JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2024MS004796

Postprocessing East African Rainfall Forecasts Using a Generative Machine Learning Model

Bobby Antonio[1,2], Andrew T. T. McRae[1], David MacLeod[3], Fenwick C. Cooper[1], John Marsham[4], Laurence Aitchison[5], Tim N. Palmer[1], and Peter A. G. Watson[2]

[1]Department of Physics, University of Oxford, Oxford, UK, [2]School of Geographical Sciences, University of Bristol, Bristol, UK, [3]School of Earth and Environmental Sciences, University of Cardiff, Cardiff, UK, [4]School of Earth and Environment, University of Leeds, Leeds, UK, [5]Machine Learning and Computational Neuroscience Unit, University of Bristol, Bristol, UK

Key Points:

- We combine generative machine learning with conventional postprocessing to improve short range precipitation forecasts over East Africa
- The postprocessed forecasts show substantial improvements compared to a strong baseline, particularly for the diurnal cycle
- Improvements persist when evaluated against an extreme rainfall season, indicating an ability to extrapolate to more extreme conditions

Correspondence to:
B. Antonio,
bobby.antonio@physics.ox.ac.uk

Citation:
Antonio, B., McRae, A. T. T., MacLeod, D., Cooper, F. C., Marsham, J., Aitchison, L., et al. (2025). Postprocessing East African rainfall forecasts using a generative machine learning model. *Journal of Advances in Modeling Earth Systems*, *17*, e2024MS004796. https://doi.org/10.1029/2024MS004796

Received 30 OCT 2024
Accepted 30 JAN 2025

Author Contributions:
Conceptualization: Peter A. G. Watson
Data curation: David MacLeod
Formal analysis: Bobby Antonio
Funding acquisition: Tim N. Palmer, Peter A. G. Watson
Investigation: Bobby Antonio
Methodology: Bobby Antonio, Andrew T. T. McRae, David MacLeod, Fenwick C. Cooper, John Marsham, Laurence Aitchison, Peter A. G. Watson
Project administration: Peter A. G. Watson
Resources: Peter A. G. Watson
Software: Bobby Antonio, Andrew T. T. McRae
Supervision: Laurence Aitchison, Tim N. Palmer, Peter A. G. Watson
Validation: Bobby Antonio

Abstract Existing weather models are known to have poor skill at forecasting rainfall over East Africa. Improved forecasts could reduce the effects of extreme weather events and provide significant socioeconomic benefits to the region. We present a novel machine learning (ML)-based method to improve precipitation forecasts in East Africa, using postprocessing based on a conditional generative adversarial network (cGAN). This addresses the challenge of realistically representing tropical rainfall, where convection dominates and is poorly simulated in conventional global forecast models. We postprocess hourly forecasts made by the European Centre for Medium-Range Weather Forecasts Integrated Forecast System at 6–18 hr lead times, at 0.1° resolution. We combine the cGAN predictions with a novel neighborhood version of quantile mapping, to integrate the strengths of ML and conventional postprocessing. Our results indicate that the cGAN substantially improves the diurnal cycle of rainfall, and improves predictions up to the 99.9th percentile (∼ 10mm/hr). This improvement extends to the March–May 2018 season, which had extremely high rainfall, indicating that the approach has some ability to generalize to more extreme conditions. We explore the potential for the cGAN to produce probabilistic forecasts and find that the spread of this ensemble broadly reflects the predictability of the observations, but is also characterized by a mixture of under- and over-dispersion. Overall our results demonstrate how the strengths of ML and conventional postprocessing methods can be combined, and illuminate what benefits ML approaches can bring to this region.

Plain Language Summary Weather forecasts over tropical areas are typically not very good at predicting how heavy rain will be and exactly when it will fall. This is particularly a problem for parts of East Africa, where heavy rainfall can lead to negative impacts such as flooding. We investigate whether forecasts over East Africa can be improved using a combination of machine learning (ML) and more traditional techniques, where forecast corrections are learned from satellite observations. We show these methods substantially improve forecasts errors, particularly in the timing of rainfall. The ML model performs well even when forecasting over an extremely wet season.

1. Introduction

East Africa experiences highly variable rainfall, which can have negative impacts on the region, such as severe droughts that cause famine (Gebremeskel Haile et al., 2019) and extreme rainfall leading to floods that significantly impact those living in the region (Kilavi et al., 2018; Wainwright et al., 2021). For example, flooding of the Shabelle river in Somalia in March 2023 affected an estimated 460,000 people (Floodlist, 2023), and an estimated 3,000–5,000 people die on Lake Victoria every year as a result of storms that capsize or damage boats (IFRC, 2014; Watkiss et al., 2020).

Accurate rainfall forecasts are therefore crucial for attempts to mitigate the adverse effects of rainfall, through enabling more accurate early warning systems, more timely disaster response, and agricultural planning. A recent study as part of the WISER project estimated that their work improving early warning systems brought benefits of 3 m/yr to coastal Tanzania alone, plus additional intangible improvements to the well-being and safety of people living in the area (Watkiss & Cimato, 2021). Heavy rainfall advisories in Kenya have been demonstrated to be effective at predicting impactful rainfall events, especially over recent years, although there is still a need for increased resolution in the forecasts in order to better enable initiatives such as forecast-based financing (MacLeod, Dankers, et al., 2021; MacLeod, Kilavi, et al., 2021).

Visualization: Bobby Antonio
Writing – original draft: Bobby Antonio
Writing – review & editing: Bobby Antonio, David MacLeod, Fenwick C. Cooper, John Marsham, Laurence Aitchison, Peter A. G. Watson

Conventional forecast products that use convection parameterization tend to predict too many low intensity rainfall events and perform poorly at predicting heavy rainfall (Bechtold et al., 2014; Chamberlain et al., 2014; Haiden et al., 2012; Vogel et al., 2018, 2020; Woodhams et al., 2018). They also do not model the diurnal cycle of rainfall well (Bechtold et al., 2004; Kim & Joan Alexander, 2013; MacLeod, Dankers, et al., 2021; MacLeod, Kilavi, et al., 2021), which is likely because of the convective parameterization schemes used in the models (Bechtold et al., 2014; Marsham et al., 2013).

In recent years, it has become computationally feasible to run "convection permitting" (CP) models at higher resolutions (∼4 km) for which the model can better capture convection processes without using a parameterization scheme (Cafaro et al., 2021; Chamberlain et al., 2014; Clark et al., 2016; Warner et al., 2023; Woodhams et al., 2018). Whilst these CP models significantly improve the diurnal-timing of rainfall (Marsham et al., 2013), they also still have remaining biases, such as a tendency to produce too intense rainfall (Woodhams et al., 2018), and still show some discrepancies in the diurnal cycle for particular areas in East Africa (Cafaro et al., 2021; Chamberlain et al., 2014). In addition, these models are currently too expensive to run in large ensembles, and so are not currently cost-effective replacements for existing ensemble models.

At the same time machine learning (ML) models have improved dramatically over the last decade at a range of tasks, including generating realistic images (Karras et al., 2019). This has inspired several attempts to train ML models to predict the weather from large amounts of historical data (Bi et al., 2023; Lam et al., 2023; Nguyen et al., 2023; Ravuri et al., 2021; Zhang et al., 2023) and for bias correction (Ben-Bouallegue et al., 2023; Chen et al., 2024; Dai & Hemri, 2021; Rasp & Lerch, 2018).

One of the most widely used ML models that can learn to sample realistic images from a target distribution is a Generative Adversarial Network or GAN (Goodfellow et al., 2014). Several works have demonstrated the effectiveness of GANs in a range of tasks; in Leinonen et al. (2020) a GAN was trained to downscale low resolution observations. Harris et al. (2022) and Price and Rasp (2022) then applied the same approach to the problem of downscaling coarse resolution forecasts toward radar observations. There are several works that have used GANs to postprocess precipitation forecasts: for example, Duncan et al. (2022) and Dai and Hemri (2021) use a GAN to postprocess the output of a ML model, Jeong and Yi (2023) used a cyclical GAN to perform corrections of precipitation forecasts over South Korea, and Chen et al. (2024) train a generative moment matching network to postprocess.

However, only a small amount of effort has been applied to developing these techniques in tropical regions, such as regions in Africa (see e.g., Vogel et al., 2020; Walz et al., 2024), for which the need for improved forecasting is often greater, and which historically have not seen the same forecast skill improvements as mid-latitude areas like Europe and North America (Youds et al., 2021). This raises the question; can we leverage recent advances in ML to improve forecasts in tropical regions such as East Africa? In this work we investigate this by training a Generative Adversarial Network to postprocess short range NWP rainfall forecasts in East Africa at lead times of 6–18 hr, to see whether this can provide an effective means of correcting existing forecasts. We use forecasts from the state-of-the-art European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) as the NWP product. In contrast to previous work on ML-based forecasting, we consider the challenging problem of forecasting rainfall in a tropical region, and evaluate whether a GAN is able to learn the properties of rainfall in a region where convection is dominant. We also combine the ML with a novel quantile mapping approach, in order to leverage the strengths of both conventional and ML methods. We compare against ECMWF IFS forecasts, with quantile mapping also applied—this gives a baseline that is stronger than unmodified forecasts as commonly used in previous work on ML-based forecasting (e.g., Bi et al., 2022), giving a stronger test of whether ML can improve upon the state-of-the-art.

This paper is structured as follows: In Section 2 we provide details of the region, data, ML model, and forecast verification metrics used. In Section 3 we evaluate our approach: Sections 3.1–3.3 present results for a year with typical weather conditions, whilst in Section 3.4 we present an analysis over a particularly heavy Long Rains season over Kenya, to evaluate the model's ability to extrapolate to events more intense than any it has seen in training, helping address the question of whether ML models can perform well for extreme events (Watson, 2022).

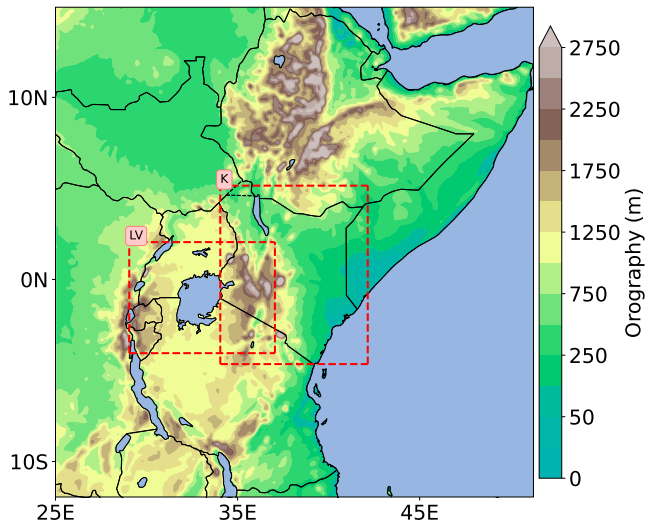

Figure 1. The region we consider in this study. The filled contours show the orography in land regions in meters. The red dashed box labeled LV outlines the Lake Victoria subdomain (4°S − 2°N and 29° − 37°E) analyzed for diurnal cycle in Section 3. The red dashed box labeled K outlines the Kenya subdomain (approx. 5°S − 5°N and 34° − 42°E) analyzed over the March–May 2018 rainy season in Section 3.4.

2. Methods

2.1. Region

The region we consider is 12°S − 15°N, and 25° − 51°E, roughly centered around Kenya and Lake Victoria, sometimes referred to as Equatorial East Africa (see Figure 1). The rainfall characteristics in this region can broadly be divided into two regions; a "summer rainfall" region (containing e.g., South Sudan, North-Western Ethiopia, Djibouti, and Coastal regions) and an "equatorial rainfall" region (containing e.g., Kenya, Uganda, Northern Tanzania and Southern Ethiopia) (Nicholson, 2017). Within the summer rainfall region the rain mainly falls in the boreal summer, whilst the equatorial rainfall region has two rainy seasons; the "long" rains in March–May, and the "short" rains in October–December. The long rains tend to be the wettest season, with less interannual variability but more intraseasonal variability, and is generally regarded as the season with lowest seasonal forecast skill (Kilavi et al., 2018; Nicholson, 2017; Walker et al., 2019).

Freshwater lakes, such as Lake Victoria and Lake Tanganyika, amongst the largest freshwater lakes in the world, have a significant impact on the surrounding weather. Around Lake Victoria there is rain for much of the year with storms frequently occurring on or near the lake (Chamberlain et al., 2014; MacLeod, Dankers, et al., 2021; MacLeod, Kilavi, et al., 2021; Woodhams et al., 2019). For our analysis later on, we define a region around Lake Victoria, as 4°S–2°N, 29°E–37°E; this is based to roughly align with the area containing storm tracks and wind anomalies in Woodhams et al. (2019), the area used in Finney et al. (2019), and the area identified in Thiery et al. (2015) as being most impacted by the presence of the lake.

There are also interesting orographic features in the area, such as Mt. Kilimanjaro and Mt. Kenya, and mountains extending along the East African rift from the Ethiopian highlands down either side of Lake Victoria (which we refer to as the Rift Valley in this work). A gap in this range, the Turkana channel, extending from Northwest Kenya to South Sudan, is another important feature in this area that affects moisture transport via the Turkana jet that flows through it (Nicholson, 2016).

2.2. Data

For observational data, we use the Integrated Multi-satellite Retrievals for GPM (IMERG) V06B data set from 2016 to 2021; this is calculated through satellite observations of microwave and infrared by the array of Global Precipitation Measurement (GPM) satellites, with subsequent calibration incorporating rain gauges (Huffman et al., 2022). The observations are half-hourly at a resolution of 0.1° × 0.1° (11 km × 11 km at the equator), with precipitation rates given in mm/hr representing the average rainfall over the entire half hour period. For our purposes the data is coarsened to hourly resolution.

Whilst the IMERG data does not perfectly represent the true rainfall, it performs reasonably well at capturing the diurnal cycle and distribution of rainfall in East Africa (Camberlin et al., 2018; Dezfuli et al., 2017; Roca et al., 2010) compared to other alternatives, particularly given the scarcity of reliable and widely available rain gauge and radar data in the area. In Ageet et al. (2022) a range of satellite-based rainfall estimates, including the IMERG product, were compared with rain gauges in an area around Uganda (including parts of Kenya, Tanzania, Sudan, and the Democratic Republic of Congo) over 17 years. Based on a combined assessment of quantile-quantile plots, correlation, and skill scores such as hit rate (HR) and false alarms, they identified the IMERG product (V06B) as the best performing at daily resolution. However, it still has biases; for example, it has a tendency to underestimate the rainfall rate, and over-predict the frequency of rainfall. There are also known issues with similar satellite products in mountainous areas (Dinku et al., 2010), which means these observations may be more unreliable over areas such as the Ethiopian Highlands and parts of the Rift Valley. A dry bias has also been observed in several studies (e.g., Vogel et al., 2018). Overall, though, it provides a good source of data with a high temporal and spatial resolution over our target region, and it has been used in other studies in this area (e.g., Cafaro et al., 2021; Finney et al., 2019; Woodhams et al., 2018).

The forecast data set used is the ECMWF IFS HRES deterministic hourly forecast (ECMWF, 2023a, 2023b, 2023c) as this tends to perform amongst the best compared to similar models (Haiden et al., 2012). IFS forecasts are provided at 00 and 12 hr and we use lead times within a 6–18 hr window, corresponding to short-range weather prediction (however it is expected that the method we use could also equally apply to longer lead times, see e.g., Yang et al., 2025). The data is interpolated from 9 km × 9 km resolution to 0.1° × 0.1° to match the grid points of the IMERG precipitation. The data starts at March 2016, after the increase in horizontal resolution for the IFS with the release of Cycle 41r2. To ensure the precipitation forecasts are reasonably consistent, we use data up until September 2021 before the upgrade to the convection parameterization scheme with the release of Cycle 47r3 in October 2021 (ECMWF, 2023a, 2023b, 2023c).

2.3. Machine Learning Model

2.3.1. Training and Evaluation Strategy

For training and evaluating the model, the data set was split up as follows:

- Training set: March 2016–February 2018 and July 2018–September 2020, excluding validation months (3.7×10^4 samples).
- Validation set: June 2018, October 2018, January 2019, and March 2019 (2.9×10^3 samples).
- Test sets: October 2020–September 2021 (8.6×10^3 samples) and March–May 2018 (2.1×10^3 samples).

We used the October 2020–September 2021 year as the primary test data set, and the 2018 long rains (March–May) as an extreme test set, since this was a season of particularly heavy rainfall (Kilavi et al., 2018) for which the March–May rainfall was significantly higher than in any other season in the full IMERG data set.

The purpose of the validation data set is to guide choice of the model structure and hyperparameters. The development process was to train different versions of the model on the training data set, then evaluate these on the validation data set to select the best version. This avoids overfitting on the test data by selecting a model that performs well by chance. The standard choice of validation set would be the period October 2019–September 2020, which spans a full year and would sit between the training and test periods. However, the rains of October–December 2019 were exceptionally high (Wainwright et al., 2021), as were the long rains of March–May 2020 (Palmer et al., 2023). So to avoid validation over an atypical year, which may have given an inaccurate assessment of the model's general performance, we chose to validate over the period of June 2018–May 2019. Rather than use a full year for validation, we also chose to maximize the amount of training data by including a month from each of the different seasons in the validation period. This sampling variability observed for this size of validation data also indicated there was no additional benefit from using a whole year.

All evaluation results reported in Sections 3.1 and 3.2 are evaluated on all of the data from the unseen test period October 2020–September 2021, with 20 ensemble members used in the example plots. The ensemble calibration results in Section 3.3 are assessed over 500 unique hours sampled uniformly from the same period with an ensemble size of 100.

For the extreme rainfall evaluation in Section 3.4, we analyze all of the hours from March to May 2018. Since much of the anomalous rainfall in this season was concentrated over Kenya leading to flooding (Kilavi et al., 2018), we restrict our analysis to this region (4.6°S − 5.1°N and 34.0° − 42.1°E, see Figure 1).

2.3.2. Model Description

Our model uses similar architecture and code to that Harris et al. (2022) used to postprocess UK rainfall forecasts. This is itself based on Leinonen et al. (2020) and a variant was developed for downscaling tropical cyclone rainfall by Vosper et al. (2023). A conditional Wasserstein GAN is trained to predict realistic rainfall patterns conditioned on several meteorological inputs together with constant inputs such as orography, using radar or satellite observations as ground truth. We use the same approach to test whether it will transfer to also perform well at postprocessing forecasts in a tropical domain.

Both the generator and discriminator of the GAN are deep neural networks, primarily made up of residual blocks, where each residual block contains two convolution layers that use square convolutional kernels of width 3 pixels (see e.g., Goodfellow et al., 2016 for background on convolutional neural networks). The generator is composed of 7 residual blocks (each with f_g filters), with a final softplus activation function, giving a total of $2 \times 7 \times f_g$

intermediate arrays each of size 200 × 200. The discriminator is made up of 3 residual blocks (each with f_d filters), and two dense layers, giving a total of $2 \times 3 \times f_d$ intermediate arrays each of size 200 × 200. Excluding the output layers, Parametric Rectified Linear Unit activation functions were used, which have a parameter α that controls how negative activations are weighted (He et al., 2015); we set the α parameter to 0.2 following Harris et al. (2022). The number of noise channels was set to 4, and the learning rates for the generator and discriminator were set equal to 1×10^{-5}, with the discriminator being trained for 5 steps for every 1 step of generator training. The batch size was set to 2 based on hardware memory constraints, and the Adam optimizer was used for training.

Following Harris et al. (2022) we use a Wasserstein GAN (Arjovsky et al., 2017), which has been demonstrated to improve training stability in many cases (Creswell et al., 2018). This modifies the GAN discriminator to output low numbers for real samples and high numbers for fake samples, rather than producing a number in the range [0, 1], and modifies the loss function to approximate the Wasserstein distance between the generated and true distributions (Gulrajani et al., 2017). This approximation is parameterized by a gradient penalty parameter γ which we set to 10 in line with Gulrajani et al. (2017).

In order to perform shorter experiments to tune hyperparameters, smaller models with $f_g = 32, f_d = 128$ were trained for 6.4×10^4 iterations, and then finally larger models with $f_g = 64, f_d = 256$ were trained for 3.2×10^5 steps, and used for evaluation; thus our largest model was smaller than the model in Harris et al. (2022) that had $f_g = 128, f_d = 512$. However, since the model is not being used for downscaling, and because we use a larger domain, the model dimensions scale differently, and so using a smaller number of channels was required to achieve a reasonable training time.

As inputs, the model uses IFS forecast variables from the same time as the target rainfall forecasts. On top of the 9 variables used in Harris et al. (2022), we used 11 extra fields, including temperature, convective precipitation, vertical velocity, and relative humidity (some of which are at several pressure levels; see Appendix A for a full table of inputs). Convective inhibition was included, with null values set to 0. We included these extra variables as they contain important information about convective processes, which are critical for forecasting in East Africa. Based on the transformations applied in Harris et al. (2022), we normalize the input variables; precipitation variables are log-normalized via $x \rightarrow \log_{10}(1 + x)$, whilst others were either divided by the maximum value, or normalized to fall within the maximum and minimum values (see Appendix A).

Model checkpoints were saved every 3,200 steps, and the best model in the last 1/3rd of checkpoints was selected based on performance on the validation set. This performance was assessed through judgment of the combined performance on CRPS, Radially Averaged Power Spectral Density (RAPSD) and mean squared error, plus visual evaluation of the samples produced (wherein the realism of the samples and colocation with observations was assessed by eye). Our batch size was limited to 2 due to the need to generate an ensemble to calculate part of the loss function (discussed in the next paragraph). All models were trained on a single Nvidia A100 GPU. Post-processing 24 hourly forecast samples using this hardware takes approximately 25 s.

One notable addition by Harris et al. (2022) is the inclusion of a "content loss" term, inspired by Ravuri et al. (2021), which penalizes GAN predictions that do not have an ensemble mean close to the observed value. Specifically, at each training step the generator produces an ensemble of predictions (set to 8 in this work), and the generator loss function includes a mean-squared error term between the observed image and the ensemble mean of the generated samples.

During validation of the models, we observed that using log normalization of the output precipitation predictions, as done by Harris et al. (2022), produced a distribution of rainfall that tended to greatly overestimate the observations at the extreme rainfall values. Removing the log normalization of the output rainfall values remedied this, and also removed the need to clip the predicted rainfall values to a given maximum, as done in Harris et al. (2022). This also required modifying the content loss parameter λ, with $\lambda = 100$ appearing to produce the best results according to a joint assessment of quantile-quantile plots, CRPS and Radially Averaged Log Spectral Distance (RALSD) (see Appendix B). Since this modification reduces the number of unphysical values produced by the model, and removes the need for a tuned maximum rainfall parameter, it is possible that this modification would also make an improvement to the UK model in Harris et al. (2022). However, we may expect the long-tail distributions of rainfall in the UK and East Africa to be different, such that the log normalization could still be an effective choice for UK models.

In Harris et al. (2022), samples are grouped into predefined bins based on the fraction of grid points exceeding a set threshold, and then at training time samples are drawn more frequently from the high rainfall bins, in order to oversample higher rainfall values and improve performance in these cases. However, due to the substantially different rainfall characteristics of the regions, the threshold used by Harris et al. (2022) for postprocessing UK rainfall could not be directly applied to the East Africa model, and our attempts to use a similar approach did not improve scores on the validation set. Therefore, we did not apply this oversampling, but given the relative scarcity of rainfall in this region a more nuanced oversampling technique could potentially help with the underprediction of the most intense rainfall in the unpostprocessed conditional generative adversarial network (cGAN).

To increase the variation in the samples seen during training, we randomly cropped the 270 × 265 images to smaller images of 200 × 200, as this has been demonstrated to improve the generalizability of deep learning models (Goodfellow et al., 2016), and produces output similar in size to that in Harris et al. (2022).

Similarly to the results in Harris et al. (2022), we observed that the model skill can vary considerably between training steps (as measured by CRPS, RAPSD, and visual inspection), so it was necessary to have a validation data set set aside in order to choose the final model, and this data was not used in the final training.

To summarize the changes made to the model relative to Harris et al. (2022); we use the model to postprocess East African rainfall forecasts rather than downscale, using different inputs that we expect to be more appropriate to forecasting convective rainfall. The output normalization has been modified to remove the problem of unrealistically high rainfall values, and the training method changed so that rainy periods are no longer oversampled.

2.4. Quantile Mapping

Since it is known that IFS forecasts with postprocessing outperform those without in this region (Vogel et al., 2020), and IFS forecasts are not specifically tuned to reproduce the properties of IMERG observations, we applied quantile mapping to the IFS forecasts (see e.g., Maraun & Widmann, 2017) to provide a stronger baseline.

Additionally, since GAN predictions are not guaranteed to precisely capture the rainfall distribution, and we observed that our GAN predictions tended to under-predict high rainfall values, we produced a variant of our model with quantile mapping applied to the output. In doing so we aimed to combine the strengths of both postprocessing methods to achieve an overall more accurate and realistic forecast. The GAN could be expected to perform well at producing predictions with realistic spatial structure, but not necessarily with realistic point frequency distributions. Quantile mapping can greatly improve the latter.

We used empirical quantile mapping rather than a distribution-based quantile mapping approach, since it has been demonstrated to work well (Gudmundsson et al., 2012), and does not require a parametric distribution. Our method is based on the well-used methods outlined in Boé et al. (2007), Déqué (2007), and Maraun and Widmann (2017), in which empirical cumulative density functions are calculated over the training period and used to create a mapping between the forecast quantiles and the observed quantiles. In general this means creating an estimate of the cumulative density functions F_f and F_o of the forecast and observations respectively, and mapping the forecast values x_f to an adjusted value $\tilde{x}_f$ according to:

$$\tilde{x}_f = F_o^{-1}\left(F_f\left(x_f\right)\right) \tag{1}$$

In Boé et al. (2007), percentiles at 1% spacing are first calculated on the training set to find an approximation to the quantile distributions. Forecast values are then mapped into quantile values relative to the training data, and then converted into adjusted forecast values using the observed quantile values (using linear interpolation when the quantile falls between the known quantile values).

Since the East African precipitation is low, there can be multiple quantiles that are 0; therefore for a forecast of 0 mm/hr there is no way to tell which quantile it belongs to. We follow the method in Boé et al. (2007) and pick one of the 0-valued forecast quantiles at random, then assign the value of the matching observational quantiles. In practise, this can lead to low level noise on the corrected forecast, but replicates the high level statistics.

In order to better match the tail of the frequency distribution in our work, the step size between the quantiles was decreased toward the higher quantiles; so a step size of 0.01 was used up to 0.99, a step size of 0.001 used from 0.99 to 0.999, and so on up to the 99.9999^{th} percentile, above which we observed significant sampling variability.

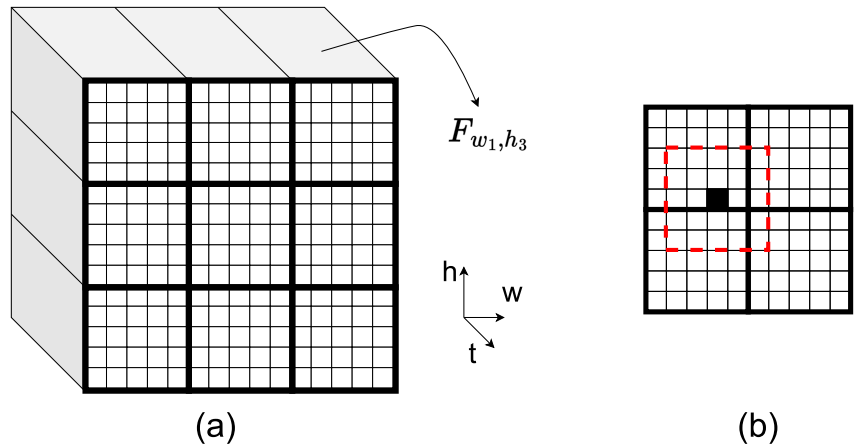

Figure 2. (a) Illustration of the general method for how grid cells are grouped together in order to estimate quantiles. In this example, the spatial domain is split into squares of 5 × 5 grid cells, giving 9 separate large square regions, and in each region the quantiles are calculated. (b) To calculate a particular quantile for a grid point, filled in black, we perform a weighted sum of the value of this quantile calculated in the square regions nearest to that point. The weighting for each large region is proportional to the number of small squares inside the red dashed square. In this example, the red dashed square covers 25 grid cells and the weightings would be $\frac{12}{25}$, $\frac{3}{25}$, $\frac{8}{25}$, and $\frac{2}{25}$.

Note that these percentiles are calculated at the grid box level, so that the number of samples available to calculate these quantiles is 4000 × 270 × 265 grid boxes for the validation and test data sets.

For data greater than the maximum value observed in training, we follow the additive uplift method (Boé et al., 2007; Déqué, 2007) and add the uplift of the highest quantile for the IMERG and IFS data. For example, if $o_{\max}, f_{\max}$ are the highest values seen in the training set for the observations and forecast respectively, then for any forecast value in the test data greater than $f_{\max}$ we add the uplift $(o_{\max} - f_{\max})$ to it.

From experiments we found that the typical approach of quantile mapping the GAN at each grid point individually was not a robust approach for the highest values. Therefore we aggregated the data into square regions to calculate quantiles (Figure 2a). The intuition is that nearby points will have similar distributions and so we can gain accuracy by grouping nearby points together.

To avoid any artifacts due to the edges of these domains, the quantiles for a given grid cell were calculated as a weighted average of the nearest square regions; specifically, the quantiles used to update the values at grid cell (m, n) are calculated as a weighted sum, where the weighting is calculated by drawing a square around (m, n) and counting the number of grid points that fall into each quantile grouping (Figure 2b). This is partly motivated by the ease of implementation, as this can be easily done by broadcasting the grouped quantiles to the same dimensions as the original grid, and using square convolutions with reflective padding to calculate the weighted versions of each quantiles. The length scale of the weighting window was chosen to be the same as the length scale of the quantile groupings, as this was empirically observed to produce reasonably smoothed values.

To decide on the optimal grouping in the spatial domain, the quantile mapping approach described above was trained on the same training data as the cGAN, and then used to perform quantile mapping on forecasts in the validation set. The best parameters were chosen by calculating the quantiles over the whole domain after quantile mapping, and comparing these to the quantiles of the IMERG data over the whole domain using mean square error (MSE) up to the 99.9999^{th} percentile. Using this method, the cGAN performed best when split into 4 square regions (each region having width 66 grid boxes), whilst the IFS forecast performed best when split into 9 regions (each region having width 30 grid boxes). We denote these quantile mapped models as cGAN-qm and IFS-qm.

2.5. Assessing Sample Variability

For many diagnostics, particularly those concerned with high rainfall events, the results can be swayed by the presence or absence of a small number of high rainfall events particular to the test year. To estimate the uncertainty due to these effects, we use bootstrapping along the time dimension (Efron & Tibshirani, 1986). To perform

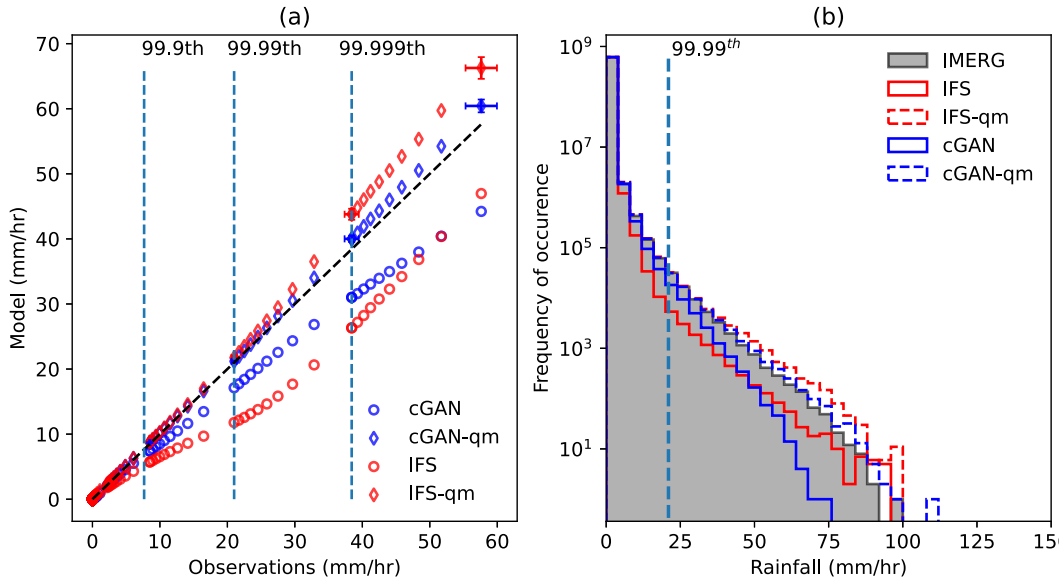

Figure 3. (a) Quantile-quantile plot, up to the 99.9999th percentile. Red circles (diamonds) indicate quantiles for the Integrated Forecast System (IFS-qm) model. Blue circles (diamonds) indicate quantiles for the conditional generative adversarial network (cGAN-qm) model. The black dashed line is the line along which a perfectly calibrated forecast would sit. The error bars indicate an estimate of 2 standard errors from 1,000 bootstrap samples (only shown for the 99.999th and 99.9999th percentiles). (b) A histogram showing the distribution of rainfall values; the vertical dashed blue line indicates the 99.99th percentile of observed rainfall.

bootstrapping for a property of interest θ calculated over a set of N hourly samples, we sample with replacement N times from the samples, and repeat this process M times, resulting in M sets of samples of size N. Then the mean and standard error of θ can then be estimated from the mean and standard deviation of θ calculated on the bootstrap samples. Since the hours are sampled uniformly at random, this method does not take into account the correlation between adjacent hours, and so it is likely that the standard error calculated from this method is an underestimate.

3. Evaluation

In this section we present results of evaluating the model on unseen data. Evaluations are performed on the primary test data set except for Section 3.4 which is evaluated on the extreme Long Rains of March–May 2018.

3.1. Climatological Properties of the Forecasts

We first assess how well the forecasts capture the distribution of rainfall, shown by a quantile-quantile plot and a histogram of rainfall distribution for 4,000 samples, in Figures 3a and 3b, respectively. The unpostprocessed model outputs are shown together with quantile-mapped outputs. From these we can see that, without post-processing, the distribution of cGAN output is an improvement upon that of the IFS output up to extremely high levels of rainfall (around 50 mm/hr) beyond which point the IFS is closer to the distribution.

After both forecasts have been quantile mapped, they are much closer to the ideal line, with deviations at high quantiles. The scale of sampling variability due to variability of samples within the test year was quantified by performing 1,000 iterations of bootstrapping (see Section 2.5) to estimate the standard error of the quantiles. These are shown in Figure 3a, where each error bar shows 2 standard errors. The extreme values of the quantile-mapped forecasts are slightly larger than those in the observations, which we attribute to the sampling variability between the training and test periods.

In order also to get a sense of how extreme these quantile values are, we plot the approximate return period in days for a range of thresholds in Figure 4, calculated over all hours in the test period. These are calculated as the average time interval between instances where at least one point in the domain exceeds the threshold. Note that the very high rainfall values (above around the 99.99th percentile, ~20 mm/hr) are mainly concentrated over Lake Victoria and parts of the sea.

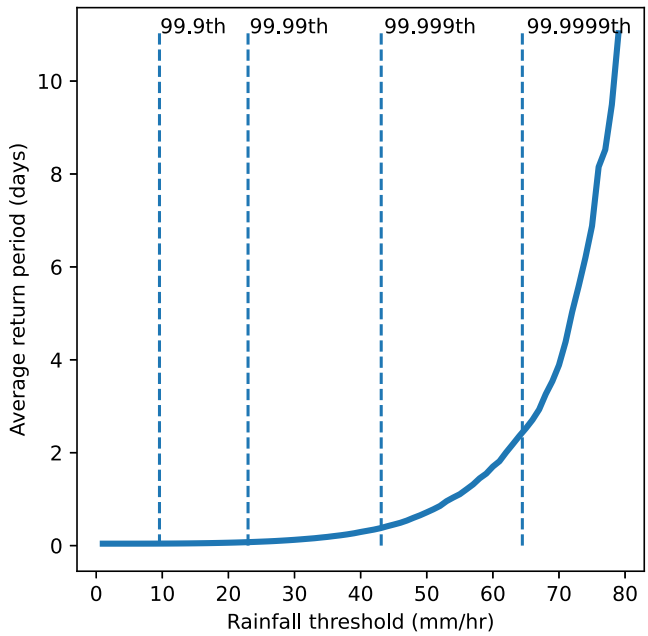

Figure 4. The approximate return periods for different rainfall thresholds to be exceeded at any grid point in the spatial domain in the training set. The dashed lines indicate the values of particular high percentiles.

We observed that the quantile-mapped models cGAN-qm and IFS-qm performed better than the unmodified cGAN and IFS forecasts across the range of diagnostics considered here, and so for the remaining diagnostics in this work we will focus on these quantile-mapped forecasts.

We plot examples of samples generated by cGAN-qm in Figure 5. The generated samples appear to have a realistic spatial structure, and the ensemble mean has substantial rainfall in places where rainfall is observed, indicating that at least some members include rainfall in those locations. In the top row example (5 p.m. on 19 February 2021), it appears to have increased the area of intense rainfall in the South East to better match that of IMERG, whilst in the second row example (9 p.m. on 21 October 2021), the cGAN-qm appears to have improved the number of showers occurring over land. In the bottom two examples (5 a.m. on 17 December 2020 and 8 a.m. on 8 May 2021) we can see cGAN-qm removing some of the excess area of low rainfall seen over the land and sea for the IFS-qm model.

To provide a quantitative picture of the realism of spatial structure in these rainfall patterns across all length scales, we plot the RAPSD (described in B1 in Appendix B) in Figure 6. The deviations from the observed RAPSD are very similar for both the IFS-qm and cGAN-qm models, with the cGAN-qm perhaps slightly closer to IMERG observations for the highest and lowest wavenumbers.

Biases in the diurnal cycle are harder to correct with conventional post-processing methods such as quantile mapping, which would require either doing the postprocessing separately for each time of day or creating a more complex statistical approach. In contrast, the cGAN has the ability to autonomously learn this correction without any separation of the data, and so here we evaluate the cGAN's performance on this task. Summary statistics of the rainfall as a function of hour across the whole spatial domain are plotted in Figure 7, from which it is clear that cGAN-qm is making significant improvements to the diurnal cycle (improvements that are also present in cGAN, not shown). The IFS-qm forecast clearly has a peak in mean rainfall that is too large and occurs too early, at around midday (Figure 7a). This bias is also prominent for the diurnal cycle of high quantiles (Figures 7b and 7c). The cGAN-qm diurnal cycle remains much closer to that of observations for the high quantiles, although there is a tendency to overpredict the dip between 6 a.m. and noon.

These results are further broken down by subdomain in Figure 8, showing that cGAN-qm gives substantial improvements over land points, ocean points, and the Lake Victoria region. However, for Lake Victoria lake points only (Figure 8c), the improvement is less pronounced.

To better quantify to what extent cGAN-qm improves the timing of the diurnal rainfall peak as well as its magnitude, Figure 9 shows the peak rainfall hour for observations, and biases in peak rainfall hour for cGAN-qm and IFS-qm. Summary results averaged over different areas are given in Table 1. Peak rainfall hour is calculated by first finding the average rainfall by hour for each grid square individually, creating a diurnal cycle for each grid square. These diurnal cycles are then smoothed along the time axis using a centered moving average of length 3 hr, before identifying the time at which this smoothed curve is maximized. The results are spatially smoothed using a uniform filter of size 3×3 grid boxes before plotting. The peak rainfall hour for cGAN-qm is a little more spatially noisy than that of IMERG. From Table 1, we can see that cGAN-qm has an overall positive bias in the peak rainfall hour, tending to predict rain about 1 hr too late, whilst IFS-qm has an overall negative bias, predicting rain too early by about the same amount, consistent with other studies (Cafaro et al., 2021). The biggest improvements in the timing of the peak made by the cGAN-qm are over the ocean and over the Lake Victoria region. Both models perform similarly over the Lake itself.

Figure 10a shows the mean biases of the models over the domain. When averaged over the whole domain, cGAN-qm shows a slightly negative bias on the test set whilst IFS-qm shows a slight positive bias. The root mean square bias averaged over the whole domain is 0.04 mm/hr for cGAN-qm and 0.07 mm/hr for IFS-qm. Figure 10b shows the biases in the standard deviation, which have root mean square 0.27 mm/hr for cGAN-qm and 0.30 for IFS-qm.

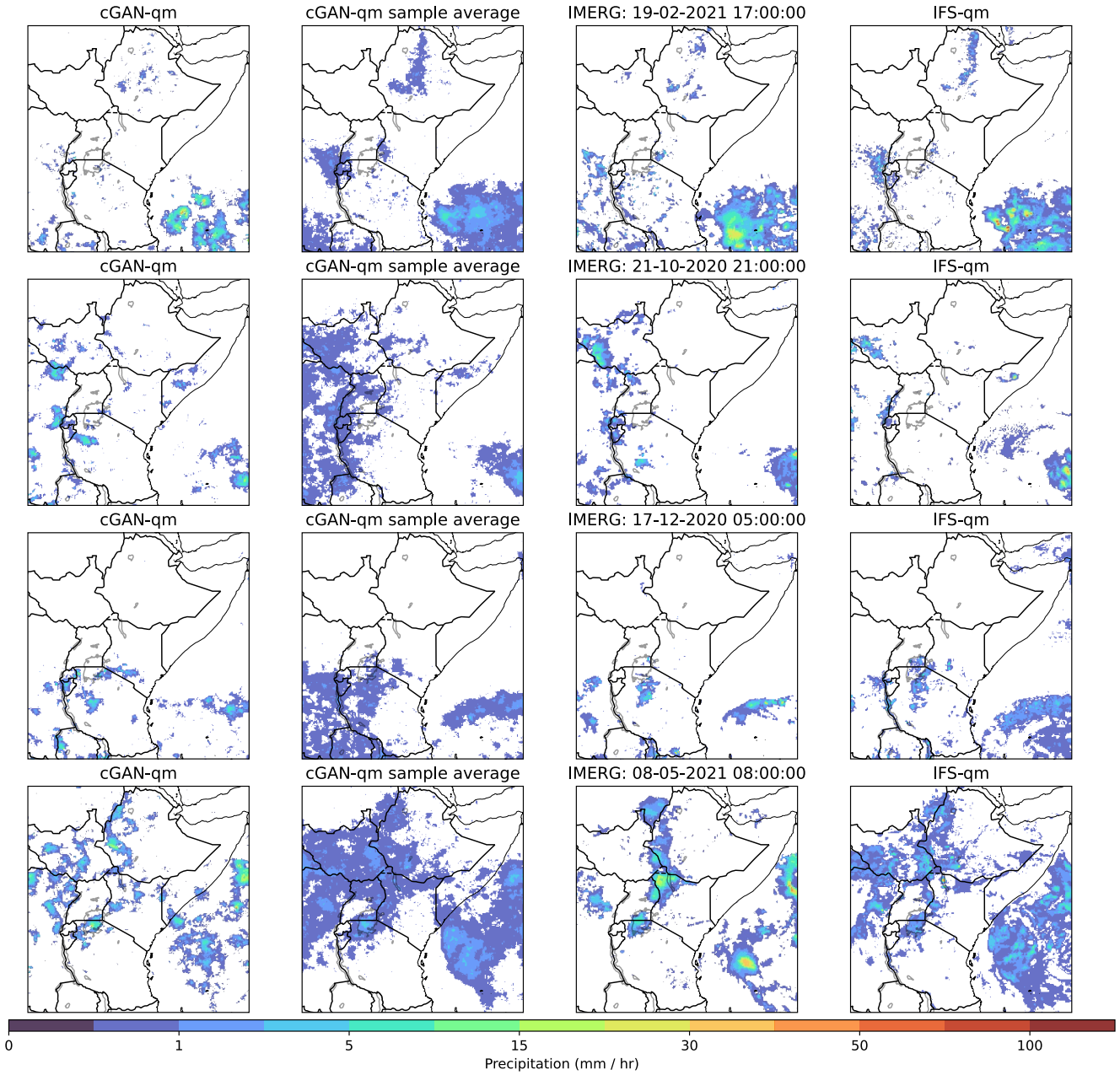

Figure 5. Example precipitation forecasts following postprocessing by the cGAN-qm model, for a selection of hours throughout the primary test data set. The examples are from randomly chosen dates, but filtered to ensure that a diverse range of months in the year are represented, and so that the examples show different behaviors for periods with high and low rainfall. The columns from left to right show: a single member of the cGAN-qm ensemble, the average of 20 cGAN-qm ensemble members, the Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG) observations, the IFS-qm forecast. Each row corresponds to one time value, shown in the title of the IMERG sample.

These biases reflect differences in the optimum quantile mapping function between the training and test data, and indicate how large biases may be expected to be in unseen years.

3.2. Forecast Skill Assessment

In this subsection we investigate the forecasting skill of both models.

The scatter plot in Figure 11 shows how well each model predicts the domain-averaged rainfall. The cGAN-qm model is more tightly clustered around the ideal diagonal line, with IFS-qm more spread out and having more

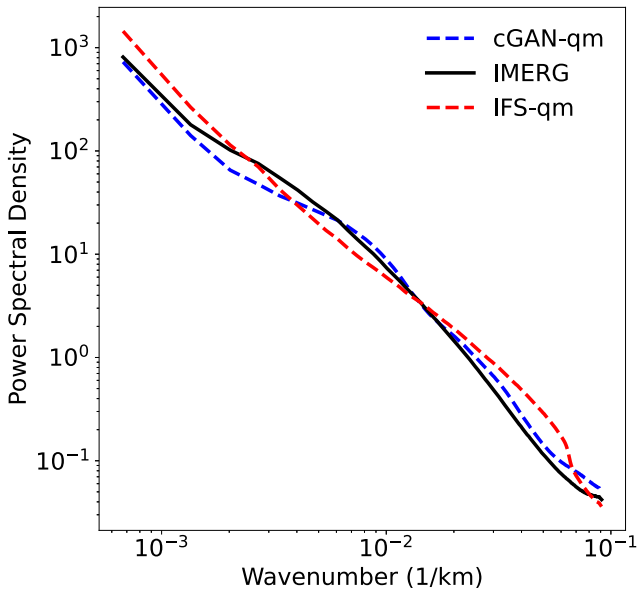

Figure 6. Radially averaged power spectral density for the quantile mapped forecasts. The black line is for the Integrated Multi-satellite Retrievals for Global Precipitation Measurement data, the blue line for cGAN-qm, and the red line for IFS-qm. See B1 in Appendix B for more details on this diagnostic.

points for which the forecast is substantially larger than observations. This is reflected in the Pearson correlation coefficients; 0.75 for the cGAN-qm and 0.60 for the IFS-qm. This suggests that cGAN-qm has improved the accuracy of predicting domain-average rainfall on at least some days. cGAN-qm appears to reduce the frequency of large overpredictions of rainfall compared to IFS-qm (though in some cases it may be preferable for a forecast to produce more false alarms than false negatives).

We now look at skill at predicting rainfall events at smaller spatial scales, focusing on occurrences of rainfall above a certain threshold. We start with the Fractions Skill Score (FSS; see B3 in Appendix B). In Figure 12 we show plots of FSS for different quantile thresholds, where the quantiles are calculated over the whole domain rather than for each grid cell individually. Domain sizes indicate the full width of the neighborhood containing each grid box, which can extend to a maximum of twice the domain width (with zero padding for grid points outside the domain), following the approach of previous studies.

First we note that, without quantile mapping, cGAN shows substantial improvements relative to the unpostprocessed IFS up to the 99.99th percentile. This improvement is reduced when comparing the quantile-mapped results. For thresholds up to around the 99.9th percentile, cGAN-qm has consistently higher FSS than IFS-qm for most neighborhood sizes. However, we can see that for the 99th and 99.9th percentiles IFS-qm achieves higher scores at low neighborhood sizes, which suggests higher skill at fine resolutions. At higher percentiles, which are dominated by events over Lake Victoria and the ocean, IFS-qm has higher FSS on average, although the effects of sampling variability are substantial (especially given that these error bars are likely an underestimate).

The asymptotic value that the FSS approaches as the neighborhood width becomes large indicates the frequency bias of the particular rainfall event (see B3 in Appendix B), and so we can see that, up to thresholds around the 99.9th percentile, cGAN-qm has lower overall frequency bias.

Overall then, these results suggest that, for up to around the 99.9th percentile cGAN-qm has higher FSS. Beyond this, the IFS-qm shows better performance on average, although neither model shows high skill at the highest percentiles.

The Equitable Threat Score (ETS) for several thresholds over the whole domain is shown in Figure 13a, where rainfall is compared at individual grid points. This shows that IFS-qm tends to perform better by this metric,

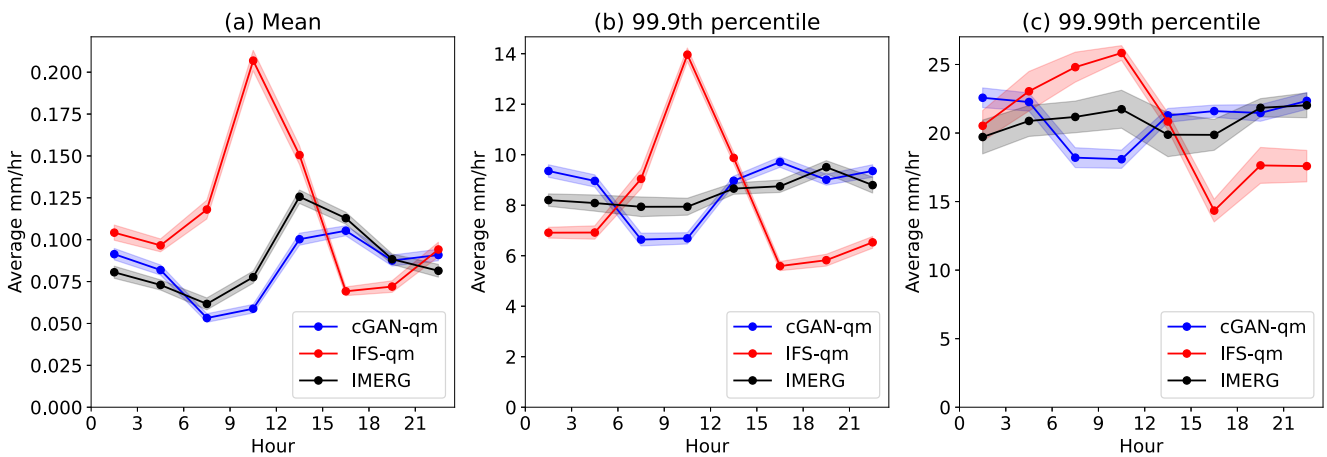

Figure 7. Evaluation of the diurnal cycle by hour (in local time) over the whole domain in; (a) mean rainfall, (b) 99.9th percentile, and (c) 99.99th percentile, where percentiles are calculated over all individual grid points. The shaded regions indicate ±2 standard errors about the mean, estimated by bootstrapping with 50 samples. Note that rainfall at the 99.99th percentile and above is mainly concentrated over Lake Victoria and the ocean.

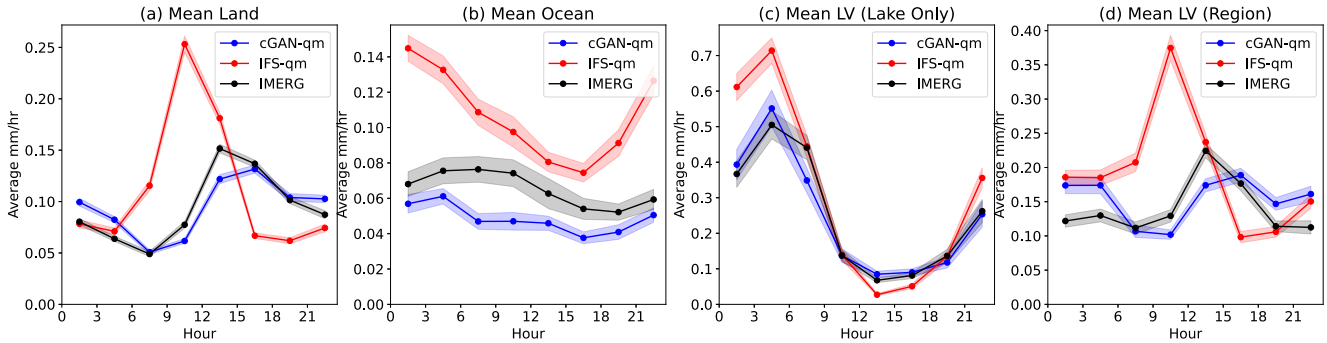

Figure 8. Evaluation of the diurnal cycle of mean rainfall over different subdomains (a) Land points only, (b) Ocean points only, (c) Lake Victoria (lake points only), and (d) Lake Victoria region as defined in Figure 1. The shaded regions indicate ±2 standard errors about the mean, estimated by bootstrapping with 50 samples.

which is in agreement with IFS-qm having higher FSS at smaller length scales. In Figure 13b the HR and False Alarm Rate (FAR) are shown, from which we can see that the difference in ETS score is driven by IFS-qm having both a slightly lower FAR and a slightly higher HR.

3.3. Assessment of Ensemble Calibration

We now focus on the probabilistic calibration of the cGAN-qm ensemble, in order to evaluate how suitable cGAN-qm samples are for use as an ensemble. Note that the IFS forecast data we have is not an ensemble so cannot be used in comparison. These results are assessed on 500 samples drawn randomly from the test year, each with 100 ensemble members.

Conceptually, the probabilistic skill depends on variability of both the deterministic and stochastic components of the forecasts. The deterministic component of cGAN-qm is identified with the population ensemble mean and the stochastic component with the distribution of individual members about the mean.

In Figure 14 we show the spread error diagnostic (see B4 in Appendix B), where the forecasts are binned into 100 different intervals of the forecast "spread," which equals the standard deviation of the forecasts. This specifically tests the calibration of the stochastic component of the predictions alone, and it is of interest how well the cGAN can learn to represent this. From the figure we can see that the spreads of forecasts from the cGAN-qm ensemble cover a substantial range and are closely correlated with the RMSE, showing that cGAN-qm skilfully distinguishes between situations with higher and lower predictability. The ensemble tends to be under-dispersed (i.e., over-confident) in relatively predictable situations (low RMSE and ensemble spread), and over-dispersed (i.e., under-confident) in relatively less predictable situations (high RMSE and ensemble spread). Whilst the ensemble

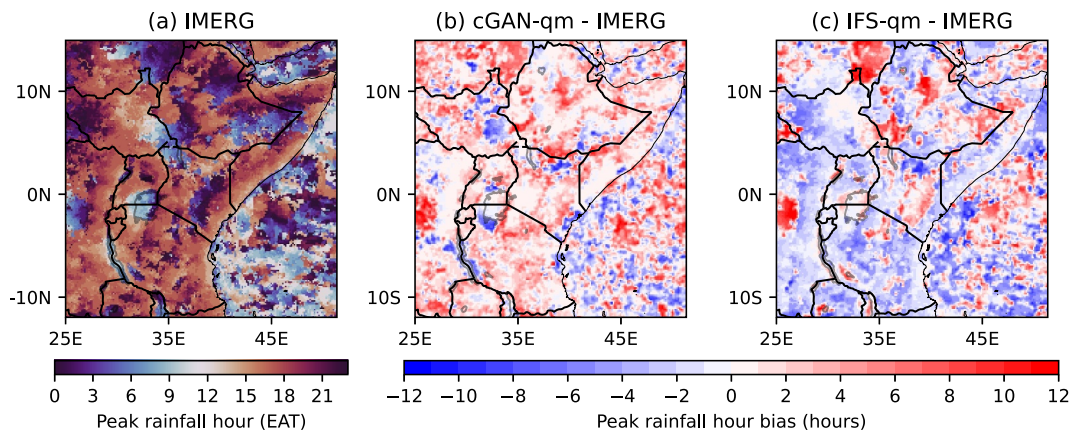

Figure 9. (a) Map of peak rainfall hour for Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG). Panels (b, c): the differences in peak rainfall hour with respect to IMERG in cGAN-qm (middle) and IMERG (right). A 3-hr moving average along the time dimension is applied to the data in each case before calculating the peak hour. Spatial smoothing using a uniform filter of size 3 × 3 grid boxes is also applied to panels (b, c).

Table 1
Bias (in Hours) of the Peak Rainfall Hour, Calculated at Each Grid Point and Averaged Over Different Subdomains of the Domain

	Whole domain	Land	Ocean	Lake Victoria (lake only)	Lake Victoria (region)
cGAN-qm	**0.9**	1.3	**0.1**	−0.1	**1.3**
IFS-qm	−1.1	**−1.2**	−1.2	−0.1	−1.5

Note. Bold values indicate the best performing model for each subdomain.

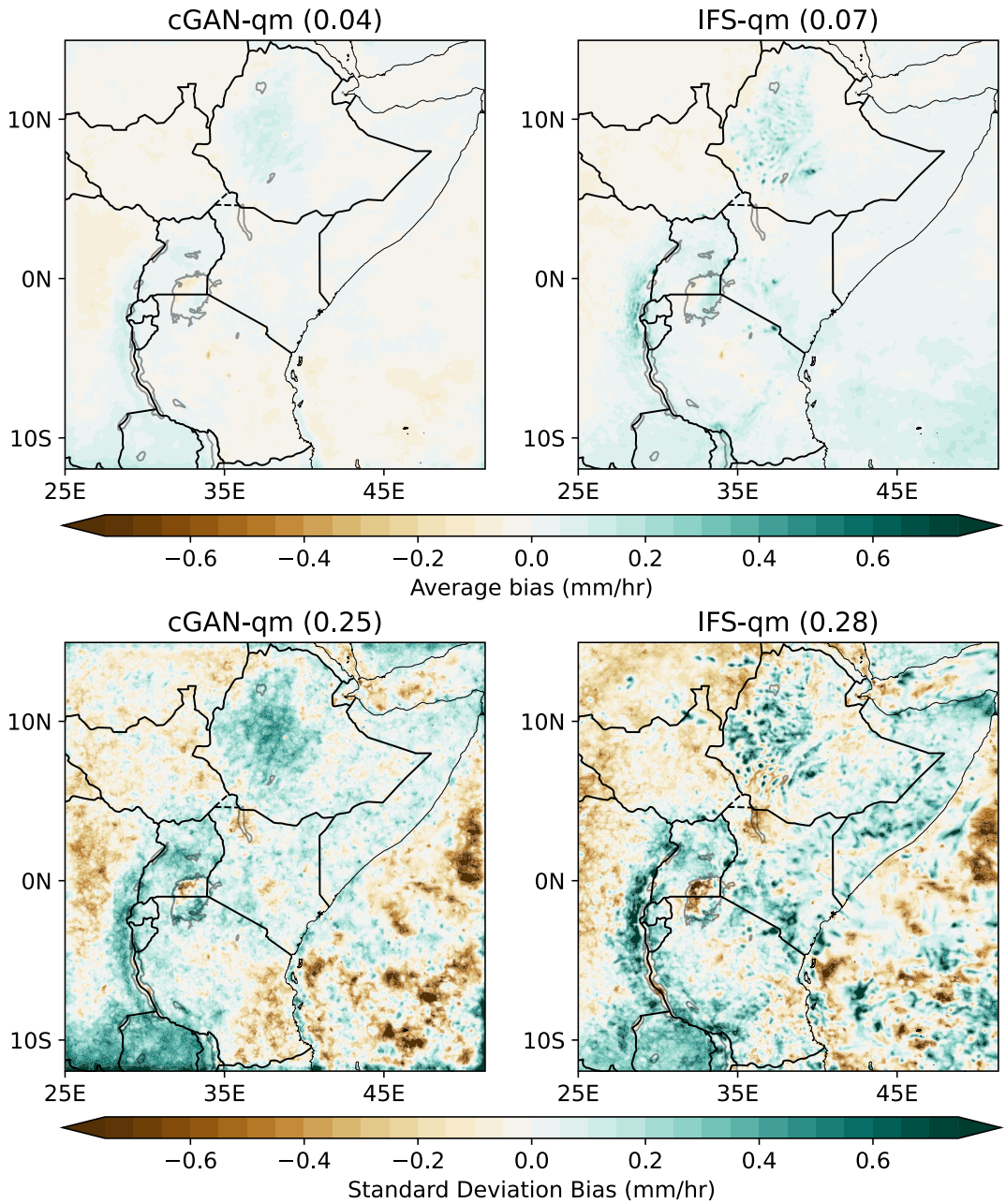

Figure 10. Mean bias (top) and bias in standard deviation (bottom) of the cGAN-qm and IFS-qm forecasts. The number in brackets in each title is the root mean square bias over the whole domain.

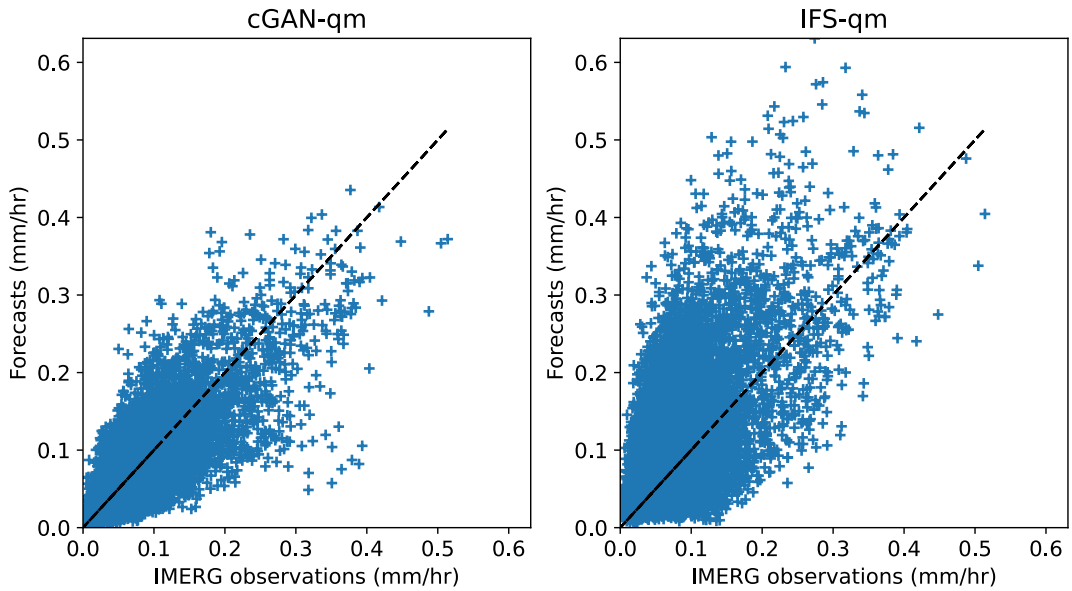

Figure 11. Scatter plots of predicted versus observed domain-average rainfall, for the cGAN-qm forecasts (left) and IFS-qm forecasts (right). The dashed black lines represents perfect forecasts. The Pearson correlation coefficient is 0.75 for cGAN-qm and 0.60 for IFS-qm.

is not calibrated perfectly, it is also reassuring to see that cGAN-qm has not fallen into the common failure mode of predominantly producing predictions close to the most likely result (Arjovsky & Bottou, 2016), which would produce forecasts with no ensemble spread.

Figure 15a shows a rank histogram of forecasts for rainfall at individual grid points. This depends on both the deterministic (ensemble mean) and stochastic components of the predictions. The ensemble deviates from being perfectly calibrated, and appears to be made up of a mixture of behaviors. In general it is hard to uniquely attribute forecast behavior from a rank histogram (Hamill, 2001); a U-shaped distribution can be indicative of under-dispersion of the ensemble, but can also be a mixture of over- and under-forecasting biases.

In order to shed light on the ensemble behavior at different rainfall intensities, we also calculate rank histograms conditioned on whether the cGAN-qm ensemble mean indicates relatively wet or dry weather. We condition on the mean because its sampling variability is lower, and so using this as the conditioning variable greatly reduces the selection bias that would occur if we conditioned on the observed rainfall. In Figures 15b and 15c we show rank histograms conditioned on the cGAN-qm ensemble mean being >0.1 mm/hr and ≤ 0.1 mm/hr, respectively. In Figure 15b we can see a clear U-shape for the wet hours, indicating under-dispersion, and the histogram is skewed to the left indicating a tendency for predictions to be higher than the observations too often. From Figure 15c we can see a slight peak at low ranks plus sharper peaks at both ends, and the histogram has a negative gradient. This suggests that at low rainfall intensity there is slight over-prediction and a mixture of under- and over-dispersive behavior.

3.4. Evaluation on Extreme Test Set

In this section we evaluate the performance of the cGAN-qm model on the March–May 2018 extreme test set in the Kenya region, which saw the largest rainfall anomalies (see Section 2.2). This is to better understand how it performs at forecasting weather during rainy periods that are more extreme than any seen in training, which is of high importance for forecast applications (Watson, 2022). The total seasonal rainfall observed in this season over the whole domain was higher than any other value in our IMERG data set (between 2001 and 2020), being 25% higher than the mean seasonal rainfall, and 9% higher than the wettest March–May season seen in training. Over the Kenya subdomain, these figures rise to 77% and 34%, respectively. Note that although the land-based rainfall for this period showed marked increases in rainfall frequencies up to extremely large rainfall values, the highest hourly values do not exceed those seen during training, so this does not test the cGAN-qm's ability to generalize to greater hourly extremes. However, the fact that the season has particularly high seasonal total rainfall indicates

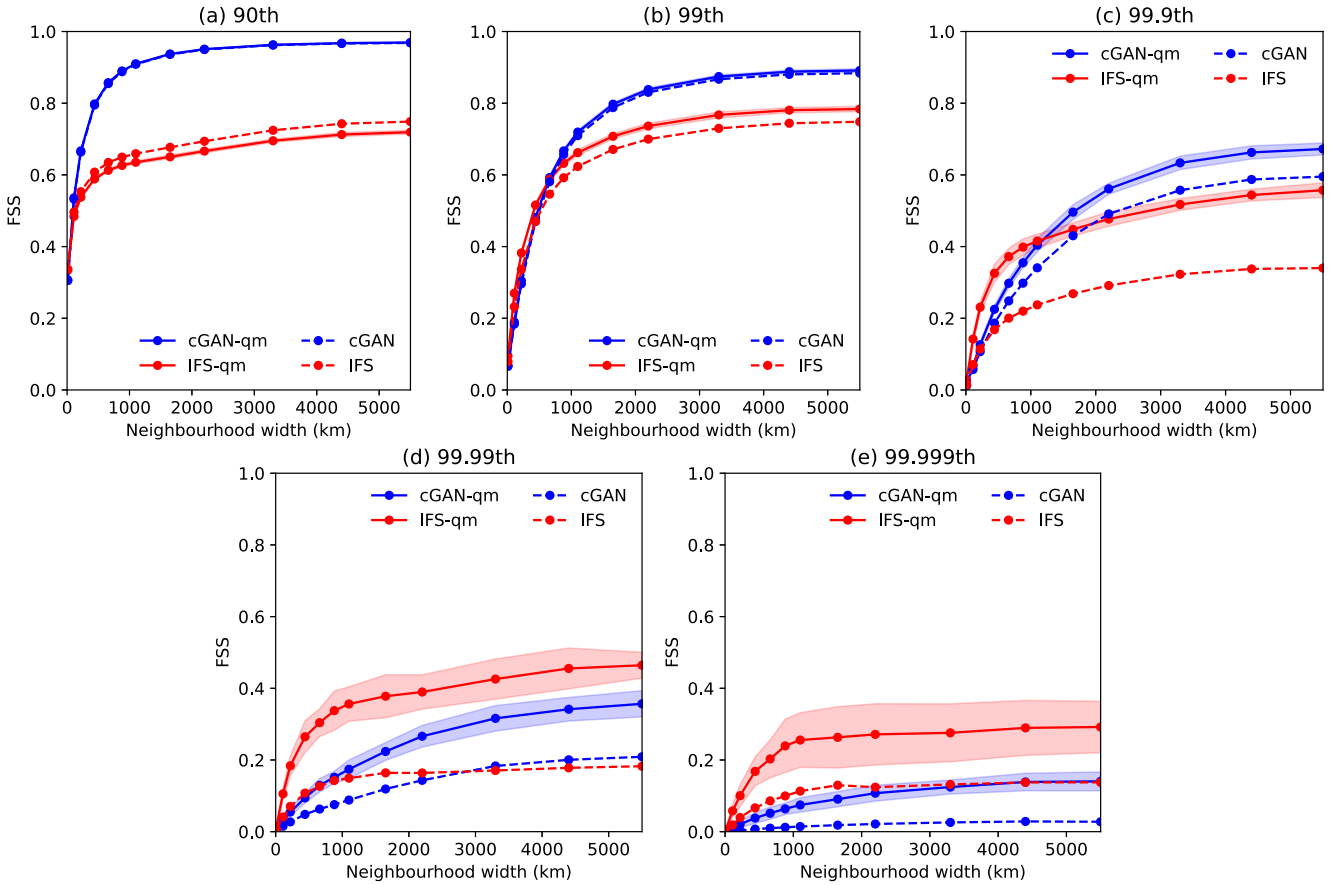

Figure 12. Fractions Skill Score results for different quantile thresholds (a) 90th percentile, (b) 99th percentile, (c) 99.9th percentile, (d) 99.99th percentile, and (e) 99.999th percentile, shown for both unpostprocessed forecasts (Integrated Forecast System and conditional generative adversarial network) and forecasts postprocessed with quantile mapping (IFS-qm and cGAN-qm). Shaded bands indicate ±2 standard errors estimated from bootstrapping with 50 samples. Note that rainfall at the 99.99th percentile and above is mainly concentrated over Lake Victoria and the ocean.

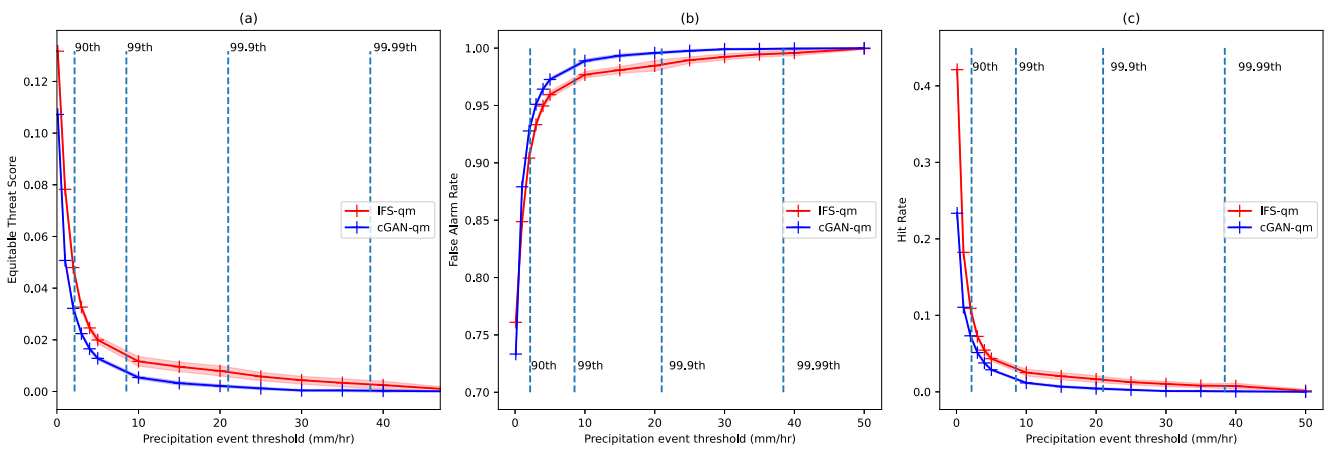

Figure 13. (a) Equitable Threat Score (ETS), (b) False Alarm Rate (FAR), and (c) hit rate (HR) for IFS-qm (red) and cGAN-qm (blue), for forecasts at individual grid points, plotted against rainfall threshold. Here an event is defined as occurrence of rainfall exceeding the threshold. For ETS and HR, higher is better, and for FAR lower is better. Shaded bands indicate ±2 standard errors estimated from bootstrapping with 50 samples. Note that rainfall at the 99.9th percentile and above is mainly concentrated over Lake Victoria and the ocean.

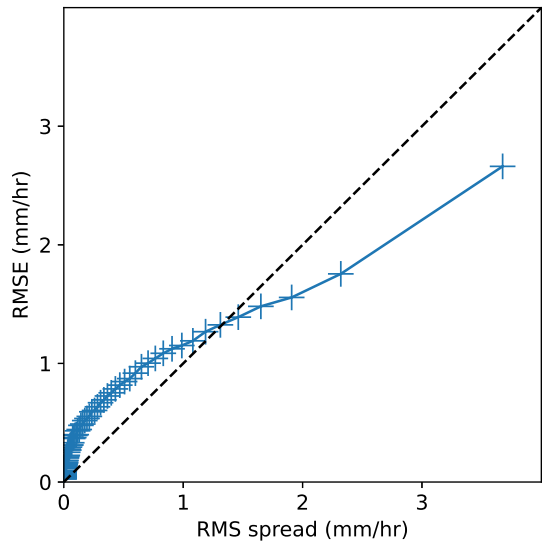

Figure 14. Spread error plot for cGAN-qm, where hourly ensemble forecasts are divided into 100 bins with different ensemble spreads. Calculated using 500 forecasts from cGAN-qm with 100 ensemble members and evaluated at individual grid points.

that it contained an unusually severe set of weather patterns, so is an indicator of how well the GAN can perform in seasons that show more extreme behavior overall than those seen in training.

The quantile-quantile plot in Figure 16 shows that the unpostprocessed cGAN hourly forecast values are reasonably close to the ideal line, the cGAN having corrected the substantial underestimation of the highest values by the IFS. Both quantile-mapped forecasts capture the magnitudes of the high quantiles well, with differences from IMERG being not much larger than sampling variability. After quantile mapping, cGAN-qm has slightly higher intensities of extreme values, and IFS-qm slightly lower.

The scatter plots of Kenya-mean hourly rainfall in Figure 17 show that, whilst the results show larger deviations from the diagonal than those for the main test data set in Figure 11, the cGAN-qm results are more tightly clustered around the ideal diagonal line than IFS-qm. cGAN-qm does however seem to predict fewer values of Kenya-mean rainfall above 1 mm/hr than are seen in IMERG; whilst 24 points in Figure 17 have observed Kenya-mean rainfall above 1 mm/hr, only 2 are predicted by cGAN-qm, compared to 32 predicted by IFS-qm. However, most of these values occur in three particular days, so there is likely considerable sampling variability. This potentially indicates some underestimation in the forecasts of the frequency of very intense rainfall values on spatial scales much larger than 0.1°.

To gain a more complete picture of how the cGAN-qm skill varies with spatial scale and precipitation intensity, we show the FSS of forecasts over the Kenya subdomain as a function of these variables in Figure 18. In contrast to the results for the normal evaluation period shown in Figure 12, cGAN-qm achieves higher scores than IFS-qm up to the 99.99th percentile, although both models have low scores beyond the 99.9th percentile. Other diagnostics (such as diurnal cycle) also followed a similar pattern to that seen in the primary test data set (not shown). Overall, these results indicate that cGAN-qm has extrapolated reasonably well to a rainfall regime outside that which it was trained on. However, a full understanding of cGAN-qm's ability to predict extreme rainfall events would require a more in depth analysis.

4. Conclusions

In this work we have investigated the use of a ML-based method, a cGAN, to postprocess the ECMWF IFS HRES forecast over equatorial East Africa at 0.1° resolution and to generate an ensemble forecast. A novel quantile mapping technique was applied to the IFS forecast (labeled "IFS-qm") to provide a strong baseline, and also to the cGAN output (labeled "cGAN-qm") in order to combine strengths from both conventional postprocessing methods and ML. This extends the application of GAN-based postprocessing of NWP precipitation forecasts to a

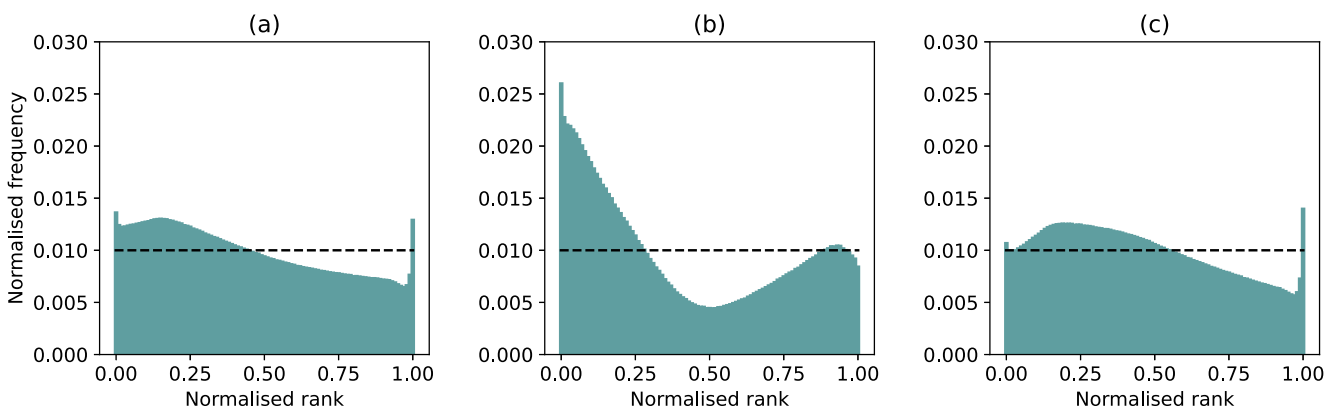

Figure 15. Rank histograms for (a) all data, (b) grid points where the cGAN-qm ensemble mean is >0.1mm/hr, and (c) grid points where the cGAN-qm ensemble mean is ≤0.1mm/hr. The dashed black lines represent the result for a perfectly calibrated ensemble.

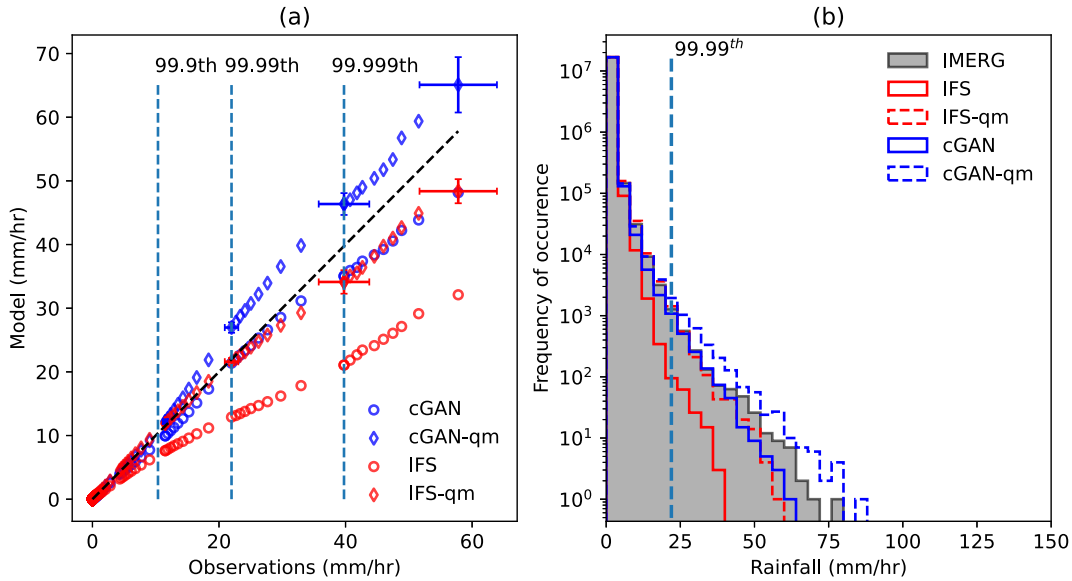

Figure 16. Quantile-quantile plot, up to the 99.9999th percentile, for the Long Rains (March–May) 2018 over Kenya. Results are plotted as in Figure 3a.

tropical domain, which presents distinct forecasting challenges to the extratropical regions where it has previously been tested, informing the use of new systems being trialled in the region (WFP et al., 2024).

The most substantial improvement produced with cGAN-qm was in the diurnal cycle (Figure 7), which is known to be particularly problematic for conventional NWP models to capture. The cGAN-qm model demonstrated a substantial correction to the timing of peak mean rainfall over the whole domain, which persisted when looking at the diurnal cycle of high quantiles. This improvement is a result of the cGAN-qm being trained directly on observations, such that it can learn a skillful statistical relationship between the forecast inputs and the observed rainfall intensity. However, there is substantial spatial noise in the cGAN-qm peak rainfall hour (Figure 9); using time as an input variable may be one way to improve the learned relationships. The frequency distributions of

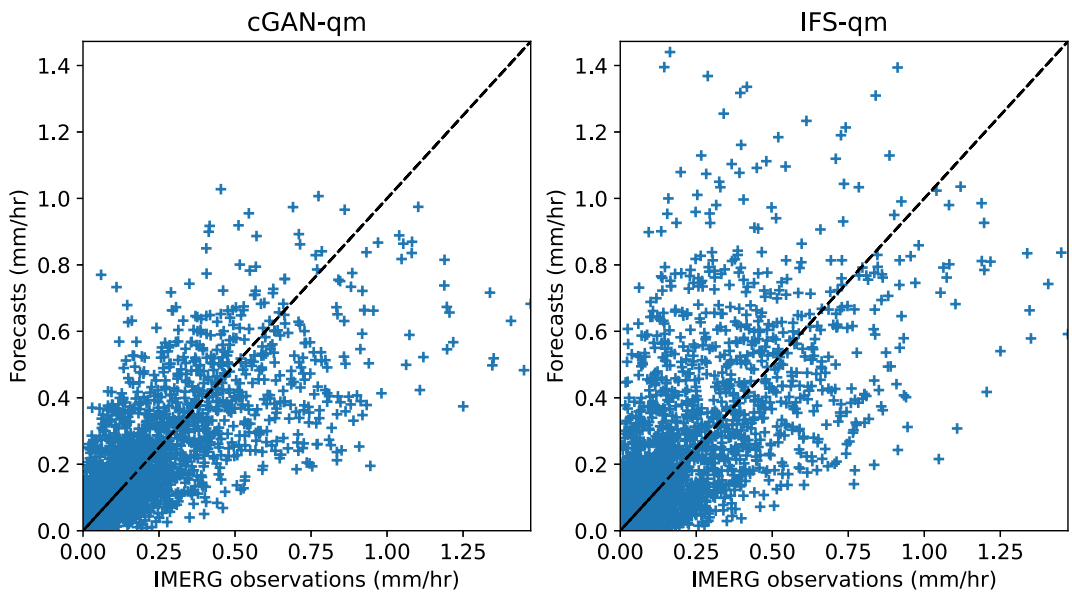

Figure 17. Scatter plots of predicted versus observed domain-average rainfall, for the cGAN-qm forecasts (left) and IFS-qm forecasts (right) during the 2018 Long Rains over Kenya. The dashed black lines represents perfect forecasts. The Pearson correlation coefficients are 0.72 for cGAN-qm and 0.59 for IFS-qm.

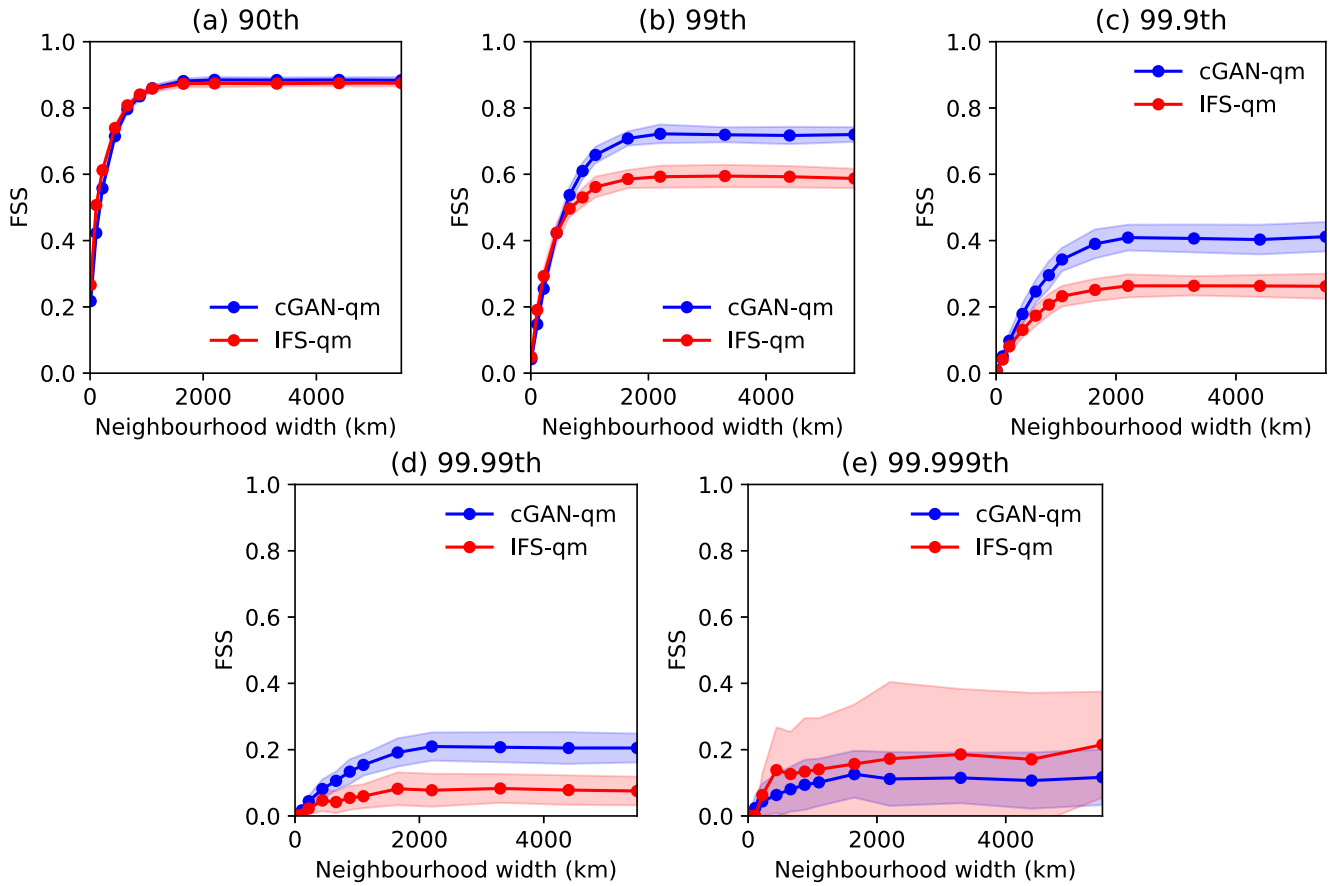

Figure 18. Fractions Skill Score in the extreme March–May 2018 season, over the Kenya subdomain, for different quantile thresholds, with quantiles calculated for this season and subdomain (a) 90th percentile, (b) 99th percentile, (c) 99.9th percentile, (d) 99.99th percentile, and (e) 99.999th percentile. Shading indicates ±2 standard errors estimated from bootstrapping with 50 samples.

rainfall for cGAN-qm and IFS-qm (Figure 3) were both comparable, with small biases which we attribute to sampling variability.

The cGAN-qm improved forecast skill scores in some respects. The cGAN-qm shows generally higher Fractions Skill Scores (Figure 12) up to a high percentile $\left(99.9^{th}\right)$, particularly at larger neighborhood sizes (above ∼500 km). For higher percentiles the IFS-qm forecast demonstrated a higher score; since the rainfall for these high percentiles is predominantly over Lake Victoria and the sea, this may largely be a reflection of the strength of the conventional forecasts over these regions. The IFS-qm model also achieved higher ETS at the grid scale at all thresholds, although the scores were nevertheless quite low (Figure 13).

Both models were also evaluated on the 2018 Long Rains (March–May), which were significantly wetter than any Long Rains season seen in training and across the whole IMERG data set. It may be expected that ML-based methods would show degraded performance on situations with characteristics outside their training data, and this is highly important to evaluate for forecasting applications (Watson, 2022). In fact, we found that unmodified cGAN forecasts actually had a more realistic frequency distribution of grid-scale rainfall than that of the unpostprocessed IFS (Figure 16), and cGAN-qm forecasts had higher skill than IFS-qm when evaluated using the FSS, albeit with a potential underestimation of the frequency of the very highest Kenya-mean rainfall amounts (Figures 17 and 18). We stress here that these only provide an indication of how the cGAN-qm may operate outside the bounds of the training data, and a more thorough evaluation would be required to have high confidence that the method will generally perform well in extreme situations.

Through understanding the regimes at which cGAN-qm can improve forecast skill, improvements in short-term rainfall forecasts can be made which we anticipate can improve applications such as flood prediction across the

region; this is further supported by the increase in skill shown by the cGAN-qm during the 2018 Long Rains over Kenya. An interesting extension of this evaluation would be to pass the output of cGAN-qm through an impacts model such as a flood inundation model to see whether it gives improved forecasts of impacts.

An important advantage that generative ML models provide over other non-generative models is the ability to create an ensemble of predictions from a single forecast. It is therefore an interesting question as to whether this ML model can provide well-calibrated probability distributions. Our assessment indicates that the spread of the model correlates well with the observed error, although it also demonstrates a mixture of under- and over-dispersive behavior (Section 3.3). Note that the stochastic component of the cGAN predictions is not temporally coherent, so that combining predictions from different times would produce a time series with hour-to-hour variability that is likely too large. This could be addressed in future work by postprocessing using models like those applied in conditional video generation (e.g., Xing et al., 2023).

Many studies on the performance of ML models compare the output of a model to the raw IFS forecast, or similar (e.g., Bi et al., 2022; Lam et al., 2023). Whilst the cGAN without quantile mapping improves on the unpostprocessed IFS forecast rainfall distribution (Figure 2), diurnal cycle, and skill scores (Figure 12), by evaluating our model against a strong baseline of quantile mapped IFS forecasts, our comparison reveals what ML can do that is not achieved by conventional postprocessing methods like quantile mapping, which is a sterner test than comparing to unprocessed forecast.

A strength of using ML to postprocess existing forecasts is that we incorporate the skill and physical understanding of physics-derived models (Watson, 2019). However, we implicitly assume that the IFS forecast captures all the useful information about phenomena such as the Indian Ocean Dipole and Madden-Julian oscillations, and our model may be improved by including indexes of these drivers and/or sea surface temperatures as additional inputs. It would be interesting as well to include the initial state of the forecast, if available, to see whether the ML model is able to correct for errors in how the IFS evolves this initial state.

Since the key improvement we observe here is an improvement in the diurnal cycle, which it is well established is improved by convection-permitting (CP) models relative to coarser resolution models (e.g., Cafaro et al., 2021), it would be valuable to compare this approach to the output of a CP model. Since CP models are much more computationally expensive, and the cGAN-qm model is relatively cheap to run, it would be interesting to explore the trade-offs of computational cost and performance between the methods. There are also other state-of-the-art ML models that would be interesting to compare with, to see if performance improvements can be made. One particularly promising approach would be diffusion models, which have recently started to be applied in weather and climate prediction (e.g., Addison et al., 2022; Leinonen et al., 2023), since these models appear to perform well and are easier to train. Another simple alternative to compare with would be a model trained using a suitable probabilistic loss function, such as the energy score (Pacchiardi et al., 2022).

Appendix A: Forecast Variables Used

The IFS variables and constant fields used to train the model are shown in Table A1, definitions taken from ECMWF (2023a, 2023b, 2023c).

The preprocessing methods mentioned in the table are as follows, using the year 2017 as the reference period:

- Minmax: calculate the minimum $d_{\min}$ and maximum $d_{\max}$ over the reference period, and then transform each value v according to $(v - d_{\min})/(d_{\max} - d_{\min})$.
- Max: calculate the and maximum $d_{\max}$ over the reference period, and then transform each value v according to $v/d_{\max}$.
- Log: Transform each value v according to $\log_{10}(1 + v)$.

Table A1
Integrated Forecast System Variables Used to Train the Model, as Well as the Normalization Applied to Each Variable

Variable name	Symbol	Pre-processing applied
2 m temperature	2t	Minmax
Convective available potential energy	cape	Log
Convective inhibition	cin	Max
Convective precipitation	cp	Log
Surface pressure	sp	Minmax
Total column cloud liquid water	tclw	Max
Total column vertically integrated water vapor	tcwv	Log
Top of atmosphere incident solar radiation	tisr	Max
Total precipitation	tp	Log
Relative humidity at 200 hPa	r200	Max
Relative humidity at 700 hPa	r700	Max
Relative humidity at 950 hPa	r950	Max
Temperature at 200 hPa	t200	Minmax
Temperature at 700 hPa	t700	Minmax
Eastward component of wind at 200 hPa	u200	Max
Eastward component of wind at 700 hPa	u700	Max
Northward component of wind at 200 hPa	v200	Max
Northward component of wind at 700 hPa	v700	Max
Vertical velocity at 200 hPa	w200	Max
Vertical velocity at 500 hPa	w500	Max
Vertical velocity at 700 hPa	w700	Max
Orography	h	Max
Land-sea mask	lsm	N/A

Note. See text for description of the different preprocessing types.

Appendix B: Forecast Verification Measures

B1. Radially Averaged Power Spectral Density (RAPSD)

In order to assess the spatial realism of the generated forecasts, we use the Radially Averaged Power Spectral Density (RAPSD) (Sinclair & Pegram, 2005). This is calculated by taking the 2D Fourier transform of the precipitation image, and averaging the power spectrum over the total wavenumber magnitude, yielding a one-dimensional series showing the distribution of weights given to different frequencies. For assessing multiple forecast images we take the mean RAPSD over the images, and for situations where we need to summarize the overall similarity of two RAPSD curves we use the Radially Averaged Log Spectral Distance defined in Harris et al. (2022).

B2. Equitable Threat Score (ETS)

The Equitable Threat Score (ETS) measures the balance between the hit rate and false alarm rate, whilst accounting for the probability of random events (Schaefer, 1990; Wilks, 2019). This score is used in operational forecast verification (Mittermaier et al., 2013) and in Manzato and Jolliffe (2017) was shown to be one of the most robust metrics with respect to random (unskillful) changes in the forecast. It is defined as:

$$\mathrm{ETS} := \frac{\mathrm{TP} - \mathrm{TP}_r}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} - \mathrm{TP}_r} \tag{B1}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives. TP_r accounts for the number of true positives we would expect to achieve by guessing at random, and is estimated from the data using the formula:

$$TP_r = \frac{(TP + FP)(TP + FN)}{N} \tag{B2}$$

B3. Fractions Skill Score (FSS)

Many forecast verification scores, such as mean square error or the ETS, do not always align with human forecasters' subjective evaluations. There have been many different approaches employed to try and remedy this problem (see e.g., Gilleland et al., 2009). One approach, usually called the neighborhood approach, is to smooth the forecasts and observations by averaging the forecast around each grid cell with a particular length scale before applying a forecast metric (Ebert, 2008). A commonly used metric in this class is the Fractions Skill Score (FSS) (N. Roberts, 2008; N. M. Roberts & Lean, 2008).

To calculate the FSS, we first choose a threshold rainfall value r, and for each grid cell of the forecast and observations we calculate the fractions F_{tij} and O_{tij} of neighboring cells for which the rainfall exceeds the threshold, where t, i, and j index the time, latitude, and longitude axes, respectively. The FSS is then defined as:

$$\mathrm{FSS}(n, r) := \frac{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} 2F_{tij}O_{tij}}{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} F_{tij}^2 + O_{tij}^2} \tag{B3}$$

We use the pySTEPS implementation of the FSS (Pulkkinen et al., 2019), which performs the averaging using square convolutions with zero-padding.

In the limit of large neighborhood size, the FSS approaches the asymptotic limit FSS_∞ where (N. M. Roberts & Lean, 2008):

$$\mathrm{FSS}_\infty = \frac{2f_o f_m}{f_o^2 + f_m^2} \tag{B4}$$

where f_o and f_m are the observed and modeled frequency of exceeding the threshold, respectively. Thus the value that the FSS reaches at large neighborhood sizes indicates the level of bias in the average number of grid boxes exceeding the threshold, with $\mathrm{FSS}_\infty = 1$ for no bias.

B4. Spread Error

In order to assess how well calibrated the probabilities of the generated forecast are, we use a spread-error plot, commonly used to assess ensemble calibration (Leutbecher & Palmer, 2008). For an ensemble of forecasts $\{f_{i,t}\}_{i=1}^{M}$ with ensemble mean μ_t, and an observation y_t, the spread s_t and error e_t are defined as:

$$s_t^2 = \frac{1}{M}\sum_{i=1}^{M}\left(f_{i,t} - \mu_t\right)^2 \tag{B5}$$

$$e_t^2 = \left(y_t - \mu_t\right)^2 \tag{B6}$$

Note that for a finite number of ensemble members M, we also include a correction to the spread:

$$\tilde{s}_t = \frac{M+1}{M-1}s_t \tag{B7}$$

To produce a spread-error plot, we first calculate the spread values for each grid cell and each time value. Then we split the paired observations and forecasts into bins (of size 100 in our case) according to this spread value, and for each bin we calculate the root mean squared spread and error values. For a perfect forecast ensemble with infinite

members, the spread of the ensemble will equal the average error between the ensemble mean and observations, so that an ideal spread-error plot is a straight line with gradient 1.

B5. Rank Histogram

Another method to assess the statistical calibration of an ensemble forecast is the use of a rank histogram (also known as a Talagrand diagram) (Wilks, 2019). To construct a rank histogram, for each sample we rank the observed value relative to the ensemble members, and then average this over all the samples. This gives a frequency of how many times the observations were seen to be in each rank, which can be plotted as a histogram. A perfectly calibrated ensemble produces a flat histogram. For an imperfect ensemble, the spread of the ensemble members may be too wide, such that the observations will rank in the middle most of the time, producing a peak in the histogram. For the reverse scenario, the spread is too narrow leading to a U-shaped histogram indicating the ensemble members are too narrowly spread. A histogram sloping to the left or right is also indicative of conditional under-forecasting or over-forecasting, respectively (Hamill, 2001; Wilks, 2019). In this work we use the pySTEPS implementation of the rank histogram (Pulkkinen et al., 2019).

Data Availability Statement

The code for the GAN and quantile mapping used in this paper is available at https://github.com/bobbyantonio/downscaling-cgan; this code was forked from the code in Harris et al. (2022). All experiments in this paper were performed within TensorFlow 2.7.0, and some of the analysis in this work utilized the PySteps package (Pulkkinen et al., 2019). The ECMWF IFS forecasts can be obtained through MARS, for which academic accounts are freely available subject to conditions; see https://www.ecmwf.int/en/forecasts/accessing-forecasts/licences-available. The IMERG satellite precipitation data (Huffman et al., 2022) is freely available after registration, see https://gpm.nasa.gov/data/directory.

Acknowledgments
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant 741112, ITHACA). BA was supported by a NERC Cross-disciplinary Research for Environmental Science award. PW was supported by a NERC Independent Research Fellowship (Grant NE/S014713/1). This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol—http://www.bristol.ac.uk/acrc/. We are grateful to Rachel James and Neil Hart for feedback on an earlier version of this work.

References

Addison, H., Kendon, E., Ravuri, S., Aitchison, L., & Watson, P. A. (2022). Machine learning emulation of a local-scale UK climate model. Retrieved from http://arxiv.org/abs/2211.16116

Ageet, S., Fink, A. H., Maranan, M., Diem, J. E., Hartter, J., Ssali, A. L., & Ayabagabo, P. (2022). Validation of satellite rainfall estimates over Equatorial East Africa. *Journal of Hydrometeorology*, *23*(2), 129–151. https://doi.org/10.1175/JHM-D-21-0145.1

Arjovsky, M., & Bottou, L. (2016). Towards principled methods for training generative adversarial networks. In *Proceedings of the 5th international conference on learning representations*. Retrieved from https://openreview.net/forum?id=Hk4_qw5xe

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th international conference on machine learning* (pp. 214–223). PMLR. Retrieved from https://proceedings.mlr.press/v70/arjovsky17a.html

Bechtold, P., Chaboureau, J. P., Beljaars, A., Betts, A. K., Köhler, M., Miller, M., & Redelsperger, J. L. (2004). The simulation of the diurnal cycle of convective precipitation over land in a global model. *Quarterly Journal of the Royal Meteorological Society*, *130 C*(604), 3119–3137. https://doi.org/10.1256/qj.03.103

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J. P., Beljaars, A., & Bormann, N. (2014). Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, *71*(2), 734–753. https://doi.org/10.1175/JAS-D-13-0163.1

Ben-Bouallegue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., & Chantry, M. (2023). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. https://doi.org/10.48550/arXiv.2303.17195

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3D high-resolution model for fast and accurate global. *Weather and Forecasting*. Retrieved from http://arxiv.org/abs/2211.02556

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Boé, J., Terray, L., Habets, F., & Martin, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *International Journal of Climatology*, *27*(12), 1643–1655. https://doi.org/10.1002/joc.1602

Cafaro, C., Woodhams, B. J., Stein, T. H., Birch, C. E., Webster, S., Bain, C. L., et al. (2021). Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa? *Weather and Forecasting*, *36*(2), 697–716. https://doi.org/10.1175/WAF-D-20-0172.1

Camberlin, P., Gitau, W., Planchon, O., Dubreuil, V., Funatsu, B. M., & Philippon, N. (2018). Major role of water bodies on diurnal precipitation regimes in Eastern Africa. *International Journal of Climatology*, *38*(2), 613–629. https://doi.org/10.1002/joc.5197

Chamberlain, J. M., Bain, C. L., Boyd, D. F., Mccourt, K., Butcher, T., & Palmer, S. (2014). Forecasting storms over Lake Victoria using a high resolution model. *Meteorological Applications*, *21*(2), 419–430. https://doi.org/10.1002/met.1403

Chen, J., Janke, T., Steinke, F., & Lerch, S. (2024). Generative machine learning methods for multivariate ensemble postprocessing. *Annals of Applied Statistics*, *18*(1), 159–183. https://doi.org/10.1214/23-aoas1784

Clark, P., Roberts, N., Lean, H., Ballard, S. P., & Charlton-Perez, C. (2016). Convection-permitting models: A step-change in rainfall forecasting. *Meteorological Applications*, *23*(2), 165–181. https://doi.org/10.1002/met.1538

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Dai, Y., & Hemri, S. (2021). Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, *149*(12), 3923–3937. https://doi.org/10.1175/MWR-D-21-0046.1

Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, *57*(1), 16–26. https://doi.org/10.1016/j.gloplacha.2006.11.030

Dezfuli, A. K., Ichoku, C. M., Huffman, G. J., Mohr, K. I., Selker, J. S., De Giesen, N. V., et al. (2017). Validation of IMERG precipitation in Africa. *Journal of Hydrometeorology*, *18*(10), 2817–2825. https://doi.org/10.1175/JHM-D-17-0139.s1

Dinku, T., Connor, S. J., & Ceccato, P. (2010). Comparison of CMORPH and TRMM-3B42 over mountainous regions of Africa and South America. In M. Gebremichael & F. Hossain (Eds.), *Satellite rainfall applications for surface hydrology* (pp. 193–204). Springer Netherlands. https://doi.org/10.1007/978-90-481-2915-7_11

Duncan, J., Subramanian, S., & Harrington, P. (2022). Generative modeling of high-resolution global precipitation forecasts. https://doi.org/10.48550/arXiv.2210.12504

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological applications*, *15*(1), 51–64. https://doi.org/10.1002/met.25

ECMWF. (2023a). Changes to the forecasting system - forecast user - ECMWF confluence wiki. Retrieved from https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system

ECMWF. (2023b). Operational configurations of the ECMWF integrated forecasting system (IFS). Retrieved from https://confluence.ecmwf.int/pages/viewpage.action?pageId=324860211

ECMWF. (2023c). Parameter database. Retrieved from https://codes.ecmwf.int/grib/param-db/

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Finney, D. L., Marsham, J. H., Jackson, L. S., Kendon, E. J., Rowell, D. P., Boorman, P. M., et al. (2019). Implications of improved representation of convection for the East Africa water budget using a convection-permitting model. *Journal of Climate*, *32*(7), 2109–2129. https://doi.org/10.1175/JCLI-D-18-0387.1

Floodlist. (2023). Somalia – flooding displaces thousands, prompts urgent humanitarian response. Retrieved from https://floodlist.com/africa/somalia-floods-may-2023

Gebremeskel Haile, G., Tang, Q., Sun, S., Huang, Z., Zhang, X., & Liu, X. (2019). Droughts in East Africa: Causes, impacts and resilience. *Earth-Science Reviews*, *193*, 146–161. https://doi.org/10.1016/j.earscirev.2019.04.015

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, *24*(5), 1416–1430. https://doi.org/10.1175/2009WAF2222269.1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Skaugen, T. E. (2012). Quantile mapping hydrology and earth system sciences discussions technical note: Downscaling RCM precipitation to the station scale using quantile mapping-a comparison of methods quantile mapping. *Hydrology and Earth System Sciences Discussions*, *9*, 6185–6201. https://doi.org/10.5194/hessd-9-6185-2012

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in neural information processing systems* (Vol. 30). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

Haiden, T., Rodwell, M. J., Richardson, D. S., Okagaki, A., Robinson, T., & Hewson, T. (2012). Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Monthly Weather Review*, *140*(8), 2720–2733. https://doi.org/10.1175/MWR-D-11-00301.1

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. https://doi.org/10.1029/2022ms003120

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. https://doi.org/10.48550/arXiv.1502.01852

Huffman, G., Bolvin, D., Braithwaite, D., Hsu, K., Joyce, R., & Xie, P. (2022). Integrated multi-satellite retrievals for GPM (IMERG), V06B [Dataset]. *NASA's Precipitation Processing Center*. Retrieved from https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmallversions/V06/YYYY/MM/DD/imerg/

IFRC. (2014). *World disasters report, focus on culture and risk* (Technical Report). International Federation of Red Cross and Red Crescent Societies. Retrieved from https://www.ifrc.org/document/world-disasters-report-2014

Jeong, C. H., & Yi, M. Y. (2023). Correcting rainfall forecasts of a numerical weather prediction model using generative adversarial networks. *The Journal of Supercomputing*, *79*(2), 1289–1317. https://doi.org/10.1007/s11227-022-04686-y

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.48550/arXiv.1812.04948

Kilavi, M., MacLeod, D., Ambani, M., Robbins, J., Dankers, R., Graham, R., et al. (2018). Extreme rainfall and flooding over Central Kenya including Nairobi City during the long-rains season 2018: Causes, predictability, and potential for early warning and actions. *Atmosphere*, *9*(12), 472. https://doi.org/10.3390/atmos9120472

Kim, J. E., & Joan Alexander, M. (2013). Tropical precipitation variability and convectively coupled equatorial waves on submonthly time scales in reanalyses and TRMM. *Journal of Climate*, *26*(10), 3013–3030. https://doi.org/10.1175/JCLI-D-12-00353.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Leinonen, J., Hamann, U., Nerini, D., Germann, U., & Franch, G. (2023). Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. Retrieved from http://arxiv.org/abs/2304.12891

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

MacLeod, D., Dankers, R., Graham, R., Guigma, K., Jenkins, L., Todd, M. C., et al. (2021). Drivers and subseasonal predictability of heavy rainfall in Equatorial East Africa and relationship with flood risk. *Journal of Hydrometeorology*, *22*(4), 887–903. https://doi.org/10.1175/JHM-D-20-0211.1

MacLeod, D., Kilavi, M., Mwangi, E., Ambani, M., Osunga, M., Robbins, J., et al. (2021). Are Kenya Meteorological Department heavy rainfall advisories useful for forecast-based early action and early preparedness for flooding? *Natural Hazards and Earth System Sciences*, *21*(1), 261–277. https://doi.org/10.5194/nhess-21-261-2021

Manzato, A., & Jolliffe, I. (2017). Behaviour of verification measures for deterministic binary forecasts with respect to random changes and thresholding. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1903–1915. https://doi.org/10.1002/qj.3050

Maraun, D., & Widmann, M. (2017). Model output statistics. In *Statistical downscaling and bias correction for climate research* (pp. 170–200). Cambridge University Press. https://doi.org/10.1017/9781107588783.013

Marsham, J. H., Dixon, N. S., Garcia-Carreras, L., Lister, G. M., Parker, D. J., Knippertz, P., & Birch, C. E. (2013). The role of moist convection in the West African monsoon system: Insights from continental-scale convection-permitting simulations. *Geophysical Research Letters*, *40*(9), 1843–1849. https://doi.org/10.1002/grl.50347

Mittermaier, M., Roberts, N., & Thompson, S. A. (2013). A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteorological Applications*, *20*(2), 176–186. https://doi.org/10.1002/met.296

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. Retrieved from http://arxiv.org/abs/2301.10343

Nicholson, S. E. (2016). The Turkana low-level jet: Mean climatology and association with regional aridity. *International Journal of Climatology*, *36*(6), 2598–2614. https://doi.org/10.1002/joc.4515

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. *Reviews of Geophysics*, *55*(3), 590–635. https://doi.org/10.1002/2016rg000544

Pacchiardi, L., Adewoyin, R., Dueben, P., & Dutta, R. (2022). Probabilistic forecasting with generative networks via scoring rule minimization. Retrieved from http://arxiv.org/abs/2112.08217

Palmer, P. I., Wainwright, C. M., Dong, B., Maidment, R. I., Wheeler, K. G., Gedney, N., et al. (2023). Drivers and impacts of Eastern African rainfall variability. *Nature Reviews Earth & Environment*, *4*(4), 254–270. https://doi.org/10.1038/s43017-023-00397-x

Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *Proceedings of the 25th international conference on artificial intelligence and statistics (AISTATS)* (Vol. 151). Retrieved from https://proceedings.mlr.press/v151/price22a/price22a.pdf

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1.0) [Software]. *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

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

Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological applications*, *15*(1), 163–169. https://doi.org/10.1002/met.57

Roberts, N., & Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Roca, R., Chambon, P., Jobard, I., Kirstetter, P. E., Gosset, M., & Bergés, J. C. (2010). Comparing satellite and surface rainfall products over West Africa at meteorologically relevant scales during the AMMA campaign using error estimates. *Journal of Applied Meteorology and Climatology*, *49*(4), 715–731. https://doi.org/10.1175/2009JAMC2318.1

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, *5*(4), 570–575. https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2

Sinclair, S., & Pegram, G. G. S. (2005). Empirical mode decomposition in 2-D space and time: A tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, *9*(3), 127–137. https://doi.org/10.5194/hess-9-127-2005

Thiery, W., Davin, E. L., Panitz, H.-J., Demuzere, M., Lhermitte, S., & Lipzig, N. v. (2015). The impact of the African great lakes on the regional climate. *Journal of Climate*, *28*(10), 4061–4085. https://doi.org/10.1175/JCLI-D-14-00565.1

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, *33*(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2020). Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics. *Weather and Forecasting*, *35*(6), 2367–2385. https://doi.org/10.1175/WAF-D-20-0082.1

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*(10), e2022JD038163. https://doi.org/10.1029/2022JD038163

Wainwright, C. M., Finney, D. L., Kilavi, M., Black, E., & Marsham, J. H. (2021). Extreme rainfall in East Africa, October 2019–January 2020 and context under future climate change. *Weather*, *76*(1), 26–31. https://doi.org/10.1002/wea.3824

Walker, D. P., Birch, C. E., Marsham, J. H., Scaife, A. A., Graham, R. J., & Segele, Z. T. (2019). Skill of dynamical and GHACOF consensus seasonal forecasts of East African rainfall. *Climate Dynamics*, *53*(7–8), 4911–4935. https://doi.org/10.1007/s00382-019-04835-9

Walz, E.-M., Knippertz, P., Fink, A. H., Köhler, G., & Gneiting, T. (2024). Physics-based vs. data-driven 24-hour probabilistic forecasts of precipitation for northern tropical Africa. *Monthly Weather Review*, *152*(9), 2011–2031. https://doi.org/10.1175/MWR-D-24-0005.1

Warner, J. L., Petch, J., Short, C. J., & Bain, C. (2023). Assessing the impact of a NWP warm-start system on model spin-up over tropical Africa. *Quarterly Journal of the Royal Meteorological Society*, *149*(751), 621–636. https://doi.org/10.1002/qj.4429

Watkiss, P., & Cimato, F. (2021). *Socio-economic benefits of the WISER programme. Synthesis of results* (Technical Report). Met Office UK. Retrieved from https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/business/international/wiser/wiser-seb-results_final-web.pdf

Watkiss, P., Powell, R., Hunt, A., & Cimato, F. (2020). *The socio-economic benefits of the HIGHWAY project* (Technical Report). Weather and Climate Information Services for Africa (WISER).

Watson, P. A. G. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, *11*(5), 1402–1417. https://doi.org/10.1029/2018MS001597

Watson, P. A. G. (2022). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, *17*(11), 111004. https://doi.org/10.1088/1748-9326/ac9d4e

WFP, Oxford, & ICPAC. (2024). AI-led science innovation protects communities hit by climate change. Retrieved from https://www.ox.ac.uk/news/2024-06-25-ai-led-science-innovation-protects-communities-hit-climate-change

Wilks, D. S. (2019). Forecast verification. In *Statistical methods in the atmospheric sciences* (pp. 369–483). Elsevier. https://doi.org/10.1016/b978-0-12-815823-4.00009-2

Woodhams, B. J., Birch, C. E., Marsham, J. H., Bain, C. L., Roberts, N. M., & Boyd, D. F. (2018). What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Monthly Weather Review*, *146*(9), 2757–2780. https://doi.org/10.1175/MWR-D-17-0396.1

Woodhams, B. J., Birch, C. E., Marsham, J. H., Lane, T. P., Bain, C. L., & Webster, S. (2019). Identifying key controls on storm formation over the lake Victoria basin. *Monthly Weather Review*, *147*(9), 3365–3390. https://doi.org/10.1175/MWR-D-19-0069.1

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., et al. (2023). A survey on video diffusion models. *ACM Computing Surveys*, *57*(2), 1–42. https://doi.org/10.1145/3696415

Yang, S., Ling, F., Bai, L., & Luo, J.-J. (2025). Improving the seasonal forecast of summer precipitation in southeastern China using a CycleGan-based deep learning bias correction method. *Advances in Atmospheric Sciences*, *42*(1), 26–35.

Youds, L. H., Parker, D. J., Adefisan, E. A., Antwi-Agyei, P., Bain, C. L., Black, E. C. L., et al. (2021). *GCRF African SWIFT and ForPAc SHEAR white paper on the potential of operational weather prediction to save lives and improve livelihoods and economies in sub-saharan Africa*. University of Leeds. https://doi.org/10.5518/100/79

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, *1–7*(7970), 526–532. https://doi.org/10.1038/s41586-023-06184-4