Towards debiasing climate simulations using unsupervised image-to-image translation networks

James D. Fulton [1] **Ben J. Clarke** [2]

Abstract

Climate models form the basis of a vast portion of earth system research, and inform our climate policy. Due to the complex nature of our climate system, and the approximations which must necessarily be made in simulating it, these climate models may not perfectly match observations. For further research, these outputs must be bias corrected against observations, but current methods of debiasing do not take into account spatial correlations. We evaluate unsupervised image-to-image translation networks, specifically the UNIT model architecture, for their ability to produce more spatially realistic debiasing than the standard techniques used in the climate community.

1. Introduction

The vast majority of research on the physical earth system is reliant on the use of general circulation models (GCMs). These are numerical simulators which evolve an ocean and/or atmospheric state forward in time according to analytically-derived equations (Washington & Parkinson, 2005). Through the work of the International Panel on Climate Change (IPCC), these GCMs have been central to informing policymakers about climate change (IPCC, 2013).

There are a multiplicity of different GCMs, developed by research centres across the world. The latest Climate Model Intercomparison Project (CMIP6) (Eyring et al., 2016) includes output from tens of GCMs. These GCMs must discretise the earth in order to simulate it, and no matter how fine-grained the simulation, there will always be unresolved physical processes. These must be approximated, and different GCMs use different parameterisations to do so. Because our climate is highly chaotic, these different parameterisations can lead to detectable differences in the GCMs' outputs (Wang et al., 2014; Maher et al., 2018).

These differences are important, as the simulations are increasingly used to assess the risks associated with extreme weather events, such as droughts, heatwaves and storms. To quantify these risks, researchers must decide which GCM(s) to use. This is a non-trivial choice, as it can cause quantitative and qualitative differences on these risks (e.g. (Kirchmeier-Young et al., 2017; Herger et al., 2018)). Further, the biases of each GCM vary by variable and geographical location (Ridder et al., 2021). This makes compound risks (Leonard et al., 2014), like simultaneous hot-dry events, even harder to assess, as they involve multiple variables and possibly multiple geographical areas. Simultaneous crop failure in multiple regions of high agricultural output (Gaupp et al., 2020) is an important risk which may be particularly affected.

GCMs are improving, but will never be perfect. They can be made more useful through bias correction with respect to observations of the real climate (Bellprat et al., 2019). The standard method used is quantile mapping (QM) (Cannon et al., 2015). This is a simple method where a single value of a variable x obtained from the GCM at a spatial location $\vec{\theta}$, is converted into a percentile using the estimated cumulative distribution function $\mathcal{F}_{GCM}$. Then an equivalent observation value $\hat{x}_{obs}$ is obtained using the inverse cumulative distribution function $\mathcal{F}^{-1}_{obs}$.

$$\hat{x}_{obs} = \mathcal{F}^{-1}_{obs}(\mathcal{F}_{GCM}(x_{GCM}; \vec{\theta}); \vec{\theta}) \tag{1}$$

This approach is limited, as it does not capture conditional relationships. The QM predicted value of $\hat{x}_{obs}$ does not use the values of x_{GCM} in neighbouring locations in space. This means, for example, if some weather event in a GCM has a different characteristic shape than in observations, then QM cannot reshape it coherently. This could be the size and shape of a cyclone, or the position of storms along the polar front.

Further, QM does not use the values of other variables at the same spatial location. So relationships between variables are severed and may become physically unrealistic. For example, if QM transforms a dry day in the GCM to a

*Equal contribution [1]School of GeoSciences, University of Edinburgh [2]School of Geography and the Environment, University of Oxford. Correspondence to: James D. Fulton <james.fulton@ed.ac.uk>.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

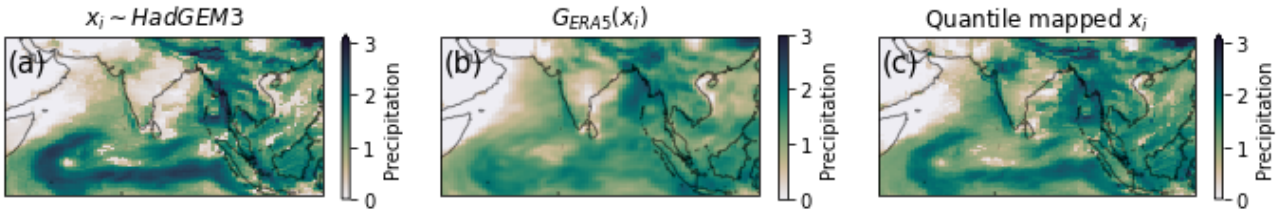

Figure 1: An example HadGEM3 precipitation field from a single day and its translations via UNIT and QM. The forth root of precipitation (mm day^{-1}) is shown.

wet day in equivalent observations, it may not have the associated cooling effect of precipitation.

Previous work has attempted to solve some of these issues, but incompletely. In (Levy et al., 2013) the authors propose stretching simulated precipitation fields to match precipitation patterns in observations. This was for monthly average precipitation, and the technique does not allow the use of daily data which is important for studying extreme event risks. Floods and fires happen on a timescale of days, not months. It also cannot easily be extended to multiple variables. In (Cannon, 2018) the authors propose a way to generalise QM to N-dimensions. This allows the user to transform multiple variables in multiple spatial locations using daily data. However, as they note, their method cannot be extended to many spatial points in many variables before it becomes computationally limited, and becomes prone to overfitting.

In this work we explore the use of unsupervised image-to-image translation networks, particularly the UNIT architecture (Liu et al., 2017), to map between GCMs and observations.

2. Unpaired image-to-image translation

Unsupervised image-to-image translation networks such as UNIT (Liu et al., 2017) and CycleGAN (Zhu et al., 2017) are designed to map between images $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ in the absence of any information about which image x_i corresponds to which image y_j. This makes the problem distinct from image-to-image translation (e.g. (Isola et al., 2017)) where image pairs $\{x_i, y_i\}_{i=1}^N$ are available.

In order to map between two GCMs, or a GCM and observations, the ability to translate without pairs is absolutely necessary. Even if two GCMs were initialised with the exact same climate state $\psi_0 = \psi_X(0) = \psi_Y(0)$, when we run them forward in time there will be no correspondence between $\psi_X(t)$ and $\psi_Y(t)$ after only a short time $t > \tau$, due to the chaotic dynamics of the earth system. This is also true between a GCM and observations. The time scale over which a GCM accurately predict the weather τ, is approximately 2 weeks (Lorenz, 1969; Zhang et al., 2019). This is a blink of an eye on time scales of climate, as our simulations often last tens or hundreds of years. But to assess risks, we need to know that the daily weather data generated by these simulations is representative of our the real climate system.

It is also not valid to train a mapping using pairs collected at times $t < \tau$, like recent debiasing for weather forecasts (Steininger et al., 2020) have done, and assume that this short time-range mapping can be applied to a long simulation. The observed initial climate state ψ_0 may be an unrealistic state for the free-running GCM, and after some time, the simulated climate state may drift (Gupta et al., 2013) and never revisit states similar to those where training examples were collected. So using the trained mapping would be predicting outside of the training data.

2.1. Unsupervised image-to-image translations networks

Unsupervised image-to-image translations networks are based on the architecture of generative adversarial networks (GANs) (Goodfellow et al., 2014). They are composed of two conditional generators which carry out mappings between the two domains, $G_Y : X \to Y$ and $G_X : Y \to X$. G_Y tries to generate images $G_Y(x)$ which are consistent with the domain Y, and has an accompanying adversarial discriminator D_Y which aims to detect the fake images created by G_Y. G_X is similar and has an adversary D_X. The GAN loss associated with one of these mappings is

$$\mathcal{L}(G_Y, D_Y, X, Y) = \mathbb{E}_{y\sim Y}[\log D_Y(y)] + \mathbb{E}_{x\sim X}[\log(1 - D_Y(x))]. \quad (2)$$

$\mathcal{L}(G_X, D_X, Y, X)$ is similar. It is also necessary that the mappings show cycle consistency, so that $G_X(G_Y(x)) \approx x$ and $G_Y(G_X(y)) \approx y$. This ensures that the network learns a coherent two-way translation. The total loss function includes a penalty associated with this comparison. We use L1 loss here.

These networks have been shown to translate between photos of horses and zebras; and between simulated city driving

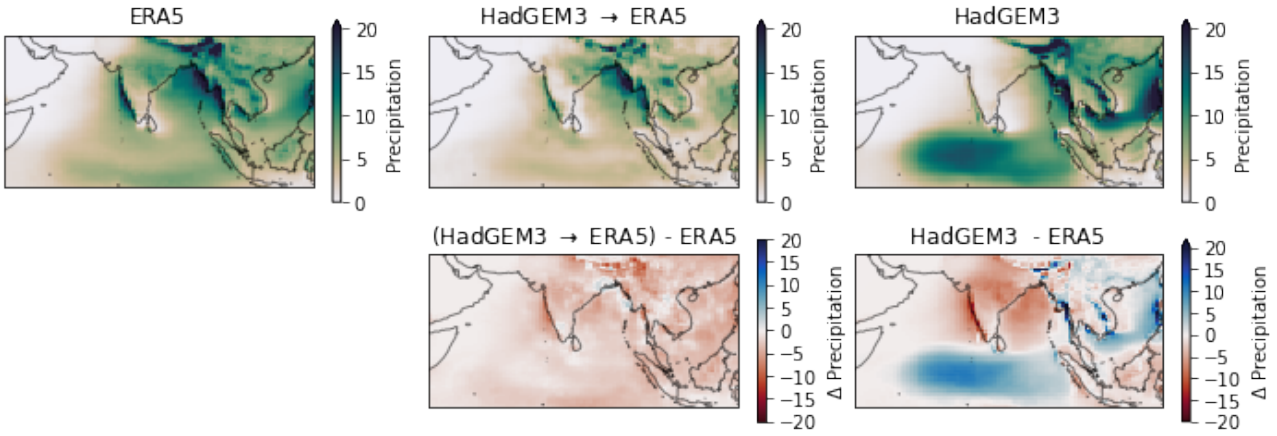

Figure 2: Mean precipitation (mm day^{-1}) during monsoon season (June-September) for the ERA5, and HadGEM3 datasets and the UNIT translation of the HadGEM3 data.

scenes and real city driving scenes. This last use case is most similar to our purpose here. We note also that in (Shrivastava et al., 2017) the authors train a related *refiner* model to make simulated eye poses more realistic.

3. Methods

We use GCM data from the Climate of the 20th Century Plus (C20C+) Project (Folland et al., 2014), particularly from the `HadGEM3-A-N216` GCM (Ciavarella et al., 2018) run under a historical recreation scenario. In this dataset the ocean temperatures are prescribed.[1] We also use `ERA5` reanalysis data (Hersbach et al., 2020) as a stand-in for perfect observations. These data are daily global gridded fields which appear like multichannel images, where the channels are 2-metre mean, minimum, and maximum temperature; daily accumulated precipitation; and z500 geopotential height.

We train a UNIT neural network to perform a mapping between the datasets in the geographical region bounded by 8°S-30°N 44°E-121°E. This region of study was chosen to capture the South Asian monsoon. More details on the networks can be found in the appendix and source paper (Liu et al., 2017).

4. Results

Figure 1 shows an example day of precipitation sampled during the South Asian summer monsoon. The top row of the figure shows the sample from the HadGEM3 dataset, and its mapping onto the ERA5 dataset using the UNIT network and QM.

A striking feature of this example is crook-shaped rain system over the Indian Ocean in panel (a). Weather systems similar to this example are common in the HadGEM3 data, but do not occur in the ERA5 data. The UNIT transform of this example removes this system, whilst quantile mapping only morphs its intensity, but does not remove its spatial structure. This leaves a precipitation signal which is physically implausible.

The South Asian monsoon is crucial to the region, so reliable bias correction is likewise imperative. This weather system is a common bias across GCMs (Bollasina & Ming, 2013; Ashfaq et al., 2017), and does not occur in observations. Figure 2 shows how this weather system impacts the mean monsoon precipitation over the region. The figure shows that the UNIT translation removes the spatial structure of the mean bias. It increases precipitation over the Indian subcontinent, and reduces it in the Indian Ocean. Note that UNIT achieves this mean correction whilst being trained only to map individual days. This is unlike QM which explicitly maps the entire univariate distributions. QM is not shown in figure 2 as it would, by definition, have the same mean precipitation as the target ERA5 dataset. However, as shown in figure 1, the daily QM fields can be left with unphysical structures. We must then ask, how are we to trust the aggregate when the individual days seem nonsensical?

It is a fiendish task to determine whether any presented weather field is realistic. We therefore cannot use the same human verification as used in other work (e.g. (Salimans et al., 2016)) to find an aggregate score of how realistic the fields are. Neither can we train an off-the-shelf classifier to verify whether our generated images contain useful features as in (Isola et al., 2017). In order to estimate how realistic the translated fields are, we take each translated field

[1]The GCM is run in a mode where the sea surface temperatures are forced to set values at given times - i.e. they are treated as boundary conditions. Meanwhile the atmosphere may freely evolve.

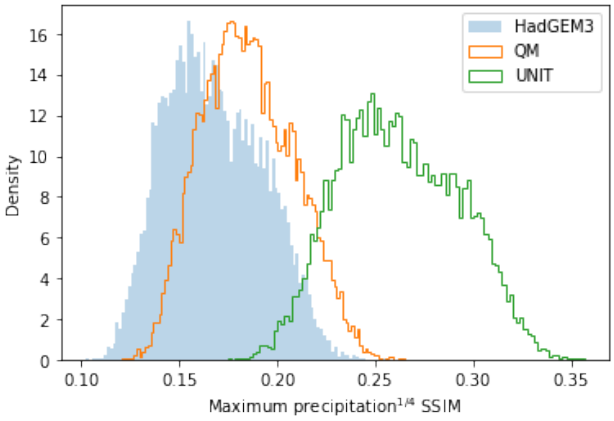

Figure 3: Distributions of the maximum SSIM score between each daily data sample from {HadGEM3, UNIT translated HadGEM3, QM translated HadGEM3} and all samples from the ERA5 dataset. The data used are the forth root of precipitation (mm day^{-1}).

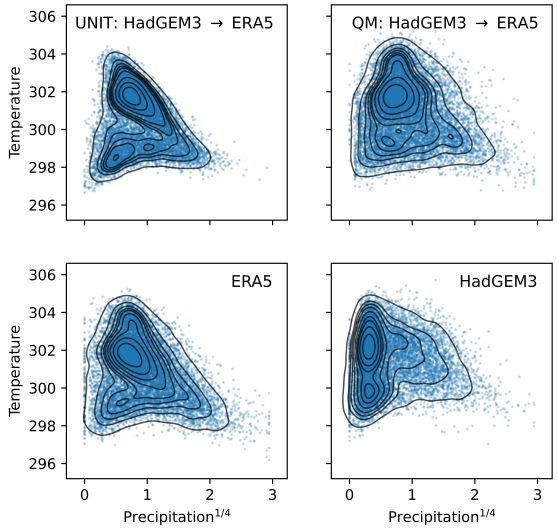

Figure 4: Joint distributions between temperature (K) and the 4th root of precipitation (mm day^{-1}). These data are sampled from a single spatial location in the southern tip of India (box centred at 8.6°N 77.9°E with longitudinal width 0.83° and latitudinal height 0.56°). In the figure, the points are the measurements on individual days. The contours are an estimate of the distribution via kernel density estimation. These distributions are shown for the UNIT translation of the HadGEM3 data to ERA5, the equivalent QM translation, the original ERA5 data, and the original HadGEM3 data.

and find the maximum structural similarity index measure (SSIM) (Wang et al., 2004) score between it and all samples of the target ERA5 dataset. In figure 3, we plot the distribution of these scores for the HadGEM3 data, and its UNIT, and QM translations. Here, the SSIM was calculated between the forth root of the precipitation fields. The forth root was used due to the extreme skew of the precipitation values. The figure shows that The UNIT translation produced fields that were more structurally similar to the ERA5 dataset than QM. This supports the example shown in figure 1 where QM introduces unphysical spatial structures.

Finally, we compare a sample of cross-variable correlations. Figure 4 shows the joint distribution between the temperature and precipitation from a single grid-point at the southern tip of India, for the original datasets and the translations. The ERA5 data appears to have two notable peaks in the joint distribution, and a strong conditional dependence between the temperature and precipitation at higher precipitation values. The QM translation smears out the features of the joint distribution, whilst the UNIT translation perhaps enhances these features too much, with less values on the perimeter of the distribution than should be the case. As the UNIT network has an adversarial component, this enhancement may be be a form of boundary distortion, as is common in generative adversarial networks (Santurkar et al., 2018). We also note that the dry bias, as shown in figure 2, was due to the same reduction in the high end of the precipitation distributions at each point. Therefore, this may be a systematic bias in the UNIT model.

5. Conclusions and further work

We have shown that unsupervised image-to-image translation networks, in particular the UNIT architecture (Liu et al., 2017), can be used to debias climate simulations with respect to observations; a task that has previously been dominated by quantile mapping. We've shown how the UNIT network can produce more spatially realistic mappings of simulation data than quantile mapping can. This is because the UNIT networks considers spatial correlations whilst quantile mapping does not.

Improvements in the power and capacity of unsupervised image-to-image translation networks are already under way (e.g. (Park et al., 2020; Wang et al., 2020)), which is very encouraging for its use in GCM debiasing. However, issues remain around how reliably the UNIT model architecture represents the boundary of its target distribution. Further developments on this issue could allow image-to-image translation networks to find an important niche in this application.

Acknowledgements Computing resources for this work were provided by a Microsoft AI for Earth grant.

Broader Impact

The authors do not foresee negative ethical consequences as a result of this work. On the contrary, if further developments were made in this area, this could be on great use to climate researchers, and could be used to better understand and predict the effects of climate change.

References

Ashfaq, M., Rastogi, D., Mei, R., Touma, D., and Leung, L. R. Sources of errors in the simulation of south asian summer monsoon in the cmip5 gcms. *Climate Dynamics*, 49(1):193–223, 2017.

Bellprat, O., Guemas, V., Doblas-Reyes, F., and Donat, M. G. Towards reliable extreme weather and climate event attribution. *Nature communications*, 10(1):1–7, 2019.

Bollasina, M. A. and Ming, Y. The general circulation model precipitation bias over the southwestern equatorial Indian Ocean and its implications for simulating the South Asian monsoon. *Climate dynamics*, 40(3):823–838, 2013.

Cannon, A. J. Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics*, 50(1):31–49, 2018.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q. Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17):6938–6959, 2015.

Ciavarella, A., Christidis, N., Andrews, M., Groenendijk, M., Rostron, J., Elkington, M., Burke, C., Lott, F. C., and Stott, P. A. Upgrade of the HadGEM3-A based attribution system to high resolution and a new validation framework for probabilistic event attribution. *Weather and climate extremes*, 20:9–32, 2018.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

Folland, C., Stone, D., Frederiksen, C., Karoly, D., and Kinter, J. The international CLIVAR Climate of the 20th Century Plus (C20C+) Project: Report of the sixth workshop. *CLIVAR Exchange*, 19:57–59, 2014.

Gaupp, F., Hall, J., Hochrainer-Stigler, S., and Dadson, S. Changing risks of simultaneous global breadbasket failure. *Nature Climate Change*, 10(1):54–57, 2020.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Gupta, A. S., Jourdain, N. C., Brown, J. N., and Monselesan, D. Climate drift in the cmip5 models. *Journal of Climate*, 26(21):8597–8615, 2013.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M. Selecting a climate model subset to optimise key ensemble properties. *Earth System Dynamics*, 9(1):135–151, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi: 10.1017/CBO9781107415324.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Jones, P. W. First-and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, 127(9):2204–2210, 1999.

Kirchmeier-Young, M. C., Zwiers, F. W., and Gillett, N. P. Attribution of extreme events in arctic sea ice extent. *Journal of Climate*, 30(2):553–571, 2017.

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., and Stafford-Smith, M. A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):113–128, 2014.

Levy, A. A., Ingram, W., Jenkinson, M., Huntingford, C., Hugo Lambert, F., and Allen, M. Can correcting feature location in simulated mean climate improve agreement on projected changes? *Geophysical research letters*, 40 (2):354–358, 2013.

Liu, M., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 700–708, 2017.

Lorenz, E. N. The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307, 1969.

Maher, P., Vallis, G. K., Sherwood, S. C., Webb, M. J., and Sansom, P. G. The impact of parameterized convection on climatological precipitation in atmospheric global climate models. *Geophysical Research Letters*, 45(8):3728–3736, 2018.

Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

Ridder, N. N., Pitman, A. J., and Ukkola, A. M. Do CMIP6 climate models simulate global or regional compound events skillfully? *Geophysical Research Letters*, 48(2): e2020GL091152, 2021.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2226–2234, 2016.

Santurkar, S., Schmidt, L., and Madry, A. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning*, pp. 4480–4489. PMLR, 2018.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116, 2017.

Steininger, M., Abel, D., Ziegler, K., Krause, A., Paeth, H., and Hotho, A. Deep learning for climate model output statistics. *CoRR*, abs/2012.10394, 2020.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R. A global perspective on CMIP5 climate model biases. *Nature Climate Change*, 4(3):201–205, 2014.

Wang, Y., Yu, L., and van de Weijer, J. DeepI2I: Enabling deep hierarchical image-to-image translation by transferring from gans. *arXiv preprint arXiv:2011.05867*, 2020.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Washington, W. M. and Parkinson, C. *Introduction to three-dimensional climate modeling*. University science books, 2005.

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., and Emanuel, K. What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091, 2019.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A. Data

We use the `All-Hist/est1` scenario from the C20C+ Project for the `HadGEM3-A-N216` GCM. This scenario represents other possible weather histories we could have observed without modifying anthropogenic emissions. This uses observed emissions, and the sea surface temperatures are prescribed using observations. We also use the `ERA5` data (Hersbach et al., 2020) as an observational product. These data are generated by weather prediction models that assimilate observations to reproduce the observed climate state, and provide a complete representation of the atmosphere. In this experiment we used only a single run from the GCM, although many exist, and single ERA5 'run'. These datasets were filtered to the time period in which they are both available, which covered 1979-2013. This gave us around 12,500 images from each source, 20% of which were held back for testing.

The raw datasets were on different grids. For ease of implementation we decided to regrid them to match the resolution of the coarser of the two in each experiment. We used conservative regridding (Jones, 1999).

We preprocessed the data so the 2-metre temperature was in Celsius and divide it by the temperature standard deviation globally. This was designed so the physically meaningful value of 0°C is still given the value of zero after preprocessing. The daily min and max temperatures were scaled identically to the mean, and then transformed to their difference from the mean i.e. $T_{max} \rightarrow (T_{max} - T_{mean})/\sigma$ and $T_{min} \rightarrow (T_{mean} - T_{min})/\sigma$. This was motivated by early experiments where negative daily temperature ranges were occasionally produced in translations. This happened solely at land-sea borders, which suggested our network was struggling to infer this boundary. Due to this we included a land-sea mask channel as an input to both the generator and the discriminator. The generator was not tasked with translating the land-sea mask, as it is a constant.

Due to the extreme skew of precipitation, we applied the transform $precip \rightarrow precip^{1/4}$ in preprocessing. We found this transformation made the distribution of precipitation values near-normal.

B. UNIT network

We trained a mostly vanilla UNIT network as described in the original paper (Liu et al., 2017).

Modifications were made to the final activation functions to enforce positive values of precipitation, $T_{max} - T_{mean}$, and $T_{mean} - T_{min}$. ReLU activations were used for these. In each convolutional layer we used replication padding.

C. Quantile mapping

We implemented empirical quantile mapping. This form of QM does not assume anything about the shape of the distribution. A cumulative distribution function (CDF) was calculated for each month of the year, spatial point, and variable. For the highest resolution dataset this meant $(12 \text{ months})(68 \text{ latitudes})(92 \text{ longitudes})(5 \text{ variables}) = 375360$ CDFs. Each CDF was estimated by finding the variable value of 100 quantiles spread equally between $(0+\epsilon)\%$ and $(100-\epsilon)\%$ where ϵ is a small value (0.1) used to make the estimate CDF more robust to outliers.

We use linear interpolation of the 100 CDF points to map a variable value x to a quantile and vice versa.