ClimForGe: A Diffusion-based Forcing–Response Climate Emulator on Daily Timescales

Jack Kai Lim
Halıcıoğlu Data Science Institute
UC San Diego
jklim@ucsd.edu

Salva Rühling Cachay
UC San Diego
sruhlingcachay@ucsd.edu

Duncan Watson-Parris
UC San Diego
dwatson@ucsd.edu

Abstract

Climate models are indispensable tools for projecting and understanding climate change. Unfortunately, their computational demands severely limit the exploration of diverse climate scenarios and the characterization of extreme events, hindering informed policy decisions. While computationally efficient climate model emulators offer a potential solution, they typically only provide monthly or annual statistics. This paper introduces ClimForGe, a diffusion-based stochastic climate emulator trained on CESM2 capable of efficiently sampling global, daily-scale weather changes under realistic climate forcings. We demonstrate that our emulator accurately reproduces both daily snapshots and long-term statistical properties of temperature and precipitation, offering a powerful tool for rapid exploration and characterization of extreme events in a changing climate.

1 Introduction

Climate models are crucial tools in modern climate analysis and policymaking; however, their high computational demands make them slow and expensive to run, limiting our ability to explore diverse climate scenarios and extreme weather events. To address these limitations, machine learning-based emulators have emerged as an efficient alternative, leveraging existing data from climate and Earth system models (ESMs). These emulators learn to capture the statistical characteristics of our climate, enabling the generation of thousands of data samples in a fraction of the time required by traditional models.

Some key approaches towards climate emulation include forcing-response emulators [9, 15, 17, 20] and autoregressive atmospheric models [12, 18, 21, 22]. Other approaches include the use of generative models such as GANs and diffusion models to perform temporal [2, 13] and spatial [1, 14, 19] super-resolution of ESM outputs. Recently, cBottle [3] demonstrated the ability of diffusion models to generate high-resolution climate samples conditioned on sea surface temperature (SST).

Unfortunately, forcing–response emulators are typically limited to coarse temporal resolutions, making them insufficient for accurately assessing future climate states at high resolution. Autoregressive models, on the other hand, have struggled to produce climate responses under realistic forcing scenarios, particularly further into the future, and they additionally require lengthy rollouts to generate long-term simulations. To address these issues, we introduce ClimForGe (Climate Force-response Generator), a novel approach utilizing the successful EDM diffusion framework [10] to generate global maps of precipitation and surface temperature at a daily resolution conditioned on CMIP6 forcings. The forcings consist of various greenhouse gases and the incident shortwave radiation at the top of the atmosphere (RSDT). ClimForGe is capable of **efficient daily-scale, global forcing–response climate emulation** generating realistic responses to the unseen SSP245 scenarios, and accurately reproducing the long-term statistics of the CESM2 climate model.

2 Related Works

Forcing-Response climate emulators have emerged as computationally efficient tools capable of capturing the complex relationships between external forcings and climate responses typically produced by Earth System Models (ESMs). ClimateBench [20] serves as a benchmark framework to systematically evaluate these emulators, highlighting several critical challenges. A prominent issue is accurately emulating non-linear precipitation responses, exacerbated by high-dimensional climate data and limited training samples. Such constraints often result in coarse spatial resolutions, such as $1.875° \times 2.5°$ reported in previous studies [9, 15, 20], and introduce substantial uncertainty that is difficult to quantify. Furthermore, these existing approaches typically simplify the emulation task to yearly or monthly timescales, an approach also prevalent in downscaling-emulator models like ClimaX [16].

Autoregressive emulators represent another popular approach to climate emulation [4, 12, 18, 21, 22]. These models condition on an atmospheric snapshot (an initial condition) with specified forcings and boundary conditions to subsequently predict the atmosphere at the next time step. However, a notable drawback of the approach is its autoregressive nature requiring lengthy sequential rollouts for generating future climate states. For example, the common resolution of 6 hours would require 14600 autoregressive iterations to simulate a climate state 10 years in the future. This causes them to be computationally expensive and at risk of accumulating errors. Another limitation of autoregressive models is their deterministic nature, which prevents them from capturing the uncertainty inherent in ensemble-based models like ESMs which is a challenge shared by many deterministic emulators.

Generative climate modeling offers a powerful means of generating accurate atmospheric states that faithfully capture both the chaotic, instantaneous spatial structure and the accurate reproduction of long-term climate statistics. Among these, DiffESM [2] adopts a diffusion-based model similar to ours, but diverges in its design as a temporal super-resolution emulator, generating realistic climate samples conditioned on monthly averaged inputs. While an effective method for temporal downscaling, its reliance on prior monthly outputs from conventional climate models inherently limits its ability to explore different scenarios. Another notable work is cBottle [3], an atmospheric emulator similar to [21, 22] but utilizing a diffusion framework. cBottle is comprised of two modules: a coarse-generation module, which produces climate states conditioned on sea surface temperature (SST) and other boundary conditions, and a super-resolution module, which applies a patch-based multi-diffusion to achieve extremely high spatial resolutions and data compression. In contrast, ClimForGe is a forcing-response emulator, being conditioned directly on external drivers such as greenhouse gases (GHG) and aerosols and trained to generate climate states directly.

3 ClimForGe: Climate Forcing-response Generator

3.1 Background

Diffusion models [8, 10] are used to generate high-quality data samples through iterative denoising starting from a simple prior distribution. In the forward process, the data is corrupted until reaching the prior, typically Gaussian, distribution. To reverse this, a neural network is trained to progressively reconstruct the original sample [8]. ClimForGe's architecture uses the EDM framework [10], using a U-Net-based denoiser network [6]. Given the inherent stochastic nature of our climate, diffusion models represent a natural and promising choice to model and sample from the implicit high-dimensional probability distribution.

3.2 Dataset and Experiment Setup

ClimForGe is conditioned on forcing data from Input4MIPs within the Coupled Model Intercomparison Project Phase 6 (CMIP6) [7] processed similarly in ClimateBench [20] and incident shortwave radiation at the top of the atmosphere (RSDT) from Community Earth System Model (CESM2) [5]. We included RSDT, at the daily temporal resolution, as it provides the crucial information about the day of year that the model is generating, effectively encoding the seasonality signal that would otherwise be missing from our annual forcings.

As both the forcing data and RSDT are originally at a coarser resolution (more details in Table 2), we reconcile the spatial discrepancy between RSDT and our other forcings by simple spatial downscaling,

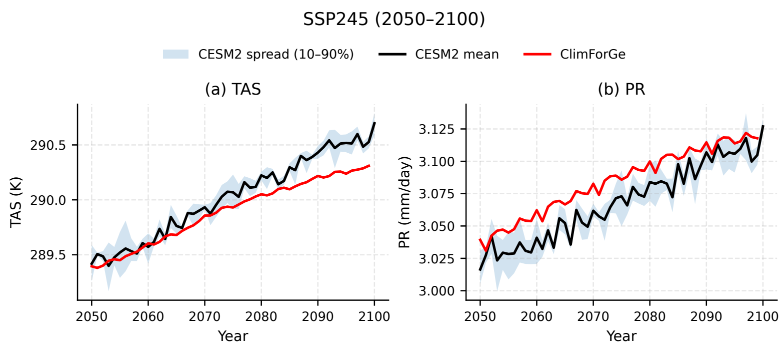


Figure 1: Comparison of ClimForGe's global annual means against CESM2 ensembles global annual means over the period 2050-2100 for our test scenario ssp245

while the temporal mismatch was addressed via a linear interpolation for details, see Appendix B.2. Making our inputs a $5 \times 96 \times 144$ tensor.

The target data is derived from the CESM2, on a spatial grid of 192×288 with a daily resolution for both temperature (tas) and precipitation (pr). Given the probabilistic nature of our model, it outputs results in $T \times E \times 192 \times 288$, where T are our output variables and E is a sampling parameter for the ensemble size to generate. To predict double the spatial resolution of the inputs, we insert an extra upsampling Unet block to the base architecture from [10]. We train on SSP126, SSP585 (2015–2100) and SSP370 (2015–2060), reserving SSP370 (2060–2100) for validation; all testing is conducted on SSP245.

3.3 Training

We trained two ClimForGe models, one for temperature and one for precipitation, both trained on the area-weighted mean squared error loss and following the noise schedule and preconditioning hyperparameters from [10]. The UNet backbone was initialized with 64 model channels and trained for 100 epochs with an effective batch size of 512. We used the Adam optimizer [11] with learning rate 4×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0, and a cosine annealing learning rate schedule. Exponential moving average (EMA) weights with decay 0.9999 were maintained during training. Models were trained stochastically over three CESM2 ensemble members (sampling one member per epoch) on 8 Tesla V100-SXM2-32GB GPUs, requiring approximately 4 days. At inference, ClimForGe can generate 50 years of full global daily temperature and precipitation fields in approximately 1h 40m on a single Nvidia A100 GPU.

3.4 Precipitation Bias Correction

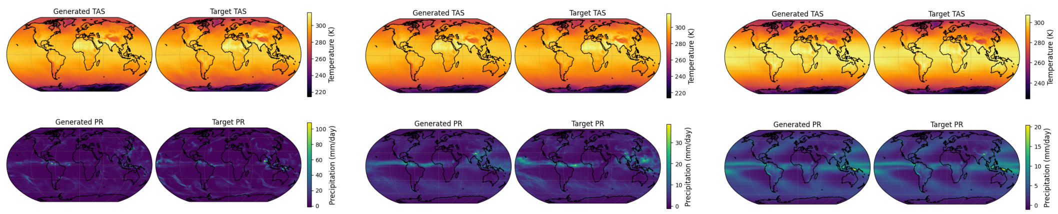
During evaluation, we observed a positive bias in regions with near-zero precipitation (pr), traced to CESM2's non-zero values in "dry" areas. This encouraged our model to generate larger and non-zero pr values. To address this, we applied post-processing before metric computation, setting all pr values below 5×10^{-10} mm/day to zero. This threshold was chosen after inspecting CESM2 precipitation distributions, with density histograms that revealed an abnormally large concentration of values below 5×10^{-10} mm/day, which we classified as "dry" regions.

4 Results

To assess how well ClimForGe reproduces the ESMs behavior, we first compute the global annual means from generated daily climate states over 2050-2100 for temperature and precipitation and we compare this with the CESM2 ensemble mean for the respective variables. Figure 1 shows that the ClimForGe generated ensemble means are close to the mean of the three CESM2 ensembles and able to closely follow the positive trend of SSP245. We do note however, that there is a slight negative bias for temperature and a slight positive bias for precipitation despite our post-processing efforts.

Table 1: RMSE and CRPS of global and time means for the years 2080-2100 of SSP245 for ClimForGe and Climatebench models. Note that because Climatebench emulates NorESM, these metrics are not directly comparable but the metrics provide a useful reference point for relative model skill. CRPS is computed based on 5 ensemble members for ClimForGe and not computed for the CNN as its a deterministic model.

	TAS		PR	
Model	RMSE (K)	CRPS (K)	RMSE (mm/day)	CRPS (mm/day)
GP	0.225	0.477	0.153	1.075
CNN	0.222	–	0.139	–
ClimForGe	0.377	0.282	0.239	0.128



(a) Daily maps (b) Monthly maps (c) 2050-2100 map

Figure 2: Global maps comparing ClimForGe outputs with CESM2 for SSP245: (a) June 1, 2099; (b) June 2099 monthly mean; (c) mean over 2050–2100.

Next we produce global maps across multiple time scales, specifically, we evaluate daily projections as well as aggregated monthly, and decadal statistics to evaluate both ClimForGe's short-term dynamics and long-term statistics against CESM2 targets. Evaluating figure 2 it highlights ClimForGe's ability to capture key spatial patterns with correct reproduction of warm areas and the line of precipitation at the tropics.

Finally, we report the area-weighted Root Mean Square Error (RMSE) and area-weighted Continuous Ranked Probability Score (CRPS) [23] computed from the 20-year time mean (see A.1), following the protocol defined in ClimateBench [20]. We benchmark ClimForGe's scores against ClimateBench's Gaussian Process (GP) model and the Convolutional Neural Network (CNN) model, as they represent the two best-performing baselines in ClimateBench [20], with the GP additionally offering ensemble generation capabilities. We see that ClimForGe delivers highly competitive RMSE scores compared to ClimateBench, despite operating at $2\times$ spatial and $365\times$ temporal resolutions, and outperforms the GP baseline in terms of CRPS. These results position ClimForGe as an effective daily forcing–response climate emulator, with the added advantage of leveraging a diffusion framework to naturally capture uncertainty and generate diverse ensembles.

5 Conclusion and Future Work

In this paper, we have demonstrated the capabilities of a diffusion-based architecture for forcing–response climate emulation. Our approach enables rapid exploration of diverse socioeconomic scenarios while accurately capturing the uncertainties and internal variability of Earth System Models (ESMs), while reproducing accurate long-term statistics in the relevant metrics for climate modeling for both temperature and precipitation. Despite these positives, there are some limitations, such as the disjoint generation of temperature and precipitation, which we aim to address in the future. Another limitation is the lack of temporal coherence between samples, as the model is i.i.d., restricting its ability to capture multi-day events.

Another promising direction for future work is the implementation of diffusion guidance. Since diffusion models aim to learn the full distribution, they should in principle capture rare extreme examples such as prolonged dry spells. These events, however, are difficult to model due to their rarity. By incorporating classifier guidance, we hope to more effectively sample from the tails of the distribution, enabling the direct generation of realistic extreme events that may occur.

References

[1] Henry Addison, Elizabeth Kendon, Suman Ravuri, Laurence Aitchison, and Peter Watson. Machine learning emulation of a local-scale uk climate model. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022. URL https://www.climatechange.ai/papers/neurips2022/21.

[2] Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. DiffESM: Conditional emulation of earth system models with diffusion models. *ICLR Workshop on Tackling Climate Change with Machine Learning*, 2023. doi:10.48550/arxiv.2304.11699.

[3] 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, 2025. URL https://arxiv.org/abs/2505.06474.

[4] William E. Chapman, John S. Schreck, Yingkai Sha, David John Gagne II, Dhamma Kimpara, Laure Zanna, Kirsten J. Mayer, and Judith Berner. CAMulator: fast emulation of the community atmosphere model, 2025. URL https://arxiv.org/abs/2504.06007.

[5] G. Danabasoglu, J.-F. Lamarque, J. Bacmeister, D. A. Bailey, A. K. DuVivier, J. Edwards, L. K. Emmons, J. Fasullo, R. Garcia, A. Gettelman, C. Hannay, M. M. Holland, W. G. Large, P. H. Lauritzen, D. M. Lawrence, J. T. M. Lenaerts, K. Lindsay, W. H. Lipscomb, M. J. Mills, R. Neale, K. W. Oleson, B. Otto-Bliesner, A. S. Phillips, W. Sacks, S. Tilmes, L. van Kampenhout, M. Vertenstein, A. Bertini, J. Dennis, C. Deser, C. Fischer, B. Fox-Kemper, J. E. Kay, D. Kinnison, P. J. Kushner, V. E. Larson, M. C. Long, S. Mickelson, J. K. Moore, E. Nienhouse, L. Polvani, P. J. Rasch, and W. G. Strand. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. doi:https://doi.org/10.1029/2019MS001916. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916. e2019MS001916 2019MS001916.

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021. doi:10.48550/arxiv.2105.05233.

[7] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. doi:10.5194/gmd-9-1937-2016. URL https://gmd.copernicus.org/articles/9/1937/2016/.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. doi:10.48550/arxiv.2006.11239.

[9] Julia Kaltenborn, Charlotte E. E. Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. ClimateSet: A large-scale climate model dataset for machine learning. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023. doi:10.48550/arXiv.2311.03721.

[10] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 2022. doi:10.48550/arxiv.2206.00364.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[12] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, July 2024. ISSN 1476-4687. doi:10.1038/s41586-024-07744-y.

[13] Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2021. doi:10.1109/TGRS.2020.3032790.

[14] Ignacio Lopez-Gomez, Zhong Yi Wan, Leonardo Zepeda-Núñez, Tapio Schneider, John Anderson, and Fei Sha. Dynamical-generative downscaling of climate model ensembles. *Proceedings of the National Academy of Sciences*, 122(17):e2420288122, 2025. doi:10.1073/pnas.2420288122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2420288122.

[15] 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*, 13(2):851–877, 2022. doi:10.5194/esd-13-851-2022. URL https://esd.copernicus.org/articles/13/851/2022/.

[16] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. *International Conference on Machine Learning*, 2023. doi:10.48550/arxiv.2301.10343.

[17] 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(3):353–369, February 2016. ISSN 0165-0009, 1573-1480. doi:10.1007/s10584-015-1509-9. URL http://link.springer.com/10.1007/s10584-015-1509-9.

[18] Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S. Bretherton, and Rose Yu. Probabilistic emulation of a global climate model with Spherical DYffusion. In *Advances in Neural Information Processing Systems*, 2024.

[19] Karen Stengel, Andrew Glaws, Dylan Hettinger, and Ryan N. King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117 (29):16805–16815, 2020. doi:10.1073/pnas.1918964117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1918964117.

[20] D. Watson-Parris, Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10), October 2022. doi:10.1029/2021ms002954.

[21] Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, James Duncan, Brian Henn, Matthew Peters, Noah D Brenowitz, Karthik Kashinath, Mike Pritchard, Boris Bonev, and Christopher Bretherton. ACE: A fast, skillful learned global atmospheric model for climate prediction. *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. doi:10.48550/arxiv.2310.02074.

[22] Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1), May 2025. ISSN 2397-3722. doi:10.1038/s41612-025-01090-0.

[23] Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–232, 2018.

Table 2: Table of Inputs, Outputs and Forcings

Input Variables (Forcings)			
Symbol	Description	Spatial Resolution	Temporal Resolution
CO_2	Carbon Dioxide	96×144	Yearly
SiO_2	Silicon Dioxide	96×144	Yearly
CH_4	Methane	96×144	Yearly
BC	Black Carbon	96×144	Yearly
Additional Input Variables			
Symbol	Description	Spatial Resolution	Temporal Resolution
RSDT	The incident shortwave radiation at the top of the atmosphere	196×288	Monthly
Output Variables			
Symbol	Description	Spatial Resolution	Temporal Resolution
tas	Surface Air Temperature	196×288	Daily
pr	Total Precipitation	196×288	Daily

Appendix

A Metrics

A.1 Time Mean

We define the *time mean* of a climate variable as the average taken over the temporal dimension, while preserving the spatial dimensions:

$$\overline{x}(i,j) = \frac{1}{T}\sum_{t=1}^{T} x(t,i,j), \tag{1}$$

where T is the total number of time steps.

Given a metric function $f(\cdot,\cdot)$ that compares two fields (e.g., RMSE, CRPS), the corresponding time mean metric is defined as

$$M = f(\overline{x}, \overline{y})\,, \tag{2}$$

where $\overline{x}$ and $\overline{y}$ are the time mean fields of the prediction and target, respectively.

B Dataset

B.1 Inputs, outputs and forcing variables

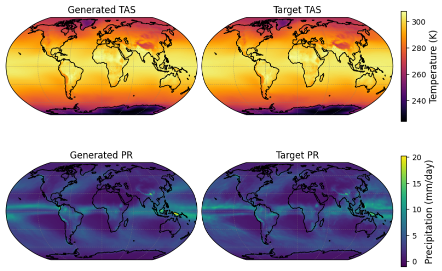
Comprehensive details for the input, output, and forcing variables are described in Table 2. Note that all models were trained using raw output variables, without applying normalization via PiControl.

Input (forcings) these variables are provided by the input4MIPS project `https://aims2.llnl.gov/search/input4mips/`. **RSDT** and **Daily Output variables** can be downloaded `https://intake-esm.readthedocs.io/en/stable/tutorials/loading-cmip6-data.html`.

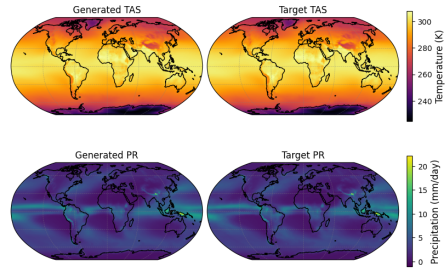
B.2 Time Interpolation

$$\begin{aligned} E_y(t) &= \frac{N-d}{N}E_{prev} + \frac{d}{N}E_{next} \quad \text{annual interpolation} \\ E_m(t) &= \frac{M-d}{M}E_{prev} + \frac{d}{M}E_{next} \quad \text{monthly interpolation} \end{aligned} \tag{3}$$

where $N = 365, M = 30$, representing the number of days in a year and an average month, respectively. d denotes the number of days elapsed from the middle of the temporal scale; (July 2^{nd}



(a) Yearly maps



(b) Decadal maps

Figure 3: Global maps comparing ClimForGe outputs with CESM2 for SSP245: (a) 2099 yearly mean (b) mean over 2090-2100.

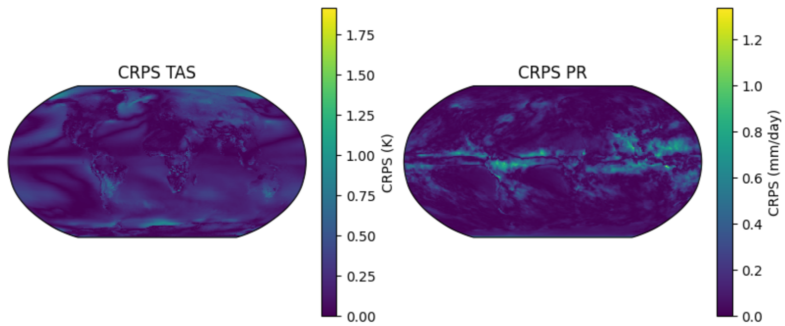
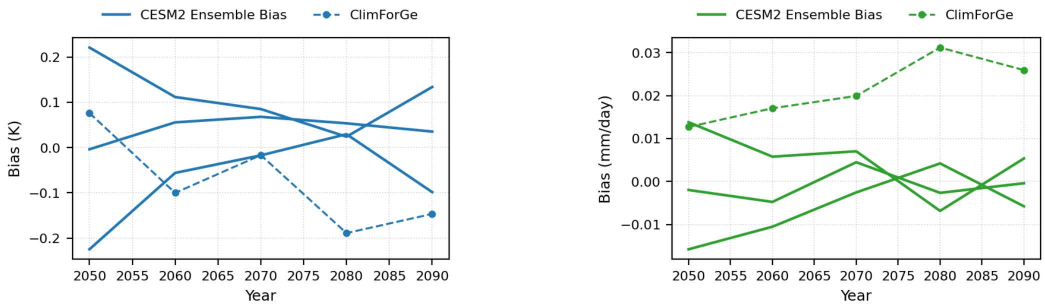


Figure 4: Grid wise CRPS Maps of Generated ClimForGe Ensembles vs CESM2 Ensembles for 2050-2100 time means.

This was done by take the CRPS at each grid cell, allowing for evaluation of our models abilities on a local scale.

for annual; $15th$ day for monthly) and E_{prev} and E_{next} correspond to variables values immediately following or preceding t.

C Additional results and figures



(a) Temperature (b) Precipitation

Figure 5: Leveraging model's ensemble generation, we quantify the emulators internal variability with it's inter-realization bias against CESM2. It can be seen that our emulator is capable of capturing a similar magnitude and structure of the internal variability present in CESM2, noting that TAS presents a high skill than for PR, but overall reproducing the patterns well

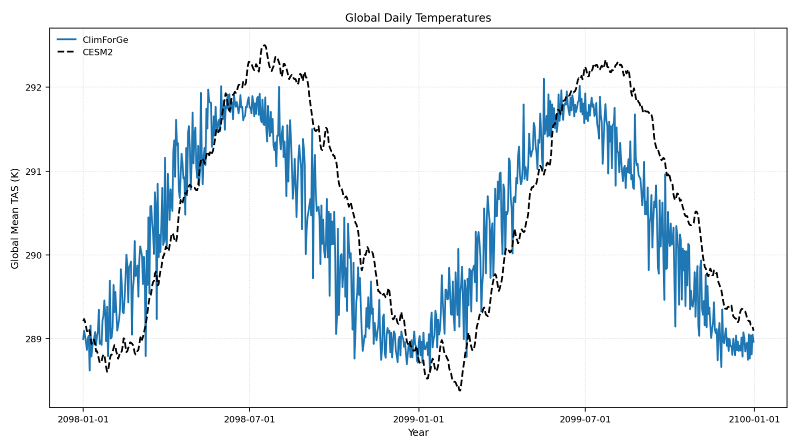


Figure 6: Global Means on a Daily Scale, Generated from ClimForGe vs CESM2 targets for temperature.

We note here that there seems to be a lag in terms of the generated values, falling slightly behind the patterns of CESM2 outputs by about a month

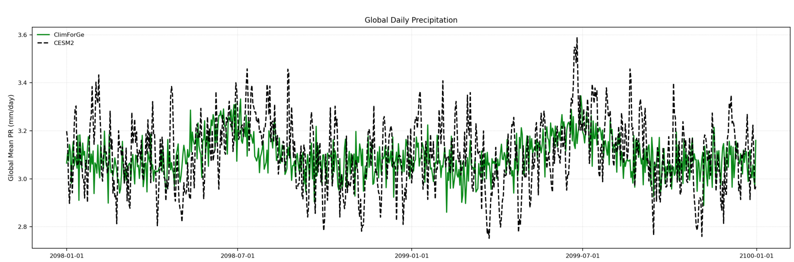


Figure 7: Global Means on a Daily Scale, Generated from ClimForGe vs CESM2 targets for precipitation.

We note here that CESM2 precipitation values oscillate more than our generated from ClimForGe