Appendix Outline

A. Related Work

A.1. Diffusion and Flow Matching Models

Diffusion models have become a cornerstone of generative modeling, framing synthesis as the reversal of a gradual noising process through learned score functions [19, 49]. Subsequent work extended this paradigm to conditional and guided settings, achieving state-of-the-art image generation [10] and later to spatiotemporal domains via video diffusion and latent video variants for efficient long-horizon synthesis [22, 60]. To further scale quality and resolution, cascaded diffusion pipelines decompose generation into multi-stage refinement processes, where coarse outputs are progressively upsampled by higher-resolution diffusion models [21, 47]. Flow-based models offer an alternative by learning invertible mappings from noise to data with exact likelihoods [11, 27], and the introduction of flow matching unified flow and diffusion training through deterministic ODE trajectories between distributions [32]. Recent hierarchical flow architectures extend this principle across scales: PyramidalFlow [24] and PixelFlow [8] use pyramids of flows at increasing spatial or spatiotemporal resolutions to more efficiently model video and image data respectively.

A.2. Generative Models for Climate

Generative AI has increasingly been applied in climate science to emulate expensive simulations and enhance spatial resolution. Diffusion models, in particular, have shown strong potential for generating realistic spatiotemporal climate fields. Several previous works have demonstrated that diffusion models can be used for downscaling, to enhance coarse reanalysis and model fields to higher resolutions Ling et al. [31], Srivastava et al. [50], Watt and Mansfield [56]. DiffESM applies conditional diffusion to produce daily climate variables consistent with coarse monthly means [2]. Climate in a Bottle (cBottle) proposes a two-stage diffusion framework that first synthesizes coarse 100 km atmospheric fields before applying a learned diffusion-based super-resolution to reach kilometer scales [4]. Spherical DYffusion introduces a weather-scale probabilistic emulator based on a dynamics-informed 'dyffusion' process paired with a spherical Fourier neural operator, enabling stable long-horizon global climate emulation of a simplified atmospheric model with fixed forcings [5].

B. Rescaling and Renoising for Probability Path Continuity

We derive the rescaling-renoising correction described in Equation 6 to handle the additional temporal dimension and support hetereogeneous resampling factors between stages.

B.1. Spatial-Only Derivation

We will start with the simpler spatial-only, homogeneous resampling setting presented in Section 2.1.3 and rederive the intermediate steps from Jin et al. [24] here for clarity. Following Jin et al. [24], we upsample the previous low-resolution endpoint using nearest-neighbor resampling, resulting in a linear combination of the inputs, which therefore follows a Gaussian distribution:

$$\mathrm{Up}_k(\hat{\boldsymbol{x}}^{e_{k+1}}) \mid \boldsymbol{x}_1 \sim \mathcal{N}\left(e_{k+1}\mathrm{Up}_k(\mathrm{Down}_{k+1}(\boldsymbol{x}_1)), (1-e_{k+1})^2\boldsymbol{\Sigma}\right), \tag{7}$$

with $\boldsymbol{\Sigma}$ is the covariance matrix induced by the upsampling operation.

To ensure continuity of the probability path between different stages of the spatial pyramid, the endpoints must have the same distributions. Eqs. (1), (2), and (7) show that the distributions of the endpoints are similar after a simple upsampling transformation:

$$\hat{\boldsymbol{x}}_{s_k} \mid \boldsymbol{x}_1 \sim \mathcal{N}\left(s_k\mathrm{Up}_k(\mathrm{Down}_{k+1}(\boldsymbol{x}_1)),\ (1-s_k)^2 I\right) \tag{8}$$

$$\mathrm{Up}_k(\hat{\boldsymbol{x}}_{e_{k+1}}) \mid \boldsymbol{x}_1 \sim \mathcal{N}\left(e_{k+1}\mathrm{Up}_k(\mathrm{Down}_{k+1}(\boldsymbol{x}_1)), (1-e_{k+1})^2\boldsymbol{\Sigma}\right). \tag{9}$$

This update rescales and re-noises the upsampled latent so that its distribution matches the start of the next stage. We

can therefore apply a linear transformation with a corrective Gaussian noise to match the distributions:

$$\hat{\boldsymbol{x}}_{s_k} = \frac{s_k}{e_{k+1}} \mathrm{Up}_k(\hat{\boldsymbol{x}}_{e_{k+1}}) + \alpha \boldsymbol{n}', \quad \text{where } \boldsymbol{n}' \sim \mathcal{N}(0, \boldsymbol{\Sigma}'). \tag{10}$$

The rescaling coefficient $\frac{s_k}{e_{k+1}}$ matches the means of these distributions, and α is the noise coefficient. To determine α and $\boldsymbol{\Sigma}'$, we need to match the covariance matrices of Eqs. (8) and (10)

$$\frac{s_k^2}{e_{k+1}^2}(1 - e_{k+1})^2 \boldsymbol{\Sigma} + \alpha^2 \boldsymbol{\Sigma}' = (1 - s_k)^2 \boldsymbol{I}. \tag{11}$$

Jin et al. [24] show that for a simple nearest neighbor upsampling operation, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ have blockwise structure that can be exploited to determine the noise correction to match the endpoint distributions.

B.2. Temporal and Hetegeneous Resampling Derivation

We now consider the setting presented in Section 2.2.1. We generalize these covariance matrices to any differing, per-stage resampling factors r_k^h, r_k^w, and r_k^t by allowing per-stage upsampling covariance matrices $\boldsymbol{\Sigma}_k$ and corrective noise covariance matrices $\boldsymbol{\Sigma}_k'$ as well as per-stage corrective noise coefficients α_k. Rewriting Eqs (10) and (11),

$$\hat{\boldsymbol{x}}_{s_k} = \frac{s_k}{e_{k+1}} \mathrm{Up}_k(\hat{\boldsymbol{x}}_{e_{k+1}}) + \alpha_k \boldsymbol{n}', \quad \text{where } \boldsymbol{n}' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k'). \tag{12}$$

$$\frac{s_k^2}{e_{k+1}^2}(1 - e_{k+1})^2 \boldsymbol{\Sigma}_k + \alpha_k^2 \boldsymbol{\Sigma}_k' = (1 - s_k)^2 \boldsymbol{I}. \tag{13}$$

Importantly, the per-stage covariance matrices remain $n_k \times n_k$ block diagonal with $n_k = r_k^h \cdot r_k^w \cdot r_k^t$.

$$(\boldsymbol{\Sigma}_k)_{\text{block}} = \mathbf{J}_n = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 1 \end{pmatrix},$$

$$\Rightarrow \quad (\boldsymbol{\Sigma}_k')_{\text{block}} = \begin{pmatrix} 1 & \gamma_k & \dots & \gamma_k \\ \gamma_k & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_k \\ \gamma_k & \dots & \gamma_k & 1 \end{pmatrix}. \tag{14}$$

where γ_k is a negative value in $[-\frac{1}{n_k - 1}, 0]$ for decorrelation, and its lower bound ensures that the covariance matrix is positive semidefinite i.e. forms an equicorrelation matrix.

We can then rewrite Eqs. (13) and (14) by equating their diagonal and non-diagonal elements respectively:

$$\frac{s_k^2}{e_{k+1}^2}(1 - e_{k+1})^2 + \alpha_k^2 = (1 - s_k)^2,$$

$$\frac{s_k^2}{e_{k+1}^2}(1 - e_{k+1})^2 + \alpha_k^2 \gamma_k = 0.$$

Since $0 < s_k, e_{k+1} < 1$, the equations are solvable with:

$$e_{k+1} = \frac{s_k \sqrt{1 - \gamma_k}}{(1 - s_k)\sqrt{-\gamma_k} + s_k \sqrt{1 - \gamma_k}}, \quad \alpha_k = \frac{1 - s_k}{\sqrt{1 - \gamma_k}}. \tag{15}$$

Intuitively, we want to preserve signals maximally at each jump point, which corresponds to minimizing the noise weight α_k. According to Eq. (15), this is equivalent to minimizing γ_k at each jump point. Substituting the minimum value $\gamma_k = -\frac{1}{n_k - 1}$ into Eq. (15) yields:

$$e_{k+1} = \frac{s_k \sqrt{n_k}}{(1 - s_k) + s_k \sqrt{n_k}}, \quad \alpha_k = (1 - s_k)\sqrt{\frac{n_k - 1}{n_k}}$$

As in Jin et al. [24], we find that $e_{k+1} > s_k$ in this generalized form, meaning the timestep is rolled back when adding the corrective noise at each jump point. Substituting this back into equation (12) yields

$$\begin{aligned} \hat{\boldsymbol{x}}_{s_k} &= \frac{s_k}{e_{k+1}} \mathrm{Up}_k(\hat{\boldsymbol{x}}_{e_{k+1}}) + \alpha_k \boldsymbol{n}' \\ &= \frac{(1 - s_k) + s_k \sqrt{n_k}}{\sqrt{n_k}} \mathrm{Up}_k(\hat{\boldsymbol{x}}_{e_{k+1}}) + (1 - s_k)\sqrt{\frac{n_k - 1}{n_k}} \boldsymbol{n}' \end{aligned}$$

matching the update rule presented in Eq. (6).

C. ClimateSuite Dataset

C.1. Climate Model and Scenario Selection

We select ten CMIP6 Earth system models to ensure broad coverage across model families, physical parameterizations, and scenario availability (Table S2). We include NorESM2-LM specifically because it is the reference model used in ClimateBench, enabling direct comparison with prior benchmarks. CESM2-WACCM and UKESM1-0-LL are two of the only ESMs for which fully coupled SAI intervention experiments have been conducted. The remaining models, namely BCC-CSM2-MR, CESM2, CMCC-CM2-SR5, CMCC-ESM2, GFDL-ESM4, IPSL-CM6A-LR, and MRI-ESM2-0, are widely used, well-validated CMIP6 models that together provide a diverse set of dynamical cores and physical parameterizations, increasing robustness to structural uncertainty. Importantly, each of these models provides simulations for historical and standard SSP scenarios (1-2.6/2-4.5/3-7.0/5-8.5) as well as multiple ensemble members. This collection therefore spans a representative

Dataset	Climate Models	Model × Scenarios	Model × Scenarios × Members	Intervention Scenarios × Members	Simulation Years	Resolution Spatial	Resolution Temporal
ClimateBench [55]	1	5	11	0	1,183	192 × 288	Monthly
ACE [57]	1	1	1	0	110	180 × 360	6-Hourly
ClimateSet [25]	21	104	108	0	10,672	192 × 288	Daily
ERA5 [18]	1	1	1	0	85	720 × 1440	Hourly
ClimateSuite (Ours)	10	66	345	69	33,739	192 × 288	Monthly

Table S1. **Comparison of ClimateSuite to existing climate-scale datasets.** We report the highest available spatial and temporal resolutions in each dataset.

Climate Model	Historical / SSP	SAI	Years
BCC-CSM2-MR	5	0	839
CESM2	5	0	2847
CESM2-WACCM	5	15	3522
CMCC-CM2-SR5	5	0	2159
CMCC-ESM2	5	0	509
GFDL-ESM4	5	0	1011
IPSL-CM6A-LR	5	0	8455
MRI-ESM2-0	5	0	4216
NorESM2-LM	5	0	2294
UKESM1-0-LL	5	1	7887
Total	**50**	**16**	**33,739**

Table S2. **Climate model and scenario breakdown in ClimateSuite.**

cross-section of CMIP6 modeling centers while supporting both standard climate scenarios and specialized SAI experiments.

We ensure broad coverage of standard climate trajectories and intervention scenarios (Table S3). We include historical simulations and four SSP scenarios spanning strong mitigation (SSP1-2.6), moderate emissions (SSP2-4.5), regional-rivalry–driven warming (SSP3-7.0), and high-emissions futures (SSP5-8.5). These scenarios collectively capture a range of plausible 21st-century forcing pathways used in CMIP6 and provide diverse data for training and evaluating models under a variety radiative environments. To study climate responses under SAI, we further incorporate historical and baseline control runs in addition to a variety of single-point, two-point, and multi-latitude injection strategies. These experiments span equatorial to high-latitude injections, fixed-mass versus controller-based deployment algorithms, and temperature targets ranging from 0.5°C to 1.5°C above pre-industrial levels. By grounding all intervention runs in a common SSP2-4.5 forcing scenario, we isolate the effect of aerosol injection while maintaining comparability across experimental designs. Together, this collection is a representative and scientifically comprehensive set of forcings that enables evaluation of models across conventional, extreme, and policy-relevant climate futures.

Despite this breadth, ClimateSuite is subject to potential biases. First, the selection of climate models, while diverse, is constrained by data availability and the existence of compatible experiments, which may overrepresent certain modeling centers, shared components, or parameterization choices within CMIP6. Second, the choice of emissions pathways and SAI intervention scenarios is not exhaustive and reflects a subset of commonly studied or readily available configurations. In particular, anchoring SAI experiments to SSP2-4.5 improves comparability but may limit coverage of interactions between interventions and alternative socioeconomic trajectories. These factors should be considered when interpreting results, as they influence the distribution of climate responses represented in the dataset.

Finally we note that ClimateSuite substantially overlaps ClimateSet in climate models, but offers broader scenario and ensemble coverage and better suitability for ML work due to its structure and preprocessing. For this reason, we use ClimateSuite not ClimateSet for our experiments.

C.2. Simulation Downloading

We acquire all datasets using publicly accessible portals and tools. We download external forcings from input4MIPs through the ESGF portal, ensuring consistency with the forcing datasets used in the original modeling center runs. For the standard CMIP6 historical and SSP simulations, we use ESMValCore [1] and acccmip6 [17] to automate search, download, and integrity checks across ten models. We use both tools to increase coverage because each exposes a partially overlapping subset of CMIP6 replica servers. For the SAI experiments, we retrieve simulations via Globus. We obtain the UKESM1-0-LL SAI outputs directly from the Met Office ARISE portal [37].

C.3. Data Processing

We process all data to standardized NetCDFs to provide an ML-friendly format while still facilitating common climate analysis. ClimateSuite provides all data at native resolution. For our experiments, we standardize the spatial resolution

Scenario	Description
historical	Standard historical simulation from 1850 to near-present using observed forcings
ssp126	Future scenario under SSP1-2.6, representing strong mitigation and low greenhouse-gas forcing
ssp245	Future scenario under SSP2-4.5, a middle-of-the-road pathway of moderate emissions
ssp370	Future scenario under SSP3-7.0 representing regional rivalry and higher forcing
ssp585	Future scenario under SSP5-8.5, representing high emissions and fossil-fuel driven development
MA-HISTORICAL	Historical baseline simulation (pre-intervention control run)
MA-BASELINE	Baseline simulation without any intervention or injection applied, under future forcing
SINGLE-POINT-INJANN0N_12Tg	Single-point injection at 0°N (equator), 12 Tg total
SINGLE-POINT-INJANN15S_12Tg	Single-point injection at 15°S, 12 Tg total
SINGLE-POINT-INJANN15N_12Tg	Single-point injection at 15°N, 12 Tg total
SINGLE-POINT-INJANN30N_12Tg	Single-point injection at 30°N, 12 Tg total
SINGLE-POINT-INJANN30S_12Tg	Single-point injection at 30°S, 12 Tg total
SINGLE-POINT-INJMAM60N_12Tg	Single-point injection at 60°N, 12 Tg total
SINGLE-POINT-INJSON60S_12Tg	Single-point injection at 60°S, 12 Tg total
SSP245-MA-GAUSS-DEFAULT	Injection with a controller-based algorithm at 15°S, 15°N, 30°S, 30°N to maintain temperature near 1.5°C above pre-industrial (PI) levels
SSP245-MA-GAUSS-LOWER-0.5	Injection with a controller-based algorithm targeting 0.5°C above PI
SSP245-MA-GAUSS-LOWER-1.0	Injection with a controller-based algorithm targeting 1.0°C above PI
SSP245-MA-GAUSS15N_15S-LOWER-0.5	Two-point injection at ±15°N/S targeting 0.5°C above PI
SSP245-MA-GAUSS30N_30S-LOWER-0.5	Two-point injection at ±30°N/S targeting 0.5°C above PI
SSP245-MA-GAUSS0N-LOWER-0.5	Single-point injection at 0°N targeting 0.5°C above PI
SAI-1.5	Multi-latitude injection targeting 1.5°C above PI

Table S3. **Standard emissions and intervention scenarios included in ClimateSuite.** Both historical simulations use CESM2-WACCM and all non-historical SAI simulations use CESM2-WACCM and SSP2-4.5 as the base forcing scenario, except for SAI-1.5 which uses UKESM1-0-LL (under SSP2-4.5).

among all climate models by regridding with bilinear interpolation to CESM2-WACCM resolution (192×288) and enforcing longitudinal periodicity which preserves large-scale spatial patterns to ensure physically smooth transitions between adjacent grid cells. We note that this selected resolution is higher than or comparable to all other model resolutions, so interpolation mostly involves mild upsampling and thus does not remove fine-grained structure

D. Evaluation Metrics

D.1. Preliminaries

Let $\mathbf{X} \in \mathbb{R}^{E\times I\times J}$ denote an ensemble of predictions (i.e. multiple samples from different starting noise, all using the same conditioning), and $\mathbf{Y} \in \mathbb{R}^{I\times J}$ the corresponding simulation targets, where E is the number of ensemble members, I the number of latitudes, and J the number of longitudes in the grid. Let $\overline{\mathbf{X}} = \frac{1}{E}\sum_{e=1}^{E}$ be the prediction averaged over ensemble members. Define $w(i)$ to be the normalized latitude-dependent area weight at latitude i, such that

$$\frac{1}{I}\sum_{i=1}^{I} w(i) = 1,$$

These weights account for the decreasing surface area of grid cells toward the poles. They are therefore used to compute spatially unbiased means and evaluation metrics, a standard practice in weather and climate prediction [5, 38, 44, 55].

D.2. Metric Definitions

We report the area-weighted average Root Mean Square Error (RMSE) and Bias of the member-averaged predictions

Model	Native Timescale	Yearly Runtime (s)	Monthly Runtime (s)
100M Parameters			
ClimaX Frozen [38]	Yearly	2	-
UNet [45]	Yearly	1	-
ClimaX [38]	Yearly	2	-
Pyramidal Flow [24]	Yearly	190	-
Multi-Yearly Flow (Ours)	Yearly	20	-
Multi-Monthly Flow (Ours)	Monthly	112	112
Pyramidal Flow [24]	Monthly	844	844
PixelFlow [8]	Yearly	60	-
SPF (Ours)	Multi	21	52
200M Parameters			
UNet [45]	Yearly	9	-
Pyramidal Flow [24]	Yearly	239	-
Multi-Yearly Flow (Ours)	Yearly	41	-
Pyramidal Flow [24]	Monthly	1054	1054
Multi-Monthly Flow (Ours)	Monthly	213	213
PixelFlow [8]	Yearly	71	-
SPF (Ours)	Multi	42	100

Table S4. **Per-timescale runtime for a single 100 year sample.**

compared to the targets as follows:

$$\text{Bias} = \frac{1}{IJ}\sum_{i,j} w(i)\left(\overline{\mathbf{X}}_{i,j} - \mathbf{Y}_{i,j}\right), \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{1}{IJ}\sum_{i,j} w(i)\left(\overline{\mathbf{X}}_{i,j} - \mathbf{Y}_{i,j}\right)^2}. \tag{17}$$

Values closer to zero are better for Bias and lower values are better for RMSE.

We additionally evaluate the full ensemble forecast using the unbiased version of the CRPS [36]:

$$\text{CRPS} = \frac{1}{IJ}\sum_{i,j} w(i)\left[\frac{1}{E}\sum_{e=1}^{E}\left|\mathbf{X}_{e,i,j} - \mathbf{Y}_{i,j}\right| - \frac{1}{2E(E-1)}\sum_{e=1}^{E}\sum_{f=1}^{E}\left|\mathbf{X}_{e,i,j} - \mathbf{X}_{f,i,j}\right|\right]. \tag{18}$$

The first term represents the accuracy (or skill) of the ensemble, while the second term quantifies its internal spread. The conventional, biased form of CRPS averages the spread using a factor of $\frac{1}{2E^2}$, which introduces bias for small ensembles. The unbiased formulation instead employs $\frac{1}{2E(E-1)}$, providing a more reliable measure of ensemble performance. Lower CRPS values indicate better forecasts. For deterministic models ($E = 1$), $\overline{\mathbf{X}}_{i,j} = \mathbf{X}_{i,j}$ and CRPS simplifies to the mean absolute error (MAE).

E. Additional Results

E.1. Long Sequence Runtimes

A major benefit of SPF's design is the capability of caching intermediate states to save computation for long sequences. We show this visually for a dummy example in Figure S1. To demonstrate this empirically, we report the runtimes for generating a 100-year sequence at both the yearly and monthly timescales (Table S4). Efficiency gains compared to the other models are further emphasized in this long sequence setting, with SPF achieving much faster results than all probabilistic models on both timescales except for Multi-Yearly flow at the yearly timescale, for which it nearly matches in runtime. Importantly, SPF can generate 100-year probabilistic samples of monthly data in 1-2 minutes, which is much faster than previous weather-scale autoregressive models (which take nearly 3 hours) and massively faster than the physical simulations (which take weeks to months).

E.2. Other Results

We report a variety of additional results, including tuning ClimaX settings (Table S5), ablating the spatial encodings (Table S6), global means of SPF compared to simulations in ClimateBench (Figures S2–S3), SPF histograms on ClimateBench (Figures S4–S5), SPF samples on ClimateBench (Figures S6–S16), multi-model global means of SPF compared to simulations in ClimateSuite (Figures S17), SPF histograms on ClimateSuite (Figures S18–S19), SPF samples of multiple models in ClimateSuite (Figures S20–S23), and an ablation measuring the impact of using variable number of ensemble members on SPF performance on ClimateBench (Figure S24).

E.3. Error Analysis

Across the yearly sample comparisons (Figs. S6–S9), we observe that SPF captures large-scale spatial patterns well for both temperature and precipitation, with errors concentrated in known challenging regions. For precipitation, biases are most pronounced over the tropical Pacific, particularly the western Pacific warm pool, where variability is high and strongly influenced by ocean-atmosphere coupling and multi-year modes such as ENSO. These regions are also known to be difficult for Earth system models themselves, suggesting that SPF may inherit intrinsic uncertainty from the underlying dynamics. For temperature, errors are generally small in magnitude but become more noticeable at high latitudes, where nonlinear feedbacks (e.g., ice–albedo) and longer equilibration timescales increase modeling difficulty. Overall, precipitation errors tend to reflect slight magnitude biases, while temperature errors exhibit higher spatial variance, especially in polar regions, as further quantified in Fig. S14.

The uncertainty estimates (Figs. S10–S13), computed as the standard deviation across multiple samples with identical forcings but different initial noise, align closely with these error patterns. Regions with higher bias like tropical precipitation zones and high-latitude temperature regions also exhibit elevated uncertainty, indicating that SPF's stochastic sampling provides a meaningful proxy for epis-

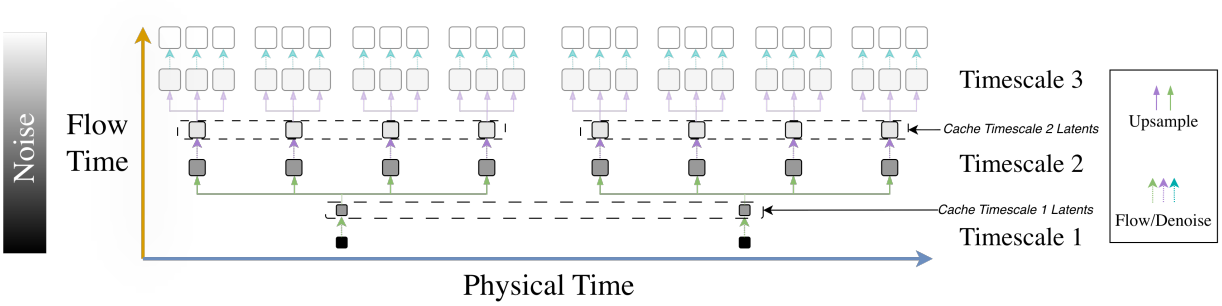

Figure S1. **Efficiency benefits from caching.** For long sequence generation, intermediate latents at coarser timescales can be cached to save compute when sampling from finer timescales. In this toy example, to generate a sequence of eight fine timescale samples, the Timescale 1 flow only needs to be run once and the Timescale 2 flow only needs to be run twice. Then, the Timescale 3 flow only needs to be run when generating each of the eight clean Timescale 3 samples as the flow can resume from the Timescale 2 latents.

Procedure	Resolution	Patch Size	RMSE ↓
Frozen	Low	2	0.811
Frozen	Low	16	0.721
Frozen	High	4	0.721
Frozen	High	16	0.619
Fine-Tuned	Low	16	0.594
Fine-Tuned	Low	2	0.577
Fine-Tuned	High	4	0.556
Fine-Tuned	High	16	0.546

Table S5. **ClimateBench yearly temporal resolution results under different ClimaX variants.**

temic and aleatoric uncertainty. Notably, precipitation uncertainty is generally lower in magnitude, consistent with the smaller forced signal relative to internal variability, yet still expands in regions of strong convection. The normalized bias and uncertainty analysis (Fig. S14) further shows that these relationships remain stable across temporal horizons. Importantly, uncertainty tends to correlate with bias in most regions, especially for precipitation.

F. Conclusion and Broader Impact

We introduce SPF, a generalization of spatial pyramid flows to joint spatiotemporal cascades with heterogeneous resampling. SPF not only outperforms prior spatial-only and autoregressive models, but also enables direct, efficient sampling at multiple timescales. Our ablations validate SPF's design, as it outperforms alternate pyramid designs and autoregressive variants. We hope our approach inspires future approaches for efficient climate emulation as well as methods for other simulation modeling tasks like weather forecasting and fluid dynamics.

Encoding	CRPS ↓	RMSE ↓
Linear(SH)	0.506	1.292
Siren(SH)	0.481	1.236
Siren(D)	0.479	1.099
Fourier (SPF)	**0.453**	**1.060**

Table S6. **Ablation comparing geographic encodings to our spatial encoding design.** SPF's encodings outperform several other variants [46], supporting SPF's design. We note geographic encodings usually help for dense grids ($\approx 10^3$) not the coarse 24×32 patchified grid we use for climate data.

Timescale	Obj.	CRPS ↓	RMSE ↓	Time
Monthly	Diff.	0.457	1.071	50s
	Flow	**0.453**	**1.060**	27s
Yearly	Diff.	0.229	0.521	10s
	Flow	**0.222**	**0.511**	6s

Table S7. **Ablation comparing flow vs. diffusion objective.** Replacing the flow matching objective of our monthly model with a diffusion objective leads to a small decrease in performance, demonstrating improvement of our flow-based approach over generative variants beyond just flow-based variants.

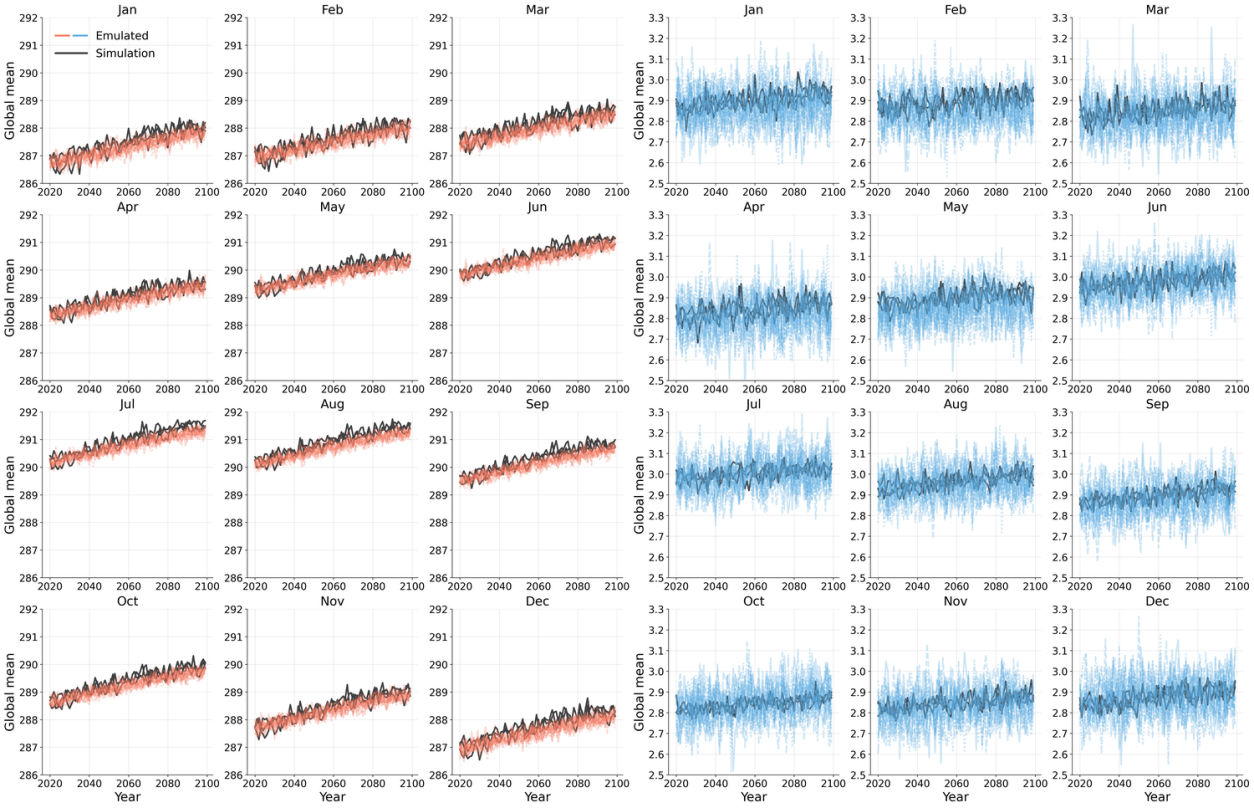

Figure S2. **Monthly latitude-weighted global means of the 200M SPF on ClimateBench SSP2-4.5.** Temperature shown in the left three columns (red) and precipitation in the right three columns (blue).

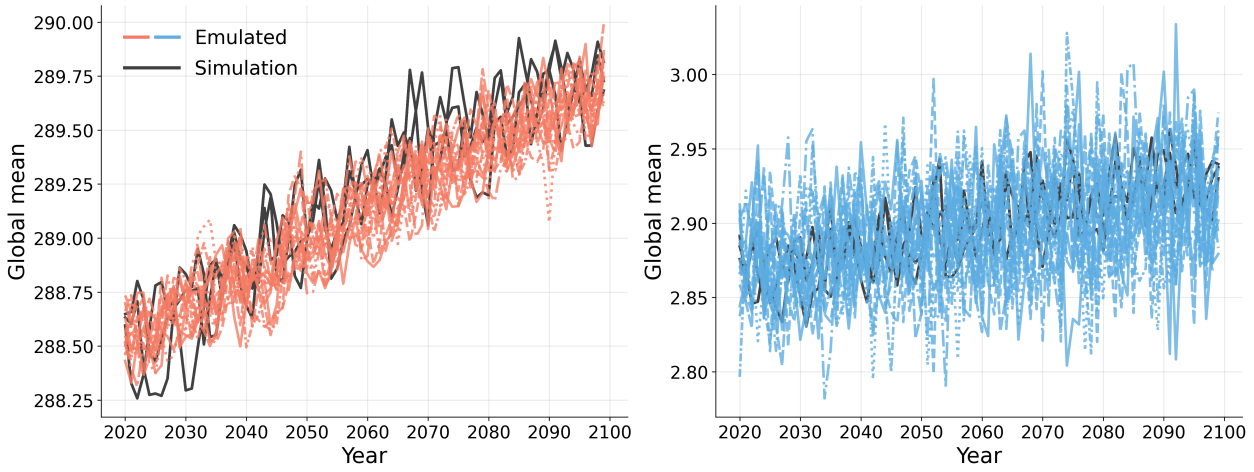

Figure S3. **Yearly latitude-weighted global means of the 200M SPF on ClimateBench SSP2-4.5.** Temperature shown in the left column (red) and precipitation in the right column (blue).

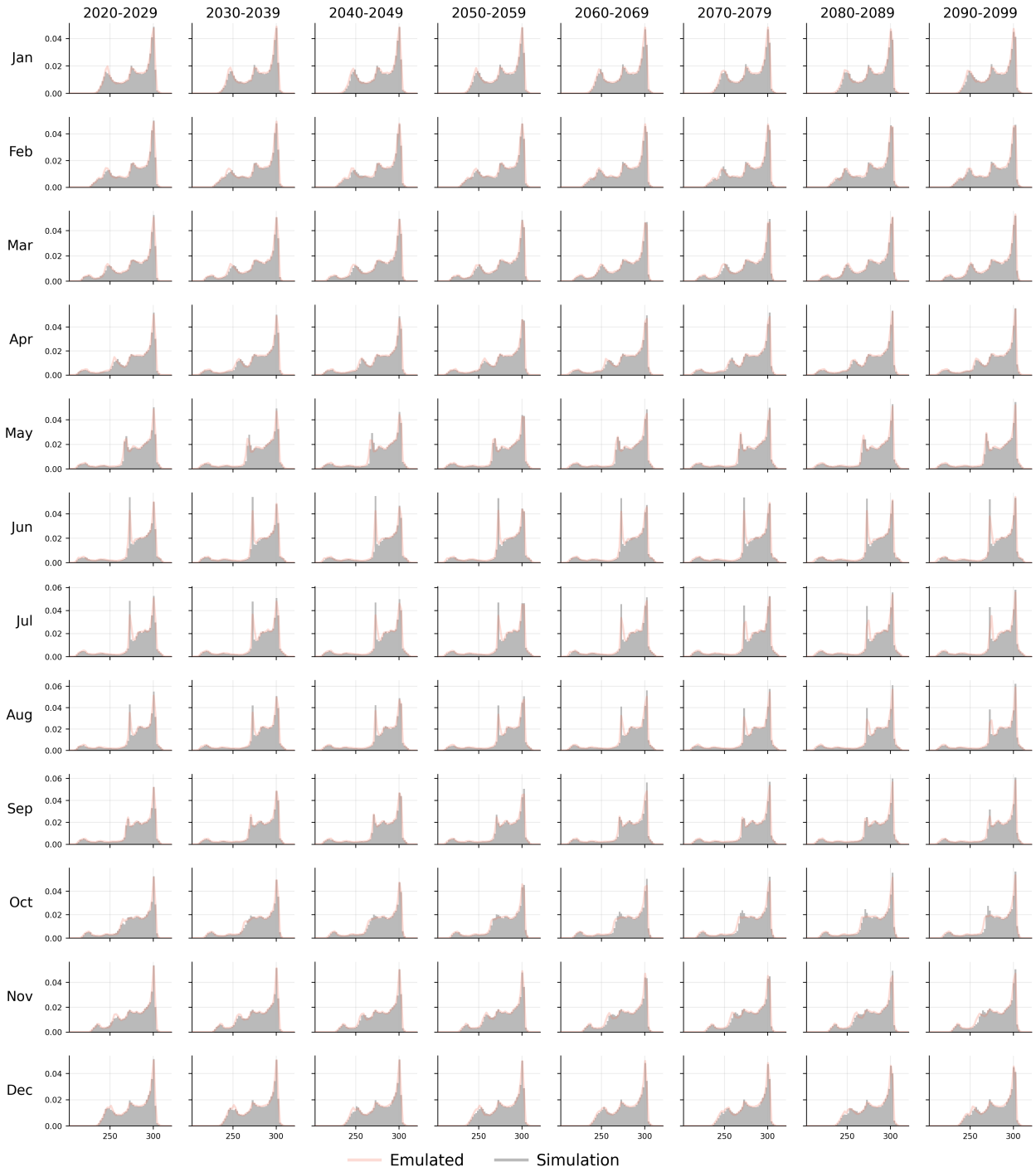

Figure S4. **Monthly temperature histograms of the 200M SPF model on ClimateBench SSP2-4.5, broken down by month and decade.**

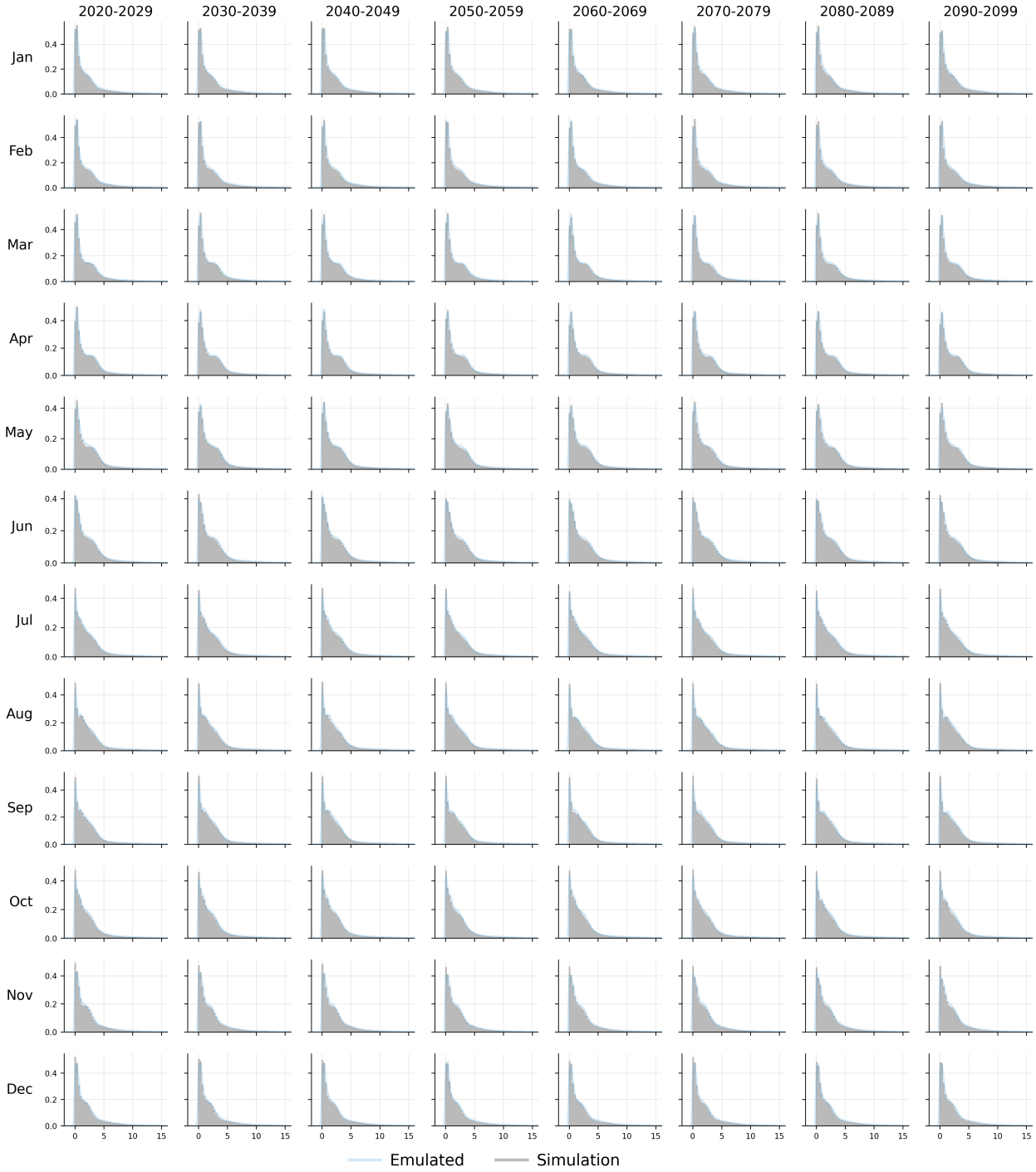

Figure S5. **Monthly precipitation histograms of the 200M SPF model on ClimateBench SSP2-4.5, broken down by month and decade.**

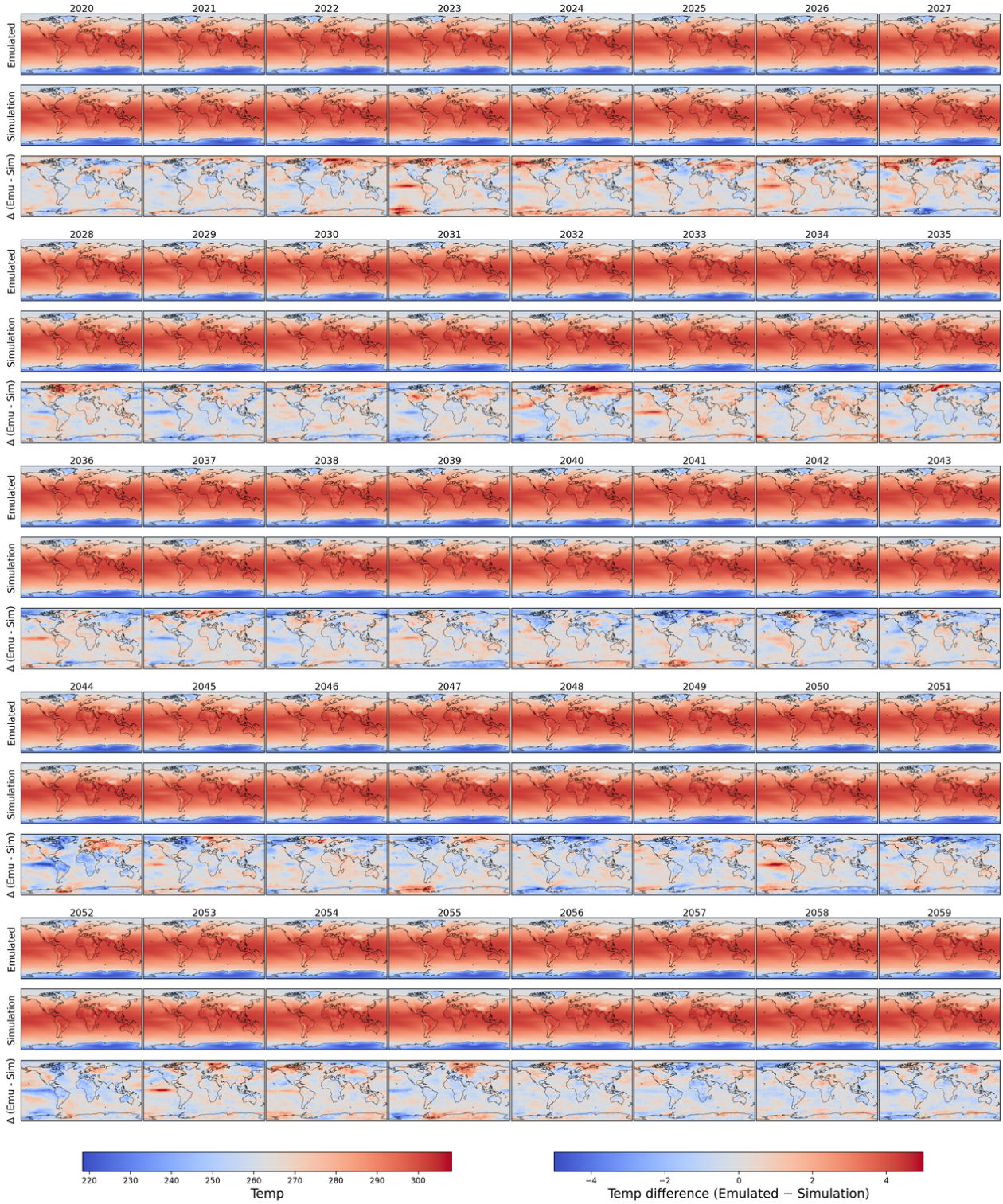

Figure S6. **ClimateBench samples of yearly emulated, simulated, and difference temperature maps (2020–2059, SSP2-4.5).**

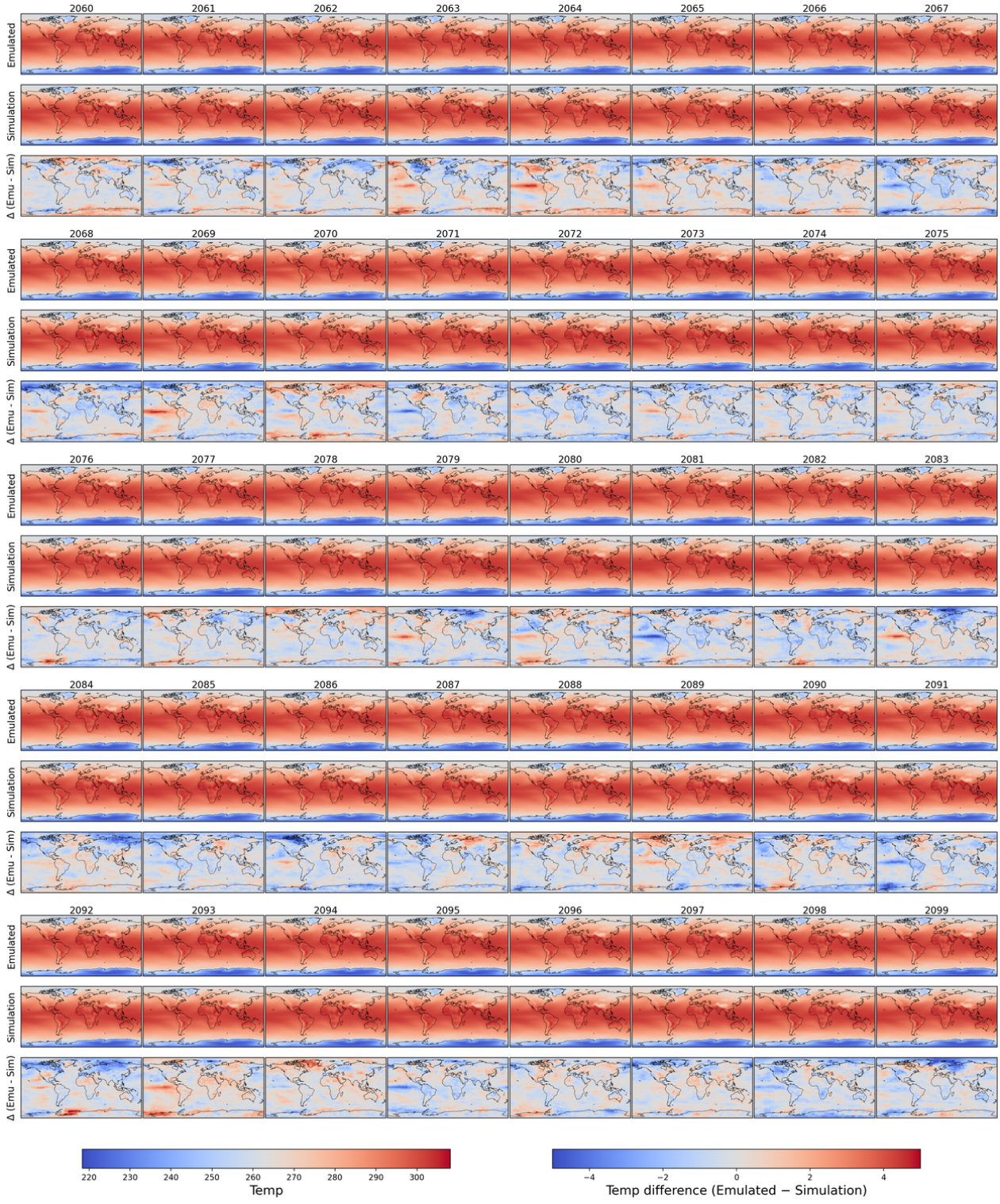

Figure S7. **ClimateBench samples of yearly emulated, simulated, and difference temperature maps (2060–2099, SSP2-4.5).**

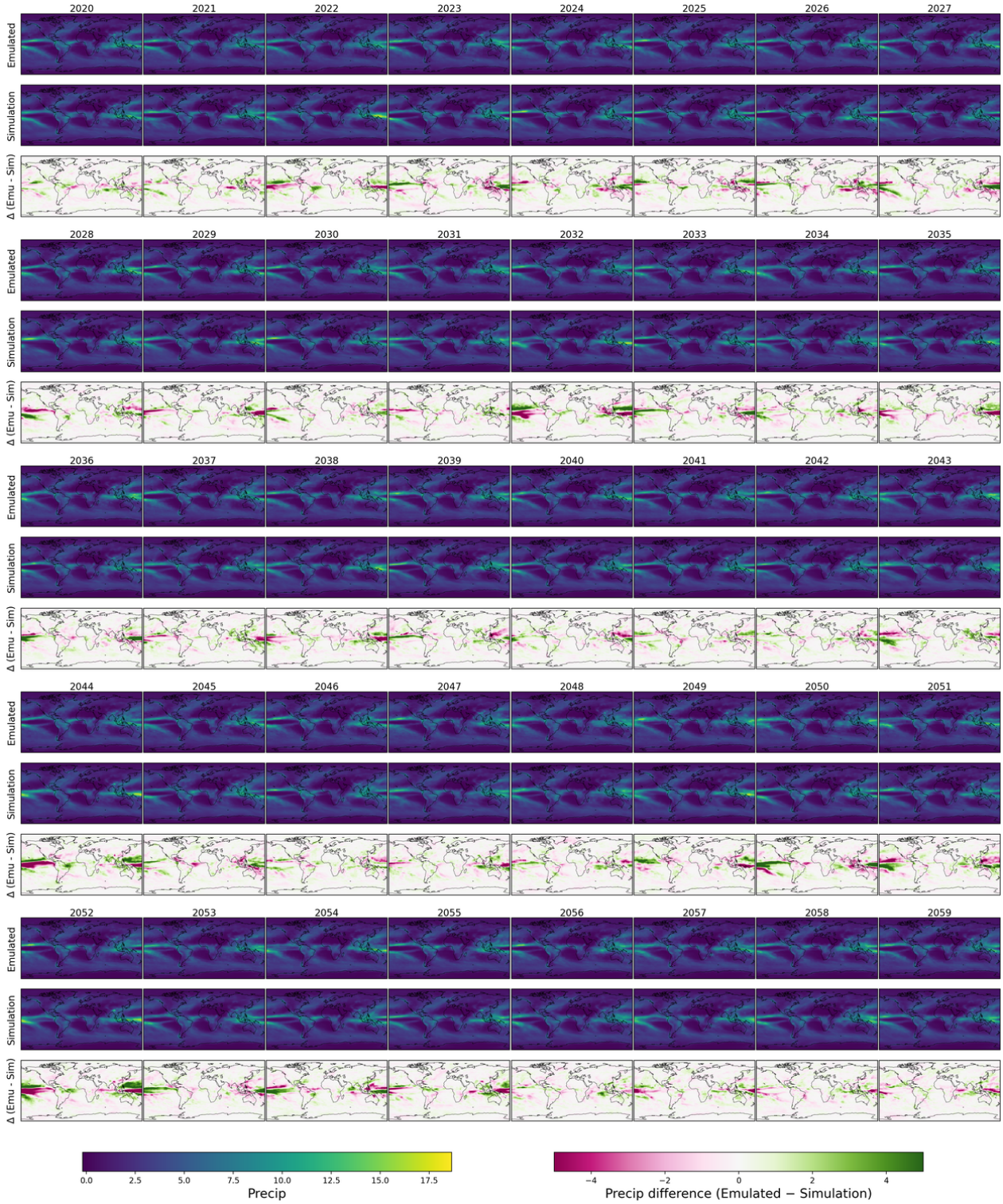

Figure S8. **ClimateBench samples of yearly emulated, simulated, and difference precipitation maps (2020–2059, SSP2-4.5).**

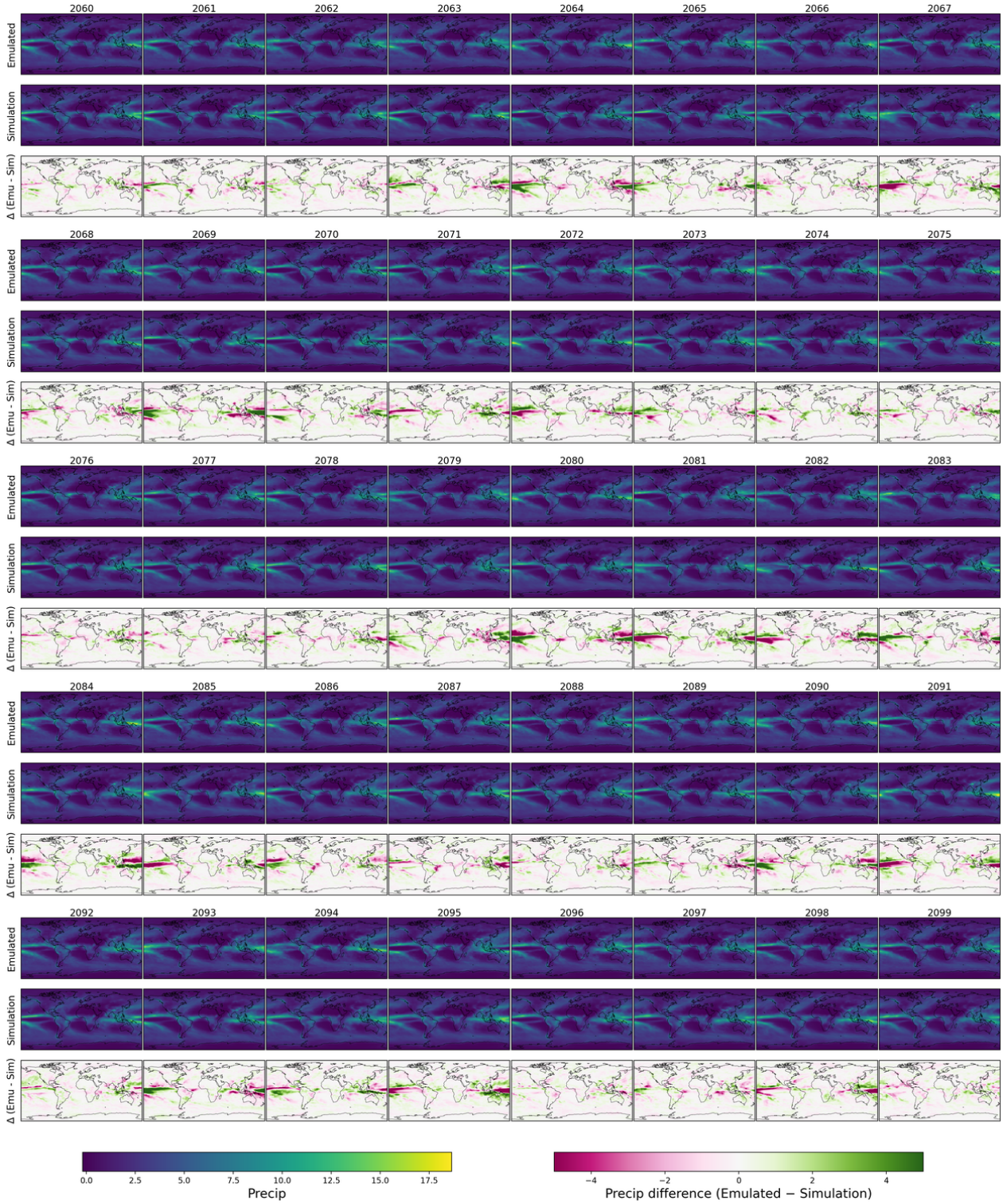

Figure S9. **ClimateBench samples of yearly emulated, simulated, and difference precipitation maps (2060–2099, SSP2-4.5).**

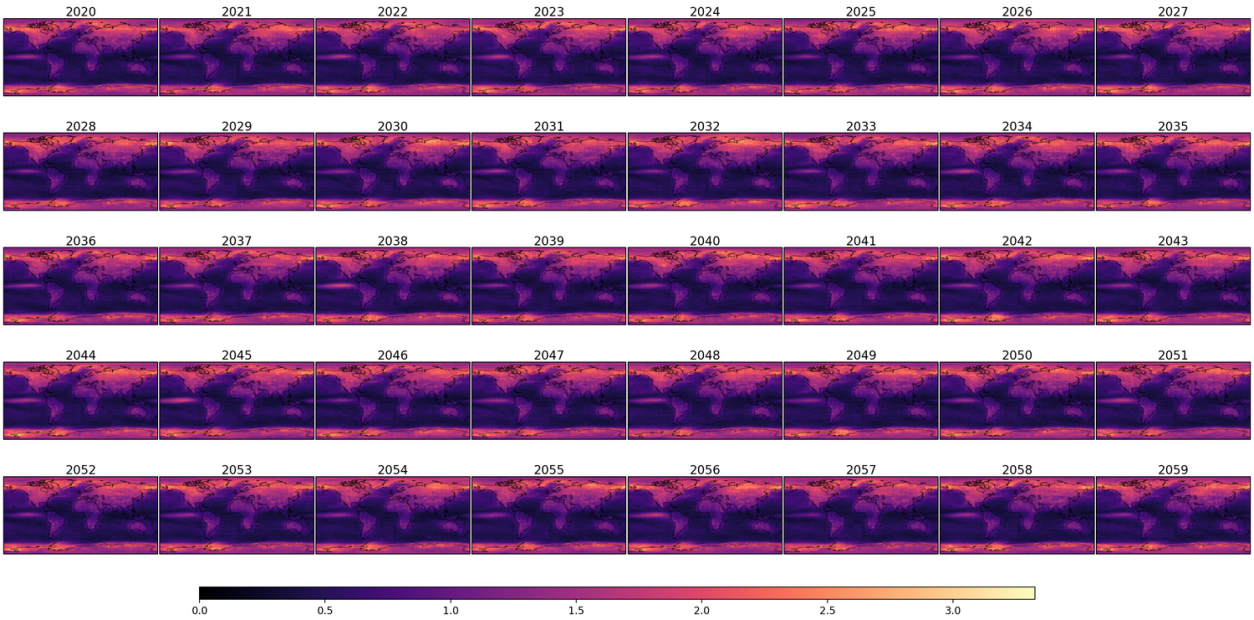

Figure S10. **SPF uncertainty for yearly temperature maps (2020–2059, SSP2-4.5).** Uncertainty is computed using the standard deviation across samples with different starting noise but the same forcings.

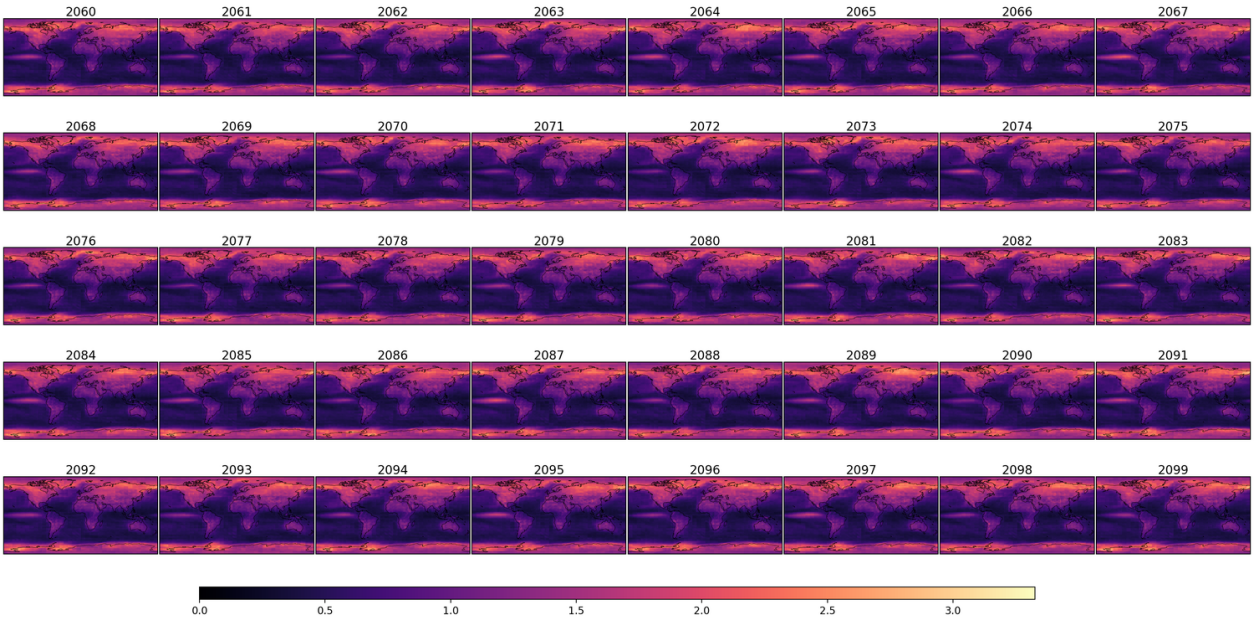

Figure S11. **SPF uncertainty for yearly temperature maps (2060–2099, SSP2-4.5).** Uncertainty is computed using the standard deviation across samples with different starting noise but the same forcings.

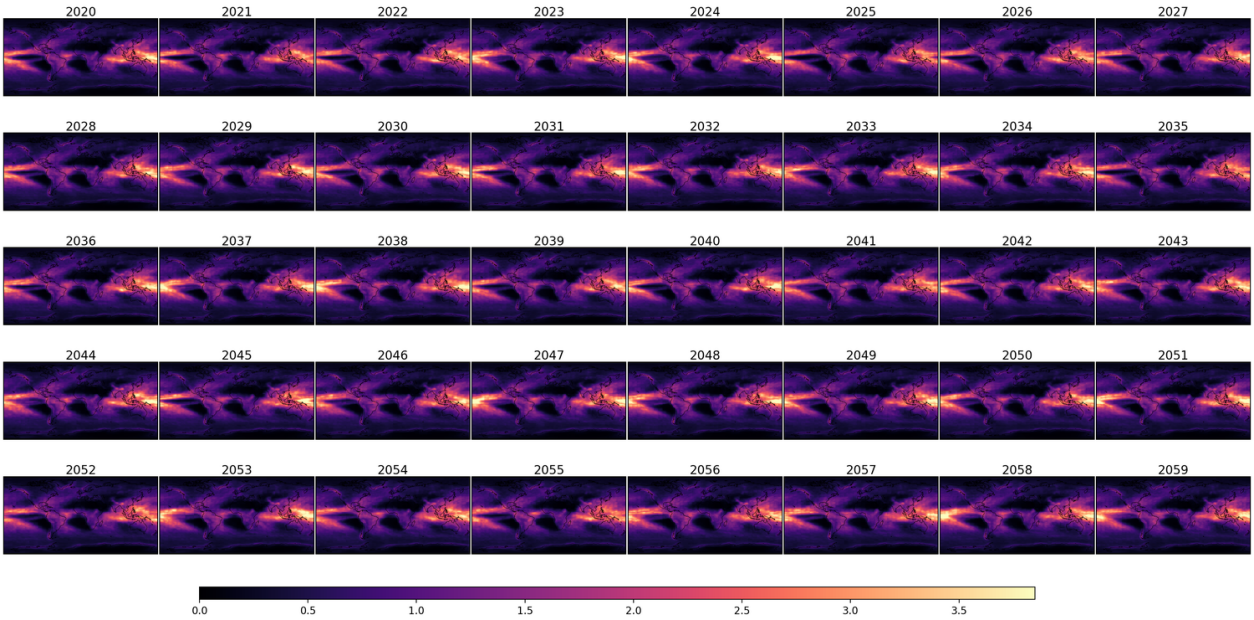

Figure S12. **SPF uncertainty for yearly precipitation maps (2020–2059, SSP2-4.5).** Uncertainty is computed using the standard deviation across samples with different starting noise but the same forcings.

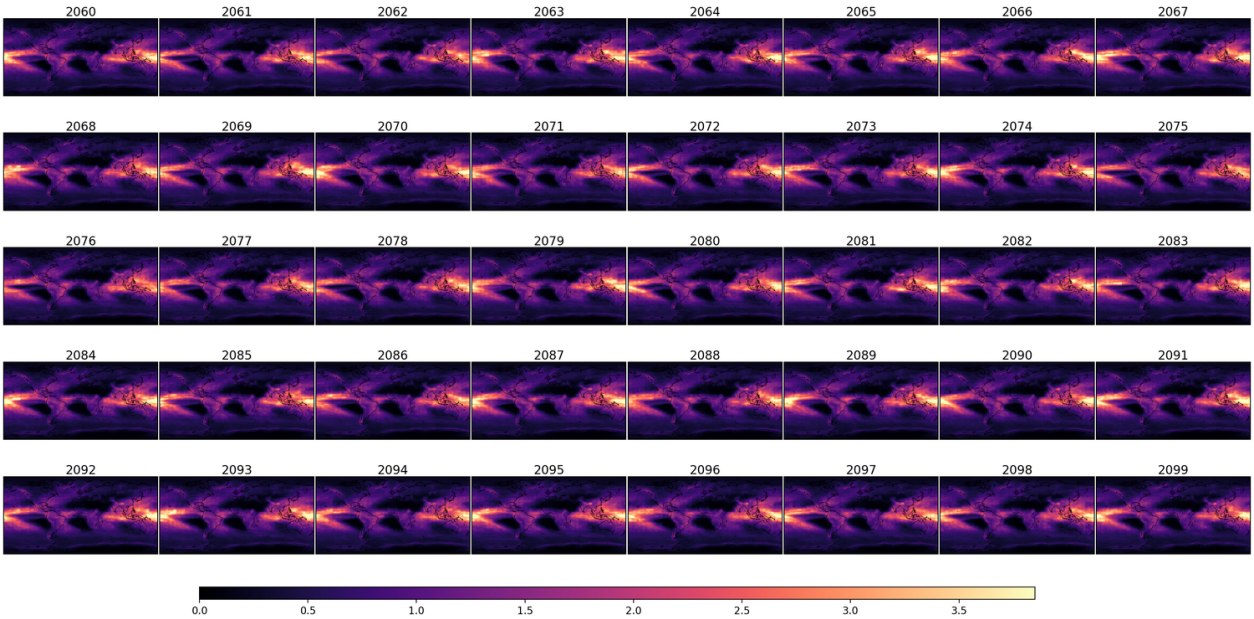

Figure S13. **SPF uncertainty for yearly precipitation maps (2060–2099, SSP2-4.5).** Uncertainty is computed using the standard deviation across samples with different starting noise but the same forcings.

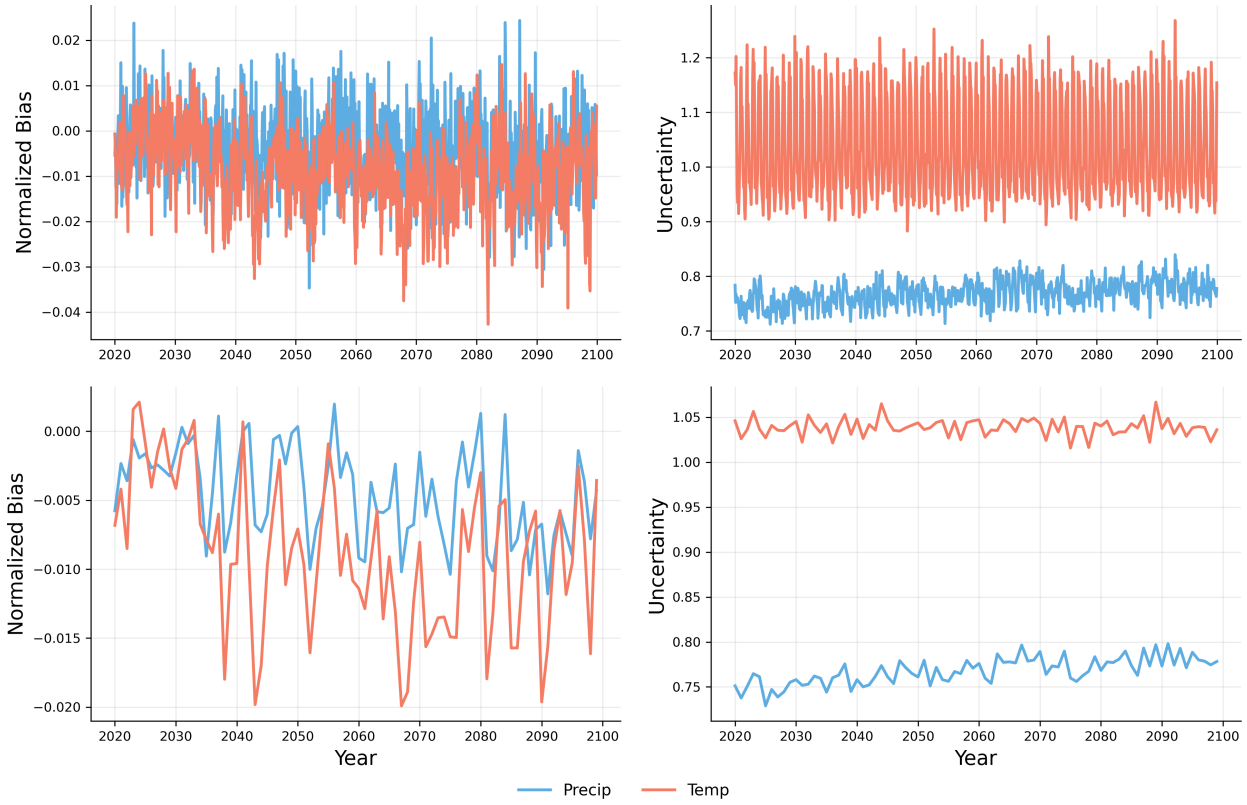

Figure S14. **Normalized bias and uncertainty across temporal resolutions.** The left column shows normalized bias (global mean prediction minus ground truth, scaled by the global standard deviation of the truth), and the right column shows normalized uncertainty (standard deviation across samples with different starting noise). The top row corresponds to monthly resolution, and the bottom row to yearly resolution.

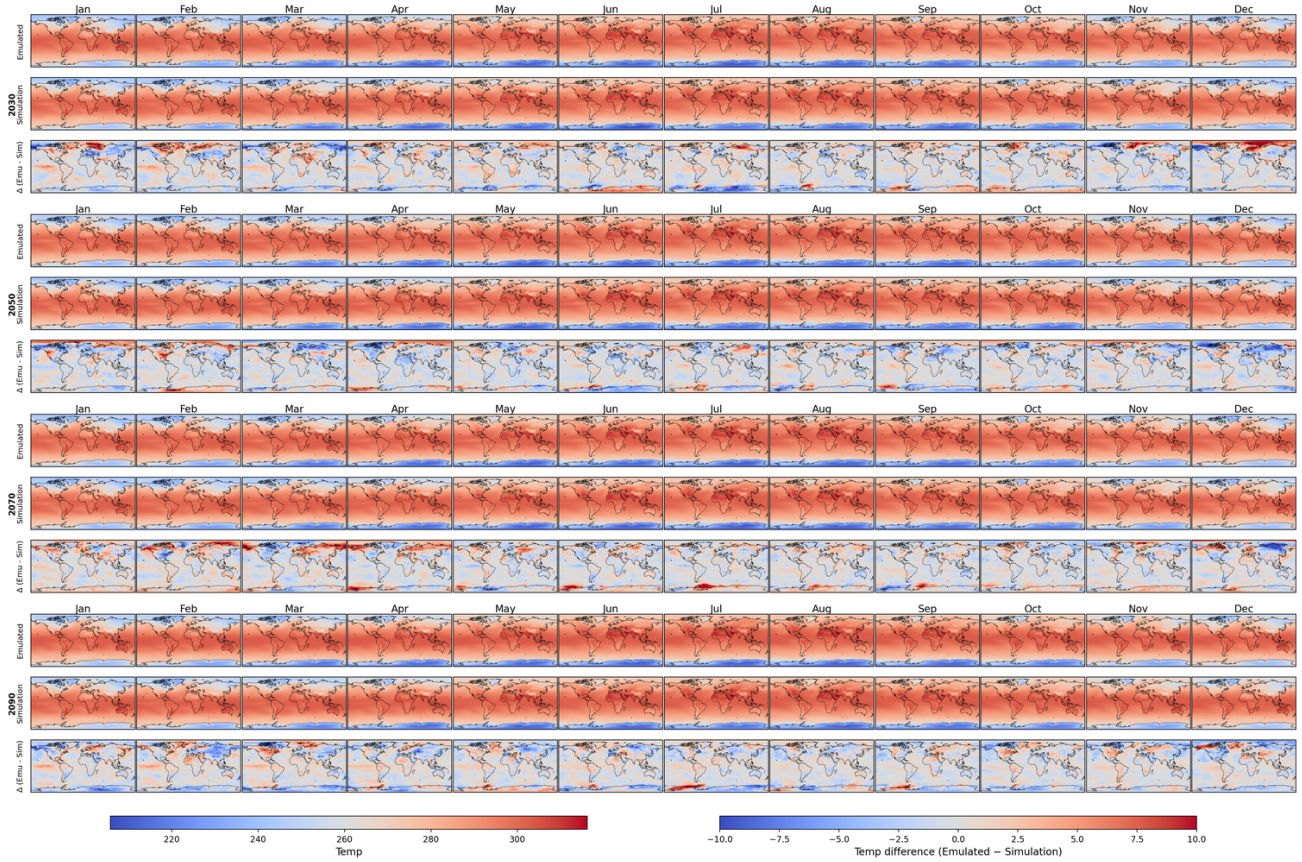

Figure S15. **ClimateBench samples of monthly emulated, simulated, and difference temperature maps (SSP2-4.5).**

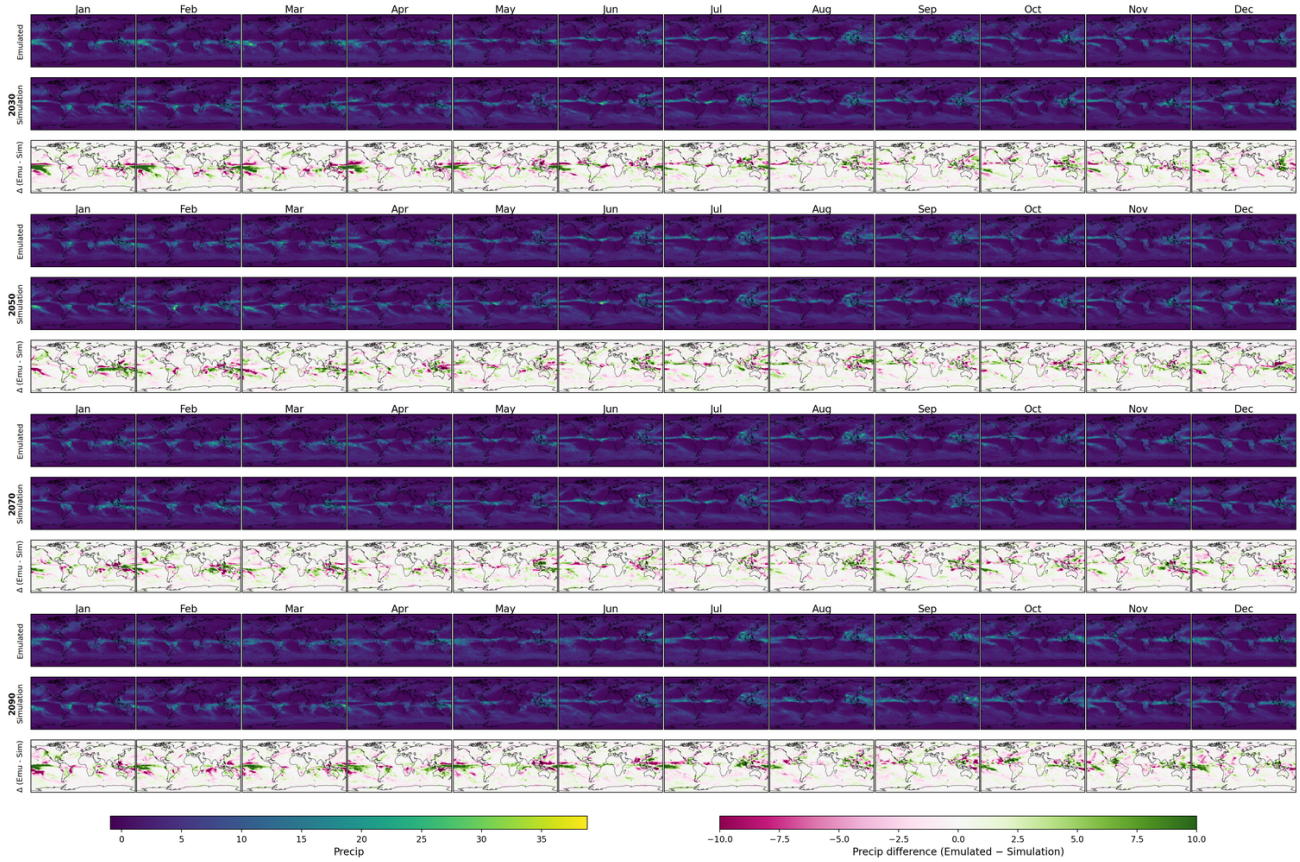

Figure S16. **ClimateBench samples of monthly emulated, simulated, and difference precipitation maps (SSP2-4.5).**

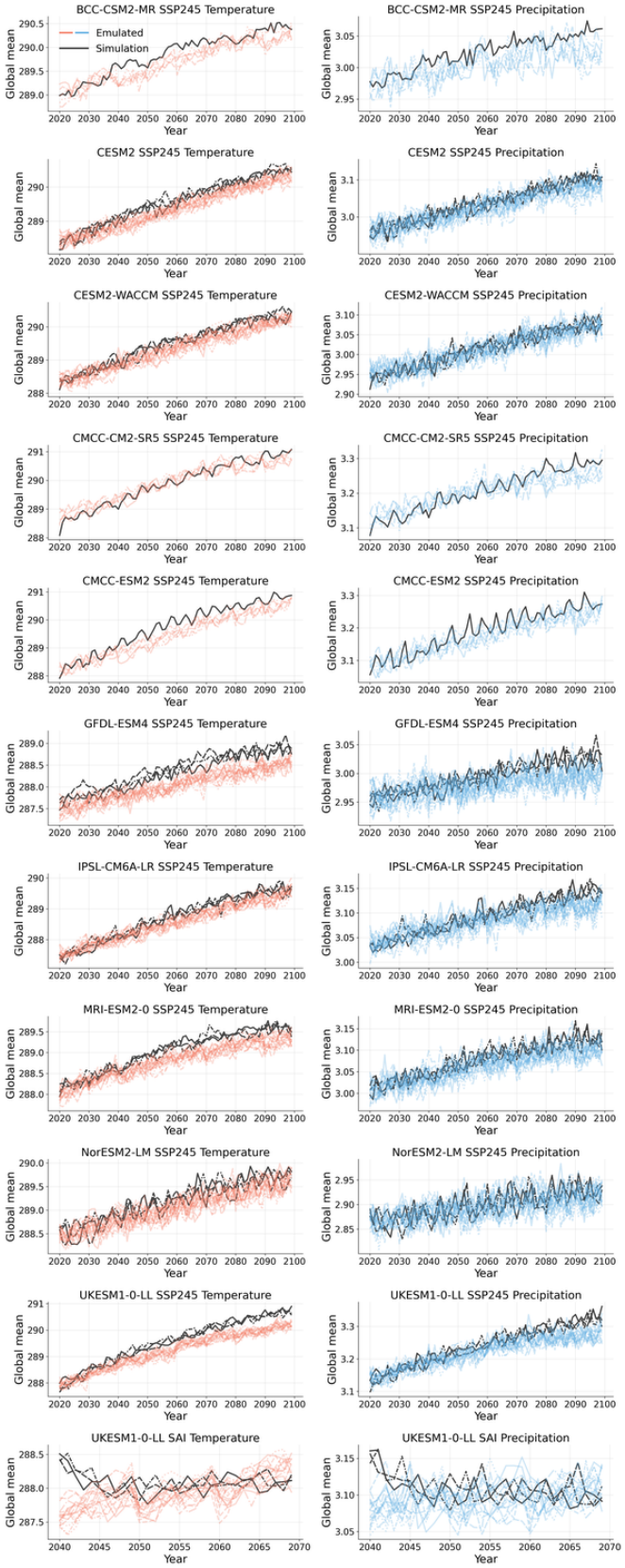

Figure S17. **Yearly latitude-weighted global means of the 600M SPF model on SSP2-4.5 across the 10 climate models and the SAI experiment on UKESM1-0-LL in ClimateSuite.**22

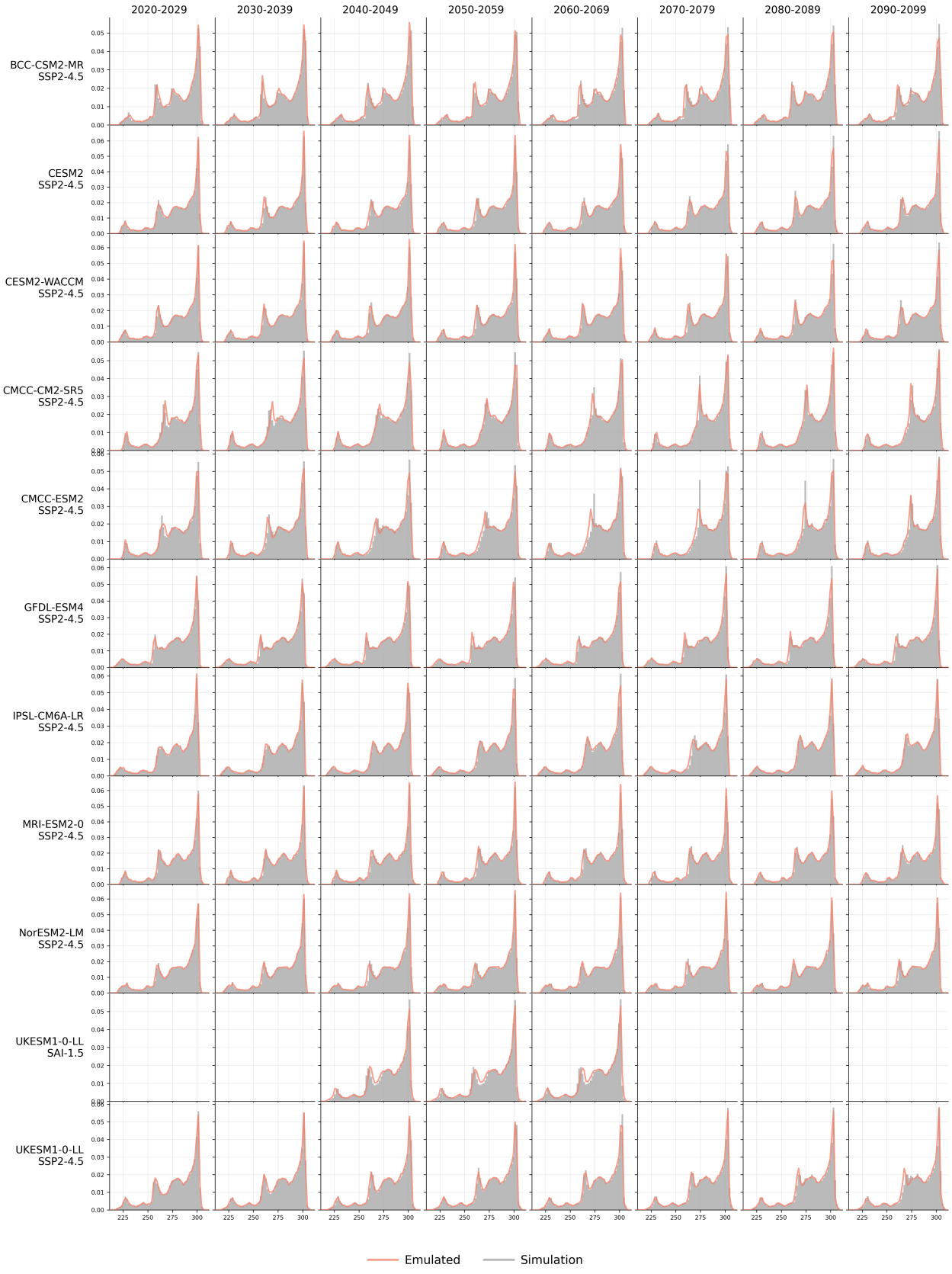

Figure S18. **Yearly temperature histograms of the 600M SPF model on SSP2-4.5 across the 10 climate models and the SAI experiment on UKESM1-0-LL, broken down by decade.**

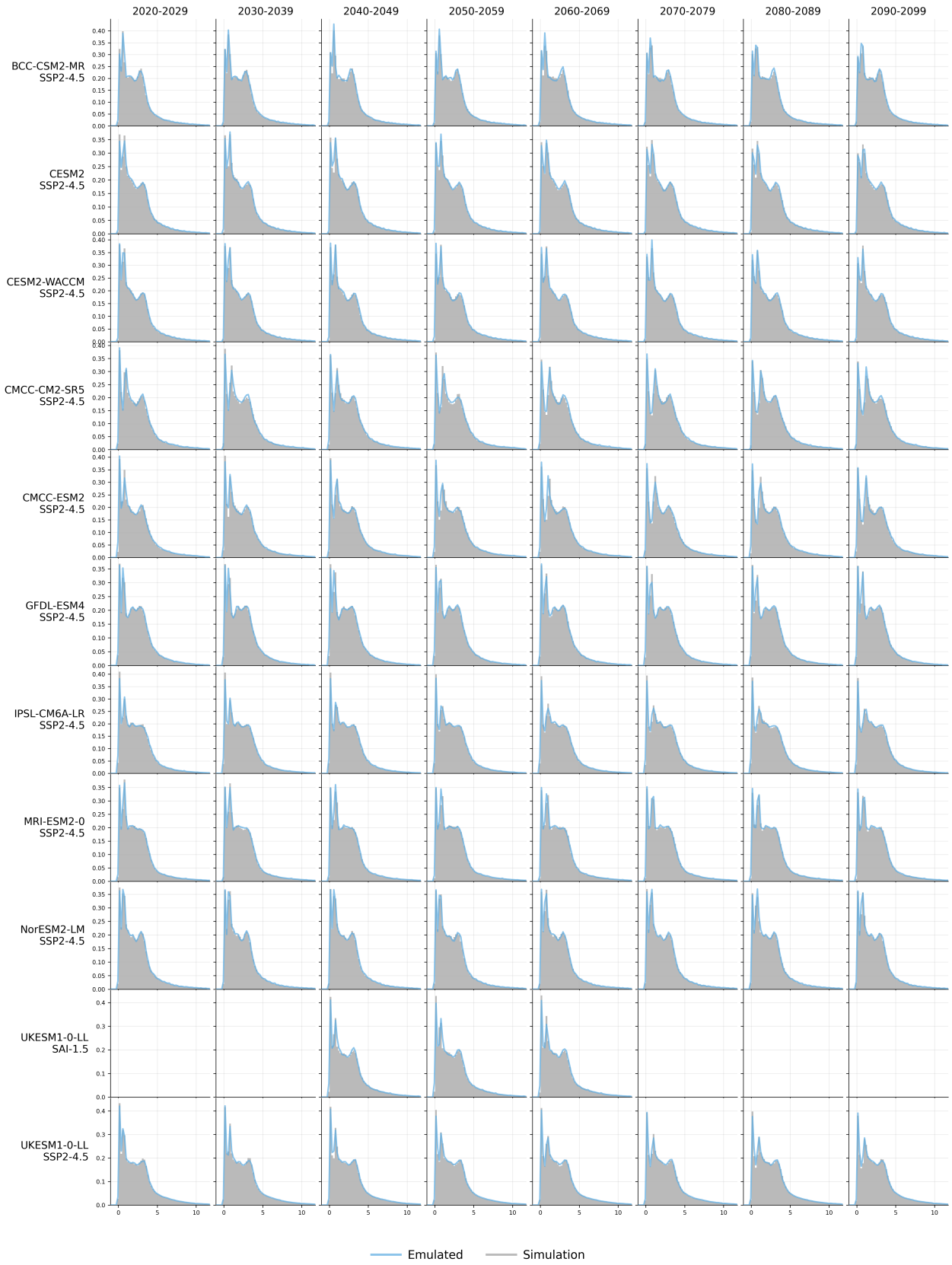

Figure S19. **Yearly precipitation histograms of the 600M SPF model on SSP2-4.5 across the 10 climate models and the SAI experiment on UKESM1-0-LL, broken down by decade.**

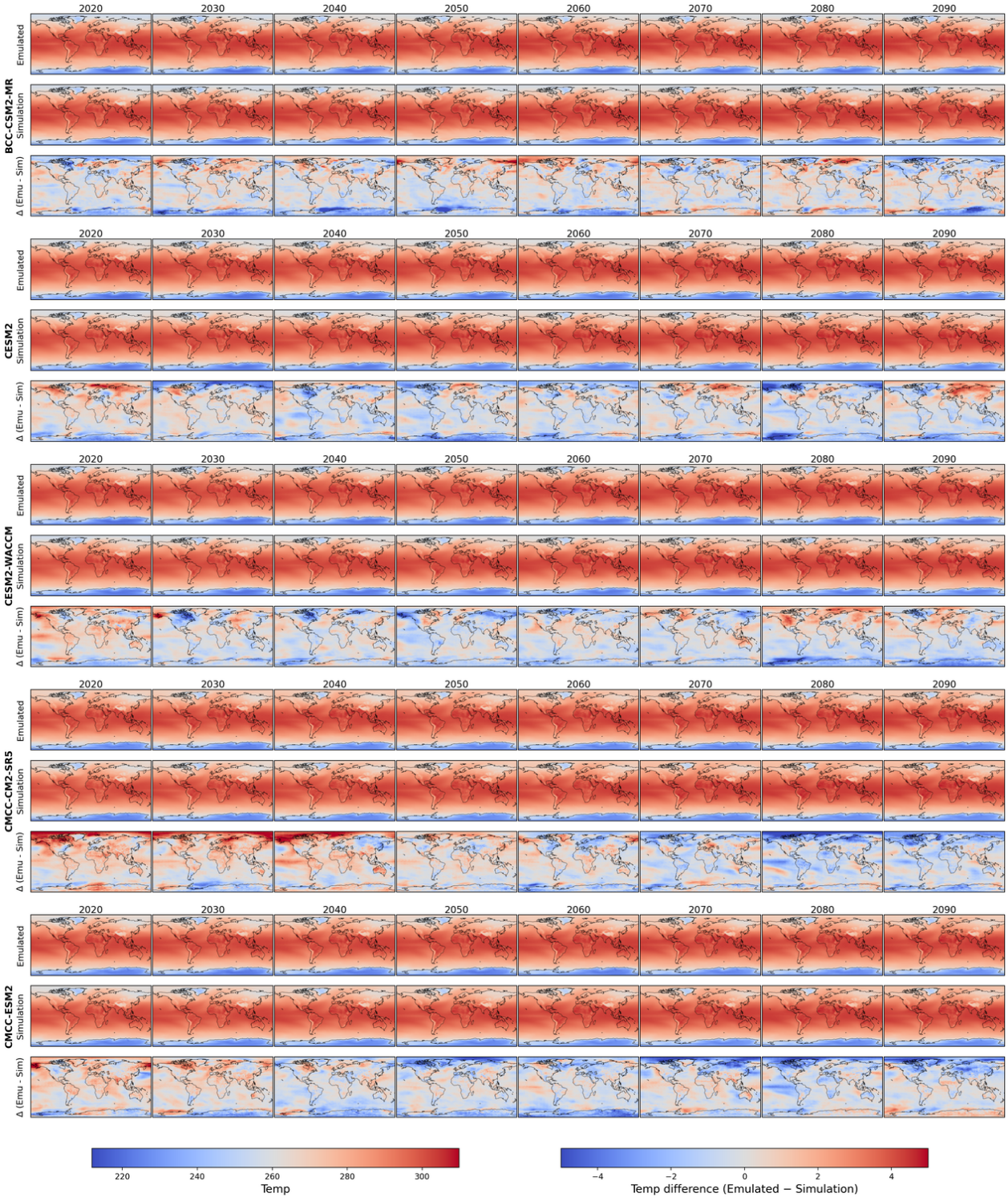

Figure S20. **Per-model ClimateSuite samples of yearly emulated, simulated, and difference temperature maps (SSP2-4.5).**

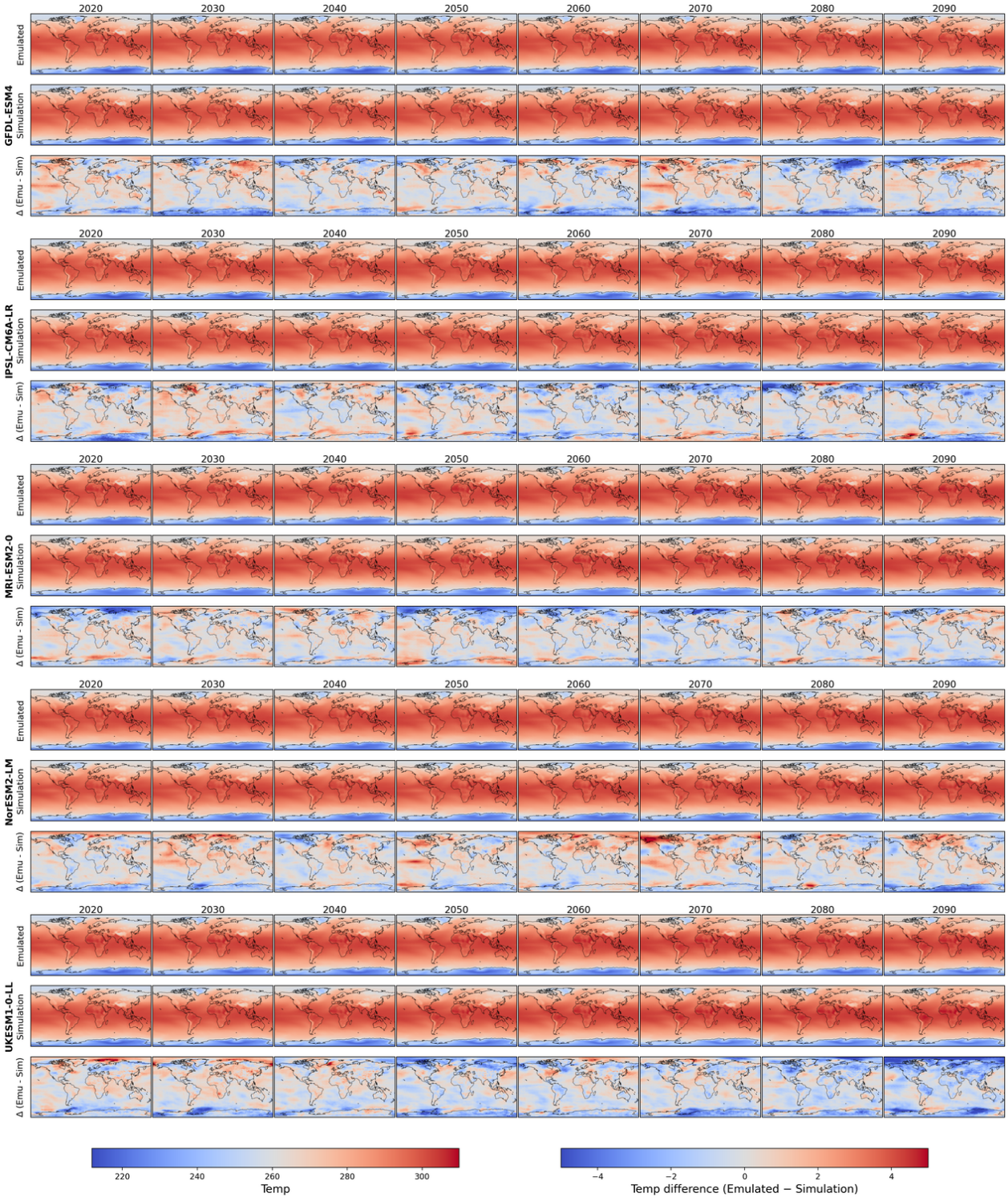

Figure S21. **Per-model ClimateSuite samples of yearly emulated, simulated, and difference temperature maps (SSP2-4.5).**

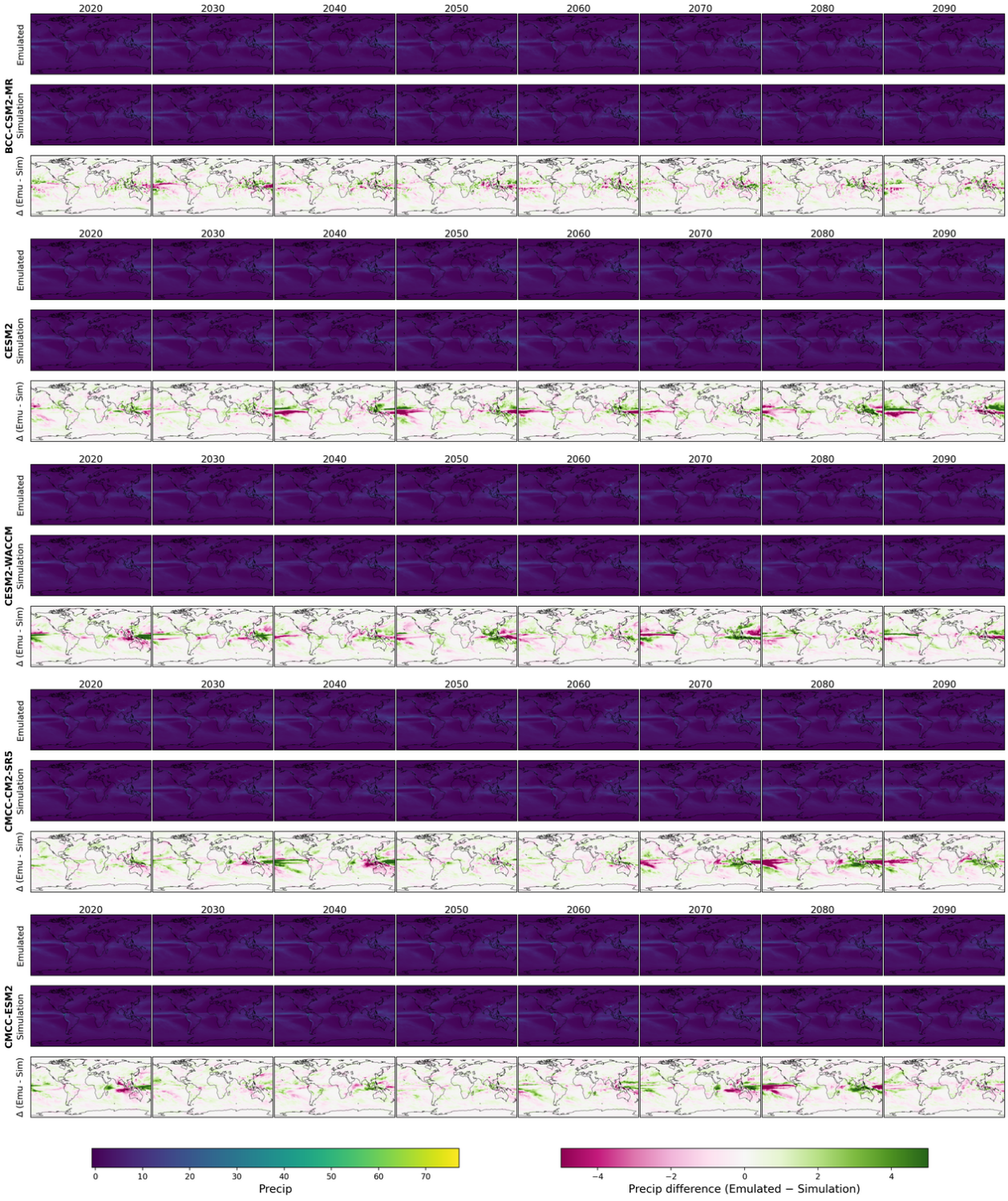

Figure S22. **Per-model ClimateSuite samples of yearly emulated, simulated, and difference precipitation maps (SSP2-4.5).**

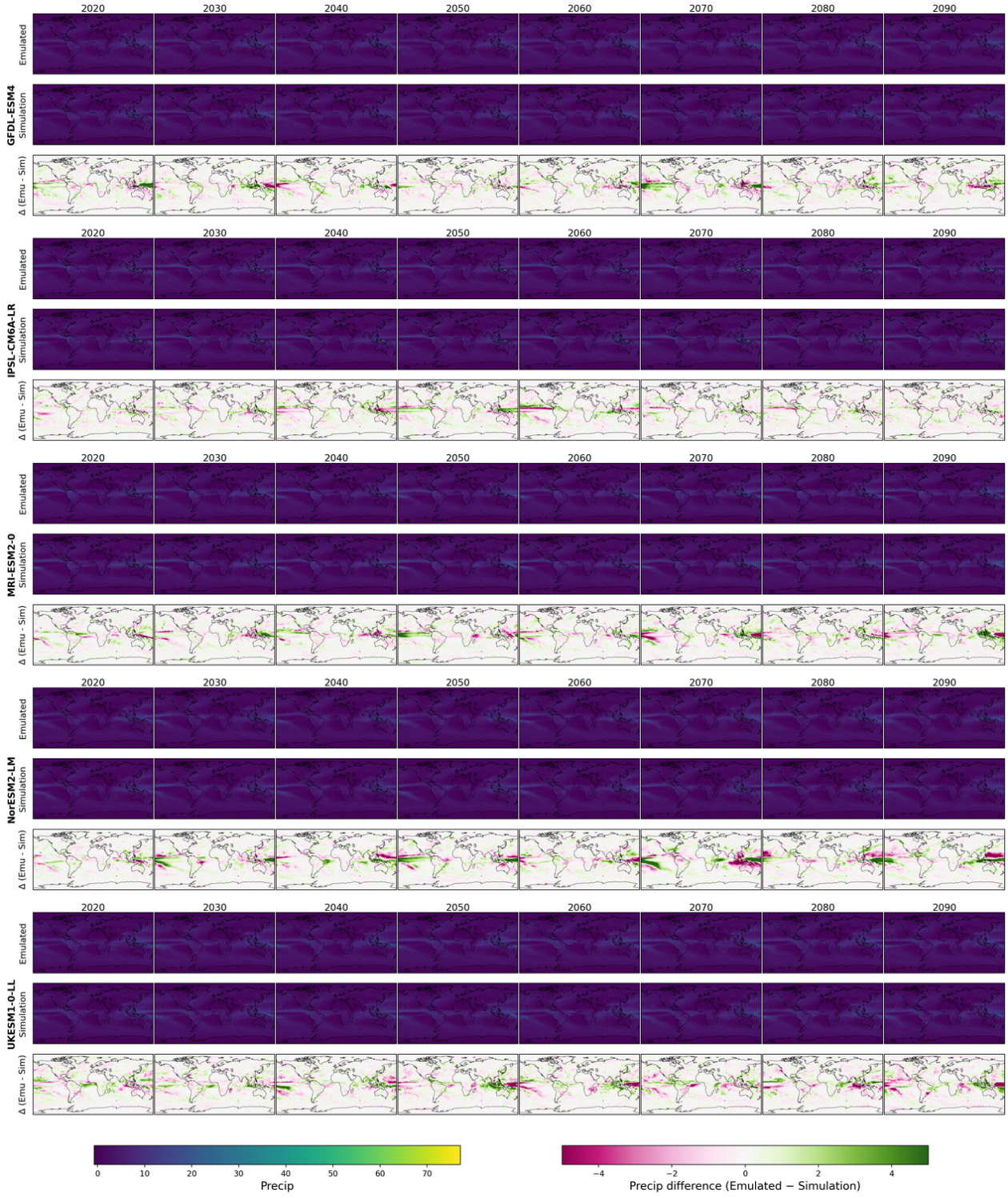

Figure S23. **Per-model ClimateSuite samples of yearly emulated, simulated, and difference precipitation maps (SSP2-4.5).**

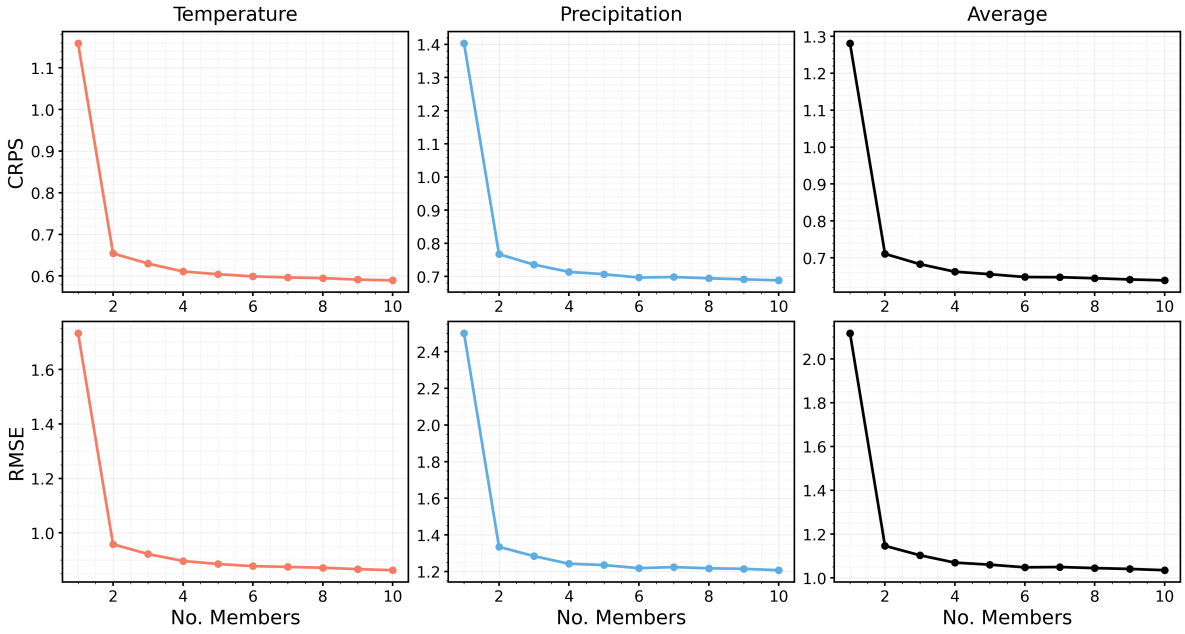

Figure S24. **Effect of number of members on SPF performance on ClimateBench SSP2-4.5.**