Postprocessing East African rainfall forecasts using a generative machine learning model

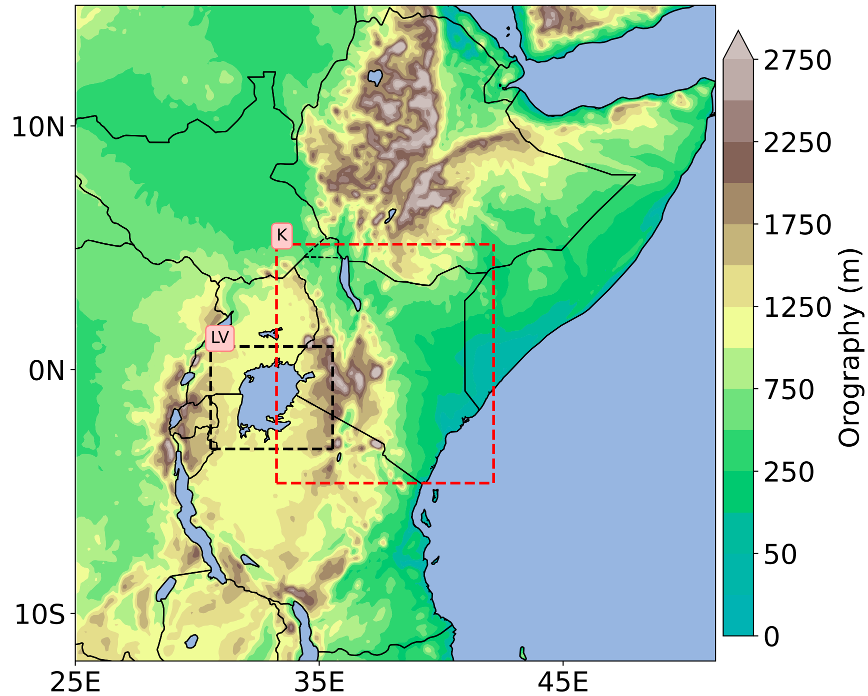
Bobby Antonio[1], Andrew T T McRae[1], David MacLeod[2], Fenwick C Cooper[1], John Marsham[3], Laurence Aitchison[4], Tim N Palmer[5], and Peter A G Watson[4]

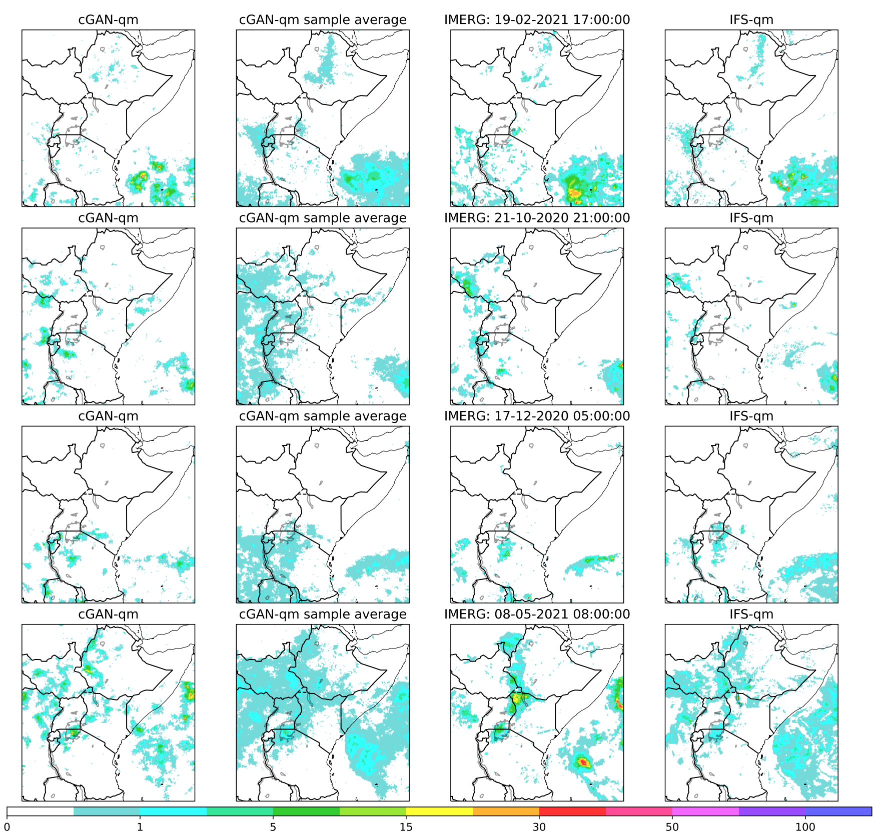
[1]University of Oxford
[2]Cardiff University
[3]University of Leeds
[4]University of Bristol
[5]Oxford University

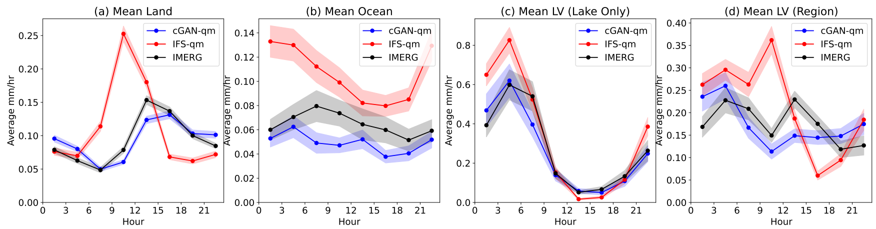
March 05, 2024

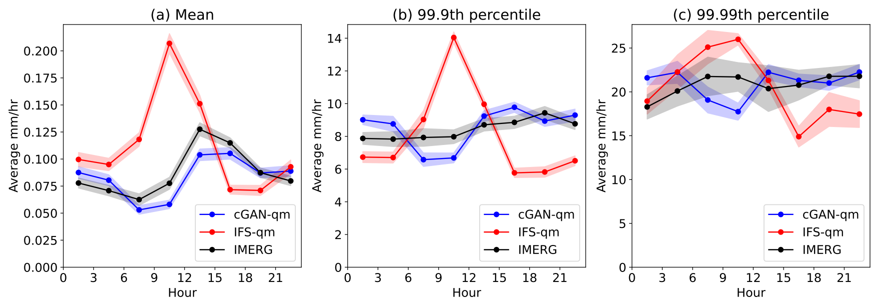
Abstract

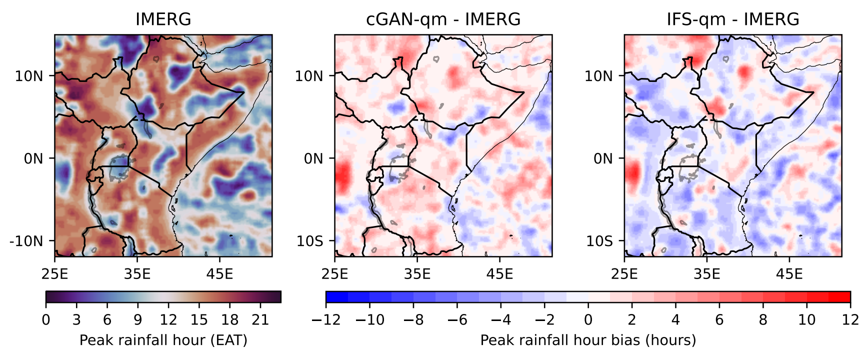
Existing weather models are known to have poor skill at forecasting rainfall over East Africa, where there are regular threats of drought and floods. Improved forecasts could reduce the effects of these extreme weather events and provide significant socioeconomic benefits to the region. We present a novel machine learning-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 6-18h lead times, at $0.1ˆ{\circ}$ resolution. We combine the cGAN predictions with a novel neighbourhood version of quantile mapping, to integrate the strengths of machine learning and conventional postprocessing. Our results indicate that the cGAN substantially improves the diurnal cycle of rainfall, and improves predictions up to the $99.9ˆ{\text{th}}$ percentile ($\sim 10 \text{mm}/\text{hr}$). This improvement extends to the 2018 March–May season, which had extremely high rainfall, indicating that the approach has some ability to generalise 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 characterised by a mixture of under- and over-dispersion. Overall our results demonstrate how the strengths of machine learning and conventional postprocessing methods can be combined, and illuminate what benefits machine learning approaches can bring to this region.

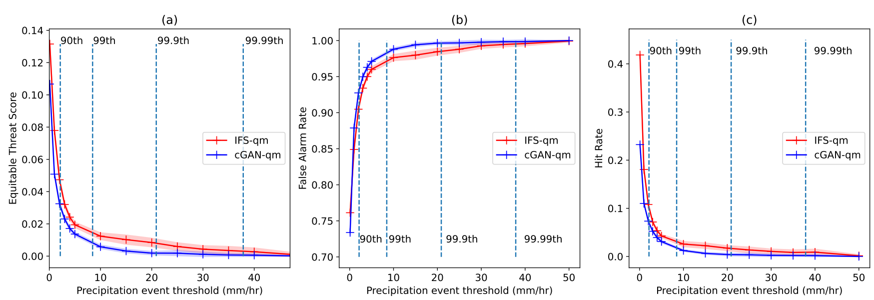

10N
0N
10S
25E
35E
45E
K
LV
2750
2250
1750
1250
750
250
50
0
Orography (m)

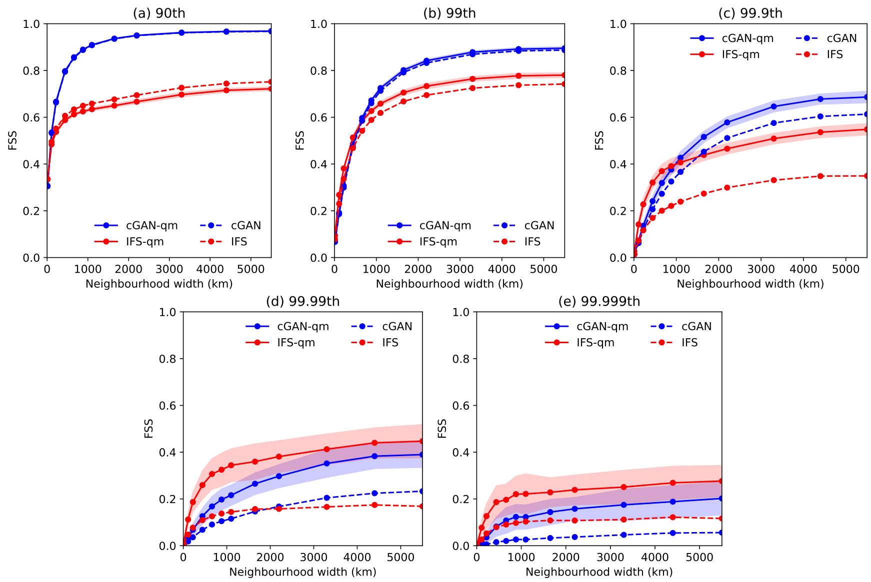

cGAN-qm
cGAN-qm sample average
IMERG: 19-02-2021 17:00:00
IFS-qm
cGAN-qm
cGAN-qm sample average
IMERG: 21-10-2020 21:00:00
IFS-qm
cGAN-qm
cGAN-qm sample average
IMERG: 17-12-2020 05:00:00
IFS-qm
cGAN-qm
cGAN-qm sample average
IMERG: 08-05-2021 08:00:00
IFS-qm
0
1
5
15
30
50
100
Precipitation (mm / hr)

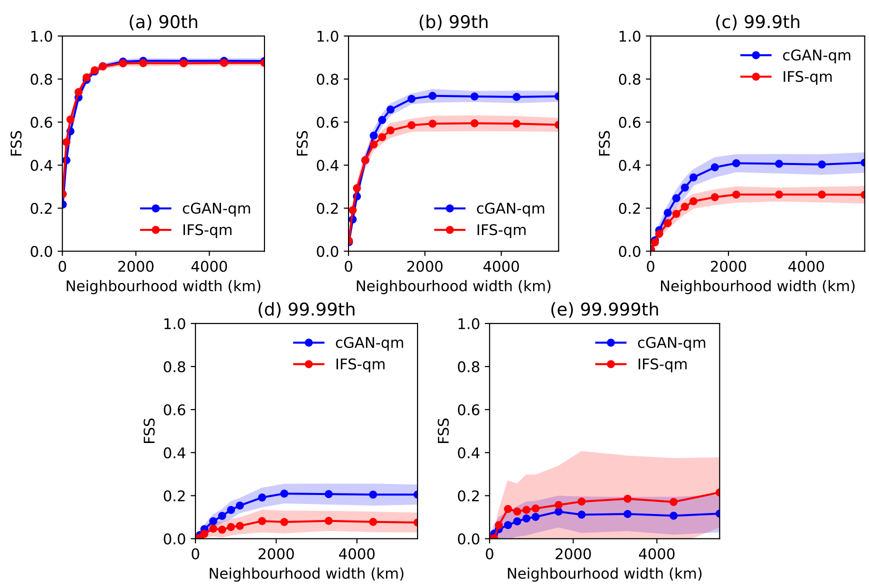

(a) Mean Land
(b) Mean Ocean
(c) Mean LV (Lake Only)
(d) Mean LV (Region)
cGAN-qm
IFS-qm
IMERG
Average mm/hr
Hour

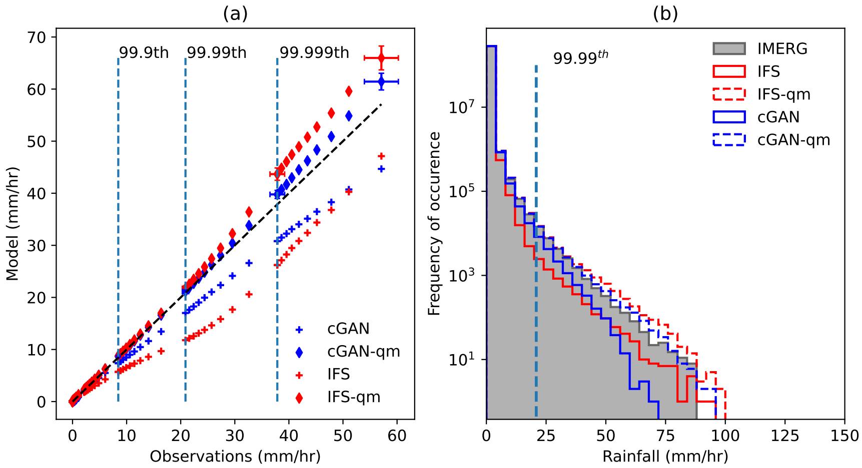

(a) Mean
(b) 99.9th percentile
(c) 99.99th percentile
cGAN-qm
IFS-qm
IMERG
Average mm/hr
Hour

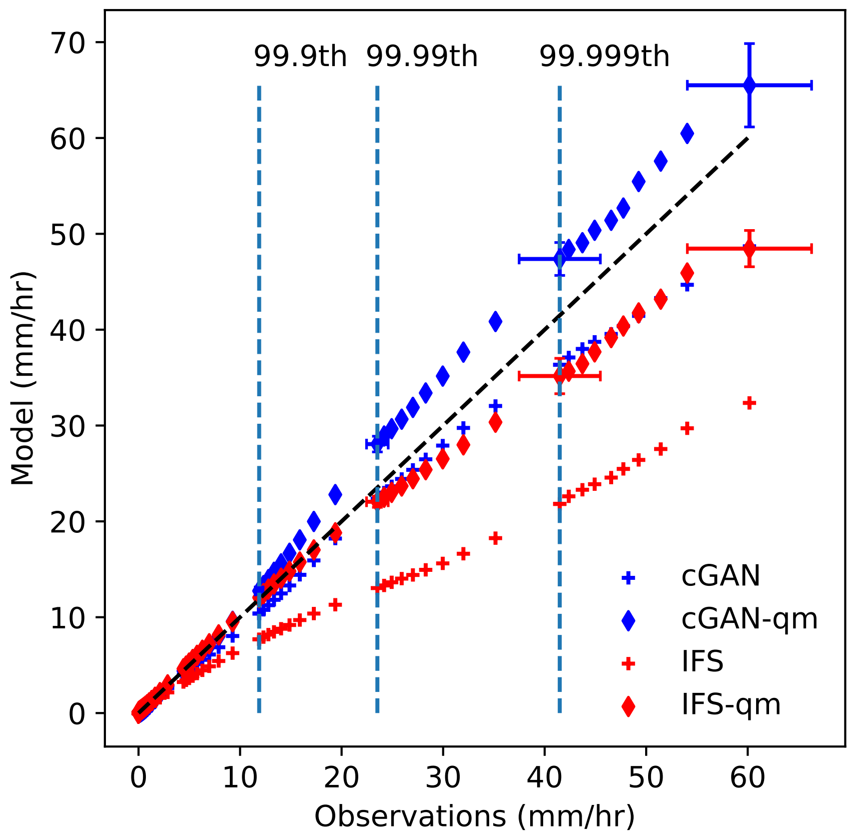

IMERG
cGAN-qm - IMERG
IFS-qm - IMERG
10N
0N
-10N
10S
25E
35E
45E
Peak rainfall hour (EAT)
Peak rainfall hour bias (hours)

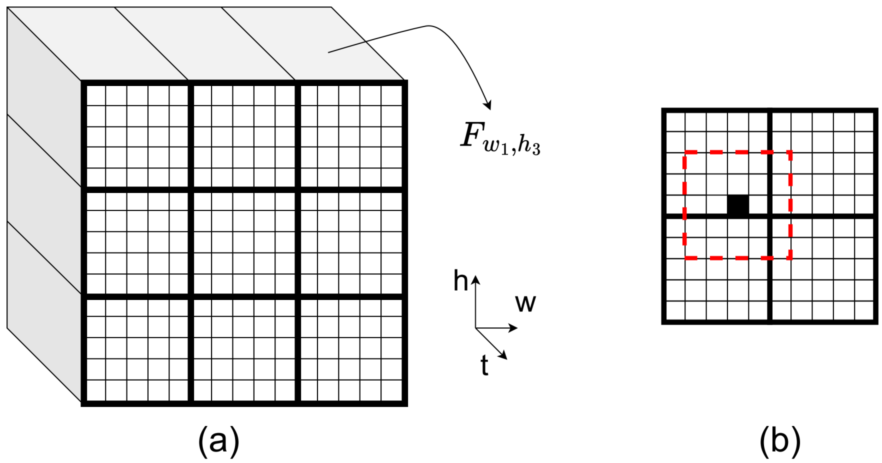

(a)
(b)
(c)
90th
99th
99.9th
99.99th
IFS-qm
cGAN-qm
Equitable Threat Score
False Alarm Rate
Hit Rate
Precipitation event threshold (mm/hr)

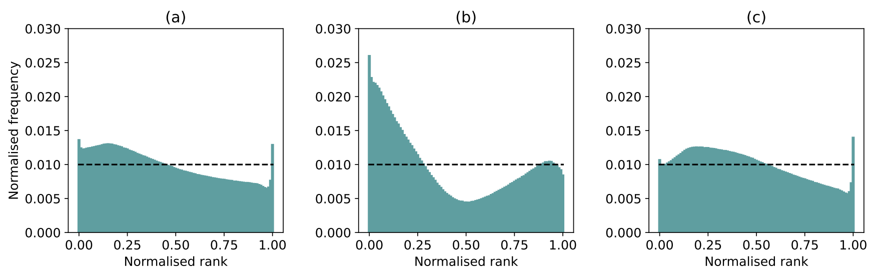

(a) 90th
(b) 99th
(c) 99.9th
(d) 99.99th
(e) 99.999th
cGAN-qm
IFS-qm
cGAN
IFS
FSS
Neighbourhood width (km)

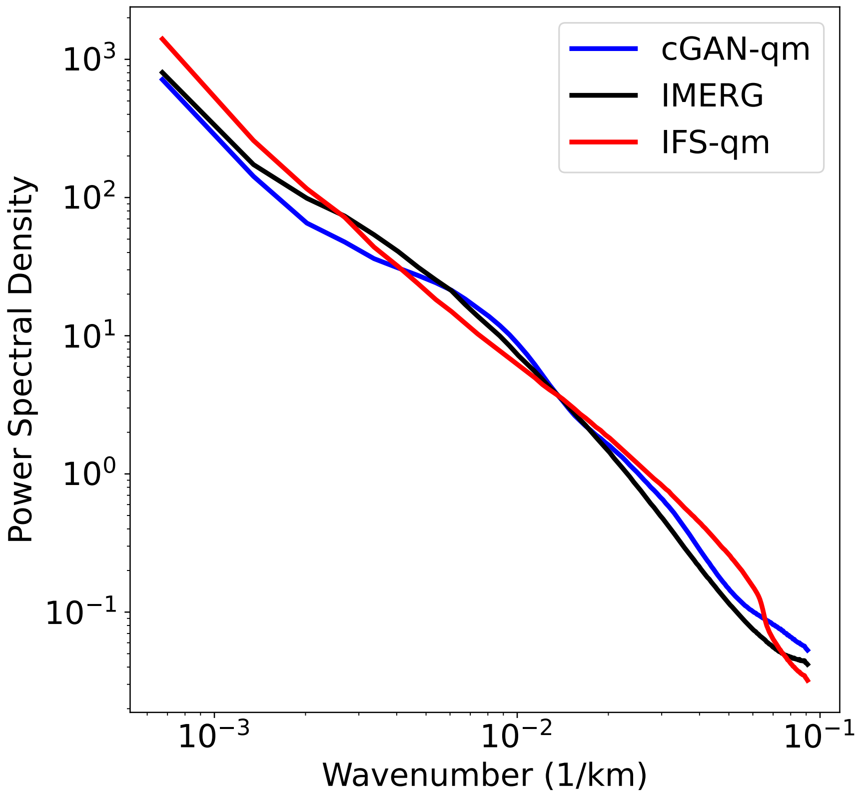

(a) 90th
(b) 99th
(c) 99.9th
(d) 99.99th
(e) 99.999th
cGAN-qm
IFS-qm
FSS
Neighbourhood width (km)


(a)
(b)
99.9th
99.99th
99.999th
Model (mm/hr)
Observations (mm/hr)
cGAN
cGAN-qm
IFS
IFS-qm
99.99^{th}
IMERG
IFS
IFS-qm
cGAN
cGAN-qm
Frequency of occurence
Rainfall (mm/hr)

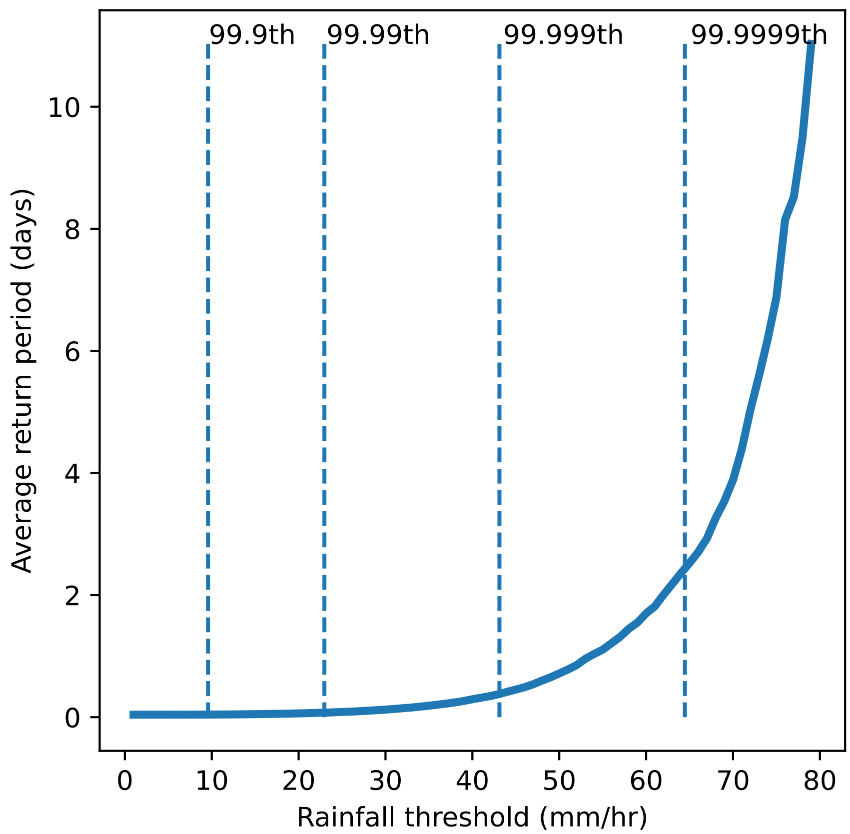

99.9th
99.99th
99.999th
Model (mm/hr)
Observations (mm/hr)
cGAN
cGAN-qm
IFS
IFS-qm

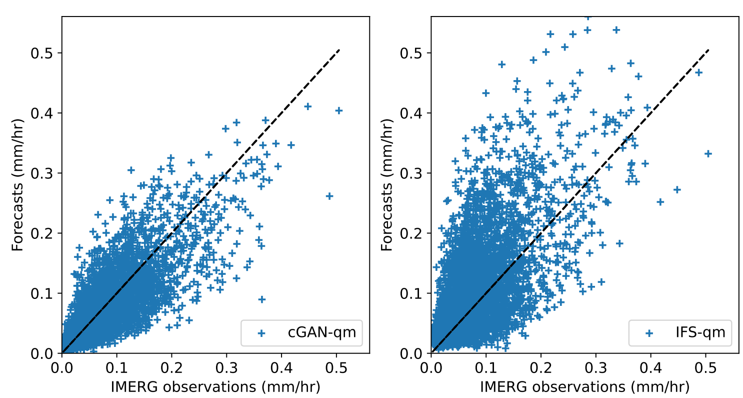

F_{w_1,h_3}
h
w
t
(a)
(b)

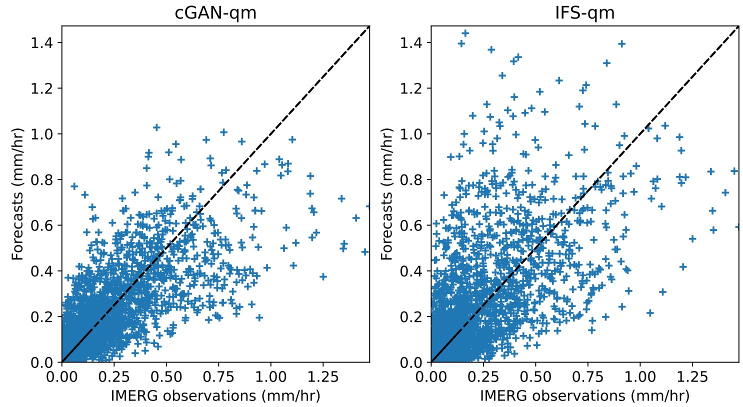

(a)
(b)
(c)
Normalised frequency
Normalised rank
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.00
0.25
0.50
0.75
1.00

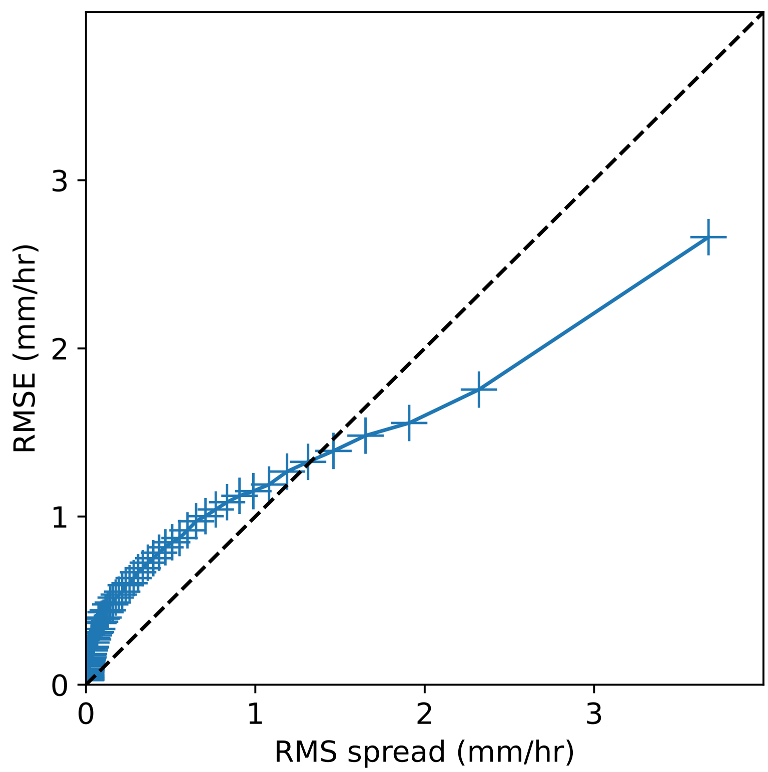


cGAN-qm
IMERG
IFS-qm
Power Spectral Density
10^3
10^2
10^1
10^0
10^{-1}
10^{-3}
10^{-2}
10^{-1}
Wavenumber (1/km)


99.9th
99.99th
99.999th
99.9999th
Average return period (days)
Rainfall threshold (mm/hr)
0
2
4
6
8
10
0
10
20
30
40
50
60
70
80


Forecasts (mm/hr)
IMERG observations (mm/hr)
cGAN-qm
IFS-qm
0.0
0.1
0.2
0.3
0.4
0.5


cGAN-qm
IFS-qm
Forecasts (mm/hr)
IMERG observations (mm/hr)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.00
0.25
0.50
0.75
1.00
1.25


RMSE (mm/hr)
RMS spread (mm/hr)
0
1
2
3

radar data in the area. In Ageet et al. (2022) a range of satellite rainfall estimating products, 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 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. Woodhams et al. (2018); Finney et al. (2019); Cafaro et al. (2021)).

The forecast dataset used is the ECMWF IFS HRES deterministic hourly forecast (ECMWF, 2023) as this tends to perform amongst the best compared to similar models (Haiden et al., 2012). IFS forecasts are provided at 00h and 12h and we use lead times within a 6-18h window, corresponding to short-range weather prediction (however it is expected that the method we use could also equally apply to longer lead times). The data is interpolated from 9km × 9km 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, 2023).

2.3 Machine Learning Model

Our model architecture uses the same architecture and code that L. 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 the IMERG data 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 $2 \times 7 \times f_{\mathrm{g}}$ 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 $2 \times 3 \times f_{\mathrm{d}}$ intermediate arrays each of size 200×200. Excluding the output layers, PReLU activation functions were used, where we set the α parameter for the PReLU to 0.2 following L. 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 optimiser was used for training.

Following L. 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 parameterised 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 $f_{\mathrm{g}} = 32, f_{\mathrm{d}} = 128$ were trained for 6.4×10^4 iterations, and then finally larger models with $f_{\mathrm{g}} = 64, f_{\mathrm{d}} = 256$ were trained for 3.2×10^5 steps, and used for evaluation; thus our largest model was smaller than the model in L. Harris et al. (2022) that had $f_{\mathrm{g}} = 128, f_{\mathrm{d}} = 512$. 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 L. 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 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 L. Harris et al. (2022), we normalise the input variables; precipitation variables are log-normalised via $x \to \log_{10}(1 + x)$, whilst others were either divided by the maximum value, or normalised to fall within the maximum and minimum values (see Appendix A).

Model checkpoints were saved every 3200 steps, and the best model in the last 1/3rd of checkpoints was selected based on judgement of the combined performance on CRPS, RAPSD and mean squared error, plus visual evaluation of the samples produced. 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.

One notable addition by L. Harris et al. (2022) is the inclusion of a 'content loss' term, inspired by Ravuri et al. (2021), which penalises 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 normalisation of the output precipitation predictions, as done by L. Harris et al. (2022), produced a distribution of rainfall that tended to greatly overestimate the observations at the extreme rainfall values. Removing the log normalisation 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 L. 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 RALSD (see Sec. 2.6).

In L. Harris et al. (2022), samples are grouped into prede ned 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 for our data, the threshold used by L. Harris et al. (2022) for postprocessing UK rainfall was not appropriate, and our attempts at using a similar approach did not give any improvements on the validation set. Therefore we did not apply this oversampling.

To increase the variation in the samples seen during training, we randomly cropped the 270 265 images to smaller images of 200200, as this has been demonstrated to improve the generalisability of deep learning models (Goodfellow et al., 2016), and produces output similar in size to that in L. Harris et al. (2022).

Similarly to the results in L. 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 dataset set aside in order to choose the nal model, and this data was not used in the nal training.

2.3.1 Training and evaluation strategy

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

- Training set: March 2016 { February 2018 and July 2018 { Sept 2020 (excluding validation months)
- Validation set: Jun 2018, Oct 2018, Jan 2019, March 2019
- Test sets: October 2020 - September 2021, March - May 2018

We used the nal year as a primary test dataset, 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 signi cantly higher than in any other season in the full IMERG dataset.

The purpose of the validation dataset is to guide choice of the model structure and hyperparameters. The development process was to train di erent versions of the model on the training dataset, then evaluate these on the validation dataset to select the best version. This avoids over tting 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 maximise the amount of training data by including a month from each of the di erent seasons in the validation period. This sampling variability observed for this size of validation data also indicated there was no additional bene t from using a whole year.

All evaluation results reported in Secs. 3.1 and 3.2 are evaluated on 4000 unique hours randomly sampled from the unseen test period October 2020 - September 2021, with 20 ensemble members used in the example plots. The ensemble calibration results in Sec. 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 Sec. 3.4, we analyse all of the hours from March-May 2018. Since much of the anomalous rainfall in this season was concentrated over Kenya, we restrict our analysis to this region (4.6°S 5.1°N, 33.2° 43.1°E, see Fig. 1).

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 speci cally tuned to reproduce the properties of IMERG observations, we applied quantile mapping to the IFS forecasts (see e.g. Maraun and 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 Bœ et al. (2007), Deque (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 $\tilde{x}_{\mathrm{f}}$ according to:

$$\tilde{x}_{\mathrm{f}} = F_{\mathrm{o}}^{-1}(F_{\mathrm{f}}(x_{\mathrm{f}})) \tag{1}$$

In Bœ 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 0mm/hr there is no way to tell which quantile it belongs to. We follow the method in Bœ 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 towards 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$^{\text{th}}$ percentile, above which we observed significant sampling variability. Note that these percentiles are calculated at the grid box level, so that the number of samples available to calculate these quantiles is 4000270 265 grid boxes for the validation and test datasets.

For data greater than the maximum value observed in training, we follow the additive uplift method (Bœ et al., 2007; Deque, 2007) and add the uplift of the highest quan-

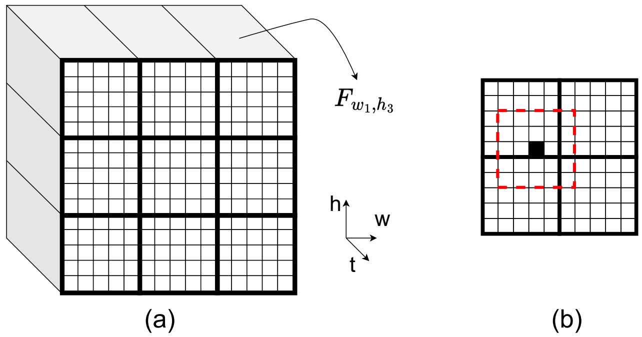


Figure 2: a) Illustration of the general method for how grid cells are grouped together in order to estimate quantiles. In this example, the spatial domain is split into squares of 4 × 4 grid cells, giving 9 separate large square regions, and in each region the quantiles are calculated. b) To calculate 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 $\frac{12}{25}$, $\frac{3}{25}$, $\frac{8}{25}$, $\frac{2}{25}$.

tile 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 (Fig. 2(a)). The intuition is that nearby points will have similar distributions and so we can gain accuracy by grouping nearby points together.

To avoid any artefacts 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 (Fig. 2(b)). 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 99999th 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 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.

2.6 Forecast verification measures

2.6.1 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; D. Harris et al., 2001). 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 summarise the overall similarity of two RAPSD curves we use the Radially Averaged Log Spectral Distance (RALSD) defined in L. Harris et al. (2022).

2.6.2 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, 2019a). 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:

$$\mathrm{ETS} := \frac{\mathrm{TP} - \mathrm{TP}_\mathrm{r}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} - \mathrm{TP}_\mathrm{r}} \tag{2}$$

where TP, FP, 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 often estimated from the data using the formula:

$$\mathrm{TP}_\mathrm{r} = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}{N} \tag{3}$$

2.6.3 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 neighbourhood 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 neighbouring 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:

$$\mathrm{FSS}(n, r) := \frac{\sum_{t=1}^{T} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} 2F_{tij} O_{tij}}{\sum_{t=1}^{T} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(F_{tij}^2 + O_{tij}^2 \right)} \tag{4}$$

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

$$\mathrm{FSS}_\infty = \frac{2 f_o f_m}{f_o^2 + f_m^2} \tag{5}$$

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

2.6.4 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:

$$s_t^2 = \frac{1}{M} \sum_{i=1}^{M} (f_{i,t} - \mu_t)^2 \tag{6}$$

$$e_t^2 = (y_t - \mu_t)^2 \tag{7}$$

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

$$\tilde{s}_t = \frac{M+1}{M-1} s_t \tag{8}$$

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.

2.6.5 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, 2019a). 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, 2019a). In this work we use the pySTEPS implementation of the rank histogram (Pulkkinen et al., 2019).

2.6.6 Continuous Ranked Probability Score (CRPS)

The Continuous Ranked Probability Score, or CRPS, is a particularly important score in assessing probabilistic forecasts, as it is a *strictly proper* score (Wilks, 2019b), which means that the score is only maximised when the forecast distribution equals the target distribution. For a cumulative forecast distribution $F(y)$, the CRPS for an observed occurrence of x (e.g. observed rainfall value) is defined as:

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left[F(y) - 1_{y \geq x}\right]^2 dy \tag{9}$$

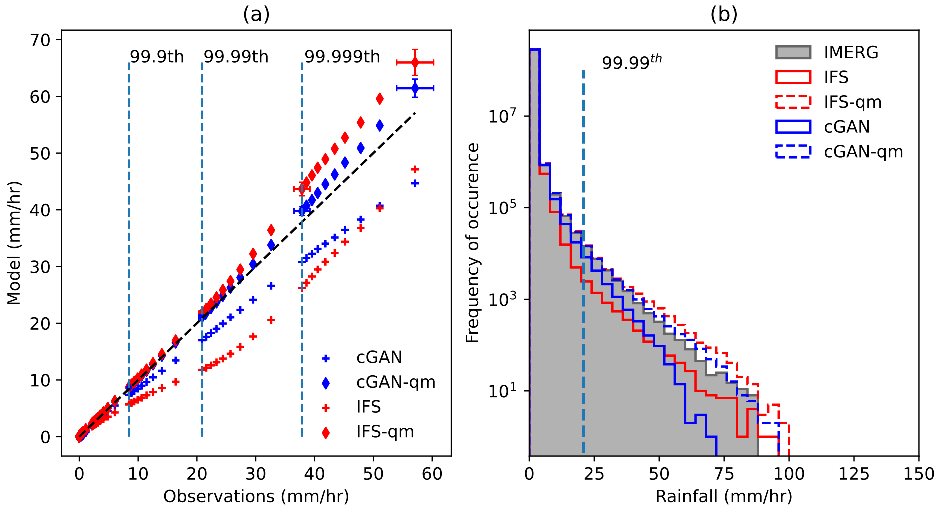


Figure 3: (a) Quantile-quantile plot, up to the 99.9999th percentile. Red circles (diamonds) indicate quantiles for the IFS (IFS-qm) model. Blue circles (diamonds) indicate quantiles for the cGAN (cGAN-qm) model. The black dashed line is the line along which a perfectly calibrated forecast would sit. The error bars indicate an estimate of 2 standard errors from 1000 bootstrap samples (only shown for the 9**9**99th and 99.9999th percentiles). (b) A histogram showing the distribution of rainfall values; the vertical dashed blue line indicates the 9999th percentile of observed rainfall

where $1_{y \geq x}$ is the Heaviside step function. The CRPS is a univariate measure, so does not properly account for spatial correlations. Whilst there is a multivariate generalisation of the CRPS, the energy score (Gneiting & Raftery, 2007), the CRPS is more commonly used and has been used in previous related works, so we use it in this work as one of many validation metrics, in order to choose the best model.

3 Evaluation

In this section we presents results of evaluating the model on unseen data. Evaluations are performed on the primary test dataset except for Sec. 3.4 which is evaluated on the extreme Long Rains of March{May 2018.

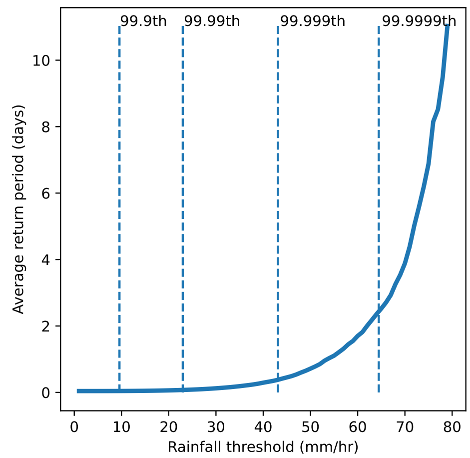


Figure 4: The approximate return periods for di erent rainfall thresholds to be exceeded at any grid point in the spatial domain in the training set. The dashed lines indicate the values of particular high percentiles.

3.1 Climatological properties of the forecasts

We rst assess how well the forecasts capture the distribution of rainfall, shown by a quantile-quantile plot and a histogram of rainfall distribution for 4000 samples, shown in Fig. 3 (a) and (b) respectively. The unpostprocessed model outputs are shown together with quantile-mapped outputs. From these we can see that, without postprocessing, the distribution of cGAN output is an improvement upon that of the IFS output up to extremely high levels of rainfall (around 50mm/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 quanti ed by performing 1000 iterations of bootstrapping (see Sec. 2.5) to estimate the standard error of the quantiles. These are shown in Fig. 3 (a), where each error bar shows 2 standard errors. The quantile-mapped forecasts' extreme values are slightly larger than in the observations, which we attribute to sampling variability between the training and test periods.

In order to also get a sense of how extreme these quantile values are, we plot the approximate return period in days for a range of thresholds in Fig. 4, calculated over all hours in the test period. These are calculated as the average time gap between instances

Figure 5: Example precipitation forecasts following postprocessing by the cGAN-qm model, for a selection of hours throughout the primary test dataset (rst column). The examples are from randomly chosen dates, but ltered to ensure that a diverse range of months in the year are represented, and so that the examples show di erent behaviours 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 IMERG observations, the IFS-qm forecast. Each row corresponds to one time value, shown in the title of the IMERG sample.

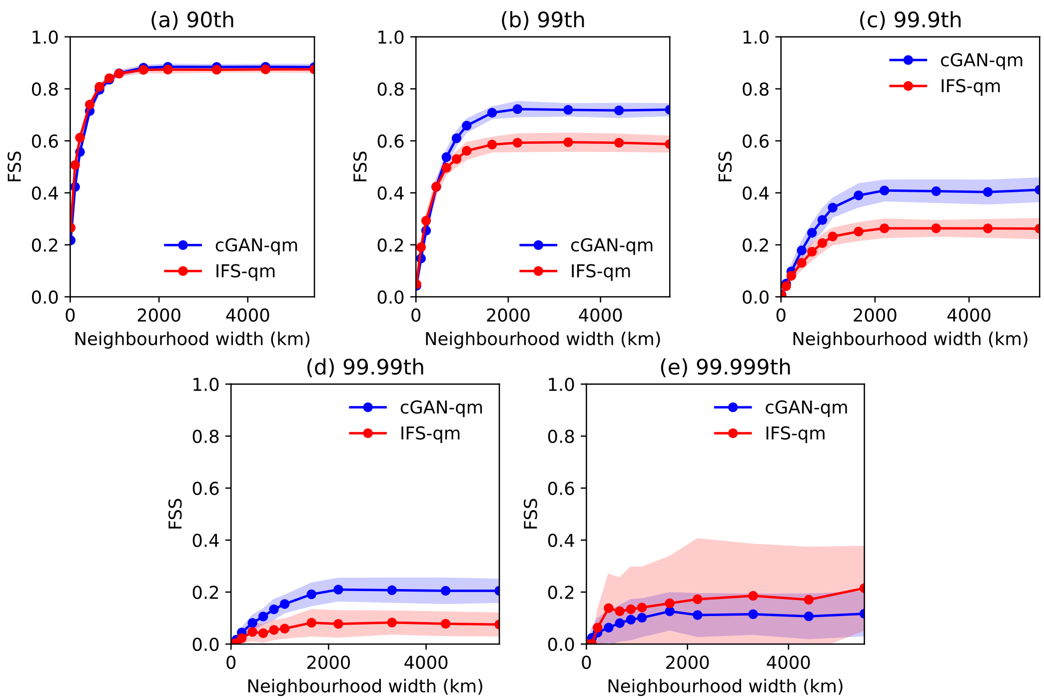


Figure 18: Fractions Skill Score in the extreme March{May 2018 season, over the Kenya subregion, for di erent quantile thresholds, with quantiles calculated for this season and subregion (a) 90^{th} percentile (b) 99^{th} percentile (c) 99.9^{th} percentile (d) 99.99^{th} percentile (e) 99.999^{th} percentile.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

els 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. However, there is substantial spatial noise in the cGAN-qm peak rainfall hour (Fig. 9). Using time as an input variable may be one way to improve the learnt relationships. The frequency distributions of rainfall for cGAN-qm and IFS-qm (Fig. 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 (Fig. 12) up to a high percentile (99^{th}), particularly at larger neighbourhood sizes (above 500km). For higher percentiles the IFS-qm forecast demonstrated a higher score. The IFS-qm model also achieved higher ETS at the grid scale at all thresholds, although the scores were nevertheless quite low (Fig. 13).

Both models were also evaluated on the 2018 Long Rains, which were significantly wetter than any Long Rains season seen in training and across the whole IMERG dataset. It may be expected that machine learning-based methods would show degraded performance on situations 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 rainfall than that of the IFS (Fig. 16), and cGAN-qm forecasts had higher skill than IFS-qm when evaluated using scatter plots and the FSS (Figs. 17 and 18). However, more evaluation work would be required to have high confidence that the method will generally perform well in extreme situations.

An important advantage that generative machine learning 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 machine learning 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 behaviour (Sec. 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 machine learning models compare the output of a model to an unpostprocessed 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 (Fig. 2), diurnal cycle, and skill scores (Fig. 12), by evaluating our model against a strong baseline of quantile mapping our comparison reveals what machine learning 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 machine learning 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 machine learning model is able to correct for errors in how the IFS evolves this initial state.

There are also other state-of-the-art machine learning models that would be interesting to compare with, to see if performance improvements can be made. Particularly promising approaches would be di usion models, which have recently started to be applied in weather and climate prediction (Addison et al., 2022; Leinonen et al., 2023, e.g.), since they appear to perform well and are easier to train, and models trained directly on a suitable loss function such as the energy score (Pacchiardi et al., 2022).

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 L. Harris et al. (2022). All experiments in this paper were performed within TensorFlow 2.7.0, and some of the analysis in this work utilised 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 (Hu man 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 No 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 No. 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/.

References

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

Ageet, S., Fink, A. H., Maranan, M., Diem, J. E., Hartter, J., Ssali, A. L., & Ayabagabo, P. (2022, February). Validation of Satellite Rainfall Estimates over Equatorial East Africa. *Journal of Hydrometeorology*, *23*(2), 129{151. Retrieved 2023-09-21, from `https://journals.ametsoc.org/view/journals/hydr/23/2/JHM-D-21-0145.1.xml` (Publisher: American Meteorological Society Section: Journal of Hydrometeorology) doi: 10.1175/JHM-D-21-0145.1

Arjovsky, M., & Bottou, L. (2016, November). Towards Principled Methods for Training Generative Adversarial Networks. In *Proceedings of the 5th International Conference on Learning Representations.* Retrieved 2023-07-05, from `https://openreview.net/forum?id=Hk4_qw5xe`

Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 214{223). PMLR. Retrieved 2023-07-05, from `https://proceedings.mlr.press/v70/arjovsky17a.html` (ISSN: 2640-3498)

Bechtold, P., Chaboureau, J. P., Beljaars, A., Betts, A. K., Kohler, M., Miller, M., & Redelsperger, J. L. (2004, October). 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. doi: 10.1256/qj.03.103

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

Ben-Bouallegue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., & Chantry, M. (2023, August). *Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers.* arXiv. Retrieved 2023-08-30, from http://arxiv.org/abs/2303.17195 (arXiv:2303.17195) doi: 10.48550/arXiv.2303.17195

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022, November). *Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast.* Retrieved from http://arxiv.org/abs/2211.02556 (arXiv: 2211.02556)

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023, July). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533{538. Retrieved 2024-01-11, from https://www.nature.com/articles/s41586-023-06185-3 (Number: 7970 Publisher: Nature Publishing Group) doi: 10.1038/s41586-023-06185-3

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

Cafaro, C., Woodhams, B. J., Stein, T. H., Birch, C. E., Webster, S., Bain, C. L., . . . Hill, P. (2021, April). Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical east africa? *Weather and Forecasting*, *36*(2), 697{716. (Publisher: American Meteorological Society) doi: 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. Retrieved 2023-06-15, from https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.5197 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5197) doi: 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. (Publisher: John Wiley and Sons Ltd) doi: 10.1002/met.1403

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. Retrieved 2024-01-23, from https://onlinelibrary.wiley.com/doi/abs/10.1002/met.1538 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.1538) doi: 10.1002/met.1538

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018, January). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53{65. (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2017.2765202

Dezfuli, A. K., Ichoku, C. M., Hu man, G. J., Mohr, K. I., Selker, J. S., De Giesen, N. V., ... Annor, F. O. (2017). Validation of IMERG Precipitation in Africa. *Journal of Hydrometeorology*, *18*(10), 2817{2825. Retrieved from https://doi.org/10.1175/ doi: 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). Dordrecht: Springer Netherlands. Retrieved 2023-09-21, from https://doi.org/10.1007/978-90-481-2915-7_11 doi: 10.1007/978-90-481-2915-7_11

Duncan, J., Subramanian, S., & Harrington, P. (2022). *Generative Modeling of High-resolution Global Precipitation Forecasts.* (arXiv:2210.12504)

Deque, M. (2007, May). 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. Retrieved 2023-06-16, from https://www.sciencedirect.com/science/article/pii/S0921818106002748 doi: 10.1016/j.gloplacha.2006.11.030

Ebert, E. E. (2008). Fuzzy veri cation of high-resolution gridded forecasts: A review and proposed framework. In *Meteorological Applications* (Vol. 15, pp. 51{64). John Wiley and Sons Ltd. (Issue: 1 ISSN: 14698080) doi: 10.1002/met.25

ECMWF. (2023, February). *Changes to the forecasting system - Forecast User - ECMWF Confluence Wiki.* Retrieved 2023-11-02, from https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system (https://con uence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system, Accessed 2nd November 2023)

ECMWF. (2023, February). *Operational configurations of the ECMWF Integrated Forecasting System (IFS).* Retrieved 2023-01-09, from https://confluence.ecmwf.int/pages/viewpage.action?pageId=324860211 (https://con uence.ecmwf.int/pages/viewpage.action?pageId=324860211. Accessed 1st September)

ECMWF. (2023, February). *Parameter database.* Retrieved 2023-09-20, from https://codes.ecmwf.int/grib/param-db/ (https://codes.ecmwf.int/grib/param-db/, Accessed February 2023)

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Con- dence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54{75. Retrieved 2023-11-02, from https://www.jstor.org/stable/2245500 (Publisher: Institute of Mathematical Statistics)

Finney, D. L., Marsham, J. H., Jackson, L. S., Kendon, E. J., Rowell, D. P., Boorman, P. M., . . . Senior, C. A. (2019, April). Implications of Improved Representation of Convection for the East Africa Water Budget Using a Convection-Permitting Model. *Journal of Climate*, *32*(7), 2109{2129. Retrieved 2023-07-27, from https://journals.ametsoc.org/view/journals/clim/32/7/jcli-d-18-0387.1.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI-D-18-0387.1

Floodlist. (2023, May). *Somalia – Flooding Displaces Thousands, Prompts Urgent Humanitarian Response.* Retrieved 2023-08-04, from https://floodlist.com/africa/somalia-floods-may-2023 (https:// oodlist.com/africa/somalia- oods-may-2023. Accessed May 2023.)

Gebremeskel Haile, G., Tang, Q., Sun, S., Huang, Z., Zhang, X., & Liu, X. (2019, June). Droughts in East Africa: Causes, impacts and resilience. *Earth-Science Reviews*, *193*, 146{161. Retrieved 2023-08-03, from https://

www.sciencedirect.com/science/article/pii/S0012825218303519 doi: 10.1016/j.earscirev.2019.04.015

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

Gneiting, T., & Raftery, A. E. (2007, March). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359{378. doi: 10.1198/016214506000001437

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., . . . Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. Retrieved 2023-07-05, 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. *Hydrol. Earth Syst. Sci. Discuss*, *9*, 6185{6201. Retrieved from www.hydrol-earth-syst-sci-discuss.net/9/6185/2012/ doi: 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). Curran Associates, Inc. Retrieved 2023-07-05, 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, August). Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Monthly Weather Review*, *140*(8), 2720{2733. doi: 10.1175/MWR-D-11-00301.1

Hamill, T. M. (2001, March). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 550{560. Retrieved 2023-09-01, from https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml (Publisher:

American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2

Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., & Levit, J. J. (2001, August). Multiscale Statistical Properties of a High-Resolution Precipitation Forecast. *Journal of Hydrometeorology*, *2*(4), 406–418. Retrieved 2023-08-03, from `https://journals.ametsoc.org/view/journals/hydr/2/4/1525-7541_2001_002_0406_mspoah_2_0_co_2.xml` (Publisher: American Meteorological Society Section: Journal of Hydrometeorology) doi: 10.1175/1525-7541(2001)002⟨0406:MSPOAH⟩2.0.CO;2

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022, April). A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10).

Hu man, G., Bolvin, D., Braithwaite, D., Hsu, K., Joyce, R., & Xie, P. (2022, October). Integrated multi-satellite retrievals for GPM (IMERG), V06B. [dataset]. *NASA's Precipitation Processing Center, accessed 1st October 2022, https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmallversions/V06/YYYY/MM/DD/imerg/*.

IFRC. (2014). *World Disasters Report, Focus on culture and risk* (Tech. Rep.). International Federation of Red Cross and Red Crescent Societies. Retrieved 2023-07-27, from `https://www.ifrc.org/document/world-disasters-report-2014`

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

Karras, T., Laine, S., & Aila, T. (2019, June). A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Kilavi, M., MacLeod, D., Ambani, M., Robbins, J., Dankers, R., Graham, R., . . . Todd, M. C. (2018, November). Extreme rainfall and ooding over Central Kenya Including Nairobi City during the long-rains season 2018: Causes, predictability, and potential for early warning and actions. *Atmosphere*, *9*(12).

(Publisher: MDPI AG) doi: 10.3390/atmos9120472

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

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., . . . Battaglia, P. (2023, November). Learning skillful medium-range global weather forecasting. *Science*, *0*(0), eadi2336. Retrieved 2023-11-22, from https://www.science.org/doi/10.1126/science.adi2336 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.adi2336

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

Leinonen, J., Nerini, D., & Berne, A. (2020, November). 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. (arXiv: 2005.10374 Publisher: Institute of Electrical and Electronics Engineers (IEEE)) doi: 10.1109/tgrs.2020.3032790

Leutbecher, M., & Palmer, T. N. (2008, March). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515{3539. (Publisher: Academic Press Inc.) doi: 10.1016/j.jcp.2007.02.014

MacLeod, D., Kilavi, M., Mwangi, E., Ambani, M., Osunga, M., Robbins, J., . . . Todd, M. C. (2021, January). Are Kenya Meteorological Department heavy rainfall advisories useful for forecast-based early action and early preparedness for ooding? *Natural Hazards and Earth System Sciences*, *21*(1), 261{277. Retrieved 2023-06-15, from https://nhess.copernicus.org/articles/21/261/2021/ (Publisher: Copernicus GmbH) doi: 10.5194/nhess-21-261-2021

Macleod, D. A., Dankers, R., Graham, R., Guigma, K., Jenkins, L., Todd, M. C., . . . Mwangi, E. (2021). Drivers and Subseasonal Predictability of Heavy Rainfall in Equatorial East Africa and Relationship with Flood Risk. *Journal of Hydrometeorology*, *22*(4), 887{903. Retrieved from https://doi.org/10.1175/

JHM-D-20- doi: 10.1175/JHM-D-20

Manzato, A., & Jolli e, I. (2017, April). Behaviour of veri cation measures for deterministic binary forecasts with respect to random changes and thresholding. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1903{1915. (Publisher: John Wiley and Sons Ltd) doi: 10.1002/qj.3050

Maraun, D., & Widmann, M. (2017, December). Model Output Statistics. In *Statistical Downscaling and Bias Correction for Climate Research* (pp. 170{200). Cambridge University Press. doi: 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, May). 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. doi: 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. (Publisher: John Wiley and Sons Ltd) doi: 10.1002/met.296

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

Nicholson, S. (2016). The Turkana low-level jet: mean climatology and association with regional aridity. *International Journal of Climatology*, *36*(6), 2598{2614. Retrieved 2023-08-08, from https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.4515 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4515) doi: 10.1002/joc.4515

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. *Reviews of Geophysics*, *55*(3), 590{635.

Pacchiardi, L., Adewoyin, R., Dueben, P., & Dutta, R. (2022, May). *Probabilistic Forecasting with Generative Networks via Scoring Rule Minimization.* arXiv. Retrieved 2023-07-20, from http://arxiv.org/abs/2112.08217 (arXiv:2112.08217 [cs, stat])

Palmer, P. I., Wainwright, C. M., Dong, B., Maidment, R. I., Wheeler, K. G., Gedney, N., . . . Turner, A. G. (2023, April). Drivers and impacts of Eastern African rainfall variability. *Nature Reviews Earth & Environment*, *4*(4), 254{270. Retrieved 2023-09-01, from `https://www.nature.com/articles/s43017-023-00397-x` (Number: 4 Publisher: Nature Publishing Group) doi: 10.1038/s43017-023-00397-x

Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *151*. doi: https://doi.org/10.48550/arXiv.2203.12297

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019, October). Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0) [software]. *Geoscientific Model Development*, *12*(10), 4185{4219. doi: 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. Retrieved from `https://doi.org/10.1175/MWR-D-18-` doi: 10.1175/MWR-D-18

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., . . . Mohamed, S. (2021, September). Skilful precipitation nowcasting using deep generative models of radar: Supplementary Information. *Nature*, *597*(7878), 672{677. Retrieved 2023-07-06, from `https://www.nature.com/articles/s41586-021-03854-z` doi: 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. In *Meteorological Applications* (Vol. 15, pp. 163{169). John Wiley and Sons Ltd. (Issue: 1 ISSN: 14698080) doi: 10.1002/met.57

Roberts, N. M., & Lean, H. W. (2008, January). Scale-selective veri cation of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78{97. doi: 10.1175/2007MWR2123.1

Roca, R., Chambon, P., Jobard, I., Kirstetter, P. E., Gosset, M., & Bergès, 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. (Publisher: American Meteorological Society) doi: 10.1175/2009JAMC2318.1

Schaefer, J. T. (1990, December). The Critical Success Index as an Indicator of Warning Skill. *Weather and Forecasting*, *5*(4), 570{575. Retrieved 2023-07-03, from https://journals.ametsoc.org/view/journals/wefo/5/4/1520-0434_1990_005_0570_tcsiaa_2_0_co_2.xml (Publisher: American Meteorological Society Section: Weather and Forecasting) doi: 10.1175/1520-0434(1990)005h0570:TCSIAAi2.0.CO;2

Sinclair, S., & Pegram, G. G. S. (2005, July). 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. Retrieved 2023-08-03, from https://hess.copernicus.org/articles/9/127/2005/ (Publisher: Copernicus GmbH) doi: 10.5194/hess-9-127-2005

Vogel, P., Knippertz, P., Fink, A. H., & Schlueter, A. (2018). Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa. *Weather and Forecasting*, *33*(2), 369{388. Retrieved from https://doi.org/10.1175/WAF-D-17- doi: 10.1175/WAF-D-17

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2020, December). Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall in the Tropics. *Weather and Forecasting*, *35*(6), 2367{2385. Retrieved 2024-01-19, from https://journals.ametsoc.org/view/journals/wefo/35/6/WAF-D-20-0082.1.xml (Publisher: American Meteorological Society Section: Weather and Forecasting) doi: 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. Retrieved 2023-10-27, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2022JD038163 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022JD038163) doi: 10.1029/2022JD038163

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

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

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. Retrieved 2023-11-02, from https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4429 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4429) doi: 10.1002/qj.4429

Watkiss, P., & Cimato, F. (2021, September). *Socio-Economic Benefits of the WISER Programme. Synthesis of Results.* (Tech. Rep.). Met Of- ce 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, September). *The Socio-Economic Benefits of the HIGHWAY project* (Tech. Rep.). 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. Retrieved 2023-09-14, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001597 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001597) doi: 10.1029/2018MS001597

Watson, P. A. G. (2022, November). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, *17*(11), 111004. Retrieved 2023-08-04, from https://dx.doi.org/10.1088/1748-9326/ac9d4e (Publisher: IOP Publishing) doi: 10.1088/1748-9326/ac9d4e

Wilks, D. S. (2019a). Forecast Veri cation. In *Statistical Methods in the Atmospheric Sciences* (pp. 369{483). Elsevier. doi: 10.1016/b978-0-12-815823-4.00009-2

Wilks, D. S. (2019b). Statistical Forecasting. In *Statistical Methods in the At-*

mospheric Sciences (pp. 235{312). Elsevier. doi: 10.1016/b978-0-12-815823-4.00007-9

Woodhams, B. J., Birch, C. E., Marsham, J. H., Bain, C. L., Roberts, N. M., & Boyd, D. F. (2018, September). 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. (Publisher: American Meteorological Society) doi: 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. (Publisher: American Meteorological Society) doi: 10.1175/MWR-D-19-0069.1

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., . . . Jiang, Y.-G. (2023). A survey on video di usion models. *arXiv preprint arXiv:2310.10647*.

Youds, L. H., Parker, D. J., Ade san, E. A., Antwi-Agyei, P., Bain, C. L., Black, E. C. L., . . . Woolnough, S. J. (2021, November). *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* [Monograph]. Retrieved 2023-07-04, from `https://eprints.whiterose.ac.uk/181045/` (Publisher: University of Leeds)

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023, July). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 1{7. Retrieved 2023-07-14, from `https://www.nature.com/articles/s41586-023-06184-4` (Publisher: Nature Publishing Group) doi: 10.1038/s41586-023-06184-4

Appendix A Forecast variables used

The IFS variables and constant elds used to train the model are shown in Table A1, de nitions taken from (ECMWF, 2023).

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

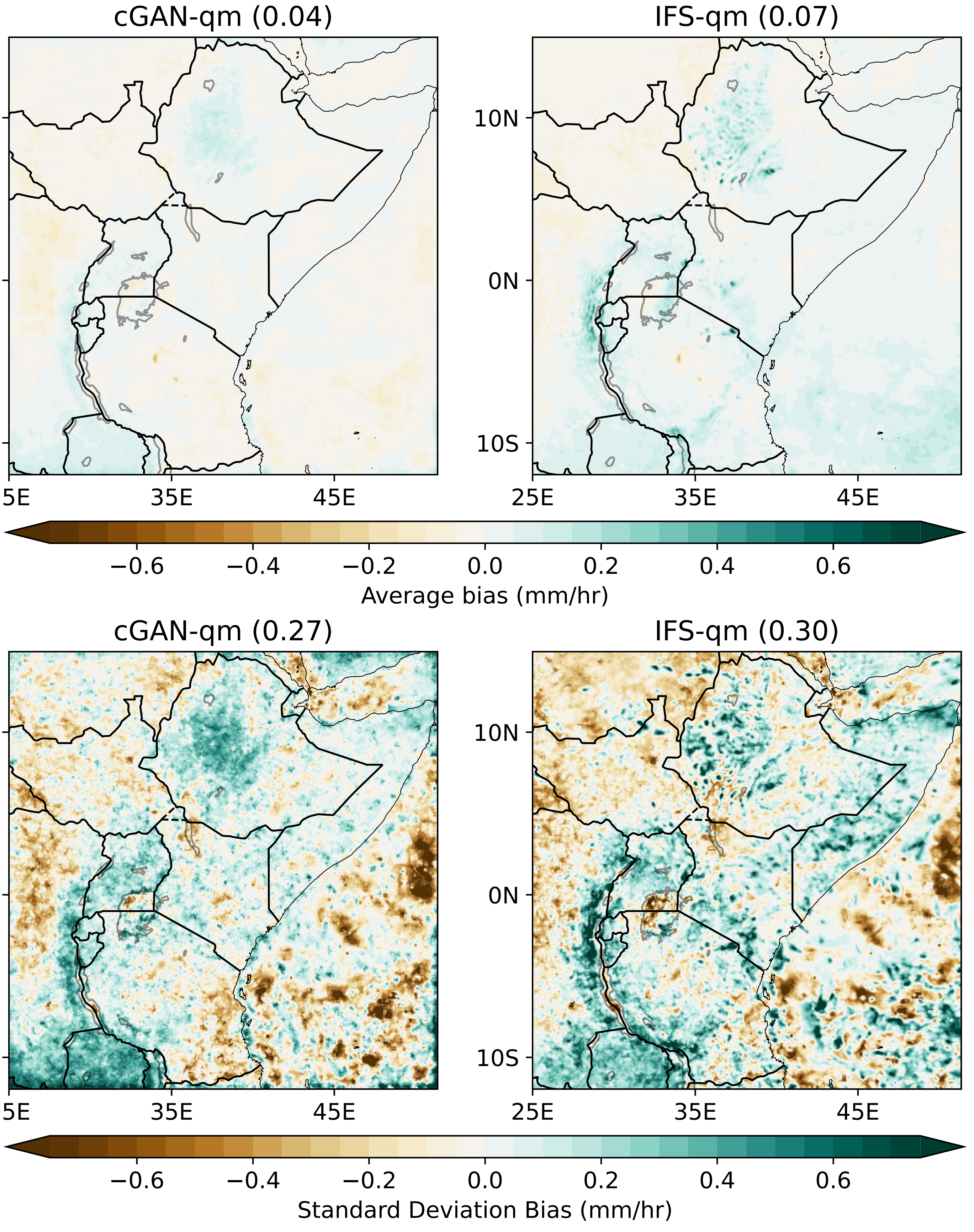
- Minmax: calculate the minimum d_{min} and maximum d_{max} over the reference period, and then transform each value v according to $(v - d_{\text{min}})/(d_{\text{max}} - d_{\text{min}})$.

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 $\log_{10}(1 + v)$.

Variable name	Symbol	Pre-processing applied
2 metre 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 vapour	tcwv	Log
Top of atmosphere incident solar radiation	tisr	Max
Total precipitation	tp	Log
Relative humidity at 200hPa	r200	Max
Relative humidity at 700hPa	r700	Max
Relative humidity at 950hPa	r950	Max
Temperature at 200hPa	t200	Minmax
Temperature at 700hPa	t700	Minmax
Eastward component of wind at 200hPa	u200	Max
Eastward component of wind at 700hPa	u700	Max
Northward component of wind at 200hPa	v200	Max
Northward component of wind at 700hPa	v700	Max
Vertical velocity at 200hPa	w200	Max
Vertical velocity at 500hPa	w500	Max
Vertical velocity at 700hPa	w700	Max
Orography	h	Max
Land-sea mask	lsm	N/A

Table A1: IFS variables used to train the model, as well as the normalisation applied to each variable. See text for description of the di erent preprocessing types.


cGAN-qm (0.04)
IFS-qm (0.07)
10N
0N
10S
25E
35E
45E
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Average bias (mm/hr)
cGAN-qm (0.27)
IFS-qm (0.30)
Standard Deviation Bias (mm/hr)