Loosely Conditioned Emulation of Global Climate Models With Generative Adversarial Networks

Alexis Ayala[1], **Christopher Drazic**[1], **Brian Hutchinson**[1,2], **Ben Kravitz**[3], **and Claudia Tebaldi**[4]
[1] Computer Science Department, Western Washington University, Bellingham, WA
[2] Computing & Analytics Division, Pacific Northwest National Laboratory, Richland, WA
[3] Earth and Atmospheric Sciences Department, Indiana University, Bloomington, IN
[4] Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD

1 Introduction

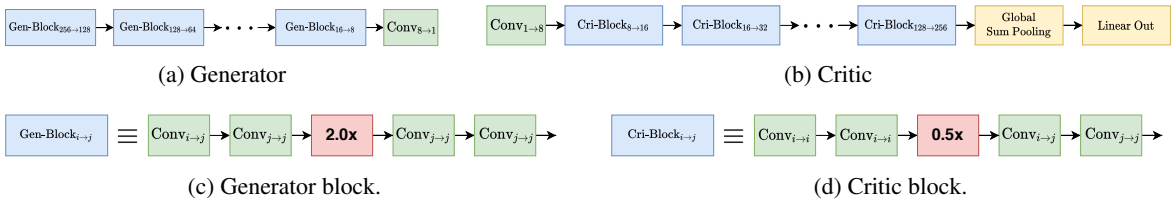
Climate models encapsulate our best understanding of the Earth system, allowing research to be conducted on its future under alternative assumptions of how human-driven climate forces are going to evolve. An important application of climate models is to provide metrics of mean and extreme climate changes, particularly under these alternative future scenarios, as these quantities drive the impacts of climate on society and natural systems [1, 2, 3]. Furthermore, efforts in integrated modeling seek to "close the loop," by having impacts on society feedback on societal conditions that drive emissions [4]. Because of the need to explore a wide range of alternative scenarios and other sources of uncertainties in a computationally efficient manner, climate models can only take us so far, as they require large computational resources, with a single simulation of the 21^{st} century taking on the order of weeks on a supercomputer. The computational requirements expand considerably when attempting to characterize extreme events, which are rare and thus demand long and numerous simulations exploring a noisy system in order to accurately represent their changing statistics.

Climate model emulators address some of these problems. Trained on climate model output, emulators are simpler, data-driven tools (often obtained through parametric fits like regressions) that are less accurate or less complex than climate models but can produce values in fractions of a second on a laptop. Their computational cost, when significant, is made upfront in the training phase. Traditionally, emulators like Pattern Scaling have been used to approximate average quantities, like annual or seasonal or monthly average temperature and precipitation [5, 6, 7, 8, 9, 10, 11, 12]. Recently the accuracy of some of these techniques for representing extremes has been documented as well [13]. "Top-down" approaches to emulation involve directly approximating metrics themselves, like the hottest or the wettest day of the year. Alternatively, a "bottom-up" approach tackles the emulation of the building blocks (in these cases, daily temperatures and precipitation during the year) and then compute the metric of interest. Examples include stochastic weather generators [14, 15, 16], which rely on parametrizing the distribution of the weather variable and randomly sample its realizations. Weather generators have been usually developed for limited domains and specific applications, rarely facing the issue of representing non-stationarities and non-linearities, which are critical for integrated modeling of impacts, arbitrary domains, and scenario generation.

Increasingly, joint efforts between climate science and machine learning are being formed to tackle some of the most complex data-driven problems [17, 18]. So far most of the applications have focused on bringing deep learning in aid of better model forecasts, model parameterizations, or in substitution of climate models [19, 20, 21, 22, 23, 24, 25, 26, 27]; of better detection of signals, from extreme events to large scale patterns of anthropogenic changes amidst the internal noise of the climate system [28, 29, 30, 31, 32, 33]; and of spatial in-filling in the case of fine-scale features that models would be too expensive, or plainly unable, to generate, or observations cannot cover [34, 35, 36, 37].

Here we use deep learning in a proof of concept that lays the foundation for emulating global climate model output for different scenarios. We train Generative Adversarial Networks (GANs) that emulate daily precipitation output from a fully coupled Earth system model. Our GANs are trained to produce samples in the form of $T \times H \times W$ tensors, where T denotes the number of timesteps (days) and H and W are the spatial height and width, respectively, of a regular grid discretizing the globe. The goal is for these samples to be statistically indistinguishable from samples of the same dimension drawn from a state-of-the-art earth system model. Our trained GAN can rapidly generate numerous realizations at a vastly reduced computational expense, compared to large ensembles of climate models [38, 39], which

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020.



(a) Generator

(b) Critic

(c) Generator block.

(d) Critic block.

Figure 1: Generator (a), critic (b), generator block (c), and critic block (d) architectures. Subscript $i \to j$ denotes the feature dimension of the input (i) and output (j). Convs in the generator block are followed by batch norm and Leaky ReLU; Convs in the critic block are followed by Leaky ReLU.

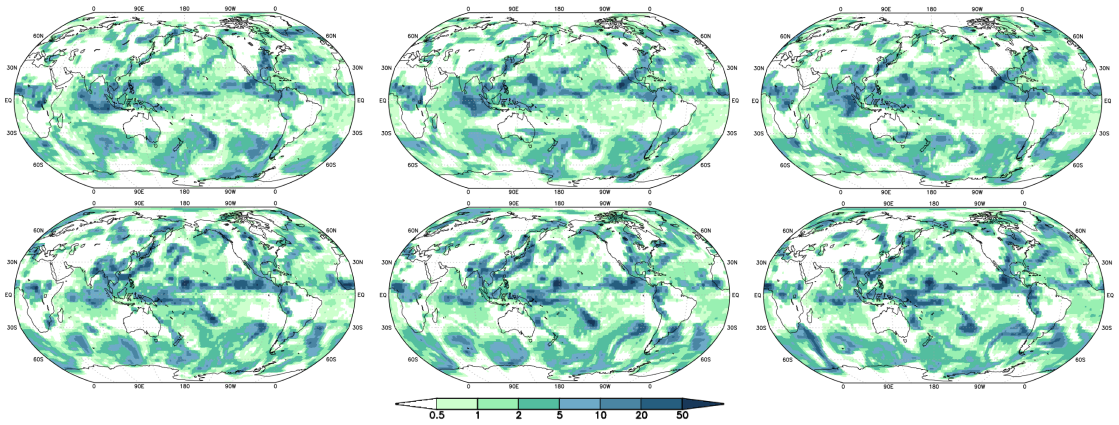


Figure 2: Generated samples. Top row: precipitation maps for three successive generated days. Bottom row: precipitation maps for three successive test days. Units: mm/day.

greatly aids in estimating the statistics of extreme events. Compared to our prior "DeepClimGAN" [40], we find that the approach proposed here produces significantly higher quality spatio-temporal samples.

2 Model

Our model is largely based off the BigGan [41] architecture, following a similar channel progression and utilizing global sum pooling, although we replace 2D convolutions with 3D and remove the residual connections. Our full generator, shown in Fig. 1a, consists of 20 convolutional layers organized into five sequential blocks, with block structure shown in 1c. It is built and trained in a block-wise progressive fashion [42]. Block i produces samples in $\mathbb{R}^{T \times H_i \times W_i}$, where T, H_i, and W_i denote the number of timesteps, height, and width, respectively. The first block takes as input a 256-dimensional noise vector and produces output in $\mathbb{R}^{32 \times 4 \times 8}$ output. Note that even the first block's output is of the full temporal resolution, $T = 32$, which initial experiments suggested performed better than progressively growing in the time dimension. The next four blocks each double the spatial resolution, yielding an overall generator output in $\mathbb{R}^{32 \times 64 \times 128}$. Every block halves the feature dimension. The critic's architecture (Fig. 1b and d) largely mirrors the generator's: it consists of five sequential blocks, each of which halves the spatial dimensions of the data and doubles the feature dimension. Batch normalization is not used in the critic, following the findings of [43]. We use a block-wise progressive training process, the details of which can be found in Appendix A. Hyperparameter details can be found in Appendix B.

3 Experimental Setup

Data We use daily output from the MIROC5 Earth System Model [44], which has fully coupled atmosphere, ocean, land, and sea ice components at a $\sim$ 140 km horizontal resolution. Although we limit the current study to precipitation, future work will generalize to other variables (e.g., temperature and humidity). We limit our study to the historical simulation (1850–2005), as described under the Coupled Model Intercomparison Project Phase 5 [45]. To account for the highly skewed distribution of daily precipitation values (mm/day), we apply $\log(1 + x)$ normalization, but undo

this normalization before computing performance metrics. We randomly split the data, by 9-year-long chunks, into training (90%), validation (5%) and test (5%) sets. We then split our data into two half-years, with different seasonal behaviors: Fall-Winter (SONDJF) and Spring-Summer (MAMJJA). For Fall-Winter, this yields 131225, 8145, and 8145 days in training, validation and test, respectively. For Spring-Summer, training, validation and test have 133400, 8280 and 8280 days, respectively. One model is trained on the Fall-Winter data, another is trained independently on the Spring-Summer data; we refer to these models as "loosely conditioned" (on the season), as opposed to a hypothetical model more tightly conditioned (e.g., on a target amount of global precipitation). Fig. 2 contains generated and test set Fall-Winter samples.

Performance Metrics We use KL divergence to measure dissimilarity between pairs of empirical distributions (e.g. those induced by samples from a generator and those by samples from the test set). Our proposed approach is to take each sample in the first set, flatten it to a vector in $\mathbb{R}^{THW}$, and fit a full-covariance multivariate normal (MVN) distribution to the set of vectors. This process is repeated for the second set, and then we compute the KL divergence between the two distributions. However, estimating the $TWH \times TWH$ covariance matrix is impractical. Instead, we break each sample up into subtensors and fit MVN distributions over these smaller vectors. We consider three ways to break the overall tensor up into smaller tensors.

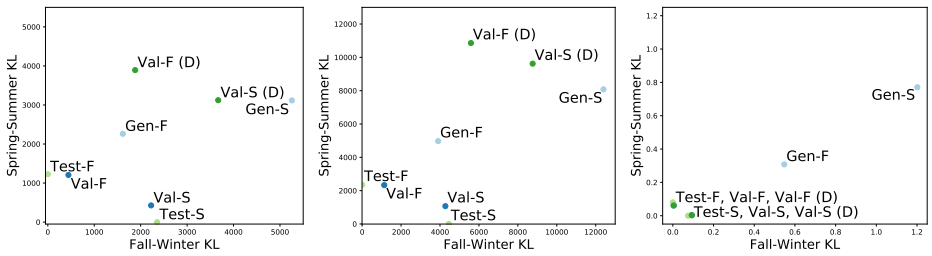
1. **Spatial KL**. We first break each sample up by time, yielding T spatial maps in $\mathbb{R}^{H \times W}$. If $H > 32$ and $W > 64$, each map is further broken into $(\frac{H}{32})(\frac{W}{64})$ maps, each in $\mathbb{R}^{32 \times 64}$.
2. **Temporal KL**. We break sample into HW temporal vectors, each in $\mathbb{R}^{T}$.
3. **Spatiotemporal KL**. Here we produce small subtensors with both time and space dimensions. Specifically, we break each sample into $(\frac{H}{4})(\frac{W}{8})$ subtensors, each in $\mathbb{R}^{T \times 4 \times 8}$.

Evaluation Process Recall that our goal is not to predict a particular outcome, but to produce a generator that defines a distribution as close to the distribution of the ESM as possible. To evaluate our "loosely conditioned" generators, we first generate two sets of N samples: one set using the generator trained on Fall-Winter data and the other the generator trained on the Spring-Summer data. We compute the three KL divergences between each of these sets of data and the two test sets (Fall-Winter and Spring-Summer). Because these values alone are not easy to interpret, we also compute and report the same performance metrics between the two half-year validation sets and the test sets. Lastly, because validation and test are both produced by the same ESM, the "performance" of validation is effectively an upper bound on our performance. We therefore also consider a degraded version of the validation data, in which zero-mean Gaussian noise is added to each element of each sample. We pick $\sigma = 0.024$ as the smallest standard deviation such that the degraded validation performance (according to the spatiotemporal KL) matches the generated data performance over both seasons.

4 Results and Discussion

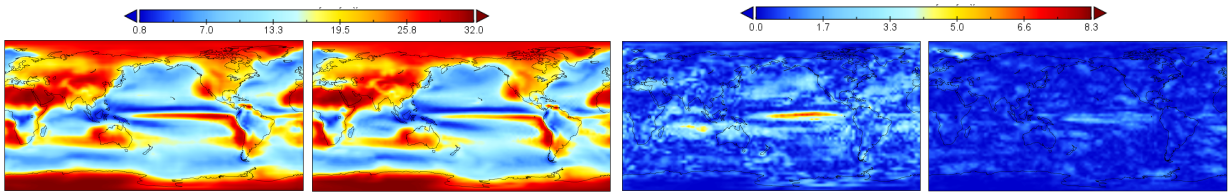
Figure 3 plots the three KL divergence metrics (a-c) for the Fall-Winter (F) and Spring-Summer (S) generated models. The x-axis coordinate represents the KL divergence between that empirical distribution and the Fall-Winter test set; likewise, the y-axis coordinate represents the KL divergence with the Spring-Summer test set. The "(D)" denotes the degraded version of validation (see above). We note that while the generated data does have higher spatial and spatiotemporal KL divergences than the clean validation data does, its performance is comparable to the validation data with only a very small amount of white noise added. By the temporal KL performance metric, the generated model is worse than the degraded validation data, which may be an result of our decision not to progressively grow the time dimension. As expected, seasonality has an effect: the half-year data almost always has lower KL divergences with the matching half-year test set data.

As an additional qualitative assessment of model performance, for each spatial location, we compute two statistics of the sets of samples: (i) the average number of dry days in each $T = 32$ sample, and (ii) the average length of the longest dry spell in each sample (longest consecutive number of dry days). Fig. 4a plots the average number of dry days in the Fall-Winter generator's samples, and Fig. 4b the average number of dry days in the Fall-Winter test data. The absolute difference between the two is plotted in panel c, and, for reference, panel d shows the absolute difference between these quantities when comparing test and validation data. Although the generated data is not quite as well matched as the validation data is, we see that most locations are no more than 1-3 days off (out of $T = 32$). There is a notable exception, where errors approach 8 days in the equatorial Pacific west of South America, which appears to be due to a tendency of the generated data to have a slightly wider and intense dry tongue than the GCM climatology. Appendix D presents analogous maps and reports similar behavior for the maximum dry spell length statistics. Future



(a) Spatiotemporal (b) Spatial (c) Temporal

Figure 3: KL divergences between Test and Validation, Validation plus noise $\sim \mathcal{N}(0, 0.024)$ (Val (D)), Test and Generated data for Fall-Winter (F) and Spring Summer (S).



(a) (b) (c) (d)

Figure 4: Maps with mean number of dry days in the (a) generated and (b) test Fall-Winter data; (c) the absolute difference between a and b; (d) the absolute difference between test and validation.

work will address this behavior, here we note that for most applications of emulated dry days the domain of interest will be land regions of the globe, where our method does not manifest any strikingly biased behavior.

5 Conclusion and Future Work

We have demonstrated the ability of a progressive GAN architecture to successfully emulate the spatio-temporal characteristics of sequences of daily precipitation as produced by an ESM. Once trained, it becomes computationally efficient to produce an arbitrary number of such sequences, akin to the output of a large initial condition ensemble. According to a number of verification methods, the GAN appears to have satisfactorily replicated the ESM's spatiotemporal behavior.

There are many ways we plan to extend the current work. The current work trains our GAN on a single earth system model and experiment; adding support for conditioning the generator would enable modeling arbitrary transient scenarios, an important next step. Established emulators of annual and monthly quantities on the basis of arbitrary scenarios of future forcing [11] could be used here as the sources of conditional information. While the current model is locked to producing a sequence of 32 days, it would be useful to develop models capable of producing any number of consecutive days. One last natural extension would be to explore the joint and sequential generation of variables (e.g. generating temperature and humidity conditioned on precipitation). These capabilities, once developed, would support a significant improvement in integrated modeling of climate change impacts by enabling a rich representation of some of the most damaging hazards and an exploration of uncertainties in scenario, model and internal variability [46] "on the fly," substituting for expensive, time consuming and necessarily constraining computational resources needed to run climate models.

Acknowledgments and Disclosure of Funding

This research was supported by the U.S. Department of Energy, Office of Science, as part of research in MultiSector Dynamics, Earth and Environmental System Modeling Program. The authors also thank the NVIDIA corporation for the donation of GPUs used in this work.

References

[1] C. Mora, D. Spirandelli, E. Franklin, J. Lynham, M. Kantar, W. Miles, C. Smith, K. Freel, J. Moy, L. V. Louis, E. W. Barba, K. Bettinger, A. Frazier, J. Colburn IX, N. Hanasaki, E. Hawkins, Y. Hirabayashi, W. Knorr, C. Little, K. Emanuel, J. Sheffield, J. Patz, and C. L. Hunter, "Broad threat to humanity from cumulative climate hazards intensified by greenhouse gas emissions," *Nature Climate Change*, vol. 8, pp. 1062–1071, 2018.

[2] G. Forzieri, A. Bianchi, F. Batista e Silva, M. A. Marin Herrera, A. Leblois, C. Lavalle, J. Aerts, and L. Feyen, "Escalating impacts of climate extremes on critical infrastructures in europe," *Global Environmental Change*, vol. 48, pp. 97 – 107, 2018.

[3] C. Raymond, R. M. Horton, J. Zscheischler, O. Martius, A. AghaKouchak, J. Balch, S. G. Bowen, S. J. Camargo, J. Hess, K. Kornhuber, M. Oppenheimer, A. C. Ruane, T. Wahl, and K. White, "Understanding and managing connected extreme events," *Nature Climate Change*, vol. 10, no. 7, pp. 611–621, Jul 2020. [Online]. Available: https://doi.org/10.1038/s41558-020-0790-4

[4] K. Calvin and B. Bond-Lamberty, "Integrated human-earth system modeling—state of the science and future directions," *Environmental Research Letters*, vol. 13, no. 6, p. 063006, jun 2018. [Online]. Available: https://doi.org/10.1088%2F1748-9326%2Faac642

[5] B. D. Santer, T. M. L. Wigley, M. E. Schlesinger, and J. Mitchell, "Developing climate scenarios from equilibrium gcm results," *Report of the Max Plank Institut f ur Meteorologie*, vol. 47, p. 29, 1990.

[6] S. Castruccio, D. J. McInerney, M. L. Stein, F. L. Crouch, R. L. Jacob, and E. J. Moyer, "Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs," *JOURNAL OF CLIMATE*, vol. 27, no. 5, pp. 1829–1844, MAR 2014.

[7] P. B. Holden, N. R. Edwards, P. H. Garthwaite, K. Fraedrich, F. Lunkeit, E. Kirk, M. Labriet, A. Kanudia, and F. Babonneau, "PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment," *GEOSCIENTIFIC MODEL DEVELOPMENT*, vol. 7, no. 1, pp. 433–451, 2014.

[8] C. Tebaldi and J. M. Arblaster, "Pattern scaling: Its strengths and limitations, and an update on the latest model simulations," *Climatic Change*, vol. 122, no. 3, pp. 459–471, Feb 2014. [Online]. Available: https://doi.org/10.1007/s10584-013-1032-9

[9] N. Herger, B. M. Sanderson, and R. Knutti, "Improved pattern scaling approaches for the use in climate impact studies," *Geophysical Research Letters*, vol. 42, no. 9, pp. 3486–3494, 2015.

[10] B. Kravitz, C. Lynch, C. Hartin, and B. Bond-Lamberty, "Exploring precipitation pattern scaling methodologies and robustness among cmip5 models," *Geoscientific Model Development*, vol. 10, no. 5, pp. 1889–1902, 2017. [Online]. Available: https://www.geosci-model-dev.net/10/1889/2017/

[11] R. Link, A. Snyder, C. Lynch, C. Hartin, B. Kravitz, and B. Bond-Lamberty, "Fldgen v1. 0: an emulator with internal variability and space–time correlation for earth system models," *Geoscientific Model Development (Online)*, vol. 12, no. 4, pp. 1477–1489, 2019.

[12] L. Beusch, L. Gudmundsson, and S. I. Seneviratne, "Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land," *EARTH SYSTEM DYNAMICS*, vol. 11, no. 1, pp. 139–159, FEB 17 2020.

[13] C. Tebaldi, A. Armbruster, H. P. Engler, and R. Link, "Emulating climate extreme indices," *Environmental Research Letters*, vol. 15, no. 7, p. 074006, jun 2020. [Online]. Available: https://doi.org/10.1088%2F1748-9326%2Fab8332

[14] M. A. Semenov and E. M. Barrow, "Use of a stochastic weather generator in the development of climate change scenarios," *Climatic Change*, vol. 35, no. 4, pp. 397–414, Apr 1997. [Online]. Available: https://doi.org/10.1023/A:1005342632279

[15] C. Kilsby, P. Jones, A. Burton, A. Ford, H. Fowler, C. Harpham, P. James, A. Smith, and R. Wilby, "A daily weather generator for use in climate change studies," *Environmental Modelling & Software*, vol. 22, no. 12, pp. 1705 – 1719, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S136481520700031X

[16] S. Fatichi, V. Y. Ivanov, and E. Caporali, "Simulation of future climate scenarios with a weather generator," *Advances in Water Resources*, vol. 34, no. 4, pp. 448 – 467, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0309170811000042

[17] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-0912-1

[18] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Körding, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio, "Tackling climate change with machine learning," *CoRR*, vol. abs/1906.05433, 2019. [Online]. Available: http://arxiv.org/abs/1906.05433

[19] J. Cohen, D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, "S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts," *WIREs Climate Change*, vol. 10, no. 2, p. e00567, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.567

[20] A. Grover, A. Kapoor, and E. Horvitz, "A deep hybrid model for weather forecasting," in *Proc. KDD*, 2015, p. 379–386.

[21] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, "Deep learning for multi-year enso forecasts," *Nature*, vol. 573, no. 7775, pp. 568–572, Sep 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1559-7

[22] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Deep learning for precipitation nowcasting: A benchmark and a new model," in *Proc. NIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5617–5627.

[23] N. Jones, "How machine learning could help to improve climate forecasts," *Nature*, vol. 548, no. 7668, pp. 379–380, Aug 2017.

[24] S. He, X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, "Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2006.07972

[25] T. Schneider, S. Lan, A. Stuart, and J. Teixeira, "Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations," *Geophysical Research Letters*, vol. 44, no. 24, pp. 12,396–12,417, 2017. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL076101

[26] T. Weber, A. Corotan, B. Hutchinson, B. Kravitz, and R. Link, "Technical note: Deep learning for creating surrogate models of precipitation in earth system models," *Atmospheric Chemistry and Physics*, vol. 20, no. 4, pp. 2303–2317, 2020. [Online]. Available: https://acp.copernicus.org/articles/20/2303/2020/

[27] V. Schmidt, M. Alghali, K. Sankaran, T. Yuan, and Y. Bengio, "Modeling cloud reflectance fields using conditional generative adversarial networks," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2022.07579

[28] Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1605.01156

[29] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1906.04928

[30] K. Klemmer, A. Koshiyama, and S. Flennerhag, "Augmenting correlation structures in spatial data using deep generative models," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1905.09796

[31] E. A. Barnes, B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, "Indicator patterns of forced change learned by an artificial neural network," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 9, 2020.

[32] B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff, "Physically interpretable neural networks for the geosciences: Applications to earth system variability," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 9, 2020.

[33] R. C. J. Wills, D. S. Battisti, K. C. Armour, T. Schneider, and C. Deser, "Pattern Recognition Methods to Separate Forced Responses from Internal Variability in Climate Model Ensembles and Observations," *Journal of Climate*, vol. 33, no. 20, pp. 8693–8719, 09 2020. [Online]. Available: https://doi.org/10.1175/JCLI-D-19-0855.1

[34] M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss, "Improving the accuracy of rainfall rates from optical satellite sensors with machine learning — a random forests-based approach applied to msg seviri," *Remote Sensing of Environment*, vol. 141, pp. 129 – 143, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425713003945

[35] F. Amato, F. Guignard, S. Robert, and M. Kanevski, "A novel framework for spatio-temporal prediction of climate data using deep learning," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2007.11836

[36] T. Vandal, E. Kodra, S. Ganguly, A. R. Michaelis, R. R. Nemani, and A. R. Ganguly, "Deepsd: Generating high resolution climate change projections through single image super-resolution," *CoRR*, vol. abs/1703.03126, 2017. [Online]. Available: http://arxiv.org/abs/1703.03126

[37] K. Stengel, A. Glaws, D. Hettinger, and R. N. King, "Adversarial super-resolution of climatological wind and solar data," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 16 805–16 815, 2020.

[38] J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. M. Arblaster, S. C. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein, "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*, vol. 96, no. 8, pp. 1333–1349, 09 2015. [Online]. Available: https://doi.org/10.1175/BAMS-D-13-00255.1

[39] F. Lehner, C. Deser, N. Maher, J. Marotzke, E. M. Fischer, L. Brunner, R. Knutti, and E. Hawkins, "Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6," *EARTH SYSTEM DYNAMICS*, vol. 11, no. 2, pp. 491–508, MAY 29 2020.

[40] A. Puchko, R. Link, B. Hutchinson, B. Kravitz, and A. Snyder, "DeepClimGAN: A high-resolution climate data generator," *CorRR*, 2020. [Online]. Available: http://arxiv.org/abs/2011.11705

[41] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1809.11096

[42] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, 2017. [Online]. Available: http://arxiv.org/abs/1710.10196

[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *CoRR*, vol. abs/1704.00028, 2017. [Online]. Available: http://arxiv.org/abs/1704.00028

[44] S. Watanabe, T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, T. Ise, and H. Sato, "Miroc-esm: model description and basic results of cmip5-20c3m experiments," *Geoscientific Model Development Discussions*, vol. 4, 05 2011.

[45] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of CMIP5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, pp. 485–498, 2012. [Online]. Available: https://doi.org/10.1175/BAMS-D-11-00094.1

[46] E. Hawkins and R. Sutton, "The potential to narrow uncertainty in regional climate predictions," *Bulletin of the American Meteorological Society*, vol. 90, no. 8, pp. 1095–1108, 2009.

[47] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

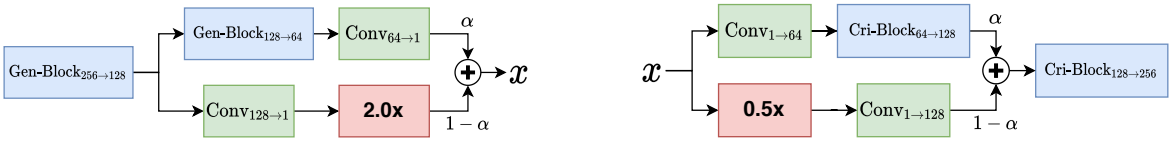
[48] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proc. NIPS*, 2012, p. 2951–2959.

A Progressive Training

We employ progressive training [42]. Specifically, we first train the first block of the generator and critic, which learn to produce and evaluate, respectively, samples with low spatial resolution. After the first block has converged, we train the first two blocks of the generator and critic. This process continues until all blocks have been trained. Importantly, when a new block is added, the training undergoes a 12,800 update *fading period*. During this fading period, the generated samples are a linear interpolation of the higher resolution but less trained output of the new block and the lower resolution but better trained output of the previous block. An equivalent process happens for the critic. The interpolation process is illustrated in Fig. 5. The interpolation coefficient, α, linearly transitions from 0 (start of fading period) to 1 (end of fading period). After this fading period, α is effectively clamped to 1, and the block continues to train for up to 64,000 additional updates, until convergence. The fading period reduces the risk of the new block's randomly initialized weights disrupting training. Progressive training means that the early training is limited to computational cheap blocks, accelerating the training process.

B Hyperparameter Tuning

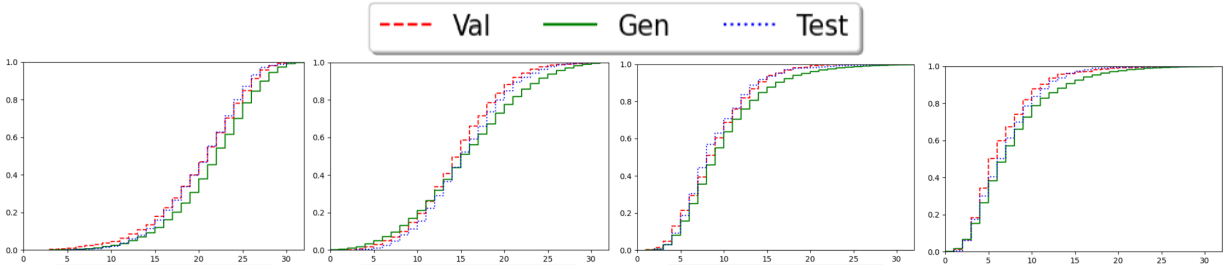
We train each block with early stopping based upon the spatiotemporal KL divergence with the validation set. We utilize a batch size of 32 and the AdamW optimizer. Our generator and critic learning rate is 8e-5, our grad penalty is 20.0, and for each iteration our critic is updated three times and the generator is updated one time. These parameters were found over a period of trial-and-error and hyperparameter searches utilizing Weights and Biases [47] sweeps, which implements Bayesian Optimization [48]. For the last block, we found that increasing the grad penalty to 40.0 greatly improved training stability.



(a) Fading of the generator model. The generator produces its low resolution sample which is spatially upscaled and weighed with the output of the new block

(b) Fading of the critic model. The critic slowly begins using the features from the new block.

Figure 5: Fading in additional layers during progressive training. $\alpha \to 1$ during the fading period.



(a) Denver (b) London (c) Denver (d) London

Figure 6: Empirical cumulative distribution functions for the validation, generated and test data at two locations. Number of dry days per sample in (a)-(b); length of longest dry spell in (c)-(d).

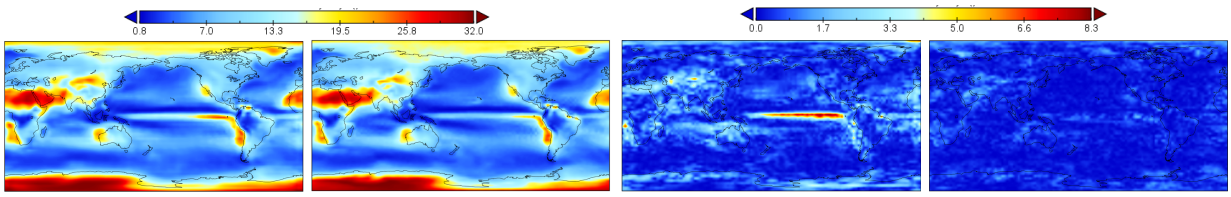
C Empirical Cumulative Distributions

Continuing our analysis on performance, an additional metric compares the empirical distributions over points of interest in the test, validation and generated data. We choose two locations roughly corresponding to the coordinates of Denver, USA and London, UK. For the corresponding grid-points we extract time series and compute the total number of dry days (defined as before to be days with $< 1mm$ precipitation) in each $T = 32$ day sample. Figs. 6a (Denver) and 6b (London) plot the empirical cumulative distribution functions (eCDFs) of the number of dry days computed over $N = 4096$ samples each from the generator, test and validation sets. Figures 6c and 6d replicate this setup, except modeling the distribution over the longest dry spell in each sample.

From the results we can assess a slight tendency of the generated samples to produce more dry days and longer dry spells than the samples from the climate model (both test and validation), but the difference between the eCDFs remains limited to very few days in all instances, and the shape of the eCDFs qualitatively similar.

D Longest Dry Spell Analysis

Figure 7 plots maps of the average length of the longest dry spell from the Fall-Winter data, analogous to results presented in Fig. 4. The map of generated data averages is plotted in panel (a), whereas the map of test set averages is plotted in panel (b). Similar trends continue in panel (c), where the equatorial Pacific west of South America still provides the generator some difficulty.



(a) (b) (c) (d)

Figure 7: Maps with mean longest dry spell length in the (a) generated and (b) test Fall-Winter data; (c) the absolute difference between a and b; (d) the absolute difference between test and validation.

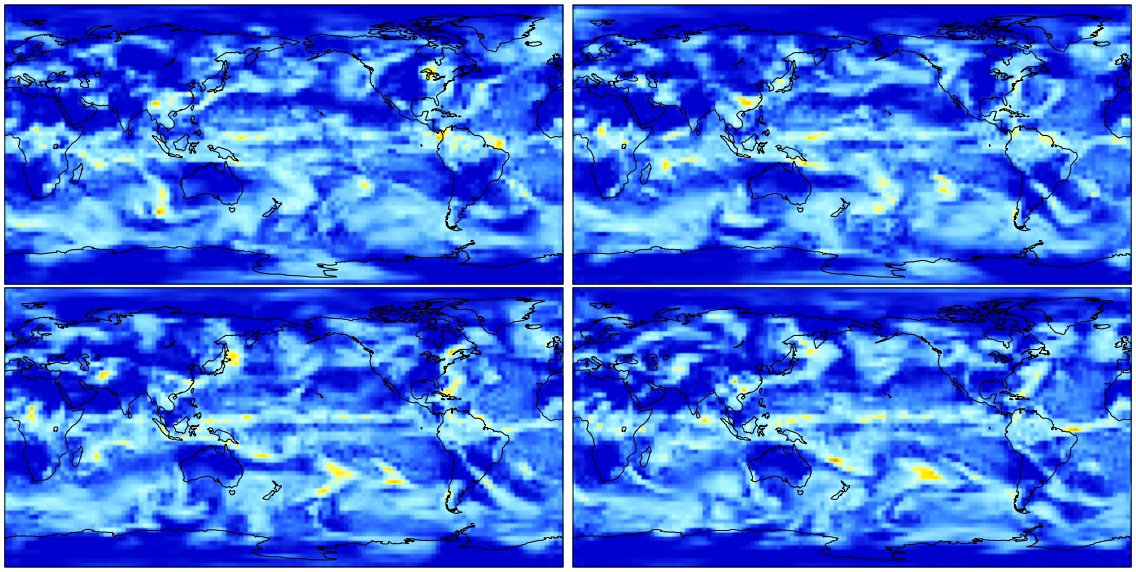


Figure 8: Precipitation maps for four successive generated days in spring and summer

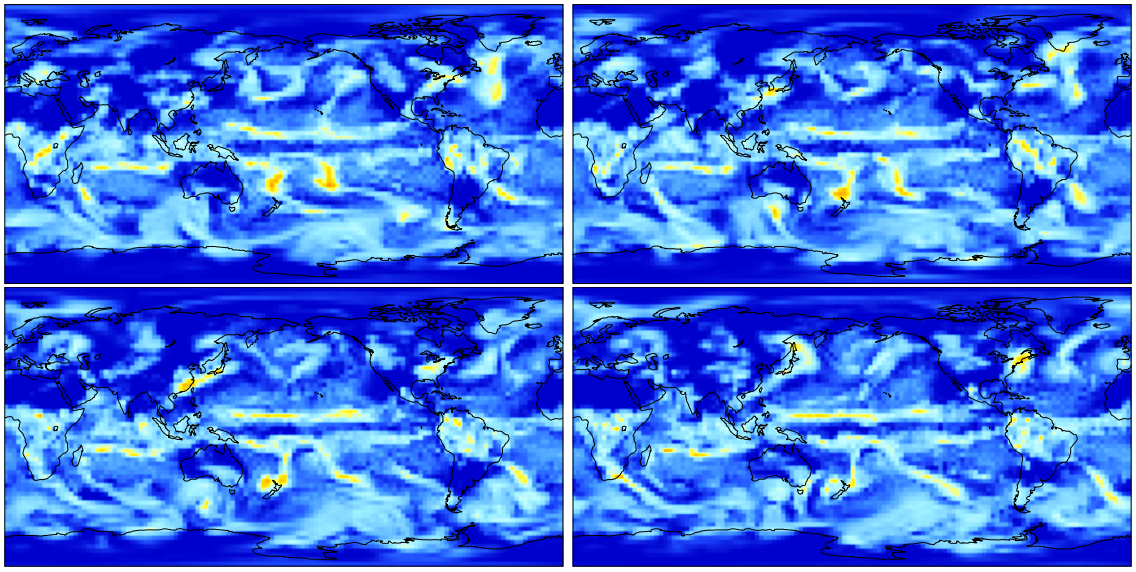


Figure 9: Precipitation maps for four successive test set days in spring and summer

E Additional Samples

Figures 8, 10, 9, and 11 are visualizations of log-normalized four-day sequences of the generated and test set precipitation data. Each sequence proceeds left-to-right, then top-to-bottom, and was randomly selected from a 32-day sample.

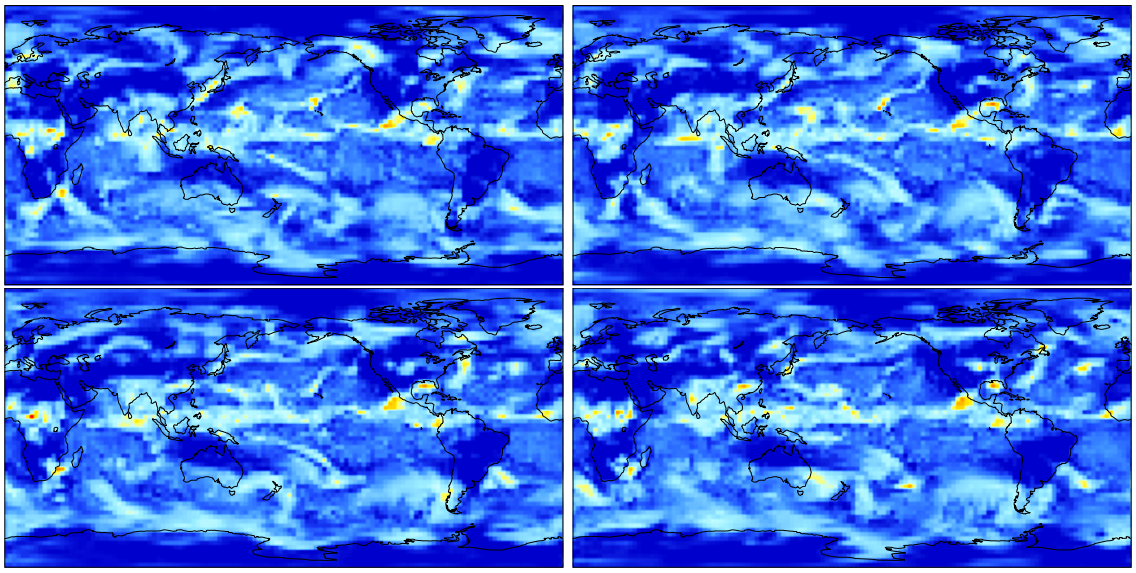


Figure 10: Precipitation maps for four successive generated days in fall and winter

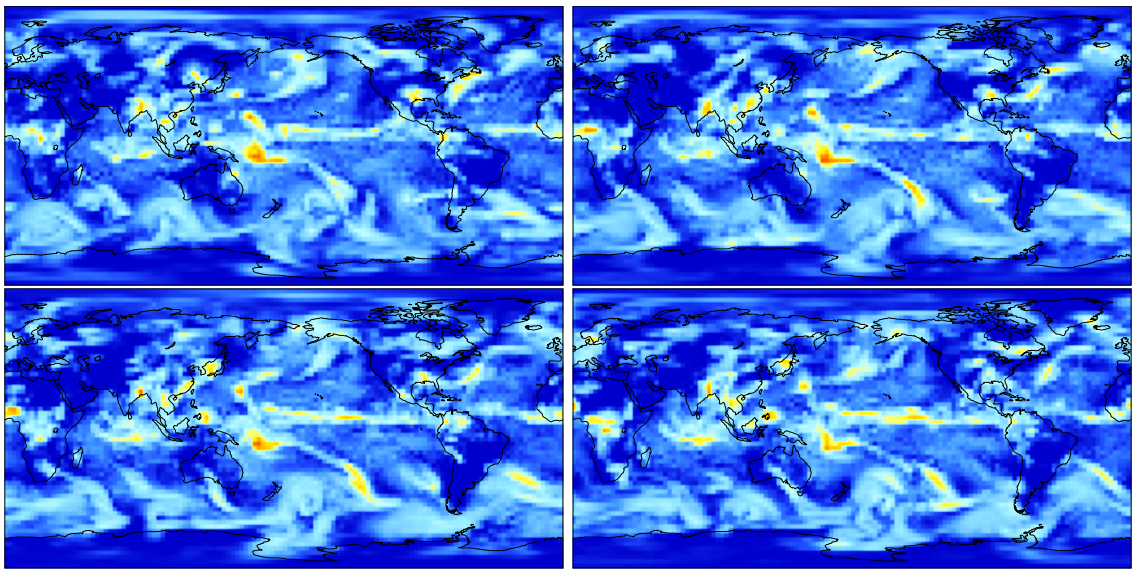


Figure 11: Precipitation maps for four successive test set days in fall and winter