THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Bias correcting climate model simulations using unpaired image-to-image translation networks

Citation for published version:
Fulton, J, Clarke, BJ & Hegerl, G 2023, 'Bias correcting climate model simulations using unpaired image-to-image translation networks', *Artificial Intelligence for the Earth Systems*. https://doi.org/10.1175/AIES-D-22-0031.1

Digital Object Identifier (DOI):
10.1175/AIES-D-22-0031.1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Artificial Intelligence for the Earth Systems

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 10. Jun. 2023

Bias correcting climate model simulations using unpaired image-to-image translation networks

D. James Fulton, [a] Ben J. Clarke,[b] Gabriele C. Hegerl, [a]

[a] *School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom*
[b] *School of Geography and the Environment, University of Oxford, Oxford, United Kingdom*

ABSTRACT: We assess the suitability of unpaired image-to-image translation networks for bias correcting data simulated by global atmospheric circulation models. We use the UNIT neural network architecture to map between data from the HadGEM3-A-N216 model and ERA5 reanalysis data in a geographical area centred on the South Asian monsoon, which has well-documented serious biases in this model. The UNIT network corrects cross-variable correlations and spatial structures but creates bias corrections with less extreme values than the target distribution. By combining the UNIT neural network with the classical technique of quantile mapping, we can produce bias corrections that are better than either alone. The UNIT+QM scheme is shown to correct cross-variable correlations, spatial patterns, and all marginal distributions of single variables. The careful correction of such joint distributions is of high importance for compound extremes research.

1. Introduction

A large portion of research into the physical earth system is reliant on the use of general circulation models (GCMs). GCMs are used for weather prediction and for climate studies on timescales of days to years, although their seamless use for both is still rare. They have been central to informing policymakers through the International Panel on Climate Change (IPCC 2013).

The IPCC report relies on a multiplicity of different GCMs, developed by research centres spread across the world. The latest Climate Model Intercomparison Project (CMIP6) (Eyring et al. 2016) includes output from tens of GCMs. These differ in how the earth system is discretised, in their approximations for and inclusion of sub-gridscale processes, and even down to integration schemes, all of which introduce uncertainty that is reflected in the spread of results across models. The climate is highly chaotic, and so these differences can lead to detectable differences in the GCM outputs (Wang et al. 2014; Maher et al. 2018).

GCM outputs are used to assess the risks associated with different weather events, such as droughts, heatwaves, floods, and wildfire risk. To quantify these risks, researchers must decide which GCMs are fit for purpose. Persistent biases in climate models exist (Eyring et al. 2021) and need to be addressed by either selecting the best models or by correcting biases. The choice of models can have a quantitative and qualitative difference on estimated climate risks (e.g. Kirchmeier-Young et al. (2017); Herger et al. (2018)). A GCM which has a low bias compared to observations in one variable and one geographical location may have a large bias in another (Ridder et al. 2021). This makes compound risks (Leonard et al. 2014), such as simultaneous heatwaves and drought, even harder to assess as they involve multiple variables and may involve multiple geographical areas. Simultaneous crop failure in multiple regions of high agricultural output (Gaupp et al. 2020) is a risk of recent study. Particularly persistent biases exist in climate model simulated rainfall patterns, with many models exhibiting a double Inter Tropical Convergence zone (ITCZ) and misplaced monsoons (Wang et al. 2020; Tian and Dong 2020).

GCMs are improving, but in the interim, we must use their outputs effectively to understand our current climate and the potential effects of global warming. Therefore we must devise methods to optimally correct biases in the output of GCMs (Bellprat et al. 2019).

Recently, modern artificial neural network architectures have been developed which can be used for bias correction and statistical downscaling (Moghim and Bras 2017; Steininger et al. 2020; Le et al. 2020; Han et al. 2021; Wang et al. 2021), and which offer some theoretical advantages over classical techniques. In this paper, we will focus on unpaired image-to-image translation networks to perform bias correction between a GCM and observations.

These neural networks use layers of convolutional filters which are applied across multiple climate variables simultaneously. This gives these architectures the capacity to 'see' and therefore correct both spatial and cross-variable relations simultaneously. This is an advantage over current methods. These networks are presented in more detail later.

The simplest classical method of bias correction is to adjust the climatological mean, yet such a method may leave extremes still biased (e.g. Hanlon et al. (2015)). Another very commonly used method is quantile mapping (QM) (Cannon et al. 2015). This is a simple method where a

Corresponding author: James Fulton, james.fulton@ed.ac.uk

single value of a variable x_{GCM} obtained from the GCM at a spatial location and time of year denoted by coordinate θ, 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 $\mathcal{F}_{obs}^{-1}$.

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

This approach doesn't capture conditional relationships. The QM predicted value of $\hat{x}_{obs}$ doesn't 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 cyclones or the position of storms along the polar front.

Further, QM doesn't 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 a pixel is translated from a dry day in the GCM to a wet day in equivalent observations, the relationships between surface temperature and precipitation (Trenberth and Shea 2005) may not be preserved.

These are limitations that modern neural network architectures could be ideally suited to improve upon. Both of these features would be required to accurately correct for the presence and strength of teleconnections (Yuan et al. 2018; Stan et al. 2017). A specific example is shown in Maraun et al. (2017), where extreme precipitation in Piura, Peru only occurs during El Niño events in observations. These extreme rainfall events did not occur in the GCM outputs, so when QM was applied to the GCM data the extreme rainfall events no longer co-occurred with El Niño. El Niño events are primarily characterised by warm sea surface temperatures in the equatorial Pacific (Timmermann et al. 2018), and so conditional bias correction which takes spatial and cross-variable information is required.

Previous work has attempted to solve these issues with classical techniques, but incompletely. In Levy et al. (2013), the authors propose optimally stretching simulated precipitation fields to match precipitation patterns in observations. They found that when precipitation features were stretched onto the correct places, human attributable change in precipitation could be more easily detected, while misplaced features lead to poor fingerprints of the expected climate change. However, this technique uses monthly average precipitation and does not allow the use of daily data, which is important for studying extreme risks. It also can't 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.

We note that this problem of bias correcting GCM outputs is more generally the problem of mapping between two empirical distributions of multi-channel images without having any one-to-one corresponding pairs. This is precisely the problem description of unpaired image-to-image translation (Liu and Tuzel 2016).

Some previous work has used applied similar neural network architectures to this problem. François et al. (2021) apply an architecture based on CycleGAN (Zhu et al. 2017) to bias correct temperature and precipitation data in a region over Paris, but only for the winter season. Pan et al. (2021) develop a similar network architecture to bias correct precipitation data over the contiguous United States. They use the *dynamical variables* (sea level pressure, geopotential height, and specific humidity at 500 hPa) to aid the translations but do not bias correct these variables themselves.

In this paper, we extend on and complement these previous studies. We first introduce a different image translation method which has not been applied in the climate domain. We apply this method to bias correct simulations from a GCM of the South Asian Monsoon across five variables, using a GCM which has substantial biases in the spatial pattern of the monsoon, and using reanalysis data as a target for the correction. This is a larger geographical region and uses more variables than previous studies, and this region has a known and very significant physical bias in the GCM. We evaluate the method's use by comparing its performance to quantile mapping and explore using these two techniques in conjunction to better represent the monsoon and other extreme events in the region. We analyse the translation results focusing on relationships between different variables as well as spatial relationships.

2. A prime use for unpaired image-to-image translation networks

These architectures, such as UNIT (Liu et al. 2017), CycleGAN (Zhu et al. 2017), and AlignFlow (Grover et al. 2020) are neural networks that incorporate the architecture of generative adversarial networks (GANs) (Goodfellow et al. 2014). The aim of these techniques is to translate between images $\{\mathbf{x}_i\}_{i=1}^{N}$ in domain X and $\{\mathbf{y}_j\}_{j=1}^{M}$ in domain Y, without requiring corresponding pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$.

These networks were initially used to translate between images of summer and winter driving scenes, and between simulated city driving scenes and real city driving scenes (Liu et al. 2017; Zhu et al. 2017). Further, in Hao et al. (2021), they use these networks to make video game images look more photorealistic. In Shrivastava et al. (2017), the authors train a related *refiner* model to create more realistic-looking images of eyes from simulated 3D models. These are all examples of statistically bias correcting

spatial fields with multiple variables, as we aim to do with GCMs.

In order to map between GCM outputs and observations, the ability to translate without pairs is absolutely necessary. Imagine an idealised GCM which captures all of the physics of the earth system almost perfectly. However, due to discretisation, it accumulates errors when integrating a climate state forwards in time. If it was initiated with the exact observed climate state $\psi_{GCM}(0) = \psi_{obs}(0)$ we should expect that after approximately two weeks (Lorenz 1969; Zhang et al. 2019) the simulation will have diverged from the observations due to chaos. Therefore image pairs from the simulation and observations collected beyond two weeks have no relation to each other. If the GCM is not perfect and hence has a bias, then the initial state matched to observations $\psi_{GCM}(0) = \psi_{obs}(0)$ may be one which the GCM does not visit often, and the simulation will drift towards its own preferred states. This means we cannot use the image pairs collected during the first two weeks to bias correct the rest of the GCM simulation, i.e. cannot use pairs $\{\psi_{GCM}(t_i), \psi_{obs}(t_i)\}_{t_i<2 \text{ weeks}}$. This would be predicting outside the limits of our training data.

This lack of corresponding pairs makes bias correction distinctly different than perfect prognosis and statistical downscaling (Maraun et al. 2010), so we cannot use the machine learning techniques employed therein. In these settings, predictions are made for short lead times and so gathering corresponding pairs of GCM predictions and observations is possible. Although the application of deep learning to these tasks and statistical nowcasting (Steininger et al. 2020; Vaughan et al. 2022; Ravuri et al. 2021; Sønderby et al. 2020) has been developing recently, it would not be valid to simply use this learned mapping for times after around two weeks.

We note that Wang and Tian (2022) attempt to bias correct temperatures from GCMs by assuming synchronised pairs between observations and simulations over long time frames. However, they admit this to be a limitation of their proposed method and that the dynamics of the GCM may be distorted. They also do not thoroughly test their assumption.

a. A brief overview of unpaired image-to-image translations networks

The UNIT network is composed of multiple subcomponents. A more precise breakdown and diagram of the components in terms of layers is available in appendix A2, and also in the original paper (Liu et al. 2017). The main backbone of the network is composed of two variational autoencoders (VAEs) (Kingma and Welling 2013). The encoder E_X is a convolutional neural network that maps images from domain X to a region in latent space Z, i.e. $p(\mathbf{z}|\mathbf{x}) = E_X(\mathbf{x})$. There is a similar encoder E_Y that maps images in domain Y into a latent space which, if trained correctly, should be the same as Z. The main purpose of this network is to learn this shared latent space of the two image domains, and learn functions to map into and out of it to the two domains. Decoders G_X and G_Y map vectors from the shared latent space Z to images in the domains X and Y respectively, i.e. $\hat{\mathbf{x}} = G_X(\mathbf{z})$. After training is completed, the composition function $I_X(\mathbf{x}) = G_X(E_X(\mathbf{x}))$ should approximate the identify function (due to the VAE bottleneck, some information will be lost in the mapping and so the identity function cannot be exact). Also after training, the mapping $F_{X\to Y}(\mathbf{x}) = G_Y(E_X(\mathbf{x}))$ maps an image from domain X into its predicted equivalent image in domain Y. Similar statements are true for functions I_Y and $F_{Y\to X}$.

In training this network we minimise the value of a compound loss function with multiple components. The loss components L_X^{recon} and L_Y^{recon} are the expected reconstruction losses associated with autoencoders (AEs) and VAEs. L_X^{cycle} and L_Y^{cycle} are the cyclic reconstruction losses associated with translating from one domain to the other domain and back again. These losses are defined

$$\begin{aligned} L_X^{recon} &= \mathbb{E}_{\mathbf{x}\sim X}[|G_X(E_X(\mathbf{x})) - \mathbf{x}|] \\ L_Y^{recon} &= \mathbb{E}_{\mathbf{y}\sim Y}[|G_Y(E_Y(\mathbf{y})) - \mathbf{y}|] \\ L_X^{cycle} &= \mathbb{E}_{\mathbf{x}\sim X}[|F_{Y\to X}(F_{X\to Y}(\mathbf{x})) - \mathbf{x}|] \\ L_Y^{cycle} &= \mathbb{E}_{\mathbf{y}\sim Y}[|F_{X\to Y}(F_{Y\to X}(\mathbf{y})) - \mathbf{y}|]. \end{aligned} \tag{2}$$

Here, $\mathbb{E}_{\mathbf{x}\sim X}[l]$ means the mean of the component l for data samples $\mathbf{x}$ sampled from X. Practically, this is estimated as the average of l over samples in a training batch. These losses encourage the latent space representation to encode as much information as possible about the images. They also ensure that the encoding is consistent across domains and that translating an input from one domain to the other and then back again reconstructs the same input. We use the L1-norm for these losses as it has been shown to support sharper translations (Zhu et al. 2017) in similar networks. We modify the L1 loss to weight the different channels with respect to each other (see appendix A3).

The following loss components are also used:

$$\begin{aligned} L_X^{KL} &= \mathbb{E}_{\mathbf{x}\sim X}[KL(E_X(\mathbf{x}))] \\ L_Y^{KL} &= \mathbb{E}_{\mathbf{y}\sim Y}[KL(E_Y(\mathbf{y}))] \\ L_X^{KL-cyc} &= \mathbb{E}_{\mathbf{x}\sim X}[KL(E_Y(F_{X\to Y}(\mathbf{x})))] \\ L_Y^{KL-cyc} &= \mathbb{E}_{\mathbf{y}\sim Y}[KL(E_X(F_{Y\to X}(\mathbf{y})))]. \end{aligned} \tag{3}$$

The loss components L_X^{KL} and L_Y^{KL} are the Kullback–Leibler (KL) divergences associated with the samples from domains X and Y encoded into the shared latent space Z. These are standard loss components used to train VAEs (Kingma and Welling 2013). The function $KL(\mathbf{z})$ computes the KL divergence between a multivariate Gaussian distribution with mean $\mathbf{z}$ and another with mean $\mathbf{0}$,

both with variance of **1**. L_X^{KL-cyc} and L_Y^{KL-cyc} are associated with the KL divergence of samples translated to the opposite domain and then encoded back into the shared latent space.

These KL divergence losses encourage the images to have a Gaussian distribution when encoded into the shared latent space. During training, a random perturbation is added in the latent space encoding during training so that $\mathbf{z} = E_X(\mathbf{x}) \mapsto E_X(\mathbf{x}) + \eta$ where η is a random vector sampled from a multivariate Gaussian distribution with unit diagonal variance. This so-called reparameterisation trick is required to train VAEs (Kingma and Welling 2013) and ensures that the learned latent space Z is smooth - i.e. that images that are similar in domain X will be translated to a similar region in domain Z.

To train the translation network, we also train two adversarial discriminator networks. Such discriminator networks have been used in recent work on short-term weather forecasting (Ravuri et al. 2021). A discriminator D_X is trained to predict the probability that an image comes from the domain X or whether it was created by the conditional generator $F_{Y\to X}$. Similarly, D_Y is trained to predict whether images in domain Y are real. These discriminator networks are trained simultaneously with the translation network. The following loss components are used to train the translation network functions $F_{Y\to X}$ and $F_{X\to Y}$:

$$\begin{aligned} L_X^{GAN} &= \mathbb{E}_{\mathbf{x}\sim X}[|1 - D_Y(F_{X\to Y}(\mathbf{x}))|^2] \\ L_Y^{GAN} &= \mathbb{E}_{\mathbf{y}\sim Y}[|1 - D_X(F_{Y\to X}(\mathbf{y}))|^2]. \end{aligned} \tag{4}$$

These loss components encourage the translations of the fields to look realistic in each domain. Any obvious imperfections or distortion of the distribution of images $\{F_{Y\to X}(\mathbf{y})\}_{\mathbf{y}\sim Y}$ should be penalised by the discriminator (Goodfellow et al. 2014) via L_X^{GAN}. The discriminator D_X is itself trained to minimise the loss function

$$\begin{aligned} L_X^{discrim} = \mathbb{E}_{\mathbf{y}\sim Y}[|D_X(F_{Y\to X}(\mathbf{y}))|^2] \\ + \mathbb{E}_{\mathbf{x}\sim X}[|1 - D_X(\mathbf{x})|^2] \end{aligned} \tag{5}$$

and the loss for D_Y is similar. This is the same loss function as is used in least-square GANs (Mao et al. 2017) and has been shown to be more stable in training than traditional GANs.

The full loss function for the translation network is

$$\begin{aligned} L = \lambda_{rec}(L_X^{recon} + L_Y^{recon}) + \lambda_{cyc}(L_X^{cycle} + L_Y^{cycle}) \\ + \lambda_{KL-rec}(L_X^{KL} + L_Y^{KL}) + \lambda_{KL-cyc}(L_X^{KL-cyc} + L_Y^{KL-cyc}) \\ + \lambda_{GAN}(L_X^{GAN} + L_Y^{GAN}) \end{aligned} \tag{6}$$

where the scalars λ_{rec}, λ_{cyc}, λ_{KL-rec}, λ_{KL-cyc}, and λ_{GAN} were set to the values used in the original work introducing UNIT. More details about these losses and the UNIT hyperparameters are presented in appendix A3.

b. What do bias corrected simulations represent?

It is important to consider what the translated sequence $F_{GCM\to obs}(\psi_{GCM}(t))$ represents, and we refer to Ehret et al. (2012) for a more thorough discussion. $\psi_{GCM}(t)$ and thus $F_{GCM\to obs}(\psi_{GCM}(t))$ is driven dynamically by the time evolution operator of the GCM, but each image $F_{GCM\to obs}(\psi_{GCM}(t))$ is mapped individually to look like it came from the observations. If we set $\psi_{GCM}(0) = F_{obs\to GCM}(\psi_{obs}(0))$ and evolve each of these states forward in time with their own time evolution operator, then we would still expect that $F_{GCM\to obs}(\psi_{GCM}(t)) \neq \psi_{obs}(t)$. Although each image from $F_{GCM\to obs}(\psi_{GCM}(t))$ should be realistic compared to observations, the entire sequence needn't be. An example of this inconsistency could come when bias correcting wind velocities and precipitation. The wind velocities may be debiased such that it should change the velocity of a precipitation system, but in consecutive frames the precipitation pattern may not reflect this.

So when using GCM output, we must assume that the GCM has realistic time dynamics, which would not be drastically affected by our moderate bias correction. This also emphasises that bias correction is no cure for a poorly performing GCM and that capturing the physics of the climate system is key.

3. Bias correcting the South Asian Monsoon

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 (`All-Hist/est1`). In this dataset, the ocean temperatures are prescribed to their observational estimates and emissions are set to the historical record. Therefore only the atmosphere component of the model is run.

We attempt to bias correct the HadGEM3 historical recreation data with respect to the ERA5 reanalysis data (Hersbach et al. 2020). Both of these datasets are daily data fields. We choose this over monthly data as extreme events like heatwaves and floods occur on the timescale of days to weeks.

We limit the geographical extent to the area bounded by 8°S-30°N 44°E-121°E. This region was chosen to capture the South Asian monsoon. Many GCMs, including HadGEM3, have a large bias in simulating the South Asian monsoon (Bollasina and Ming 2013; Ashfaq et al. 2017), which is usually placed too far south over the Indian Ocean and leaves the landmass drier than reality (shown in figure 1). This bias remains when the model is run with prescribed sea surface temperatures. We chose this region as a hard case to solve for bias correction. We consider the

daily 2-metre mean, minimum, and maximum temperature; accumulated precipitation; and mean 500hPa geopotential height at all grid locations. The first 4 of these variables are important for climate impact studies and are of primary interest. The geopotential height is a dynamical variable which can aid the accuracy of bias correction of the other variables (Pan et al. 2021), although we allow the network to correct this variable also.

Before bias correction, we conservatively regrid (Jones 1999) the ERA5 data onto the coarser grid of the HadGEM3 data, which has resolution of 0.56°latitude and 0.83°longitude. This gave a region of size 68×92 gridpoints. We also limit the two datasets to the time period in which they overlap; this is 1979 to 2013 inclusive, giving us 35 complete years and approximately 13 000 daily fields. We split this data into train and test sets, using the odd-numbered years for training each method and the even-numbered years to test on. This rather extreme split of using nearly 50% of the data test was required to perform adequate analysis of the spatial and cross-variable statistics of the results. All figures to follow were based on the test set, which was not used in training or choosing parameters for any method.

We applied quantile mapping (equation (1)) to both datasets by fitting a 100-point empirical cumulative distribution function to each gridpoint for each month of the year and for each variable. This equates to ~ 75 million points estimated from the data in order to perform the bias correction with quantile mapping. For comparison, the UNIT translation network had ~ 38 million parameters. The form of QM we use is generally applied in situations where the data is stationary, i.e. where there is no distribution shift due to global warming. The neural network approach we choose similarly assumes stationarity. Since we limit our datasets to only a 35 year period and split alternative years into train and test data, we do not expect this assumption to have a significant impact on the results we show. Especially as we are considering daily variability rather than monthly averages, so changes in the mean are relatively small compared to variability. The extensions that are added to quantile mapping to allow it to approximately debias data that is not stationary (Cannon et al. 2015) could be applied to the UNIT neural network approach with little modification.

In the following section, we compare these translations and the original datasets. In particular, we study in detail how they address the large biases in simulations of the South Asian monsoon. To test this more broadly, we also consider their performance in correcting a variety of other extreme events of societal relevance to the region.

4. Results

Early in our experimentation with the UNIT network, we noticed that it was biased against producing extreme values in its translations. This limitation may be due to the fact that UNIT inherits features of its architecture from GANs, and GANs are known to reduce the distribution at its boundaries (Dionelis et al. 2020; Bau et al. 2019; Arora and Zhang 2017). Notably, Ravuri et al. (2021) found that using a conditional GAN for precipitation nowcasting also reduced extreme values in precipitation fields. UNIT also inherits features VAEs, which blur the image reconstructions (Snell et al. 2017) and thus reduce the extreme values. The VAE blurring is a result of the VAE architecture. These extreme values are important for climate research, so we propose to improve the UNIT network performance by following it up with QM. In the combined UNIT+QM method, we take the trained UNIT network and use it to translate the entire HadGEM3 dataset; then we train QM between this dataset and the ERA5 dataset. The same trained UNIT instance is used in both the UNIT and the UNIT+QM results presented.

a. South Asian Monsoon

Figure 1 shows the means and biases of precipitation in the peak monsoon months (June-September) in ERA5, HadGEM3, and the UNIT-corrected data. This figure shows the key bias present in HadGEM3's simulated monsoon. It also shows the effects of the reduction in extremes generated by UNIT, where the average precipitation is reduced everywhere. The results for QM and UNIT+QM are not shown. By the definition of QM, their biases under this plot would be near-zero (non-zero only due to sampling error between train and test data), and their means would be the same as the ERA5 field. We confirmed this by plotting.

Figure 2 shows three non-consecutive days from the HadGEM3 validation data and their bias corrections using the three different methods. These examples show that UNIT can coherently bias correct the structure of the fields and shows the consequences of QM's lacking in this ability. In example day 1, it removes the precipitation in HadGEM3 which is characteristic of the HadGEM3 monsoon bias. It also removes the associated minor depression in geopotential height. QM maintains the spatial structure of the fields but simply adjusts the intensity; hence the QM field is left with a muted form of the HadGEM3 monsoon precipitation. In the QM data, the spatial structure is still intact, and the field produced is unrealistic. In example day 2, UNIT removes the monsoon precipitation feature and a stronger geopotential height anomaly. In this case, the dynamical conditions simulated by the GCM are truly biased and do not happen in the observations. In this case, it makes sense to bias correct the dynamical feature of geopotential height as well as the precipitation. This is an advantage over previous studies which assume the dynamical features are correct and use them without scrutiny to aid bias correcting other variables. Example day 3 shows a case where UNIT adds a precipitation system over the

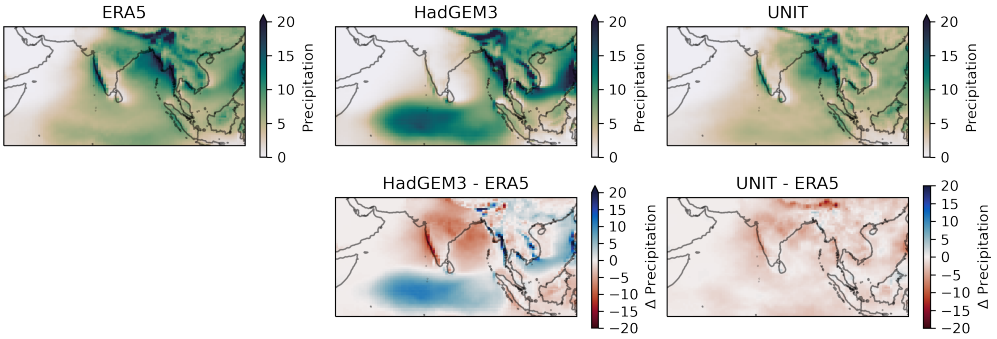

FIG. 1. Mean precipitation (mm/day) during peak monsoon months (June-September) for each dataset and their bias with respect to ERA5.

Bay of Bengal, showing the UNIT network can add meteorological features as well as remove them.

In figure 3, we show the distribution of daily accumulated precipitation from a single gridpoint. This gridpoint is located on the southern tip of India. The datasets of the three bias correction techniques of UNIT, UNIT+QM, and QM are created by taking this HadGEM3 data and debiasing it with each method.

This figure shows the typical behaviour of these bias correction methods. There is a significant bias between the ERA5 and HadGEM3 datasets, with HadGEM3 showing much more light drizzle (Takahashi et al. 2021) with precipitation values below 0.1 mm/day. Since this is a marginal distribution, QM is expected to perform well. If the datasets were infinitely large, then by definition, QM would perform perfectly here (Maraun et al. 2017). The same is true of the combined UNIT+QM method. The UNIT translation performs reasonably well on the marginal distribution, although it reduces the occurrence of both high and low extremes and shifts more of the distribution towards ERA5's central peak. Note that QM is explicitly designed to match these one-dimensional distributions at each gridpoint, whilst UNIT matches the distributions only as an emergent feature of optimising its loss function.

Figure 4 shows the joint distribution between temperature and precipitation at the same grid location at the southern tip of India (we also sampled several others and found similar qualitative results). This plot shows that although UNIT underdisperses the data, it can capture the shape of the joint distribution well. We see that quantile mapping does not correct the joint distribution between precipitation and temperature well. UNIT+QM performs the best of the three techniques and captures both the joint distribution and the dispersion of the data. The UNIT step in UNIT+QM corrects the correlation structure, whilst the QM part corrects the marginal distributions.

In order to quantitatively assess the similarities of these joint distributions, we use Jensen–Shannon (JS) divergence. This is a common statistical measure of how different two distributions are; a lower value is better. We estimate this empirically by binning the 2D (temperature and precipitation at a gridpoint) data into bins $\{\xi_i\}$ and counting the frequency in each bin. The JS divergence is computed

$$D_{JS}(P||Q) = \frac{1}{2}\left(D_{KL}(P||M) + D_{KL}(Q||M)\right) \quad (7)$$

where P and Q are the two distributions in question, and $M = (P+Q)/2$ is their mean. $D_{KL}(P_1||P_2)$ computes the KL divergence between distributions P_1 and P_2, and is estimated empirically via

$$D_{KL}(P_1||P_2) = \sum_{\xi_i} P_1(\xi_i) \log\left(\frac{P_1(\xi_i)}{P_2(\xi_i)}\right). \quad (8)$$

We compute the JS divergence between each of the distributions in figure 4 and the ERA5 distribution. We form a 2D histogram by splitting the data into 20 bins of equal width in both the temperature and precipitation$^{1/4}$ dimensions. We calculate JS divergences of 0.189, 0.053, 0.020, and 0.025 for the HadGEM3, UNIT, UNIT+QM ,and QM distributions with respect to the ERA5 dataset. The results were comparatively similar for a range of reasonable choices of the number of bins - see appendix A4. These JS divergence values are all significantly distinct from each other - see appendix A5.

Continuing this analysis, figure 5 shows maps of the JS divergence of precipitation and temperature at all gridpoints. UNIT performed poorly due to how it underdisperses the data, whilst UNIT+QM performed the best. UNIT+QM appears to adjust for the correlations in the data whilst also maintaining the dispersion.

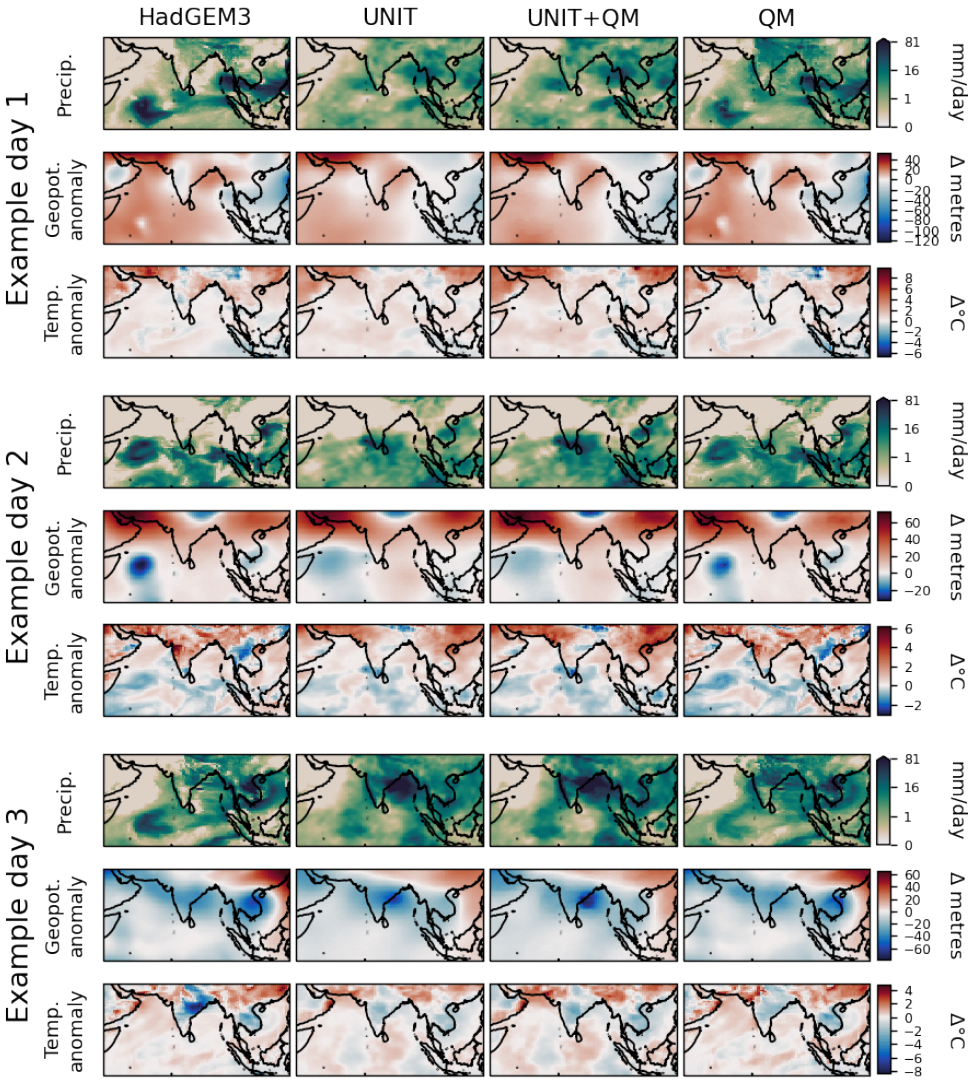

FIG. 2. Three samples of daily fields from the HadGEM3 data and their bias corrections with three different methods. The fields shown for each day are the accumulated precipitation (plotted with power-law colour intensity), 500hPa geopotential height, and 2-metre temperature. The geopotential height and temperature are shown as anomalies and have had their monthly means subtracted. Each row of the figure shares the colourbar plotted at its end.

In order to assess the spatial structures of the data and translations, we start locally, looking at the joint distribution of precipitation at two neighbouring gridpoints. We choose the same familiar gridpoint on the southern tip of India and the cell directly eastwards of it. Again, in figure 6, we see that UNIT+QM performs the best, improving the

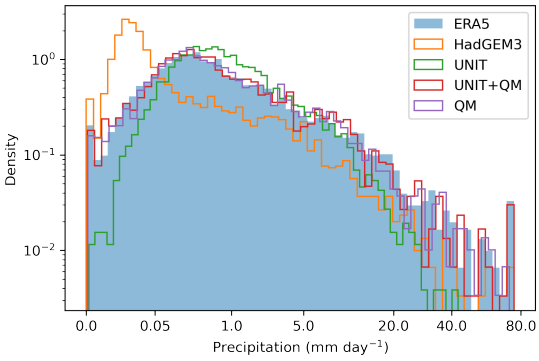

FIG. 3. Distributions of daily accumulated precipitation from a single gridpoint located at the southern tip of India (centred at 8.6°N 77.9°E). Five distributions for the same location are plotted from the five different data sources. The normalised histogram is on an x-scale of the fourth root of precipitation so that the differences in the distributions can be seen. It is also clipped at an upper value of 75 mm/day.

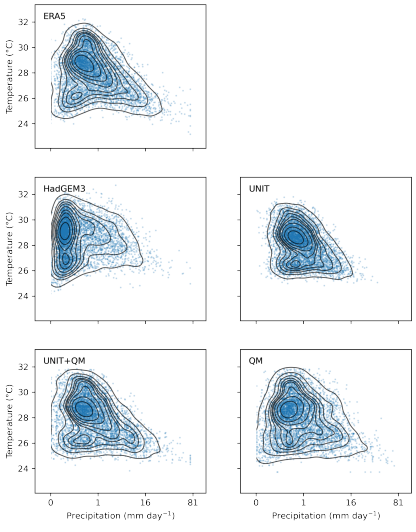

FIG. 4. Joint distributions of daily accumulated precipitation and daily mean temperature from a single gridpoint located at the southern tip of India (centred at 8.6°N 77.9°E) for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x-scale in all subplots is the 4th root of precipitation. This was chosen so that the differences in the distributions can be seen. It is also clipped at an upper value of 75 mm/day.

structure of the joint distribution and also the dispersion of the data. UNIT reduces the dispersion, and QM does not adequately correct the correlation structure.

We wish to assess whether the daily fields produced by each method of bias correction are realistic and whether the spatial structures observed are biased. In order to do so, we use the mean structural similarity index measure (SSIM) (Wang et al. 2004), which is a metric designed to measure how structurally similar two images are. SSIM takes into account the difference in mean values, the contrast in the image between high and low values, and the structure in the image - i.e. whether high and low values are in the same locations. As is common, we use the mean SSIM, and these comparisons are made for each gridpoint using an 11×11 pixel Gaussian sliding window. Then the average is taken over the image. A larger value shows a closer match between images, with a maximum possible similarity of 1.

In order to assess whether the datasets produce fields that are spatially realistic, we propose running the following computation:

Take one daily field of a single variable from a dataset $\chi \in \{\text{HadGEM3, UNIT, UNIT+QM, QM}\}$ and find the daily field from the ERA5 test dataset which is most similar using SSIM. This is a similar matching algorithm to flow-analogues as developed in Yiou et al. (2007). Store this optimal value of SSIM. Repeat for all days in dataset χ.

Then we plot the distribution of these best-match SSIM values. Figure 7 shows the distributions calculated by matching on geopotential height, mean daily temperature, and the fourth root of precipitation. We chose to use the fourth root of precipitation as the raw marginal distribution has a very long tail, and we wish to avoid extreme values dominating the comparison. Instead, we are trying to focus on the overall spatial structure. In the figure, the further the distribution is to the right the more similar the fields in the dataset were to the ERA5 fields. In the figure, we also include the results of comparing fields from the ERA5 train dataset to the test dataset to set a baseline and aid interoperability. The results in this figure tell us that the UNIT translations produced fields that were closer to ERA5 fields than the other translation methods, with UNIT+QM having the next most similar fields.

There are some limitations to this matching method which should be noted and can explain why UNIT outperforms UNIT+QM here. The UNIT translations are underdispersed, as we have seen in previous figures. This means each UNIT field lies more towards the centre of the ERA5 distribution. Fields from any dataset χ which lie towards the centre of the ERA5 distribution are more likely to find an ERA5 field that matches them closely than fields from χ which are towards the tails of the ERA5 distribution. This is simply because there are more ERA5 fields to choose from in dense regions. The best-match SSIM between a field $\mathbf{x}$ and the ERA5 data is positively associated with the density of the ERA5 data $p_{\text{ERA}}(\mathbf{x})$ around the location $\mathbf{x}$. This explains why UNIT performs best in this analysis. We

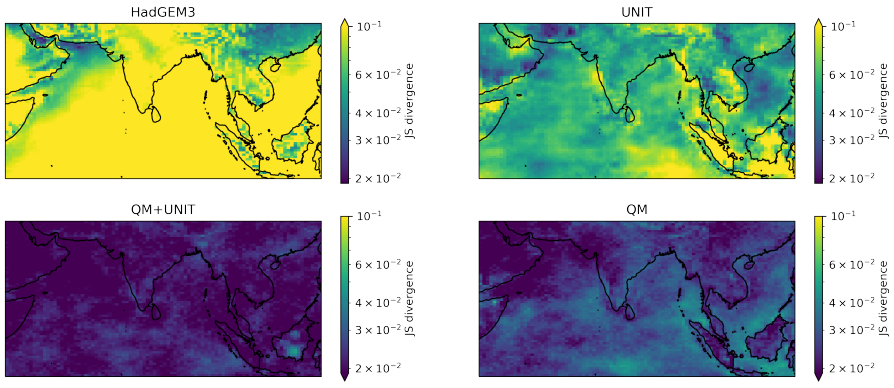

FIG. 5. A map showing estimated of JS divergence (equation (7)) of the HadGEM3, UNIT, UNIT+QM and QM datasets with respect to the ERA5 dataset. The JS divergence is calculated at each gridpoint between the 2-dimensional precipitation and temperature distributions using 20 bins in each dimension.

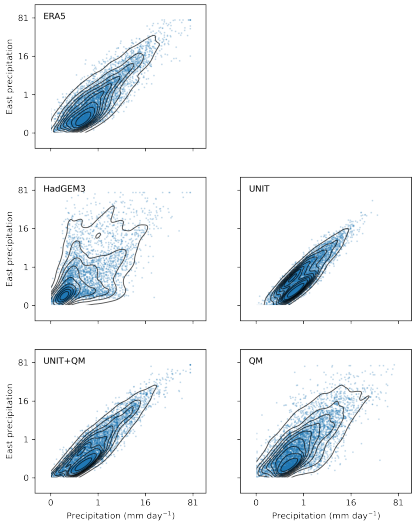

FIG. 6. Joint distributions of daily accumulated precipitation at a gridpoint located at the southern tip of India (centred at 8.6°N 77.9°E) and its neighbouring cell directly eastwards (centred at 78.8°E) for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x and y-scale in all subplots is the 4th root of precipitation. This was chosen so that the differences in the distributions can be seen. It is also clipped at an upper value of 75 mm/day.

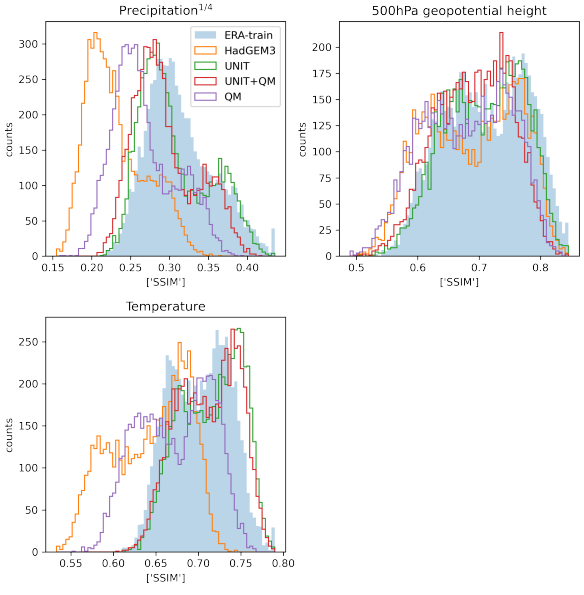

FIG. 7. Distributions of the SSIM between each dataset field and its most similar ERA5 field. Repeated for the mean 2 metre temperature, mean 500hPa geopotential height and the fourth root of daily precipitation.

note that UNIT+QM performed second best and the distribution of UNIT+QM is not underdispersed like UNIT. In appendix A9, we perform extra analysis which confirms that UNIT's exaggerated performance in this analysis is likely due to underdispersion.

We repeated the computation above using mean absolute difference and mean square difference as alternative metrics to SSIM and achieved qualitatively similar results.

Furthermore, we examine if the characteristics of these methods hold when we spatially aggregate the data. We aggregate to a few river basins in this region, using data

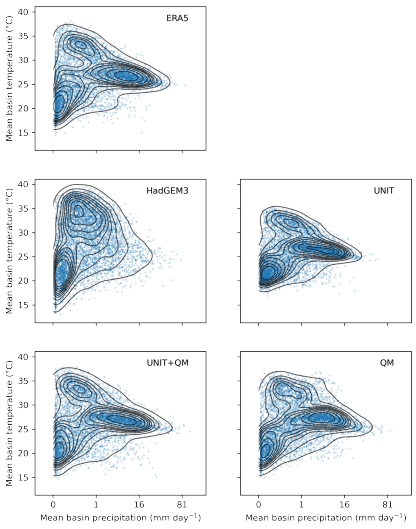

FIG. 8. Joint distributions of the mean daily accumulated precipitation and daily mean temperature aggregated over the Mahanadi river basin for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x-axis has power law scaling. This was chosen so that the differences in the distributions can be seen. It is also clipped at an upper value of 75 mm/day.

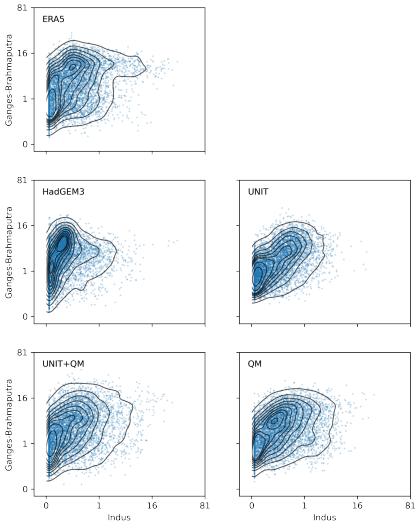

FIG. 9. Joint distributions of mean daily accumulated precipitation across the Ganges-Brahmaputra and the Indus basin for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x-axis and y-axis have power law scaling. This was chosen so that the differences in the distributions can be seen. It is also clipped at an upper value of 75 mm/day.

from the World Bank data catalogue (The World Bank 2019) to define the basin boundaries. We aggregate to the portion of these basins that lie within the spatial extent of our data, and therefore two of the basins are clipped (see appendix A6 for plotted basin masks).

Figure 8 shows the joint distribution between the spatial mean precipitation and temperature in the Mahanadi river basin. This basin was chosen as there is a big difference between the HadGEM3 and ERA5 joint distributions, and therefore the bias correction method has a lot to correct. Debiasing this distribution involves correcting spatial correlations (since it is a spatial aggregate) and cross-variable correlations simultaneously. Once again we can see that UNIT+QM performs well, correcting the cross-variable correlations and the dispersion.

Finally, in figure 9, we plot the joint distribution between the spatial mean of precipitation in the Ganges-Brahmaputra and Indus basins. These are two important basins, both impacted by the South Asian monsoon and geographically separated. This kind of long-range bias correction could be important for assessing the risks of multiple breadbasket failure. None of the methods performed particularly well in this analysis.

b. Other Extremes

In the second part of the results section, we study the performance of these bias correction techniques for other extreme events of societal relevance in the South Asia region. In particular, we study whether the combined UNIT+QM improves the representation of physical extremes. We consider three cases, testing the cross-correlation of different variables and spatial areas.

First, we analyse the relationship between temperature and pressure on the hottest day of each year over a region in central India (8°-28°N 72°-85°E). We select the day with the highest average maximum daily temperature over this region in each year. South Asia experiences some of the most extreme humid heat on the planet (Raymond et al. 2020), and combined with high population density and vulnerability to such hazards, this poses a severe threat to human health (Im et al. 2017). Furthermore, even in spite of the cooling influence of anthropogenic aerosols over the region, recent events such as the heatwave of 2015 have been amplified by climate change (Wehner et al. 2016). Thus, the understanding of such events is becoming increasingly pertinent.

Figure 10 shows the joint distributions between daily maximum temperature and 500hPa geopotential height at all gridpoints across all annual hottest days. In this case,

Model	Hot days (fig. 10)	Wet periods (fig. 11)	Two basins SPI (fig. 12)
HadGEM3	0.1383	0.2531	0.1417
UNIT	0.0712	0.1856	0.1100
UNIT+QM	0.0442	0.0310	0.0811
QM	0.0462	0.0511	0.0844

TABLE 1. JS divergence values for each model relative to ERA5 across three joint distributions of extreme events: temperature and pressure on the hottest days of the year over central India (figure 10); temperature and precipitation over the wettest continuous 30 days of each year over the Ganges-Brahmaputra basin (figure 11); the Standardised Precipitation Index in the Ganges-Brahmaputra and Indus river basins (figure 12). These JS divergence values are all significantly distinct from each other - see appendix A5

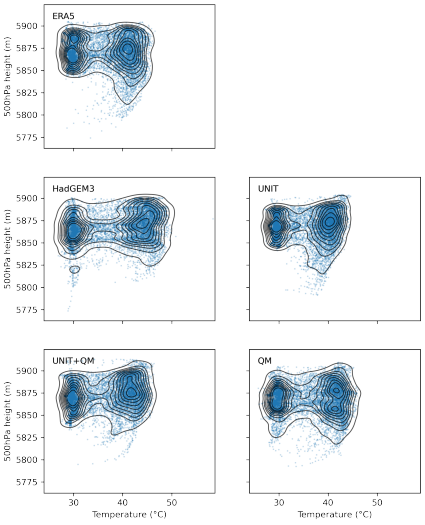

FIG. 10. Joint distributions of the daily maximum temperature and 500 hPa geopotential height at each gridpoint over central India, bounded by 8°-28°N 72°-85°E, on the hottest day of each year (defined by the spatially averaged daily maximum temperature) for the five different datasets. Each point is from a single day and gridpoint, and the contours show the joint density as estimated by kernel density estimation.

UNIT+QM appears to perform the best, though with only 17 annually hottest days to analyse, so this could be quite sensitive to noise.

Results of the JS divergence for all of the various extreme distributions are presented in table 1. In this first case, UNIT+QM performs best, though not significantly better than QM.

Second, we analyse the relationship between temperature and precipitation over the Ganges-Brahmaputra river basin during the wettest continuous 30-day period of each year. Flooding, driven partly by rainfall excesses or

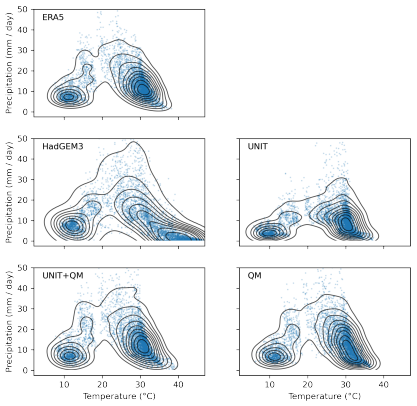

FIG. 11. Joint distributions of the daily maximum temperature and accumulated precipitation at each gridpoint over the Ganges-Brahmaputra river basin over the wettest 30-day period each year, defined by the spatially-averaged accumulated precipitation, for the five different datasets. Each point is from a single gridpoint, and the contours show the joint density as estimated by kernel density estimation.

deficits, has severe impacts on the region. Between 2000-2020, floods caused over USD 100 billion in damages and the deaths of more than 49 000 people - almost half of the global flood mortality in the period - according to disaster database EM-DAT (Guha-Sapir et al. 2014). While this is largely driven by the behaviour of the monsoon, impactful rainfall extremes also occur outside of this season and are also influenced by anthropogenic climate change (Rimi et al. 2019). When widespread flooding occurs, simultaneous extreme heat may result in compounded impacts, such as through disrupted water supplies and water- and insect-borne disease (Levy et al. 2016; Moors et al. 2013).

Figure 11 shows the joint distributions between the 30-day averages of daily maximum temperature and precipitation, for the wettest 30-day period of each year, at all gridpoints in the Ganges-Brahmaputra basin and for all years.

Across the basin there are two temperature regimes due to the change in altitude from low-lying Bangladesh and northeast India to the Tibetan Plateau. Over the low-lying warmer region, HadGEM3 shows severe discrepancies in rainfall, likely owing to its poor monsoon representation and its tendency to drizzle. While UNIT underdisperses the translated data, both QM and UNIT+QM perform well. The results of JS divergence (table 1) show that combining the two techniques captures the reanalysis data most effectively, with UNIT+QM significantly outperforming the other methods.

Third, we analyse the relationship between precipitation in the Indus and Ganges-Brahmaputra river basins, following on from earlier analysis (figure 9), to measure

not only spatial but also temporal variation between the basins. Combined, these rivers are of crucial importance to agriculture and thus food security in the region. Accurately simulating the possibility of co-occurring flooding or flash drought in the two is therefore pivotal. To measure excesses and deficits in rainfall over time, we use the Standardised Precipitation Index, which is commonly used to monitor drought and flood hazards (Chandrasekara et al. 2021; Aadhar and Mishra 2017; Tirivarombo et al. 2018). This index is defined by

$$SPI_T = \frac{P_T - P_*}{\sigma_{P_T}}, \tag{9}$$

where P_T is the mean precipitation over a time period length T. P_* is the mean value of all P_T across the dataset, and σ_{P_T} is its standard deviation. 30 days is used as the baseline time period to represent medium-term extremes in precipitation, relevant for both flooding and subseasonal flash drought (Christian et al. 2021; Mishra et al. 2021).

Figure 12 shows the joint distributions between co-occurring SPI values in each river basin for each model and bias correction system. We note that the data points in this plot originate from a 30 day rolling time window instead of selecting independent 30 day periods. To turn our rolling window data into independent samples would mean selecting 1/30th of the time indices, i.e. indices $\{30i + s\}_{i\in[0,1,2,3,...]}$ where s is the starting index $0 < s < 29$. However, we are only interested in estimating the joint density, and all starting indices s are equally valid. So we decide not to filter to independent periods in the density plot and effectively marginalise over s when performing kernel density estimation in the figure. However, we note that the datapoints are oversampled, so we must avoid overinterpreting point clusters. An independently-sampled version of this plot is presented in the appendix (figure A8) and shows the same overall structure in each case.

HadGEM3 underestimates the co-occurrence of very wet events where the Ganges-Brahmaputra SPI >1 and Indus SPI >2. All translations perform reasonably well at correcting for these edge cases, but again UNIT+QM performs best. This is evident from the JS divergence values and visually; while UNIT underdisperses the data, using UNIT+QM provides an accurate correction for more extreme events.

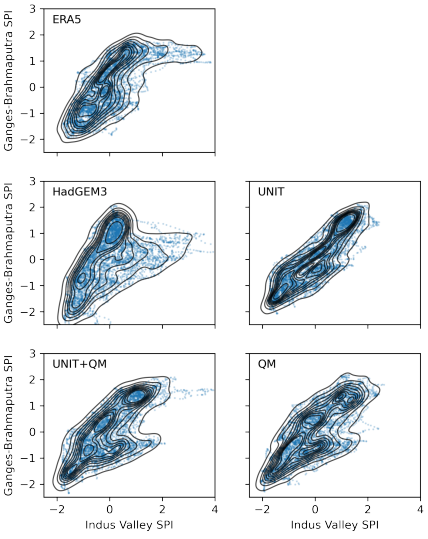

FIG. 12. Joint distributions of the Standardised Precipitation Index at each gridpoint over the Ganges-Brahmaputra and Indus river basins using a 30-day time window for the five different datasets. Each point is from a single gridpoint, and the contours show the joint density as estimated by kernel density estimation.

5. Conclusion

In this study, we examined the appropriateness of unpaired image-to-image translation networks to bias correct climate data. We found that the UNIT neural network architecture was not sufficient by itself for bias correcting data. This was because it reduced the dispersion of the data and therefore led to less extreme values than expected. We showed that the bias correction produced by UNIT did have desirable properties, such as coherently bias correcting cross-variables correlations and spatial structures. We proposed to combine the UNIT translation with quantile mapping, a more traditional bias correcting technique - a combination which is similar to previous work (François et al. 2021). We found that applying these techniques in sequence made up for shortcomings in each. UNIT was able to bias correct the spatial and cross-variable correlations, whilst QM corrected UNIT's tendency to underdisperse the data.

When applied to bias correct the daily minimum, mean, and maximum temperature, the precipitation, and geopotential height, we found that UNIT+QM was able to simultaneously correct all variables. This is an advantage of previous work where the dynamical feature of geopotential height is assumed to be correct and unbiased and thus is used as a key part of regularising the bias correction (Pan et al. 2021). In our method, we leave it to the discriminator network to ensure that the temperatures and precipitation are consistent with each other and the dynamical variable. This allows the network to bias correct the dynamical variable where it is appropriate.

Designing bias correction methods such as UNIT+QM, which can bias correct many variable joint distributions, is crucial to study and make risk assessments of compound extreme events. Such cross-variable corrections are also important when the output of GCMs are used as boundary

conditions of regional climate models (White and Toumi 2013). We have shown that incorporating modern machine learning methods alongside classical techniques can provide us with more powerful tools for bias correction.

Further work on this topic may be needed in several aspects. First, to consider the performance of this approach compared to newer developments in statistical bias correction such as multivariate quantile mapping (Cannon 2018), alongside computational demands, as well as how the approaches could be further combined to address the issues inherent in both. Second, to consider bias correction of non-stationary data. We note that this is also an ongoing challenge for classical techniques. Many of the extensions to classical techniques used for non-stationary data, such as detrending either additively or multiplicatively (Cannon et al. 2015), could also be used with UNIT+QM.

Acknowledgments. D.J.F. was supported by a NERC Doctoral Training Partnership grant NE/L002558/1. B.J.C. was supported by a NERC Doctoral Training Partnership grant NE/L002612/1. G.C.H. was funded by NERC large grant GloSAT (NE/S015698/1).

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

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modelling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF.

We would also like to thank Simon Tett, Massimo Bollasina, and Friederike Otto for helpful discussions.

Data availability statement. All the data used in this work comes from the Climate of the 20th Century Plus (C20C+) Project (Folland et al. 2014), and is freely available. Code will be made available in a public repository upon publication.

APPENDIX A

A1. Preprocessing the data

As is ubiquitous in training machine learning (ML) models, we preprocessed the data before training the UNIT network.

The daily mean temperature fields originally in Kelvin were converted to Celsius and then divided by the mean of the standard deviation of temperature across all latitudes and longitudes. This maintained the physically-significant value of 0°C whilst avoiding excessive gradients when updating the model parameters. The daily maximum and minimum temperatures were converted to degrees Celsius above or below the daily mean temperature. This allowed us to choose ReLU activation functions which enforced $T_{min} < T_{mean} < T_{max}$. They were also divided by the spatially mean standard deviation in temperature.

Precipitation values are extremely skewed and non-Gaussian. This can be challenging for an ML model to learn. In keeping with ML best practices, we transformed the data so that the distribution became closer to Gaussian. We found that taking the fourth root of the daily accumulated precipitation appeared reasonably dispersed. Then we divided the data by its standard deviation. We note that the transforms used needn't make the data exactly Gaussian, merely less extremely distributed.

We simply standardised the 500hPa geopotential height by subtracting the mean and dividing by the standard deviation. The means and standard deviations used were not calculated point-wise, instead a single value was used globally.

A2. UNIT network structure

Figure A1 shows a diagram of the UNIT architecture used here. We use colour and shape to represent different layers and sub-components of the network.

The encoders $E(\mathbf{x})$ are comprised of 7 convolutional layers with convolutional downsampling to reduce the spatial dimension. The decoders $G(\mathbf{z})$ are comprised of 9 convolutional layers with nearest neighbour upsampling to increase spatial dimension. Nearest neighbour upsampling was used over transpose convolutions to avoid checkerboard artefacts Odena et al. (2016) which are present in the original UNIT network. Some of these layers are organised into residual skip connections (He et al. 2016), which allow more routes for gradient backpropagation to earlier layers.

We use a multi-scale discriminator in each domain. Isola et al. (2017) found that such multi-scale discriminators produce images which have more realistic large-scale features and sharper small-scale features. The multi-scale discriminator is composed of 3 discriminator components which asses the plausibility of the translations at separate spatial scales. We use spatial scales which are 1×, 2×, and 4×

coarser than the input image. Each discriminator component had 6 convolutional layers. For the discriminators coarser than the original image, the input image is coarsened by mean pooling.

Batch normalisation and leaky ReLU activation functions were used throughout these components, as shown in the diagram. The final layer of the decoder $G(\mathbf{z})$ used different activation functions for the different channels. The mean temperature and geopotential height used none (i.e. $x \mapsto x$); the precipitation, and maximum and minimum temperatures (expressed as difference from mean temperature) used ReLU activation functions. Where padding was used, it was replication padding. The number of filters used in each layer is shown in the diagram.

The bottom panel of the diagram shows the overall structure of the UNIT network. It also shows the path of an image $\mathbf{x}$ from domain X through the network components to generate its reconstruction ($\hat{\mathbf{x}}_x$), its translation to domain Y ($\hat{\mathbf{y}}_x$), and its cyclic reconstruction ($\hat{\mathbf{x}}_{xy}$). This panel also shows where in the path the translation is passed into the multi-scale discriminator.

This panel also shows the use of the land mask. Early in the study we noticed that the UNIT network was producing unphysical translations at land-sea boundaries, such as negative daily temperature ranges (this was noticed before we added some of the preprocessing and activation constraints mentioned in A1). This occurred exclusively on land-sea border pixels. To aid the network in translating these border regions effectively, the encoder $E(\mathbf{x})$ and discriminator $D(\mathbf{x})$ networks were given the concatenated weather fields and binary land-sea mask $[\mathbf{x}, \text{mask}]$ as input. The decoder $G(\mathbf{z})$ was only trained to reconstruct images $\hat{\mathbf{x}}_{GCM/obs}$ from the latent encoding $\mathbf{z}$ and ignored the land-sea mask.

A3. Training and hyperparameters

The hyperparameters used in training this network were taken from those used in the original UNIT network and adjusted manually a little using only the results of the training data. These were not highly optimised, so there may be room for improvement.

We trained the network from scratch using a batch size of 8 and the Adam optimiser (Kingma and Ba 2014) with learning rate of 5×10^{-4} for both the translation network and the discriminators. In the Adam optimiser, we also used weight decay, β_1 and β_2 values of 0.0001, 0.5, and 0.999. No explicit regularisation was used in training. The only regularisation present in the network is the implicit regularisation associated with the reparameterisation step in the VAEs.

We trained the network for 208,000 iterations. As with many adversarial networks, training is quite unstable, and so the model was checkpointed regularly (every 2000 iterations) throughout training. We chose the 208,000 iteration checkpoint manually although training continued to 570,000 iterations. We chose this checkpoint as the loss was reasonably low and stable at this point. The network took a few days to train on a single NVIDIA Tesla T4 GPU.

The loss coefficients used in equation (6) are shown in table A1. These are similar to those used in the original UNIT network.

Parameter	Value	Parameter	Value
λ_{rec}	10	λ_{KL-rec}	0.005
λ_{cyc}	10	λ_{KL-cyc}	0.005
λ_{GAN}	1		

TABLE A1. The coefficients used in the UNIT loss function

As noted in the main text, we use a weighted L1-norm loss function for the image reconstructions in equation (2). This was motivated to emphasise the precipitation fields in the translation as these have the most complex distributions. We weight the L1 loss calculated separately for each channel such that

$$L_1 = \frac{\sum_c w_c L_{1c}}{\sum_c w_c} \tag{A1}$$

where L_1 is the total weighted L1 loss for a sample, L_{1c} is the L1 loss for a channel c, and w_c is the weight for that channel. We set the weight of the precipitation channel to 5 and the weight of the other four channels to 1.

A4. Empirically calculated JS divergence

In this study, we have used an approximation of JS divergence (equation (7) that can be used on binned data.

Figure A2 shows the sensitivity to the number of bins of the spatial mean of the JS divergence between precipitation and temperature calculated at each gridpoint (i.e. the mean value across figure 5 for each dataset). This shows that our results are qualitatively robust to the choice in the number of bins in this estimation.

A5. Significance of JS divergence values

In the main text, we quote the values of the JS divergence calculated between each of the distributions in figure 4. We estimate the significance of these values via bootstrapping.

We randomly sample N fields with replacement from the ERA5 dataset, which is also of length N. We sample from each dataset $\chi \in$ {HadGEM3, UNIT, UNIT+QM, QM} similarly (we also sample again independently from ERA5 for results in table A2). Then we calculate the JS divergence between these bootstrapped samples. We repeat this bootstrapping calculation 200 times. The 5th, 50th and 95th percentiles, as well as the values quoted in the main

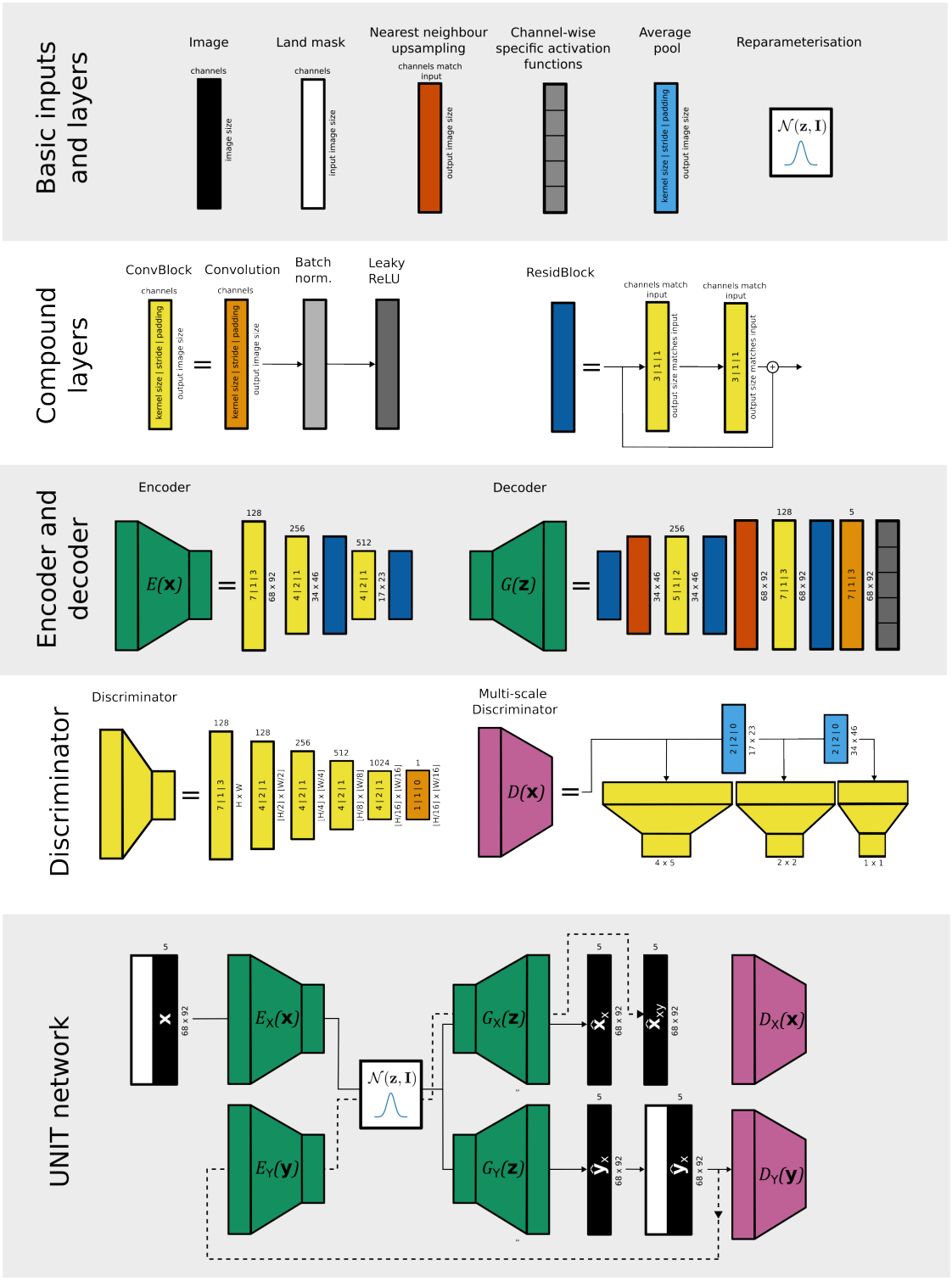

FIG. A1. Architecture of the UNIT network, configured as we use in this study.

text which did not use bootstrapping, are collated in table A2.

We note that the JS value calculated using all the data is consistently lower than even the 5% value using boot-

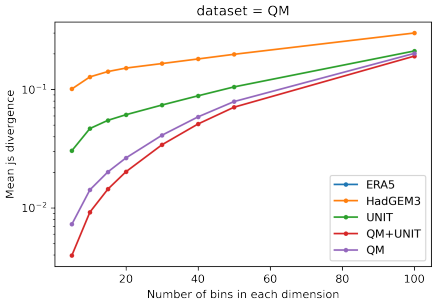

FIG. A2. The spatial mean of the JS divergence between temperature and precipitation at all gridpoints.

		Bootstrapped percentile		
Data source	value	5%	50%	95%
HadGEM3	0.1890	0.1903	0.1975	0.2048
UNIT	0.0527	0.0564	0.0605	0.0649
UNIT+QM	0.0199	0.0255	0.0288	0.0316
QM	0.0247	0.0311	0.0342	0.0381

TABLE A2. The values of the JS divergence calculated between each of the distributions in figure 4 as mentioned in the main text and their bootstrapped intervals.

		Bootstrapped percentile		
Data source	value	5%	50%	95%
Hot Days Temperature-Pressure (figure 10)				
HadGEM3	0.1383	0.1396	0.1451	0.1520
UNIT	0.0712	0.0738	0.0776	0.0811
UNIT+QM	0.0442	0.0482	0.0510	0.0545
QM	0.0462	0.0501	0.0533	0.0569
Wet Periods Temperature-Precipitation (figure 11)				
HadGEM3	0.2531	0.2523	0.2593	0.2670
UNIT	0.1856	0.1911	0.1971	0.2054
UNIT+QM	0.0310	0.0426	0.0462	0.0497
QM	0.0511	0.0628	0.0678	0.0734
SPI in two basins (figure 12)				
HadGEM3	0.1417	0.1413	0.1468	0.1522
UNIT	0.1100	0.1132	0.1186	0.1240
UNIT+QM	0.0811	0.0848	0.0886	0.0931
QM	0.0844	0.0891	0.0936	0.0989

TABLE A3. Bootstrapped confidence intervals calculated using the same method as those in table A2, but for JS divergence values presented in table 1.

strapped samples. This is simply because the bootstrapped samples are less diverse than the full dataset, and so it can only be expected that the match between the distributions will get worse. We consider one data source to be significantly better than the other if the 50% bootstrapped JS divergence of that dataset is lower than the 5% of the other data source. This suggests that UNIT+QM is a significant improvement on the other methods in this statistic.

Table A3 shows similarly calculated confidence intervals calculated for the JS divergences presented in table 1.

A6. Basins used

Figure A3 shows the areas used for each basin used in this study. The mask was created by checking whether the centre of each gridpoint is contained within the boundaries of the basin shape downloaded from the World Bank data catalogue (The World Bank 2019).

As noted, two of these basins are clipped due to the boundaries chosen (8°S-30°N 44°E-121°E). In this study, we only use the results of aggregating these basins as a demonstration of how each method of bias correction performs when aggregated over a spatial area. Therefore clipping these basins to the area in our domain does not affect the contents of the work presented here.

A7. Plots repeated as copulas

In order to assess the joint distribution of the translations without the influence of the marginal distributions,

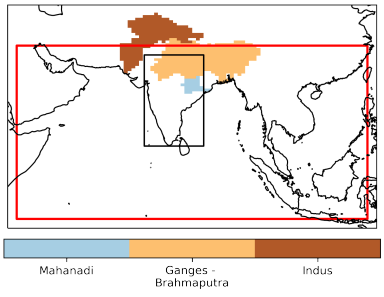

FIG. A3. Geographical extent of the key areas and basins used in this study. The red line denotes the overall region used in this study. For the basins which extend outside this red box, we have aggregated to only the area of the basin which lies inside this region. The black bounding box denotes the area bounded by 8°-28°N 72°-85°E, used to study extreme heat events as in figure 10.

we recreate some of our figures from the main text after transforming the data values into quantiles. This produces an empirical estimate of the copula of the two-dimensional datasets plotted. Figures A4, A5, A6, and A7 are copula alternatives to figures 4, 6, 8, and 9 respectively.

The results of figures A4-A7 support what we observed in the figures in the main text. The UNIT+QM method performs the best across the four figures. The UNIT+QM copula is more similar to the ERA5 copula than those of the other methods in all the figures. QM does not adequately

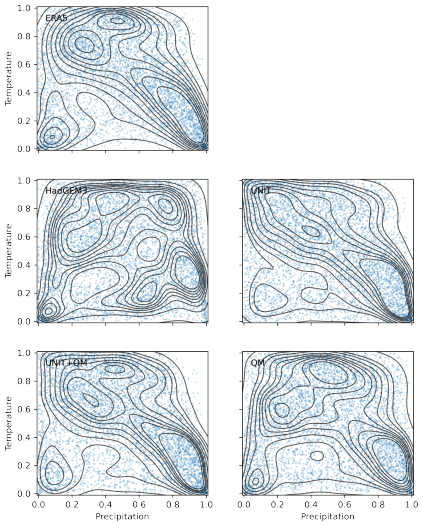

FIG. A4. Copula decomposition of figure 4 - Joint distributions of daily accumulated precipitation and daily mean temperature from a single gridpoint located at the southern tip of India (centred at 8.6°N 77.9°E) for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x and y-scales show the precipitation and temperature as expressed as quantiles calculated with respect to each dataset.

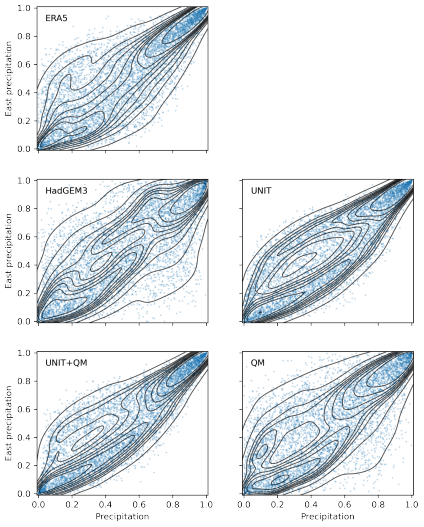

FIG. A5. Copula decomposition of figure 6 - Joint distributions of daily accumulated precipitation at a gridpoint located at the southern tip of India (centred at 8.6°N 77.9°E) and its neighbouring cell directly eastwards (centred at 78.8°E) for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x and y-scales show the precipitation at the two gridpoints as expressed as quantiles calculated with respect to each dataset.

correct the copula of the HadGEM3 data in any of the figures.

When a dataset is transformed via regular QM, its copula is unaffected. Therefore, we might expect the copulas of the HadGEM3 and QM datasets to be identical in each of the figures. Similarly, we might expect the UNIT and UNIT+QM data to have identical copulas. However, we apply quantile mapping to each month separately. This makes the QM we apply a conditional QM. This method can therefore modify the copula, not just the marginal distributions when considered over the whole year.

In figures A6 and A7 there are no data points in the gap between 0 and around 0.1-0.2 on the x-axis. This is because around 10-20% of the basin average precipitation for the Mahanadi and Indus basins there is exactly zero. Hence these 10-20% of data points are assigned a quantile of zero. The Ganges-Brahmaputra basin is large enough that almost no days have exactly zero precipitation across the basin, so no gap is observed in the y-axis in figure A7.

A8. Independently sampled SPI events

Figure A8 shows independently sampled values of SPI from the Ganges-Brahmaputra and Indus river basins. The overall structure of the distributions is very similar to figure 12, in which values are calculated from a rolling time window.

A9. Alternative matching algorithm

We examine the opposite match algorithm to the one used to produce figure 7. Instead of taking fields from dataset $\chi \in \{\text{HadGEM3, UNIT, UNIT+QM, QM}\}$ and finding the closest match in ERA5, we match the opposite way. We take each ERA5 field and find the closest match in dataset χ. The results of this are shown in figure A9. This figure shows fairly similar results to figure 7, with UNIT and UNIT+QM performing the best.

From the SSIM match distributions, we may be left asking how often each individual sample is chosen as the best match. If the translation is done well, we would expect that many samples as chosen as the best match rather than the same sample always being matched to. Figures A10 (based on matching precipitation$^{1/4}$) and A11 (based on matching temperature) show the results of this analysis. We count how many times each sample is chosen as the best match and then plot the frequencies (y-axis) of these best match counts (x-axis). For example, in figure A10, we can see that when we take HadGEM3 fields and select the best match amongst the ERA5 data, around 1500 of

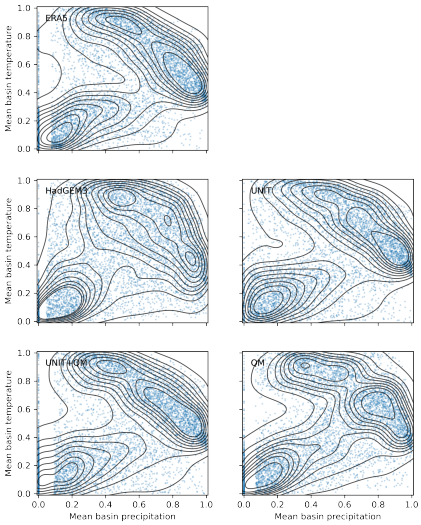

FIG. A6. Copula decomposition of figure 8 - Joint distributions of the mean daily accumulated precipitation and daily mean temperature aggregated over the Mahanadi river basin for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x and y-scales show the precipitation and temperature as expressed as quantiles calculated with respect to each dataset.

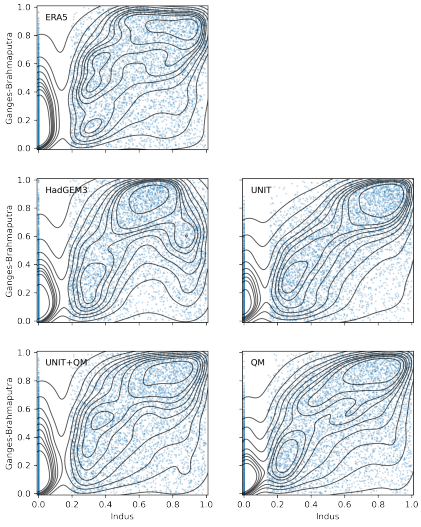

FIG. A7. Copula decomposition of figure 9 - Joint distributions of mean daily accumulated precipitation across the Ganges-Brahmaputra and the Indus basin for the five different datasets. Each point is from a single day, and the contours show the joint density as estimated by kernel density estimation. The x and y-scales show the precipitation in the two basins as expressed as quantiles calculated with respect to each dataset.

the ERA5 fields were chosen as the best match exactly one time. A little over 400 ERA5 fields were chosen as the best match exactly twice, and so on. In the figures, the black lines marked as ERA5→ERA5 are the results of matching the training ERA5 data to the validation ERA5 data. Each panel of these figures also shows the fraction of data which was matched to at least once.

In figure A10, we see evidence of the underdispersion of UNIT. When UNIT is matched to ERA, a few of the ERA5 fields are chosen many times, with one ERA5 field being the best match to around 400 UNIT fields.

In the UNIT+QM panel, we see that the match frequencies are similar to the ERA-to-ERA baseline regardless of whether we match ERA5 to UNIT+QM or UNIT+QM to ERA.

Figure A11 shows similar features to figure A10 but with these features less pronounced. This may be because temperature has a less complex spatial joint distribution than precipitation.

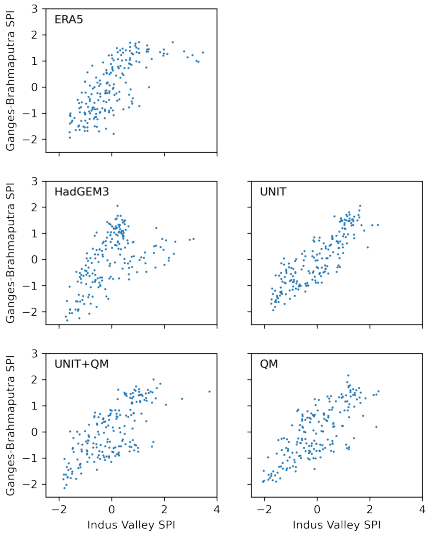

FIG. A8. Joint distributions of the Standardised Precipitation Index at each gridpoint over the Ganges-Brahmaputra and Indus river basins using a 30-day time window for the five different datasets. Each point is from a single gridpoint and a completely independent 30-day time window.

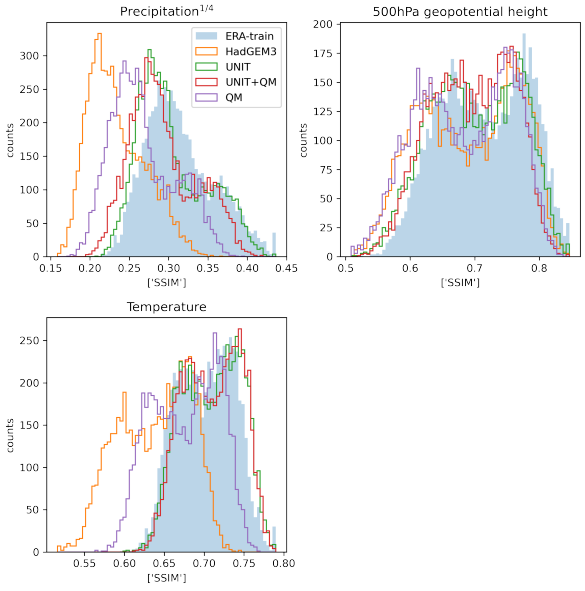

FIG. A9. Distributions of the SSIM between each ERA5 field and its best-matched dataset field. Repeated for the mean 2 metre temperature, mean 500hPa geopotential height and the fourth root of daily precipitation.

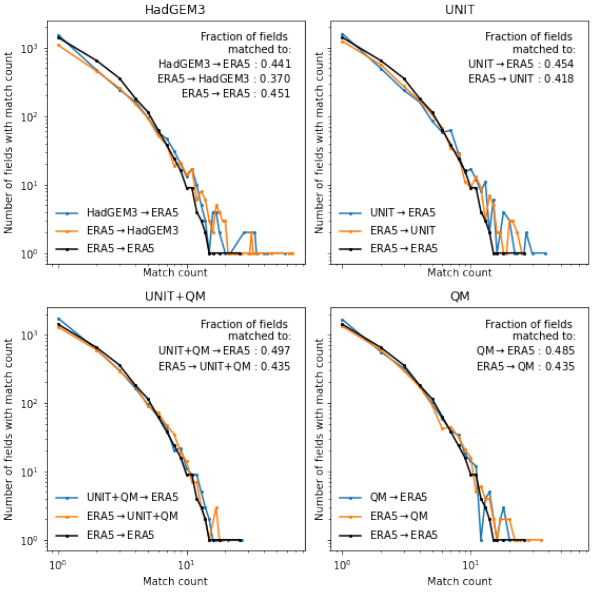

FIG. A10. Distribution of the number of times matched to fields were chosen as the best match using SSIM to compare precipitation$^{1/4}$. The x-axis is the number of times a field was chosen as the best match (the match count). The y-axis is the number of fields with this match count. The figure shows the results of matching in both directions between ERA5 and the bias correction methods. The black line is the result of performing the same SSIM matching based on matching the ERA5 training set to the ERA5 test set and is included for comparison.

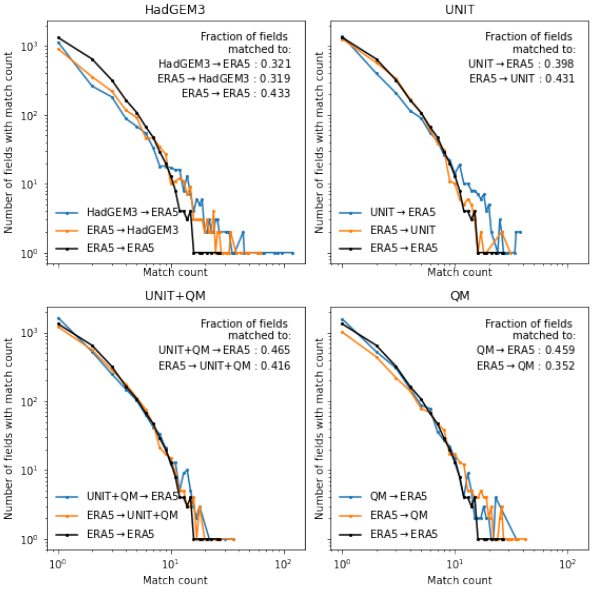

FIG. A11. The same as figure A10 but based on matching temperature.

References

Aadhar, S., and V. Mishra, 2017: High-resolution near real-time drought monitoring in south asia. *Scientific Data*, **4 (1)**, 1–14.

Arora, S., and Y. Zhang, 2017: Do GANs actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*.

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

Bau, D., J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba, 2019: Seeing what a GAN cannot generate. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4502–4511.

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

Bollasina, M. A., and Y. Ming, 2013: 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.

Cannon, A. J., 2018: 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.

Cannon, A. J., S. R. Sobie, and T. Q. Murdock, 2015: 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.

Chandrasekara, S. S., H.-H. Kwon, M. Vithanage, J. Obeysekera, and T.-W. Kim, 2021: Drought in south asia: A review of drought assessment and prediction in south asian countries. *Atmosphere*, **12 (3)**, 369.

Christian, J. I., J. B. Basara, E. D. Hunt, J. A. Otkin, J. C. Furtado, V. Mishra, X. Xiao, and R. M. Randall, 2021: Global distribution, trends, and drivers of flash drought occurrence. *Nature communications*, **12 (1)**, 1–11.

Ciavarella, A., and Coauthors, 2018: 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.

Dionelis, N., M. Yaghoobi, and S. A. Tsaftaris, 2020: Tail of distribution gan (tailgan): Generativeadversarial-network-based boundary formation. *2020 Sensor Signal Processing for Defence Conference (SSPD)*, IEEE, 1–5.

Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert, 2012: HESS Opinions "should we apply bias correction to global and regional climate model data?". *Hydrology and Earth System Sciences*, **16 (9)**, 3391–3404.

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

Eyring, V., and Coauthors, 2021: *2021: Human Influence on the Climate System*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

François, B., S. Thao, and M. Vrac, 2021: Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, **57 (11)**, 3323–3353.

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

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

Grover, A., C. Chute, R. Shu, Z. Cao, and S. Ermon, 2020: Alignflow: Cycle consistent learning from multiple domains via normalizing flows. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 4028–4035.

Guha-Sapir, D., R. Below, and P. Hoyois, 2014: Em-dat: International disaster database–www. emdat. be. université catholique de louvain. *Brussels: Belgium*.

Han, L., M. Chen, K. Chen, H. Chen, Y. Zhang, B. Lu, L. Song, and R. Qin, 2021: A deep learning method for bias correction of ecmwf 24–240 h forecasts. *Advances in Atmospheric Sciences*, **38 (9)**, 1444–1459.

Hanlon, H., G. Hegerl, S. Tett, and D. Smith, 2015: Near-term prediction of impact-relevant extreme temperature indices. *Climatic Change*, **132 (1)**, 61–76.

Hao, Z., A. Mallya, S. Belongie, and M.-Y. Liu, 2021: GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. *ICCV*.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049.

Im, E.-S., J. S. Pal, and E. A. Eltahir, 2017: Deadly heat waves projected in the densely populated agricultural regions of south asia. *Science advances*, **3 (8)**, e1603 322.

IPCC, 2013: *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, 1535 pp., https://doi.org/10.1017/CBO9781107415324.

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

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

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and M. Welling, 2013: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Le, X.-H., G. Lee, K. Jung, H.-u. An, S. Lee, and Y. Jung, 2020: Application of convolutional neural network for spatiotemporal bias correction of daily satellite-based precipitation. *Remote Sensing*, **12 (17)**, 2731.

Leonard, M., and Coauthors, 2014: A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, **5 (1)**, 113–128.

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

Levy, K., A. P. Woster, R. S. Goldstein, and E. J. Carlton, 2016: Untangling the impacts of climate change on waterborne diseases: a systematic review of relationships between diarrheal diseases and temperature, rainfall, flooding, and drought. *Environmental science & technology*, **50 (10)**, 4905–4922.

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

Liu, M., and O. Tuzel, 2016: Coupled generative adversarial networks. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 469–477.

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

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

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, 2017: Least squares generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, **48 (3)**.

Maraun, D., and Coauthors, 2017: Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, **7 (11)**, 764–773.

Mishra, V., S. Aadhar, and S. S. Mahto, 2021: Anthropogenic warming and intraseasonal summer monsoon variability amplify the risk of future flash droughts in india. *Npj Climate and Atmospheric Science*, **4 (1)**, 1–10.

Moghim, S., and R. L. Bras, 2017: Bias correction of climate modeled temperature and precipitation using artificial neural networks. *Journal of Hydrometeorology*, **18 (7)**, 1867–1884.

Moors, E., T. Singh, C. Siderius, S. Balakrishnan, and A. Mishra, 2013: Climate change and waterborne diarrhoea in northern india: Impacts and adaptation strategies. *Science of the Total Environment*, **468**, S139–S151.

Odena, A., V. Dumoulin, and C. Olah, 2016: Deconvolution and checkerboard artifacts. *Distill*, https://doi.org/10.23915/distill.00003.

Pan, B., G. J. Anderson, A. Goncalves, D. D. Lucas, C. J. Bonfils, J. Lee, Y. Tian, and H.-Y. Ma, 2021: Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, **13 (10)**, e2021MS002 509.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597 (7878)**, 672–677, https://doi.org/10.1038/s41586-021-03854-z, URL https://doi.org/10.1038/s41586-021-03854-z.

Raymond, C., T. Matthews, and R. M. Horton, 2020: The emergence of heat and humidity too severe for human tolerance. *Science Advances*, **6 (19)**, eaaw1838.

Ridder, N. N., A. J. Pitman, and A. M. Ukkola, 2021: Do CMIP6 climate models simulate global or regional compound events skillfully? *Geophysical Research Letters*, **48 (2)**, e2020GL091 152.

Rimi, R. H., K. Haustein, M. R. Allen, and E. J. Barbour, 2019: Risks of pre-monsoon extreme rainfall events of bangladesh: Is anthropogenic climate change playing a role? *Bulletin of the American Meteorological Society*, **100 (1)**, S61–S65.

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

Snell, J., K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, 2017: Learning to generate images with perceptual similarity metrics. *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 4277–4281.

Sønderby, C. K., and Coauthors, 2020: Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*.

Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of tropical-extratropical teleconnections on intraseasonal time scales. *Reviews of Geophysics*, **55 (4)**, 902–937.

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

Takahashi, H., A. Bodas-Salcedo, and G. Stephens, 2021: Warm cloud evolution, precipitation, and their weak linkage in hadgem3: New process-level diagnostics using a-train observations. *Journal of the Atmospheric Sciences*, **78 (7)**, 2075–2087.

The World Bank, 2019: WB data catalog - major river basins of the world. https://datacatalog.worldbank.org/search/dataset/0041426.

Tian, B., and X. Dong, 2020: The double-itcz bias in cmip3, cmip5, and cmip6 models based on annual mean precipitation. *Geophysical Research Letters*, **47 (8)**, e2020GL087 232.

Timmermann, A., and Coauthors, 2018: El Niño–southern oscillation complexity. *Nature*, **559 (7715)**, 535.

Tirivarombo, S., D. Osupile, and P. Eliasson, 2018: Drought monitoring and analysis: standardised precipitation evapotranspiration index

(spei) and standardised precipitation index (spi). *Physics and Chemistry of the Earth, Parts A/B/C*, **106**, 1–10.

Trenberth, K. E., and D. J. Shea, 2005: Relationships between precipitation and surface temperature. *Geophysical Research Letters*, **32 (14)**.

Vaughan, A., W. Tebbutt, J. S. Hosking, and R. E. Turner, 2022: Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, **15 (1)**, 251–268.

Wang, B., C. Jin, and J. Liu, 2020: Understanding future change of global monsoons projected by cmip6 models. *Journal of Climate*, **33 (15)**, 6471–6489.

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

Wang, F., and D. Tian, 2022: On deep learning-based bias correction and downscaling of multiple climate models simulations. *Climate Dynamics*, 1–18.

Wang, F., D. Tian, L. Lowe, L. Kalin, and J. Lehrter, 2021: Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, **57 (4)**, e2020WR029 308.

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

Wehner, M., D. Stone, H. Krishnan, K. AchutaRao, and F. Castillo, 2016: 16. the deadly combination of heat and humidity in india and pakistan in summer 2015. *Bulletin of the American Meteorological Society*, **97 (12)**, S81–S86.

White, R., and R. Toumi, 2013: The limitations of bias correcting regional climate model inputs. *Geophysical Research Letters*, **40 (12)**, 2907–2912.

Yiou, P., R. Vautard, P. Naveau, and C. Cassou, 2007: Inconsistency between atmospheric dynamics and temperatures during the exceptional 2006/2007 fall/winter and recent warming in europe. *Geophysical Research Letters*, **34 (21)**.

Yuan, X., M. R. Kaplan, and M. A. Cane, 2018: The interconnected global climate system—a review of tropical–polar teleconnections. *Journal of Climate*, **31 (15)**, 5765–5792.

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

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