



Predictability of Rainfall over Equatorial East Africa in the ECMWF Ensemble Reforecasts on Short- to Medium-Range Time Scales

SIMON AGEET,^{a,b} ANDREAS H. FINK,^a MARLON MARANAN,^a AND BENEDIKT SCHULZ^c

^a *Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany*

^b *Uganda National Meteorological Authority, Kampala, Uganda*

^c *Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany*

(Manuscript received 31 May 2023, in final form 10 October 2023, accepted 11 October 2023)

ABSTRACT: Despite the enormous potential of precipitation forecasts to save lives and property in Africa, low skill has limited their uptake. To assess the skill and improve the performance of the forecast, validation and postprocessing should continuously be carried out. Here, we evaluate the quality of reforecasts from the European Centre for Medium-Range Weather Forecasts over equatorial East Africa (EEA) against satellite and rain gauge observations for the period 2001–18. The 24-h rainfall accumulations are analyzed from short- to medium-range time scales. Additionally, 48- and 120-h rainfall accumulations were also assessed. The skill was assessed using an extended probabilistic climatology (EPC) derived from the observations. Results show that the reforecasts overestimate rainfall, especially during the rainy seasons and over high-altitude areas. However, there is potential of skill in the raw forecasts up to 14-day lead time. There is an improvement of up to 30% in the Brier score/continuous ranked probability score relative to EPC in most areas, especially the higher-altitude regions, decreasing with lead time. Aggregating the reforecasts enhances the skill further, likely due to a reduction in timing mismatches. However, for some regions of the study domain, the predictive performance is worse than EPC, mainly due to biases. Postprocessing the reforecasts using isotonic distributional regression considerably improves skill, increasing the number of grid points with positive Brier skill score (continuous ranked probability skill score) by an average of 81% (91%) for lead times 1–14 days ahead. Overall, the study highlights the potential of the reforecasts, the spatiotemporal variation in skill, and the benefit of postprocessing in EEA.


KEYWORDS: Africa; Rainfall; Forecast verification/skill; Hindcasts; Postprocessing


1. Introduction

Equatorial East Africa (EEA) is among the regions of the world most vulnerable to weather- and climate-related extremes, mainly in the form of floods and drought. The IPCC (2022) estimates the regions' death rate due to these disasters to be 15-fold more than that of the less vulnerable regions of the world. Currently, many countries in EEA are experiencing the longest and most severe drought leaving about 70 million people at risk of starvation and death (Toreti et al. 2022). On the other hand, most flood-prone regions of EEA have seen an

increase in death and displacement of people due to floods and landslides. OCHA (2020) estimates that between December 2019 and January 2020 alone, about 3.4 million people were affected in the region. The impact of the disasters can, and should, be mitigated through issuing early warnings informed by accurate weather and climate forecasts. Unfortunately, the potential of these forecasts to save lives and property has not yet been realized for EEA and Africa at large (Youds et al. 2021).

The major reason for the unrealized potential could be the forecasts' lack of skill in most of sub-Saharan Africa. Vogel et al. (2018) demonstrated that over sub-Saharan Africa, forecasts from nine global forecasting centers, including the European Centre for Medium-Range Weather Forecasts (ECMWF), hardly beat climatology, even after postprocessing. Haiden et al. (2012) showed that a forecast with 1-day lead time in the tropics is equivalent to a 6-day lead-time forecast in the extratropics. On the bright side, with continued research and increasing computational resources, forecast models now better represent climate drivers such as the Madden–Julian oscillation (MJO), El Niño–Southern Oscillation (ENSO), and Indian Ocean dipole (IOD), leading to improved predictability in the

 Denotes content that is immediately available upon publication as open access.

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-23-0093.s1>.

Corresponding author: Simon Ageet, simon.ageet@kit.edu

DOI: 10.1175/WAF-D-23-0093.1

© 2023 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by KARLSRUHE INSTITUTE F. TECHNOL. | Unauthenticated | Downloaded 12/14/23 07:54 PM UTC

region, especially at subseasonal time scale (Li and Robertson 2015; Vitart et al. 2017; de Andrade et al. 2021). However, these gains are not homogeneous over the globe due to a number of reasons, e.g., difference in geography, climate, and quality of initial conditions. Therefore, there is a need to validate forecasts to ascertain their performance and to calibrate to correct for miscalibration if any.

A number of studies over EEA have assessed the quality of rainfall forecasts from different global forecasting centers. At subdaily and daily time scales, Woodhams et al. (2018) and Cafaro et al. (2021) showed that a convection-permitting (CP) version of the Met Office (UKMO) model performed better than the global model over East Africa, especially for extreme events. However, they also noted that the CP model often overestimates the extreme rainfall compared to observations. Stellingwerf et al. (2021) showed that the ECMWF model performed best when compared to models from other centers. At the subseasonal time scale, de Andrade et al. (2021) evaluated reforecasts from three global forecasting centers over Africa and found that rainfall predictions, accumulated over 7 days, were more skillful over East Africa compared to those of other regions in Africa and that ECMWF reforecasts performed best compared to those from the UKMO and the National Centers for Environmental Prediction (NCEP). Based on weekly precipitation accumulation over EEA, Macleod et al. (2021) showed that reforecasts from ECMWF and UKMO are skillful even after the second week. Endris et al. (2021), using monthly and weekly accumulations, support the superior performance of ECMWF forecasts over the East African region compared to other forecasts from the different centers in the Subseasonal-to-Seasonal (S2S) prediction project database (Vitart et al. 2017). There is also a general consensus that aggregating rainfall over several days gives higher skill than 1-day rainfall accumulation (Vogel et al. 2020; Stellingwerf et al. 2021). The higher skill in EEA compared to other regions has been attributed to a better representation of the teleconnections between the rainfall and climate drivers such as ENSO, IOD, and MJO (de Andrade et al. 2021; Macleod et al. 2021; Endris et al. 2021).

Although there is promising skill in the forecasts over EEA, the studies also showed limitations in the models leading to uncalibrated and biased forecasts. Macleod et al. (2021) noted that using these forecasts may lead to wrong actions almost 50% of the time and that the model could not capture the magnitude of wet precipitation events. The shortcomings of models in representing the magnitudes of convective rainfall over the tropics were also reported in Vogel et al. (2020). These deficiencies have been mainly attributed to inadequate parameterization of convection (Vogel et al. 2018), and the poor quality or lack of observations going into the model (Zagar 2017). One common way to deal with the deficiencies in the forecasts is by postprocessing to correct for systematic errors such as biases and miscalibration. Using the ensemble model output statistics (EMOS; Gneiting et al. 2005) approach, Vogel et al. (2020) showed that for ECMWF forecasts over tropical Africa, postprocessing increased the number of grid cells, where the forecasts performed better than a climatological benchmark in terms of probabilistic evaluation

metrics by 50% and 78%, for rainfall occurrence and amounts, respectively. Similar improvements were seen by Stellingwerf et al. (2021) who showed that bias correction using quantile-to-quantile mapping improved the Brier skill score (BSS) of forecasts over Ethiopia. Even by taking simple observed regression patterns into account, as done in de Andrade et al. (2021), an improvement in the forecast quality was obtained.

Most of the validation studies listed above either considered longer temporal aggregations (de Andrade et al. 2021; Endris et al. 2021), did not postprocess the forecasts (Endris et al. 2021), or used different postprocessing approaches (Stellingwerf et al. 2021). Moreover, the studies, with the exception of Vogel et al. (2018, 2020), use the classical reference forecast in the form of observation climatology to assess skill. This study focuses on Uganda and the adjacent regions including Congo where Vogel et al. (2020) showed that rainfall predictability at 1- and 5-day accumulation periods is one of the poorest in the entire tropics. We intend to add to the growing body of knowledge in EEA by assessing the predictability of rainfall in both raw and postprocessed versions of the ECMWF reforecast. With the postprocessing, we test a new generic method, which to the best of our knowledge has not been applied in our study domain. In addition to satellite rainfall estimates, we leverage gauge observations that are not publicly accessible to analyze the performance of the reforecasts. The analysis is done for different domains with different terrain characteristics within EEA, from 1- to 14-day lead times and for 24-, 48-, and 120-h accumulations. We also use a slightly different reference forecast in the form of the extended probabilistic climatology (EPC; Vogel et al. 2018; Walz et al. 2021) generated from the observation.

The study is structured as follows. Section 2 gives a brief description of the datasets and explains the methods used. We also explain the method used to postprocess the raw reforecasts. In section 3, we present our findings starting with results based on satellite observations followed by gauge observations. Finally, the discussion and conclusions are presented in section 4.

2. Data and methods

a. Data

1) ECMWF RAINFALL REFORECASTS

The study uses precipitation reforecasts from the ECMWF model available from the S2S prediction project database (Vitart et al. 2017). The reforecasts are generated on the fly with respect to the near-real-time forecasts, initialized twice a week, have a lead time of 46 days, and have a spatial (temporal) resolution of 1.5° (6 h). The reforecasts used here were generated using the model version dates from 2020, which were based on CY46R1 and CY47R1 cycles of the Integrated Forecasting System (IFS) model. For the 20-yr period, 2000–19, we have 105 initialization dates per year resulting from two forecasts in each of the 52.5 weeks. The reforecast is made up of 11 ensemble members (1 control and 10 perturbed).

2) SATELLITE RAINFALL ESTIMATES

The first set of precipitation observations used for the verification of the reforecasts are satellite rainfall estimates, specifically, the final daily product of Integrated Multi-satellite Retrieval for Global Precipitation Measurement (GPM) v6B (IMERG; Huffman et al. 2020). Being a satellite-based product, IMERG was chosen because it offers a more complete spatiotemporal coverage for the data sparse regions like EEA (Diem et al. 2014; Dinku 2019) compared to ground-based observation, e.g., rain gauges or radars. Moreover, it has been shown to be among the best-performing products for this region at daily to monthly time scales (Ageet et al. 2022). It should also be noted that IMERG is not a perfect dataset and has shortcomings, some of which are pointed out in the results and discussion sections. IMERG is available from June 2000 to date at a temporal resolution of 30 min and a spatial resolution of 0.1° . Since the reforecasts are available at a spatial resolution of 1.5° , IMERG rainfall estimates are regridded to the same resolution using first-order conservative remapping (Jones 1999).

3) RAIN GAUGE RAINFALL

In situ rainfall observations from rain gauges in Uganda available from the Karlsruhe African Surface Station-Database (KASS-D; Vogel et al. 2018) were used. Because of the relatively coarse resolution of the gridded products, the analysis with gauges was done at nearest grid points, with a requirement that the grid point is nearest to at least four stations. Additionally, the stations considered should have at least 95% daily data availability in the period of study. Three grid points satisfied all the above conditions. Due to their regional location, we named these grid points Lake Victoria, Lake Kyoga, and western Uganda regions (Fig. 1). The period 2001–18 offered the most complete record and was therefore used for the analysis. The mean of the stations at a grid point was used for the validation of the forecasts. Because the daily gauge rainfall in this region is accumulated from 0600 to 0600 UTC of the following day, the reforecast and IMERG rainfall were aggregated in the same manner. All the analyses were done for the common period of 2001–18.

b. Methods

1) VERIFICATION METHODS

Forecast verification primarily assesses how “well” a forecasting system predicts the target variable based on the observed values. Because the goodness of a forecast may depend on more than one attribute and the purpose of the verification (Murphy 1993), a single verification method is not sufficient to assess the predictive performance of the forecast. Hence, several methods should be taken into account. To assess different aspects of the ECMWF reforecasts, we assess the ensemble in both probabilistic and deterministic terms, the latter in the form of the ensemble median. While the ensemble is used to assess the predictive performance of the reforecasts in terms of scores, calibration, and discrimination, the ensemble median is used to assess the accuracy and bias of a point forecast

derived from the ensemble, which we refer to as a deterministic forecast. The association of the deterministic forecasts and the observations is also assessed using Pearson’s correlation coefficient r . Further, we derive probability forecasts for the occurrence of rainfall from the ensemble by calculating the fraction of ensemble members that predicts rainfall. The different metrics are briefly explained below.

For the deterministic forecast, the mean error (ME) is computed as the mean difference between the ensemble median of the reforecasts and the observations, with positive (negative) values indicating overestimation (underestimation). Additionally, the mean absolute error (MAE) is computed for different rainy day thresholds (i.e., when the observations have a precipitation accumulation of more than a given threshold value), in this case from 0 up to 10 mm, to show how the error changes for higher rain rates. The correlation coefficient describes the linear relationship between the deterministic forecasts and observation, where perfect positive (negative) association is indicated by +1 (−1). If f_i and o_i are the point forecast and observation at time i , respectively, \bar{o} and \bar{f} are their means, and N is the sample size, then the metrics above are defined as follows:

$$\text{ME} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i), \quad (1)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |f_i - o_i|, \quad (2)$$

$$r = \frac{\sum_{i=1}^N (o_i - \bar{o})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^N (o_i - \bar{o})^2} \sqrt{\sum_{i=1}^N (f_i - \bar{f})^2}}. \quad (3)$$

To assess the discrimination ability or the potential prediction ability of probability forecasts for the occurrence of rainfall, the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are suitable tools (Wilks 2011). The ROC curve is generated by plotting the hit rate against the false alarm rate at different probability thresholds. The reforecasts have no discrimination ability if the curve falls on the diagonal, and a perfect discrimination is obtained if the curve passes at the top-left corner. The area between the diagonal and the curve gives the AUC, with values between 0.5 (no discrimination) and 1 (perfect discrimination). By comparing to a reference, EPC (EPC15, hereinafter, as observations in a ± 15 -day window around the date of interest are considered), we compute the so-called AUC skill score (AUCS) like in Walz et al. (2021). For every forecast date, we generated the EPC15 by taking past observations on this date and the 30 days around it, yielding an ensemble of 527 members for the training dataset, i.e., 31 members times 17 years (1 year less due to cross validation). Details and code to compute the EPC are given in Walz et al. (2021).

Most probabilistic forecasts do not quantify the forecast uncertainty adequately, meaning they are miscalibrated or unreliable (Wilks 2011). To check the reforecasts for calibration,

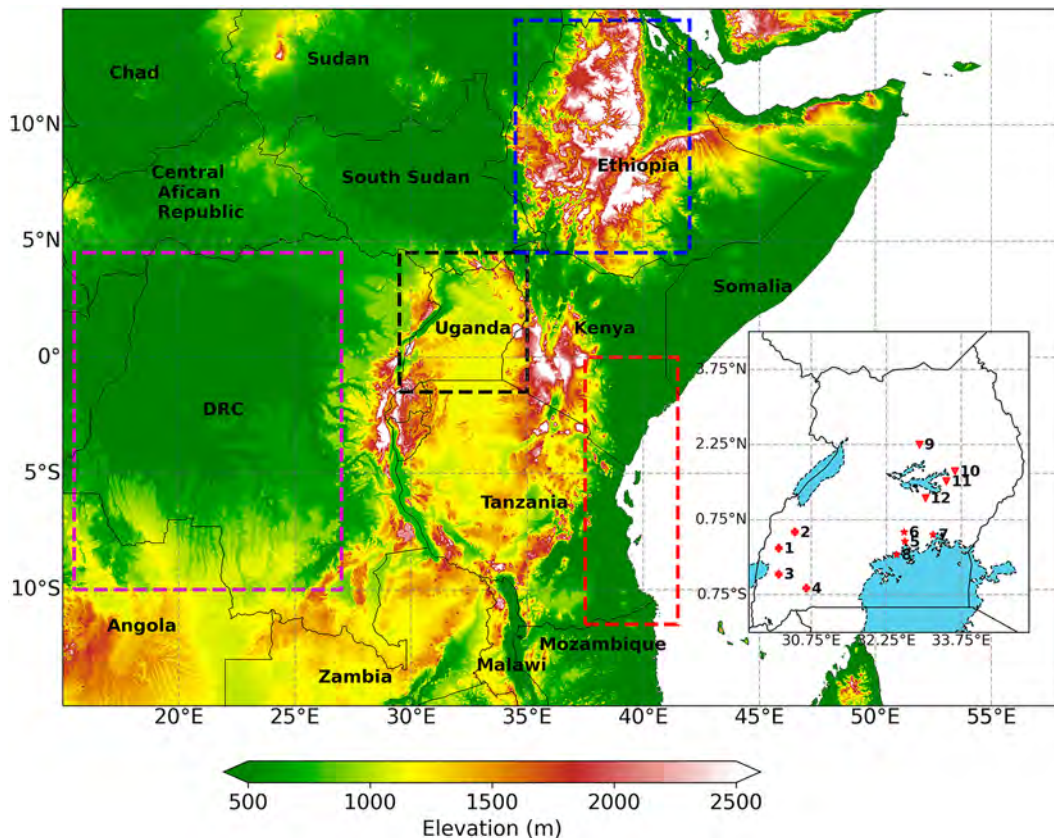


FIG. 1. Study domain with four regions: Uganda (black), Congo basin (magenta), Ethiopian highlands (blue), and East African coast (red). The shading shows the terrain elevation as given by the Global Land 1-km Base Elevation project (GLOBE; Hastings et al. 1999). The map inset has the location of the 12 synoptic stations, grouped into four clusters depending on the nearest 1.5° grid they fall in. These stations are representative of western Uganda regions (1–4), Lake Victoria region (5–8), and Lake Kyoga region (9–12). The station names are listed in Table S1.

we use standard tools from forecasting methodology (Gneiting and Katzfuss 2014). Rank histograms are used to assess the calibration of the ensemble forecasts, and probability integral transform (PIT) histograms are used for the postprocessed forecasts and EPC15. Note that the PIT histograms for EPC15 are calculated as described in Vogel et al. (2018, 2020) and Schulz and Lerch (2022). Both rank and PIT histograms can be interpreted analogously, where a flat histogram corresponding to a uniform distribution indicates that the forecasts are calibrated, while a U-shaped (hump-shaped) histogram indicates underdispersed (overdispersed) reforecasts; that is, the forecasts are overconfident (underconfident). The calibration of probability forecasts is checked via reliability diagrams, which show the calibration curve that plots the conditional event probability of the dichotomous event against the associated forecast probabilities. If the curve is close to the diagonal, the forecast is said to be calibrated or reliable. Here, we generate the reliability diagrams using the CORP approach, which ensures consistency, optimality, and reproducibility and is based on the pool-adjacent-violators algorithm (Dimitriadis et al. 2021). A major advantage of this approach is that it generates “optimally binned, reproducible,

and statistically consistent reliability diagrams” (Dimitriadis et al. 2021).

A quantitative evaluation of the forecast performance is done using proper scoring rules, which yield the best scores (in expectation) when we forecast the “true” underlying distribution of the observation (Gneiting and Raftery 2007; Wilks 2011). For assessing probability forecasts of rainfall occurrence, the most common metric is the Brier score (BS; Brier 1950), while for the rainfall amounts, the continuous ranked probability score (CRPS; Gneiting and Raftery 2007) is used. The mean BS for a sample of size N is defined as

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2, \quad (4)$$

where p_i is the probability forecast at time i and y_i is the corresponding observation, which takes the value 1 for rainfall occurrence and 0 otherwise. If F_i is the cumulative distribution function (CDF) of a precipitation forecast at time i , and o_i is the corresponding observation, the mean CRPS of a sample of size N is defined as

$$\text{CRPS} = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{\infty} [F_i(x) - \mathbb{1}\{x \geq o_i\}]^2 dx. \quad (5)$$

Both scores are negatively oriented meaning that smaller values indicate superior predictive performance. The CRPS is in the unit of the observation, mm, in this case. For scores of a method, the corresponding skill scores, i.e., BSS, and continuous ranked probability skill score (CRPSS) were computed relative to EPC15. Negative skill scores indicate that the method performs worse than the reference forecast, a skill score of 0 indicates equal performance, and a positive skill score indicates that the method outperforms the reference with 1 corresponding to perfectly forecasting the observed values. For a detailed discussion of these, and the other metrics used in this manuscript, we refer the reader to [Schulz and Lerch \(2022\)](#).

To investigate the sources of the strengths and weaknesses of the forecast, we use the CORP approach to decompose the BS in a miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) component. MCB quantifies the degree of miscalibration of the forecasts (smaller values are preferred), while DSC evaluates the ability to discern between events and nonevents (larger values are better), and finally, UNC, which is dependent only on the observations, quantifies the inherent uncertainty in the underlying forecasting problem (larger values indicate more uncertainty). To check if any observed differences in performance between the reforecasts and EPC15 are significant, we applied the (two-sided) Diebold–Mariano test (DM test; [Diebold and Mariano 1995](#)) to forecast EPC15 score pairs (BS or CRPS) at each grid point. The DM test checks the hypothesis whether the (raw or postprocessed) reforecasts and EPC15 have the same expected score, that is, equal predictive performance. Because we are testing multiple grid points independently, there is a need to account for the possibility of the false discovery rate ([Wilks 2016](#)). We therefore applied a [Benjamini and Hochberg \(1995\)](#) procedure which controls the proportion of falsely rejected hypothesis at a chosen significance level, 0.05 in this case.

2) POSTPROCESSING

To remove or reduce systematic errors of the ensemble forecasts such as biases and dispersion errors, methods from statistical postprocessing are typically used to correct for them (see, e.g., [Gneiting et al. 2005](#); [Vannitsem et al. 2018](#); [Schulz and Lerch 2022](#)). Here, we apply the nonparametric isotonic distributional regression (IDR; [Henzi et al. 2021](#)) method. The method assumes an isotonic relationship (i.e., an increase of the predictor variable yields an increase of the predictand) between the forecasts and observations. In this case, the ensemble members were used as the predictors, following the component-wise partial order. The advantages of this approach over other postprocessing methods such as EMOS are that 1) it is a generic method that can be applied directly as it does not require any prior conditioning or tuning and 2) it estimates a flexible data-driven forecast distribution based only on the assumption of isotonicity. The method has also been

shown to work well in other postprocessing applications (e.g., [Maier-Gerber et al. 2021](#); [Schulz and Lerch 2022](#)).

For the postprocessing, the data were divided into a training set and a test set. As noted by [Henzi et al. \(2021\)](#), IDR, being a nonparametric method, requires quite a large training period for the model to sufficiently learn the forecast–observation relationship. We therefore divided the data into 17 years of training and 1 year for testing. Although we use a local approach, whereby we train and apply the model for each grid point separately, data from eight grid points surrounding the grid point of interest are incorporated for training the model. This increased the size of the training dataset and is reasonable, given the similarity in rainfall climatology of neighboring grid points.

3) SPATIOTEMPORAL CONSIDERATIONS

Because the performance of the forecast varies with lead time and is influenced by the underlying topography, we assess the forecasts at different temporal aggregations and in four different locations in the region (cf. [Fig. 1](#)). The different regions are 1) Uganda, characterized by a mixture of mountains, large water bodies, and flat land, 2) Congo basin, a vast area of mainly low-lying and forested region, 3) East African coast, a coastal region along the shores of the Indian Ocean, and 4) Ethiopian highlands. For the temporal aggregations, we consider 24-, 48-, and 120-h accumulations and increasing lead time, that is, 1–14 days ahead. The longer aggregations are important to certain economic sectors. For example, to farmers, the exact timing of rain may not be so crucial, but how much rain falls in a particular period is more important. A seasonal perspective was also analyzed given that the region has distinct dry and wet seasons.

3. Results

In the first part of the results section, we present the analysis based on IMERG observations, starting with the deterministic and then probabilistic verification. The analysis against gauges will be presented at the end of the section.

a. Deterministic verification

As shown in [Fig. 2](#), the study domain has four distinct seasons: December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON). The mean seasonal daily rainfall intensity based on IMERG (dotted-blue contours) shows the rainfall maximum in the south in DJF, in the north in JJA, and in the vicinity of the equator in MAM and SON. The highest mean daily rainfall is 8 mm day^{−1} over Lake Victoria in MAM, in the southwestern part of the domain in DJF, and over the Ethiopian highlands in JJA. During each season, some land grid points, especially in the northern part of the domain, are dry, defined here as any grid point with a seasonal mean daily rainfall of less than 1 mm day^{−1}. These grid points are masked out in [Fig. 2](#) and were excluded in the subsequent analysis. The reforecasts are biased, with overestimation of rainfall over mountainous regions and underestimation in low-lying regions. In JJA when the seasonal

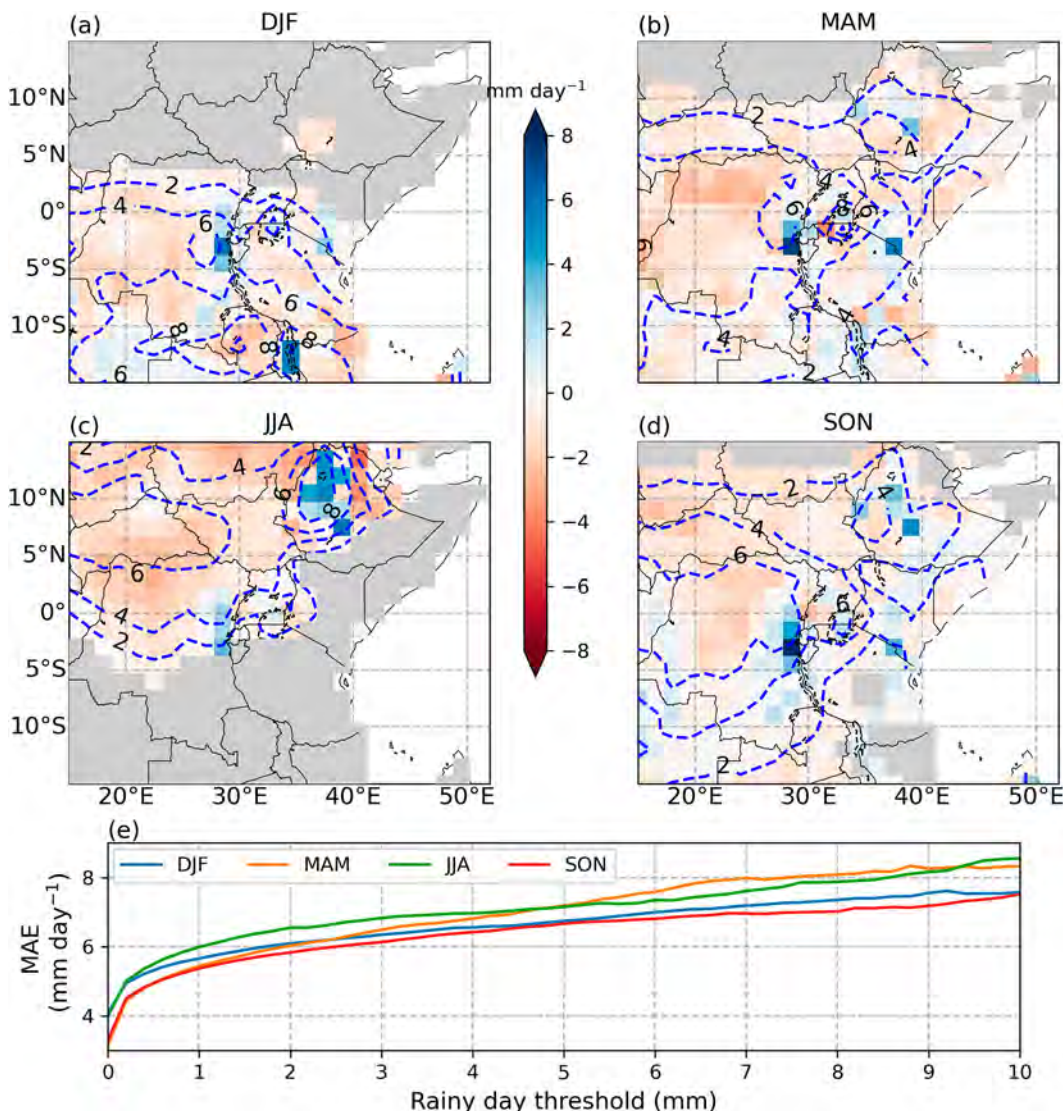


FIG. 2. (a)–(d) Seasonal daily rainfall mean error of the reforecasts relative to IMERG, computed by considering only days in the respective seasons, for the period 2001–18. The mean error is computed for a lead time of 1 day, and only for hits, i.e., days when both the reforecasts and IMERG recorded 0.2 mm or more of rainfall. The dotted blue contours represent the seasonal mean daily rainfall computed from the IMERG. Note also that dry grid points (seasonal mean daily rainfall amounts less than 1 mm day⁻¹) are masked out (gray shading). (e) The variation of the domain-averaged seasonal mean absolute error with rainy day threshold. The absolute error is also computed for a lead time of 1 day, and only for hits, with thresholds ranging between 0 and 10 mm day⁻¹ at 0.2-mm intervals.

rains are concentrated north of the equator, there is an overestimation of ~ 6 mm day⁻¹ over the Ethiopian highlands (Fig. 2c), while in the DJF season, the overestimation is predominant over the elevated terrain south of the equator accordingly. A feature independent of the season considered is the overestimation over highlands and mountains, e.g., the mountainous Congo–Uganda/Rwanda borders and Mount Kilimanjaro in all the seasons (Fig. 2). However, when the errors are normalized with their seasonal means, the overestimation in the rainy seasons scales down. Rather, there are large normalized mean absolute errors (NMAEs)

in the dry seasons (see Fig. S1 in the online supplemental material), probably due to higher frequency of very low rainfall intensities (< 0.2 mm) in IMERG. The accuracy of the reforecasts reduces with an increase in rainy day threshold (see Fig. 2e). The MAE increases from about 3–4 mm day⁻¹ for a rainy day threshold of > 0 mm day⁻¹, to reach about 7.5–8 mm day⁻¹, depending on the season, at a threshold of > 10 mm day⁻¹. The absolute error initially sharply rises from the initial value reaching ~ 5 mm at a threshold of about 0.2 mm day⁻¹ and then continues increasing almost linearly, but at a lower rate. The largest inaccuracy is recorded in JJA,

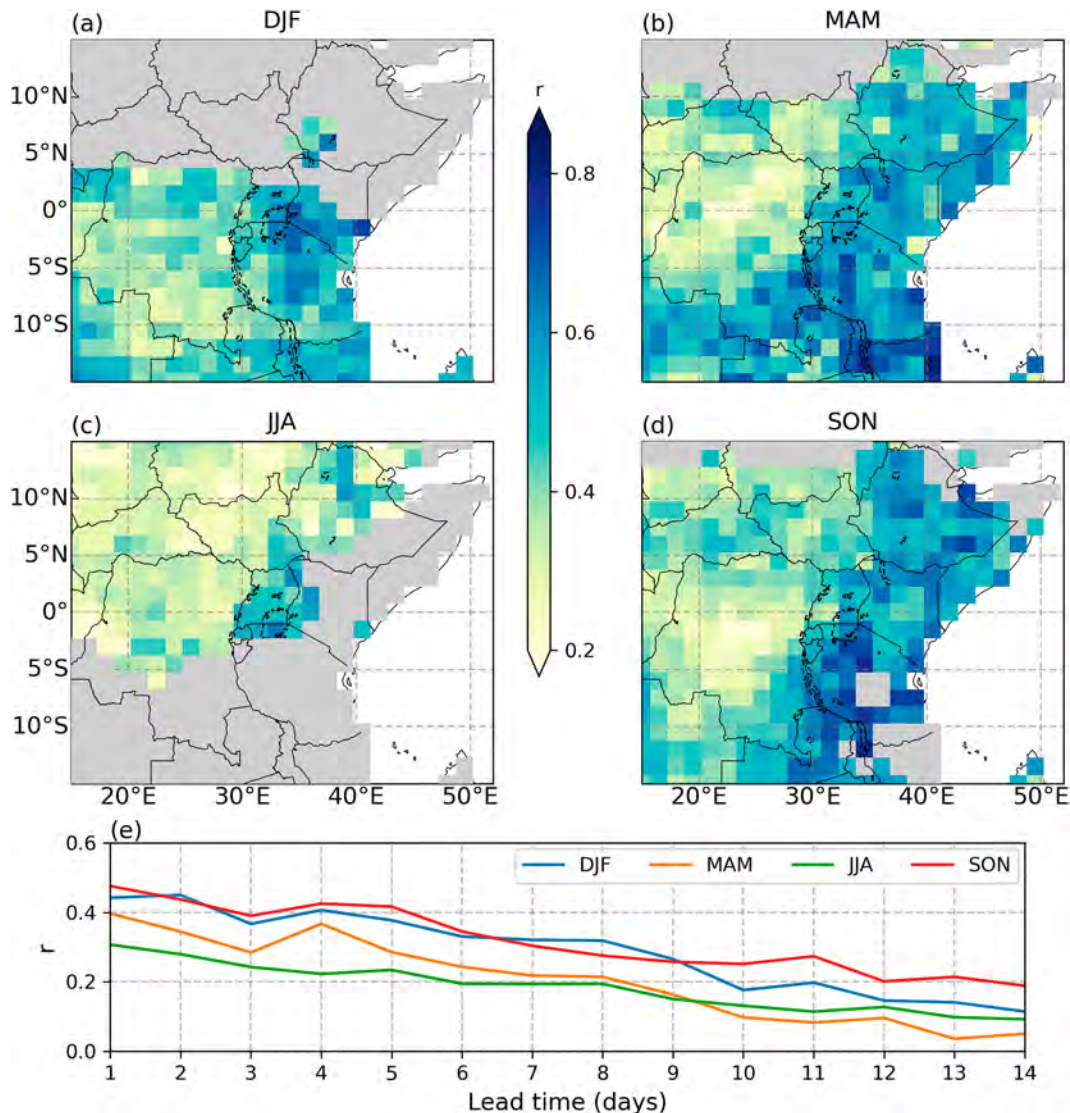


FIG. 3. (a)–(d) Association between the reforecasts and IMERG for a lead time of 1 day ahead. The Pearson correlation coefficient (r) at each grid point is computed by considering only days in the respective seasons and for the period 2001–18. The r values in all grid points are statistically significant at the 95% level, based on the Student's t test. (e) The mean seasonal correlation for the different lead times is shown. For each season, the mean is calculated by concatenating all the wet grid points.

followed by MAM, DJF, and SON, based on the domain-averaged MAE and NMAE. These biases can be partially attributed to model deficiencies. However, caution should be exercised here because the observation dataset, IMERG, has been known to exhibit a dry bias over high-altitude topography and for high-intensity rain rates (O and Kirstetter 2018; Ageet et al. 2022), possibly accounting for the larger biases.

Despite the biases in Figs. 2a–d, we see many grid points with zero or close to zero ME values, indicating low biases at these points. The fact that the model mostly overestimates rainfall at locations of maximum observed rainfall accumulation (blue dotted contours) also suggests that it

captures the seasonal cycle of the rainfall, putting rainfall in the right locations, albeit with wrong amounts and/or frequency. On the other hand, the model underestimates precipitation in some locations, especially over low lands like the Congo basin in all the seasons. This can be explained by the fact that the model struggles to represent convective rainfall at the mesoscale, often leading to very frequent low-intensity rainfall (Marsham et al. 2013; Birch et al. 2014; Vogel et al. 2018).

The association between the reforecasts and observation in all the seasons is generally positive (Fig. 3). However, the degree of association varies with season, being moderately positive in most grid points in DJF, MAM, and SON and

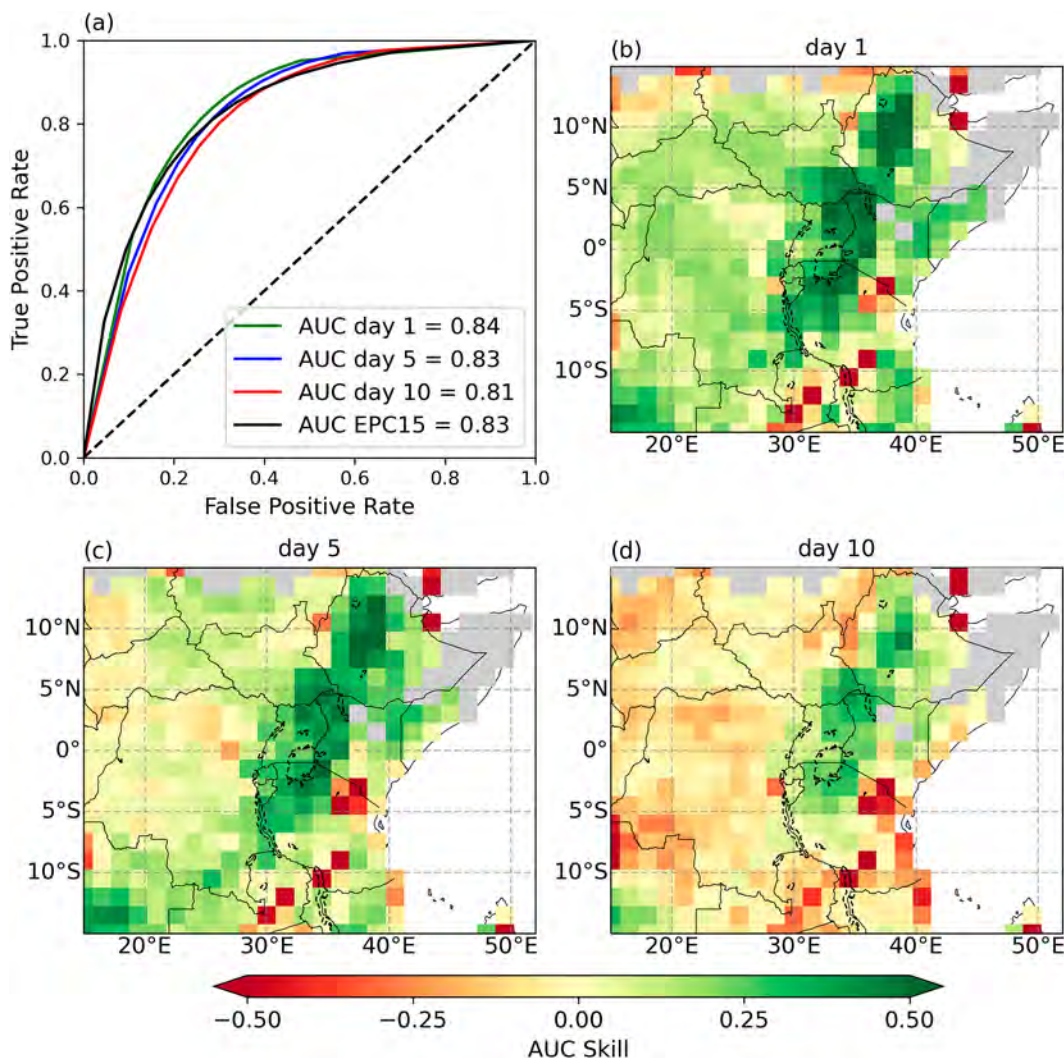


FIG. 4. Full domain-averaged receiver operating characteristic (ROC) curve (a) for day 1 (green curve), day 5 (blue curve), day 10 (red curve) lead times, and EPC15 (black curve) at a rainy day threshold of 0.2 mm. (b)–(d) The area under the curve skill (AUCS) computed relative to EPC15. Dry grid points (in this case, grid points with an annual mean daily rainfall of less than 1 mm day^{-1}) are masked out. Although only the 1-, 5-, and 10-day lead times are shown here, the analysis was repeated for all the other lead times, i.e., up to day 14.

positively weak in JJA (Figs. 3a–d). Over the eastern sector and high-altitude regions like the East African and Ethiopian highlands (cf. Fig. 1), the correlation is higher than over lower altitude regions like the Congo basin. These findings are in agreement with those in de Andrade et al. (2021), who also showed a better association in DJF, MAM, and SON compared to JJA and over eastern Africa compared to the Congo basin. As expected, the correlation decreases as the lead time increases (Fig. 3e). The differences among the seasons can also be seen, with generally DJF and SON being best, followed by MAM and then JJA across most lead times. The poor association between the reforecasts and observations over the Congo basin region has been partly attributed to the models' inefficiency in representing the north–south convection migration during the year in this region (de Andrade et al. 2021).

b. Probabilistic verification

1) RAW REFORECASTS

Over the entire study period, the reforecasts are able to distinguish between rain and nonrain events at 0.2 mm rainy day threshold up to 14 days ahead for some grid points (Fig. 4). The curves in the ROC plot are to the left of the diagonal (Fig. 4a), with shorter lead times being the furthest to the left. This implies that the discrimination ability of the reforecasts decreases with lead time, since the model gradually drifts away from the truth as memory of the initial conditions is eroded with time. EPC15 (black curve) compares favorably with the model (colored curves), i.e., the curves are close together. The AUC computed as the area between the diagonal and the curve in Fig. 4a, and consequently, the AUCS

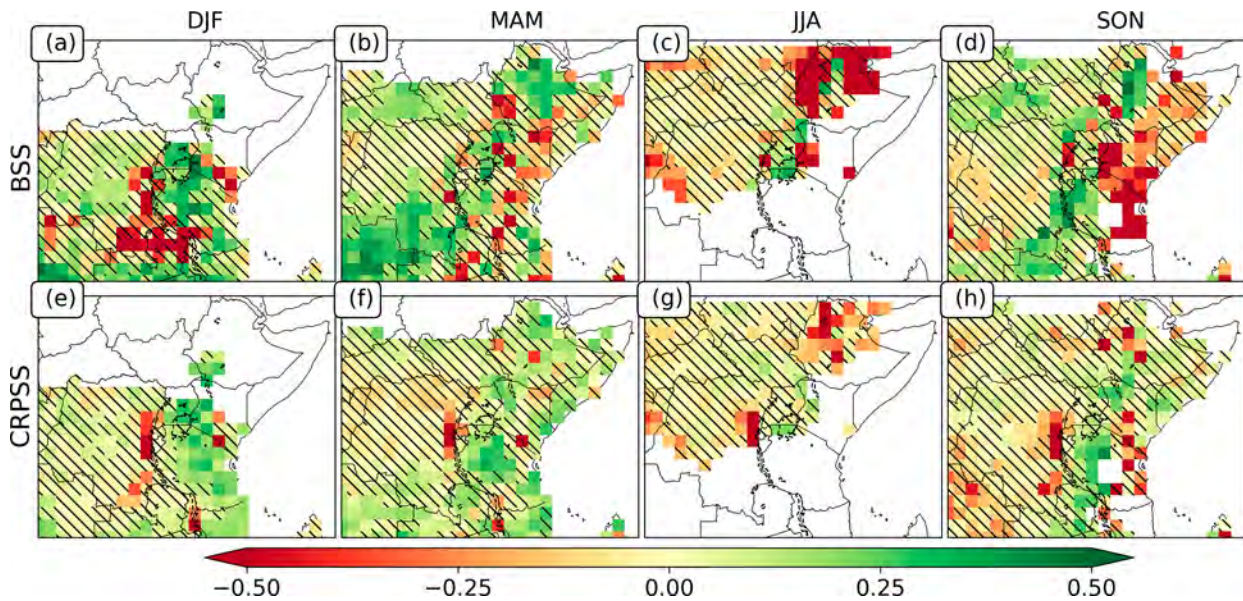


FIG. 5. BSS and CRPSS of raw hindcast relative to EPC15 at 1-day lead time. The skill scores are calculated for each of the four seasons and only for wet ($>1.0 \text{ mm day}^{-1}$ in the season) grid points. The hatching shows grids where the reforecasts are not significantly different from EPC15 based on the BS/CRPS values according to the DM test.

computed relative to EPC15 (Figs. 4b–d) confirms the potential skillfulness of the reforecasts in most parts of the study domain. At 1-day lead time, most of the grid points have positive AUCS with the strongest values over Uganda, the Ethiopian highland, and the southwestern regions. The distribution of the AUCS has an almost southwest–northeast orientation with the better scores over the elevated terrain (cf. Fig. 1). Arguably, the model has a better ability to represent orographically triggered precipitation over highlands. Furthermore, but likely of secondary importance, moist static instability over the highlands is not high such that more moisture and higher instability need to be present, e.g., from larger-scale convection signals such as MJO (Pohl and Camberlin 2006). This means that over the highlands, the triggers of convection are at larger scales, which are known to have better predictability (e.g., Vitart 2017; de Andrade et al. 2021; Specq and Batté 2022), and improve the model performance. Over the low-lying regions like the Congo basin, a low convective inhibition (CIN), and medium convective available potential energy (CAPE) environment, there are more stochastic triggers which the model largely fails to represent (Rasheeda Satheesh et al. 2023). This result corroborates the low correlation seen over the Congo basin (cf. Fig. 3). The AUCS deteriorates with an increase in lead time (Figs. 4c,d). The positive AUCS especially at day 1 has also been previously shown in Walz et al. (2021), although they used the operational forecasts from the earlier version of the ECMWF model.

The raw reforecasts have skill in predicting rainfall occurrence relative to EPC15 depending on the time of the year and location (Figs. 5a–d). Based on the BSS values, the reforecasts are able to capture a rainy day in most grid points in MAM, being up to 40% better than the EPC15. The strongest positive BSS lies along the northeast–southwest stretch in the

study domain confirming the potential skill previously seen (cf. Fig. 4). In DJF and JJA, most grid points, especially in the locations of maximum seasonal rainfall occurrence, have negative BSS values; i.e., the ECMWF model performs worse than EPC15 in predicting a rainy day. This is most likely linked to the fact that the hindcasts have a tendency to overestimate the frequency of precipitation during the rainy season (cf. Fig. 2). In SON, the model fails to correctly forecast rainy days along the East African coast (EAC) at all the grid points. Indeed, in all the seasons, the BSS is mostly negative along the coast. This may be due to the model's deficiencies in reproducing the sea–land–breeze effect, although caution should be exercised here since IMERG struggles retrieving warm rain along the coast (Vogel et al. 2020). Based on the DM test, the ability of the raw reforecasts to detect rain or no-rain events is significantly different from that of the EPC15 in only 38%, 43%, 31%, and 34% of the grid points in DJF, MAM, JJA, and SON, respectively (Table 1). Taking the percentage of grid points with positive BSS values into consideration, we conclude that, although the BS improves

TABLE 1. Percentage of grid points where the BSS and CRPSS are positive (skill) and the score of the reforecasts is significantly different from that of EPC15 (DM) for each of the seasons. The table corresponds to the results shown in Fig. 5.

	BSS		CRPSS	
	Skill (%)	DM (%)	Skill (%)	DM (%)
DJF	69	38	77	38
MAM	65	43	76	33
JJA	22	31	26	29
SON	53	34	45	22

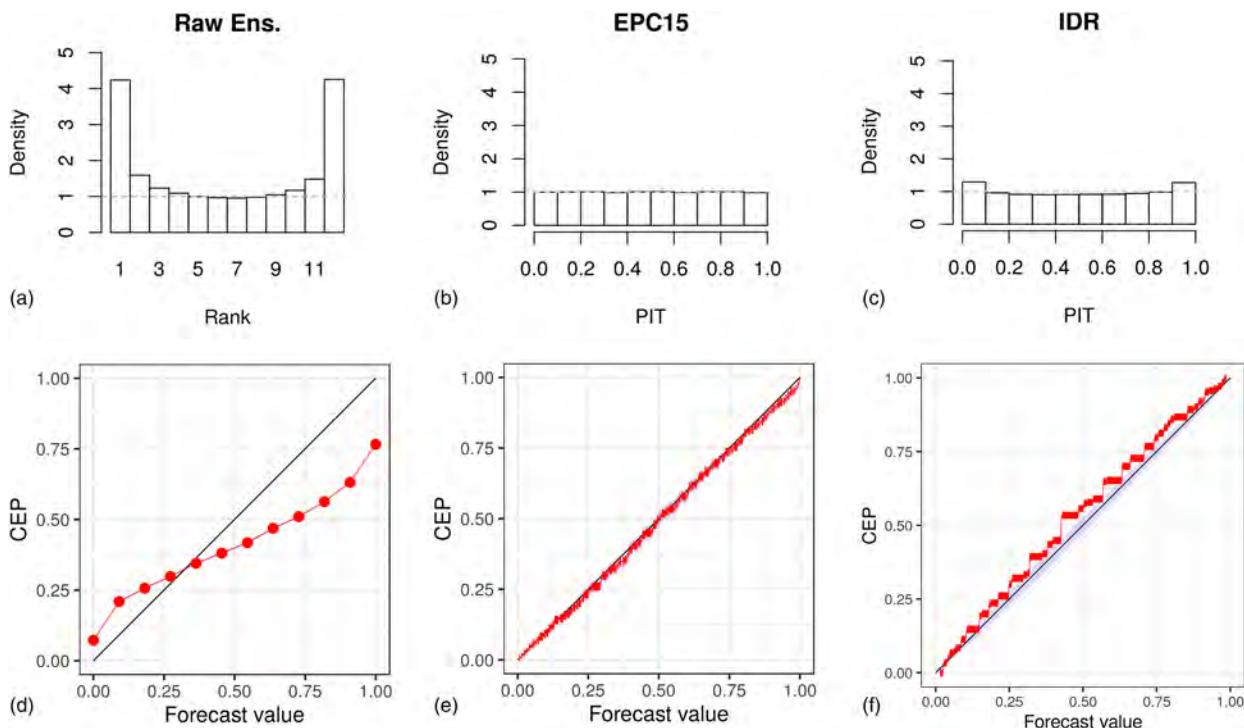


FIG. 6. (a) Rank histogram at a lead time of 1 day for the raw ensemble reforecast, (b) the PIT for the reference forecast, and (c) IDR-postprocessed reforecast for the period 2001–18. The bottom plots are the corresponding reliability diagrams. CEP in the y label stands for conditional event probability.

for many grid points relative to EPC15 in all the seasons, with the exception of JJA, the scores are not significantly better in all the grid points.

The performance of the full distribution of the raw reforecast relative to EPC15 is assessed using CRPSS values shown in Figs. 5e–h. The performance of the reforecasts varies spatiotemporally and is generally superior to EPC15; however, the CRPSS values are weaker than the BSS values for the same seasons and grid points. In a few places, e.g., along the coast, the CRPSS values are higher than the BSS values, similar to the general observation in Vogel et al. (2020), who also found that the ECMWF operational model performed better for rainfall amounts. The CRPSS is again best in the DJF and MAM seasons where most of the grid points have positive values. However, the scores are not significantly different from those of EPC15 in most grid points (Table 1). The worst performance is seen in the JJA season with most of the grid points having negative CRPSS values. The weaker performance for the CRPSS compared to BSS for the same places, e.g., southwest of the domain and Congo basin in MAM, suggests that the reforecast can differentiate between a rainy day but has biases in the intensities. The opposite is true for some locations like over the western border of Tanzania in DJF.

The reforecasts are miscalibrated, as shown in the underdispersed histograms (Fig. 6a). The observation frequently ranks lowest or highest in comparison with the ensemble; i.e., the observation falls outside of the ensemble range more often than expected. The miscalibration can also be seen in the

reliability diagrams for probability forecasts of rainfall occurrence (Fig. 6d). The reforecasts are overconfident; that is, the event happens less (or more) frequently than expected when forecasting high (or low) probabilities. As expected, EPC15 is well calibrated (Figs. 6b,e) with the PIT histograms being uniformly distributed and the calibration curve following the diagonal almost perfectly.

Figures 7a–c show the CORP decomposition of the raw BS. The worst MCB is in parts of Eastern Kenya and around the Mount Elgon area. Mount Kilimanjaro region and southwestern Ethiopia also have relatively high MCB values. The best DSC is in the southeastern parts of the study domain, stretching from central Tanzania to southern Congo. The model also has relatively high discrimination ability in the northern region, parts of southern Sudan and Ethiopia. The Congo basin features the lowest, i.e., worst, DSC values in the study region, explaining the lower correlation and AUCS values discussed earlier.

2) POSTPROCESSED REFORECASTS

Generally, postprocessing using the IDR method reduces the miscalibration. The underdispersion is substantially reduced and the curve is closer to the diagonal in the reliability diagram (Figs. 6c,f). Hence, the BSS and CRPSS of the reforecasts improve in almost all the grid points in all the seasons (Figs. 8a–h and Table 2). The BSS values are considerably better, especially in MAM, reaching 0.5 in some areas, e.g., the Ethiopian highlands, eastern Uganda, and southeastern

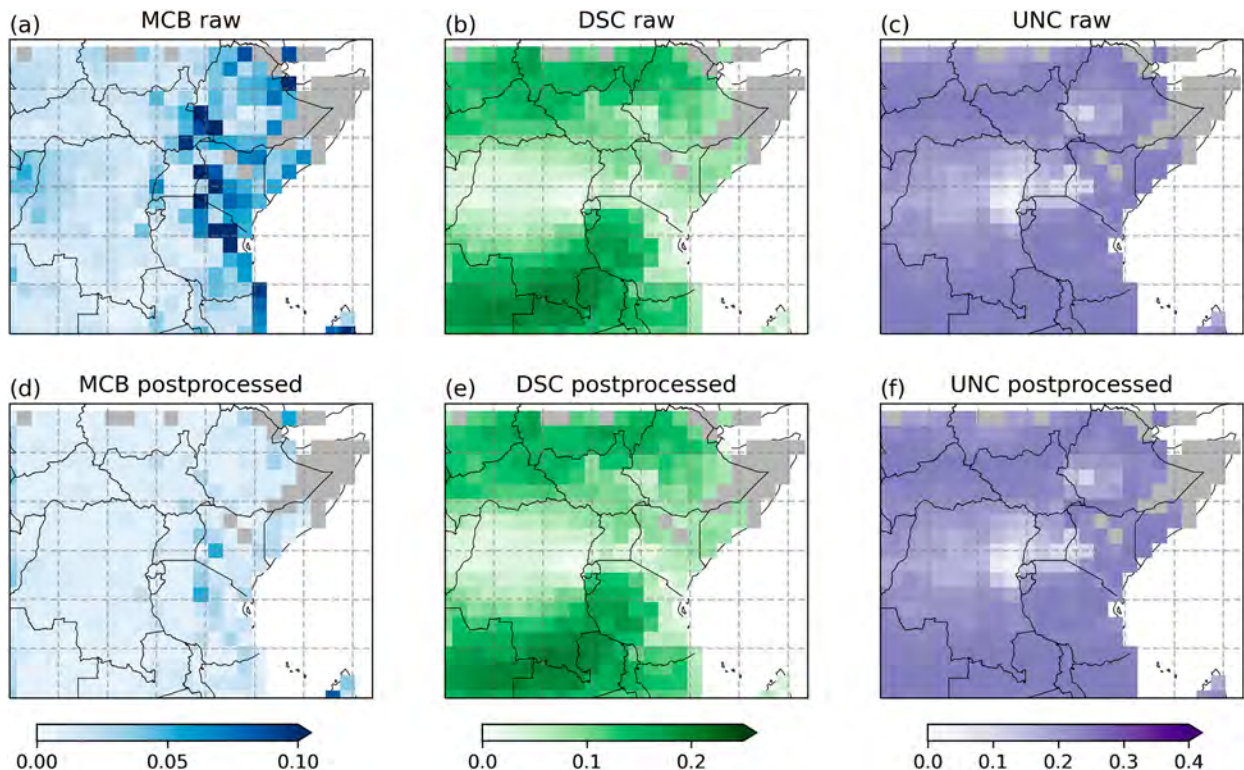


FIG. 7. Decomposition of the CORP score into the miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components for (a)–(c) raw and (d)–(f) IDR-postprocessed reforecasts. The metrics are averaged over all 18 years (2001–18), including the entire year (i.e., 105 days). The dry area (as defined in Fig. 4) is masked out as before. Note that the DSC and UNC for raw and postprocessed reforecasts are equal as postprocessing only corrects for miscalibration in a forecast.

Congo. The scores are again mostly higher at elevated terrain, whereas over the flatter regions, especially the Congo forest basin, they are still low and not significantly different from those of the EPC15. The performance of the postprocessed compared to the raw reforecasts is better in SON too, with only a few areas having high negative BSS values, e.g., over mountainous regions in Uganda (Elgon and Rwenzori), the raised areas in the southeast of Kenya, and the mountains over the Rwanda/Burundi–Congo borders. In DJF and JJA seasons, the performance is lower compared to the other two seasons. This is clear in the southeastern Congo and the western parts of Tanzania in DJF, where we see high negative BSS values. The same can be seen in JJA over the Ethiopian highlands. A detailed analysis of these two regions revealed frequent cases of “forecast busts” (almost all ensemble members forecasted rain, but it did not rain). These busts resulted in very high BS values for these days, which negatively skewed the BSS in these areas. The reduction in hatched grid points further highlights the improvement after postprocessing in all the seasons, although to a very small extent in JJA.

The skill with respect to the CRPSS of the postprocessed reforecasts for rainfall amounts also improved (Figs. 8e–h). Similar to the BSS, the CRPSS is better in the DJF and MAM and worst in the JJA. Contrary to the BSS in the DJF season, CRPSS around the southeastern Congo and the western border of Tanzania is positive. This means that the model is

better in forecasting rainfall amounts than in distinguishing between rainy and dry days in these parts of the study domain. This feature of the model could be important for some applications like agriculture where the focus is not really on the timing but rather the amount of rainfall in a particular area. In MAM and SON seasons, the lower scores over low-lying areas like the Congo basin region are also apparent. Despite the improvement, the DM test suggests that the performance is not significantly different from that of EPC15 in most grid points, especially in JJA and SON (Table 2). Of note also is that the largest gains in skill after postprocessing are in regions like the EAC which initially have a high degree of miscalibration (cf. Figs. 7a,d). For places like the Congo basin where the source of error was the low discrimination ability, the benefit of the postprocessing is almost zero (Figs. 8i–p) as DSC was not affected by postprocessing (cf. Figs. 7b,e).

Just like in the raw reforecasts, not all the positive BSS and CRPSS are significantly better than the EPC15, given the higher percentages for “skill” compared to the “DM” column (Table 2). Directly comparing Tables 1 and 2, there is an average improvement of 82% (67%) in the number of grid points with positive BSS (CRPSS) after postprocessing at a lead time of 1 day. The number of grid points where the reforecasts are significantly different from EPC15, at a lead time of 1 day, also increased by an average of 50% and 45% for BSS and CRPSS, respectively. This analysis was extended to the

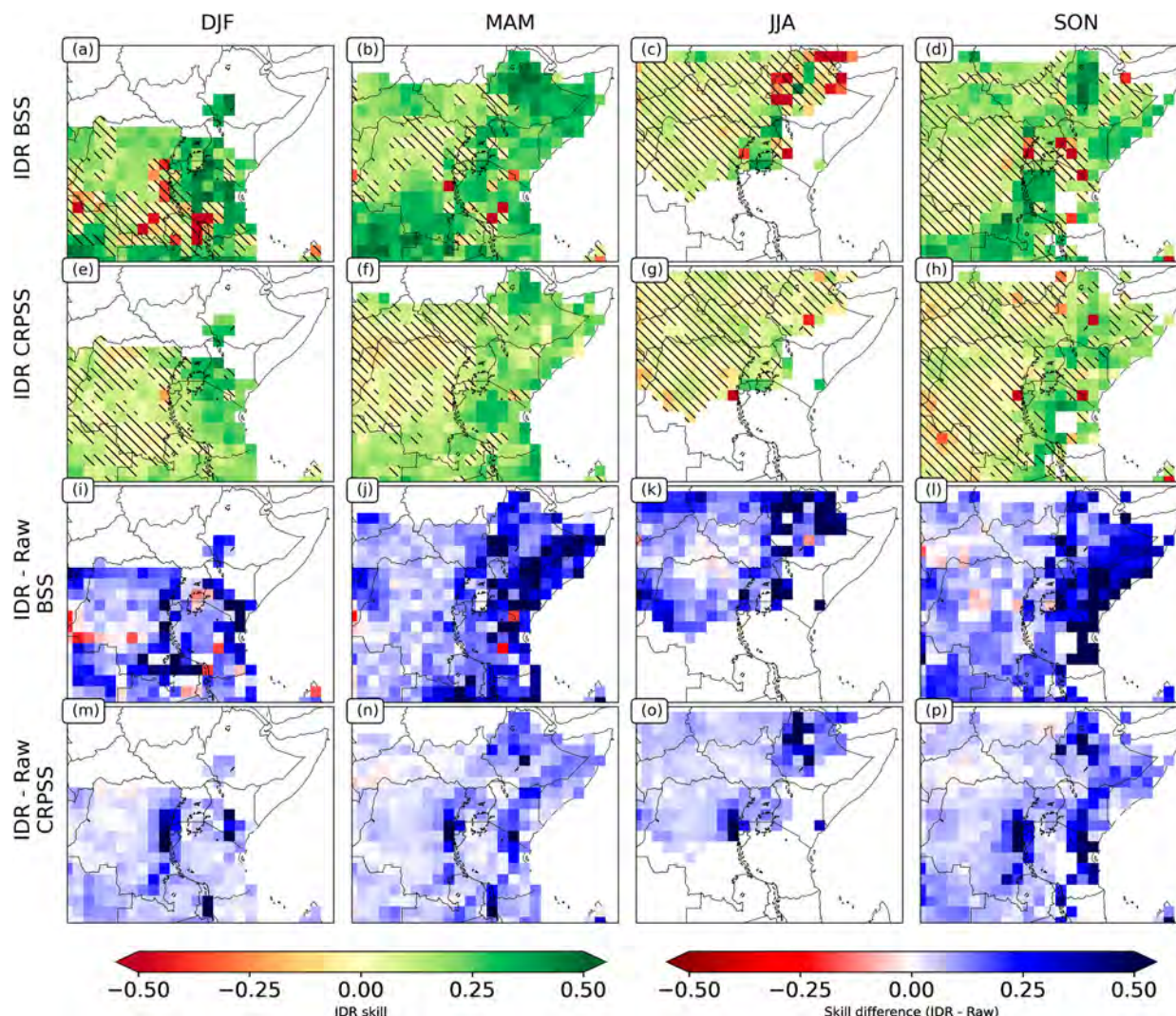


FIG. 8. (a)–(h) As in Fig. 5, but for IDR-postprocessed reforecasts. (i)–(p) The difference in skill between the postprocessed and raw reforecasts.

other lead times, i.e., 2–14 days ahead (Table 3). The improvement fluctuates, varying between 60% (66%) and 100% (158%), with an average over all the lead times of 81% (91%) for the BSS (CRPSS). Similar improvements were seen in Vogel et al. (2020) for rainfall occurrence over the tropics. This increase coupled with the positive anomalies in Figs. 8i–p implies that the absolute values of the BSS and CRPSS are also generally improved relative to EPC15.

TABLE 2. As in Table 1, but for the postprocessed reforecasts. The table corresponds to the results shown in Figs. 8a–h.

	BSS		CRPSS	
	Skill (%)	DM (%)	Skill (%)	DM (%)
DJF	81	60	92	61
MAM	95	83	93	63
JJA	66	26	66	20
SON	88	56	78	35

From Table 4, the best BS (CRPS) performance is in DJF (JJA) season for the raw and postprocessed forecasts and EPC15. The generally better scores in the dry seasons do not necessarily translate to better skill. This is highlighted by the worst skill scores relative to EPC15 in the JJA season (e.g., cf. Table 1). Indeed, in JJA, even after postprocessing, the BS of the reforecast is only equal to the score of EPC15, highlighting the limited predictability in this season. Only in DJF and MAM are the raw reforecasts either equal or better than the EPC15, and the scores are further improved after postprocessing. In SON, even though the scores are initially worse than those of the reference forecast, the situation is reversed after postprocessing.

3) SPATIOTEMPORAL VARIATION OF SKILL

The skill scores of the ensemble reforecasts vary with location and lead time (Fig. 9). In all the regions (cf. Fig. 1), the performance of the reforecasts for occurrence and amounts

TABLE 3. Improvement after postprocessing computed by comparing the percentages of grids with positive skill before and after postprocessing.

Lead time (days)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Mean
BSS (%)	82	60	100	83	87	107	75	81	85	72	58	71	98	73	81
CRPSS (%)	67	83	88	83	86	66	101	68	83	85	108	158	103	—	91

for 24-h accumulations degrades with lead time as expected. Both of the two approaches used here: (i) counting the number of grid points with positive BSS/CRPSS and (ii) displaying the absolute BSS/CRPSS with boxplots, show this characteristic of the model. We also see the improvement after postprocessing with the skill scores of the postprocessed (red color) being higher than those of the raw reforecasts (black color) in all the regions (Fig. 9). In terms of the absolute skill score, we find that, although the postprocessing improves the BSS and CRPSS for all the regions, the median scores (yellow line in the boxplots) are low positive or negative values, highlighting the poor performance of forecasting systems in the tropics. For some regions, the skill scores are always negative, e.g., EAC for rainfall occurrence detection and the Congo basin for rainfall amounts. This worrying level of skill in the tropics has previously been shown by Haiden et al. (2012) and Vogel et al. (2018, 2020).

The best skill scores for both occurrence and amounts of rain are over the Ethiopian highland, followed by the Uganda region. These two regions have relatively higher altitudes, consistent with earlier results which suggested the model performed best over high-altitude areas. Over the Congo basin, the results for occurrence, especially for the postprocessed reforecasts, are positive up to day 5 (based on the median scores in the boxplots) and over 50% of the grid points have positive BSS up to day 6 (Fig. 9b). However, for rainfall amounts, the skill scores are always negative after day 1 (Fig. 9f). This region, as explained earlier, is one where the model notably underestimated rainfall and has poor discrimination ability. The performance at the EAC is the worst being negative all the time for the raw reforecasts. After postprocessing, this region shows a large improvement in BSS, being positive up to day 8 (Figs. 9c,g). The fact that raw reforecast in this region had the highest miscalibration explains this high improvement after postprocessing. Similar plots to Fig. 9 for higher accumulation of rainfall, i.e., 48 and 120 h (Figs. S2 and S3), show similar trends in the different regions. It is worth noting that with larger aggregation, the BS (CRPS) values get better (worse) as the occurrence mismatches (absolute errors) reduce (grow) (Table S2).

TABLE 4. Domain-averaged mean seasonal BS and CRPS of raw, IDR-postprocessed reforecasts corresponding to Fig. 5. The numbers in bold font are the best scores, and the numbers in italic font are the worst scores for each season.

	BS				CRPS (mm day ⁻¹)			
	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
Raw	<i>0.11</i>	0.15	<i>0.16</i>	<i>0.16</i>	1.81	2.04	<i>1.70</i>	<i>2.06</i>
EPC15	<i>0.11</i>	<i>0.16</i>	0.12	0.15	<i>1.91</i>	<i>2.12</i>	1.56	1.97
IDR	0.08	0.13	0.12	0.12	1.68	1.89	1.53	1.86

Generally, longer temporal aggregations of precipitation show better BSS values (Fig. 10a). This is expected as the error due to rainfall timing mismatches between the reforecasts and observations is gradually reduced. However, it should be noted that this is not the case in all regions, for example, over Uganda and Ethiopian highlands, where the skill scores of the 120-h raw and postprocessed reforecasts rank lowest in performance (Fig. S4). Similar findings were seen in Vogel et al. (2020), especially in dry areas. They reasoned that 5-day accumulations increased the number of 5-day dry periods in the observation while doing the opposite in the forecasts. We concur with their reasoning that accumulating precipitation for longer time range increases the observation–reforecast mismatches for that accumulation in some regions. Similarly, the model's ability for rainfall amounts also generally improves with an increase in the temporal aggregation (Fig. 10b).

c. Assessment based on conventional rain gauges

Because rain gauge measurements are often regarded as the “truth,” we analyze the performance of the reforecast against available gauge data over Uganda. Three grid points (cf. Fig. 1) satisfied the set conditions, i.e., contained four stations, and the rainfall data availability of these stations was at least 95% for the period 2001–18. The stations used are listed in Table S1. The statistical methods applied are similar to those used in the previous section.

The ensemble forecasts more often than not predict the occurrence of a rainy day with certainty, i.e., probability of precipitation (PoP) value of 1.0, especially for the Lake Victoria region (Fig. 11a). Based on the ensemble mean, it rains more in the ensemble reforecasts than in IMERG and the gauges (numbers in the top left of Fig. 11). For most days, the observations agree on the occurrence of rain (highest frequency of gray shading). However, IMERG has a higher frequency of rainy days compared to the gauges (higher frequency of blue compared to red-shaded days). This is further highlighted in the reference forecasts; the EPC15 based on IMERG (blue dotted lines) has higher PoP values compared to the EPC15 based on gauges (red dotted lines) in all the three regions. Despite the difference in magnitude of the PoP, both reference forecasts reproduce well the annual seasonal cycle of the rainfall in the different regions.

Postprocessing using IDR modifies the PoP of the forecast, bringing the values toward the EPC15 curves. The IMERG-based postprocessed forecast (green line) shows very little difference though, with most days having a PoP value of 1.0, especially during the rainy seasons in the Lake Victoria region. The similarity between the raw and postprocessed forecasts may be due to the fact that, just like in the reforecasts, it rains very frequently in IMERG too, i.e., 87%, 77%, and 89%

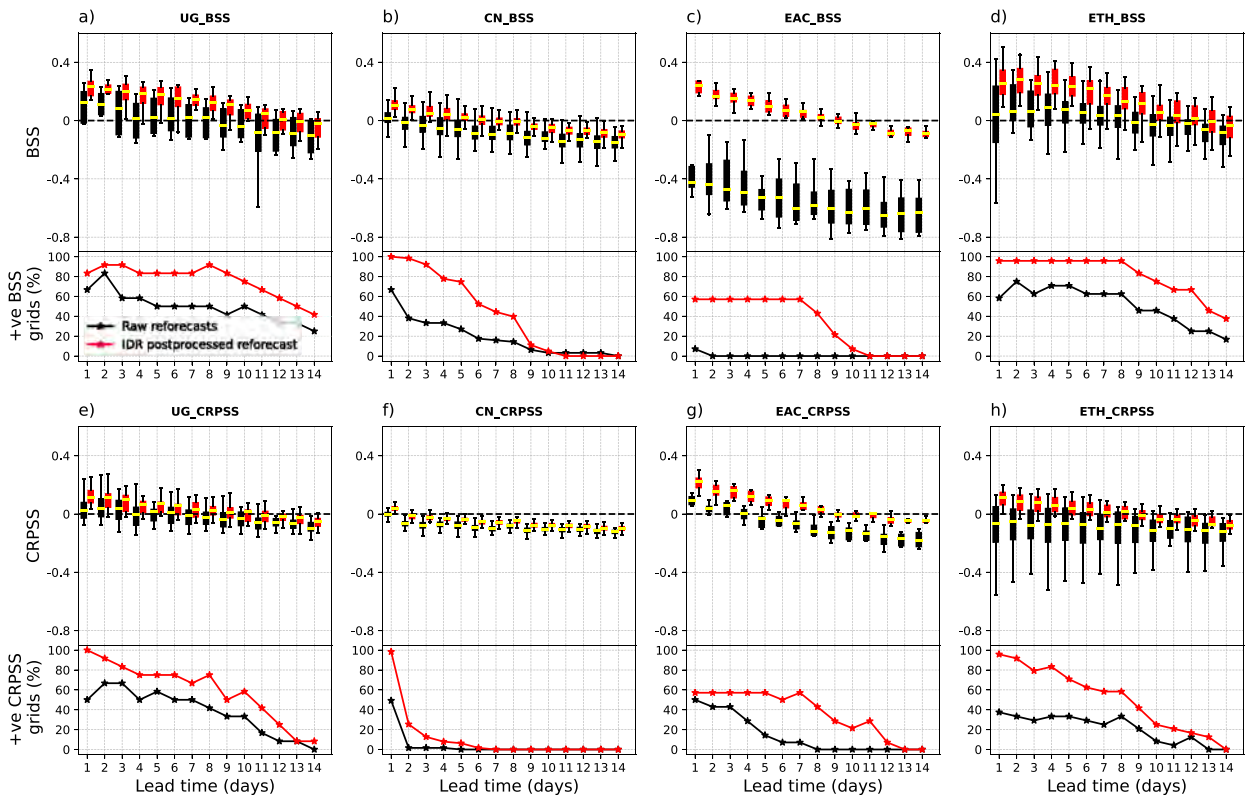


FIG. 9. Progression of the BSS and CRPSS over lead time for 24-h accumulation of rainfall for the four regions (cf. Fig. 1) in EEA. The line plots show the percentage of grid points with positive skill scores relative to all the grids in the particular region, with the black and red being for the raw and postprocessed reforecasts, respectively. The boxplots show the distribution of the actual skill scores in the region, with the yellow line denoting the median skill score for the raw reforecast and the red line denoting the median skill score for the postprocessed reforecast. There are 12, 63, 8, and 24 grid points in Uganda (UG), Congo basin (CN), East African coast (EAC), and Ethiopian highlands (ETH), respectively.

of the time in Lake Victoria, Lake Kyoga, and western Uganda regions, respectively, for the year 2001. The gauge-based postprocessing (orange line) modifies the raw forecast to a larger extent compared to IMERG, likely because of the

larger differences between the occurrence and amounts of precipitation in the observation and the forecast. We also note that the observation and forecast converge with an increase in the number of stations at a grid point. Note that

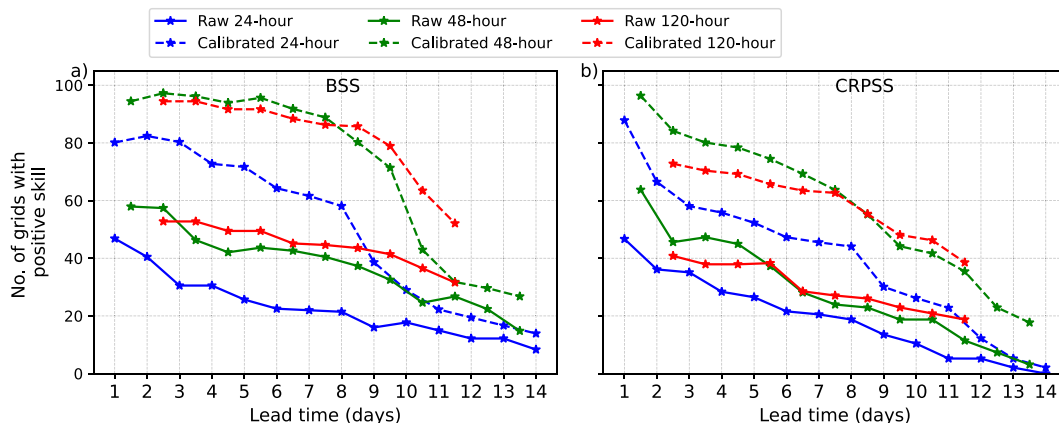


FIG. 10. Percentage of grid points with positive BSS and CRPSS in the raw and postprocessed reforecast for 24-, 48-, and 120-h accumulations of rainfall averaged over the whole study domain for the period 2001–18 and increasing lead time. Note that the x axis shows 24-h intervals. For the longer aggregations, the data points lie in the middle of the 24-h ticks, e.g., the day-1–2 accumulation data point is located at 1.5 days (the middle point of day 1–day 2), and day-1–5 accumulation data point is at 2.5 days (the middle point of day 1–day 5).

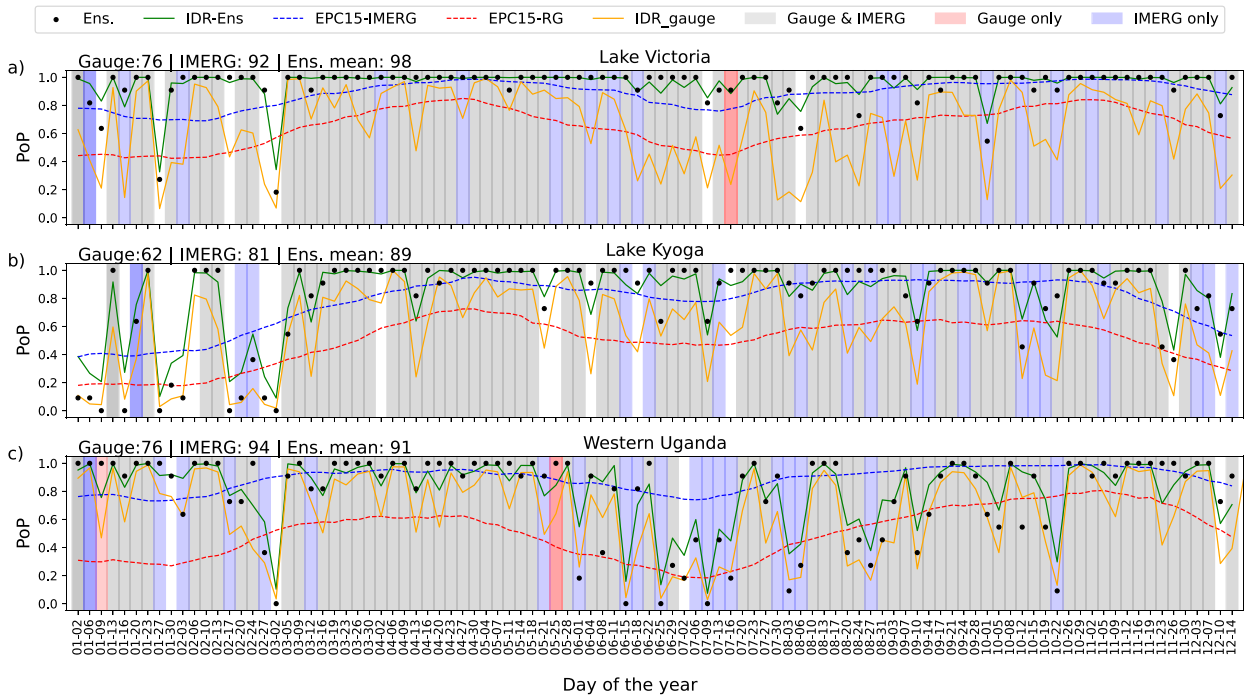


FIG. 11. Probability of precipitation (PoP) for 24-h rainfall accumulation evaluated at three grid points with four gauges (cf. inset of Fig. 1) at a rainy day threshold of 0.2 mm for the different datasets: EPC15 created from IMERG (blue dotted line) and gauges (red dotted line), IDR-postprocessed forecasts using IMERG (green line) and gauge (orange line) observations, and the raw ensemble forecast (black dots). The shading shows when a rainy day was observed in the gauge only (red), IMERG only (blue), and in both gauge and IMERG (gray). The numbers in the top-left corner are the counts of the rainy days in the gauge, IMERG, and the ensemble mean. Note that only the first 100 out of the 105 reforecasts for the year 2001 are shown.

here we show only the PoP of 2001 as it is representative of the other years (i.e., 2002–18).

The BSS and CRPSS of the raw and postprocessed forecasts are variables in the three regions, with one common characteristic, that is, the decrease with lead time (Fig. 12). The best (worst) skill scores are depicted in the Lake Kyoga (Lake Victoria) region for both occurrence and amount of rainfall. The scores are positive, especially the BSS in Lake Kyoga and western Uganda regions, being 35% better than EPC15 at 1-day lead time. The reforecasts perform better for rain occurrence detection compared to accuracy of amounts (i.e., $BSS > CRPSS$). The CORP decomposition of the BS shows that the miscalibration of the raw forecasts is worst for the Lake Victoria region and the discrimination ability of the forecast is best in the Lake Kyoga region (Fig. S5). After post-processing, the largest improvement in BSS/CRPSS is unsurprisingly seen in the Lake Victoria region because IDR mainly reduces the miscalibration which was largest in this region. In Lake Kyoga and western Uganda, the improvement is smaller due to the initially low miscalibration in these regions. Similar to the findings in Fig. 10, the skill for the larger aggregations, i.e., 48- and 120-h rainfall accumulations, increases (decreases) for BSS (CRPSS) in all the three regions. However, although this is generally true for most parts of our study domain, for some regions, e.g., Uganda and Ethiopia, the BSS (CRPSS) decreased (increased) with increasing rainfall accumulation.

4. Discussion and conclusions

The study evaluated the skill of rainfall reforecasts from ECMWF against IMERG and gauge observations over EEA for the period 2001–18. The analysis was done on multiple spatiotemporal aggregations. The reforecasts were analyzed using several verification methods in deterministic and probabilistic terms. The ME and MAE are used to assess the bias and accuracy in the deterministic forecasts, while BS and CRPS assess the predictive performance of the ensemble reforecasts in probabilistic terms. The reliability diagrams, ROC curves, and AUC values were used to assess calibration and discrimination ability of the reforecasts. The skill was assessed using the BSS, CRPSS, and AUCS, all computed relative to EPC15. Further, the raw reforecasts were postprocessed to correct the miscalibration, and the resulting forecasts were evaluated using the same verification methods as for the raw reforecasts. The main findings of the study are as follows:

- 1) The reforecasts are biased, with overestimation of rainfall amounts over mountainous regions. This overestimation is more pronounced during the rainy season. Moreover, the absolute error increases in all seasons with an increase in the rainy day threshold. The overestimation of rainfall observed in the ECMWF reforecasts, especially during the rainy seasons and over raised topography, agrees with other past studies. de Andrade et al. (2021) and Endris et al. (2021) also showed overestimation over most parts of Africa

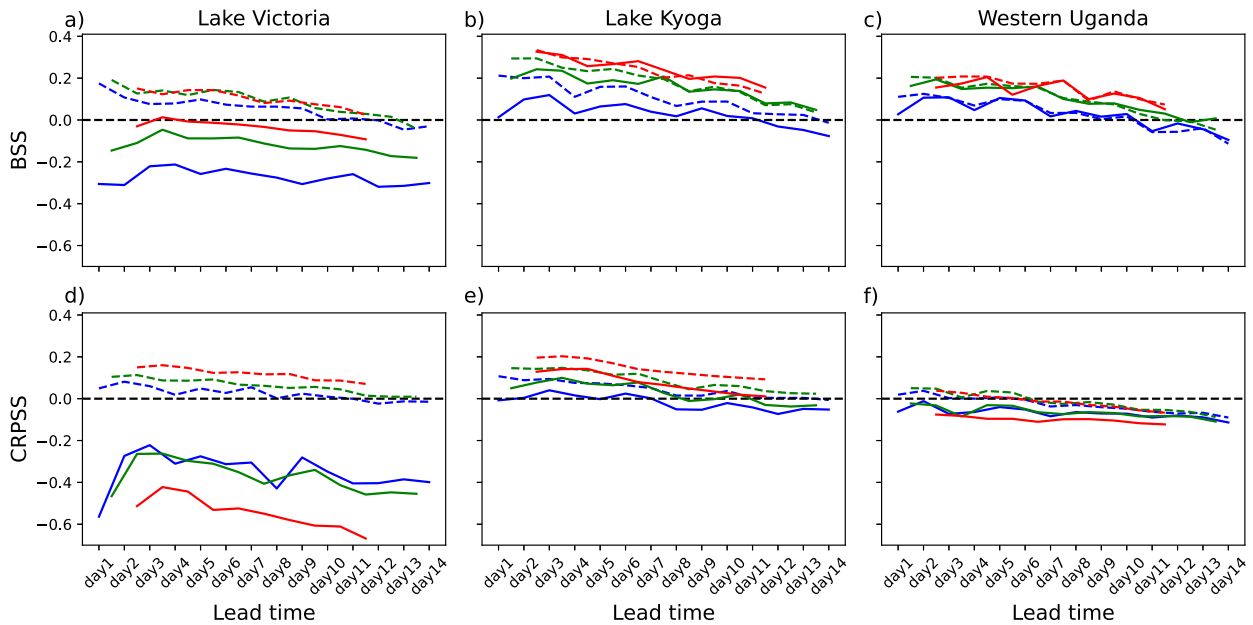


FIG. 12. Skill scores for raw (solid lines) and IDR-postprocessed (dashed lines) reforecasts for 24- (blue line), 48- (green line), and 120-h (red line) rainfall accumulations in the three regions of Uganda relative to EPC created with gauge precipitation data.

and the Greater Horn of Africa, respectively. They also noted that the overestimation was most pronounced during the rainy season. Overestimation was also shown by [Stellingwerf et al. \(2021\)](#) over Ethiopia, especially for the higher intensity amounts, although they used the ECMWF operational forecasts. As pointed out earlier, the overestimation may not be solely due to model errors since IMERG has also been shown to underestimate warm rain in this region.

- 2) The raw reforecasts are potentially skillful, being able to discriminate between events and nonevents up to day 14, depending on location. This potential skill is translated into positive skill, especially over land with positive BSS and CRPSS, depending on location, lead time, and temporal aggregation. The improvement of the raw reforecasts relative to EPC15 is up to 30% (i.e., BSS and CRPSS values of 0.3) in some areas. However, the BS and CRPS values are largely not significantly different from those of the EPC15. The fact that ECMWF reforecasts are skillful confirms results from previous studies which showed that precipitation in this region has higher predictability compared to other regions in Africa and the model is better compared to other models over the region ([de Andrade et al. 2021](#)). We also note that the skill is dependent on the season, being best in the DJF and MAM, followed by SON, and worst in JJA. This temporal dependence of skill has also been shown by other verification studies in the region (e.g., [de Andrade et al. 2021](#); [Endris et al. 2021](#)). We also see that although the skill for rainfall amounts (CRPSS) is highest in the MAM and SON, the magnitude is lower than that of the BSS in corresponding seasons. However, we also note that on average, about 49% (44%) of the grid points had negative BSS (CRPSS) relative to EPC15. This low or negative skill supports findings in the tropics ([Haiden et al. 2012](#); [Vogel et al. 2018, 2020](#)).

- 3) The reforecasts are subject to biases and calibration errors. Postprocessing using IDR substantially reduces the miscalibration, hence boosting the skill with a 50% improvement relative to EPC15 for most grid points especially for rainy day occurrence in MAM, DJF, and SON seasons. Past studies suggested and showed the benefit of postprocessing ([Vogel et al. 2018, 2020](#); [Schulz and Lerch 2022](#)). We also find that postprocessing the reforecast using IDR considerably increased the number of grids with positive BSS and CRPSS over EEA by an average of 81% and 91%, respectively. The largest improvements occur in regions with the highest miscalibration.
- 4) The analysis against gauges confirms overconfidence of the reforecasts and the improvement after postprocessing. However, it is clear that rainfall in the model and IMERG occurs more frequently compared to the gauges. Increasing the gauge network helps reduce the bias, pointing to the common problem of point versus gridded dataset comparisons. In our case, we saw an improvement in skill when the number of gauges was increased. We used four gauges in a grid point which is far from ideal given the $1.5^\circ \times 1.5^\circ$ resolution of the reforecasts. However, given the general lack of good and consistent gauge networks in this region, the four gauges in a grid point is a fair number and provided valuable insights into the skill of the reforecasts.
- 5) The skill of the reforecasts varies spatially in the study domain, supporting the findings of previous studies ([Endris et al. 2021](#); [de Andrade et al. 2021](#)). The analysis at the different regions revealed that the best performance was over the raised areas of Ethiopian highlands and Uganda. This is partly due to the model being able to represent orographically triggered rainfall and the fact that the convection here is often connected to larger-scale signals like

MJO (Pohl and Camberlin 2006) which have higher predictability (e.g., Vitart 2017; de Andrade et al. 2021; Specq and Batté 2022). The lowest skill scores were over the East African coast, mainly due to the poor calibration. Over the Congo basin, the skill was also poor, owing to the low discrimination ability of the model. The low skill at low-lying regions like the Congo basin has been attributed to the difficulties of the model to represent convective rainfall (Marsham et al. 2013; Birch et al. 2014; Vogel et al. 2018). The rainfall triggers are stochastic (Rasheeda Sathesh et al. 2023), limiting the predictability. We acknowledge that although the poor skill especially at the coast is partly due to the model struggles, it has been suggested that IMERG, the observation dataset, struggles with warm rain retrieval at the coast (Vogel et al. 2020) and over high altitudes like mountains over East Africa (e.g., Diem et al. 2014; Monsieurs et al. 2018; Ageet et al. 2022). The analysis against gauges over the Uganda domain further highlights how variable the performance of forecasts can be even over small domains. This variation in skill emphasizes the need for validation studies to ascertain how the model performs in specific regions and not generalize.

This study has highlighted that raw reforecasts have skill especially over high-altitude areas which is potentially beneficial to meteorological services in the region. However, because the forecasts are biased and uncalibrated, postprocessing is necessary if the forecasts are to offer more meaningful information. Here, we used only one reforecast from the ECMWF center as this has been shown to be one of, if not the best, in the region. However, studies have suggested that using the multiforecast ensemble mean provides the best outcome (Stellingwerf et al. 2021). This would have the benefit of increasing the ensemble size, especially for reforecasts which often have a limited number of members (e.g., only 11 in our case), hence increasing the spread of the reforecast. We tested one novel postprocessing approach, but other classical approaches like ensemble model output statistics (EMOS) or Bayesian model averaging can also be applied. Additionally, machine learning approaches have the potential to further improve the quality of the postprocessed forecasts with the ability to incorporate more information than ensemble forecasts of precipitation (Schulz and Lerch 2022). Given the coarse resolution of the reforecasts, it would also be interesting to see how the skill compares if more than the four stations used here are included in the verification of a grid point. Studies like Macleod et al. (2021), de Andrade et al. (2021), and Specq and Batte (2022) have shown that skill is regime-dependent in this region. As a next step, we intend to stratify the skill shown here based on known sources of predictability in the region, namely, MJO, IOD, and Kelvin waves.

Acknowledgments. S. A. was supported by a DAAD Ph.D. fellowship. A. H. F. and B. S. acknowledge support from the Transregional Collaborative Research Centre SFB/TRR 165 “Waves to Weather” (www.wavestoweather.de) funded by the German Research Foundation (DFG). We also thank Eva-Maria Walz for providing the code for calculating the EPCs.

Data availability statement. The reforecast data and the satellite observations used in this work are publicly available at <https://apps.ecmwf.int/datasets/data/s2s/levtype=sfc/type=cf/> and <https://daac.gsfc.nasa.gov/datasets>, respectively. The rain gauge data are available from the second author upon request.

REFERENCES

- Ageet, S., A. H. Dink, M. Maranan, J. E. Diem, J. Hartter, A. L. Ssali, and P. Ayabagabo, 2022: Validation of satellite rainfall estimates over equatorial East Africa. *J. Hydrometeorol.*, **23**, 129–151, <https://doi.org/10.1175/JHM-D-21-0145.1>.
- Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Birch, G. E., D. J. Parker, J. H. Marsham, and D. Copsey, 2014: A seamless assessment of the role of convection in the water cycle of the West African monsoon. *J. Geophys. Res. Atmos.*, **119**, 2890–2912, <https://doi.org/10.1002/2013JD020887>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Cafaro, C., and Coauthors, 2021: Do convective-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa? *Wea. Forecasting*, **36**, 697–716, <https://doi.org/10.1175/WAF-D-20-0172.1>.
- de Andrade, F. M., M. P. Young, D. MacLeod, L. C. Hirons, S. J. Woolnough, and E. Black, 2021: Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. *Wea. Forecasting*, **36**, 265–284, <https://doi.org/10.1175/WAF-D-20-0054.1>.
- Diebold, X. F., and S. R. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.
- Diem, J. E., J. Hartter, S. J. Ryan, and M. W. Palace, 2014: Validation of satellite rainfall products for western Uganda. *J. Hydrometeorol.*, **15**, 2030–2038, <https://doi.org/10.1175/JHM-D-13-0193.1>.
- Dimitriadis, T., T. Gneiting, and A. I. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proc. Natl. Acad. Sci. USA*, **118**, e2016191118, <https://doi.org/10.1073/pnas.2016191118>.
- Dinku, T., 2019: Challenges with availability and quality of climate data in Africa. *Extreme Hydrology and Climate Variability*, A. M. Melesse, W. Abtew, and G. Senay, Eds., Elsevier, 71–80, <https://doi.org/10.1016/B978-0-12-815998-9.00007-5>.
- Endris, H. S., L. Hirons, Z. T. Segele, M. Gudoshava, S. Woolnough, and G. A. Artan, 2021: Evaluation of the skill of monthly precipitation forecasts from Global Prediction Systems over the Greater Horn of Africa. *Wea. Forecasting*, **36**, 1275–1298, <https://doi.org/10.1175/WAF-D-20-0177.1>.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- , A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.

- Haiden, T., M. J. Rodwell, D. S. Richardson, A. Okagaki, T. Robinson, and T. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using SEEPS score. *Mon. Wea. Rev.*, **140**, 2720–2733, <https://doi.org/10.1175/MWR-D-11-00301.1>.
- Hastings, D. A., and Coauthors, 1999: The Global Land One-kilometer Base Elevation (GLOBE) digital elevation model, version 1.0. National Oceanic and Atmospheric Administration, National Geophysical Data Center, accessed 23 July 2021, <http://www.ngdc.noaa.gov/mgg/topo/globe.html>.
- Henzi, A., J. F. Ziegel, and T. Gneiting, 2021: Isotonic distributional regression. *J. Roy. Stat. Soc.*, **83**, 963–993, <https://doi.org/10.1111/rssb.12450>.
- Huffman, G., and Coauthors, 2020: NASA Global Precipitation Measurement (GPM) Integrated Multi-satellite Retrievals for GPM (IMERG). Algorithm Theoretical Basis Doc., 39 pp. https://gpm.nasa.gov/sites/default/files/2020-05/IMERG_ATBD_V06.3.pdf.
- IPCC, 2022: Summary for policymakers. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*, H.-O. Pörtner et al., Eds., Cambridge University Press, 3–34, <https://doi.org/10.1017/9781009325844.001>.
- Jones, P. W., 1999: First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Wea. Rev.*, **127**, 2204–2210, [https://doi.org/10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2).
- Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889, <https://doi.org/10.1175/MWR-D-14-00277.1>.
- MacLeod, D. A., and Coauthors, 2021: Drivers and subseasonal predictability of heavy rainfall in equatorial East Africa and relationship with flood risk. *J. Hydrometeorol.*, **22**, 887–903, <https://doi.org/10.1175/JHM-D-20-0211.1>.
- Maier-Gerber, M., A. H. Fink, M. Riemer, E. Schoemer, C. Fischer, and B. Schulz, 2021: Statistical-dynamical forecasting of subseasonal North Atlantic tropical cyclone occurrence. *Wea. Forecasting*, **36**, 2127–2142, <https://doi.org/10.1175/WAF-D-21-0020.1>.
- Marshall, J. H., N. S. Dickson, L. Garcia-Carreras, G. M. S. Lister, D. J. Parker, P. Knippertz, and C. E. Birch, 2013: The role of moist convection in the West African monsoon system: Insights from continental-scale convection-permitting simulations. *Geophys. Res. Lett.*, **40**, 1843–1849, <https://doi.org/10.1002/grl.50347>.
- Monsieurs, E., and Coauthors, 2018: Evaluating TMPA rainfall over the sparsely gauged East African Rift. *J. Hydrometeorol.*, **19**, 1507–1528, <https://doi.org/10.1175/JHM-D-18-0103.1>.
- Murphy, H. A., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- O, S., and P.-E. Kirstetter, 2018: Evaluation of diurnal variation of GPM IMERG-derived summer precipitation over the contiguous US using MRMS data. *Quart. J. Roy. Meteor. Soc.*, **144**, 270–281, <https://doi.org/10.1002/qj.3218>.
- OCHA, 2020: Eastern Africa Region: Regional floods and locust outbreak snapshot January 2020. Accessed 31 March 2023, <https://reliefweb.int/report/south-sudan/eastern-africa-region-regional-floods-and-locust-outbreak-snapshot-january-2020>.
- Pohl, B., and P. Camberlin, 2006: Influence of the Madden-Julian Oscillation on East African rainfall. I: Intraseasonal variability and regional dependency. *Quart. J. Roy. Meteor. Soc.*, **132**, 2521–2539, <https://doi.org/10.1256/qj.05.104>.
- Rasheeda Satheesh, A., P. Knippertz, A. H. Fink, E.-M. Walz, and T. Gneiting, 2023: Sources of predictability of synoptic-scale rainfall during the West African summer monsoon. *Quart. J. Roy. Meteor. Soc.*, <https://doi.org/10.1002/qj.4581>, in press.
- Schulz, B., and S. Lerch, 2022: Machine learning methods for post-processing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, <https://doi.org/10.1175/MWR-D-21-0150.1>.
- Specq, D., and L. Batté, 2022: Do subseasonal forecasts take advantage of Madden-Julian Oscillation windows of opportunity. *Atmos. Res.*, **23**, e1078, <https://doi.org/10.1002/asl.1078>.
- Stellingwerf, S., E. Riddle, T. M. Hopson, J. C. Kniviel, B. Brown, and M. Gebremichael, 2021: Optimizing precipitation forecasts for hydrological catchment in Ethiopia using statistical bias correction and multi-modeling. *Earth Space Sci.*, **8**, e2019EA000933, <https://doi.org/10.1029/2019EA000933>.
- Toreti, A., and Coauthors, 2022: Drought in East Africa August 2022. Tech. Rep., Publications Office of the European Union, Luxembourg, 28 pp., https://edo.jrc.ec.europa.eu/documents/news/GDODroughtNews202208_East_Africa.pdf.
- Vannitsem, S., D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 347 pp., <https://doi.org/10.1016/C2016-0-03244-8>.
- Vitart, F., 2017: Madden-Julian Oscillation and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220, <https://doi.org/10.1002/qj.3079>.
- , and Coauthors, 2017: The Subseasonal to Seasonal Prediction (S2S) Project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting, 2018: Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Wea. Forecasting*, **33**, 369–388, <https://doi.org/10.1175/WAF-D-17-0127.1>.
- , —, —, —, and —, 2020: Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics. *Wea. Forecasting*, **35**, 2367–2385, <https://doi.org/10.1175/WAF-D-20-0082.1>.
- Walz, E., M. Maranan, R. van der Linden, A. H. Fink, and P. Knippertz, 2021: An IMERG-based optimal extended probabilistic climatology (EPC) as a benchmark ensemble forecast for precipitation in the tropics and subtropics. *Wea. Forecasting*, **36**, 1561–1573, <https://doi.org/10.1175/WAF-D-20-0233.1>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- , 2016: “The stippling shows statistically significant grid points”: How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, <https://doi.org/10.1175/BAMS-D-15-00267.1>.
- Woodhams, J. B., E. C. Birch, H. J. Marshall, L. B. Bain, M. N. Roberts, and F. A. D. Boyd, 2018: What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Mon. Wea. Rev.*, **146**, 2757–2780, <https://doi.org/10.1175/MWR-D-17-0396.1>.
- Youds, L., and Coauthors, 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 Rep., 63 pp., <https://eprints.whiterose.ac.uk/181045/>.
- Zagar, N., 2017: A global perspective of the limits of prediction skill of NWP models. *Tellus*, **16A**, 1317573, <https://doi.org/10.1080/16000870.2017.1317573>.