JGR Atmospheres

RESEARCH ARTICLE
10.1029/2022JD038163

Deep Learning for Downscaling Tropical Cyclone Rainfall to Hazard-Relevant Spatial Scales

Emily Vosper[1,2], Peter Watson[1,3], Lucy Harris[4], Andrew McRae[4], Raul Santos-Rodriguez[2], Laurence Aitchison[2], and Dann Mitchell[1,3]

[1]School of Geographical Sciences, University of Bristol, Bristol, England, [2]School of Computer Science, University of Bristol, Bristol, England, [3]Cabot Institute for the Environment, University of Bristol, Bristol, England, [4]Department of Physics, University of Oxford, Oxford, England

Key Points:

- We find that the Wasserstein Generative Adversarial Network (WGAN) can be used to downscale tropical cyclone rainfall
- The WGAN reproduces the fine-scale spatial structure of tropical cyclones both visually and in its power spectra
- The WGAN is able to extrapolate to storms more extreme than seen in training

Correspondence to:

E. Vosper,
emily.vosper@bristol.ac.uk

Citation:

Vosper, E., Watson, P., Harris, L., McRae, A., Santos-Rodriguez, R., Aitchison, L., & Mitchell, D. (2023). Deep learning for downscaling tropical cyclone rainfall to hazard-relevant spatial scales. *Journal of Geophysical Research: Atmospheres*, *128*, e2022JD038163. https://doi.org/10.1029/2022JD038163

Received 9 NOV 2022
Accepted 2 MAY 2023

Author Contributions:

Conceptualization: Emily Vosper, Peter Watson, Laurence Aitchison, Dann Mitchell
Data curation: Emily Vosper
Formal analysis: Emily Vosper
Investigation: Emily Vosper, Lucy Harris, Andrew McRae
Methodology: Emily Vosper, Lucy Harris, Andrew McRae
Software: Emily Vosper, Lucy Harris, Andrew McRae
Supervision: Peter Watson, Raul Santos-Rodriguez, Laurence Aitchison, Dann Mitchell
Validation: Emily Vosper
Visualization: Emily Vosper
Writing – original draft: Emily Vosper
Writing – review & editing: Emily Vosper, Peter Watson, Lucy Harris, Andrew McRae, Raul Santos-Rodriguez, Laurence Aitchison, Dann Mitchell

Abstract Flooding, driven in part by intense rainfall, is the leading cause of mortality and damages from the most intense tropical cyclones (TCs). With rainfall from TCs set to increase under anthropogenic climate change, it is critical to accurately estimate extreme rainfall to better support short-term and long-term resilience efforts. While high-resolution climate models capture TC statistics better than low-resolution models, they are computationally expensive. This leads to a trade-off between capturing TC features accurately, and generating large enough simulation data sets to sufficiently sample high-impact, low-probability events. Downscaling can assist by predicting high-resolution features from relatively cheap, low-resolution models. Here, we develop and evaluate a set of three deep learning models for downscaling TC rainfall to hazard-relevant spatial scales. We use rainfall from the Multi-Source Weighted-Ensemble Precipitation observational product at a coarsened resolution of ∼100 km, and apply our downscaling model to reproduce the original resolution of ∼10 km. We find that the Wasserstein Generative Adversarial Network is able to capture realistic spatial structures and power spectra and performs the best overall, with mean biases within 5% of observations. We also show that the model can perform well at extrapolating to the most extreme storms, which were not used in training.

Plain Language Summary Tropical cyclones (TCs) are often associated with intense winds, but it is actually their associated rainfall and flooding that cause the majority of mortality and damages. A warmer atmosphere is able to hold more water vapor and therefore we expect to see increases in rainfall from TCs with global warming. To better support resilience efforts, it is critical to model current and future TC rainfall, but climate models at standard resolution struggle to do this accurately. Running climate models at very high resolution produces better results, though this requires significant computational resources meaning that fewer high-impact, low-probability TCs can be generated. Other methods, called downscaling models, are used to provide a computationally cheaper alternative by generating high-resolution TC-specific data rather than an entire global climate data set. In this study, we develop a set of deep learning models which can generate high-resolution rainfall data from low-resolution rainfall data. To do this, we train our models on data from observational data sets that have data for the period 1979–2020. We find that the Wasserstein Generative Adversarial Network performs the best over the metrics studied and is able to reproduce the most extreme storms that were not used in training.

1. Introduction

Tropical cyclones (TCs) are among the most impactful natural disasters, with associated flood hazards frequently being the leading cause of mortality and damages (Mitchell et al., 2022; Rezapour & Baldock, 2014). Current research indicates that each degree of ocean warming will likely raise average TC rainfall by 7%, while the proportion of category 4–5 TCs might also increase with global warming (Knutson et al., 2019). The World Meteorological Organization aims to reduce the impact of extreme flood hazards on society and has called for an urgent scaling up of funding to improve adaptation and resilience. To reduce the impact of TCs on society, it is important to understand global warming's impact on not only TC statistics but also associated flood risks, which in turn requires high-resolution rainfall predictions. Policymakers will require reliable risk quantification of TC rainfall and flood risks to make informed decisions and ensure resources are allocated effectively. With decision-making guided by thorough scientific insight, adaptation and resilience efforts can be improved in a targeted way.

Future climate risk can be quantified using output from general circulation models (GCMs), but at their standard resolution, these models struggle to accurately reproduce TC statistics, especially for the most extreme storms (Murakami et al., 2015). While increasing the horizontal resolution can improve precipitation estimates (Roberts et al., 2020; Zhang et al., 2021), these methods are computationally expensive. Therefore, there is often a trade-off between accuracy and generating large enough ensembles to evaluate sufficient high-impact, low-probability events for regional flood risk analysis. One solution to this is to employ a downscaling technique. In short, downscaling models use the outputs from another data set, such as a GCM, as boundary conditions to generate TC statistics over a specific basin or region and offer a computationally cheaper alternative while conserving—or even improving—performance over certain metrics compared to high-resolution GCMs (Jing et al., 2021).

Downscaling methods can either be statistical, or dynamical. Statistical downscaling methods use the historical record to generate TC statistics. For example, Tuleya et al. (2007) developed a rainfall climatology and persistence (R-CLIPER) model which generates the TC rainfall rate and accumulates it along either the forecast or observed storm track. Tuleya et al. (2007) link hourly rain gauge and satellite data from the Tropical Rainfall Measuring Mission to historical TC tracks and calculate an estimation of rain rate as a function of distance from the storm center and proportional to wind intensity. A topographic effect and asymmetric rainfall intensity field were later added in Lonfat et al. (2007) named the parametric hurricane rain model (PHRaM). Evaluations of T-CLIPER and PHRaM found they do not perform as well as their dynamical counterparts and underestimate TC rainfall (Brackins & Kalyanapu, 2020; Langousis & Veneziano, 2009).

Dynamical downscaling approaches can involve running a standalone TC-specific model or by running a regional climate model at resolutions as high as 1.5 km. For example, Steptoe et al. (2021) use European Centre for Medium-Range Weather Forecasting fifth generation Re-Analysis (ERA5) data as boundary conditions to the regional configuration of the Met Office Unified Model to generate 4.4-km and 1.5-km resolution TC data affecting Bangladesh. Their study was limited to only 12 TCs and nine ensemble members due to this method being computationally expensive. Alternatively, dynamical downscaling can otherwise involve seeding a low-resolution GCM with vortices that strengthen to form TCs should the right conditions be met (Emanuel et al., 2008). Once these seed vortices form TCs, the generated wind fields can then be used to calculate TC rainfall for thousands of synthetic TCs (Emanuel, 2017; Feldmann et al., 2019; Lu et al., 2018; Zhu et al., 2013). This method is significantly faster than regional climate models and can generate the wind profiles of thousands of synthetic TCs within a much shorter period of time. These high-resolution wind profiles can be used to calculate TC rainfall using a physics-based model that brings together major rainfall mechanisms including surface friction, topography, wind shear, and vortex stretching (Zhu et al., 2013). This rainfall model allows the user to evaluate TC rainfall risk over countries or regions and has been shown to be sensitive to topography, so it requires local calibration, much of which has been done on the US coast (Vosper et al., 2020). Therefore, applying this model on a global scale would likely require further calibration and come at a substantial computational cost.

For precipitation downscaling applied to other weather phenomena, studies have started to adopt deep learning techniques from a field of Computer Vision called image superresolution to achieve comparable results at a fraction of the computational cost. Initially, convolutional neural networks (CNNs) showed promise in precipitation downscaling (Adewoyin et al., 2021; Huang, 2020; Kumar et al., 2021; Sha et al., 2020; Wang et al., 2021). Many of these CNNs are formulated from a U-Net architecture (Ronneberger et al., 2015). U-Nets downscale by learning the relationship between the low-resolution input data and its high-resolution counterpart through a "U" shape deep learning architecture. Specifically, the low-resolution input data goes through a set of downsampling layers where it decreases in resolution before going through a series of upsampling layers to reach the desired high-resolution output. Wang et al. (2021) employed a Superresolution Deep Residual Network (SRDRN) to downscale temperature and precipitation data. The architecture of the SRDRN differs from the U-Net allowing it to be more computationally and memory efficient. U-Nets and basic CNNs are deterministic, meaning they only produce one prediction per set of inputs, and without adversarial training, they do not reproduce the fine spatial variability of rainfall.

More recently, the use of generative models has been explored following the success of Generative Adversarial Networks (GANs) in classical image superresolution (Ledig et al., 2017; Radford et al., 2016). GANs consist of two competing CNNs, a generator that generates high-resolution data from low-resolution input and a discriminator which determines whether the generated output data is realistic. Both models compete and gradual improvement is seen in the generated data as the generator is tested by an improving discriminator. Generative approaches

can be run stochastically as well as deterministically helping to account for model uncertainty, while the adversarial training likely allows models to reproduce the fine spatial detail and stochastic nature of precipitation (Harris et al., 2022a; Leinonen et al., 2020; Price & Rasp, 2022; Ravuri et al., 2021).

For example, Leinonen et al. (2020) implemented the Wasserstein Generative Adversarial Network (WGAN) with gradient penalty (Gulrajani et al., 2017) for rainfall downscaling. WGANs differ from classical GANs in that the discriminator uses a different function to measure the distance between the model and target distributions. Instead of using JS-divergence, WGANs use 1-Wasserstein distance. This modification results in training that is more stable and less sensitive to architecture choices and hyperparameter configurations (Arjovsky et al., 2017). Moreover, the WGAN implementation was sufficiently different to the original GAN idea that it motivated changing the name of "discriminator" to "critic," which will be used hereafter. Leinonen et al. (2020) found that the WGAN generated realistic, temporally consistent superresolution sequences for two data sets: one consisting of radar-measured precipitation from Switzerland; the other of cloud optical thickness derived from the Geostationary Earth Observing Satellite 16 (GOES-16). The WGAN produced close to the correct amount of variability in its outputs. Harris et al. (2022a) modified the work of Leinonen et al. (2020) such that the temporal component of the model was removed and the model focused on downscaling UK rainfall data. They trained the WGAN on nine data fields, including low-resolution rainfall data, from the Integrated Forecast System using 1-km Resolution UK Composite Rainfall Data from the Met Office NIMROD System (Met Office, 2003) as the high-resolution target data. In addition to the WGAN, Harris et al. (2022a) developed a Variational Autoencoder GAN (VAEGAN) which replaces the generator with a variational autoencoder. They showed that the WGAN and VAEGAN matched the statistical properties of state-of-the-art pointwise postprocessing methods while creating high-resolution, spatially coherent precipitation maps.

In Leinonen et al. (2020), the inputs are a coarsened version of the target data, whereas Harris et al. (2022a) use inputs and outputs that come from different domains. The latter is often a more difficult challenge, as there can be biases present between the two data sets. One novel approach to reduce the bias present in input data from a different domain was developed by Price and Rasp (2022). They used a novel two-stage architecture to downscale precipitation from one data set at 32 km resolution to a distinct data set at 4 km resolution over the Continental US. First, the coarse precipitation forecast is mapped to a corrected distribution based on information about the weather situation. Then, this corrected distribution is mapped to a distribution of high-resolution, plausible predictions. Model performance was close to that of a dynamical model and even outperformed it when comparing reliable extreme rainfall estimates.

In our study, similar to Leinonen et al. (2020), we work with input data that is a coarsened version of the target data so the model learns the pure superresolution relationship. Therefore, at this stage, we do not need to implement the methodology used in Price and Rasp (2022). Here, we further develop the WGAN and VAEGAN of Leinonen et al. (2020) and Harris et al. (2022a) and, for the first time, apply it to downscaling rainfall data from TCs. We will analyze the model predictions on a set of unseen extreme samples. This is to test the models' abilities to generalize predictions to events more intense than seen in training, which is important for considering application to future individual storms that could potentially generate more rainfall than any in history. Our overarching aim is to produce an efficient downscaling model that can generate ensembles of realistic, high-resolution TC rainfall data at low cost, based on low-resolution rainfall predictions. Our work is a step toward showing that this approach can be a very valuable tool for climate risk analysis. We hope that this technique will be complementary to other downscaling methods while operating at a fraction of the computational cost.

2. Data and Methods

2.1. Data Processing

We first constructed a training set of paired samples of low-resolution and high-resolution TC rainfall. To generate the samples, we cross-referenced the historical storm locations from the International Best track Archive for Climate Stewardship (IBTrACs) tracking data set (Knapp et al., 2010, 2018) with the 10-km resolution global rainfall Multi-Source Weighted-Ensemble Precipitation (MSWEP) data set (Beck et al., 2017). MSWEP is at 3-hourly resolution from 1979 to the present, while IBTrACs contains track information every 3–6 hr, depending on basin, from 1852 to the present. For this study, we look at storms that overlap both data sets, i.e., 1979 to present, and to maximize training data we use both the 3-hourly and 6-hourly time steps available in IBTrACs.

We select only tropical storms that have reached at least category 1 strength (or equivalent) which leaves us with 1,870 tropical storms.

For each time step, we selected a 100 × 100 grid cell (1,000 km across) array surrounding the center of the storm. This process generated ~90,000 rainfall samples, which corresponds to ~50 time steps per storm. We then used conservative interpolation to generate the corresponding coarse 10 × 10 arrays, thereby mimicking a lower-resolution data set, with a typical resolution to that of many current climate models. As TCs in the northern hemisphere spin in opposite direction to those in the southern hemisphere, we flipped the northern hemisphere TCs vertically so that all TC images were rotating in the same direction. We are left with a data set whereby each TC sample consists of a high-resolution "truth" array with its low-resolution "input" counterpart.

It is important to evaluate machine learning methods on the most extreme events, including the ability to generalize to events more extreme than those seen in training (Watson, 2022). In order to test this ability in the models evaluated here, we created separate "extreme" validation and test sets. The extreme test set consists of the 100 storms with the highest peak rainfall values on the coarse grid and the storms with the next 100 peak values formed the extreme validation set. We ranked samples by the peak rainfall on the low-resolution input grid rather than on the high-resolution grid which exhibits more small-scale variability and noise. Doing this would reduce the influence of this unpredictable small-scale variability on model training, since the machine learning models could not be expected to predict when large positive stochastic peaks would occur. We chose peak rainfall rather than another diagnostic like total rainfall to focus more on how the models would deal with extreme values of individual inputs. To avoid data leakage, whereby the model is advantaged by training with samples from the most extreme storms at earlier or later time steps, we removed all TC samples along each extreme storm's trajectory from the training, validation, and test sets. The extreme test set is therefore made up of the 100 most intense TC samples, and the next 100 most intense make up the extreme validation set.

We then randomly assigned the remaining TCs to the training set (46,000 samples) and regular validation and test sets (16,000 samples each). We tuned the hyperparameters of our models based on results from the validation and extreme validation sets. Having an extreme validation set allowed us to check that the models could perform well on events more extreme than seen in training before final evaluation on the test sets. Model development was completed and final evaluation was performed using the test sets, the results of which are shown below. To our knowledge, this is the first study applying machine learning to meteorology that put the most extreme cases in a test set and did not look at these in model development, thereby giving assurance that we have not overfitted extreme situations. This helps to quantify what errors may be expected if the model is used in situations more extreme than those used in training (Watson, 2022).

2.2. Model Development and Setup

2.2.1. Baselines

Here, we use two baselines: a bilinear interpolation of the low-resolution data and a U-Net baseline adapted from Ronneberger et al. (2015) to suit a superresolution task. The U-Net consists of a contracting path and an expansive path connected via long skip connections. The layers in the contracting path are skip connected and concatenated with layers in the expansive path which allows the U-Net to use the information learned in the contracting path to construct predictions in the expanding path. Sha et al. (2020) also adapted Ronneberger et al. (2015) for rainfall downscaling by a factor of 8, we make similar changes to downscale by a factor of 10. The original U-Net from Ronneberger et al. (2015) has four downsampling blocks in the contracting path, which perform a downsampling by a factor of 16. As we are downsampling 10 × 10 images, we amended this to two downsampling blocks as the original grid is too coarse to downscale beyond two steps. Similarly, instead of four upsampling blocks in the expansive path, we have three upsampling blocks to bring the output to 100 × 100. We reduced the number of output channels from 64, 128, 256, 512, 1,024 to 128, 256, 512. Like Sha et al. (2020), we also replaced the PReLU activation function on the output layer with ReLU to account for not having negative rainfall and use mean absolute error as the loss function.

2.2.2. VAEGAN and WGAN

In this study, we use the WGAN originally developed by Leinonen et al. (2020) and adapted by Harris et al. (2022a) and VAEGAN developed in Harris et al. (2022a). We made changes to both models in order to adapt them to

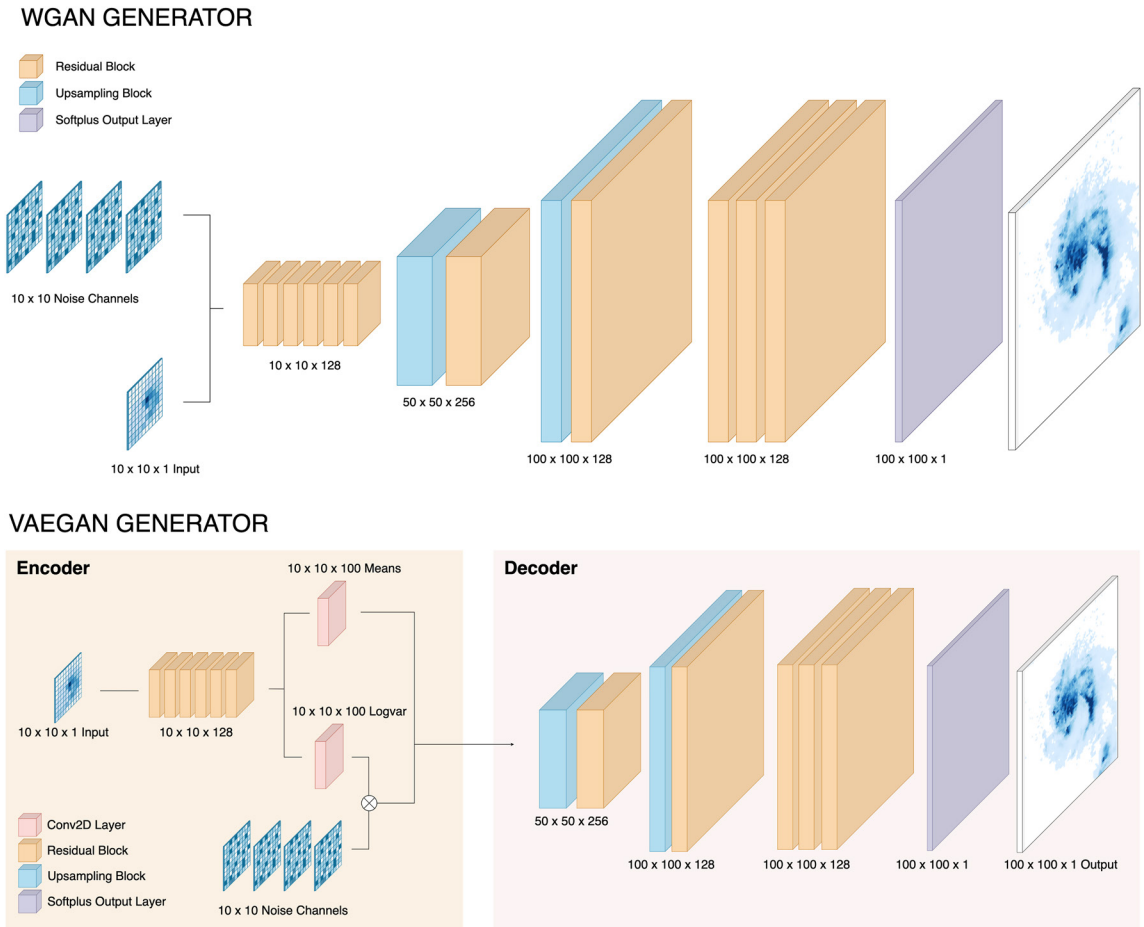

Figure 1. Generator architecture for the WGAN and VAEGAN. The main difference between generators is where the noise channels are fed to the network.

work for TC rainfall and to improve performance on our validation data sets. These changes include adding an additional three residual layers in the generators, to allow the model to further diversify the noise arrays, this improved the ensemble predictions by giving them more variability. The full generator architecture is shown in Figure 1, the critic remains the same as described in Harris et al. (2022a). It is important to note that Leinonen et al. (2020) have a time dependence that was removed in Harris et al. (2022a); in our study, each time point is treated independently and coherence across time is not modeled.

Both the VAEGAN and the WGAN train adversarially, with distinct generators and an identical critic network. The critic concatenates the coarse-resolution input, which has been passed through two residual blocks, to the high-resolution truth data, which has passed through two downsampling residual blocks. The concatenated data passes through two more residual blocks, a global average pooling layer and finally a dense output layer.

Both generators are fully convolutional without any dense layers allowing them to be size-agnostic. The benefit of this is that they can be fed input data of any size and trained for downscaling by a factor of 10. Therefore, we did not have to make any changes to the architecture in Harris et al. (2022a) to account for using differently sized input and output data. Both generators also benefit from using residual blocks, which are described in detail in Harris et al. (2022a). Residual blocks are similar to the long skip connections used in the U-Net; however, they work on a shorter scale by passing information from previous layers directly into deeper layers about 2–3 steps away. In a network without residual block, the network learns a mapping $H(x)$, residual networks learn another mapping $F(x) := H(x) - x$ and the original mapping is redefined as $H(x) := F(x) + x$ where the $F(x)$ acts like a residual, hence the name residual block. It is easier for the network to optimize for the residual mapping compared to the original mapping and thus residual networks tend to show better performance and help to address the problem of vanishing and exploding gradients. A softplus activation function is used in the output layer of the

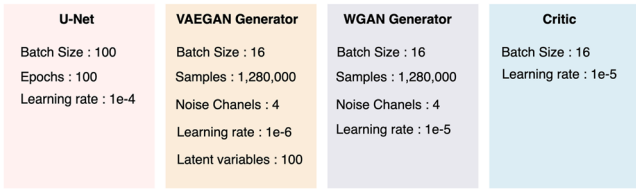

Figure 2. Model hyperparameters.

generators, this prevents the output from having an artificially constrained maximum, which would be the case with a sigmoid activation function, e.g., Harris et al. (2022a) found that without an artificially constrained maximum, the softplus activation allowed the network to produce more realistic extreme rainfall.

In the WGAN generator, shown in Figure 1, we combine our coarse rainfall samples with four noise channels before passing through six residual blocks (instead of three blocks as in Harris et al. (2022a)). We then upsample to high resolution using two upsampling residual blocks. We then pass the data through three residual blocks and lastly a softplus output layer.

The VAEGAN generator is made up of an encoder and a decoder network. The key difference between the VAEGAN and the WGAN is where the noise variables are introduced to the network. Instead of concatenating the input and noise in the initial step, we pass our input through six residual blocks (instead of three as in Harris et al. (2022a)). Then the noise is introduced when the latent variable distributions are sampled, this makes up the encoder part of the network. The decoder part of the VAEGAN is identical to the rest of the WGAN generator. Both models use convolutions with half padding.

The model hyperparameters selected after fine-tuning on the validation set are shown in Figure 2. Despite training with various content loss terms, predictions were disappointing and "blurry." Results were greatly improved when the content loss was set to zero. We found that a learning rate of 1e−6 for the VAEGAN was optimal, as larger learning rates resulted in unstable training often with the VAEGAN predicting no rainfall. The U-Net is a much more efficient network and therefore we could use a high batch size, we found that a larger batch size produced much better results overall, this was a similar case with the GANs and that a batch size of 16 was optimal to balance memory usage and performance.

We quantify the time taken to generate TC samples as 14 min to generate 16,000 samples which corresponds to roughly 320 TCs using a single NVIDIA A100 GPU.

3. Results

The results section is organized into three subsections. The first evaluates the model prediction's visual and spatial patterns to determine if generated data behaves like observed TC rainfall. We do this by plotting samples of predictions, evaluating the power spectra and examining the bias and standard deviation. In the second section, we concentrate on the difference in model performance at the extremes, this was most evident through analysis on quantile-quantile plots and looking at the spread-error relationship. Lastly, we evaluate the best performing model on accumulated rainfall, to determine how valuable this model could be for flood risk applications.

3.1. Visual and Spatial Analysis

Figure 3 shows model predictions of rainfall from six notable TCs chosen for their distinct visual patterns: Hurricane Maria (2017; North Atlantic), Typhoon Haiyan (2013; West Pacific), Hurricane Barbara (1995; East Pacific), Cyclone Vayu (2019; South Indian), Cyclone Ursula (1998; South Pacific), and Typhoon Brendan (1985; West Pacific Ocean). The first and second three are from the regular and extreme test sets, respectively, those from the extreme test set are taken at the point where maximum rainfall occurs at coarse resolution. For the less intense TCs Maria (2017), Haiyan (2013), and Barbara (1995), both the VAEGAN and the WGAN produce realistically similar results to the high-resolution observations, and are a clear improvement over the U-Net model. The U-Net baseline clearly struggles with fine detail of the rainfall data, producing much smoother images overall. Both the VAEGAN and the WGAN have a stochastic component and adversarial training, which has led to them making predictions with much finer detail, whereas the U-Net model is deterministic and is constructed to minimize root mean squared error (RMSE).

At the time points represented in Figure 3, TCs Vayu (2019), Ursula (1998), and Brendan (1985) exhibit their highest rainfall values at coarse resolution: these are 77, 69, and 79 mm/hr, respectively. These high values all lie outside of the model's training set where input samples do not exceed 65 mm/hr. The highest rainfall values at high resolution are 117 mm/hr for Cyclone Vayu (2019), 96 mm/hr for Cyclone Ursula (1998), and 149 mm/hr

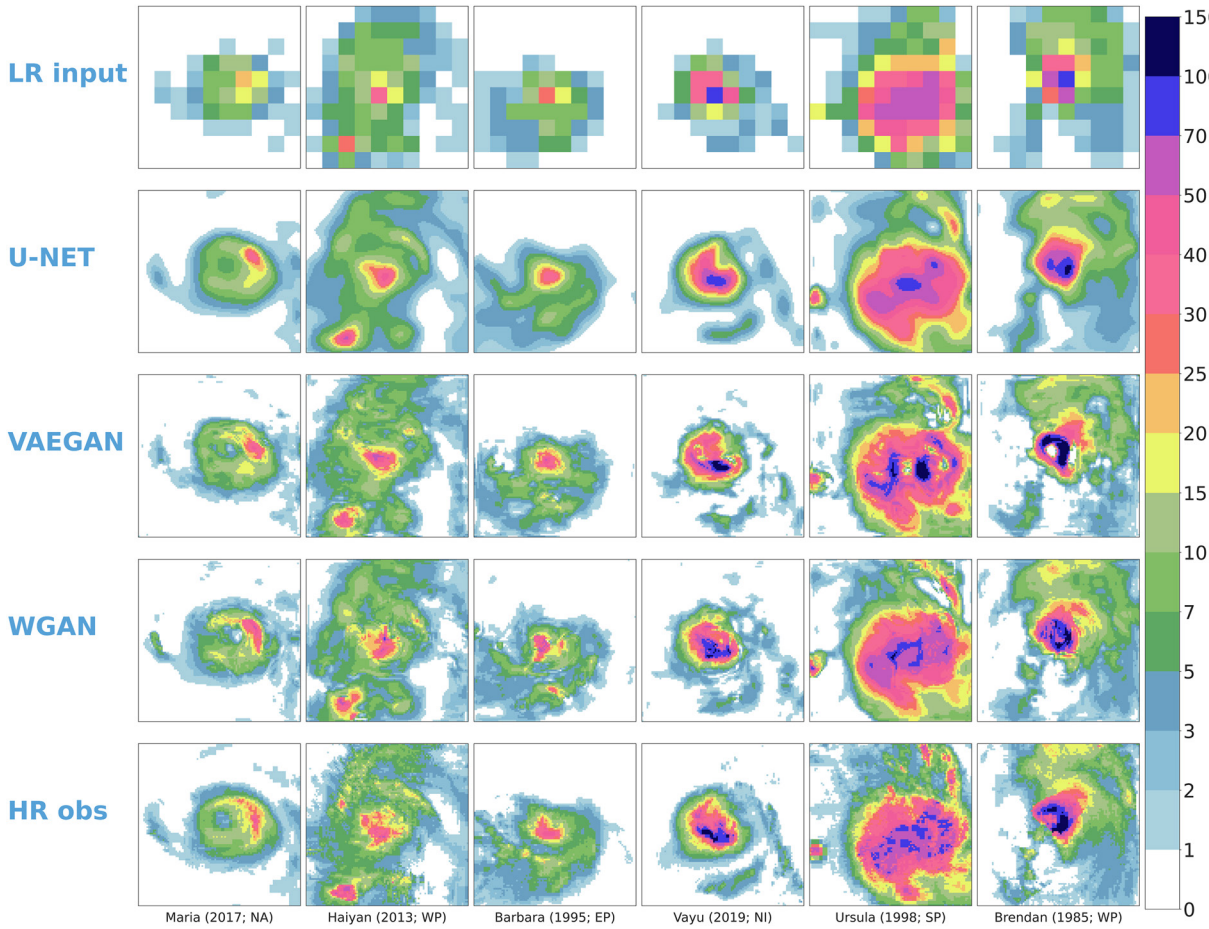

Figure 3. Using only precipitation data at 100-km resolution (top row), the models predict the corresponding high-resolution rainfall field which increases resolution by a factor of 10. Target high-resolution observed rainfall of six tropical cyclones over five different ocean basins (North Atlantic, West Pacific, East Pacific, South Pacific, and South Indian) are shown in the bottom row. The middle rows show high-resolution predictions made using the U-Net, VAEGAN, and WGAN models. The left three columns correspond to TCs from the regular test set and the right three columns to TCs from the extreme test set where the maximum peak rainfall occurs.

for Typhoon Brendan (1985). For all three storms, the U-Net struggles to generate these high values reaching 88, 86, and 108 mm/hr, respectively. On average, the U-Net underpredicts the peak values by 70 mm/hr over the 100 most extreme time points in the extreme test set. In these three examples, the VAEGAN consistently over-predicted the maximum rainfall: 170, 161, and 182 mm/hr. Please refer to Table 1 for a summary of the observed versus model-generated peak rainfall values for TCs Vayu, Ursula, and Brendan. We also see that the 2D rainfall patterns begin to deteriorate with the more extreme samples (Figure 3). On average, the VAEGAN overpredicts maximum rainfall by 9,000 mm/hr. This surprising error occurs in predictions from the 10 most extreme samples, when we remove these anomalous 10 predictions, the average peak rainfall error becomes 102 mm/hr. Similarly, the WGAN has a mean absolute error of 224 mm/hr over the 100 most extreme samples, this reduces to 33 mm/hr when removing the top 10 where peak coarse rainfall values range from 72 to 190 mm/hr. The WGAN predicts

Table 1
The Maximum Rainfall Values for Three Intense Tropical Cyclones Shown in Low-Resolution and High-Resolution Observation Arrays and in Predictions From the U-Net, VAEGAN, and WGAN

Storm	LR observations	HR observations	U-Net	VAEGAN	WGAN
Vayu (2019)	77	117	88	170	127
Ursula (1998)	69	96	86	161	110
Brendan (1985)	79	149	108	182	146

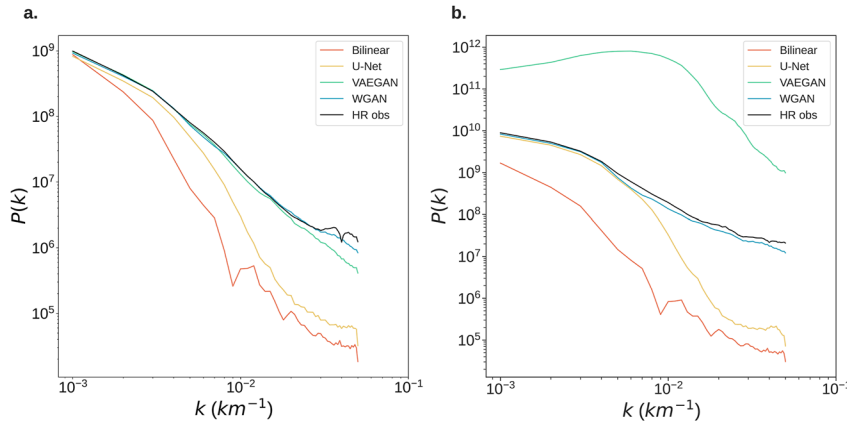

Figure 4. Spectral power against 1/wavenumber of the predictions from bilinear interpolation and machine learning models on the regular test set (panel a) and the 100 extreme test set samples (panel b). The *x*-axis is in units of km and represents the reciprocal of the wavenumber multiplied by the number of pixels (100) and the distance between each pixel (10 km).

peak values of 127, 110, and 146 mm/hr, respectively, for the three extreme storms. Unlike the VAEGAN, the WGAN is able to retain realistic 2D rainfall features for the extreme examples shown in Figure 3, giving us confidence that the WGAN will work for events more extreme than seen in training up until peak coarse resolution of 72 mm/hr.

All three machine learning models struggle with reproducing the finer rotational features observed in TCs. The fine rotational aspect of Cyclone Brendan (the rain band) looks artificial for model predictions which have struggled to replicate the behavior consistent with a rotating storm. This is unlikely to be due to the opposite rotation of TCs in the southern and northern hemispheres as they were flipped in data processing to all rotate clockwise, they were subsequently flipped back to the correct rotation for analysis. The architecture we use is fully convolutional, which makes it a size-agnostic network and therefore able to perform downscaling of a factor of 10 regardless of the dimensions of the input data. However, this feature also means that the model does not associate image features with a particular location in the image, i.e., it does not know that the upper right quadrant is distinct to the lower left quadrant which is evident in a rotating image. Adding a fully connected layer could improve the predictions of rotational features as the image quadrants differ from each other.

To quantify the quality of simulated spatial structures of the rainfall predictions, we analyze the power spectra for each of the four models. Power spectra show the total variance associated with features at different spatial scales. To calculate the power spectra, we use Fourier transforms to decompose the image into a series of decreasing spatial scales, then we calculate the variance at each scale. Figure 4 shows the power spectra, P, versus wave number, k, of predictions of images in the regular test set (4a) and the 100 most extreme samples in the extreme test set (4b). In the observations of Figure 4a, we see that most power is located at large scales (small wave numbers) meaning that the variance is dominated by large structures such as the rain bands and the eye of the storms. Toward smaller scales (larger wave numbers), the power drops off significantly. This is an expression of the complex nature of the TC: with the image having a distinctive overall shape with finer detail being present mainly at the edges of the structure, and not a dominant feature of the image. The larger features in images generated from bilinear interpolation and the U-Net remain at a similar power to observations, this is unsurprising given that the input data is a coarsened version of the target data and therefore preserving the large scale features, though the difference becomes stark for medium and small-scale features, <300 km across ($k > 3$), showing that the U-Net has not added fine-structured detail. This is evident through the smoothness of images generated by the U-Net, shown in Figure 3, the model clearly struggles to produce the noisy, fine-scale details. Both the VAEGAN and the WGAN exhibit similar power spectra to observations for wave numbers up to around 20 which corresponds to the 50-km scale, where the VAEGAN deviates more than the WGAN which continues to perform well at even finer scales.

Figure 4b assesses the spectral power against wave number for the most extreme 100 samples. Here, we see a much stronger deviation from observations in the VAEGAN predictions. This behavior is likely dominated by a few rogue predictions, where the VAEGAN has generated very unrealistic values (discussed later), but we

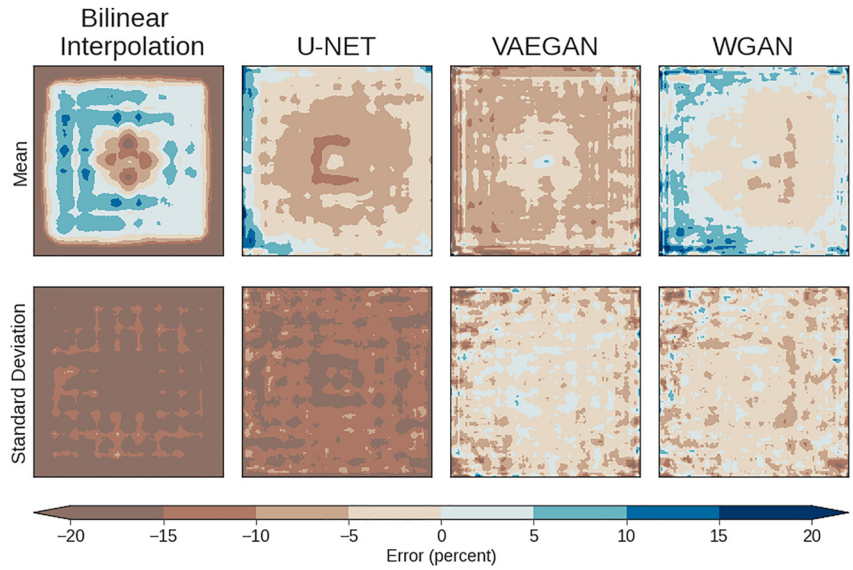

Figure 5. Biases of the mean (top row) and standard deviation (bottom row) of predictions from the regular test set for each of the four models. Horizontal and vertical axes indicate position relative to the storm track location, which corresponds to the center of each panel. Edge errors in the bilinear interpolation method are large because absolute values are small, these small errors cause large percentage errors.

also see from Figure 3 that the spatial structure in VAEGAN predictions is less well represented in the extreme cases. The U-Net actually performs slightly better than it did for regular TCs, only beginning to deviate from observations for values of $k > 6$ or for features around 160 km across. The WGAN remains in line with observed spectral power for all wave numbers, indicating that it produces consistently realistic images of both regular TCs and extreme samples which lie outside of the original training set. This analysis also indicates the importance of evaluating data on unseen extreme cases, as the VAEGAN performed well on the original training distribution, but its predictions broke down when given samples that lie outside of this distribution.

Figure 5 shows the percentage bias in the mean and standard deviation composited over each sample in the predictions versus the high-resolution observations in the regular test set. Here, the storm track position is at the center of each panel which is roughly 1,000 km across. For each model, we calculated the mean and standard deviation at each grid box and subtracted the same quantity calculated using the high-resolution observational data. It is important to note that the mean and standard deviation will have distinct structures to the individual TCs as shape and physical extent will vary between storms. Likewise, if we were to calculate the bias over a region, the mean and standard deviations would possibly highlight a more defined structural bias.

The magnitude of the mean and standard deviation biases is greatest with predictions from bilinear interpolation. The bilinear interpolation has a distinctive dry bias of up to 20% in the center of the mean predictions, this is surrounded by a wet bias of up to 15%. This pattern is not seen in results from the U-Net, where almost all regions have a dry bias of between 5% and 15%.

Results from the VAEGAN suggest a dry bias in the mean in all regions apart from the center. Bias as high as 15%–20% is seen on the edges of the image, which could be in part due to the model responding to padding, where data are lost at the edges when performing convolutions. The WGAN exhibits the smallest bias of the four models, with the large majority of results within ±5% of observations. The WGAN overpredicts rainfall around the edge of the TCs, where rainfall is on average lower though features such as rain bands occasionally bring in high rainfall values. The WGAN also overpredicts rainfall in the eye of the TC, this is typically where very little rainfall occurs compared to the eye wall. The WGAN underpredicts rainfall in the eye wall, where the majority of rain from TCs is distributed.

Both bilinear interpolation and the U-Net predictions exhibit less spread than observations, shown by a large negative bias in the standard deviation which mostly exceeds 10%. The VAEGAN and the WGAN have a smaller negative standard deviation bias of between 0% and 5%.

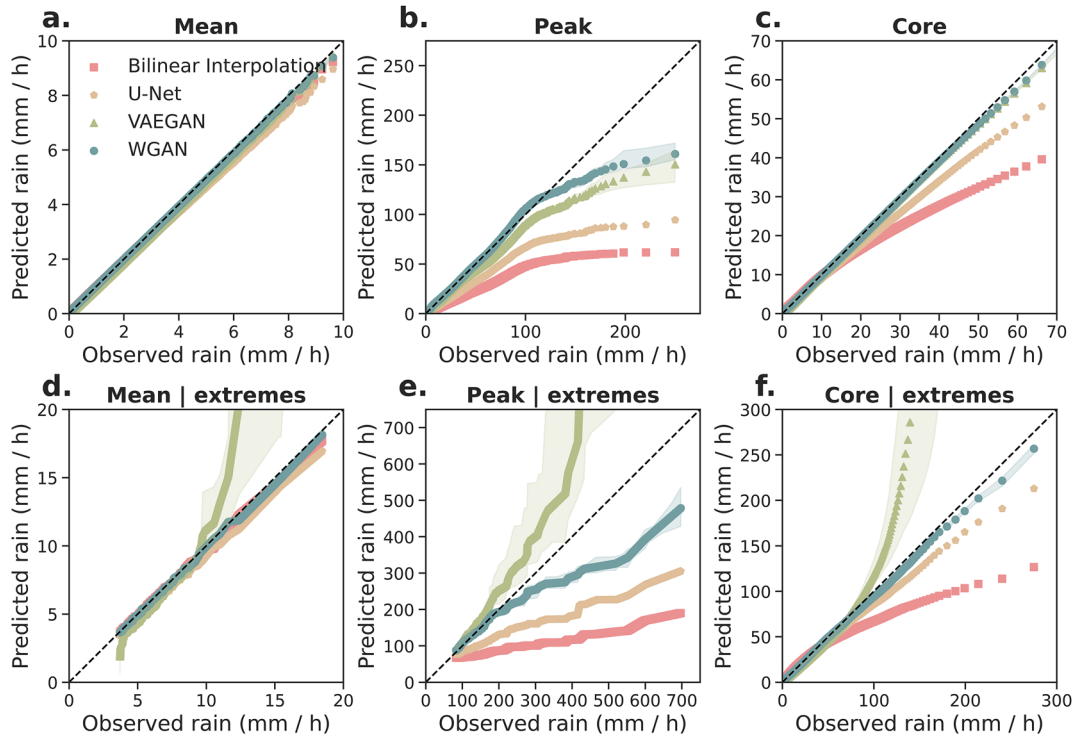

Figure 6. Quantile-quantile plots of observed versus predicted values of, for each sample, the mean, peak value, and mean value over the core of the TC—i.e., values spanning a 250-km radius. Results are shown for both the regular test set and the extreme test set of 100 most extreme time points. Shading for the WGAN and VAEGAN represents the ensemble spread. Quantiles in bins of 0.1 are plotted spanning from the 0.1th quantile to the 99.9th quantile. In the extreme cases, high values from the VAEGAN are out of plot range, and not shown, reaching 255 mm/hr in the mean, 46,684 mm/hr in the peak, and 5,051 mm/hr in the core for the 99.9th percentile values.

Across all four comparisons, there is a similar bias between predictions on the test set and the extreme test set (not shown), with biases on the extreme test set exhibiting similar spatial patterns but with higher biases that exceed ±20% in places. All plots show a differing degree of "hatching" in the bias and standard deviation, this is likely to be a feature of the various upsampling steps in model architecture. Adding more residual layers (or double convolutional layers for the CNN) between upsampling steps could act to smooth out these artifacts.

3.2. How Do the Models Handle Extreme Events?

Figure 6 shows quantile-quantile plots of model predictions against observations across four metrics: the mean rainfall across the 100 × 100 output domain, the peak rainfall over the domain, and the mean rainfall around the core of the TC (radius 250 km). Figure 6 shows results for both the regular and extreme test set.

There is very little error in quantile estimates in the mean rainfall, with all models reproducing mean values well in results on the regular test set (Figure 6a). When looking at the mean representation on the extreme test set (Figure 6d), the VAEGAN begins to deviate from the observations for rainfall values larger than 10 mm/hr, which lie out of the range of the regular test set values in Figure 6a.

There is much greater variation in results between models when analyzing the peak rainfall quantiles. In the nonextreme test set (Figure 6b), the VAEGAN and WGAN models begin to produce lower values than observations for values greater than 100 mm/hr. They both perform better than the bilinear interpolation method and U-Net which deviate from observations at much lower values. In Figure 6e, the VAEGAN deviates exponentially away from observations of extreme TCs for values exceeding 400 mm/hr, until this point the WGAN and VAEGAN are equal distance away from the 1–1 line. The WGAN remains on the 1–1 line up until peak rainfall values of around 200 mm/hr, it then begins to underpredict rainfall for more extreme values. The WGAN is able to reproduce core values of the extreme TCs Figure 6f extremely well, with the 99.9th percentile only around

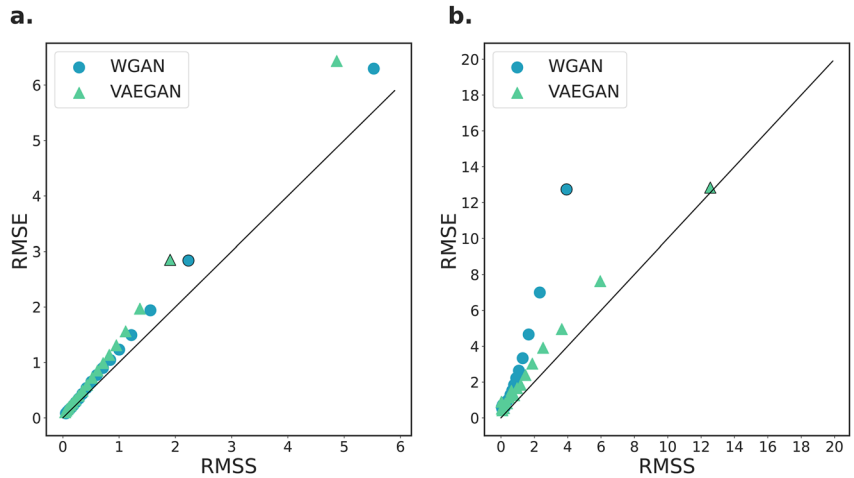

Figure 7. Spread-error of stochastic models WGAN and VAEGAN on the test set predictions (panel a) and the extreme test set predictions (panel b). Ensemble root mean squared error (RMSE) and root mean square spread (RMSS) are calculated over percentile bins five percentile values in width. Scatter points outlined in black represent the 90–95th percentile bin. The results for the extreme test over the 95–100th bin are not shown here as they are out of plot range: the VAEGAN has RMSS of 1,210 and RMSE of 524 and the WGAN has RMSS of 12 and an RMSE of 27.

20 mm/hr lower than observations. The VAEGAN deviates exponentially away from observations for rainfall values above 100 mm/hr.

Overall, the quantiles of the WGAN predictions remain the closest to observations, while bilinear interpolation performs the worst consistently. Both the GAN and VAEGAN do well at reproducing core average TC rainfall, though for the more extreme TCs the VAEGAN begins to overpredicts core rainfall exponentially.

The VAEGAN and WGAN are stochastic models which allow us to generate an ensemble of predictions for each low-resolution input; here, we have generated 20 ensemble members. It is important to test whether the degree of stochasticity is appropriate and approximately matches the size of the errors of the mean predictions. Figure 7 shows the spread-error relationship (Leutbecher & Palmer, 2008), based on 20-member ensembles generated from these models. This is shown for the nonextreme and extreme test sets. Ideally, the truth data are not distinguishable from samples from the models, and the mean distance between the truth and the ensemble-mean equals that between individual members and the mean. Hence, the mean over samples of the RMSE of the ensemble-mean would be equal to that of the ensemble root mean square spread (RMSS) for an infinite ensemble. This would also be the case for subsamples with the spread falling into different bins five percentiles in width, with higher average RMSE for samples with higher average RMSS, i.e., predicted spread is higher in situations that are harder to predict, giving typically higher error. Therefore, desirable models would give points falling on the 1–1 lines in Figure 7. With ensemble sizes lower than 50, it is good practice to introduce a correction term of $n/(n-1)$ where n is the ensemble size applied to the RMSS and $n/(n+1)$ applied to the RMSE. In our case, the RMSS correction term is 1.05 and the RMSE correction term is 0.95 and including these correction terms mean that the points in Figure 7 should be on the 1–1 line.

Under the initial setup from Harris et al. (2022a), the VAEGAN struggled producing ensemble spread with the ensemble predictions being almost identical. To tackle this, we increased the number of latent variables to 100 and added three extra residual blocks before the upsampling step which allowed the model more steps to diverge the noise arrays. This dramatically increased ensemble variation of the VAEGAN and the additional residual blocks also improved the spread-error relationship for the WGAN. Figure 7a shows that overall, the WGAN displays more spread, closer to the size of the error, compared to the VAEGAN. Both models have lower spread compared to the corresponding RMSE, indicating that the variation of predictions is slightly too small. The WGAN has an RMSE on average 0.18 lower than the RMSS and the RMSE of the VAEGAN is 0.31 lower than the RMSS on average. When analyzing results from the extreme test set in Figure 7b, we see that the WGAN performance is poorer, especially for the larger percentile bins—meaning that the WGAN's ensemble predictions for the extreme cases are more underspread compared to the RMSE than they are for more regular TCs. Conversely, the VAEGAN has a much better spread-error relationship for the extreme cases compared to the regular test set in Figure 7a, though

for the most extreme cases—the 95–100th percentile bin—the VAEGAN has an RMSE of 1,210 and an RMSS of 524. While the ensemble predictions for the VAEGAN might have more appropriate spread, the predictions at the extremes are generally lower quality. This is also shown in Figures 6d–6f where predictions deviate exponentially away from observations. While the WGAN has a worse spread-error relationship in the extreme case, overall its behavior is more consistent with observations for the extreme cases than the VAEGAN across the other metrics.

Overall, the results indicate that the WGAN and VAEGAN produce predictions that are slightly underspread for storms in the regular test set, with larger errors in the extreme test set, so direct use of the samples would give an underestimate of uncertainty. This is important to test in GANs, given the potential that they will exhibit mode collapse and have too low a spread amongst samples. To our knowledge, this is the first test of this in a meteorological application.

Having an unseen extreme test set outside of the training distribution allowed us to notice that in a handful of cases, both GANs produce very high rainfall values that would not usually be observed. One possible reason for the unphysical rainfall predictions could be due to a combination of the log transform used in normalization, and the unbounded softplus activation in the output layer. This would require further investigation to be certain this is the case, which will not be explored here. We use a log transform of $\log10(1 + \text{rainfall})$ to normalize the rainfall field. The reason for this is that Harris et al. (2022a) showed it was beneficial in predicting the extreme values of rainfall, especially with sparse data sets, e.g., rainfall over the UK. Without the log normalization, the models would struggle to predict extreme rainfall and would tend to predict much drier scenarios than reality, especially with sparse data sets. However, a drawback of the log normalization is that it can be sensitive to small changes in predictions as it makes extreme rainfall values a lot closer to normal values as they otherwise would be and can predict unphysical rainfall values from small deviations in predictions. This could be exacerbated by using a softplus activation at the output layer as softplus is unbounded. These two things would make it possible for the GANs to produce the unphysically high rainfall values noticed in the extreme test set predictions. Though these high values might seem concerning, this only became noticeable when evaluating very extreme storms that the GANs have never seen before and otherwise the models are well equipped to deal with most TC rainfall values. One of the reasons, why this is less of an issue for TC rainfall, is that this data is not sparse unlike rainfall over the UK. This could allow extreme values in the predictions to be closer to usual values such that the log transform is less sensitive to small changes in predictions.

It is difficult to determine precisely why the WGAN demonstrates better overall performance compared to the VAEGAN, particularly shown in Figures 6 and 7 though it is possibly due to a few key differences in the generator networks. Figure 1 highlights this key difference in architecture. The noise channels of the WGAN are immediately concatenated with the input variables and fed into the network whereas in the case of the VAEGAN this happens after the input has passed through six residual blocks. It is plausible that passing the noise channels through the residual blocks has allowed the network more time to diversify its predictions leading to a better spread-error relationship on the standard test set. On the extreme test set, the WGAN has a lower spread compared to the VAEGAN, though this is not crucial to a good overall result, it would affect the interpretation of the spread of extreme samples from the WGAN. The WGAN and VAEGAN also have different loss functions, the WGAN uses the Wasserstein metric as a loss function, whereas the VAEGAN uses an additional Kullback-Leibler divergence term. The choice of loss function could possibly be a contributing factor to the disparity in results seen at the extremes. We observe that the Wasserstein loss of the WGAN could result in a linear relationship between input and output for out-of-training samples, which when evaluated at the extremes, the predictions align well with how precipitation behaves at higher intensities. This cannot be said for the VAEGAN, which seems to predict a nonlinear and potentially exponential relationship between inputs and outputs for the unseen extremes.

This could indicate that using the Wasserstein distance as a loss function results in more stable predictions of extreme precipitation. Further, testing is required to determine this with certainty, and is outside the scope of the original study.

3.3. WGAN Performance at Predicting Accumulated Rainfall

It is also important that models make good predictions of accumulated rainfall over the total lifespan of TCs. Given that the WGAN performed best overall for the aforementioned diagnostics, we now focus solely on that model, with results shown in Figure 8. To generate these plots, we align each TC rainfall sample from a given

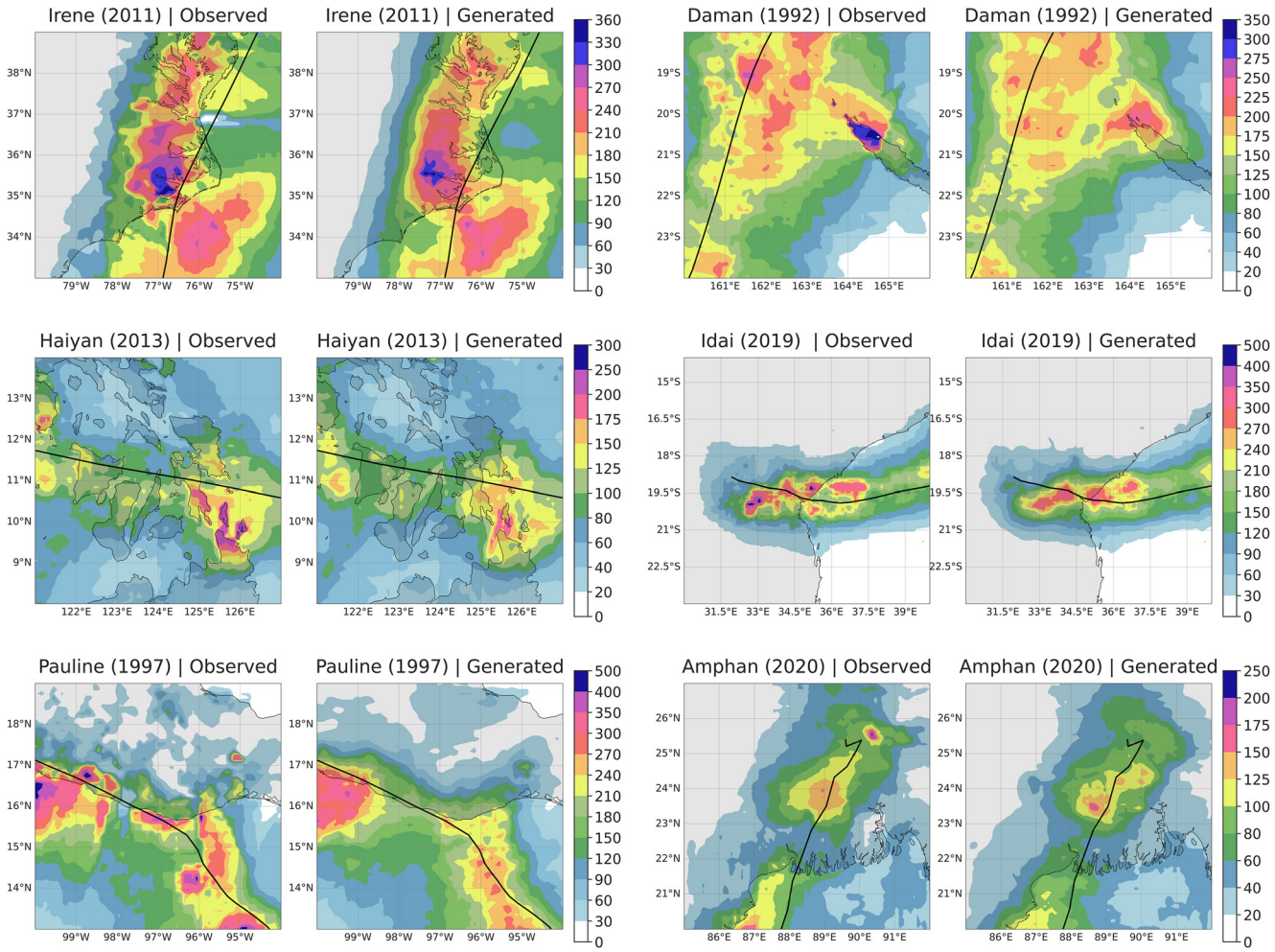

Figure 8. Accumulated rainfall estimates from Multi-Source Weighted-Ensemble Precipitation (MSWEP) observations (first and third columns) and Wasserstein Generative Adversarial Network (WGAN) predictions from the first ensemble member (second and fourth columns) for six notable tropical cyclones (TCs). The storms were chosen on the basis of being outside of the training set and across six different TC basins: Hurricane Irene (2011) over North Carolina in the North Atlantic Basin; Typhoon Haiyan (2013) over Visayas and Mindanao (The Philippines) in the West Pacific Basin; Hurricane Pauline (1997) over Oaxaca (Mexico) in the East Pacific Basin; Cyclone Daman (1992) over New Caledonia in the South Pacific Basin; Cyclone Idai (2019) over Mozambique in the South Indian Ocean; and Cyclone Amphan (2020) over Bangladesh in North Indian Ocean.

storm with its corresponding center latitude and longitude from IBTrACs tracking data and sum up rainfall over a particular region. We chose to study these six TCs as they were intense and devastating storms that were most notable for heavy rains and associated flood impacts. We chose one storm from six out of seven of the TC basins and ensured that these storms were not part of the original training data set.

The highest rainfall values are underestimated by around 50 mm in the case of Typhoon Haiyan and Cyclone Daman. New Caledonia is a mountainous region and Cyclone Daman passed by the northernmost tip, near Mont Panié (1,628 m), in 1992. Typhoon Haiyan passed through the Visayas and Mindanao islands in the Philippines. The Northern region of Mindanao, where the majority of the heavy rainfall fell, is also mountainous with elevation around 1,600 m on average.

In the MSWEP observations of Cyclone Idai, we see two pockets of high accumulated rainfall (over 400 mm) one at 33°E, 20°S and the other at 35°E, 19.25°S. The first high intensity rainfall patch falls over Chimanimani Mountains and directly over Monte Binga (2,436 m). The second patch occurs near the region of Papozo, which is a much flatter region, slowly inclining to around 500 m from the ocean. The WGAN struggles to reach these high accumulated rainfall values of above 350 mm, but has estimated the spatial patterns of rainfall up to 350 mm to a good degree.

Table 2
Total Rainfall and CRPS Skill Score to Evaluate the WGAN Rainfall Predictions for the Lifetime of Six Tropical Cyclones Taken From Outside the Original Training Set

Storm	CRPS	CRPS (land only)	Total rain (mm; MSWEP)	Total rain (mm, WGAN)	Normalized CRPS	Normalized CRPS (land only)
Amphan	3.10	0.84	204,658	206,849	2.17×10^{-6}	4.10×10^{-6}
Idai	3.78	1.71	339,491	342,965	3.72×10^{-6}	5.04×10^{-6}
Daman	2.19	0.10	25,071	31,829	1.21×10^{-6}	4.08×10^{-6}
Pauline	6.03	2.65	231,167	231,140	5.08×10^{-6}	1.14×10^{-5}
Haiyan	1.35	0.35	378,025	392,969	5.74×10^{-7}	9.23×10^{-7}
Irene	2.20	0.90	625,263	633,363	9.90×10^{-7}	1.45×10^{-6}

Note. Total rain is calculated for land only. Normalized CRPS is calculated using the predicted rainfall over all grid points and for land only, respectively.

Hurricane Pauline (1997) was an Eastern Pacific Hurricane which passed by the coast of Oaxaca (Mexico) in October 1997. The WGAN underpredicts the high accumulated rainfall totals around 98.75°W, 16.75°N near the city of St Luis Acatlán. This region is relatively flat (200 m), with peaks of around 400 m immediately surrounding the city. There are two mountain ranges within 20 km of the city one to the North West (∼1,200–1,600 m) and another (∼400–800 m) to the North East.

Cyclone Amphan made landfall as a category 2 TC near the border of India and Bangladesh, with winds of 155 km/hr (95 mph) (Mitchell et al., 2022). The WGAN misses the high rainfall patch at 90.3°E, 25.5°N which corresponds to Nokrek National Park (India) which has a mountain range (elevation ∼1,200 m) after the relatively flat land of Bangladesh (Figure 8, final 2 panels).

As a further comparison, we take our first prediction in Figure 8, Hurricane Irene, and compare with a published result of the same event, but using a different and more computationally expensive way of downscaling (Xi et al., 2020). In that study, a dynamical and statistical tropical cyclone risk model (TCRM) is developed to generate high-resolution rainfall estimates from TC wind profile data. The authors apply their model to Hurricane Irene (2017), among others, and compare their modeled estimations to the 4 km, hourly NCEP Stage IV quantitative precipitation estimation (Lin & Mitchel, 2005). Our results are plotted using the same color scale and intervals for ease of comparison. NCEP IV uses rain gauge and radar data to estimate precipitation data. In observations of Hurricane Irene from Xi et al. (2020), a maximum of 500 mm of rainfall is reached, whereas using the MSWEP data set, we only estimate up to 360 mm rainfall. Both our MSWEP analysis and Xi et al. (2020) report a rainfall high at 77°W, 35.5°N near Greenville, North Carolina. This is an area of relatively flat land not exceeding 200 m in elevation, which differs from rainfall peaks for the other storms which typically occurred over areas of high elevation relative to their surroundings. The WGAN is able to reproduce the rainfall high for this storm, estimating values above 330 mm in line with MSWEP observations, while the TCRM also obtains a maximum of around 330 mm it does not match corresponding NCEP stage IV observations of over 450 mm (Xi et al., 2020). Figure 8 shows that, in the absence of large topographic features, the WGAN produces peak predictions similar to those in the MSWEP data set. If trained on the NCEP stage IV data set, the WGAN may produce higher peak predictions that more closely match the results from that data set.

We seek to quantify the findings discussed above, by calculating the Continuous Ranked Probability Score (CRPS) of the WGAN for these six storms over all grid points and on land only, shown in Table 2. Here, we see that the CRPS for the six storms in the WGAN 20-member ensemble is systematically higher for land and ocean combined, compared to just land. Moreover, storms with higher rainfall in general have a higher CRPS value indicating that CRPS is proportional to total rainfall. Therefore, to remove this proportionality with total rainfall, we divided the CRPS by total rainfall on land and ocean combined and just land, respectively, in order to normalize for total rainfall. This allows us to see that, in general, the normalized CRPS is higher (approximately double) on land compared to ocean and land combined.

Overall, the WGAN is able to reproduce rainfall estimates for all six storms well, away from locations of peak accumulated rainfall. The peak rainfall is often underestimated, especially over mountainous land where there is systematic underestimation by the WGAN. This is evident in five of the six storms analyzed here.

4. Conclusion

In this study, we developed a series of deep learning models to increase the spatial resolution of rainfall predictions from TCs by a factor of 10. We add to the growing body of evidence that deep learning methods show potential in climate science downscaling applications by showing that such models can produce high-resolution TC rainfall predictions that have a realistic spatial structure and frequency distribution. These models could be applied to create output that is more readily usable for predicting other hazards as well as precipitation, such as storm surge or floods.

Overall, the Wasserstein GAN, based on that of Harris et al. (2022a), performed better than other methods, the VAEGAN, U-Net, and bilinear interpolation, across all diagnostics explored here. We showed that for regular TCs the WGAN had the most realistic power spectra for all wave numbers, closely followed by the VAEGAN which only deviated for scales of around five pixels or fewer. The U-Net and bilinear interpolation methods both reproduced power spectra poorly compared to observations, with significant differences present from wave numbers greater than 3. We found that the WGAN had the lowest mean bias overall with errors around the core of the TC within $\pm$5% error, while the VAEGAN had a dry bias of over 5% outside of the inner core region. Both models had a low negative bias in standard deviation of between 0% and 5%. Evaluating the quantile-quantile relationship showed how both the WGAN and VAEGAN did well at representing high percentiles of means across the core region (250-km radius) and whole storm area, but underpredicted the highest quantiles of peak precipitation values. The spread-error relationship of the WGAN was slightly better than the VAEGAN for nonextreme TCs, with both models slightly underpredicting uncertainty in their predictions.

When looking at the 100 most extreme samples, beyond the intensity of storms used in training the models, the WGAN is able to produce results of similar quality to those for TCs of intensities used in training, except for predictions having too low spread. This indicates that if the WGAN was trained on the full observational data set, it could perform well for storms more intense than those previously observed, which is important for judging the model's robustness (Watson, 2022). Conversely, the power spectra of the VAEGAN became more unrealistic and predictions more artificial. There were some very large errors present in VAEGAN at the upper end of the extreme test set which demonstrates the importance of evaluating models on the most extreme, unseen, cases. This study highlights the importance of evaluating models on an unseen extreme distribution subsequent to any model development, which is something that to the best of our knowledge, has not previously been studied in this context. With rainfall from TCs likely to increase with global warming (Knutson et al., 2019), it is important to evaluate the extreme events when considering the application to future individual storms that could potentially generate more rainfall than any in history. As such, we have high confidence in the WGAN making out-of-distribution predictions for coarse-resolution input values up to 72 mm/hr.

When using the WGAN to estimate TC rainfall accumulated over land, we found that it struggled the most at reproducing rainfall over mountainous regions, often failing to achieve the highest rainfall values over land above 1,000 m in elevation. Over flat land, the WGAN was able to reproduce accumulated rainfall peaks well. This could likely be improved in future by using topography as an input variable to give more accurate representations of TC rainfall over land. In general, having additional variables would likely improve the robustness of the models. For example, variables such as wind shear, wind profile, and specific humidity would provide the models with additional information and possibly lead to a tangible improvement. This study was designed to explore model capability for a classical image superresolution problem, which is why we only address rainfall-to-rainfall downscaling. Future work aims to investigate which additional variables add the most value and would be best represented in climate models.

It is not clear to what extent using input data from a different source, rather than being a coarsened version of the high-resolution data, would impact model robustness. It is plausible that using low-resolution input data from a different data set could result in poorer performance overall as biases between different data sets present an additional challenge for learning. Though not explored here, techniques exist to correct for biases between different data sets such as the two-step downscaling method developed in Price and Rasp (2022). These would be worthy of exploring in future work should data be used from different sources.

In future work, we will explore whether this method could be valuable in estimating high-resolution TC rainfall using GCM data as input, as this could allow us to analyze sufficient extreme events for risk quantification and generate regional hazard maps for any region in the world. The direction this takes would be conditional on how

well the GCMs simulate the spatial distribution and magnitude of TC rainfall. Overall, these developments will improve climate change attribution and global TC risk quantification. This will help a more efficient distribution of funding to the most at-risk areas and increase resilience to climate change with decision-making that is informed.

Data Availability Statement

The trained WGAN used for downscaling tropical cyclone rainfall (Vosper et al., 2023a) is preserved at https://zenodo.org/deposit/7305257 (Vosper et al., 2023b), available via the Creative Commons Attribution 4.0 International license and developed openly at https://github.com/vosps/tropical_cyclone. For access to the public version of the WGAN (dsrnngan) developed in Harris et al. (2022a) please see https://github.com/ljharris23/public-downscaling-cgan (Harris et al., 2022b).

Acknowledgments
This project was supported by a NERC Independent Research Fellowship (Grant NE/S014713/1). Emily Vosper was supported by the UKRI Centre for Doctoral Training in Interactive Artificial Intelligence under Grant EP/S022937/1. Raul Santos-Rodriguez was funded by the UKRI Turing AI Fellowship EP/V024817/1. The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement 741112).

References

Adewoyin, R. A., Dueben, P., Watson, P., He, Y., & Dutta, R. (2021). TRU-NET: A deep learning approach to high resolution prediction of rainfall. *Machine Learning*, *110*(8), 2035–2062. https://doi.org/10.1007/s10994-021-06022-6

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *Paper presented at 34th International Conference on Machine Learning (ICML 2017)* (Vol. 1, pp. 298–321).

Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & De Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Brackins, J. T., & Kalyanapu, A. J. (2020). Evaluation of parametric precipitation models in reproducing tropical cyclone rainfall patterns. *Journal of Hydrology*, *580*, 124255. https://doi.org/10.1016/j.jhydrol.2019.124255

Emanuel, K. (2017). Assessing the present and future probability of Hurricane Harvey's rainfall. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(48), 12681–12684. https://doi.org/10.1073/pnas.1716222114

Emanuel, K., Sundararajan, R., & Williams, J. (2008). Hurricanes and global warming: Results from downscaling IPCC AR4 simulations. *Bulletin of the American Meteorological Society*, *89*(89), 347–368. https://doi.org/10.1175/BAMS-89-3-347

Feldmann, M., Emanuel, K., Zhu, L., & Lohmann, U. (2019). Estimation of Atlantic tropical cyclone rainfall frequency in the United States. *Journal of Applied Meteorology and Climatology*, *58*(8), 1853–1866. https://doi.org/10.1175/JAMC-D-19-0011.1

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. *Paper presented at Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 5769–5779). https://doi.org/10.3997/2214-4609.201405839

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022a). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003120. https://doi.org/10.1029/2022ms003120

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022b). A generative deep learning approach to stochastic downscaling of precipitation forecasts [Software]. Retrieved from https://github.com/ljharris23/public-downscaling-cgan

Huang, X. (2020). Deep-learning based climate downscaling using the super-resolution method: A case study over the western US. *Geoscientific Model Development Discussions*, 1–18.

Jing, R., Lin, N., Emanuel, K., Vecchi, G., & Knutson, T. R. (2021). A comparison of tropical cyclone projections in a high-resolution global climate model and from downscaling by statistical and statistical-deterministic methods. *Journal of Climate*, *34*(23), 9349–9364. https://doi.org/10.1175/JCLI-D-21-0071.1

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., & Schreck, C. J. (2018). *International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4*. NOAA National Centers for Environmental Information. https://doi.org/10.25921/82ty-9e16

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The International Best Track Archive for Climate Stewardship (IBTrACS). *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Knutson, T., Camargo, S. J., Chan, J. C., Emanuel, K., Ho, C. H., Kossin, J., et al. (2019). Tropical cyclones and climate change assessment. *Bulletin of the American Meteorological Society*, *100*(10), 1987–2007. https://doi.org/10.1175/BAMS-D-18-0189.1

Kumar, B., Chattopadhyay, R., Singh, M., Chaudhari, N., Kodari, K., & Barve, A. (2021). Deep learning-based downscaling of summer monsoon rainfall data over Indian region. *Theoretical and Applied Climatology*, *143*(3–4), 1145–1156. https://doi.org/10.1007/s00704-020-03489-6

Langousis, A., & Veneziano, D. (2009). Theoretical model of rainfall in tropical cyclones for the assessment of long-term risk. *Journal of Geophysical Research*, *114*, D02106. https://doi.org/10.1029/2008JD010080

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. Paper presented at *Proceedings at 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)* (pp. 105–114). https://doi.org/10.1109/CVPR.2017.19

Leinonen, J., Nerini, D., & Berne, A. (2020). Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211–7223. https://doi.org/10.1109/tgrs.2020.3032790

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Lin, Y., & Mitchel, K. (2005). The NCEP stage II/IV hourly precipitation analyses: Development and applications. Retrieved from https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm

Lonfat, M., Rogers, R., Marchok, T., & Marks, F. D. (2007). A parametric model for predicting hurricane rainfall. *Monthly Weather Review*, *135*(9), 3086–3097. https://doi.org/10.1175/MWR3433.1

Lu, P., Lin, N., Emanuel, K., Chavas, D., & Smith, J. (2018). Assessing hurricane rainfall mechanisms using a physics-based model: Hurricanes Isabel (2003) and Irene (2011). *Journal of the Atmospheric Sciences*, *75*(7), 2337–2358. https://doi.org/10.1175/JAS-D-17-0264.1

Mitchell, D., Hawker, L., Savage, J., Bingham, R., Lord, N. S., Khan, M. J. U., et al. (2022). Increased population exposure to Amphan-scale cyclones under future climates. *Climate Resilience and Sustainability*, *1*(2), 1–16. https://doi.org/10.1002/cli2.36

Murakami, H., Vecchi, G. A., Underwood, S., Delworth, T. L., Wittenberg, A. T., Anderson, W. G., et al. (2015). Simulation and prediction of category 4 and 5 hurricanes in the high-resolution GFDL HiFLOR coupled climate model. *Journal of Climate*, *28*(23), 9058–9079. https://doi.org/10.1175/JCLI-D-15-0216.1

Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *Machine Learning*. http://arxiv.org/abs/2203.12297

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *Paper presented at 4th International Conference on Learning Representations (ICLR 2016)—Conference Track Proceedings* (pp. 1–16).

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Rezapour, M., & Baldock, T. E. (2014). Classification of hurricane hazards: The importance of rainfall. *Weather and Forecasting*, *29*(6), 1319–1331. https://doi.org/10.1175/WAF-D-14-00014.1

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vannière, B., et al. (2020). Projected future changes in tropical cyclones using the CMIP6 HighResMIP multimodel ensemble. *Geophysical Research Letters*, *47*, e2020GL088662. https://doi.org/10.1029/2020GL088662

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, *9351*, 234–241. https://doi.org/10.1109/ACCESS.2021.3053408

Sha, Y., Gagne, D. J., West, G., & Stull, R. (2020). Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *Journal of Applied Meteorology and Climatology*, *59*(12), 2075–2092. https://doi.org/10.1175/JAMC-D-20-0058.1

Steptoe, H., Savage, N. H., Sadri, S., Salmon, K., Maalick, Z., & Webster, S. (2021). Tropical cyclone simulations over Bangladesh at convection permitting 4.4 km & 1.5 km resolution. *Scientific Data*, *8*, 62. https://doi.org/10.1038/s41597-021-00847-5

Tuleya, R. E., DeMaria, M., & Kuligowski, R. J. (2007). Evaluation of GFDL and simple statistical model rainfall forecasts for U.S. landfalling tropical storms. *Weather and Forecasting*, *22*(1), 56–70. https://doi.org/10.1175/WAF972.1

Vosper, E., Mitchell, D., & Emanuel, K. (2020). Extreme hurricane rainfall affecting the Caribbean mitigated by the Paris agreement goals. *Environmental Research Letters*, *15*(10), 104053. https://doi.org/10.1088/1748-9326/ab9794

Vosper, E., Watson, P., Harris, L., McRae, A., Santos-Rodriguez, R., Aitchison, L., & Mitchell, D. (2023a). WGAN for downscaling tropical cyclone rainfall[Software]. Retrieved from https://github.com/vosps/tropical_cyclone

Vosper, E., Watson, P., Harris, L., McRae, A., Santos-Rodriguez, R., Aitchison, L., & Mitchell, D. (2023b). WGAN for downscaling tropical cyclone rainfall (pre-trained)[Software]. Retrieved from https://zenodo.org/deposit/7305257

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, *57*(4), 1–21. https://doi.org/10.1029/2020WR029308

Watson, P. (2022). Machine learning applications for weather and climate need greater focus on extremes [eprint]. arXiv:2207.07390(MI)

Xi, D., Lin, N., & Smith, J. (2020). Evaluation of a physics-based tropical cyclone rainfall model for risk assessment. *Journal of Hydrometeorology*, *21*(9), 2197–2218. https://doi.org/10.1175/JHM-D-20-0035.1

Zhang, W., Villarini, G., Scoccimarro, E., Roberts, M., Vidale, P. L., Vanniere, B., et al. (2021). Tropical cyclone precipitation in the High-ResMIP atmosphere-only experiments of the PRIMAVERA Project. *Climate Dynamics*, *57*(1–2), 253–273. https://doi.org/10.1007/s00382-021-05707-x

Zhu, L., Quiring, S. M., & Emanuel, K. A. (2013). Estimating tropical cyclone precipitation risk in Texas. *Geophysical Research Letters*, *40*, 6225–6230. https://doi.org/10.1002/2013GL058284