Global Precipitation Correction Across a Range of Climates Using CycleGAN

Jeremy J McGibbon[1], Spencer Koncius Clark[2], Brian Henn[3], Anna Kwa[1], Oliver Watt-Meyer[1], W. Andre Perkins[1], and Christopher S. Bretherton[1]

[1]Allen Institute for Artificial Intelligence
[2]Allen Institute for Artificial Intelligence / NOAA-GFDL
[3]Allen Institute for Artificial Intelligence (AI2)

July 8, 2023

Abstract

Accurate precipitation simulations for various climate scenarios are critical for understanding and predicting the impacts of climate change. This study employs a Cycle-generative adversarial network (CycleGAN) to improve global 3-hour-average precipitation fields predicted by a coarse grid (200˜km) atmospheric model across a range of climates, morphing them to match their statistical properties with reference fine-grid (25˜km) simulations. We evaluate its performance on both the target climates and an independent ramped-SST simulation. The translated precipitation fields remove most of the biases simulated by the coarse-grid model in the mean precipitation climatology, the cumulative distribution function of 3-hourly precipitation, and the diurnal cycle of precipitation over land. These results highlight the potential of CycleGAN as a powerful tool for bias correction in climate change simulations, paving the way for more reliable predictions of precipitation patterns across a wide range of climates.

Global Precipitation Correction Across a Range of Climates Using CycleGAN

J. McGibbon[1*], S. K. Clark[1,2†], B. Henn[1‡], A. Kwa[1§], O. Watt-Meyer[1¶], W. A. Perkins[1‖], C. S. Bretherton [1]**

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[2]Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Key Points:

- A Cycle-generative adversarial network (CycleGAN) can debias precipitation across a range of climate forcings
- The model is able to debias data from intermediate forcings not present in training data
- The model is able to correct tails of the precipitation distribution without the use of quantile mapping

---

*Contribution: Conceptualization, Investigation, Methodology, Formal analysis, Software, Visualization, Writing - original draft, Writing - review and editing

†Contribution: Data curation, Methodology, Software, Writing - original draft, Writing - review and editing

‡Contribution: Software, Writing - review and editing

§Contribution: Software, Writing - review and editing

¶Contribution: Software, Writing - review and editing

‖Contribution: Software, Resources, Writing - review and editing

**Contribution: Project administration, Supervision, Conceptualization, Formal Analysis, Writing - original draft, Writing - review and editing

Corresponding author: Jeremy McGibbon, `mcgibbon@uw.edu`

Abstract

Accurate precipitation simulations for various climate scenarios are critical for understanding and predicting the impacts of climate change. This study employs a Cycle-generative adversarial network (CycleGAN) to improve global 3-hour-average precipitation fields predicted by a coarse grid (200 km) atmospheric model across a range of climates, morphing them to match their statistical properties with reference fine-grid (25 km) simulations. We evaluate its performance on both the target climates and an independent ramped-SST simulation. The translated precipitation fields remove most of the biases simulated by the coarse-grid model in the mean precipitation climatology, the cumulative distribution function of 3-hourly precipitation, and the diurnal cycle of precipitation over land. These results highlight the potential of CycleGAN as a powerful tool for bias correction in climate change simulations, paving the way for more reliable predictions of precipitation patterns across a wide range of climates.

Plain Language Summary

Using CycleGAN, a machine learning technique, we can remove key biases in precipitation simulated by a fast, coarse-grid atmospheric model. This method morphs maps of the output precipitation to match typical characteristics of a slower but more accurate fine-grid configuration, correcting systematic errors in both long-term average spatial precipitation patterns and 3-hourly precipitation variations. It retains skill in intermediate climate states unseen in training, making it a useful tool for climate change simulations.

1 Introduction

Throughout the history of atmospheric model development, results from fine-grid models that resolve important physical processes like cloud and precipitation formation or flow over mountain ranges have been used to improve biased climates in coarse-grid models that do not. For instance, scientists have used large-eddy simulations as a testbed for calibrating analytic turbulence and cloud parameterizations, e.g. Bogenschutz et al. (2010). This process relies heavily on expert knowledge to develop appropriate models of sub-grid behaviors, often heavily influenced by analysis of a few archetypical cases.

More recently, machine learning (ML) has been used to correct coarse-resolution models' behavior across the full range of conditions over historical periods where observational analysis is available. For example, Watt-Meyer et al. (2021) trained a corrective tendency for a 200 km grid atmospheric general circulation model (AGCM) using ML based on nudging tendencies towards observational analysis. The ML correction reduced annual-mean precipitation biases by 20%. An ML approach based on reservoir computing (Arcomano et al., 2023) and ERA5 reanalysis (Hersbach et al., 2020) halved the global root mean square bias of annual-mean precipitation in an even coarser (400 km grid) AGCM.

We require a different strategy when training a model to generalize across future climate forcings, as when training with observational analyses one can only learn the climate represented by this data. One method is to use finer-grid AGCM simulations as training targets. Such simulations are computationally expensive, but they more accurately simulate societally-important aspects of present-day climate such as means and extremes of land surface precipitation and temperature than do coarse-grid AGCMs (Flato et al., 2013; Wehner et al., 2010). Because they resolve much more detail of deep convective storm systems, orography and land surface characteristics, they are less sensitive to uncertain parameterizations of deep convection and orographic drag, making them potentially a more robust simulation tool for generalizing to future climates. S. K. Clark et al. (2022) used the same nudging approach as Watt-Meyer et al. (2021) to ML-correct a 200 km model to behave like its 25 km analogue across four climates forced by adding specified uniform sea-surface temperature (SST) increments to observed SST patterns. They were able to correct spatial patterns of precipitation over land by 10-30% in multi-year simulations across all four climates.

This bias reduction is encouraging, but to get full advantage from high-fidelity reference data, corrective ML should enable both the weather and climate to have much reduced bias (much less than 50% of a no-ML baseline simulation) vs. this reference, both for means and extremes of salient quantities such as precipitation. Yet fundamental challenges persist. This type of hybrid ML, which couples a bias correction model trained offline with a pre-existing AGCM, can induce online simulation biases due to feedbacks between these two components (Brenowitz et al., 2020). The machine learning goal of minimizing prediction error for each sample can lead to difficulties in accurately representing small-scale stochastic behaviors such as deep convection, leading e.g. to an inaccurate representation of the frequency of extreme precipitation (Kwa et al., 2023).

To further reduce bias in the simulated space-time distribution of precipitation vs. a reference climatology, we turn to a different form of ML, the Cycle-generative adversarial network or CycleGAN (Zhu et al., 2017), which is a promising tool for translation of image data between two unpaired domains. Unlike the above hybrid ML approaches, this is a post-processing approach which cannot easily be analyzed in terms of physical process errors in the coarse-grid model, and only corrects selected model fields (precipitation, in our case). In the past, cycle-generative networks have been effectively used for offline bias correction, but it has been necessary to augment the cycle-generative network with quantile mapping to achieve accurate probability distributions of precipitation (François et al., 2021; Pan et al., 2021; Fulton et al., 2023). These works focused on translating one or more model output variables towards an observational analysis over a historical period for a subset of the globe, the annual cycle, or both, and each corrected daily-mean precipitation.

Our work expands on these efforts. We use the original CycleGAN architecture of Zhu et al. (2017) to correct the output of the FV3GFS atmospheric model with a C48 cubed-sphere grid (with approximately 200 km horizontal spacing) to behave like the coarsened output of the same model run on a C384 cubed-sphere grid (25 km spacing). We demonstrate the ability to improve both the spatial distribution of annual-mean precipitation and the cumulative distribution function (CDF) of 3-hourly precipitation up to the 99.999th percentile across a range of climate forcings, without the need for quantile mapping. This method is capable of correcting data at intermediate climate forcings not used during model training, enabling its application to climate change simulations.

2 Dataset

We generate all training data using the FV3GFS atmospheric model (Putman & Lin, 2007; Harris & Lin, 2013; Zhou et al., 2019) as described in McGibbon et al. (2021), run on a cubed-sphere grid with 63 vertical levels. Annually-repeating cycles of sea surface temperature (SST) and sea ice are defined based on the observational monthly means time-averaged from 1982 to 2012 from the 1/12° Real Time Global Sea Surface Temperature (Thiébaux et al., 2003) and 0.5° Climate Forecast System Reanalysis (Saha et al., 2014) datasets, respectively. We perturb the SSTs by adding globally-constant offsets of -2 K, 0 K, +2 K, and +4 K to produce four different sets of forcings while maintaining the present-day annual cycle of sea ice and carbon dioxide concentration, analogous to S. K. Clark et al. (2022). We train using simulations with spacing between SST offsets of 2 K out of concern that precipitation may be too different between forcings at the larger 4 K spacing used in S. K. Clark et al. (2022) for the trained model to accurately generalize to intermediate forcings, though this has not been tested.

For each of these SST forcings, a simulation was performed at C48 resolution for 9 years, 1 month. Eight 1 year, 1 month simulations are performed at C384 resolution beginning with the C48 model snapshot state 1 year into the C48 run as well as the state every year thereafter; the C48 snapshots were converted to C384 initial conditions using the `chgres_cube` tool of `UFS_UTILS` (Gayno et al., 2020). For each of these C384 simulations, the first month of simulation time is discarded as a model spin-up period. This yields 8 years of useful simulation data from each climate, from which we take the first 5 years as training and the last 3 years as validation data.

During these simulations we accumulate and store the 3-hourly mean precipitation rate. We use 3-hourly precipitation instead of daily mean to test the ability of the model to correct biases in the diurnal cycle. At each output time, the C384 precipitation fields are coarsened to the C48 grid by horizontal averaging so that they can be directly compared with coarse-grid precipitation fields.

We also perform "ramping" simulations at both C48 and C384 resolution, which begin with a present-day initial condition and three month spin-up period with 0K forcing, and then enter a period where the forcing is linearly increased from 0K to +2K over the course of 4 years. This data is withheld during training and hyperparameter tuning, and is used for model evaluation only. It tests whether the CycleGAN can skillfully interpolate mean and extreme precipitation patterns between climates on which it was trained.

3 Model formulation and training

The model architecture in Zhu et al. (2017) is used with minimal modifications to allow processing of cubed-sphere data. Specifically, convolution is performed using halo updates on the cubed sphere, where missing corners are filled with zero values. This is numerically identical to the convolution approach used in Weyn et al. (2020), except that data in the corner of each tile domain is set to zero rather than copying and rotating data from the polar tile face. We do not find evidence of corner imprinting despite this choice.

The performance of this model is improved by concatenating spatiotemporal geometric features to the input of the generator and discriminator models. These features are the x, y, and z positions of each grid cell on a stationary unit sphere in 3-dimensional Euclidean space (spatial features), as well as the x and y positions of each grid cell on a unit sphere in Euclidean space as it rotates with a period of one rotation per day (time features). These time features can also be thought of as the x and y positions of an hour hand on a 24-hour clock indicating the local time, multiplied by cos(latitude) to avoid discontinuity at the poles. These are used only as inputs of these models, and are not output by the generators. The discriminator is given identical geometric features to the generator which produced the image being evaluated.

The training dataset includes 58400 3-hourly global snapshots, split evenly across the four climate forcings. Each epoch, we randomly sample 40000 snapshots with replacement, training with a batch size of 1. This data only contains two-dimensional cubed-sphere surface precipitation rate along with a UTC time, which is used to generate the spatiotemporal geometric features.

Notably, the climate forcing itself is absent from the training data, as we were able to achieve excellent bias correction without it. When we included the SST perturbation as input context, as was done for diurnal features, several performance metrics worsened without any clear improvements (compare red vs. steel-blue colors in Figures S2 and S3).

The model was trained with an exponential learning rate decay. Starting with a high learning rate and eventually reducing it further in training is a widely used technique in machine learning (Li et al., 2019). The best results shown here were achieved with an initial learning rate of 10^{-4} and a decay factor of 0.63 (a tenfold decrease every 5 epochs). Training converged (in terms of our precipitation bias metrics) by epoch 14 and was run for 16 epochs (Figure S1).

Otherwise, the hyperparameters are the same as in the 6-layer network of Zhu et al. (2017), but with twice as many filters in the generator and discriminator. We did not attempt to tune the number of layers, activation functions, or choice of optimizer, and we found that increasing the number of filters beyond the value used used did not improve the model.

4 Results

Figure 1 shows the behavior of the generative model on a single sample, taken from the ramping simulation in a climate state distinct from any in the CycleGAN training dataset. We are most concerned with the translation of C48 data into C384 (ML) data (upper left vs. upper right panels), but it is also illuminating to see the inverse generation from C384 to C48 (ML) (lower left vs. upper right panels). The model introduces finer scale features when translating into the C384 domain, especially in lightly precipitating marine boundary layer cloud regimes. It strengthens precipitation over land, introducing precipitation into areas which have none in the C48 input, for example over the South American continent.

The translation substantially improves the mean precipitation climatology vs. C48 simulations for all four SST offsets, as shown in Figure 2, with metrics reported in Figure 3. Here and throughout this analysis, time-mean statistics for fixed-SST simulations such as for the 0K climate are computed on the 3 years of validation data. Statistics for the ramping climate are computed on years 2 and 3 of the 4-year simulation linearly ramping from 0K to plus-2K forcings, to highlight the range of SST offsets that are further from the fixed-SST training data and hence provide a more rigorous out-of-sample test. The bias reductions seen in the ramping simulation are comparable to those in the target climates; the bias of mean precipitation averaged over all land is reduced over 85%, and the standard deviation of the geographic pattern of time-mean bias is reduced over 75% to values around 0.5 mm/d. A significant portion of the biases in each target climate is explained by differences in precipitation between the validation and training datasets, as shown by the "train" bars. Thus, we anticipate further bias reduction with larger training and validation datasets.

Both the 0 K and ramping simulations have smaller precipitation pattern biases than reported for a current-climate case by Arcomano et al. (2023). They reported that their hybrid reservoir computing ML reduced the standard deviation of precipitation pattern bias nearly 50% from 1.2 mm/d in their no-ML baseline model to a value of 0.63 mm/d. Our mean precipitation biases also much smaller than the bias shown in Figure 1 of Fulton et al. (2023) for the South Asian monsoon region. A direct comparison with François et al. (2021) and Pan et al. (2021) is difficult because they considered France and the continental United States, respectively, both of which have much smaller biases than the rest of the globe in our model.

Figure 4 shows that the translated data has a 3-hourly probability distribution and a diurnal cycle of land precipitation that much more closely match the C384 reference data across all climates, including the ramping simulation. We might expect the CycleGAN to struggle to represent extreme precipitation events and their sensitivity to climate forcing because they appear infrequently in the training data. Nevertheless, the 58400 precipitation fields, each containing 13824 atmospheric column, comprise almost 10^9 atmospheric columns, perhaps enough to learn how to translate even highly unusual precipitation events. Indeed, the CycleGAN improves the accuracy of the CDF of precipitation up to the 99.999th percentile. Only at the 99.9999th percentile and only for

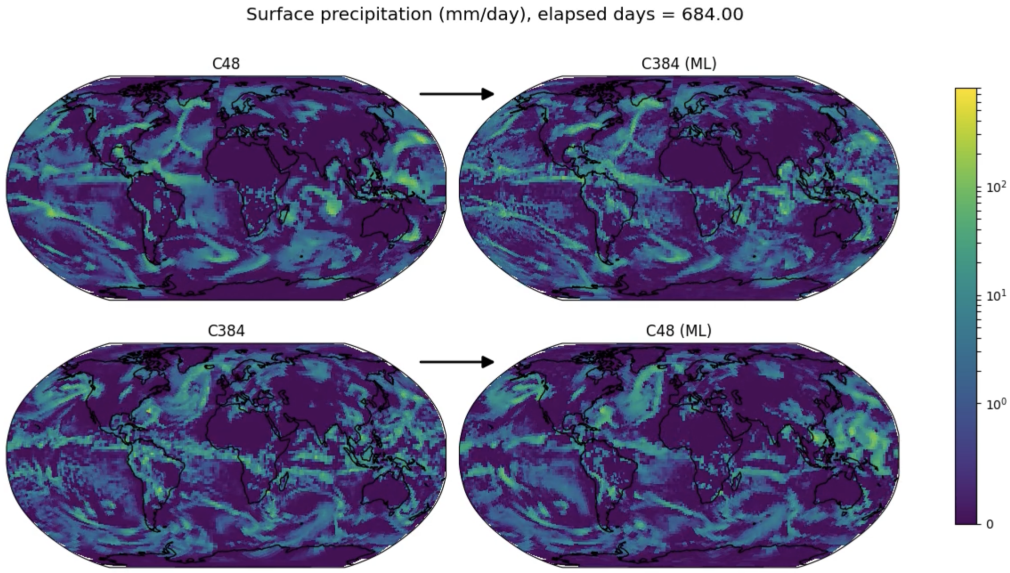

Figure 1. Inputs and outputs of the CycleGAN for one timestep during the ramping simulation. Precipitation data on the left was used as input to generate precipitation data on the right. Snapshot was selected to illustrate a common feature, significantly stronger precipitation over South America in C384 (replicated by the GAN) than in C48. All snapshots for this simulation can be viewed in the supplementary data (Movie S1).

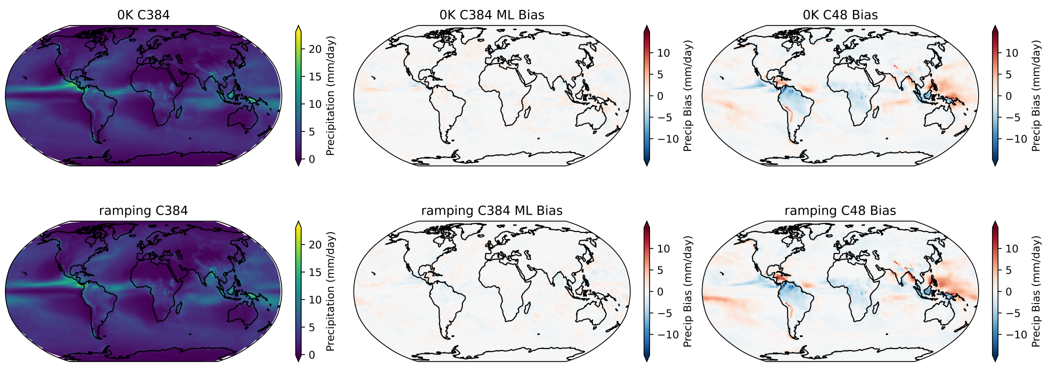

Figure 2. Annual-mean precipitation from C384 reference run (left column) and precipitation biases from the C48 simulation (right column) and from the GAN applied to this C48 simulation (C384 ML). Bias values are differences from the C384 reference.

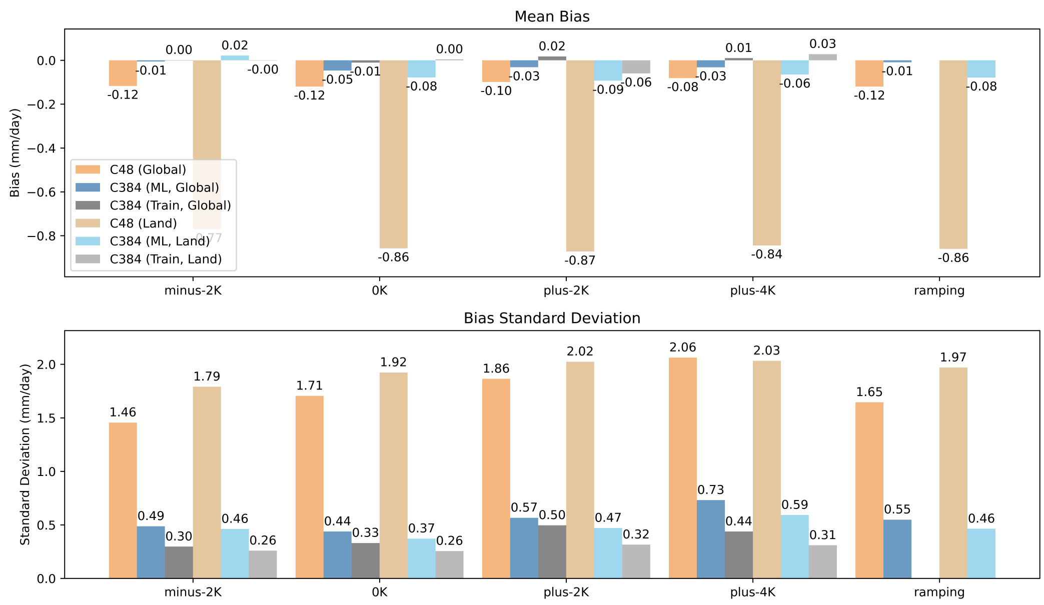

Figure 3. Metrics of time-average precipitation bias against validation and testing data. Mean bias refers to the area-weighted horizontal mean bias across all samples, or over land samples only. Bias standard deviation refers to the square root of the area-weighted mean square bias, averaged over the horizontal either globally or over land samples only. These statistics are derived from bias maps as shown in Figure 2. "Train" indicates the comparison of the training data itself against the validation data. Training data is not available for the ramping simulation.

the -2 K forcing, the CycleGAN slightly increases the error over the input C48 reference data. Surprisingly, the distribution of ML outputs is, if anything, over-dispersive in the tails. The shape of the diurnal cycle of precipitation over land is also improved across all climate forcings, with a stronger trough and sharper increase in precipitation from 6:00 to 15:00 local solar time, and more sustained precipitation through the 21:00-24:00 bin.

Here, the diurnal cycle over land was computed by determining the local solar time in each land-based grid cell for each sample based on its longitude, and then binning the data across local time before taking an area-weighted mean.

5 Sensitivity Studies

This section describes sensitivity studies that help motivate some of our model design choices. We initially trained the CycleGAN model with less data, but found the global maps of time-averaged precipitation vary significantly from year to year, resulting in significant biases in the trained model as a result of under-sampling the long-term climate. When we train the model using only the first year of data from each climate and evaluate on the same 3 years of validation data, the model has significantly worse time-mean biases (Figure S2, compare light purple bar to darker blue bar), and does a significantly worse job predicting the output CDFs, over-predicting the extremes of each climate's precipitation distribution (Figure S3).

Adding spatiotemporal features defining the diurnal cycle as context to the input of the generator and discriminator was crucial for correcting the shape of the mean di-

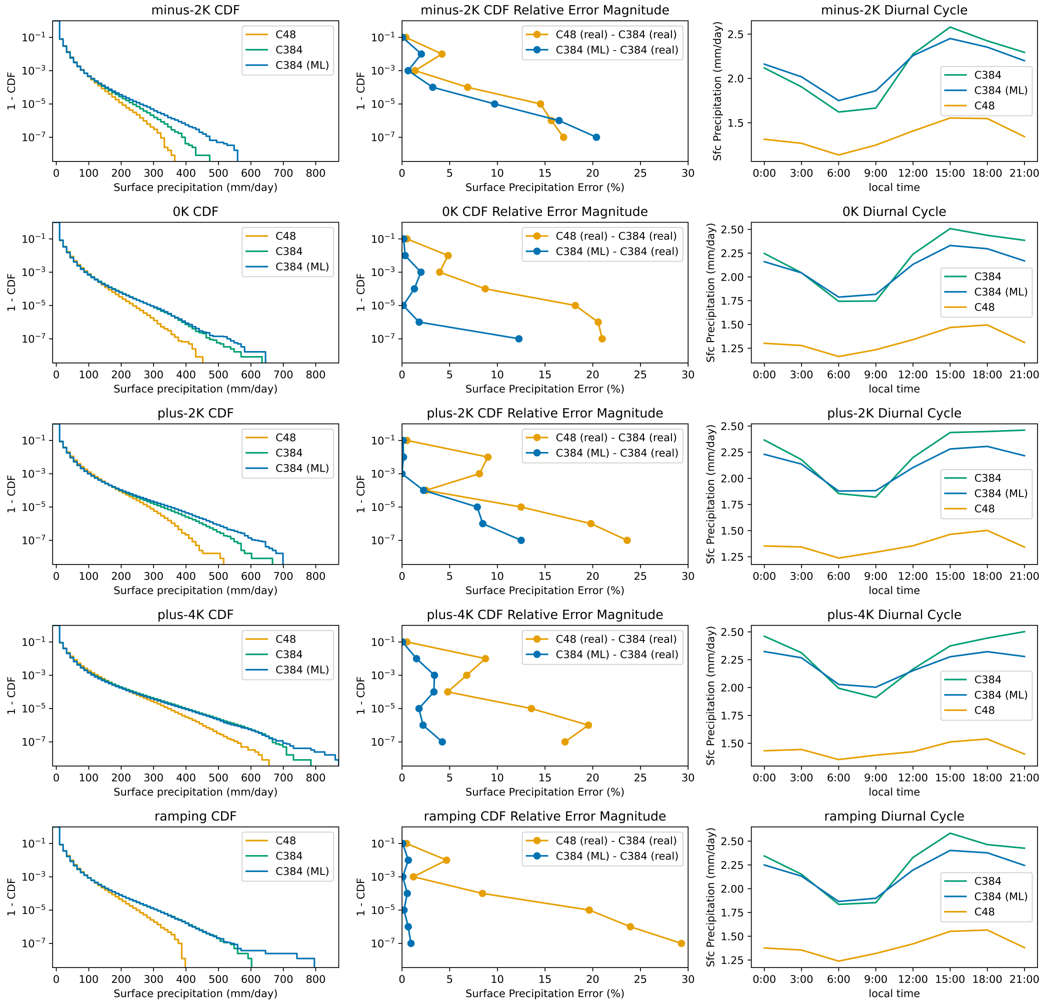

Figure 4. CDF metrics and diurnal cycle of precipitation over land for the reference C384 run, the C48 simulation, and the CycleGAN applied to the C48 output (C384 ML). The left column shows the CDF of precipitation for each climate. The center column shows the relative magnitude of errors of the values of the C48 and CycleGAN CDFs in the first column across a range of percentiles, computed as a percentage of the C384 (real) value. The right column shows the diurnal cycle of precipitation over land, with the x-axis indicating the starting local time of the 3-hour bin.

urnal cycle of precipitation over land. Without these features, the land diurnal cycle of the C384 (ML) output data is significantly improved because we have corrected the correct mean and variance, though the shape of the cycle (light green and light blue) is more similar to the C48 values (orange). Surprisingly, the inclusion of these features has little impact on the standard deviation of the geographically-resolved time-mean bias (Figure S2).

In Zhu et al. (2017), an identity loss was included for certain translation tasks to avoid unnecessary modification of the color scheme during translation. We find removing this identity loss generally degrades model performance. It leads to increased pattern bias and land-mean bias in precipitation (Figure S2) and has a neutral effect on the CDF and land diurnal cycle (Figure S3).

6 Discussion

Unlike previous works using cycle-generative architectures to bias-correct precipitation (François et al., 2021; Pan et al., 2021; Fulton et al., 2023), we could match the PDF of 3-hourly precipitation without quantile mapping. Pan et al. (2021) claimed that quantile mapping is needed because “GANs are trained to produce individual trust-worthy samples, not accurate probability distribution estimations”, due e.g. to mode collapse (Bau et al., 2019), despite the claim in Goodfellow et al. (2014) that their training Algorithm 1 is designed to “converge to a good estimator of [the probability distribution of the data], if given enough capacity and training time”.

Many methodological differences might explain why we were able to better simulate the probability of extreme precipitation events without quantile mapping. We correct only precipitation, without using other model output fields as dynamical constraints (Pan et al., 2021) or additional fields to be corrected (François et al., 2021; Fulton et al., 2023). Fully sampling the variability and covariability within more fields requires significantly more data, owing to the curse of dimensionality. In addition, our model is trained on more data than the previous studies. We used 58,400 timesteps each with 13,824 gridcells, resulting in 807M precipitation samples, while François et al. (2021); Pan et al. (2021) and Fulton et al. (2023) used 7.42M, 247M, and 40.6M samples respectively. The character of the corrections is different, in particular because of the use of 3-hourly versus daily data and the use of global data instead of limited regions. Our training methodology also differs in the introduction of a learning rate schedule, which could play a role, and Fulton et al. (2023) used the UNIT architecture (Liu et al., 2017) as opposed to CycleGAN.

While this CycleGAN significantly improves the climate of individual samples from a spun-up C48 model state, it should not be used to correct weather simulations run at C48 which are initialized from a coarsened C384 state. We trained the CycleGAN only on samples which are far into a C48 simulation, whose climate contains more significant biases than a hypothetical dataset containing samples from the first week of a C48 simulation initialized from coarsened C384 data. One could remove this input bias effect by training a CycleGAN model to correct model biases at one particular forecast lead time, and using coarse and fine-grid examples at that particular lead time. One could also train a conditional CycleGAN with forecast lead time as a model input capable of correcting a variety of lead times, similar to what was done in this work for time-of-day.

7 Conclusions

We found that CycleGAN with little modification can accurately translate 3-hourly precipitation simulated by a 200 km grid global atmospheric model across a range of climate forcing to have similar statistics as output from a reference fine-grid 25 km model, as measured by its time-mean geographically-resolved pattern, its CDF and its mean diurnal cycle over land. These biases are much reduced compared to previous online correction approaches, but because CycleGAN is a post-processing approach, this comes at the expense of interpretability. The CycleGAN generalizes well to a ramped-SST simulation with intermediate forcings not present in the training dataset. With a small set of expensive fine-grid simulations, the CycleGAN can thus quickly debias precipitation fields predicted by a fast coarse-grid model across a broad range of climates.

8 Open Research

The code used to train and evaluate the machine learning models and produce the figures in this study is available on Zenodo via https://doi.org/10.5281/zenodo.8070950 with MIT and BSD licenses (Brenowitz et al., 2023). The coarse-resolution and coarsened high-resolution model output used for training, validation, and testing are available on Zenodo via https://doi.org/10.5281/zenodo.8070973 with a Creative Commons Attribution 4.0 International License (S. Clark et al., 2023). Figures were made with Matplotlib version 3.7.1 (Caswell et al., 2023), available under the matplotlib license at https://matplotlib.org/. Our machine learning code uses Pytorch version 1.12.1 (Paszke et al., 2019), available under a BSD-3 license at https://pytorch.org/.

Acknowledgments

We thank the Allen Institute for Artificial Intelligence for supporting this work and NOAA-GFDL for providing access to the Gaea computing system used to produce the data on which our model is trained and evaluated. We thank Jun-Yan Zhu for making publicly available a high-quality PyTorch implementation of the CycleGAN model. We used GPT-4 to help create a first draft of the Abstract and Plain Language Summary of this manuscript, which were then significantly revised to correct details.

References

Arcomano, T., Szunyogh, I., Wikner, A., Hunt, B. R., & Ott, E. (2023). A hybrid atmospheric model incorporating machine learning can capture dynamical processes not captured by its physics-based component. *Geophysical Research Letters*, *50*(8), e2022GL102649. doi: https://doi.org/10.1029/2022GL102649

Bau, D., Zhu, J., Wulff, J., Peebles, W. S., Strobelt, H., Zhou, B., & Torralba, A. (2019). Seeing what a GAN cannot generate. *CoRR*, *abs/1910.11626*. doi: https://doi.org/10.48550/arXiv.1910.11626

Bogenschutz, P. A., Krueger, S. K., & Khairoutdinov, M. (2010). Assumed probability density functions for shallow and deep convection. *Journal of Advances in Modeling Earth Systems*, *2*(4). doi: https://doi.org/10.3894/JAMES.2010.2.10

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., ... Bretherton, C. S. (2020). *Machine learning climate model dynamics: Offline versus online performance.* doi: https://doi.org/10.48550/arXiv.2011.03081

Brenowitz, N. D., Kwa, A., Perkins, W. A., Henn, B., McGibbon, J., Clark, S., ... Nabizadeh, E. (2023, June). *ai2cm/fv3net: Commit used for CycleGAN initial submission.* Zenodo [Software]. doi: https://doi.org/10.5281/zenodo.8070950

Caswell, T. A., Lee, A., de Andrade, E. S., Droettboom, M., Hoffmann, T., Klymak, J., ... Moad, C. (2023, March). *matplotlib/matplotlib: Rel: v3.7.1.* Zenodo [Software]. doi: https://doi.org/10.5281/zenodo.7697899

Clark, S., McGibbon, J., & Bretherton, C. (2023, June). *Data for "Global Precipitation Correction Across a Range of Climates Using CycleGAN".* Zenodo [Dataset]. doi: https://doi.org/10.5281/zenodo.8070973

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A.,

... Harris, L. M. (2022). Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003219. doi: https://doi.org/10.1029/2022MS003219

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., ... Rummukainen, M. (2013). Evaluation of climate models. In T. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (p. 741–866). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: https://doi.org/10.1017/CBO9781107415324.020

François, B., Thao, S., & Vrac, M. (2021, July). Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, *57*(11-12), 3323–3353. doi: https://doi.org/10.1007/s00382-021-05869-8

Fulton, D. J., Clarke, B. J., & Hegerl, G. C. (2023). Bias correcting climate model simulations using unpaired image-to-image translation networks. *Artificial Intelligence for the Earth Systems*, 1 - 56. doi: https://doi.org/10.1175/AIES-D-22-0031.1

Gayno, G., Beck, J., Reames, L., programmer, G., Wright, D., Hu, M., ... Gerheiser, K. (2020, August). *UFS_UTILS.* Unified Forecast System (UFS).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.* doi: https://doi.org/10.48550/arXiv.1406.2661

Harris, L. M., & Lin, S.-J. (2013). A two-way nested global-regional dynamical core on the cubed-sphere grid. *Monthly Weather Review*, *141*(1), 283 - 306. doi: https://doi.org/10.1175/MWR-D-11-00201.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. doi: https://doi.org/10.1002/qj.3803

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., ... Bretherton, C. S. (2023). Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003400. doi: https://doi.org/10.1029/2022MS003400

Li, Y., Wei, C., & Ma, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. *CoRR*, *abs/1907.04595*. doi: https://doi.org/10.48550/arXiv.1907.04595

Liu, M., Breuel, T. M., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *CoRR*, *abs/1703.00848*. doi: https://doi.org/10.48550/arXiv.1703.00848

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S., Davis, E. C., ... Fuhrer, O. (2021). fv3gfs-wrapper: a python wrapper of the fv3gfs atmospheric model. *Geoscientific Model Development*, *14*(7), 4401–4409. doi: https://doi.org/10.5194/gmd-14-4401-2021

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., Lee, J., ... Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509. doi: https://doi.org/10.1029/2021MS002509

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, *abs/1912.01703*. doi: https://doi.org/10.48550/arXiv.1912.01703

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55 - 78. doi: https://doi.org/10.1016/j.jcp.2007.07.022

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., ... Becker, E. (2014, March). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208. doi: https://doi.org/10.1175/JCLI-D-12-00823.1

Thiébaux, J., Rogers, E., Wang, W., & Katz, B. (2003, May). A New High-Resolution Blended Real-Time Global Sea Surface Temperature Analysis. *Bulletin of the American Meteorological Society*, *84*(5), 645–656. doi: https://doi.org/10.1175/BAMS-84-5-645

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: https://doi.org/10.1029/2021GL092555

Wehner, M. F., Smith, R. L., Bala, G., & Duffy, P. (2010). The effect of horizontal resolution on simulation of very extreme us precipitation events in a global atmosphere model. *Climate Dynamics*, *34*(2), 241–247. doi: https://doi.org/10.1007/s00382-009-0656-y

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109. doi: https://doi.org/10.1029/2020MS002109

Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., & Rees, S. L. (2019). Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, *100*(7), 1225 - 1243. doi: https://doi.org/10.1175/BAMS-D-17-0246.1

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, *abs/1703.10593*. doi: https://doi.org/10.48550/arXiv.1703.10593

Global Precipitation Correction Across a Range of Climates Using CycleGAN

J. McGibbon[1*], S. K. Clark[1,2†], B. Henn[1‡], A. Kwa[1§], O. Watt-Meyer[1¶], W. A. Perkins[1‖], C. S. Bretherton [1]**

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[2]Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Key Points:

- A Cycle-generative adversarial network (CycleGAN) can debias precipitation across a range of climate forcings
- The model is able to debias data from intermediate forcings not present in training data
- The model is able to correct tails of the precipitation distribution without the use of quantile mapping

[*]Contribution: Conceptualization, Investigation, Methodology, Formal analysis, Software, Visualization, Writing - original draft, Writing - review and editing
[†]Contribution: Data curation, Methodology, Software, Writing - original draft, Writing - review and editing
[‡]Contribution: Software, Writing - review and editing
[§]Contribution: Software, Writing - review and editing
[¶]Contribution: Software, Writing - review and editing
[‖]Contribution: Software, Resources, Writing - review and editing
[**]Contribution: Project administration, Supervision, Conceptualization, Formal Analysis, Writing - original draft, Writing - review and editing

Corresponding author: Jeremy McGibbon, `mcgibbon@uw.edu`

Abstract

Accurate precipitation simulations for various climate scenarios are critical for understanding and predicting the impacts of climate change. This study employs a Cycle-generative adversarial network (CycleGAN) to improve global 3-hour-average precipitation fields predicted by a coarse grid (200 km) atmospheric model across a range of climates, morphing them to match their statistical properties with reference fine-grid (25 km) simulations. We evaluate its performance on both the target climates and an independent ramped-SST simulation. The translated precipitation fields remove most of the biases simulated by the coarse-grid model in the mean precipitation climatology, the cumulative distribution function of 3-hourly precipitation, and the diurnal cycle of precipitation over land. These results highlight the potential of CycleGAN as a powerful tool for bias correction in climate change simulations, paving the way for more reliable predictions of precipitation patterns across a wide range of climates.

Plain Language Summary

Using CycleGAN, a machine learning technique, we can remove key biases in precipitation simulated by a fast, coarse-grid atmospheric model. This method morphs maps of the output precipitation to match typical characteristics of a slower but more accurate fine-grid configuration, correcting systematic errors in both long-term average spatial precipitation patterns and 3-hourly precipitation variations. It retains skill in intermediate climate states unseen in training, making it a useful tool for climate change simulations.

1 Introduction

Throughout the history of atmospheric model development, results from fine-grid models that resolve important physical processes like cloud and precipitation formation or flow over mountain ranges have been used to improve biased climates in coarse-grid models that do not. For instance, scientists have used large-eddy simulations as a testbed for calibrating analytic turbulence and cloud parameterizations, e.g. Bogenschutz et al. (2010). This process relies heavily on expert knowledge to develop appropriate models of sub-grid behaviors, often heavily influenced by analysis of a few archetypical cases.

More recently, machine learning (ML) has been used to correct coarse-resolution models' behavior across the full range of conditions over historical periods where observational analysis is available. For example, Watt-Meyer et al. (2021) trained a corrective tendency for a 200 km grid atmospheric general circulation model (AGCM) using ML based on nudging tendencies towards observational analysis. The ML correction reduced annual-mean precipitation biases by 20%. An ML approach based on reservoir computing (Arcomano et al., 2023) and ERA5 reanalysis (Hersbach et al., 2020) halved the global root mean square bias of annual-mean precipitation in an even coarser (400 km grid) AGCM.

We require a different strategy when training a model to generalize across future climate forcings, as when training with observational analyses one can only learn the climate represented by this data. One method is to use finer-grid AGCM simulations as training targets. Such simulations are computationally expensive, but they more accurately simulate societally-important aspects of present-day climate such as means and extremes of land surface precipitation and temperature than do coarse-grid AGCMs (Flato et al., 2013; Wehner et al., 2010). Because they resolve much more detail of deep convective storm systems, orography and land surface characteristics, they are less sensitive to uncertain parameterizations of deep convection and orographic drag, making them potentially a more robust simulation tool for generalizing to future climates. S. K. Clark et al. (2022) used the same nudging approach as Watt-Meyer et al. (2021) to ML-correct a 200 km model to behave like its 25 km analogue across four climates forced by adding specified uniform sea-surface temperature (SST) increments to observed SST patterns. They were able to correct spatial patterns of precipitation over land by 10-30% in multi-year simulations across all four climates.

This bias reduction is encouraging, but to get full advantage from high-fidelity reference data, corrective ML should enable both the weather and climate to have much reduced bias (much less than 50% of a no-ML baseline simulation) vs. this reference, both for means and extremes of salient quantities such as precipitation. Yet fundamental challenges persist. This type of hybrid ML, which couples a bias correction model trained offline with a pre-existing AGCM, can induce online simulation biases due to feedbacks between these two components (Brenowitz et al., 2020). The machine learning goal of minimizing prediction error for each sample can lead to difficulties in accurately representing small-scale stochastic behaviors such as deep convection, leading e.g. to an inaccurate representation of the frequency of extreme precipitation (Kwa et al., 2023).

To further reduce bias in the simulated space-time distribution of precipitation vs. a reference climatology, we turn to a different form of ML, the Cycle-generative adversarial network or CycleGAN (Zhu et al., 2017), which is a promising tool for translation of image data between two unpaired domains. Unlike the above hybrid ML approaches, this is a post-processing approach which cannot easily be analyzed in terms of physical process errors in the coarse-grid model, and only corrects selected model fields (precipitation, in our case). In the past, cycle-generative networks have been effectively used for offline bias correction, but it has been necessary to augment the cycle-generative network with quantile mapping to achieve accurate probability distributions of precipitation (François et al., 2021; Pan et al., 2021; Fulton et al., 2023). These works focused on translating one or more model output variables towards an observational analysis over a historical period for a subset of the globe, the annual cycle, or both, and each corrected daily-mean precipitation.

Our work expands on these efforts. We use the original CycleGAN architecture of Zhu et al. (2017) to correct the output of the FV3GFS atmospheric model with a C48 cubed-sphere grid (with approximately 200 km horizontal spacing) to behave like the coarsened output of the same model run on a C384 cubed-sphere grid (25 km spacing). We demonstrate the ability to improve both the spatial distribution of annual-mean precipitation and the cumulative distribution function (CDF) of 3-hourly precipitation up to the 99.999th percentile across a range of climate forcings, without the need for quantile mapping. This method is capable of correcting data at intermediate climate forcings not used during model training, enabling its application to climate change simulations.

2 Dataset

We generate all training data using the FV3GFS atmospheric model (Putman & Lin, 2007; Harris & Lin, 2013; Zhou et al., 2019) as described in McGibbon et al. (2021), run on a cubed-sphere grid with 63 vertical levels. Annually-repeating cycles of sea surface temperature (SST) and sea ice are defined based on the observational monthly means time-averaged from 1982 to 2012 from the 1/12° Real Time Global Sea Surface Temperature (Thiébaux et al., 2003) and 0.5° Climate Forecast System Reanalysis (Saha et al., 2014) datasets, respectively. We perturb the SSTs by adding globally-constant offsets of -2 K, 0 K, +2 K, and +4 K to produce four different sets of forcings while maintaining the present-day annual cycle of sea ice and carbon dioxide concentration, analogous to S. K. Clark et al. (2022). We train using simulations with spacing between SST offsets of 2 K out of concern that precipitation may be too different between forcings at the larger 4 K spacing used in S. K. Clark et al. (2022) for the trained model to accurately generalize to intermediate forcings, though this has not been tested.

For each of these SST forcings, a simulation was performed at C48 resolution for 9 years, 1 month. Eight 1 year, 1 month simulations are performed at C384 resolution beginning with the C48 model snapshot state 1 year into the C48 run as well as the state every year thereafter; the C48 snapshots were converted to C384 initial conditions using the `chgres_cube` tool of `UFS_UTILS` (Gayno et al., 2020). For each of these C384 simulations, the first month of simulation time is discarded as a model spin-up period. This yields 8 years of useful simulation data from each climate, from which we take the first 5 years as training and the last 3 years as validation data.

During these simulations we accumulate and store the 3-hourly mean precipitation rate. We use 3-hourly precipitation instead of daily mean to test the ability of the model to correct biases in the diurnal cycle. At each output time, the C384 precipitation fields are coarsened to the C48 grid by horizontal averaging so that they can be directly compared with coarse-grid precipitation fields.

We also perform "ramping" simulations at both C48 and C384 resolution, which begin with a present-day initial condition and three month spin-up period with 0K forcing, and then enter a period where the forcing is linearly increased from 0K to +2K over the course of 4 years. This data is withheld during training and hyperparameter tuning, and is used for model evaluation only. It tests whether the CycleGAN can skillfully interpolate mean and extreme precipitation patterns between climates on which it was trained.

3 Model formulation and training

The model architecture in Zhu et al. (2017) is used with minimal modifications to allow processing of cubed-sphere data. Specifically, convolution is performed using halo updates on the cubed sphere, where missing corners are filled with zero values. This is numerically identical to the convolution approach used in Weyn et al. (2020), except that data in the corner of each tile domain is set to zero rather than copying and rotating data from the polar tile face. We do not find evidence of corner imprinting despite this choice.

The performance of this model is improved by concatenating spatiotemporal geometric features to the input of the generator and discriminator models. These features are the x, y, and z positions of each grid cell on a stationary unit sphere in 3-dimensional Euclidean space (spatial features), as well as the x and y positions of each grid cell on a unit sphere in Euclidean space as it rotates with a period of one rotation per day (time features). These time features can also be thought of as the x and y positions of an hour hand on a 24-hour clock indicating the local time, multiplied by cos(latitude) to avoid discontinuity at the poles. These are used only as inputs of these models, and are not output by the generators. The discriminator is given identical geometric features to the generator which produced the image being evaluated.

The training dataset includes 58400 3-hourly global snapshots, split evenly across the four climate forcings. Each epoch, we randomly sample 40000 snapshots with replacement, training with a batch size of 1. This data only contains two-dimensional cubed-sphere surface precipitation rate along with a UTC time, which is used to generate the spatiotemporal geometric features.

Notably, the climate forcing itself is absent from the training data, as we were able to achieve excellent bias correction without it. When we included the SST perturbation as input context, as was done for diurnal features, several performance metrics worsened without any clear improvements (compare red vs. steel-blue colors in Figures S2 and S3).

The model was trained with an exponential learning rate decay. Starting with a high learning rate and eventually reducing it further in training is a widely used technique in machine learning (Li et al., 2019). The best results shown here were achieved with an initial learning rate of 10^{-4} and a decay factor of 0.63 (a tenfold decrease every 5 epochs). Training converged (in terms of our precipitation bias metrics) by epoch 14 and was run for 16 epochs (Figure S1).

Otherwise, the hyperparameters are the same as in the 6-layer network of Zhu et al. (2017), but with twice as many filters in the generator and discriminator. We did not attempt to tune the number of layers, activation functions, or choice of optimizer, and we found that increasing the number of filters beyond the value used used did not improve the model.

4 Results

Figure 1 shows the behavior of the generative model on a single sample, taken from the ramping simulation in a climate state distinct from any in the CycleGAN training dataset. We are most concerned with the translation of C48 data into C384 (ML) data (upper left vs. upper right panels), but it is also illuminating to see the inverse generation from C384 to C48 (ML) (lower left vs. upper right panels). The model introduces finer scale features when translating into the C384 domain, especially in lightly precipitating marine boundary layer cloud regimes. It strengthens precipitation over land, introducing precipitation into areas which have none in the C48 input, for example over the South American continent.

The translation substantially improves the mean precipitation climatology vs. C48 simulations for all four SST offsets, as shown in Figure 2, with metrics reported in Figure 3. Here and throughout this analysis, time-mean statistics for fixed-SST simulations such as for the 0K climate are computed on the 3 years of validation data. Statistics for the ramping climate are computed on years 2 and 3 of the 4-year simulation linearly ramping from 0K to plus-2K forcings, to highlight the range of SST offsets that are further from the fixed-SST training data and hence provide a more rigorous out-of-sample test. The bias reductions seen in the ramping simulation are comparable to those in the target climates; the bias of mean precipitation averaged over all land is reduced over 85%, and the standard deviation of the geographic pattern of time-mean bias is reduced over 75% to values around 0.5 mm/d. A significant portion of the biases in each target climate is explained by differences in precipitation between the validation and training datasets, as shown by the "train" bars. Thus, we anticipate further bias reduction with larger training and validation datasets.

Both the 0 K and ramping simulations have smaller precipitation pattern biases than reported for a current-climate case by Arcomano et al. (2023). They reported that their hybrid reservoir computing ML reduced the standard deviation of precipitation pattern bias nearly 50% from 1.2 mm/d in their no-ML baseline model to a value of 0.63 mm/d. Our mean precipitation biases also much smaller than the bias shown in Figure 1 of Fulton et al. (2023) for the South Asian monsoon region. A direct comparison with François et al. (2021) and Pan et al. (2021) is difficult because they considered France and the continental United States, respectively, both of which have much smaller biases than the rest of the globe in our model.

Figure 4 shows that the translated data has a 3-hourly probability distribution and a diurnal cycle of land precipitation that much more closely match the C384 reference data across all climates, including the ramping simulation. We might expect the CycleGAN to struggle to represent extreme precipitation events and their sensitivity to climate forcing because they appear infrequently in the training data. Nevertheless, the 58400 precipitation fields, each containing 13824 atmospheric column, comprise almost 10^9 atmospheric columns, perhaps enough to learn how to translate even highly unusual precipitation events. Indeed, the CycleGAN improves the accuracy of the CDF of precipitation up to the 99.999th percentile. Only at the 99.9999th percentile and only for

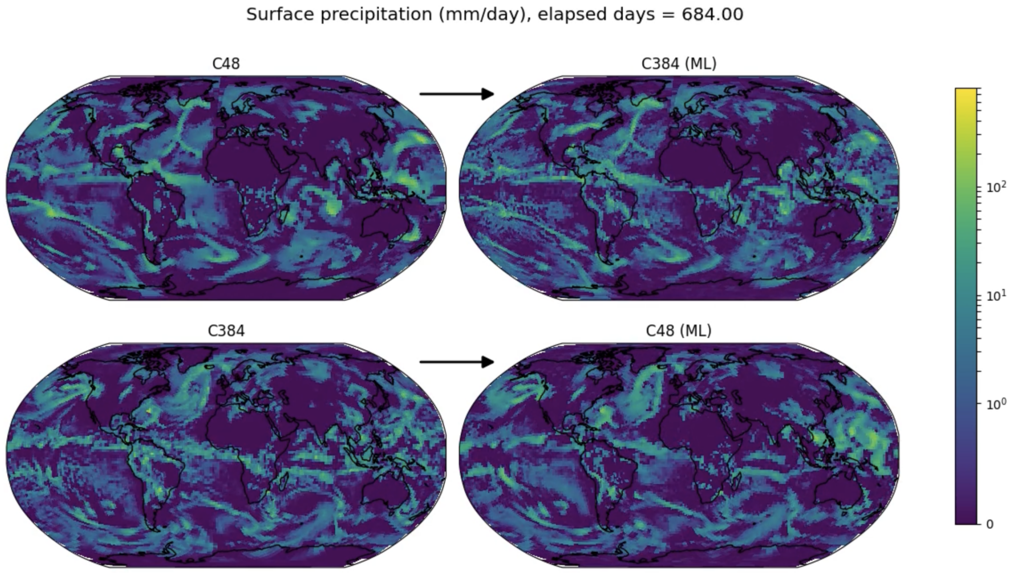

Figure 1. Inputs and outputs of the CycleGAN for one timestep during the ramping simulation. Precipitation data on the left was used as input to generate precipitation data on the right. Snapshot was selected to illustrate a common feature, significantly stronger precipitation over South America in C384 (replicated by the GAN) than in C48. All snapshots for this simulation can be viewed in the supplementary data (Movie S1).

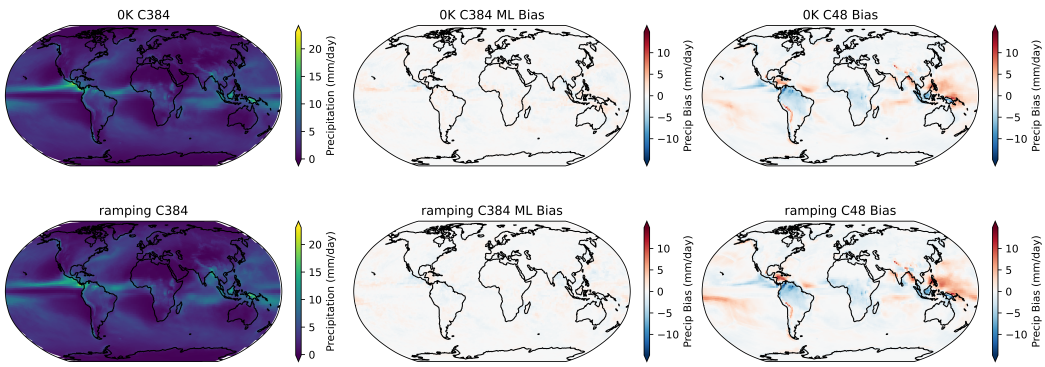

Figure 2. Annual-mean precipitation from C384 reference run (left column) and precipitation biases from the C48 simulation (right column) and from the GAN applied to this C48 simulation (C384 ML). Bias values are differences from the C384 reference.

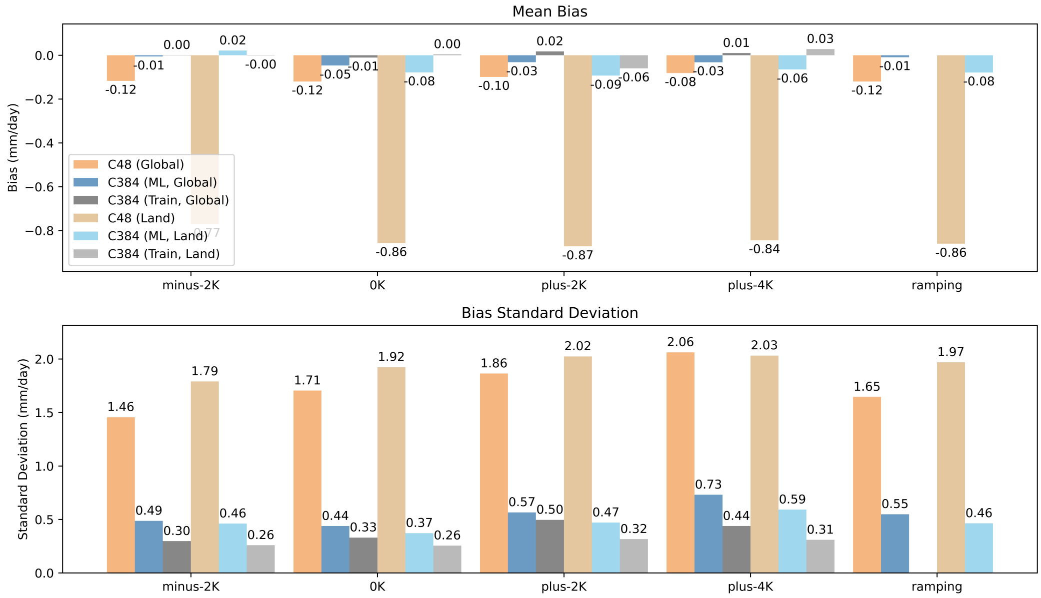

Figure 3. Metrics of time-average precipitation bias against validation and testing data. Mean bias refers to the area-weighted horizontal mean bias across all samples, or over land samples only. Bias standard deviation refers to the square root of the area-weighted mean square bias, averaged over the horizontal either globally or over land samples only. These statistics are derived from bias maps as shown in Figure 2. "Train" indicates the comparison of the training data itself against the validation data. Training data is not available for the ramping simulation.

the -2 K forcing, the CycleGAN slightly increases the error over the input C48 reference data. Surprisingly, the distribution of ML outputs is, if anything, over-dispersive in the tails. The shape of the diurnal cycle of precipitation over land is also improved across all climate forcings, with a stronger trough and sharper increase in precipitation from 6:00 to 15:00 local solar time, and more sustained precipitation through the 21:00-24:00 bin.

Here, the diurnal cycle over land was computed by determining the local solar time in each land-based grid cell for each sample based on its longitude, and then binning the data across local time before taking an area-weighted mean.

5 Sensitivity Studies

This section describes sensitivity studies that help motivate some of our model design choices. We initially trained the CycleGAN model with less data, but found the global maps of time-averaged precipitation vary significantly from year to year, resulting in significant biases in the trained model as a result of under-sampling the long-term climate. When we train the model using only the first year of data from each climate and evaluate on the same 3 years of validation data, the model has significantly worse time-mean biases (Figure S2, compare light purple bar to darker blue bar), and does a significantly worse job predicting the output CDFs, over-predicting the extremes of each climate's precipitation distribution (Figure S3).

Adding spatiotemporal features defining the diurnal cycle as context to the input of the generator and discriminator was crucial for correcting the shape of the mean di-

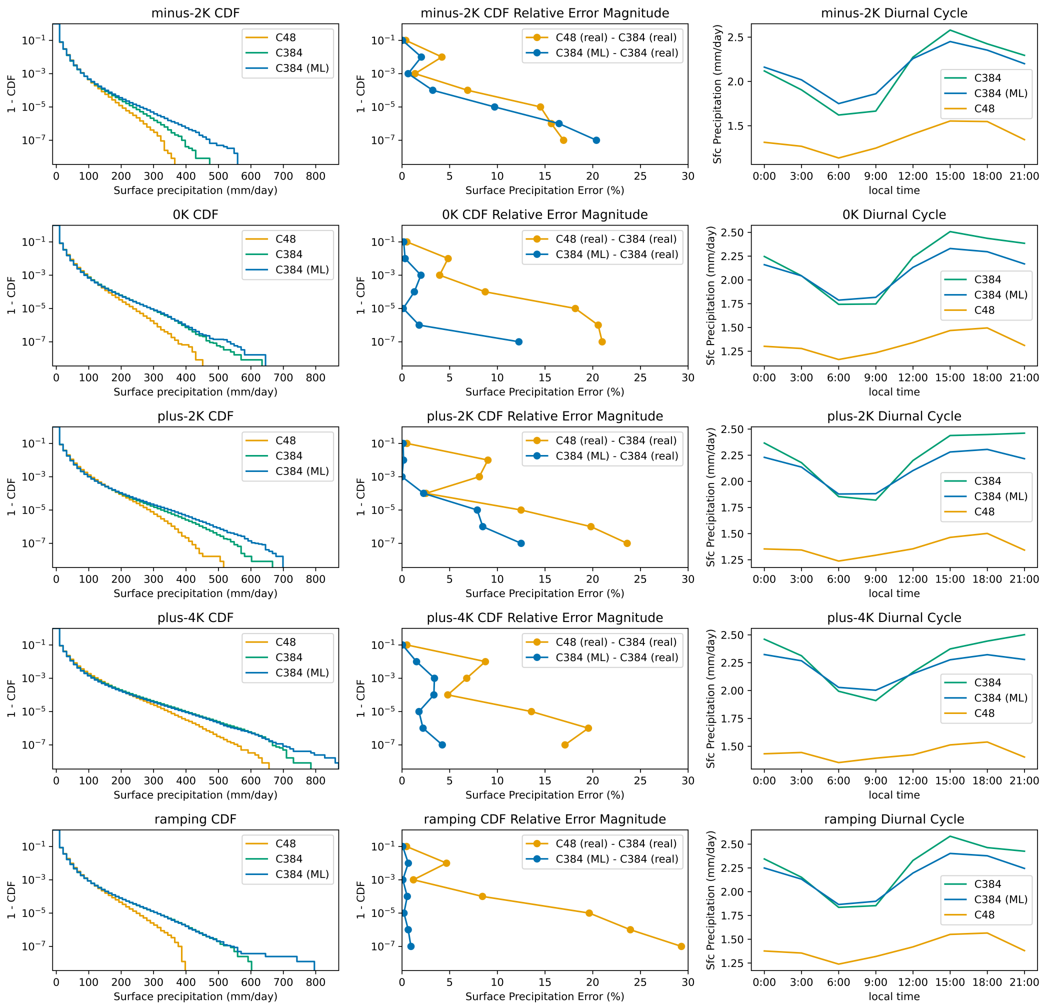

Figure 4. CDF metrics and diurnal cycle of precipitation over land for the reference C384 run, the C48 simulation, and the CycleGAN applied to the C48 output (C384 ML). The left column shows the CDF of precipitation for each climate. The center column shows the relative magnitude of errors of the values of the C48 and CycleGAN CDFs in the first column across a range of percentiles, computed as a percentage of the C384 (real) value. The right column shows the diurnal cycle of precipitation over land, with the x-axis indicating the starting local time of the 3-hour bin.

urnal cycle of precipitation over land. Without these features, the land diurnal cycle of the C384 (ML) output data is significantly improved because we have corrected the correct mean and variance, though the shape of the cycle (light green and light blue) is more similar to the C48 values (orange). Surprisingly, the inclusion of these features has little impact on the standard deviation of the geographically-resolved time-mean bias (Figure S2).

In Zhu et al. (2017), an identity loss was included for certain translation tasks to avoid unnecessary modification of the color scheme during translation. We find removing this identity loss generally degrades model performance. It leads to increased pattern bias and land-mean bias in precipitation (Figure S2) and has a neutral effect on the CDF and land diurnal cycle (Figure S3).

6 Discussion

Unlike previous works using cycle-generative architectures to bias-correct precipitation (François et al., 2021; Pan et al., 2021; Fulton et al., 2023), we could match the PDF of 3-hourly precipitation without quantile mapping. Pan et al. (2021) claimed that quantile mapping is needed because "GANs are trained to produce individual trust-worthy samples, not accurate probability distribution estimations", due e.g. to mode collapse (Bau et al., 2019), despite the claim in Goodfellow et al. (2014) that their training Algorithm 1 is designed to "converge to a good estimator of [the probability distribution of the data], if given enough capacity and training time".

Many methodological differences might explain why we were able to better simulate the probability of extreme precipitation events without quantile mapping. We correct only precipitation, without using other model output fields as dynamical constraints (Pan et al., 2021) or additional fields to be corrected (François et al., 2021; Fulton et al., 2023). Fully sampling the variability and covariability within more fields requires significantly more data, owing to the curse of dimensionality. In addition, our model is trained on more data than the previous studies. We used 58,400 timesteps each with 13,824 gridcells, resulting in 807M precipitation samples, while François et al. (2021); Pan et al. (2021) and Fulton et al. (2023) used 7.42M, 247M, and 40.6M samples respectively. The character of the corrections is different, in particular because of the use of 3-hourly versus daily data and the use of global data instead of limited regions. Our training methodology also differs in the introduction of a learning rate schedule, which could play a role, and Fulton et al. (2023) used the UNIT architecture (Liu et al., 2017) as opposed to CycleGAN.

While this CycleGAN significantly improves the climate of individual samples from a spun-up C48 model state, it should not be used to correct weather simulations run at C48 which are initialized from a coarsened C384 state. We trained the CycleGAN only on samples which are far into a C48 simulation, whose climate contains more significant biases than a hypothetical dataset containing samples from the first week of a C48 simulation initialized from coarsened C384 data. One could remove this input bias effect by training a CycleGAN model to correct model biases at one particular forecast lead time, and using coarse and fine-grid examples at that particular lead time. One could also train a conditional CycleGAN with forecast lead time as a model input capable of correcting a variety of lead times, similar to what was done in this work for time-of-day.

7 Conclusions

We found that CycleGAN with little modification can accurately translate 3-hourly precipitation simulated by a 200 km grid global atmospheric model across a range of climate forcing to have similar statistics as output from a reference fine-grid 25 km model, as measured by its time-mean geographically-resolved pattern, its CDF and its mean diurnal cycle over land. These biases are much reduced compared to previous online correction approaches, but because CycleGAN is a post-processing approach, this comes at the expense of interpretability. The CycleGAN generalizes well to a ramped-SST simulation with intermediate forcings not present in the training dataset. With a small set of expensive fine-grid simulations, the CycleGAN can thus quickly debias precipitation fields predicted by a fast coarse-grid model across a broad range of climates.

8 Open Research

The code used to train and evaluate the machine learning models and produce the figures in this study is available on Zenodo via https://doi.org/10.5281/zenodo.8070950 with MIT and BSD licenses (Brenowitz et al., 2023). The coarse-resolution and coarsened high-resolution model output used for training, validation, and testing are available on Zenodo via https://doi.org/10.5281/zenodo.8070973 with a Creative Commons Attribution 4.0 International License (S. Clark et al., 2023). Figures were made with Matplotlib version 3.7.1 (Caswell et al., 2023), available under the matplotlib license at https://matplotlib.org/. Our machine learning code uses Pytorch version 1.12.1 (Paszke et al., 2019), available under a BSD-3 license at https://pytorch.org/.

Acknowledgments

We thank the Allen Institute for Artificial Intelligence for supporting this work and NOAA-GFDL for providing access to the Gaea computing system used to produce the data on which our model is trained and evaluated. We thank Jun-Yan Zhu for making publicly available a high-quality PyTorch implementation of the CycleGAN model. We used GPT-4 to help create a first draft of the Abstract and Plain Language Summary of this manuscript, which were then significantly revised to correct details.

References

Arcomano, T., Szunyogh, I., Wikner, A., Hunt, B. R., & Ott, E. (2023). A hybrid atmospheric model incorporating machine learning can capture dynamical processes not captured by its physics-based component. *Geophysical Research Letters*, *50*(8), e2022GL102649. doi: https://doi.org/10.1029/2022GL102649

Bau, D., Zhu, J., Wulff, J., Peebles, W. S., Strobelt, H., Zhou, B., & Torralba, A. (2019). Seeing what a GAN cannot generate. *CoRR*, *abs/1910.11626*. doi: https://doi.org/10.48550/arXiv.1910.11626

Bogenschutz, P. A., Krueger, S. K., & Khairoutdinov, M. (2010). Assumed probability density functions for shallow and deep convection. *Journal of Advances in Modeling Earth Systems*, *2*(4). doi: https://doi.org/10.3894/JAMES.2010.2.10

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., … Bretherton, C. S. (2020). *Machine learning climate model dynamics: Offline versus online performance.* doi: https://doi.org/10.48550/arXiv.2011.03081

Brenowitz, N. D., Kwa, A., Perkins, W. A., Henn, B., McGibbon, J., Clark, S., … Nabizadeh, E. (2023, June). *ai2cm/fv3net: Commit used for CycleGAN initial submission.* Zenodo [Software]. doi: https://doi.org/10.5281/zenodo.8070950

Caswell, T. A., Lee, A., de Andrade, E. S., Droettboom, M., Hoffmann, T., Klymak, J., … Moad, C. (2023, March). *matplotlib/matplotlib: Rel: v3.7.1.* Zenodo [Software]. doi: https://doi.org/10.5281/zenodo.7697899

Clark, S., McGibbon, J., & Bretherton, C. (2023, June). *Data for "Global Precipitation Correction Across a Range of Climates Using CycleGAN".* Zenodo [Dataset]. doi: https://doi.org/10.5281/zenodo.8070973

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A.,

… Harris, L. M. (2022). Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003219. doi: https://doi.org/10.1029/2022MS003219

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., … Rummukainen, M. (2013). Evaluation of climate models. In T. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (p. 741–866). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: https://doi.org/10.1017/CBO9781107415324.020

François, B., Thao, S., & Vrac, M. (2021, July). Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, *57*(11-12), 3323–3353. doi: https://doi.org/10.1007/s00382-021-05869-8

Fulton, D. J., Clarke, B. J., & Hegerl, G. C. (2023). Bias correcting climate model simulations using unpaired image-to-image translation networks. *Artificial Intelligence for the Earth Systems*, 1 - 56. doi: https://doi.org/10.1175/AIES-D-22-0031.1

Gayno, G., Beck, J., Reames, L., programmer, G., Wright, D., Hu, M., … Gerheiser, K. (2020, August). *UFS_UTILS.* Unified Forecast System (UFS).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). *Generative adversarial networks.* doi: https://doi.org/10.48550/arXiv.1406.2661

Harris, L. M., & Lin, S.-J. (2013). A two-way nested global-regional dynamical core on the cubed-sphere grid. *Monthly Weather Review*, *141*(1), 283 - 306. doi: https://doi.org/10.1175/MWR-D-11-00201.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., … Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. doi: https://doi.org/10.1002/qj.3803

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., … Bretherton, C. S. (2023). Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003400. doi: https://doi.org/10.1029/2022MS003400

Li, Y., Wei, C., & Ma, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. *CoRR*, *abs/1907.04595*. doi: https://doi.org/10.48550/arXiv.1907.04595

Liu, M., Breuel, T. M., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *CoRR*, *abs/1703.00848*. doi: https://doi.org/10.48550/arXiv.1703.00848

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S., Davis, E. C., … Fuhrer, O. (2021). fv3gfs-wrapper: a python wrapper of the fv3gfs atmospheric model. *Geoscientific Model Development*, *14*(7), 4401–4409. doi: https://doi.org/10.5194/gmd-14-4401-2021

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., Lee, J., … Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509. doi: https://doi.org/10.1029/2021MS002509

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, *abs/1912.01703*. doi: https://doi.org/10.48550/arXiv.1912.01703

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55 - 78. doi: https://doi.org/10.1016/j.jcp.2007.07.022

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., ... Becker, E. (2014, March). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208. doi: https://doi.org/10.1175/JCLI-D-12-00823.1

Thiébaux, J., Rogers, E., Wang, W., & Katz, B. (2003, May). A New High-Resolution Blended Real-Time Global Sea Surface Temperature Analysis. *Bulletin of the American Meteorological Society*, *84*(5), 645–656. doi: https://doi.org/10.1175/BAMS-84-5-645

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: https://doi.org/10.1029/2021GL092555

Wehner, M. F., Smith, R. L., Bala, G., & Duffy, P. (2010). The effect of horizontal resolution on simulation of very extreme us precipitation events in a global atmosphere model. *Climate Dynamics*, *34*(2), 241–247. doi: https://doi.org/10.1007/s00382-009-0656-y

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109. doi: https://doi.org/10.1029/2020MS002109

Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., & Rees, S. L. (2019). Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, *100*(7), 1225 - 1243. doi: https://doi.org/10.1175/BAMS-D-17-0246.1

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, *abs/1703.10593*. doi: https://doi.org/10.48550/arXiv.1703.10593

Supporting Information for "Global Precipitation Correction Across a Range of Climates Using CycleGAN"

J. McGibbon[1], S. K. Clark[1,2], B. Henn[1], A. Kwa[1], O. Watt-Meyer[1], W. A. Perkins[1], C. S. Bretherton[1]

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA

[2]Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Contents of this file

1. Figures S1 to S4

Additional Supporting Information (Files uploaded separately)

1. Caption for Movie S1

Introduction

Movie S1. Four-year ramping simulations, depicting the real input C48 and C384 precipitation, and the generated C384 (ML) and C48 (ML) precipitation based on these inputs for each 3-hourly sample.

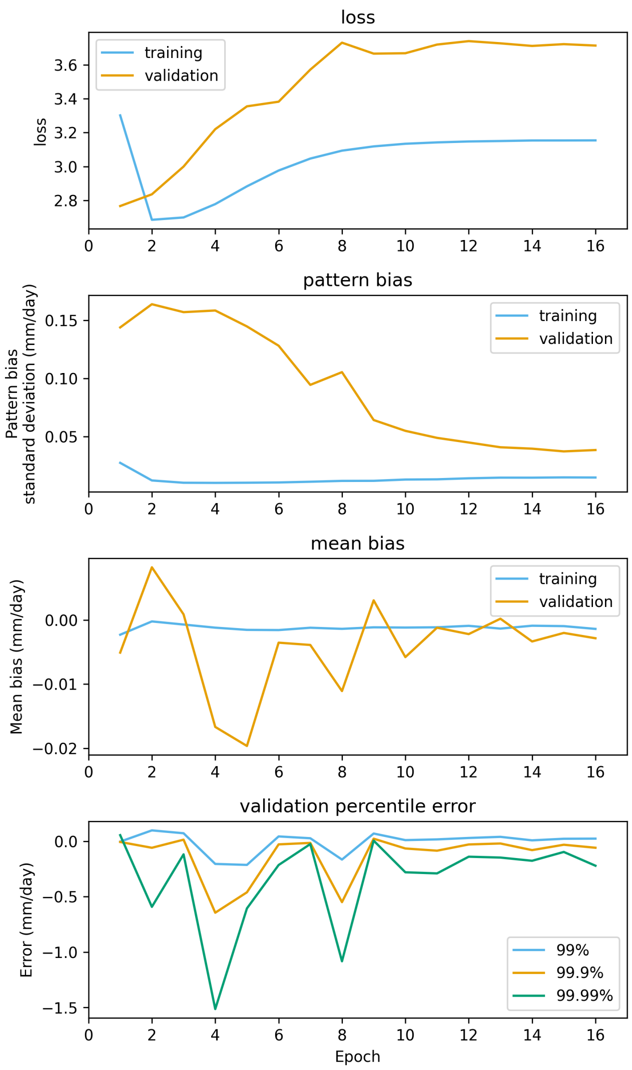

Figure S1. Metrics recorded while training the best-case CycleGAN model which were used to determine convergence. Note that when training GANs the model can improve as the loss increases, due to compensating increases in skill of the generator and discriminator models. Pattern bias for training is computed by first aggregating the time-mean of predicted outputs for each training batch. Loss indicates total loss optimized during training. The dataset considered in these losses includes data from all four training climates.

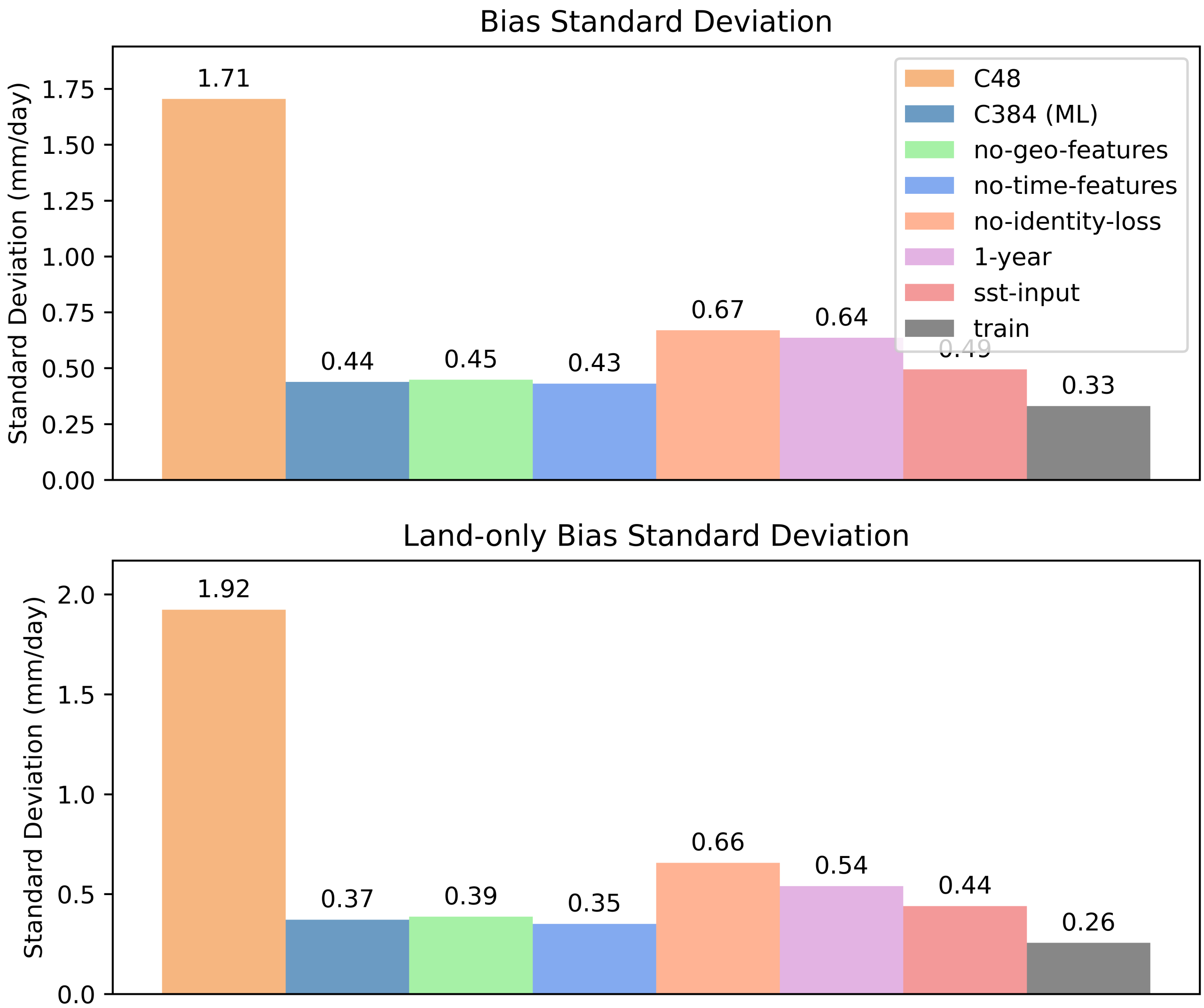

Figure S2. Area-weighted root mean-squared error of time-mean precipitation for ablation studies in the "0K" climate, both globally and over land only. C384 (ML) refers to the "best" model presented in the main text. no-geo-features and no-time-features are the models where we excluded all five geographic features and the two diurnally-varying geographic features, respectively. no-identity-loss is the model where the identity loss was excluded from training. 1-year is a model trained on only the first year of training data. sst-input is the model where the SST perturbation is provided as input context to the generator and discriminator networks. train is the result when using the coarsened C384 training data instead of model output.

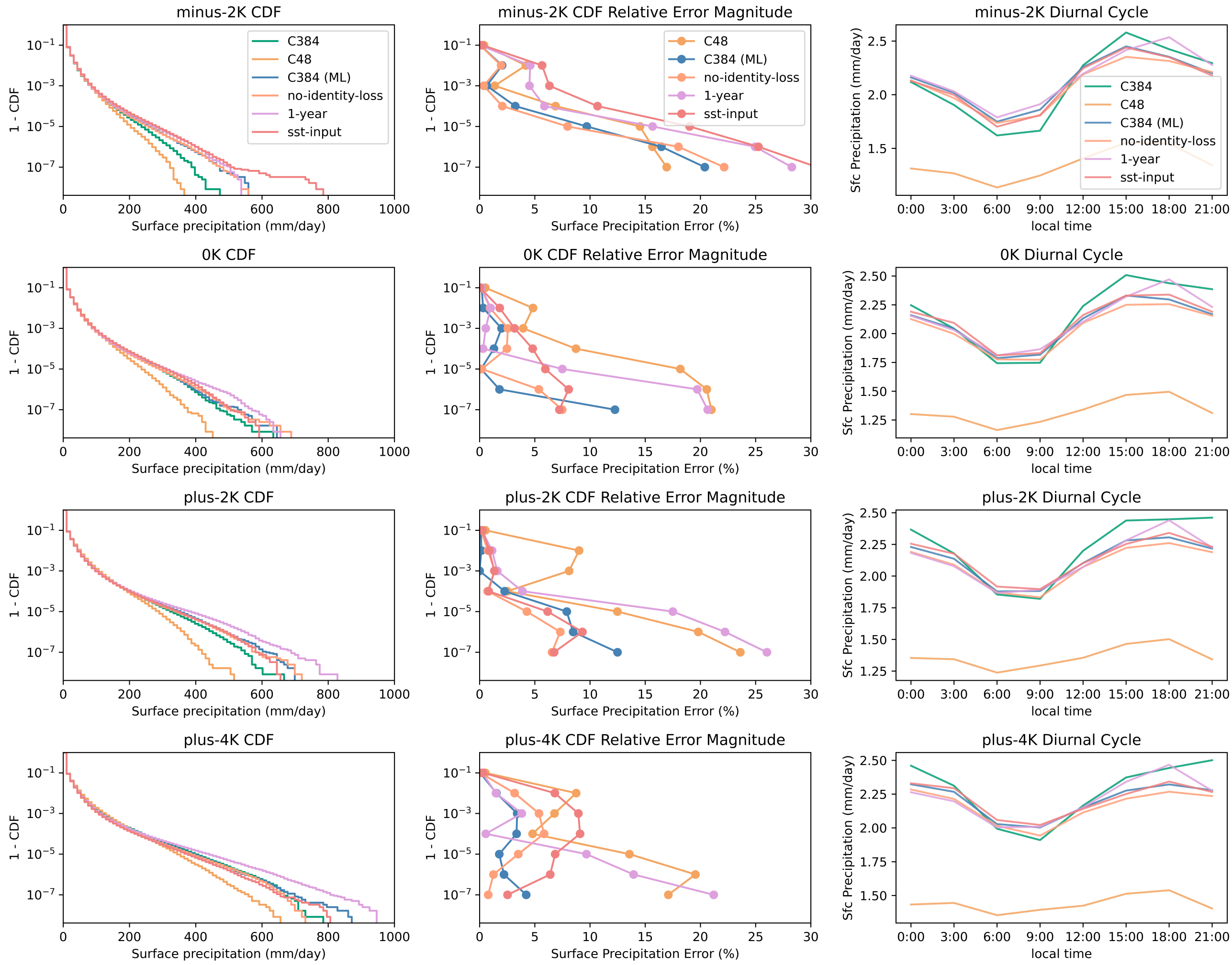

Figure S3. CDF and land-only diurnal cycle metrics for ablation study models which don't involve modifying the geographic features. Labels are as in Figure S2.

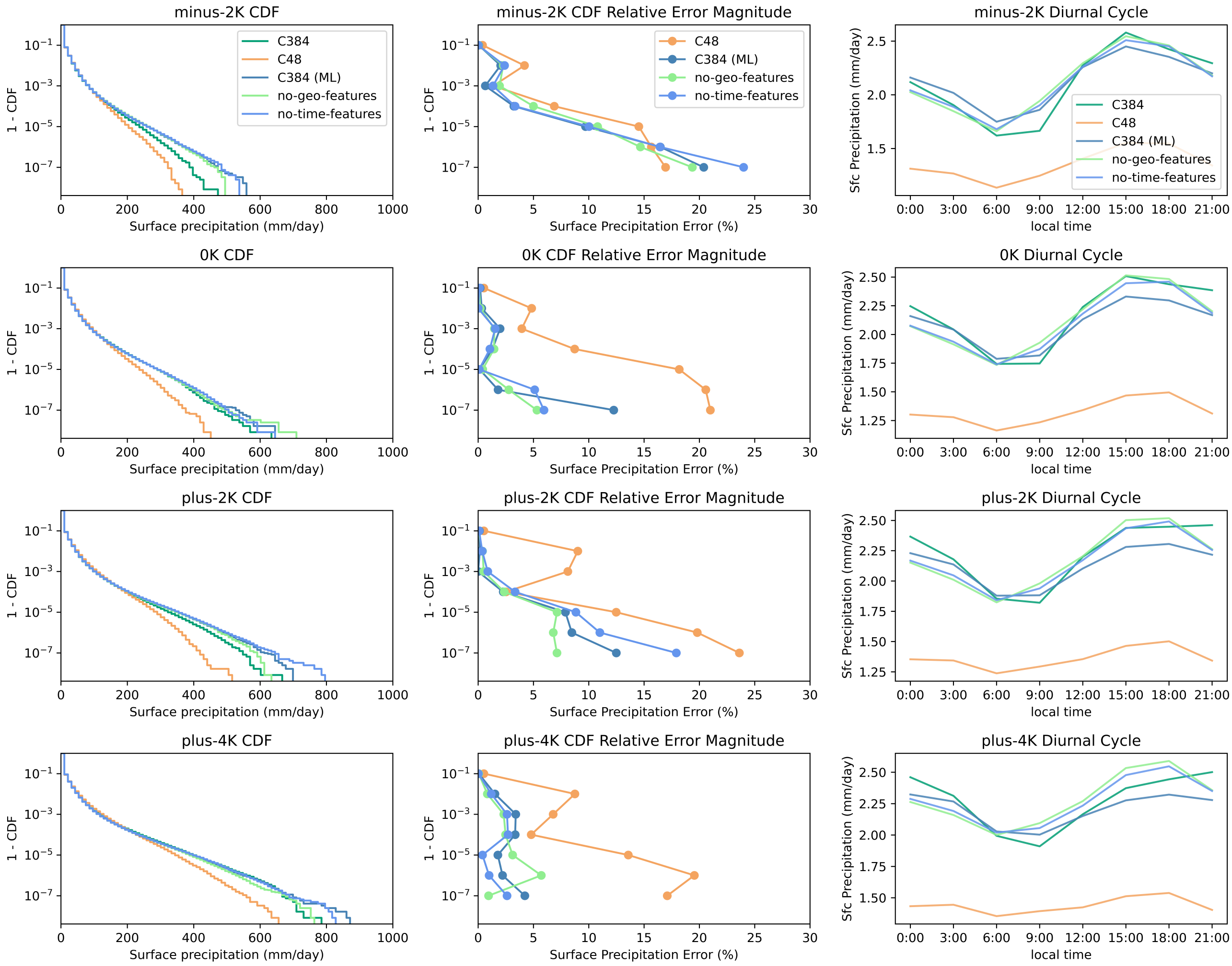

Figure S4. CDF and land-only diurnal cycle metrics for ablation study models which involve modifying the geographic features. Labels are as in Figure S2.