CLIMGEN: LEARNING THE FORCING-RESPONSE RELATIONSHIP IN CLIMATE SYSTEM

Tse-Chun Chen, Kooloth Parvathi, Jian Lu, Jason Z. Hou
Pacific Northwest National Laboratory
Richland, WA 99354, USA
{Tse-Chun.Chen, Kooloth.Parvathi, Jian.Lu, Zhangshuan.Hou}@pnnl.gov

ABSTRACT

Solar Radiation Management (SRM) is emerging as a potential geoengineering strategy to address the anthropogenic impact on climate, but its effective implementation requires an iterative and large ensemble of highly accurate and efficient climate projections. Traditional climate projections rely on executing computationally demanding and time-consuming numerical climate models. Recent advances in machine learning (ML) aim to enhance these approaches by emulating traditional methods. In this work, we propose a novel framework for directly learning the relationship between solar radiation flux at the top of the atmosphere and the corresponding surface temperature response. To evaluate the feasibility of this direct ML-based projection, we developed a dataset using an intermediate complexity model, incorporating a comprehensive suite of different forcing patterns and evaluation metrics to rigorously assess the ML model's performance. We introduce a Conditional Denoising Diffusion Probabilistic Model (cDDPM) for this task, which demonstrates encouraging skill in representing climate statistics under previously unseen forcing patterns. This approach provides a promising pathway for direct climate projections by accurately learning the forcing-response relationship, with a wide range of applications in impact mitigation, emissions policy design, and SRM strategies.

1 INTRODUCTION

Reliable prediction of climate system response under external forcing and the uncertainty thereof has emerged as a central focus of climate research, especially in the context of Solar Radiation Management (SRM) as a potential geoengineering strategy to mitigate anthropogenic impact. Initial efforts were limited to assessing the equilibrium climate sensitivity (ECS) that characterizes the long-term mean global temperature response in response to anthropogenic forcing from a doubling of the atmospheric CO_2 concentration (Roe (2009); Knutti et al. (2017)). More recent efforts have improved our understanding of the regional heterogeneity of the climate response from a linear perspective (Dong et al. (2019); Liu et al. (2022)). However, understanding and accurately predicting the full nonlinear climate response has remained challenging. This is further compounded by the chaotic nature of the Earth's climate system, its hierarchical structure, and the high dimensionality of the state space (Ghil & Lucarini (2020)). While running large ensemble Earth system models is the most reliable technique for climate prediction, it is compute-intensive. This puts a strong constraint on the number of climate scenarios and ensemble members that can be simulated and the number of parameters that can be perturbed for uncertainty quantification.

We have compiled data from a large suite of Green's function solar perturbation experiments that systematically probe the quantitative forcing-response relationship in an intermediate complexity climate model. Here, we use a denoising diffusion model to learn the emergent dynamic response function for global surface temperature conditioned on the applied solar forcing pattern. The temperature responses generated by the diffusion model also capture the inter-annual variability in the responses. Our approach aims to provide a cheap surrogate that can rapidly generate large ensembles of climate projections under various forcing scenarios while accurately capturing the internal variability in the climate response. Additionally, our trained model could be used for transfer learn-

ing by fine-tuning using minimal data to emulate the climate-forcing response in fully coupled Earth system models.

1.1 Related Work

1.1.1 Learning linear forcing-response relationship in climate

Our study falls into the broad category of climate emulators, specifically those that predict climate responses to external forcing. Many of the conventional methods used for climate emulation are based on simple pattern-scaling (Tebaldi & Arblaster (2014)), 1D impulse response function (MacMartin & Kravitz (2016)), and linear regression (Liu et al. (2022)). More importantly, these methods rely on the strong assumption that the climate response is sufficiently linear and time-invariant.

1.1.2 ML for atmospheric modeling

Recently, deep learning-based autoregressive models have made significant progress in achieving efficient and accurate medium-range weather predictions (e.g., Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Kochkov et al., 2024). However, transferring this skill to long-term climate projections remains challenging due to error accumulation that emerges in forecast roll-outs beyond a two-week timescale. Efforts have been made to extend stable forecast horizons to up to 10 years under current climate conditions (Weyn et al., 2021; Watt-Meyer et al., 2023; Guan et al., 2024; Cachay et al., 2024). For instance, NeuralGCM (Kochkov et al., 2024) has demonstrated promising skill in capturing global warming trends over several decades by prescribing historical sea surface temperature (SST) patterns. However, under strong SST forcing, this model exhibits climate drift. Consequently, the generalizability of these autoregressive climate emulators to different climate-forcing backgrounds has yet to be demonstrated.

An alternative to autoregressive climate emulation focuses on directly learning the forcing-response relationship. Several efforts have employed machine learning techniques—such as random forests, Gaussian processes, and neural networks—to predict the full nonlinear climate response to potential future anthropogenic emission pathways (Watson-Parris et al., 2022). In this approach, crucial inputs like CO_2 and CH_4 are represented as globally averaged scalars, which overlooks the spatial forcing patterns associated with anthropogenic greenhouse gas (GHG) emissions. This limitation hinders the ability of these models to capture the full complexity of climate responses. Recently, ClimateSet (Kaltenborn et al., 2023) was developed as a more comprehensive benchmark dataset. It contains data from 36 climate models corresponding to five different emission scenarios, including the spatial patterns associated with GHGs and aerosols. Although these datasets are highly suitable for climate projection studies under anthropogenic forcing, they have limited relevance for studies exploring the relationship between shortwave forcing via solar radiation management and the associated climate responses.

1.2 Contributions

This study presents several novel contributions to climate modeling and projections. First, it utilizes a comprehensive range of forcing configurations generated from an intermediate-complexity climate model, as opposed to the limited scenarios typically derived from more complex models (Watson-Parris et al. (2022)). Second, it represents the first attempt to directly project the equilibrium state of the climate system from a forcing field using a generative AI technique, enabling a distributional learning of climate responses. This would allow for not only climate projection of mean but also extreme states. Our approach also learns the spatial relationship between the forcing and response patterns. Lastly, the model was validated with an additional test case using a realistic climate model with CO_2 forcing, where it successfully produced a consistent global mean temperature distribution, demonstrating its potential applicability to more complex and realistic climate scenarios.

2 Results

Here we evaluate the performance of the trained cDDPM on two independent test cases: uniform $2Wm^{-2}$ and -2 Wm^{-2} forcing perturbations.

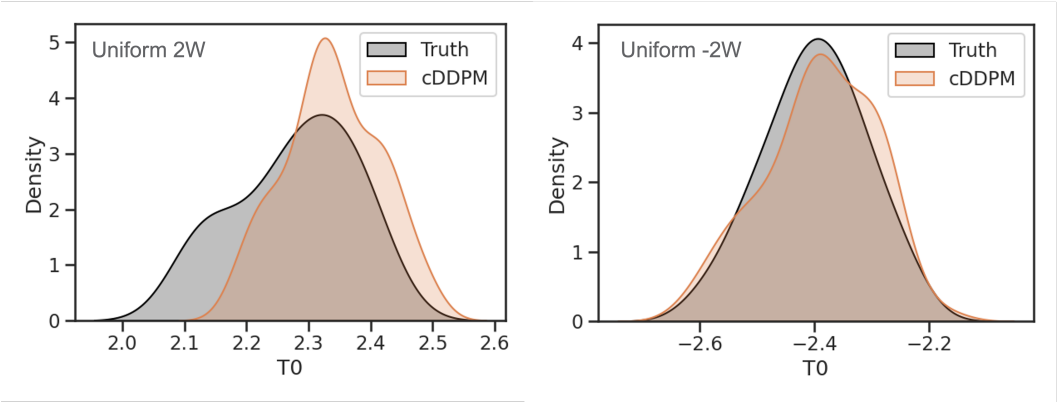

Figure 1: Comparison of the cDDPM-generated with the true (from PlaSim) distribution of global mean temperature changes $[K]$ under uniform $\pm 2 W m^{-2}$ forcing perturbations.

To better assess the performance of cDDPM and the quality of the generated samples, we employed 3 temperature metrics introduced in Kravitz et al. (2017): the global mean surface temperature (T_0), the interhemispheric surface temperature gradient (T_1), and the equator-to-pole surface temperature gradient (T_2). These metrics are defined as follows: $T_0 = \frac{1}{A}\int_{-\pi/2}^{\pi/2} T(\psi) dA$, $T_1 = \frac{1}{A}\int_{-\pi/2}^{\pi/2} T(\psi)\sin(\psi) dA$, and $T_2 = \frac{1}{A}\int_{-\pi/2}^{\pi/2} T(\psi)\frac{1}{2}(3\sin^2(\psi) - 1) dA$.

where $dA = 2\pi R^2 \cos(\psi) d\psi$ is the area of a latitudinal band, and $A = 2\pi R^2 \int_{\psi=-\pi/2}^{\psi=\pi/2} dA$ is the total surface area of the Earth.

We compare the distributions of the cDDPM-generated and the true changes in these temperature metrics relative to the reference climate (without forcing perturbation) under the 2 independent forcing scenarios. Figure 1 compares cDDPM-generated and true distributions of global mean temperature changes under the two forcing scenarios. The similarity in both the peak and width of the generated and the true distributions indicates that the cDDPM can produce samples with representative probability distributions for global mean temperature changes in response to forcing perturbations.

Notably, the responses to positive and negative forcing perturbations of the same magnitude are 2.27K and -2.40K, respectively, and are not symmetric around zero. The differing peak locations and distribution shapes further emphasize the nonlinear nature of the climate system, demonstrating that the cDDPM demonstrates reasonable skill in capturing this inherent nonlinearity.

We also compare the joint and marginal distributions of T_1 and T_2 in Fig 2. In the uniform +2 Wm^{-2} scenario, the spread of both temperature metrics aligns well between the cDDPM-generated and true distributions, although the cDDPM underestimates the peak of T_1. In the uniform -2 Wm^{-2} scenario, the cDDPM generates a representative width, but the peak location of both T_1 and T_2 shows a more noticeable underestimation. Despite these differences, the cDDPM is still capable of generating representative samples under the given forcing perturbations.

To establish a baseline, we also trained a Gaussian Process regressor to predict the distribution of the climate responses given the forcing. Specifically, we used Sparse Gaussian Process Regression (SGPR) to tackle the high-dimensionality and the large training sample size (Matthews et al. (2017)). The SGPR predicts a null response to the 2W/m^2 uniform forcing test case, which is perhaps not surprising since the Gaussian Process regressor is known to be unsuitable for predicting high dimensional outputs (Vershynin (2018)).

3 Conclusions and Future Work

In this study, we developed a conditional denoising diffusion model with a U-Net backbone to learn the relationship between forcing perturbations and climate system responses. This proof-of-concept study with a minimalist cDDPM demonstrates the potential for training a direct distributional pro-

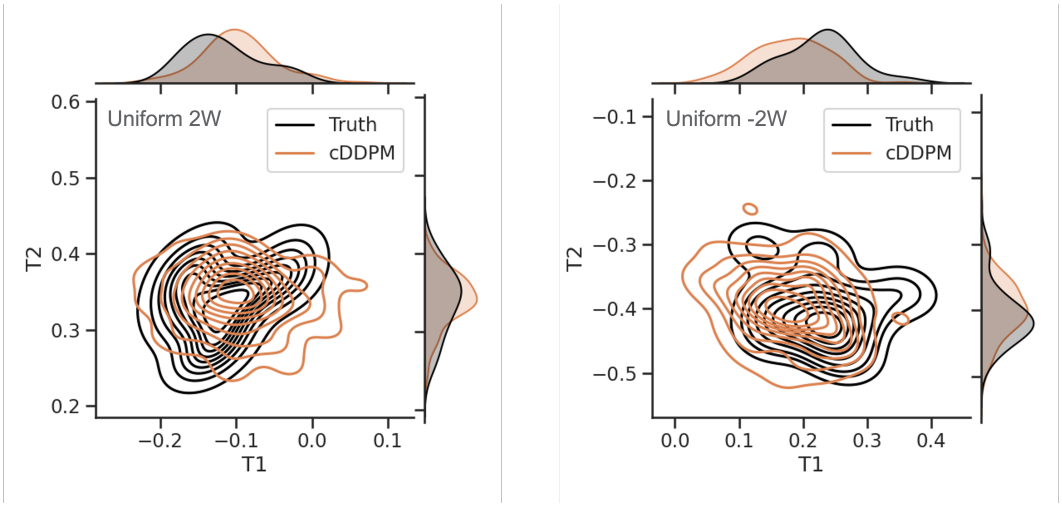

Figure 2: Comparison of the cDDPM-generated and true (from PlaSim) joint and marginal distributions of T_1 and T_2 under uniform $\pm 2 W m^{-2}$ forcing perturbations

jection of global temperature responses based on the context of a forcing perturbation field as input. Notably, this work represents the first attempt to train a machine learning model for climate projection using a comprehesive set of forcing configurations generated by an intermediate-complexity climate model. The generated samples were validated against independent test cases using various temperature metrics, showing that the cDDPM is capable of producing samples that are representative of both the intermediate-complexity model and the even more realistic and computationally intensive CMIP models (see Appendix D).

This endeavor will prove invaluable for a wide range of applications, including impact mitigation efforts, the development of effective emissions policy designs, and the exploration of SRM strategies. By providing a more accurate and efficient means of projecting climate responses to various forcing scenarios, this approach can inform policymakers and scientists in crafting data-driven strategies to reduce greenhouse gas emissions, implement adaptation measures, and assess the potential risks and benefits of geoengineering techniques.

Future work will focus on exploring alternative backbone architectures to enhance performance further. We also aim to extend the model's capabilities to predict precipitation, which poses an even greater challenge due to its larger nonlinearity compared to temperature. Moreover, one potential challenge in directly projecting climate responses from forcing perturbations is the limited training data available in terms of the variety of forcing configurations. However, we suggest that the dataset used in this study could serve as a valuable resource for pre-training foundational models like the cDDPM we developed, which could then be fine-tuned using outputs from more realistic climate models.

We also highlight the importance of considering tipping points in the climate system in future work. Tipping points refer to critical thresholds at which a small perturbation can significantly alter the state or development of the system. It is essential to explicitly evaluate ClimGen's ability to represent tipping points. Incorporating scenarios that lead up to, pass through, and extend beyond known tipping points into the training data will help ensure that these critical thresholds are adequately taken into account.

ACKNOWLEDGMENTS

The research described herein was funded by the Generative AI for Science, Energy, and Security Science & Technology Investment under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. This work was also supported by the Center for AI and Center for Cloud Computing at PNNL.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3. URL https://www.nature.com/articles/s41586-023-06185-3. Publisher: Nature Publishing Group.

Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S. Bretherton, and Rose Yu. Probabilistic Emulation of a Global Climate Model with Spherical DYffusion, June 2024. URL http://arxiv.org/abs/2406.14798. arXiv:2406.14798 [physics, stat] version: 1.

Yue Dong, Cristian Proistosescu, Kyle C Armour, and David S Battisti. Attributing historical and future evolution of radiative feedbacks to regional warming patterns using a green's function approach: The preeminence of the western pacific. *Journal of Climate*, 32(17):5471–5491, 2019.

Klaus F Fraedrich, Heiko Jansen, Edilbert Kirk, Ute Luksch, and Frank Lunkeit. The planet simulator: Towards a user friendly model. *Meteorologische Zeitschrift*, 14(3):299–304, 2005.

Michael Ghil and Valerio Lucarini. The physics of climate variability and climate change. *Reviews of Modern Physics*, 92(3):035002, 2020.

Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik. LUCIE: A Lightweight Uncoupled ClImate Emulator with long-term stability and physical consistency for O(1000)-member ensembles, May 2024. URL http://arxiv.org/abs/2405.16297. arXiv:2405.16297 [physics].

Pedram Hassanzadeh and Zhiming Kuang. The linear response function of an idealized atmosphere. part i: Construction using green's functions and applications. *Journal of the Atmospheric Sciences*, 73(9):3423–3439, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Julia Kaltenborn, Charlotte 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. *Advances in Neural Information Processing Systems*, 36:21757–21792, 2023.

Ryan Keisler. Forecasting Global Weather with Graph Neural Networks, February 2022. URL http://arxiv.org/abs/2202.07575. arXiv:2202.07575 [physics].

Reto Knutti, Maria AA Rugenstein, and Gabriele C Hegerl. Beyond equilibrium climate sensitivity. *Nature Geoscience*, 10(10):727–736, 2017.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Ben Kravitz, Douglas G. MacMartin, Michael J. Mills, Jadwiga H. Richter, Simone Tilmes, Jean-Francois Lamarque, Joseph J. Tribbia, and Francis Vitt. First Simulations of Designing Stratospheric Sulfate Aerosol Geoengineering to Meet Multiple Simultaneous Climate Objectives. *Journal of Geophysical Research: Atmospheres*, 122(23):12,616–12,634, 2017. ISSN 2169-8996. doi: 10.1002/2017JD026874. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/2017JD026874.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, December 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/10.1126/science.adi2336. Publisher: American Association for the Advancement of Science.

Fukai Liu, Jian Lu, and L Ruby Leung. Neutral mode dominates the forced global and regional surface temperature response in the past and future. *Geophysical Research Letters*, 49(15): e2022GL098788, 2022.

Jian Lu, Fukai Liu, L Ruby Leung, and Huan Lei. Neutral modes of surface temperature and the optimal ocean thermal forcing for global cooling. *npj Climate and Atmospheric Science*, 3(1):9, 2020.

Douglas G MacMartin and Ben Kravitz. Dynamic climate emulators for solar geoengineering. *Atmospheric Chemistry and Physics*, 16(24):15789–15799, 2016.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, apr 2017. URL http://jmlr.org/papers/v18/16-537.html.

Gerald A. Meehl, Catherine A. Senior, Veronika Eyring, Gregory Flato, Jean-Francois Lamarque, Ronald J. Stouffer, Karl E. Taylor, and Manuel Schlund. Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models. *Science Advances*, 6(26):eaba1981, 2020. doi: 10.1126/sciadv.aba1981. URL https://www.science.org/doi/abs/10.1126/sciadv.aba1981.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, February 2022. URL http://arxiv.org/abs/2202.11214. arXiv:2202.11214 [physics].

Gerard Roe. Feedbacks, timescales, and seeing red. *Annual Review of Earth and Planetary Sciences*, 37(1):93–115, 2009.

M. Schlund, A. Lauer, P. Gentine, S. C. Sherwood, and V. Eyring. Emergent constraints on equilibrium climate sensitivity in cmip5: do they hold for cmip6? *Earth System Dynamics*, 11(4): 1233–1258, 2020. doi: 10.5194/esd-11-1233-2020. URL https://esd.copernicus.org/articles/11/1233/2020/.

Claudia Tebaldi and Julie M Arblaster. Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, 122:459–471, 2014.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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):e2021MS002954, 2022.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. ACE: A fast, skillful learned global atmospheric model for climate prediction, December 2023. URL http://arxiv.org/abs/2310.02074. arXiv:2310.02074 [physics].

Jonathan A. Weyn, Dale R. Durran, Rich Caruana, and Nathaniel Cresswell-Clay. Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, 13(7):e2021MS002502, 2021. ISSN 1942-2466. doi: 10.1029/2021MS002502. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002502. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002502.

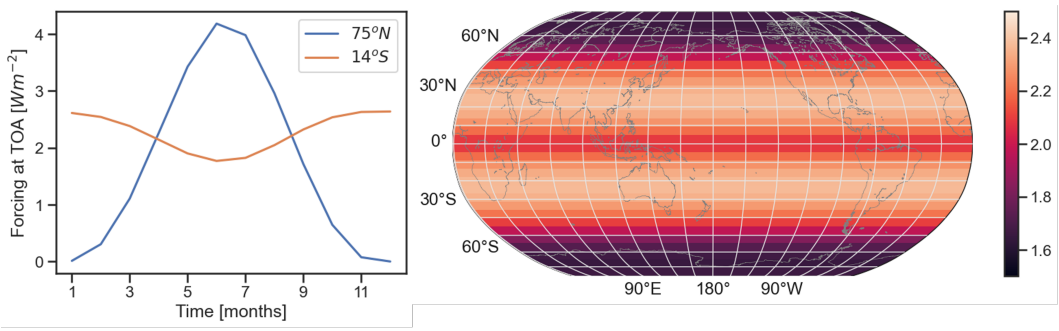

Figure 3: Configuration of the Green's function experiments: The right panel illustrates the 256 local patches distributed around the globe on top of uniform $2Wm^{-2}$ TOA forcing field, while the left panel displays the TOA forcing over two different latitudinal patches over the course of a year.

A Climate Forcing and Response Dataset

The data used in our study is generated using a widely used intermediate complexity climate model called Planet Simulator (PlaSim) (Fraedrich et al. (2005)). The dynamical core of PlaSim is a simplified global circulation model (GCM) of the atmosphere with parameterization schemes for modeling physical processes such as diffusion, radiative processes, moist processes including cloud formation and precipitation, and dry convective adjustment. It also includes linearized representations of important Earth system components such as a slab ocean with sea ice and a land surface with biosphere. The grid resolution for the simulations is chosen to be $5.6^\circ \times 5.6^\circ$.

To probe the patterned forcing-response relationship in the climate system, a large number of Green's function experiments, in the vein of Hassanzadeh & Kuang (2016); Liu et al. (2022), are conducted by perturbing the incoming solar radiation at the top of the atmosphere (TOA) over localized patches (16×16) distributed around the globe as shown in Fig 3. Note that applied forcing is seasonally varying as shown in the schematic such that the annual mean of the forcing is roughly the same across all the patches irrespective of the location. We use six different forcing levels ($\pm 15\ Wm^{-2}$, $\pm 30\ Wm^{-2}$, and $\pm 60\ Wm^{-2}$). Using multiple negative and positive forcing levels allows us to quantify the nonlinearities in the surface temperature response systematically. A preindustrial control simulation is carried out for 150 years and the data from the last 100 years is used as the baseline reference (unforced) state. The control data also provides a robust estimate of the internal variability in the climate. The forced simulations were started from the 100^{th} year of the control simulation and ran for another 60 years. The data from the last 20 years of the forced runs is used for training and testing the cDDPM to ensure sufficient equilibration of the climate. The dataset also includes uniformly forced runs at $\pm 2\ Wm^{-2}$, $\pm 4\ Wm^{-2}$, and $\pm 8\ Wm^{-2}$ (see for example Figure 3).

The diffusion model was trained using data from the Green's function runs and four of the uniformly forced runs with values of $\pm 4\ Wm^{-2}$ and $\pm 8\ Wm^{-2}$. The $\pm 2\ Wm^{-2}$ uniformly forced cases were reserved for independent testing. This decision was made because the $2Wm^{-2}$ is more pertinent to current and future warming scenarios. The final model was selected based on its performance in representing the forced $\pm 2\ Wm^{-2}$ temperature responses, evaluated using various error metrics (see Appendix F).

We would like to emphasize here that due to the nonlinearity in climate system, the prediction task is far more complicated then linear interpolation. The response changes in both magnitude and pattern under different radiative forcing scenarios.

B Conditional Diffusion Model for Climate Projection

We employ a conditional denoising diffusion probabilistic model (cDDPM) to generate an ensemble of climate projections conditioned on the applied forcing pattern.

B.1 DENOISING DIFFUSION PROBABILISTIC MODEL

The denoising diffusion probabilistic model introduced in Ho et al. (2020) belongs to the class of latent variable models; it involves a forward process that gradually adds noise to the data and a reverse process that learns to retrieve the original data through systematic denoising. The forward diffusion is a Markov process with a fixed number of steps n, during which Gaussian noise is iteratively added to the input $\mathbf{x}_0$. At the i^{th} step, $\mathbf{x}_i$ is obtained by noising the preceding iterate $\mathbf{x}_{i-1}$ based on a prescribed variance schedule ($\beta_i \in (0, 1)$) such that,

$$q(\mathbf{x}_i|\mathbf{x}_{i-1}) = \mathcal{N}(\mathbf{x}_i; \sqrt{1-\beta_i}\mathbf{x}_{i-1}, \beta_i\mathbf{I}). \tag{1}$$

Equivalently, $\mathbf{x}_i$ can be directly obtained from $\mathbf{x}_0$ as

$$q(\mathbf{x}_i|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_i; \sqrt{1-\bar{\alpha}_i}\mathbf{x}_0, \bar{\alpha}_i\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_i = \prod_{s=1}^{i} \alpha_s$ and $\alpha_s = 1 - \beta_s$.

The reverse process is also a Markov chain such that $\mathbf{x}_{i-1}$ at the i^{th} denoising step is obtained as

$$p_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_{i-1}; \mu_\theta(\mathbf{x}_i, i), \Sigma_\theta(\mathbf{x}_i, i)). \tag{3}$$

Here, $\mu_\theta(\mathbf{x}_i, k)$ is the learned mean from a trained neural network and $\Sigma_\theta(\mathbf{x}_i, i)$ is the covariance matrix. For simplicity, the reverse process employs a fixed covariance matrix ($\sigma_i^2\mathbf{I}$ with $\sigma_i^2 = \beta_i$) that mirrors the forward diffusion. The neural net is a function approximator that predicts the noise ϵ from x_i, $\epsilon_\theta(\mathbf{x}_i, i)$ where the subscript θ represents the trained network parameters. The mean $\mu_\theta(\mathbf{x}_i, i)$ is then computed as

$$\mu_\theta(x_i, i) = \frac{1}{\sqrt{\alpha_i}}\left(\mathbf{x}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\epsilon_\theta(\mathbf{x}_i, i)\right). \tag{4}$$

The parameters θ are obtained by minimizing the loss $\mathcal{L}_\theta = \mathbb{E}_{i,\mathbf{x}_0,\epsilon}\left(\|\epsilon - \epsilon_\theta(\mathbf{x}_i, i)\|^2\right)$.

Note that in our study, we use a uniform variance schedule with $n = 401$ steps with $(\beta_0, \beta_n) = (10^{-4}, 0.02)$ with a constant step size.

B.2 CLASSIFIER-FREE GUIDANCE

To generate ensembles of the climate response for a given solar forcing pattern, we added classifier-free guidance to the DDPM (Ho & Salimans (2022)). This is achieved by modifying (3) to account for the condition $\mathbf{c}$ and is given by,

$$p_\theta(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{i-1}; \mu_\theta(\mathbf{x}_i, i, \mathbf{c}), \Sigma_\theta(\mathbf{x}_i, i)). \tag{5}$$

Here, the condition $\mathbf{c}$ is an embedding generated by a fully connected neural network for a given forcing pattern $\mathbf{F}$ as shown in Figure 5. Similarly, the loss function to be optimized becomes $\mathcal{L}_\theta = \mathbb{E}_{i,\mathbf{x}_0,\epsilon}\left(\|\epsilon - \epsilon_\theta(\mathbf{x}_i, i, \mathbf{c})\|^2\right)$ and the mean $\mu_\theta(x_i, i)$ is computed as

$$\mu_\theta(x_i, i) = \frac{1}{\sqrt{\alpha_i}}\left(\mathbf{x}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\epsilon_\theta(\mathbf{x}_i, i, \mathbf{c})\right). \tag{6}$$

B.3 U-NET BACKBONE

We use a standard U-Net architecture composed primarily of five ResNet blocks, as the backbone for our cDDPM. The details of the U-Net backbone used in our study, along with the time and context embedding blocks are shown in Figures 4 and 5. We also tested several variants of this architecture by varying the location of the context and time embedding and by introducing a self-attention block at the bottleneck of the U-Net; the results comparing the performance of these variants are described in Appendix F.

B.3.1 CONTEXT EMBEDDING

For conditioning, we utilize a natural reduced-order representation of the forcing pattern that reflects the resolution at which the forcing-response relationship is represented in the Green's function experiments described in Appendix A. The full forcing field with a resolution of 32×64 is reduced

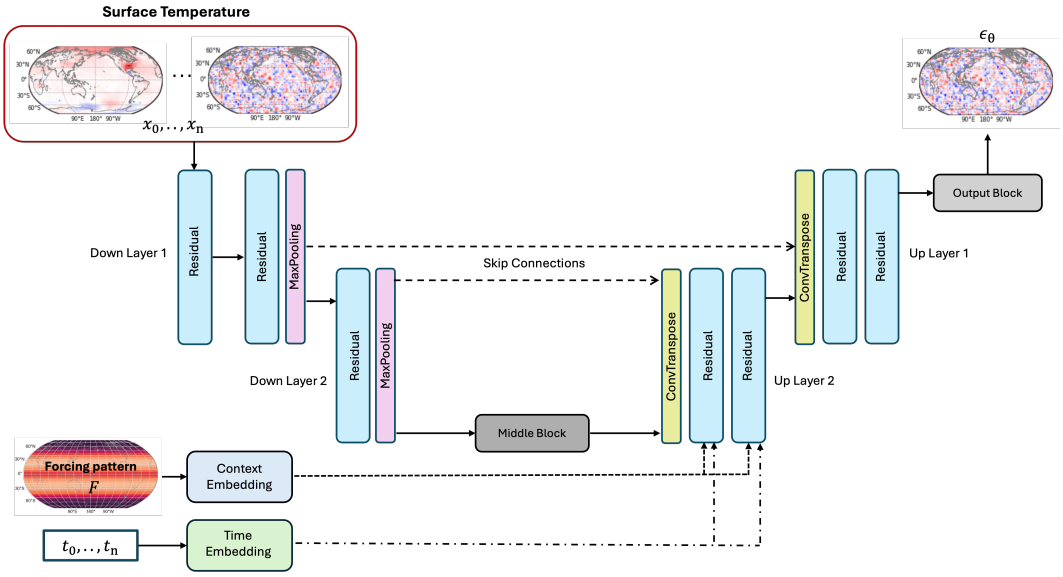

Figure 4: Schematic of the U-Net

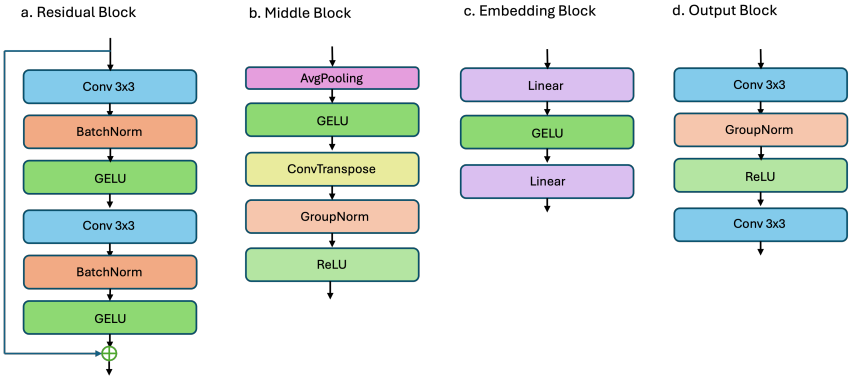

Figure 5: Detailed structure of the component blocks of the U-Net backbone.

to a 16×16 matrix with each entry representing the annual mean of the applied solar perturbation over the Green's function patches shown in the Figure 3. Direct conditioning on the forcing field at full resolution (32×64) was intractable and resulted in poor prediction of the temperature response.

The context embedding is achieved via an embedding block that consists of two linear transforms and a GELU activation as shown in Figure 5c.

B.3.2 Self-attention

We also tested a variant of the basic architecture that includes a simple self-attention block at the bottleneck of the U-Net; specifically at the beginning of the middle block shown in Figure 5.

C ClimGen Generated Samples for Uniform $\pm 2 W m^{-2}$ Forcing Perturbations

Figure 6 presents 12 generated samples for each forcing scenario. It is evident that while the large-scale patterns across the samples are quite similar, each sample retains its own inter-annual variability under the same forcing.

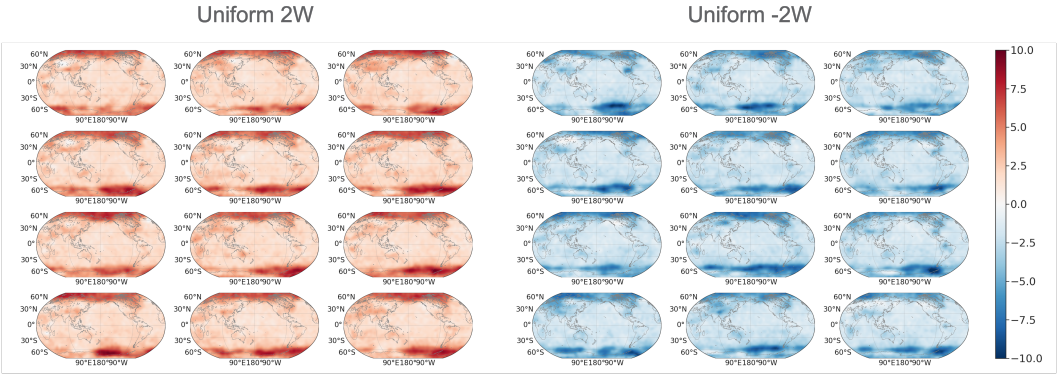

Figure 6: Ensemble of cDDPM-generated surface temperature responses to uniform $\pm 2 W m^{-2}$ forcing perturbations.

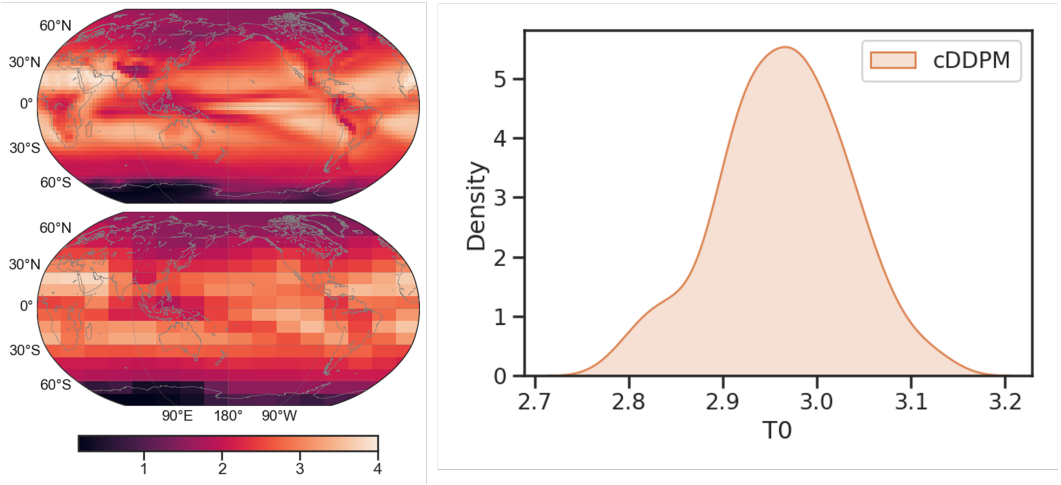

Figure 7: Left panel shows the forcing field from a CESM doubling CO_2 warming scenario. We coarsen the resolution from the original size of 48x96 (top) to the 16x16 context input size (bottom) for cDDPM. Right panel shows the distribution of an ensemble of the cDDPM-generated mean surface temperature response.

D Validation of Equilibrium Climate Sensitivity

We assess the performance of the cDDPM by validating its equilibrium climate sensitivity (ECS), defined as the global mean surface temperature response to a doubling of atmospheric CO_2. The necessary forcing field was obtained from the Community Earth System Model (CESM). Figure 7 (left) shows the original forcing perturbation field alongside the coarsened version used as the cDDPM context input, while Figure 7 (right) presents the cDDPM-generated T_0 distribution.

Although there is no definitive ground truth for ECS, we compared the cDDPM's response with estimates from state-of-the-art climate models, specifically the CMIP5 and CMIP6 simulations. The model range of ECS for CMIP5 is 2.1 to 4.7K, and for CMIP6, it is 1.8 to 5.6K (Schlund et al., 2020; Meehl et al., 2020). The ECS range generated by the cDDPM falls between approximately 2.7 to 3.2K, well within the ranges provided by these advanced climate models. This comparison suggests that, despite being trained on data from an intermediate-complexity climate model, the cDDPM is capable of producing realistic global mean temperature responses consistent with more complex and computationally intensive climate models.

	$cntx_{all}$	$attn\_cntx_{all}$	$cntx_{up}$	$attn\_cntx_{up}$	$cntx_{dn}$	$attn\_cntx_{dn}$
Self-attention block	No	Yes	No	Yes	No	Yes
Downward-layer context embedding	Yes	Yes	No	No	Yes	Yes
Upward-layer context embedding	Yes	Yes	Yes	Yes	No	No
RMSE	0.500	0.428	0.418	0.491	**0.336**	0.347
NRMSE	0.00173	0.00148	0.00145	0.00170	**0.00116**	0.00120
MAE	0.367	0.305	0.315	0.382	**0.229**	0.234
ACC	0.980	0.985	0.988	0.986	**0.991**	0.990
Bias (2W)	0.100	**0.0242**	0.232	0.180	0.0604	0.0652
Bias (-2W)	0.0203	**-0.000822**	-0.0608	-0.296	0.00797	-0.0446
RMSE (linear)	0.365	0.303	0.329	0.394	**0.245**	0.265
RMSE (nonlinear)	0.347	0.302	0.259	0.294	0.232	**0.226**
ACC (linear)	0.989	0.992	0.993	0.992	**0.995**	0.994
ACC (nonlinear)	0.348	0.378	0.505	0.572	0.589	**0.604**

Table 1: Performance scoreboard for independent test cases comparing cDDPM variants.

E RUN-TIME COMPARISON

In this section, we compare the runtime of a traditional climate model (PlaSim) with our machine learning-based alternative, ClimGen. This comparison is non-trivial because the climate model operates in an autoregressive manner, where each snapshot is generated by running the model using a previous snapshot as the initial condition. In contrast, ClimGen, our proposed diffusion model-based alternative, simultaneously generates multiple snapshots. Additionally, PlaSim was run on a CPU, while ClimGen was executed on a GPU. Despite these differences, we present the respective runtimes to illustrate how ClimGen could serve as an efficient alternative to traditional climate models like PlaSim.

When prescribed a radiative forcing field at the TOA, ClimGen directly generates an ensemble of 100 annually averaged temperature responses in 47.85 seconds using a single NVIDIA H100 GPU. In contrast, PlaSim requires 40 years of simulated time to reach an equilibrium climate state, followed by an additional 100 years of simulated time to obtain 100 annually averaged temperature responses. This combined 140 years of simulated time takes 41,107.98 seconds using a single Intel(R) Xeon(R) CPU E5-2697 v4.

F ABLATION STUDIES

To identify the best-performing model, we experimented with several variants of the cDDPM by modifying the backbone U-Net architecture. These variants differed in the location of the context embedding within the U-Net and the inclusion of a self-attention block at the bottleneck.

Table 1 outlines the architectural differences among the variants and compares their performance using a comprehensive set of error metrics: RMSE (root mean squared error), NRMSE (normalized root mean squared error), MAE (mean absolute error), and ACC (anomaly correlation coefficient). Additionally, biases for the positive and negative forcing cases were analyzed, and we further assessed RMSE and ACC for both linear and nonlinear components of the temperature response as defined in Lu et al. (2020). Overall, the $cntx_{dn}$ variant demonstrated the best performance across most error metrics and consistently ranked as the second best even when it did not achieve the top score.