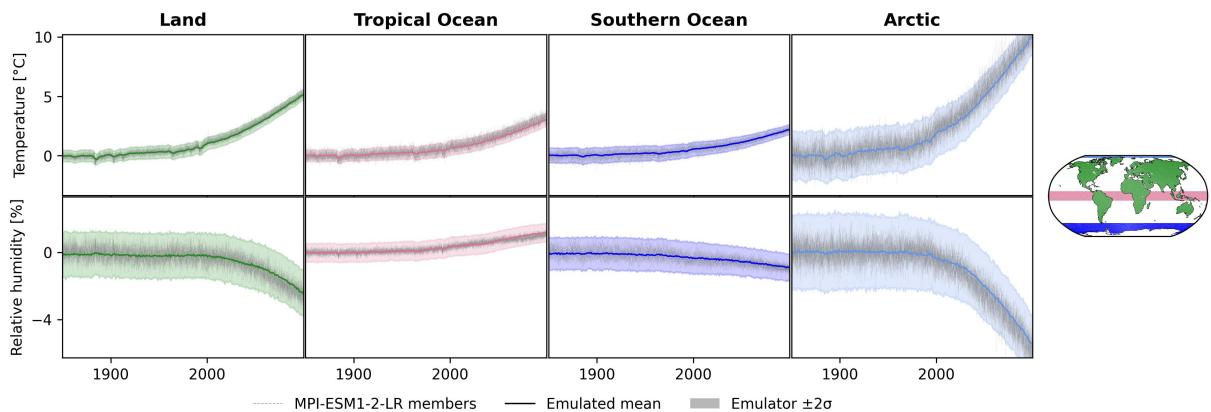


Figure 9: 1850-2100 evolution of the emulated mean annual anomaly for temperature (top) and relative humidity (bottom) in different world regions under SSP3-7.0, overlayed on top of the MPI-ESM2-1-LR ensemble members (gray). Spread represents 2 standard deviations from the emulator mean.

In Figure 9, the mean emulated trend is shown together with the range defined by two standard deviations above and below the mean. They are overlayed on the ESM ensemble member trajectories for the same region. We find that the emulator reproduces the trends associated with the simulated warming fingerprints of each region, in agreement with the ESM trajectories. The emulator can adapt to differences in magnitude and sign of regional fingerprints depending on the ESM on which it is trained. For example, it successfully reproduces the opposite trends in Southern Ocean relative humidity anomalies between the MPI-ESM1-2-LR and MIROC6 projections (shown in the supplementary).

The figure also reveals a modest underestimation of mean temperature and relative humidity anomalies by the end of the century in some regions. This effect is most pronounced in the Arctic for ACCESS-ESM1-5 and MIROC6, where the emulator does not fully capture the late-century anomaly range. For relative humidity, the emulator further tends to overestimate variability. Nonetheless, the ESM ensemble members are generally well captured within two standard deviations of the emulated large ensemble. This shows the emulator has skill in reproducing the statistics of forced trends for a scenario absent from its training data.

Precipitation trends For precipitation, we focus on the tropical belt (30°N – 30°S), where most of the variations in global mean precipitation arise, in association with the circulation changes that dominate future hydro-climate projections (Trenberth, 2011). Figure 10 shows a longitude-time view of the annual mean tropical precipitation anomaly under SSP3-7.0, comparing the ESM ensemble mean with the emulated ensemble mean.

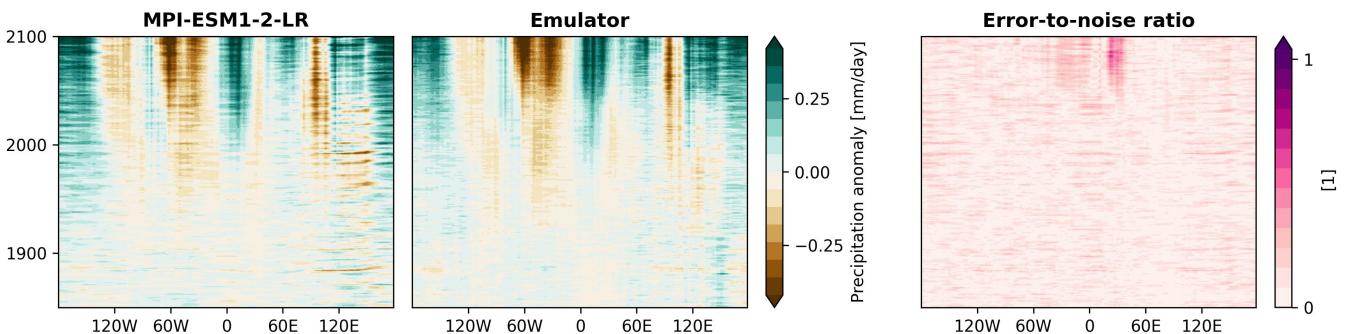


Figure 10: Hovmöller diagrams of 30°N – 30°S annual mean precipitation anomalies under SSP3-7.0 for MPI-ESM1-2-LR (left) and the emulated large ensemble (center). The right panel shows the error normalized by the MPI-ESM1-2-LR standard deviation across ensemble members.

The emulator reproduces the large-scale anomaly patterns seen in the ESM, with wetter conditions over the tropical Pacific (120°E – 120°W) and drier conditions over the Amazon rainforest ($\sim 60^{\circ}\text{W}$) (Lee et al., 2021). This reflects the “wet-get-wetter/dry-get-drier” mechanism (Held and Soden, 2006) in which regions already prone to heavy rainfall experience more precipitation, while evaporating areas such as the Amazon basin tend to dry. Note that part of this drying signal may reflect a known bias of the MPI-ESM1-2-LR model, which underestimates rainfall over tropical land (Mauritsen and Roeckner, 2020). Some discrepancies in magnitude and smaller-scale structure of the emulated and ESM projected precipitation anomaly persist. To gauge their importance, we compute the absolute difference between the two anomaly fields and normalize it by the noise magnitude. The noise is taken as the ESM internal variability estimated across ensemble members. The resulting error-to-noise map in Figure 10 shows ratios close to zero across times and longitudes. This suggests that the mismatch in the emulated mean forced trend is small compared to natural variability, and would be hard to detect in any single realization.

Wind speed trends We compute the evolution of the zonally averaged surface wind speed anomaly from the ESM output and emulated large ensemble. Zonal averages emphasize shifts in the westerlies and trade winds. The Southern Hemisphere westerlies are expected to shift poleward, in association with a weakening of low-level winds over the southern parts of South America, southern Africa, and Australia (Lee et al.,

2021). Arctic wind speed is projected to strongly increase as surface roughness decreases with ice melting and the boundary layer becomes warmer and less stratified (Mioduszewski et al., 2018). On the other hand, the response in Antarctica is expected to be less robust, as near-surface wind changes are more complex and regionally variable (Davrinche et al., 2025). Figure 11 shows that, like for precipitation, the emulated wind speed anomaly trends match well the trends computed from the ESM outputs, and are largely able to reproduce the aforementioned trends. Notably, the emulator reproduces shifts in wind patterns despite being conditioned on pattern scaling, which assumes a fixed pattern of change for surface temperatures.

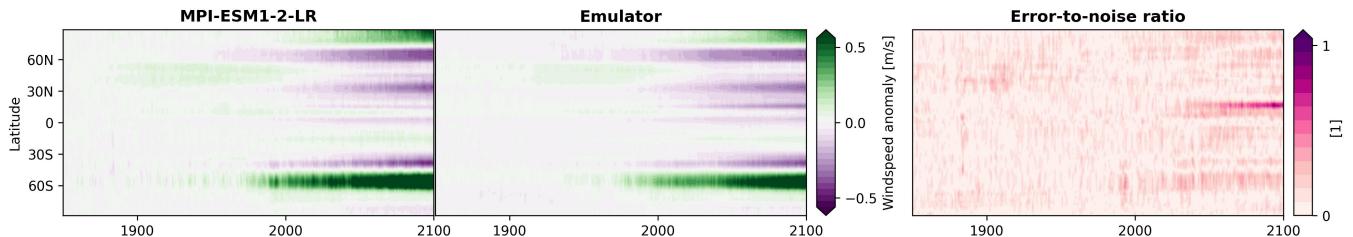


Figure 11: 1850-2100 evolution of the annual zonal mean wind speed anomalies under SSP3-7.0 for MPI-ESM1-2-LR (left) and the emulated large ensemble (center). The right panel shows the error normalized by the MPI-ESM1-2-LR standard deviation across ensemble members.

4.2 Time of Emergence

The time of emergence (ToE) refers to the moment when a forced climate change signal becomes detectable and rises above the background variability (Hawkins and Sutton, 2012). Traditionally, this has been quantified using the signal-to-noise ratio, where the signal is the long-term trend from a single model’s projection and the noise is internal variability estimated from pre-industrial control runs or large ensembles of the same model (Barnett and Schlesinger, 1987; Giorgi and Bi, 2009). This is an important concept for impact assessment as it signals when the Earth’s climate becomes unequivocally different from what was previously experienced. A recurrent assumption in the aforementioned studies of emergence is that the variance of internal variability is approximately stationary.

To verify whether the emulator can reproduce the ToE simulated by the ESM, we select SSP3-7.0 as a scenario with a sufficiently strong greenhouse gas forcing so that we can expect the projection simulated by the ESM to exhibit regional emergence for all four variables considered. We emulate a large ensemble under the SSP3-7.0 GMST trajectory and compute for each variable the first year from which the signal-to-noise ratio becomes greater than 2 for three years in a row. The noise magnitude is estimated as the anomaly standard deviation under the ESM piControl run.

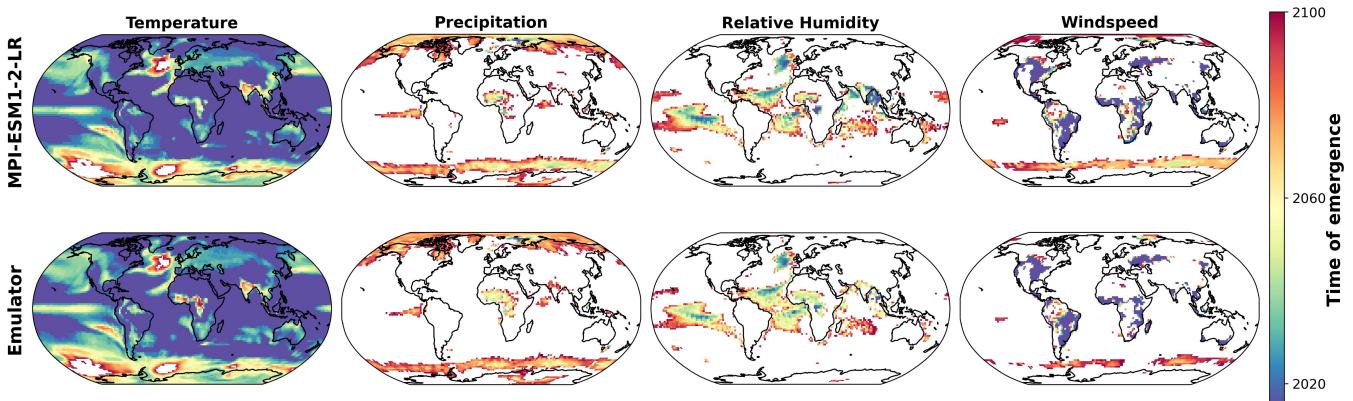


Figure 12: Time of emergence of the annually averaged anomalies computed for the MPI-ESM1-2-LR simulated ensemble and for the emulated ensemble under SSP3-7.0

Figure 12 shows that the emulated ensemble is largely capable of recovering a pattern of emergence that matches the ESM, with later emergence of the temperature signal in the Antarctic due to higher variability, emergence of precipitation at high latitudes and in some tropical regions (Ranasinghe et al., 2021), or emergence from the poleward shift in southern westerlies. We include in the supplementary results for SSP2-4.5, which shows that as we emulate a lower forcing scenario, emergence for all variables either occurs at a later year or never occurs, in accordance with the ESM response. While our definition of ToE is somewhat arbitrary, and more rigorous estimates of emergence have been proposed (Li et al., 2017; Rivoire et al., 2024), this diagnostic already reflects how good the emulator is at reproducing the signal-to-noise ratio from the climate model output.

4.3 Distributional change with warming

It is expected that not only the mean response may shift with warming, but higher-order features of climate variables' anomaly distributions can change with warming. We propose to examine how well the emulator can reproduce these changes with warming. Figure 13 compares anomaly distributions for the four emulated variables in two states: the pre-industrial control climate and the end of the century under SSP2-4.5. The distributions are plotted for a selection of regions where the climate change signal has emerged in the ESM outputs under SSP2-4.5. The emulator seems to reproduce the main distributional shifts seen in the ESM output. While some bias is visible for emulated temperature anomalies over tropical Africa, the bulk of each distribution appears well captured by the end of the century for the regions considered. In particular, the emulation of the change in skewness for wind speed over South America and relative humidity over India is appreciable, highlighting its ability to reproduce higher-order distributional shifts that Gaussian distributions cannot represent.

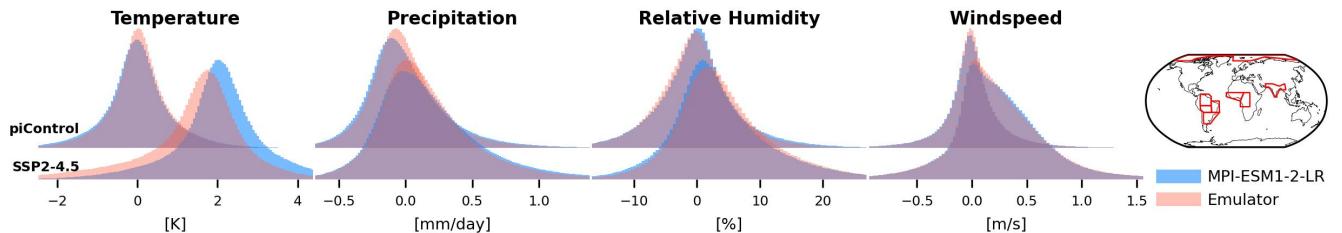


Figure 13: Distributional change of the climate model output and emulated large ensemble between pre-industrial era and end of century (2080-2100) under SSP2-4.5. Distributions are binned using monthly data over a selection of AR6 regions outlined in red on the right: central and western Africa for temperature, Arctic Ocean for precipitation, south Asia for relative humidity, and South America for wind speed.

In Figure 13, the probability masses that need to be rearranged to bring ESM and emulator projections into agreement correspond to the areas shaded in blue and red. Visually, they appear small in comparison with the range covered by the variability. To provide a comprehensive assessment of this, we compute — separately for each ESM, AR6 region, season, and for mid-century and end-century periods under SSP2-4.5 — the EMD-to-noise ratio between the emulated and the ESM output distribution. A map of AR6 regions (Iturbide et al., 2020) with acronyms is included in the supplementary. The results in Figure 14 support that in a majority of cases, the distributions align well relative to the scale of internal variability, with an EMD-to-noise ratio well below 0.5. We find that tropical Africa for the MPI-ESM1-2-LR model, including the bordering Atlantic ocean region, is in fact one of the few exceptions. It is also a region where errors in emulating this ESM have been reported in previous work (Nath et al., 2022; Geogdzhayev et al., 2024). This is possibly linked to dynamic vegetation feedbacks in the MPI-ESM1-2-LR land model, which can display a nonlinear response on the near-surface climate in Africa that does not easily correlate with GMST (Baudena et al., 2015; Reick et al., 2021). Additional mismatches in distribution are found in the Arctic Ocean for MIROC6, and the highest discrepancies generally occur at the end of the century, when SSP2-4.5 is most different from the training data.

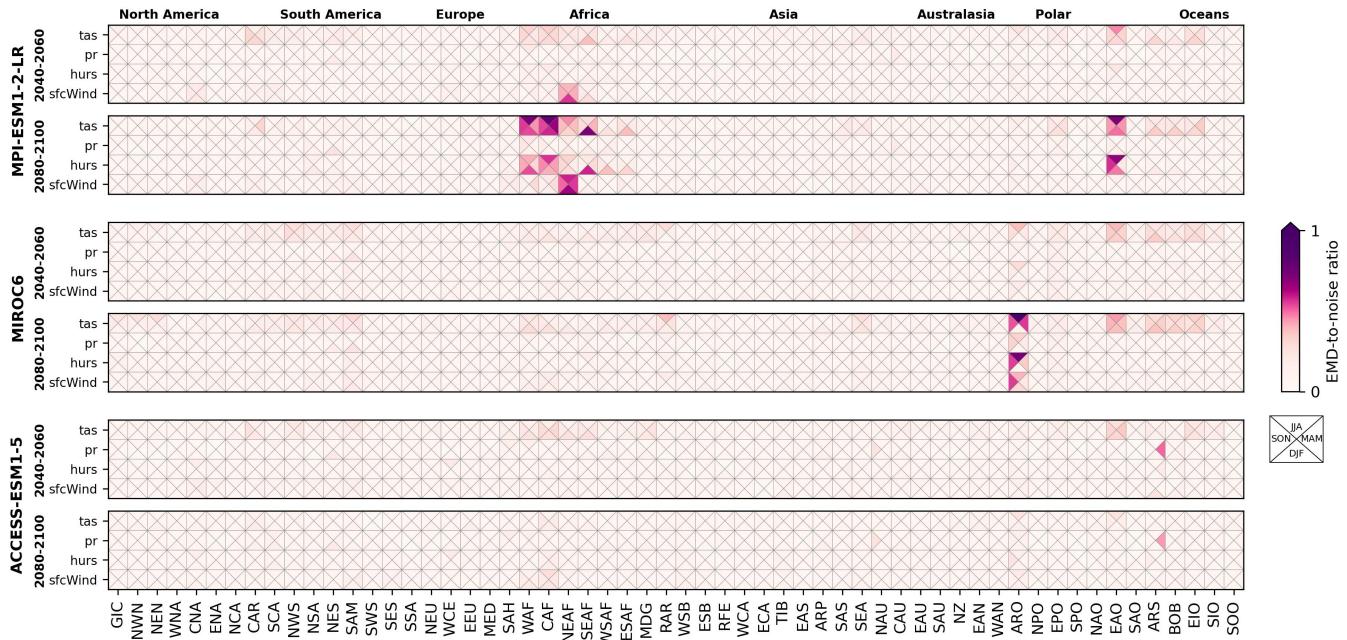


Figure 14: EMD-to-noise ratio under SSP2-4.5 for each ESM, variable, season and AR6 region. Results are shown for two periods of the scenario: mid-century (2040–2060) and end-century (2080–2100). The reference standard deviation is floored to 0.1 to prevent spurious inflation of the ratio.

4.4 Extreme tails with warming

Record-shattering events, whether observed or simulated, sit in the far tail of the climate distribution for any given year. They contribute to natural disasters that incur enormous human and economic losses. Although ESMs still have limitations in representing these extremes, they remain a useful indicator of future forced changes in extremes, while future advances in Earth system modeling will likely improve the representation of extremes in climate projection. Therefore, it is desirable for a climate model output emulator to accurately reproduce the most extreme events projected by ESMs under climate change.

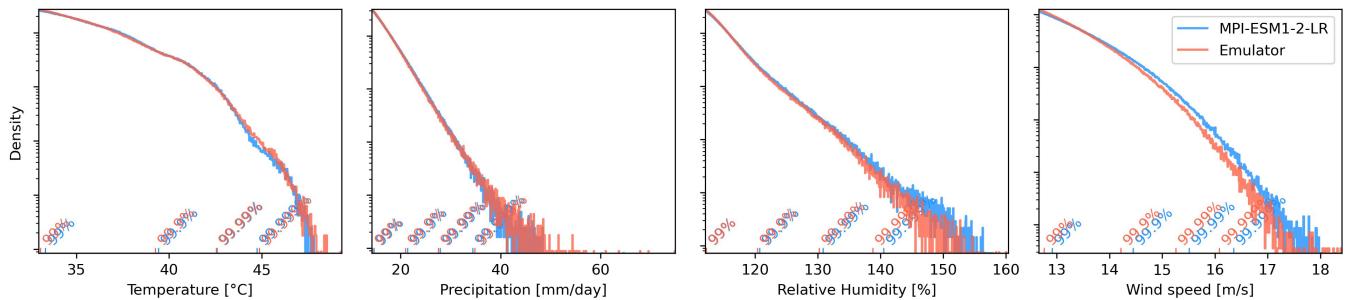


Figure 15: Extreme tails of the MPI-ESM1-2-LR output distribution and the distribution from the emulated large ensemble. Tails are binned from all grid cell values, ensemble members, and months over the 2080–2100 period under SSP2-4.5. Positions of the 99%, 99.9%, 99.99%, and 99.999% quantiles are reported on the x-axis for the ESM (blue) and emulator (red). Near-surface relative humidity values can spuriously exceed 100% in ESM outputs (Ruosteenoja et al., 2017).

Because extreme tail events carry little probability mass, the EMD is insensitive to their misalignment. We therefore prefer to evaluate biases in extreme tail quantiles at the end of the century. We convert back the anomalies into absolute values and compare the tails of the climate model output distribution and emulated ensemble distributions over 2080–2100 under SSP2-4.5. Distributions are binned by taking all grid cell values, ensemble members, and months, and we report in Figure 15 their tails beyond the 99%

quantile. The emulator shows good agreement with the ESM in the highest extremes for all variables, with modest underestimation of the most extreme quantiles for relative humidity and wind speed. The only notable bias we find is an underestimation of the extreme tails in relative humidity for ACCESS-ESM1.5. Working with monthly averages helps with emulating the tails of the distributions: time-averaging pulls extremes closer to the bulk of the distribution, making them easier for the emulator to match. Monthly extremes have relevance for seasonal risk assessments, for example, in simple crop-yield models or monthly water-balance models (Ray et al., 2015; Bock et al., 2017), and these results show the proposed emulator has skill to support them. Extreme tails are likely to become more challenging to emulate at daily or hourly resolution, which are the scales at which impact assessments matter most, and we leave this challenge for future work.

The histogram in Figure 15 bins values from all grid cells and months indiscriminately. This raises the question of whether the model gets the most extreme values right, at the right location, and in the right month. To assess this, we compute the relative deviation in tail quantiles for each AR6 region and each season under SSP2-4.5 and report the results in Figure 16. We find that the bias in extreme tail quantiles from the emulator is mostly bounded within 10% of the ESM corresponding quantile value. For temperature, relative humidity, and wind speed, the emulator matches closely the ESM’s tails in most regions and seasons, with occasional deviations as we move to more extreme quantile levels. As expected, bias grows at higher quantiles since rarer events provide less training signal. This is most visible for relative humidity in ACCESS-ESM1.5, where the relative bias increases most at the upper tails. Precipitation proves to be more difficult: it displays the largest underestimation of bias in the 99% quantile for multiple regions, and upper tails show pronounced overestimation for a large number of regions and seasons. We test the statistical significance of biases in tail quantile estimates relative to their variability and find that they are significant in a majority of cases, with detailed results included in the supplementary. This suggests that even if errors in the extreme tails are small in magnitude, they still constitute a systematic error introduced by the emulator in its representation of tails.

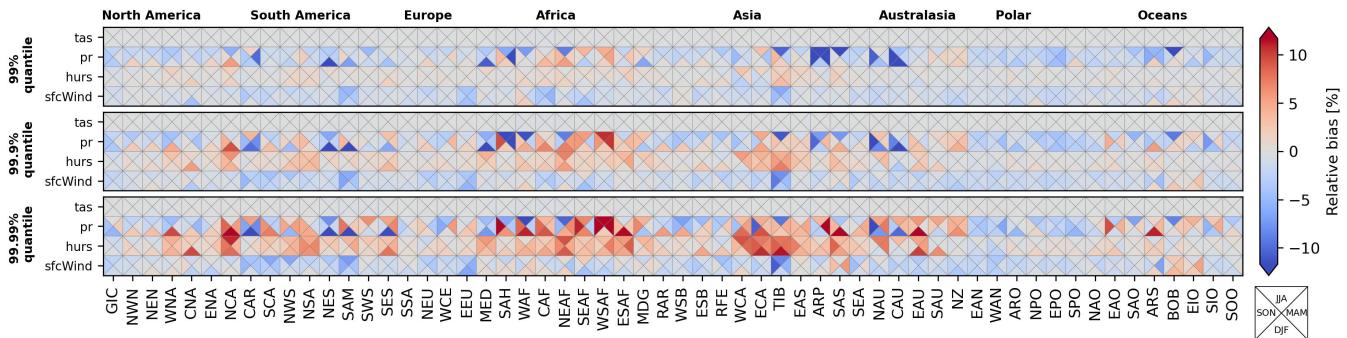


Figure 16: Relative bias in the extreme tail quantile between the MPI-ESM1-2-LR output and the emulated large ensemble for each AR6 region and season. Values are reported as relative deviations with respect to the ESM quantile estimate, where red/blue is an overestimation/underestimation by the emulator.

5 Discussion

5.1 On the utility of emulating biased climate models

One critique of climate model emulation is that many ESMs exhibit known biases, particularly in their representation of tails and fine-scale variability — hence the large research effort devoted to model tuning and bias correction, and the constant drive to increase models’ resolution. Therefore, if parts of the simulated distribution are known to be unreliable, it is reasonable to question whether emulating them is even useful. We argue that this critique is misplaced for two reasons.

First, while imperfect, ESMs have undeniably allowed us to advance our understanding of the climate system and of what the future may look like, in that some behaviors are well represented and robustly emerge across different climate models. Where biases exist, many applications already treat ESM outputs with caution. For example, bias correction, downscaling, and multi-model ensembles are standard tools in impact assessment workflows to adjust inputs and explicitly represent uncertainty, so that conclusions are not misled by ESM errors (Falloon et al., 2014; Maraun, 2016; Lange, 2019). In practice, this means emulators can be used directly where ESMs are reliable, and postprocessed through the same existing pipelines where they are not.

Second, the development of climate model emulators is also largely a methodological endeavor that is ESM-agnostic. As such, when better climate models are available, the same frameworks can be used to emulate them with minimal adjustment. In fact, we can already envision using gridded reanalysis products to train emulators that better match observed climate distributions than ESM outputs (Brenowitz et al., 2025). For methodological development, as in this work, evaluating an emulator’s ability to capture future changes requires future projections, which justifies training and validation on CMIP6 archives. Once trained on CMIP6 simulations, however, the emulator can be fine-tuned on reanalysis products or available observations using transfer learning or nudging approaches to reduce model bias before practical use (Immorlano et al., 2025; Wang et al., 2025).

5.2 On the GMST forcing assumption

In this work, we make the simplifying assumption that GMST anomalies, through pattern scaling, provide a sufficient predictor for the sources of forcing relevant to project climate change. This design, recurrent in the emulator community, enables straightforward coupling of the emulator with simple climate models, thereby allowing ESM outputs to be emulated directly from global emissions pathways (Beusch et al., 2021; Mathison et al., 2025). However, this assumption can introduce two important sources of error (Womack et al., 2025b).

First, it neglects the system’s memory by assuming that responses depend only on the instantaneous GMST. While this approximation may hold under stronger forcing scenarios (Giani et al., 2024), it potentially introduces biases in overshoot pathways, where hysteresis effects may prevent regional climates from returning to past states (Womack et al., 2025a). Second, it introduces a hidden variable error by excluding the influence of forcings not directly encoded in GMST. For instance, although global precipitation is expected to decrease with reductions in GMST, it also exhibits a direct radiative cooling response to atmospheric CO₂ such that a temporary increase in precipitation may accompany its decline (O’Gorman et al., 2012). More evidently, the GMST cannot capture the impact of regional forcers, such as aerosols or ozone, which have the potential to play a significant and very localized role (Williams et al., 2022). Directions to overcome these deficiencies include replacing pattern scaling with an impulse-response model that accounts for system memory and allows for warming patterns to evolve over time, and expanding the conditioning signal to include emissions data alongside surface temperature, with spatially resolved maps of emission maps of regional short-lived forcers.

From a different standpoint, the instantaneous nature of this forcing also neglects the dynamical nature of the system. As a result, the emulator cannot generate temporally consistent samples, limiting its ability to study the influence of seasonal and multi-annual variability in projections. A way forward would be to adopt an autoregressive strategy, conditioning the emulator not only on forcing but also on past states, to generate entire spatio-temporal sequences.

5.3 On the practical use to support impact assessment

Impact modeling often requires climate projections at daily resolution and spatial scales of 0.5° (Warszawski et al., 2014; Maraun, 2016) — in particular for extreme events — which is much finer than the resolution presented in this work. The current emulator may already be useful for some applications that use long-term averages, for example, in ecology (Tabor and Williams, 2010; Mahony et al., 2022), or for simple models of

crop yields or water balance (Ray et al., 2015; Bock et al., 2017). However, we prefer to view it as a building block toward higher temporal and spatial resolution approaches. We chose to work with monthly averages to keep data volumes manageable on our limited computing resources. However, going to higher resolutions and frequencies is well within the scope of our intended work. In practice, the emulator could already be paired with statistical or dynamical downscaling techniques to provide projections on scales suitable for impact assessment ((e.g. Schillinger et al., 2025)).

Another key consideration is the computational cost. The more efficient an emulator is, the more accessible it becomes — a common claim is that emulators should run on standard personal computers. While our emulator is indeed computationally efficient, it still requires GPU acceleration to achieve this performance. An important focus in our development was to keep the network lightweight, but reducing computational requirements further remains an important goal for future work. At the same time, while it may be desired, full portability may not always be necessary. As for large language models, which are often hosted on servers and made accessible through web interfaces, modeling centers could provide GPU-backed services to run emulators of their own ESMs, making them available to the community without requiring local hardware.

When evaluating an emulator, it is natural to aim for the closest possible match to its reference ESM. However, perfect agreement is not always necessary. First, given the known deficiencies of ESMs, a perfect reproduction of all aspects of their output distribution may have limited practical value for physical risk assessment. Some deviation should be tolerated where ESM projections are uncertain or biased. Second, although the relevant statistics may vary across impact models, many rely on specific quantities — such as means, variability, or extremes — rather than the full climate distribution. In these cases, what matters is that the emulator reproduces those relevant statistics with sufficient skill, not that it matches every feature of the ESM output. Finally, even when the ESM is reliable and the statistics of interest are well defined, we argue that small discrepancies between the emulator and the ESM are unlikely to affect downstream impact assessment, provided such deviations remain within the envelope of the ESM internal variability. In this sense, an emulator need not stringently mimic its reference ESM to be practically useful; it only needs to be accurate enough for the task it supports.

6 Conclusion

We have introduced a score-based generative emulator of monthly averaged climate model output anomalies conditioned on GMST anomalies. The emulator targets four near-surface variables of relevance for impact assessment: temperature, precipitation, relative humidity, and wind speed. It provides a computationally efficient procedure to draw samples from a distribution that approximates the joint distribution of climate model outputs. To evaluate performance, we proposed diagnostics comparing statistical features of emulator-generated data against ESM outputs under both unforced and forced regimes. In particular, we assessed whether discrepancies between emulator and ESM distributions are meaningful relative to internal variability.

Our results provide evidence that on many fronts, the proposed emulator is capable of producing distributions that closely match the ones from ESM outputs, showing promise to reproduce key statistical features of the reference ESM. These statistical features include higher-order moments, cross-variable correlations, and the extreme tails. Although the emulator may not be a perfect match to the ESM, we argue it has potential to be useful to support impact assessment given the range of variability in future projections. Our analysis also reveals important failure cases, such as difficulties in representing strong seasonal distributional shifts and a tendency to overfit portions of the training data. These findings highlight both the promise of deep generative approaches for climate model output emulation and the challenges ahead. Future work should address these modeling limitations while advancing the emulator as a practical tool for impact assessment, with priorities including finer spatial and temporal resolution, reduced computational cost, and bias correction or transfer learning to bring outputs closer to observational distributions.

Acknowledgments

This work acknowledges support by Schmidt Sciences, LLC, through the Bringing Computation to the Climate Challenge (BC3), an MIT Climate Grand Challenge Project. We also acknowledge the MIT *Engaging* cluster supported by the Office of Research Computing and Data, and MIT *Svante* cluster supported by the Center for Sustainability Science and Strategy for computing resources. We are grateful to Paolo Giani, Björn Lutjens, Christopher Womack, Noelle Eckley Selin, and Claudia Tebaldi for insightful discussions and thorough feedback, which have helped shape this work and improve the quality of the manuscript. We thank Laura Battaglia for valuable discussions that greatly contributed to the development of this work.

Open research

All data used in this work are publicly available through the World Climate Research Program (WCRP) Coupled Model Intercomparison Project 6 (CMIP6) and were retrieved through the Earth System Grid Federation interface. Code to reproduce results is published here <https://github.com/shahineb/climemu>. Pretrained model weights are published here <https://huggingface.co/shahineb/climemu>. Data is processed using Xarray (Hoyer and Hamman, 2017), SciPy (Virtanen et al., 2020) and NumPy (Harris et al., 2020). Figures were made with Matplotlib (Caswell et al., 2020; Hunter, 2007), Cartopy (Met Office, 2010 - 2015) and Seaborn (Waskom, 2021). Models are implemented using Equinox (Kidger and Garcia, 2021), Optax (DeepMind et al., 2020), DiffraX (Kidger, 2021), using the JAX ecosystem (Bradbury et al., 2018).

References

Detlef van Vuuren, Brian O'Neill, Claudia Tebaldi, Louise Chini, Pierre Friedlingstein, Tomoko Hasegawa, Keywan Riahi, Benjamin Sanderson, Bala Govindasamy, Nico Bauer, et al. The Scenario Model Intercomparison Project for CMIP7 (ScenarioMIP-CMIP7). *EGUsphere*, 2025:1–38, 2025. 1

Colin G Jones, Fanny Adloff, Ben BB Booth, Peter M Cox, Veronika Eyring, Pierre Friedlingstein, Katja Frieler, Helene T Hewitt, Hazel A Jeffery, Sylvie Joussaume, et al. Bringing it all together: science priorities for improved understanding of Earth system change and to support international climate policy. *Earth System Dynamics*, 15(5):1319–1351, 2024. 1

Claudia Tebaldi, NE Selin, R Ferrari, and G Flierl. Emulators of climate model output. *Annual Review of Environment and Resources*, 50, 2025. 1, 2

Lila Warszawski, Katja Frieler, Veronika Huber, Franziska Piontek, Olivia Serdeczny, and Jacob Schewe. The inter-sectoral impact model intercomparison project (ISI-MIP): project framework. *Proceedings of the National Academy of Sciences*, 111(9):3228–3232, 2014. 1, 20

Benjamin D Santer, Tom ML Wigley, Michael E Schlesinger, and John FB Mitchell. Developing climate scenarios from equilibrium GCM results. 1990. 2, 5

Claudia Tebaldi and Julie M Arblaster. Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, 2014. 2, 5

Nadja Herger, Benjamin M Sanderson, and Reto Knutti. Improved pattern scaling approaches for the use in climate impact studies. *Geophysical Research Letters*, 42(9):3486–3494, 2015. 2

Valerio Lucarini, Francesco Ragone, and Frank Lunkeit. Predicting climate change using response theory: Global averages and spatial patterns. *Journal of Statistical Physics*, 166(3):1036–1064, 2017. 2

Lyssa M Freese, Paolo Giani, Arlene M Fiore, and Noelle E Selin. Spatially resolved temperature response functions to CO₂ emissions. *Geophysical Research Letters*, 51(15):e2024GL108788, 2024. 2

Christopher B Womack, Paolo Giani, Sebastian D Eastham, and Noelle E Selin. Rapid emulation of spatially resolved temperature response to effective radiative forcing. *Journal of Advances in Modeling Earth Systems*, 17(1):e2024MS004523, 2025a. 2, 20

Alexander J Winkler and Carlos A Sierra. Towards a new generation of impulse-response functions for integrated Earth system understanding and climate change attribution. *Geophysical Research Letters*, 52(8):e2024GL112295, 2025. 2

Marit Sandstad, Norman Julius Steinert, Susanne Baur, and Benjamin Mark Sanderson. METEORv1.0.1: A novel framework for emulating multi-timescale regional climate responses. *EGUphere*, 2025:1–49, 2025. 2

Lea Beusch, Lukas Gudmundsson, and Sonia I Seneviratne. Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*, 2020. 2, 5

S. Nath, Q. Lejeune, L. Beusch, S. I. Seneviratne, and C.-F. Schleussner. MESMER-M: an Earth system model emulator for spatially resolved monthly temperature. *Earth System Dynamics*, 2022. 2, 4, 5, 17

Gosha Geogdzhayev, Andre Nogueira Souza, Raffaele Ferrari, and Glenn R Flierl. A Statistical Emulator Design for Averaged Climate Fields. *Authorea Preprints*, 2024. 2, 4, 17

Camilla Mathison, Eleanor J Burke, Gregory Munday, Chris D Jones, Chris J Smith, Norman J Steinert, Andy J Wiltshire, Chris Huntingford, Eszter Kovacs, Laila K Gohar, et al. A rapid-application emissions-to-impacts tool for scenario assessment: Probabilistic Regional Impacts from Model patterns and Emissions (PRIME). *Geoscientific Model Development*, 18(5):1785–1808, 2025. 2, 5, 20

Abigail Snyder, Robert Link, Kalyn Dorheim, Ben Kravitz, Ben Bond-Lamberty, and Corinne Hartin. Joint emulation of Earth System Model temperature-precipitation realizations with internal variability and space-time and cross-variable correlation: fldgen v2. 0 software description. *Plos one*, 14(10):e0223542, 2019. 2

Shruti Nath, Julie Carreau, Kai Kornhuber, Peter Pfleiderer, Carl-Friedrich Schleussner, and Philippe Naveau. MERCURY: A fast and versatile multi-resolution based global emulator of compound climate hazards. *arXiv preprint arXiv:2501.04018*, 2024. 2, 4

Sarah Schöngart, Lukas Gudmundsson, Mathias Hauser, Peter Pfleiderer, Quentin Lejeune, Shruti Nath, Sonia Isabelle Seneviratne, and Carl-Friedrich Schleussner. Introducing the MESMER-M-TPv0. 1.0 module: spatially explicit Earth system model emulation for monthly precipitation and temperature. *Geoscientific Model Development*, 17(22):8283–8320, 2024. 2, 4, 5

Stefano Castruccio, David J McInerney, Michael L Stein, Feifei Liu Crouch, Robert L Jacob, and Elisabeth J Moyer. Statistical emulation of climate model projections based on precomputed GCM runs. *Journal of Climate*, 27(5):1829–1844, 2014. 2

Robert Link, Abigail Snyder, Cary Lynch, Corinne Hartin, Ben Kravitz, and Ben Bond-Lamberty. Fldgen v1. 0: an emulator with internal variability and space-time correlation for Earth system models. *Geoscientific Model Development*, 2019. 2, 5

Philip Goodwin, Martin Leduc, Antti-Ilari Partanen, H Damon Matthews, and Alex Rogers. A computationally efficient method for probabilistic local warming projections constrained by history matching and pattern scaling, demonstrated by WASP-LGRTC-1.0. *Geoscientific Model Development*, 2020. 2

Yann Quilcaille, Lukas Gudmundsson, Lea Beusch, Mathias Hauser, and Sonia I Seneviratne. Showcasing MESMER-X: Spatially Resolved Emulation of Annual Maximum Temperatures of Earth System Models. *Geophysical Research Letters*, 49(17):e2022GL099012, 2022. 2, 5

Shahine Bouabid, Dino Sejdinovic, and Duncan Watson-Parris. FaIRGP: A Bayesian energy balance model for surface temperatures emulation. *Journal of Advances in Modeling Earth Systems*, 16(6): e2023MS003926, 2024. 2

Jakob Zscheischler, Olivia Martius, Seth Westra, Emanuele Bevacqua, Colin Raymond, Radley M Horton, Bart van den Hurk, Amir AghaKouchak, Aglaé Jézéquel, Miguel D Mahecha, et al. A typology of compound weather and climate events. *Nature reviews earth & environment*, 1(7):333–347, 2020. 2

Camilla Mathison, Eleanor Burke, Andrew J Hartley, Douglas I Kelley, Chantelle Burton, Eddy Robertson, Nicola Gedney, Karina Williams, Andy Wiltshire, Richard J Ellis, et al. Description and evaluation of the JULES-ES set-up for ISIMIP2b. *Geoscientific Model Development*, 16(14):4249–4264, 2023. 2

Claudia Tebaldi, Abigail Snyder, and Kalyn Dorheim. STITCHES: creating new scenarios of climate model output by stitching together pieces of existing simulations. *Earth System Dynamics Discussions*, 2022: 1–58, 2022. 2

Edward Byers, Michaela Werning, Mahé Perrette, Niklas Schwind, Volker Krey, Keywan Riahi, and Carl-Friedrich Schleussner. Fast climate impact emulation for global temperature scenarios with the rapid impact model emulator (RIME). *Environmental Research: Climate*, 4(3):035011, 2025. 2

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, Brian Henn, James Duncan, Noah D Brenowitz, Karthik Kashinath, Michael S Pritchard, Boris Bonev, et al. ACE: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023. 2

Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025. 2, 5, 6, 10, 20

Brian C O'Neill, Claudia Tebaldi, Detlef P Van Vuuren, Veronika Eyring, Pierre Friedlingstein, George Hurtt, Reto Knutti, Elmar Kriegler, Jean-Francois Lamarque, Jason Lowe, et al. The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9(9):3461–3482, 2016. 3

Keywan Riahi, Detlef P Van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C O'Neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, et al. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change*, 42:153–168, 2017. 3

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. 3

Martin Schupfner, Karl-Hermann Wieners, Fabian Wachsmann, Sebastian Milinski, Christian Steger, Matthias Bittner, Johann Jungclaus, Barbara Früh, Klaus Pankatz, Marco Giorgetta, Christian Reick, Stephanie Legutke, Monika Esch, Veronika Gayler, Helmuth Haak, Philipp de Vrese, Thomas Raddatz, Thorsten Mauritsen, Jin-Song von Storch, Jörg Behrens, Victor Brovkin, Martin Claussen, Traute Crueger, Irina Fast, Stephanie Fiedler, Stefan Hagemann, Cathy Hohenegger, Thomas Jahns, Silvia Kloster, Stefan Kinne, Gitta Lasslop, Luis Kornblueh, Jochem Marotzke, Daniela Matei, Katharina Meraner, Uwe Mikolajewicz, Kameswarao Modali, Wolfgang Müller, Julia Nabel, Dirk Notz, Karsten Peters-von Gehlen, Robert Pincus, Holger Pohlmann, Julia Pongratz, Sebastian Rast, Hauke Schmidt, Reiner Schnur, Uwe Schulzeida, Katharina Six, Bjorn Stevens, Aiko Voigt, and Erich Roeckner. DKRZ MPI-ESM1.2-LR model output prepared for CMIP6 ScenarioMIP, 2021. URL <https://doi.org/10.22033/ESGF/CMIP6.15349>. 3

Björn Lütjens, Raffaele Ferrari, Duncan Watson-Parris, and Noelle Selin. The impact of internal variability on benchmarking deep learning climate emulators. *arXiv preprint arXiv:2408.05288*, 2024. 3, 5

Karl-Hermann Wieners, Marco Giorgetta, Johann Jungclaus, Christian Reick, Monika Esch, Matthias Bitzner, Stephanie Legutke, Martin Schupfner, Fabian Wachsmann, Veronika Gayler, Helmuth Haak, Philipp de Vrese, Thomas Raddatz, Thorsten Mauritsen, Jin-Song von Storch, Jörg Behrens, Victor Brovkin, Martin Claussen, Traute Crueger, Irina Fast, Stephanie Fiedler, Stefan Hagemann, Cathy Hohenegger, Thomas Jahns, Silvia Kloster, Stefan Kinne, Gitta Lasslop, Luis Kornblueh, Jochem Marotzke, Daniela Matei, Katharina Meraner, Uwe Mikolajewicz, Kameswarrao Modali, Wolfgang Müller, Julia Nabel, Dirk Notz, Karsten Peters-von Gehlen, Robert Pincus, Holger Pohlmann, Julia Pongratz, Sebastian Rast, Hauke Schmidt, Reiner Schnur, Uwe Schulzweida, Katharina Six, Bjorn Stevens, Aiko Voigt, and Erich Roeckner. MPI-M MPIESM1.2-LR model output prepared for CMIP6 CMIP, 2019. URL <https://doi.org/10.22033/ESGF/CMIP6.742>. 3

Hiroaki Tatebe and Masahiro Watanabe. MIROC MIROC6 model output prepared for CMIP6 CMIP, 2018. URL <https://doi.org/10.22033/ESGF/CMIP6.881>. 3

Hideo Shiogama, Manabu Abe, and Hiroaki Tatebe. IPCC DDC: MIROC MIROC6 model output prepared for CMIP6 ScenarioMIP, 2023. URL <https://doi.org/10.26050/Wdcc/Ar6.C6SPMIMI>. 3

Tilo Ziehn, Matthew Chamberlain, Andrew Lenton, Rachel Law, Roger Bodman, Martin Dix, Yingping Wang, Peter Dobrohotoff, Jhan Srbinovsky, Lauren Stevens, Peter Vohralik, Chloe Mackallah, Arnold Sullivan, Siobhan O'Farrell, and Kelsey Druken. CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 CMIP, 2019a. URL <https://doi.org/10.22033/ESGF/CMIP6.2288>. 3

Tilo Ziehn, Matthew Chamberlain, Andrew Lenton, Rachel Law, Roger Bodman, Martin Dix, Yingping Wang, Peter Dobrohotoff, Jhan Srbinovsky, Lauren Stevens, Peter Vohralik, Chloe Mackallah, Arnold Sullivan, Siobhan O'Farrell, and Kelsey Druken. CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 ScenarioMIP, 2019b. URL <https://doi.org/10.22033/ESGF/CMIP6.2291>. 3

Yann Quilcaille, Lukas Gudmundsson, and Sonia I Seneviratne. Extending MESMER-X: a spatially resolved Earth system model emulator for fire weather and soil moisture. *Earth System Dynamics*, 14(6):1333–1362, 2023. 4, 5

Mengze Wang, Andre Nogueira Souza, Raffaele Ferrari, and Themistoklis Sapsis. Stochastic Emulators of Spatially Resolved Extreme Temperatures of Earth System Models. *Authorea Preprints*, 2024. 4

M. Meinshausen, S. C. B. Raper, and T. M. L. Wigley. Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, 2011. 4

Thomas Gasser, Philippe Ciais, Olivier Boucher, Yann Quilcaille, Maxime Tortora, Laurent Bopp, and Didier Hauglustaine. The compact Earth system model OSCAR v2. 2: description and first results. 2017. 4

N. J. Leach, S. Jenkins, Z. Nicholls, C. J. Smith, J. Lynch, M. Cain, T. Walsh, B. Wu, J. Tsutsui, and M. R. Allen. FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, 2021. 4

Sonia Seneviratne, Yann Quilcaille, Michael Windisch, Lukas Gudmundsson, Bianca Biess, Felix Jaeger, Matthias Hauser, and Martin Hirschi. Using regional ESM emulators to assess climate feedbacks to IAMs: The “FASTMIP” experimental protocol. In *EGU General Assembly Conference Abstracts*, page 18731, 2024. 4

Timothy J Osborn, Craig J Wallace, Jason A Lowe, and Dan Bernie. Performance of pattern-scaled climate projections under high-end warming. Part I: surface air temperature over land. *Journal of Climate*, 31(14):5667–5680, 2018. 5

Paolo Giani, Raffaele Ferrari, and Noelle Eckley Selin. Origin and limits of invariant warming patterns in climate models. *in preparation*, 2024. 5, 14, 20

Ben Kravitz, Cary Lynch, Corinne Hartin, and Ben Bond-Lamberty. Exploring precipitation pattern scaling methodologies and robustness among CMIP5 models. *Geoscientific Model Development*, 10(5):1889–1902, 2017. 5

June-Yi Lee, Jochem Marotzke, Govindasamy Bala, Long Cao, Susanna Corti, John P Dunne, Francois Engelbrecht, Erich Fischer, John C Fyfe, Christopher Jones, et al. Future global climate: scenario-based projections and near-term information. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 2021. 5, 15

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. ClimateBench v1. 0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022. 5

Timothy J Osborn, Craig J Wallace, Ian C Harris, and Thomas M Melvin. Pattern scaling using ClimGen: monthly-resolution future climate scenarios including changes in the variability of precipitation. *Climatic Change*, 134:353–369, 2016. 5

C Tebaldi, A Armbruster, HP Engler, and R Link. Emulating climate extreme indices. *Environmental Research Letters*, 15(7):074006, 2020. 5

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 5

Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024. 5, 13

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025. 5

Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive and uncertainty-aware downscaling of Earth system model fields with generative machine learning. *Nature Machine Intelligence*, pages 1–11, 2025. 5

Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 6

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 6

Christopher Williams, Fabian Falck, George Deligiannidis, Chris C Holmes, Arnaud Doucet, and Saifuddin Syed. A unified framework for U-Net design and analysis. *Advances in Neural Information Processing Systems*, 36:27745–27782, 2023. 6

Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthias Bartelmann. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005. 6

Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R Durran, Raul A Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V Butz. Advancing parsimonious deep learning weather prediction using the HEALPix mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, 2024. 6

Filippo Santambrogio. Optimal transport for applied mathematicians. 2015. 7

Bin Xu Wang and John J Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *arXiv preprint arXiv:2412.09726*, 2024. 11

Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37: 134614–134644, 2024. 11

Yilun Xu, Gabriele Corso, Tommi Jaakkola, Arash Vahdat, and Karsten Kreis. DisCo-Diff: Enhancing Continuous Diffusion Models with Discrete Latents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. 11

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022. 11

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18434–18443, 2023. 11

Pete Falloon, Andy Challinor, Suraje Dessai, Lan Hoang, Jill Johnson, and Ann-Kristin Koehler. Ensembles and uncertainty in climate change impacts. *Frontiers in Environmental Science*, 2:33, 2014. 13, 20

Douglas Maraun. Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, 2(4):211–220, 2016. 13, 20

Stefan Lange. Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1. 0). *Geoscientific Model Development*, 12(7):3055–3070, 2019. 13, 20

Michael P Byrne and Paul A O’Gorman. Understanding decreases in land relative humidity with global warming: Conceptual model and GCM simulations. *Journal of Climate*, 29(24):9045–9061, 2016. 14

Michael P Byrne and Paul A O’Gorman. Trends in continental temperature and humidity directly linked to ocean warming. *Proceedings of the National Academy of Sciences*, 115(19):4863–4868, 2018. 14

Michael Previdi, Karen L Smith, and Lorenzo M Polvani. Arctic amplification of climate change: a review of underlying mechanisms. *Environmental Research Letters*, 16(9):093003, 2021. 14

Kyle C Armour, John Marshall, Jeffery R Scott, Aaron Donohoe, and Emily R Newsom. Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nature Geoscience*, 9(7):549–554, 2016. 14

Yue Dong, Andrew G Pauling, Shaina Sadai, and Kyle C Armour. Antarctic ice-sheet meltwater reduces transient warming and climate sensitivity through the sea-surface temperature pattern effect. *Geophysical Research Letters*, 49(24):e2022GL101249, 2022. 14

Kevin E Trenberth. Changes in precipitation with climate change. *Climate research*, 47(1-2):123–138, 2011. 15

Isaac M Held and Brian J Soden. Robust responses of the hydrological cycle to global warming. *Journal of climate*, 19(21):5686–5699, 2006. 15

Thorsten Mauritsen and Erich Roeckner. Tuning the MPI-ESM1. 2 global climate model to improve the match with instrumental record warming by lowering its climate sensitivity. *Journal of Advances in Modeling Earth Systems*, 12(5):e2019MS002037, 2020. 15

John Mioduszewski, Stephen Vavrus, and Muyin Wang. Diminishing Arctic sea ice promotes stronger surface winds. *Journal of Climate*, 31(19):8101–8119, 2018. 16

Cécile Davrinche, Anaïs Orsi, Charles Amory, Christoph Kittel, and Cécile Agosta. Future changes in Antarctic near-surface winds: regional variability and key drivers under a high-emission scenario. *EGU-sphere*, 2025:1–32, 2025. 16

Ed Hawkins and Rowan Sutton. Time of emergence of climate signals. *Geophysical Research Letters*, 39(1), 2012. 16

Tim P Barnett and Michael E Schlesinger. Detecting changes in global climate induced by greenhouse gases. *Journal of Geophysical Research: Atmospheres*, 92(D12):14772–14780, 1987. 16

Filippo Giorgi and Xunqiang Bi. Time of emergence (TOE) of GHG-forced precipitation change hot-spots. *Geophysical Research Letters*, 36(6), 2009. 16

R. Ranasinghe, A.C. Ruane, R. Vautard, N. Arnell, E. Coppola, F.A. Cruz, S. Dessai, A.S. Islam, M. Rahimi, D. Ruiz Carrascal, J. Sillmann, M.B. Sylla, C. Tebaldi, W. Wang, and R. Zaaboul. Climate Change Information for Regional Impact and for Risk Assessment. In V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1767–1926. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.014. 17

Jingyuan Li, David WJ Thompson, Elizabeth A Barnes, and Susan Solomon. Quantifying the lead time required for a linear trend to emerge from natural climate variability. *Journal of Climate*, 30(24):10179–10191, 2017. 17

Louis Rivoire, Marianna Linz, and Jingyuan Li. Observational limitations to the emergence of climate signals. *Geophysical Research Letters*, 51(14):e2024GL109638, 2024. 17

Maialen Iturbide, José M Gutiérrez, Lincoln M Alves, Joaquín Bedia, Ruth Cerezo-Mota, Ezequiel Cimadevilla, Antonio S Cofiño, Alejandro Di Luca, Sergio Henrique Faria, Irina V Gorodetskaya, et al. An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*, 12(4):2959–2970, 2020. 17

Mara Baudena, Stefan C Dekker, Peter M van Bodegom, Barbara Cuesta, Steven I Higgins, Veiko Lehsten, Christian H Reick, Max Rietkerk, Simon Scheiter, Zun Yin, et al. Forests, savannas, and grasslands: bridging the knowledge gap between ecology and Dynamic Global Vegetation Models. *Biogeosciences*, 12(6):1833–1848, 2015. 17

Christian H Reick, Veronika Gayler, Daniel Goll, Stefan Hagemann, Marvin Heidkamp, Julia EMS Nabel, Thomas Raddatz, Erich Roeckner, Reiner Schnur, and Stiig Wilkenskjeld. JSBACH 3-The land component of the MPI Earth System Model: documentation of version 3.2. 2021. 17

Kimmo Ruosteenoja, Kirsti Jylhä, Jouni Räisänen, and Antti Mäkelä. Surface air relative humidities spuriously exceeding 100% in CMIP5 model output and their impact on future projections. *Journal of Geophysical Research: Atmospheres*, 122(18):9557–9568, 2017. 18

Deepak K Ray, James S Gerber, Graham K MacDonald, and Paul C West. Climate variation explains a third of global crop yield variability. *Nature communications*, 6(1):5989, 2015. 19, 21

Andrew R Bock, Lauren E Hay, Steven L Markstrom, Christopher Emmerich, and Marian Talbert. The US Geological Survey monthly water balance model futures portal. Technical report, US Geological Survey, 2017. 19, 21

Francesco Immorlano, Veronika Eyring, Thomas le Monnier de Gouville, Gabriele Accarino, Donatello Elia, Stephan Mandt, Giovanni Aloisio, and Pierre Gentine. Transferring climate change physical knowledge. *Proceedings of the National Academy of Sciences*, 122(15):e2413503122, 2025. 20

Mengze Wang, Benedikt Barthel Sorensen, and Themistoklis Sapsis. GEN2: A Generative Prediction-Correction Framework for Long-time Emulations of Spatially-Resolved Climate Extremes. *arXiv preprint arXiv:2508.15196*, 2025. 20

Lea Beusch, Zebedee Nicholls, Lukas Gudmundsson, Mathias Hauser, Malte Meinshausen, and Sonia I Seneviratne. From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: MAGICC (v7. 5.1)–MESMER (v0. 8.1) coupling. *Geoscientific Model Development Discussions*, 2021:1–26, 2021. 20

Christopher B Womack, Glenn Flierl, Shahine Bouabid, Andre N Souza, Paolo Giani, Sebastian D Eastham, and Noelle E Selin. A framework for assessing and understanding sources of error in Earth System Model emulation. *EGUsphere*, 2025:1–51, 2025b. 20

Paul A O’Gorman, Richard P Allan, Michael P Byrne, and Michael Previdi. Energetic constraints on precipitation under climate change. *Surveys in geophysics*, 33(3):585–608, 2012. 20

Andrew IL Williams, Philip Stier, Guy Dagan, and Duncan Watson-Parris. Strong control of effective radiative forcing by the spatial pattern of absorbing aerosol. *Nature Climate Change*, 12(8):735–742, 2022. 20

Karyn Tabor and John W Williams. Globally downscaled climate projections for assessing the conservation impacts of climate change. *Ecological Applications*, 20(2):554–565, 2010. 20

Colin R Mahony, Tongli Wang, Andreas Hamann, and Alex J Cannon. A global climate model ensemble for downscaled monthly climate normals over North America. *International Journal of Climatology*, 42(11):5871–5891, 2022. 20

Maybritt Schillinger, Maxim Samarin, Xinwei Shen, Reto Knutti, and Nicolai Meinshausen. EnScale: Temporally-consistent multivariate generative downscaling via proper scoring rules. *arXiv preprint arXiv:2509.26258*, 2025. 21

S. Hoyer and J. Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1), 2017. doi: 10.5334/jors.148. URL <https://doi.org/10.5334/jors.148>. 22

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert

Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. 22

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL <https://doi.org/10.1038/s41586-020-2649-2>. 22

Thomas A Caswell, Michael Droettboom, Antony Lee, John Hunter, Eric Firing, Elliott Sales De Andrade, Tim Hoffmann, David Stansby, Jody Klymak, Nelle Varoquaux, et al. matplotlib/matplotlib: Rel: v3.3.1. *Zenodo*, 2020. 22

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007. 22

Met Office. *Cartopy: a cartographic python library with a matplotlib interface*. Exeter, Devon, 2010 - 2015. URL <http://scitools.org.uk/cartopy>. 22

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL <https://doi.org/10.21105/joss.03021>. 22

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021. 22

DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kaputowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL <http://github.com/google-deepmind>. 22

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021. 22

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/jax-ml/jax>. 22