

# DIFFUSION-BASED JOINT TEMPERATURE AND PRECIPITATION EMULATION OF EARTH SYSTEM MODELS

**Katie Christensen<sup>1</sup>, Lyric Otto<sup>1</sup>, Seth Bassetti<sup>2</sup>, Claudia Tebaldi<sup>3</sup>, Brian Hutchinson<sup>1,4</sup>**

<sup>1</sup>Department of Computer Science, Western Washington University, Bellingham, WA

<sup>2</sup>Computer Science Department, Utah State University, Logan, UT

<sup>3</sup>Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD

<sup>4</sup>AI and Data Analytics Division, Pacific Northwest National Laboratory, Richland, WA

{chris90, ottol, brian.hutchinson}@wwu.edu,  
seth.bassetti@usu.edu, claudia.tebaldi@pnnl.gov

## ABSTRACT

Earth system models (ESMs) are the principal tools used in climate science to generate future climate projections under various atmospheric emissions scenarios on a global or regional scale. Generative deep learning approaches are suitable for emulating these tools due to their computational efficiency and ability, once trained, to generate realizations in a fraction of the time required by ESMs. We extend previous work that used a generative probabilistic diffusion model to emulate ESMs by targeting the joint emulation of multiple variables, temperature and precipitation, by a single diffusion model. Joint generation of multiple variables is critical to generate realistic samples of phenomena resulting from the interplay of multiple variables. The diffusion model emulator takes in the monthly mean-maps of temperature and precipitation and produces the daily values of each of these variables that exhibit statistical properties similar to those generated by ESMs. Our results show the outputs from our extended model closely resemble those from ESMs on various climate metrics including dry spells and hot streaks, and that the joint distribution of temperature and precipitation in our sample closely matches those of ESMs.

## 1 INTRODUCTION

Earth system models (ESMs) simulate large scale phenomena and extreme weather events and provide insights into the effects of human activity on Earth’s climate at the global and regional scales. These models can offer decision-makers crucial insights for addressing the impacts of future climate scenarios on various systems, including energy and land use. Due to the rarity of extreme events, ESMs must be run numerous times to obtain enough realizations for robust statistics of rare events. This poses an issue when considering ESMs’ significant computational demand and multiple experimental uses. Emulators address this issue by producing realistic future climate projections across various emissions scenarios while significantly reducing the time and computational resources needed, since, once trained, they can generate many realizations in a computational efficient way (Kasim et al., 2021). Generative deep learning approaches, in particular, have emerged as strong candidates for developing such emulators (Addison et al., 2022; Saharia et al., 2022; Ho et al., 2022b).

We extend the DiffESM model (Bassetti et al., 2023a;b), which emulates ESMs with generative, probabilistic diffusion models. We integrate multiple variables, specifically temperature and precipitation, into the DiffESM pipeline (Bassetti et al., 2024) to model the variables’ joint spatio-temporal trends. Our model generates realizations of future behavior that mimics ESM output, and from which statistically robust estimates of metrics related to extreme weather events can be efficiently computed. We apply our emulator to one ESM under a scenario of future greenhouse gas emissions. In the original, univariate version, DiffESM produces month-long samples of either the daily mean temperature or daily total precipitation. The monthly mean maps used for its conditioning can be generated from other low-cost emulators, such as STITCHES (Tebaldi et al., 2022) or MESMER (Nath et al., 2022); DiffESM complements these existing emulator approaches. The emulation of

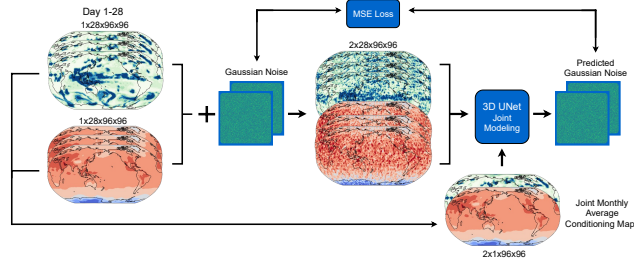


Figure 1: Training loop, illustrating the input and output channel  $C = 2$  for both daily temperature and precipitation.

single variables, however, may fail to represent the coherent relationship between variables, particularly important for the severity of extreme events that result from the combination of multiple factors, like droughts made more severe by high temperatures, or vice-versa. In our extension, DiffESM is trained to emulate these two variables, producing realistic month-long samples of both daily precipitation and daily average temperatures that capture their covariance and are consistent with their monthly means.

Additional applications of machine learning techniques in the atmospheric and climate sciences include weather forecasting (Wang et al., 2019; Scher & Messori, 2018; Rasel et al., 2018) and downscaling, i.e., increasing the spatial resolution of ESM output by borrowing information from regional models or observations on a finer scale (Hobeichi et al., 2023; Babalousmail et al., 2021; Jebeile et al., 2021). Generative adversarial networks (GANs) have shown their value in this field of application (Puchko et al., 2020; Ayala et al., 2021; Kashinath et al., 2021). For example, Hess et al. (2022) uses GANs to improve the spatial resolution of precipitation generation in finer grained locations that were previously susceptible to bias in ESMs. However, GANs have proven to be relatively difficult to train, and the easier-to-train diffusion models have become a more popular choice for generative tasks (Dhariwal & Nichol, 2021). While the statistical community has long addressed the problem of modeling the joint distribution of climate variables, their efforts have predominately focused on more limited scales, both temporally and spatially, for example, through the use of copula models (Bevacqua et al., 2017; Li et al., 2022; Sarhadi et al., 2018).

## 2 METHODS

### 2.1 DIFFESM BACKGROUND

The original DiffESM in Bassetti et al. (2023a; 2024) is a denoising generative probabilistic model. It generates samples through iterative denoising steps from a known Gaussian distribution to the target’s unknown distribution (Ho et al., 2020). The model architecture of DiffESM is influenced by Video Diffusion (Ho et al., 2022b) and Imagen Video (Ho et al., 2022a). It contains a fully convolutional U-Net (Ronneberger et al., 2015) for each denoising step, and temporal and spatial convolution layers. The input shape of the denoising step is  $T \times H \times W$ , whose elements contain the variable of interest (modelled individually).  $T = 28$  is a 28-day sequence length treated as a “month,” and  $H = 96$  and  $W = 96$  are the number of grid-boxes in the longitude and latitude dimension, i.e., the ESM grid resolution (corresponding to about 350 km in the longitude dimension at the equator, and half that length in the latitude dimension). Each denoising step maintains the same shape as its training samples, i.e., consists of  $96 \times 96$  grids of daily temperature (or precipitation) along a 28-day coherent temporal sequence. The dataset used to train DiffESM consist of daily mean temperatures (Celsius) and daily precipitation (mm) from the Institut Pierre-Simon Laplace Earth System Model (IPSL-CM5A) (Dufresne et al., 2013). DiffESM was also trained and validated, separately for the homologous output from the Community Earth System Model (CESM1-CAM5) (Kay et al., 2015) whose resolution is higher (256 by 128, corresponding to just over 100 km). To prepare the data, the temperature units are converted to degrees Celsius and normalized by dividing all values by twenty, which suffices to mitigate numerical instability. Additionally, logarithmic normalization ( $\log(x + 1)$ ) is applied to the precipitation values, managing extremes.

## 2.2 MULTIPLE VARIABLE MODELING

We extend the work of DiffESM by integrating multiple variables into the generative process (both in the conditioning and generated samples). In this extension, the samples being denoised are  $C \times T \times H \times W$ , where  $C = 2$  is a channel dimension, containing both daily temperature and precipitation variables. Figure 1 shows the training loop of our model, where the input consists of both the daily precipitation and average temperature sequence, and where the model outputs the generated 28-day month-long sequences for each variable. Once trained, the diffusion model takes as input two consistent monthly mean maps of temperature and precipitation and generates two 28-day sequences of the variables that will (a) maintain joint, spatial and temporal coherence, learned from the ESM behavior, and (b) be consistent with the input maps. In order to facilitate comparison to prior work, we train and evaluate this extended work on the IPSL-CM5A ESM run under the RCP scenarios of CMIP5 (Moss et al., 2010), with intentions to move to CMIP6 (O’Neill et al., 2016) for future developments. The  $T$ ,  $H$ , and  $W$  are the same as reported in Bassetti et al. (2023a; 2024), and our data normalization (see Sec. 2.1) is the same, as well. We use a learning rate of 0.0001, a batch size of 64 across four GPUs, and we train for 10 epochs using the Adam optimizer (Kingma & Ba, 2015).

## 3 EXPERIMENTS

Our dataset consists of both daily precipitation and daily mean temperature outputs from the IPSL-CM5A ESM on a  $96 \times 96$  grid. We use a total of 6 ESM realizations per dataset, each of them consisting of both temperature and precipitation daily output. The ESM experiments cover the period from pre-industrial times to 2100, using “historical” anthropogenic forcings from 1850 to 2006, and the RCP8.5 scenario (the most extreme emissions scenario) from 2006 to 2100. We use four of the six bivariate realizations as the training set, one realization as the first held out (Held Out 1) realization, and the remaining one realization as the second held out (Held Out 2) realization (see below for the usage of the latter two).

### 3.1 METRICS

After training, we use the Held Out 1 realization to create pairs of monthly mean maps by averaging temperature and precipitation over each 28-day “month” in the range 2080-2100. We feed each pair of monthly mean maps to the diffusion model to generate one 28-day bivariate sample of daily temperature and precipitation per month. We rely upon these monthly means to preserve the temporal coherence from month to month, while the diffusion model learns to preserve the intra-month coherence of the daily timeseries. We repeat this process using the Held Out 2 realization and generate an additional 28-day bivariate sample per month. In the absence of specification, reference to a “generated sample” pertains to a sample conditioned on Held Out 1 monthly means.

We compute temperature and precipitation metrics (summaries of daily behavior within each sequence of 28 days) per sample for each spatial location, then average over the months to produce metric maps. Likewise, we produce metric maps for the Held Out 1 and Held Out 2 sets during the same time range. We evaluate the performance of the emulator by comparing these metric maps; specifically, we take the difference of the generated samples and Held Out 2, and compare that to the difference between Held Out 1 and Held Out 2. The latter, comparing two realizations from the ESM that are different only because of internal noise gives us a measure of the unavoidable difference between realizations. To summarize the discrepancies between the generated or Held Out 1 samples and the Held Out 2 samples, we also produce superimposed histograms of the differences at all the spatial locations in these difference maps. We repeat this process using the generated samples from the Held Out 2 monthly means.

For three individual locations, representative of different climates (the ESM grid points closest to Honolulu, Hawaii; Melbourne, Australia; and Novosibirsk, Russia) we also produce contour maps to show the correlation between the two variables using the generated sample and compare to the samples from the Held Out 1 and Held Out 2 realizations. We separate the samples into dry and wet days where precipitation is  $< 1.00$  mm and  $\geq 1.00$  mm, respectively. We then calculate the deciles of the precipitation and temperature data from the Held Out 2 sample and use these to partition the days where the temperature and precipitation fall within each decile, and calculate the

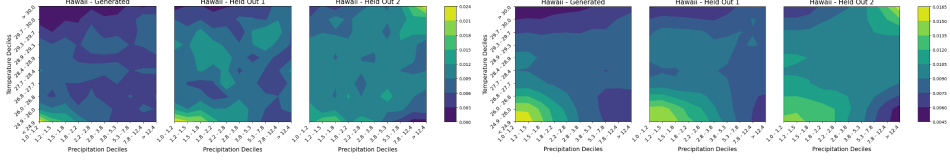


Figure 2: Joint variables discretized into temperature and precipitation deciles computed from the Held Out 2 realization. The distribution is computed only over wet days (when precipitation  $\geq 1.00$  mm), for location Hawaii. The sets of contour maps contain no smoothing (left) and average smoothing with a 3 by 3 kernel (right). Each set compares the distributions of the generated (left), Held Out 1 (middle), and Held Out 2 (right). Distributions are computed from 252 28-day samples.

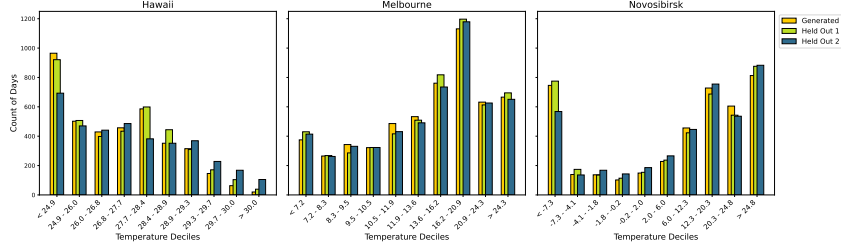


Figure 3: Joint distribution of variables discretized into temperature deciles computed from the Held Out 2 realization. The distribution is computed only over dry days (when precipitation  $< 1.00$  mm), for locations Hawaii (left), Melbourne (middle), Novosibirsk (right). The three bars of the figure compare the distributions of the generated realizations (yellow) to that of the Held Out 1 (green) and Held Out 2 (blue). Distributions are computed on the basis of 252 28-day samples.

fraction of days within each partition relative to the total number of wet days in the map. We plot these empirical joint distributions as contour maps and heatmaps to show the difference between the variable correlation of the generated sample compared to the correlations found in Held Out 1 and Held Out 2. We also produce histograms of the temperature distribution of dry days over the same temperature deciles from the wet days of Held Out 2.

### 3.2 RESULTS

Our results demonstrate the coherent relationship between variables from the generated sample of the IPSL-CM5A dataset. Figure 2 reveals these trends across wet days where precipitation is  $\geq 1.00$  mm. The generated sample is shown on the left, Held Out 1 in the middle, and Held Out 2 on the right for the location Hawaii. Appendix A shows the additional locations Melbourne (top) and Novosibirsk (bottom) in Figure 5, as well as the heatmaps across all three locations in Figure 6. The contour maps exhibit similar peak and valley patterns between the generated, Held Out 1, and Held Out 2, reflecting agreement in the bivariate distribution. Figure 3 shows the distribution of dry days where precipitation is  $< 1.00$  mm. It is noticeable that the model marginally overpredicts temperature on cool, dry days and slightly underpredicts temperature on warm, dry days. It is apparent that these variables are interdependent as we observe the correlation between high temperatures and low precipitation, as well as low temperatures and high precipitation.

We also plot four pairs of difference maps in Figure 4 for the precipitation metrics average monthly dry days and dry spell, and the temperature metrics average monthly hot streak and hot days. For each pair, the generated minus Held Out 2 is shown on the left and the Held Out 1 minus Held Out 2 is on the right. We also compare the difference distributions and find these to be similar between the generated minus Held Out 2 (orange) and the Held Out 1 minus Held Out 2 (blue). We plot these metrics in Figure 4 for the bivariate generated sample conditioned on the Held Out 1 monthly means. We do expect there to be some variability between these, and our results show a similar magnitude of variability between the generated sample and Held Out 2 to the naturally occurring variability between Held Out 1 and Held Out 2. These trends are reflected across the climate metrics hot streak and dry spell, demonstrating the model’s capability of capturing the intra-month temporal coherence

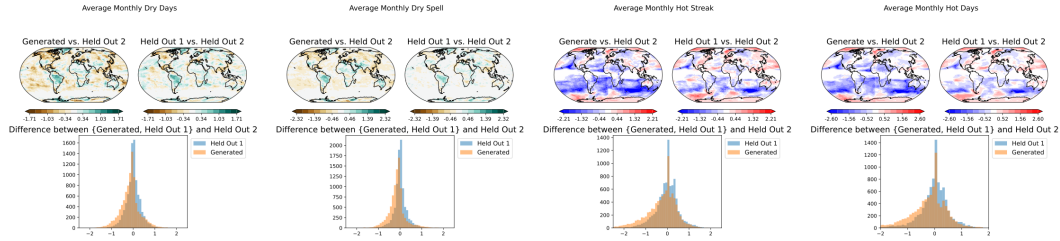


Figure 4: Pairs of difference maps between generated and Held Out 2 (left) and Held Out 1 and Held Out 2 realizations (right), and the corresponding superimposed grid-box error histograms of generated and Held Out 2 (orange) and Held Out 1 and Held Out 2 (blue) for two precipitation metrics (left) and two temperature metrics (right) for the bivariate sample conditioned on the Held Out 1 monthly means.

of ESMs. Appendix B discusses additional analyses, comparing to the univariate generated sample and the bivariate generated sample conditioned on Held Out 2.

## 4 CONCLUSIONS AND FUTURE WORK

We show that integrating joint temperature and precipitation emulation into the DiffESM generative diffusion model effectively produces joint temperature-precipitation samples that not only match the ESM in the marginal distributions, but also in their joint distribution. Specifically, we observe results similar to those reported in previous work Bassetti et al. (2023a; 2024) on single variable generation for various climate metrics based on daily behavior, including dry spells and hot days. We further demonstrate the ability to recreate the interrelationship between temperature and precipitation. This is a promising result as there are clear advantages in incorporating multiple variables into the ESM emulator and producing consistent joint realizations; this represents a step forward in the generation of realistic samples of climate scenarios. The joint behavior of variables is particularly important for addressing the emulation of extreme events that depend on both variables exacerbating hazardous conditions, like heat extremes made more dangerous by humidity, droughts made more severe by heat, and vice-versa heat domes made more persistent by dryness of the land surface. Along these lines, further experimentation includes integrating additional variables such as daily relative humidity, and daily high and low temperatures. We also plan to train and evaluate on additional ESMs including the CESM1-CAM5 ESM and the MIROC ESM Watanabe et al. (2011). Finally, we are working on emulating multiple ESMs with a single diffusion model, capable of producing samples in the style of any given constituent model, but potentially benefiting from having been trained across a larger set of ESMs, realizations, and scenarios.

## ACKNOWLEDGMENTS

This research was supported by the U.S. Department of Energy, Office of Science, under the MultiSector Dynamics, Earth and Environmental System Modeling Program. The Pacific Northwest National Laboratory, operated by Battelle Memorial Institute under contract DE-AC05-76RL01830, facilitated this research. The authors express gratitude to the World Climate Research Programme’s Working Group on Coupled Modelling for CMIP and extend appreciation to the climate modeling team at Institut Pierre Simon Laplace (France) for sharing its model outputs. Coordinating support for CMIP was provided by the U.S. Department of Energy’s Program for Climate Model Diagnosis and Intercomparison, in collaboration with the Global Organization for Earth System Science Portals, which also contributed to the development of software infrastructure.

## REFERENCES

- Henry Addison, Elizabeth Kendon, Suman Ravuri, Laurence Aitchison, and Peter AG Watson. Machine learning emulation of a local-scale uk climate model. *arXiv preprint arXiv:2211.16116*, 2022.
- Alexis Ayala, Christopher Drazic, Brian Hutchinson, Ben Kravitz, and Claudia Tebaldi. Loosely conditioned emulation of global climate models with generative adversarial networks. *arXiv preprint arXiv:2105.06386*, 2021.
- Hassen Babaousmail, Rongtao Hou, Gnim Tchalim Gnitou, and Brian Ayugi. Novel statistical downscaling emulator for precipitation projections using deep convolutional autoencoder over northern africa. *Journal of Atmospheric and Solar-Terrestrial Physics*, 218:105614, 2021. ISSN 1364-6826. doi: <https://doi.org/10.1016/j.jastp.2021.105614>. URL <https://www.sciencedirect.com/science/article/pii/S1364682621000754>.
- Seth Basseti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of earth system models with diffusion models. *arXiv preprint arXiv:2304.11699*, 2023a.
- Seth Basseti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. DiffESM, December 2023b. URL <https://github.com/sethbasseti/diff-esm>.
- Seth Basseti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. DiffESM: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. January 2024. doi: 10.22541/essoar.170612214.43012845/v1. URL <http://dx.doi.org/10.22541/essoar.170612214.43012845/v1>.
- Emanuele Bevacqua, Douglas Maraun, Ingrid Hobaek Haff, Martin Widmann, and Mathieu Vrac. Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in ravenna (italy). *HYDROLOGY AND EARTH SYSTEM SCIENCES*, 21(6):2701–2723, JUN 8 2017. ISSN 1027-5606. doi: 10.5194/hess-21-2701-2017.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- J-L Dufresne, M-A Foujols, Sébastien Denvil, Arnaud Caubel, Olivier Marti, Olivier Aumont, Yves Balkanski, Slimane Bekki, Hugo Bellenger, Rachid Benshila, et al. Climate change projections using the ipsl-cm5 earth system model: from cmip3 to cmip5. *Climate dynamics*, 40:2123–2165, 2013.
- Philipp Hess, Markus Drüke, Stefan Petri, Felix M. Strnad, and Niklas Boers. Physically constrained generative adversarial networks for improving precipitation fields from earth system models, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf).
- Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022b.
- Sanaa Hobeichi, Nidhi Nishant, Yawen Shao, Gab Abramowitz, Andy Pitman, Steve Sherwood, Craig Bishop, and Samuel Green. Using machine learning to cut the cost of dynamical downscaling. *Earth’s Future*, 11(3):e2022EF003291, 2023.
- Julie Jebeile, Vincent Lam, and Tim Räz. Understanding climate change with statistical downscaling and machine learning. *Synthese*, 199:1877–1897, 2021.

- Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, A Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021.
- Muhammad Firmansyah Kasim, D Watson-Parris, L Deaconu, S Oliver, P Hatfield, Dustin H Froula, Gianluca Gregori, M Jarvis, S Khatiwala, J Korenaga, et al. Building high accuracy emulators for scientific simulations with deep neural architecture search. *Machine Learning: Science and Technology*, 3(1):015013, 2021.
- Jennifer E Kay, Clara Deser, A Phillips, A Mai, Cecile Hannay, Gary Strand, Julie Michelle Arblaster, SC Bates, Gokhan Danabasoglu, James Edwards, et al. The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96(8): 1333–1349, 2015.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015*, 2015. URL <http://arxiv.org/abs/1412.6980>. Conference Track Proceedings.
- Kailong Li, Guohe Huang, Shuo Wang, Saman Razavi, and Xiaoyue Zhang. Development of a joint probabilistic rainfall-runoff model for high-to-extreme flow projections under changing climatic conditions. *Water Resources Research*, 58(6):e2021WR031557, 2022. doi: <https://doi.org/10.1029/2021WR031557>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021WR031557>. e2021WR031557 2021WR031557.
- Richard H. Moss, Jae A. Edmonds, Kathy A. Hibbard, Martin R. Manning, Steven K. Rose, Detlef P. van Vuuren, Timothy R. Carter, Seita Emori, Mikiko Kainuma, Tom Kram, Gerald A. Meehl, John F. B. Mitchell, Nebojsa Nakicenovic, Keywan Riahi, Steven J. Smith, Ronald J. Stouffer, Allison M. Thomson, John P. Weyant, and Thomas J. Wilbanks. The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282):747–756, Feb 2010. ISSN 1476-4687. doi: 10.1038/nature08823. URL <https://doi.org/10.1038/nature08823>.
- 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/>.
- B. C. O’Neill, C. Tebaldi, D. P. van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi, and B. M. Sanderson. The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482, 2016. doi: 10.5194/gmd-9-3461-2016. URL <https://gmd.copernicus.org/articles/9/3461/2016/>.
- Alexandra Puchko, Robert Link, Brian Hutchinson, Ben Kravitz, and Abigail Snyder. Deepclimgan: A high-resolution climate data generator. *arXiv preprint arXiv:2011.11705*, 2020.
- Risul Islam Rasel, Nasrin Sultana, and Phayung Meesad. An application of data mining and machine learning for weather forecasting. In Phayung Meesad, Sunantha Sodsee, and Herwig Unger (eds.), *Recent Advances in Information and Communication Technology 2017*, pp. 169–178, Cham, 2018. Springer International Publishing. ISBN 978-3-319-60663-7.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

- Ali Sarhadi, María Concepción Ausín, Michael P. Wiper, Danielle Touma, and Noah S. Diffenbaugh. Multidimensional risk in a nonstationary climate: Joint probability of increasingly severe warm and dry conditions. *Science Advances*, 4(11):eaau3487, 2018. doi: 10.1126/sciadv.aau3487. URL <https://www.science.org/doi/abs/10.1126/sciadv.aau3487>.
- Sebastian Scher and Gabriele Messori. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2830–2841, 2018. doi: <https://doi.org/10.1002/qj.3410>. URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3410>.
- Claudia Tebaldi, Abigail Snyder, and Kalyn Dorheim. Stitches: creating new scenarios of climate model output by stitching together pieces of existing simulations. *Earth System Dynamics*, 13(4): 1557–1609, 2022.
- Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, and Guangquan Zhang. Deep uncertainty quantification: A machine learning approach for weather forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19, pp. 2087–2095, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330704. URL <https://doi.org/10.1145/3292500.3330704>.
- S. Watanabe, T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, T. Ise, H. Sato, E. Kato, K. Takata, S. Emori, and M. Kawamiya. Miroc-esm 2010: model description and basic results of cmip5-20c3m experiments. *Geoscientific Model Development*, 4(4):845–872, 2011. doi: 10.5194/gmd-4-845-2011. URL <https://gmd.copernicus.org/articles/4/845/2011/>.

## A ADDITIONAL CONTOUR PLOTS AND HEATMAPS

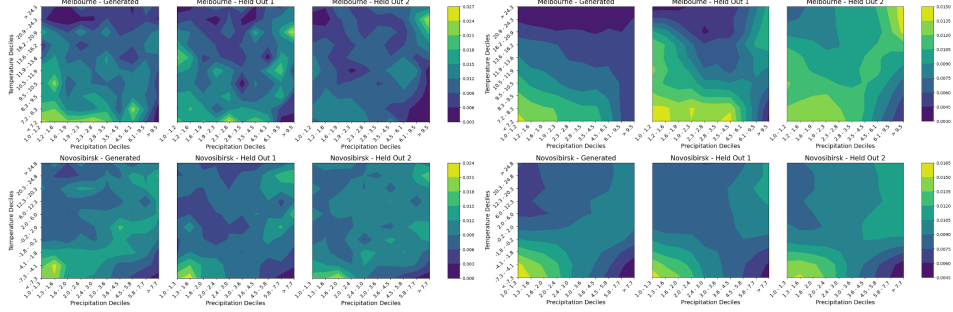


Figure 5: Joint distribution of temperature and precipitation discretized into temperature and precipitation deciles computed from the Held Out 2 realization over wet days (when precipitation  $\geq 1.00$  mm), for locations Melbourne (top), Novosibirsk (bottom). The set of contour maps on the left contain no smoothing, the set on the right contain average smoothing with a 3 by 3 kernel. The three columns of the figure compare the distributions of the generated realizations (left) to that of Held Out 1 (middle) and Held Out 2 (right). Distributions are computed on the basis of 252 28-day samples.

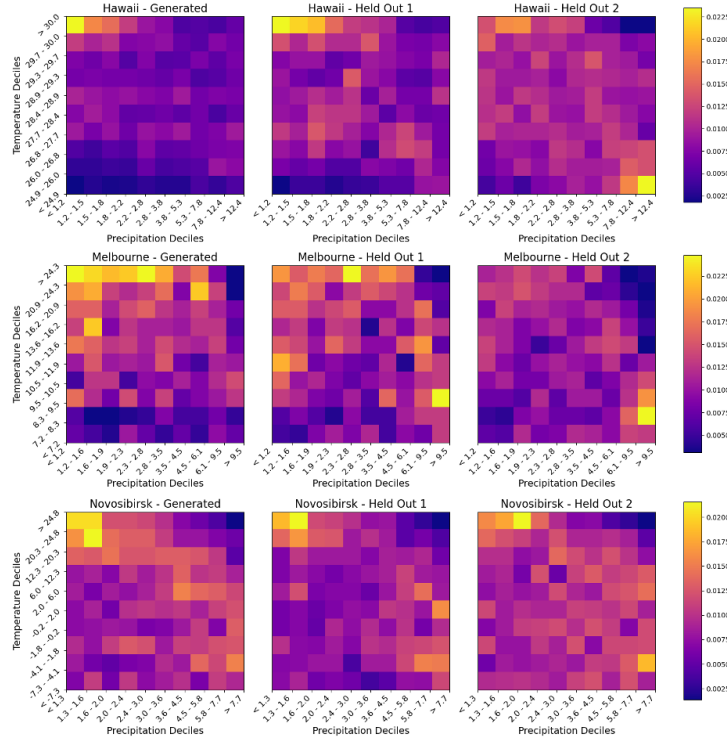


Figure 6: Joint distribution of temperature and precipitation discretized into temperature and precipitation deciles computed from the Held Out 2 realization over wet days (when precipitation  $\geq 1.00$  mm), for locations Hawaii (top), Melbourne (middle), Novosibirsk (bottom). The three columns of the figure compare the distributions of the generated realizations (left) to that of Held Out 1 (middle) and Held Out 2 (right). Distributions are computed on the basis of 252 28-day samples.

## B ADDITIONAL DIFFERENCE MAPS

Figure 7 replicates the metric map experiment using a univariate DiffESM model (top) and the bi-variate model, but conditioned on Held Out 2 (bottom). For the bottom row, the histograms compare the difference distributions of the generated on Held Out 2 minus Held Out 2 (purple), and the Held Out 1 minus Held Out 2 (blue). We expect the generated on Held Out 2 minus Held Out 2 to exhibit a narrower distribution than Held Out 1 minus Held Out 2 due to the variability between the two held out realizations, while still maintaining a degree of variability owing to the randomness of the generative process. These expectations are consistent with the results in Figure 7.

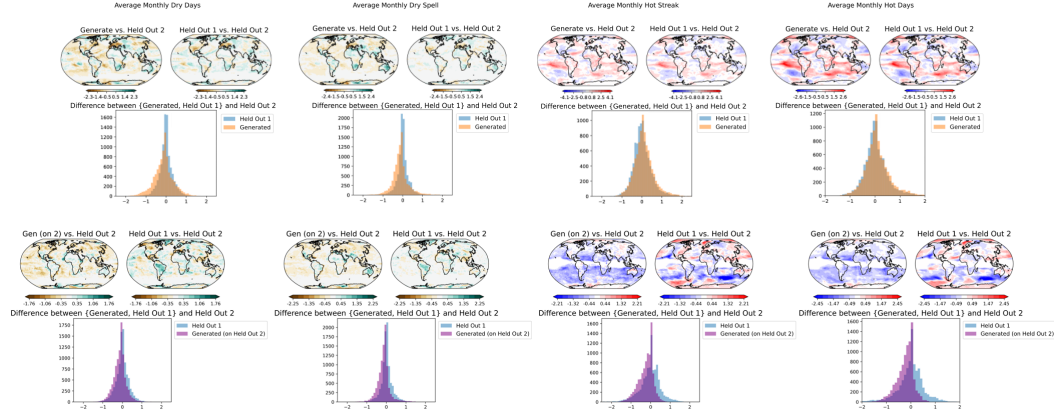


Figure 7: Pairs of difference maps between generated and Held Out 2 (left) and Held Out 1 and Held Out 2 realizations (right), and the corresponding superimposed grid-box error histograms of generated on Held Out 1 and Held Out 2 (orange), or generated on Held Out 2 and Held Out 2 (purple), and Held Out 1 and Held Out 2 (blue), for two precipitation metrics (left) and two temperature metrics (right) for the single variable samples (top), and the generated samples conditioned on the Held Out 2 monthly means (bottom).