Physics-Based vs Data-Driven 24-Hour Probabilistic Forecasts of Precipitation for Northern Tropical Africa

EVA-MARIA WALZ,[a,b] PETER KNIPPERTZ,[c] ANDREAS H. FINK,[c] GREGOR KÖHLER,[d] AND TILMANN GNEITING[b,a]

[a] *Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany*
[b] *Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*
[c] *Institute for Meteorology and Climate Research Tropospheric Research (IMKTRO), Karlsruhe Institute of Technology, Karlsruhe, Germany*
[d] *German Cancer Research Center (DKFZ), Heidelberg, Germany*

(Manuscript received 10 January 2024, in final form 21 May 2024, accepted 4 June 2024)

ABSTRACT: Numerical weather prediction (NWP) models struggle to skillfully predict tropical precipitation occurrence and amount, calling for alternative approaches. For instance, it has been shown that fairly simple, purely data-driven logistic regression models for 24-h precipitation occurrence outperform both climatological and NWP forecasts for the West African summer monsoon. More complex neural network–based approaches, however, remain underdeveloped due to the non-Gaussian character of precipitation. In this study, we develop, apply, and evaluate a novel two-stage approach, where we train a U-Net convolutional neural network (CNN) model on gridded rainfall data to obtain a deterministic forecast and then apply the recently developed, nonparametric Easy Uncertainty Quantification (EasyUQ) approach to convert it into a probabilistic forecast. We evaluate CNN+EasyUQ for 1-day-ahead 24-h accumulated precipitation forecasts over northern tropical Africa for 2011–19, with the Integrated Multi-satellitE Retrievals for GPM (IMERG) data serving as ground truth. In the most comprehensive assessment to date, we compare CNN+EasyUQ to state-of-the-art physics-based and data-driven approaches such as monthly probabilistic climatology, raw and postprocessed ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF), and traditional statistical approaches that use up to 25 predictor variables from IMERG and the ERA5 reanalysis. Generally, statistical approaches perform about on par with postprocessed ECMWF ensemble forecasts. The CNN+EasyUQ approach, however, clearly outperforms all competitors in terms of both occurrence and amount. Hybrid methods that merge CNN+EasyUQ and physics-based forecasts show slight further improvement. Thus, the CNN+EasyUQ approach can likely improve operational probabilistic forecasts of rainfall in the tropics and potentially even beyond.

SIGNIFICANCE STATEMENT: Precipitation forecasts in the tropics remain a great challenge despite their enormous potential to create socioeconomic benefits in sectors such as food and energy production. Here, we develop a purely data-driven, machine learning–based prediction model that outperforms traditional, physics-based approaches to 1-day-ahead forecasts of rainfall occurrence and rainfall amount over northern tropical Africa in terms of both forecast skill and computational costs. A combined data-driven and physics-based (hybrid) approach yields further (slight) improvement in terms of forecast skill. These results suggest new avenues to more accurate and more resource-efficient operational precipitation forecasts in the Global South.

KEYWORDS: Africa; Precipitation; Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Statistical forecasting; Neural networks

1. Introduction

Despite the continuous improvement of numerical weather prediction (NWP) models, precipitation forecasts in the tropics remain a great challenge. Several studies (Haiden et al. 2012; Vogel et al. 2020) have shown that NWP models have difficulties in outperforming climatological forecasts. A possible explanation is the exceptionally high degree of convective organization over tropical Africa (Nesbitt et al. 2006; Roca et al. 2014), a process that is difficult to capture with the convective parameterization of NWP models (Vogel et al. 2018) although recent developments show some promise (Becker et al. 2021). Statistical postprocessing, spatial averaging, or temporal aggregation leads to improvements in the skill of raw NWP ensemble gridpoint forecasts in tropical Africa (Vogel et al. 2020; Stellingwerf et al. 2021; Gebremichael et al. 2022; Ageet et al. 2023), yet in regions of particularly poor performance of the operational forecast systems, viz., West and central equatorial Africa, the forecast gain over climatology is limited.

The overall poor performance of current operational systems motivates the development of alternative approaches. Vogel et al. (2020) implement a fairly simple purely data-driven logistic regression model for 24-h precipitation occurrence, which outperforms climatology and NWP forecasts for the summer

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/MWR-D-24-0005.s1.

Corresponding author: Eva-Maria Walz, eva-maria.walz@kit.edu

DOI: 10.1175/MWR-D-24-0005.1

© 2024 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).

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

monsoon season in West Africa. The predictor variables are designed by exploiting spatial–temporal coherence patterns as developed and investigated further in Rasheeda Satheesh et al. (2023). To this end, the rainfall at each grid point is correlated with the rainfall at all other locations from 1, 2, and 3 days before using the coefficient of predictive ability (CPA) measure (Gneiting and Walz 2022). The locations showing highest CPA for 1, 2, and 3 days before, respectively, are selected as predictor variables in the logistic regression model. The good performance of this simple logistic model, which is related to coherent, tropical-wave-driven spatial propagation of precipitation features in West Africa (Rasheeda Satheesh et al. 2023), motivates the development of more sophisticated data-driven models and the usage of additional weather quantities linked to rainfall occurrence and amount.

Vogel et al. (2021) and Rasheeda Satheesh et al. (2023) have only investigated the skill of probability forecasts for the binary problem of precipitation occurrence. In this paper, the more challenging problem of producing accurate probabilistic forecasts for accumulated precipitation, a nonnegative real-valued variable, is considered. Precipitation accumulation is generally considered the "most difficult weather variable to forecast" (Ebert-Uphoff and Hilburn 2023). Indeed, precipitation accumulation follows a mixture distribution with a point mass at zero—namely, for no precipitation—and a continuous part on the positive real numbers. Therefore, despite the sweeping rise of data-driven weather prediction (Ben Bouallègue et al. 2024) and rapid progress in data-driven nowcasting of precipitation (Ayzel et al. 2020; Lagerquist et al. 2021; Ravuri et al. 2021; Schroeder de Witt et al. 2021; Espeholt et al. 2022; Zhang et al. 2023), the development of machine learning–based methods for data-driven probabilistic quantitative precipitation forecasts—at least for lead times larger than 12 h—has been lagging. For example, precipitation was "not investigated" (Bi et al. 2023, p. 537) by the Pangu-Weather team and "left out of the scope" of the GraphCast development because "precipitation is sparse and non-Gaussian and would have possibly required different modeling decisions than the other variables" (Lam et al. 2023, p. 1421). We address these challenges by developing a novel two-stage CNN+EasyUQ approach, where we first train a U-Net convolutional neural network (CNN) model to obtain a single-valued deterministic forecast and then use the Easy Uncertainty Quantification (EasyUQ) approach developed by Walz et al. (2024) to convert the deterministic forecast into a probabilistic forecast.

The paper is structured as follows. Section 2 introduces the data used in the analysis and reviews the methods and metrics used throughout the paper. Then, an overview of weather quantities which are known to be linked to precipitation and thus are candidates for predictor variables is provided in section 3. Different types of forecasting models are described in section 4. Importantly, we compare the CNN+EasyUQ forecasts to a comprehensive suite of state-of-the-art methods that include physics-based raw NWP ensemble forecasts, post-processed NWP forecasts, data-driven statistical forecasts based on logistic regression and distributional (single) index models (DIMs), and combined statistical–dynamical (hybrid) approaches. Results from this comparison are presented in section 5 with the main conclusions and outlook in section 6.

2. Data and methods

In this study, we use data from three different sources. The arguably best currently available high-resolution, gauge-calibrated, gridded precipitation product, the Integrated Multi-satellitE Retrievals for GPM (IMERG; Huffman et al. 2020), serves as ground truth for precipitation. The fifth major global reanalysis produced by European Centre for Medium-Range Weather Forecasts (ECMWF) (ERA5; Hersbach et al. 2020) is used to obtain estimates of other weather quantities. Finally, NWP forecasts, namely, the high-resolution (HRES) run and the full ECMWF ensemble prediction system (EPS), are available from the ECMWF's Meteorological Archival and Retrieval System (MARS; ECMWF 2018).

The evaluation domain, visualized in Fig. 1, is northern tropical Africa, represented by 19×61 grid boxes centered at 0°–18°N and 25°W–35°E, respectively, similar to the setup in Vogel et al. (2020) and Rasheeda Satheesh et al. (2023). Five distinct seasons are considered as identified previously (Fink et al. 2017; Maranan et al. 2018): December–February (DJF), which is the dry season with occasional showers along the Guinea Coast; the March–April (MA) period, which features highly organized mesoscale convective systems (MCSs) at the Guinea Coast and the coastal hinterland; May–June (MJ), the major rainy season along most parts of the Guinea Coast; July–September (JAS), the major rainy season in the Sahel and the little dry season at the coast; and October–November (ON), the second, weaker rainy season at the Guinea Coast. To avoid cutting seasonal periods at the beginning or the end of the time period under investigation, the time period considered starts 1 December 2000 and ends 30 November 2019, with 24-h forecasts of precipitation amount and precipitation occurrence for 1 December 2010–30 November 2019 being evaluated. Importantly, the analysis and evaluation are performed over land only, and we identify a grid box with the grid point at its center. From now on, when we refer to grid boxes or grid points, we only mean boxes or points on land.

a. GPM IMERG rainfall data

We use the GPM IMERG V06B final version (Hou et al. 2014; Huffman et al. 2020) to calculate 24-h accumulated precipitation from 0600 to 0600 UTC for the period under investigation. GPM IMERG has a temporal resolution of 30 min and a spatial resolution of $0.1° \times 0.1°$. The data were regridded to a resolution of $1° \times 1°$ using first-order conservative remapping. As we also consider 24-h rainfall occurrence, we threshold at 0.2 mm to obtain a binary event variable representing precipitation occurrence.

The GPM IMERG algorithm uses both radar-calibrated microwave radiance from polar-orbiting satellites and infrared radiance from geostationary satellites. In the final version, the precipitation totals are calibrated with rain gauge measurements provided by the Global Precipitation Climatology Centre (GPCC; Schneider et al. 2015). The degree to which the original estimates are adjusted by the gauge calibration process within a given region is

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

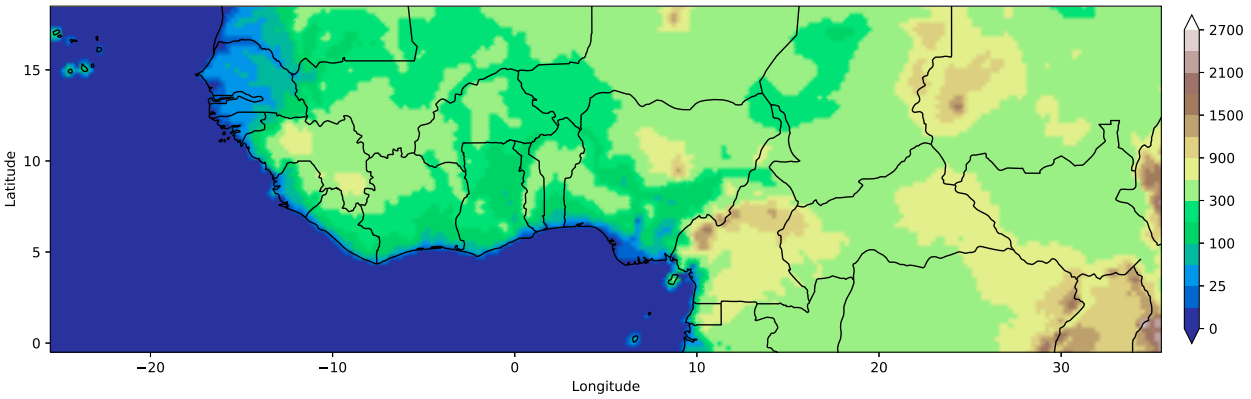

FIG. 1. Overview of the study area. Following Rasheeda Satheesh et al. (2023), we consider an evaluation domain over northern tropical Africa that comprises 19 × 61 grid boxes with centers spanning from 0° to 18°N in latitude and 25°W to 35°E in longitude. The analysis is over land only, and shading indicates altitude in meters, based on the ERA5 land–sea mask.

generally determined by the number of available rain gauges, which is highly variable across the tropics.

b. *Predictor variables from ERA5*

Our study considers a range of meteorological variables, specified in section 3b, as predictor variables for statistical models. Specifically, we use the ERA5 reanalysis (Hersbach et al. 2020), which provides a complete and consistent coverage of the study domain by combining model data with observations. For this study, the resolution of the data is 1° × 1° just like for GPM IMERG. In contrast to 24-h accumulated precipitation, the considered ERA5 weather quantities are instantaneous values at 0000 UTC, thus 6 h before the 24-h accumulation period for GPM IMERG starts. This way, observed ambient conditions well before the rainfall begins get considered. For an operational implementation of the respective statistical methods, operational analysis data would need to be used, as ERA5 is not available in near–real time. We do not expect this to make a big difference to our results since the operational analysis and ERA5 are both produced with ECMWF's IFS and therefore are quite similar. Furthermore, while it is true that NWP models can react sensitively to changes in initial conditions, our statistical models mostly pick up on "well behaved" smooth fields such as column water vapor.

c. *Physics-based forecasts from ECMWF*

We now describe the NWP forecasts used in this study, namely, the ECMWF HRES model and EPS (Molteni et al. 1996). Owing to the high resolution and the initialization with the most accurate analysis product, the HRES model is arguably the leading global deterministic NWP forecast available. As an operational product, HRES has changed considerably over time in frequent updates (ECMWF 2023a,b). The ECMWF EPS consists of one control run and 50 perturbed members. Like the HRES model, the control run is based on the most accurate initial state of the atmosphere. The perturbed members start from slightly different initial conditions and use perturbed physics options. We use the operational EPS rather than the ECMWF reforecast ensemble, as the latter is only available twice a week and has fewer members (ECMWF 2021).

The forecasts are available from MARS in a grid resolution of 0.25° × 0.25° and are first order conservatively remapped to a resolution of 1° × 1°. HRES forecasts for total precipitation are obtained by summing forecasts for large-scale precipitation and convective precipitation, which are available since April 2001. For the EPS, total precipitation is available since April 2006. To cover an equal number of seasons, we use data starting in December 2001 and December 2006, respectively. To obtain forecasts for 24-h precipitation amount, the difference between forecasts of accumulated precipitation initialized at 0000 UTC with lead times of 30 and 6 h is computed. To compute the EPS forecast probability for the occurrence of precipitation, the member forecasts are thresholded at 0.2 mm, and the respective binary outcomes are averaged.

We proceed to review statistical methods and evaluation metrics used throughout the paper.

d. *EasyUQ*

Forecasts ought to take the form of probability distributions to account for uncertainty. In NWP, probabilistic forecasts have become common practice with the operational implementation of ensemble systems (Molteni et al. 1996; Bauer et al. 2015). To quantify uncertainty in very general settings, Walz et al. (2024) introduced EasyUQ, an easy-to-implement method which transforms real-valued deterministic model output into calibrated statistical distributions. EasyUQ is trained on pairs of deterministic forecasts and corresponding outcomes and is thus independent of the type of model used to generate the single-valued forecasts. In particular, EasyUQ can be applied to the output of any NWP, statistical, or machine learning model that generates deterministic forecasts. The EasyUQ forecast distributions are discrete and have mass exclusively at the observation values in the training set. Precipitation accumulation follows a mixture distribution with a point mass at zero and a continuous part on the positive real numbers. However, rainfall amounts typically are reported in small but fixed increments, so strictly speaking, the distribution of the observation

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

values on the positive real numbers is discrete as well. The EasyUQ forecast distributions adapt naturally to the level of discretization in the observation values in the training set, without any need for tuning.

In its basic form, which we use in this study, EasyUQ is a special case of isotonic distributional regression (IDR; Henzi et al. 2021). In contrast to NWP ensemble systems, which have high computational costs and require the use of supercomputers (Bauer et al. 2015), the application of EasyUQ to deterministic model output has obvious advantages in terms of the efficient usage of computational resources (Walz et al. 2024, section 3.4).

e. Evaluation metrics

In our study, we compare methods for probabilistic forecasts of precipitation amount, where the outcome is real-valued, and probability forecasts of precipitation occurrence, where the outcome is binary. In both cases, we follow extant practice and use proper scoring rules (Gneiting and Raftery 2007).

In the setting of probability forecasts, we use the Brier score (BS) to quantify predictive performance based on a collection of pairs $(p_1, y_1), \ldots, (p_n, y_n)$ of predictive probabilities and associated binary outcomes. Specifically, we compute the mean score as follows:

$$\overline{\mathrm{BS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{BS}(p_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} (p_i - y_i)^2. \tag{1}$$

In the case of precipitation amount, we use the continuous ranked probability score (CRPS) for an assessment based on a collection of pairs $(F_1, y_1), \ldots, (F_n, y_n)$ of probabilistic forecasts and associated real-valued outcomes. Comparisons are in terms of the mean score

$$\overline{\mathrm{CRPS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} [F_i(z) - 1\{z \geq y_i\}]^2 \, dz, \tag{2}$$

where F_i is interpreted as a cumulative distribution function. To facilitate the assessment of forecast performance relative to a baseline, skill scores can be used, defined as the quantity $(\overline{S}_{\mathrm{base}} - \overline{S}_{\mathrm{fcst}})/\overline{S}_{\mathrm{base}}$, where $\overline{S}_{\mathrm{fcst}}$ is the mean score of the forecast at hand and $\overline{S}_{\mathrm{base}}$ is the mean score of the baseline. A positive Brier or CRPS skill score corresponds to predictive performance better than the baseline; a negative Brier or CRPS skill score corresponds to predictive performance worse than the baseline.

For a more informative, diagnostic comparison between forecast methods, we apply the consistency, optimalilty, reproducibility, and pool-adjacent-violators algorithm-based (CORP) decomposition of Dimitriadis et al. (2021) and the isotonicity-based decomposition of Arnold et al. (2023) to a mean score $\overline{S}$ from Eq. (1) or Eq. (2), respectively. The decompositions express $\overline{S}$ in terms of interpretable components, in that

$$\overline{S} = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC}, \tag{3}$$

where the miscalibration (MCB) component quantifies the (lack of) calibration or reliability of the forecasts (the lower, the better), the discrimination (DSC) term refers to the discrimination ability or resolution of the forecasts (the higher, the better), and the uncertainty (UNC) component is independent of the forecasts and a property of the outcomes only. For details, we refer to the original work of Dimitriadis et al. (2021) and Arnold et al. (2023).

3. Predictor variables for data-driven forecasts

In this section, we discuss and analyze potential predictor variables for data-driven forecasting methods. We distinguish predictor variables computed from IMERG data based on spatiotemporal rainfall correlation and predictor variables based on ERA5. The initial selection of the variables stems from meteorological expertise.

a. Correlated rainfall predictors from IMERG

Vogel et al. (2021) introduced a logistic regression model to produce probability forecasts for the binary outcome of precipitation occurrence. As predictors, they used precipitation data with a lag of 1 and 2 days at locations with maximum positive and minimum negative Spearman's rank correlation coefficient. Rasheeda Satheesh et al. (2023) noted that due to propagating rainfall systems, positive dependencies carry the most useful information, occasionally reaching 3 days backward in time. Moreover, they suggested a replacement of Spearman's rank correlation coefficient by the recently developed CPA (Gneiting and Walz 2022) measure. In general, CPA is asymmetric, with the predictor variable and the outcome taking clearly identified roles, as for the classical area under the receiver operating characteristic (ROC) curve (AUC) measure, to which CPA reduces when the outcomes are binary. When both the predictor variable and the outcome are continuous variables, CPA becomes symmetric and equals Spearman's rank correlation coefficient, up to a linear transformation (Gneiting and Walz 2022). AUC or CPA values above 0.5 correspond to positive dependencies, and values below 0.5 correspond to negative dependencies.

Given these insights, the statistical models in section 4c use three correlated precipitation predictor variables, by identifying the grid points with maximum CPA at a temporal lag of 1, 2, and 3 days, respectively. Following Rasheeda Satheesh et al. (2023), correlated locations are identified within an enlarged region that comprises 68°W–50°E and 0°–20°N, as compared to the evaluation domain depicted in Fig. 1, which ranges from 25°W to 35°E and from 0° to 18°N.

b. Predictor variables from ERA5 reanalysis

In addition to the correlated precipitation information, various meteorological variables from ERA5 are considered predictors (Table 1). For a summary of how environmental conditions affect convection, see Maranan et al. (2018). Unless noted otherwise, the variables are instantaneous quantities at 0000 UTC. The first four variables in Table 1 are vertically integrated measures of water in different forms. Total column water vapor (TCWV) has been shown to be a promising predictor for precipitation by Lafore et al. (2017a); Schroeder de Witt et al. (2021) use cloud information such as

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

TABLE 1. Predictor variables from ERA5, all at 0000 UTC.

Meteorological variable	Acronym
Total column water vapor	TCWV
Vertically integrated moisture divergence	VIMD
Total column cloud liquid water	TCLW
Total cloud cover	TCC
Convective available potential energy	CAPE
Convective inhibition	CIN
K index	KX
2-m temperature	T2
2-m dewpoint temperature	D2
24-h surface pressure tendency	SPT
Temperature at 850 hPa	T850
Temperature at 500 hPa	T500
Specific humidity at 925 hPa	Q925
Specific humidity at 700 hPa	Q700
Specific humidity at 600 hPa	Q600
Specific humidity at 500 hPa	Q500
Relative humidity at 500 hPa	R500
Relative humidity at 300 hPa	R300
Shear	SHR
Streamfunction at 700 hPa	Ψ700

total column cloud liquid water (TCLW) and total cloud cover (TCC) in their global statistical model. The second group comprises the three classical measures of convective instability: convective available potential energy (CAPE; the theoretical maximum of thermodynamic energy that can be converted into kinetic energy of vertical motion), convective inhibition (CIN; the energy barrier that needs to be overcome to reach the level of free convection), and *K* index (KX; based on dry static vertical stability in the 850–500-hPa layer, absolute humidity at 850 hPa, and relative humidity at 700 hPa). CAPE and CIN have a complex relationship with precipitation and should be considered together and in concert with other parameters (Lafore et al. 2017b). Galvin (2010) demonstrates the usefulness of KX in assessing convective rainfall probability in relation to African easterly waves (AEWs).

The third group [2-m temperature (T2), 2-m dewpoint temperature (D2), and 24-h surface pressure tendency (SPT)] represents near-surface conditions. The former two are closely related to the equivalent potential temperature of a starting convective air parcel, thereby influencing the level of cumulus condensation and free convection and thus CIN and CAPE, and have been shown to impact the intensity of convection in West Africa (Nicholls and Mohr 2010). SPT, the tendency from 0000 UTC of the day for which the prediction is made to 0000 UTC of the previous day, can be related to AEW propagation and rainfall (Regula 1936; Hubert 1939). The fourth group characterizes thermodynamic conditions in the boundary layer and free troposphere between 925 and 300 hPa. For temperature, we consider 850 and 500 hPa representing lower-tropospheric stability (as in KX). As moisture generally shows complex vertical structures, 925, 700, 600, and 500 hPa are chosen for specific humidity. For relative humidity, the mid- to upper-tropospheric levels of 500 and 300 hPa were selected to indicate deep moistening, which facilitates cloud formation and reduces the detrimental effects of entrainment on convective development. Midtropospheric relative humidity controls both rainfall enhancement by slow-moving tropical waves (Schlueter et al. 2019) and evaporation of rainfall and thus convective downdrafts and mesoscale organization of convection (Klein et al. 2021). The last two entries in Table 1 are the circulation-related variables shear (SHR; normalized difference of horizontal wind at 600 and 925 hPa) and streamfunction at 700 hPa (Ψ700) representing midtropospheric streamlines. SHR influences the potential for mesoscale organization and longevity through separating the areas of convective updrafts and downdrafts as well as the generation of cold pools (Rotunno et al. 1988; Lafore et al. 2017a). Anomalies in Ψ700 indicate variations in the African easterly jet (AEJ), e.g., passages of troughs and ridges of AEWs (Kiladis et al. 2006).

c. *Statistical analysis of predictor variables*

Thus far, the selection of predictor variables has been based on meteorological expertise and findings from other publications. Here, we use the aforementioned AUC (for rainfall occurrence) and CPA (for amount) measures of Gneiting and Walz (2022) (see section 3a) for a deeper analysis. In Figs. 2 and 3, we show AUC and CPA values for the 20 ERA5 variables from Table 1. Both are computed in a collocated fashion for each grid point in the evaluation domain (Fig. 1), and the resulting distributions are represented by boxplots.

Figure 2a shows AUC values for the dry season DJF. Given the overall low precipitation amount during this period, the boxplots often stretch over large ranges, indicating marked differences between grid points, and also large differences between the variables. Stable positive relations (i.e., AUC above 0.5) are found for moisture [TCWV and specific humidity at 500, 600, and 700 hPa (Q500, Q600, and Q700, respectively); and relative humidity at 500 hPa (R500)], cloud (TCLW and TCC), and instability variables (KX and CAPE), demonstrating a clear dependence on midtropospheric conditions, while low-level [specific humidity at 925 hPa (Q925) and D2] and upper-level [relative humidity at 300 hPa (R300)] variables show a more ambiguous behavior. Other well-defined relations are positive with T2 and negative with temperature at 500 hPa (T500) and vertically integrated moisture divergence (VIMD). As the variables are taken at 0000 UTC, the relation to T2 may reflect warmer nights under moister and cloudier skies. CIN, SPT, and Ψ700 show weak AUC values close to 0.5. AUC values for temperature at 850 hPa (T850) cover a wide range and stretch across 0.5, indicating that its impact depends strongly on the situation.

The corresponding analysis for MA (Fig. 2b) shows an overall less noisy behavior and AUC values more in line with the spatially averaged annual value of CPA that determines the order of the variables in all panels of Figs. 2 and 3. Compared to DJF, a more stable relation to low-level moisture (Q700, Q925, and D2) is visible. There is a stronger relation to CAPE with little changes in CIN. Other remarkable changes are less dependence on cold T500 and even more ambiguous relations to T850 and Ψ700. The premonsoon season MJ (Fig. 2c), when rainfalls begin to move inland, shows many similarities to MA, but the point-to-point variability is

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

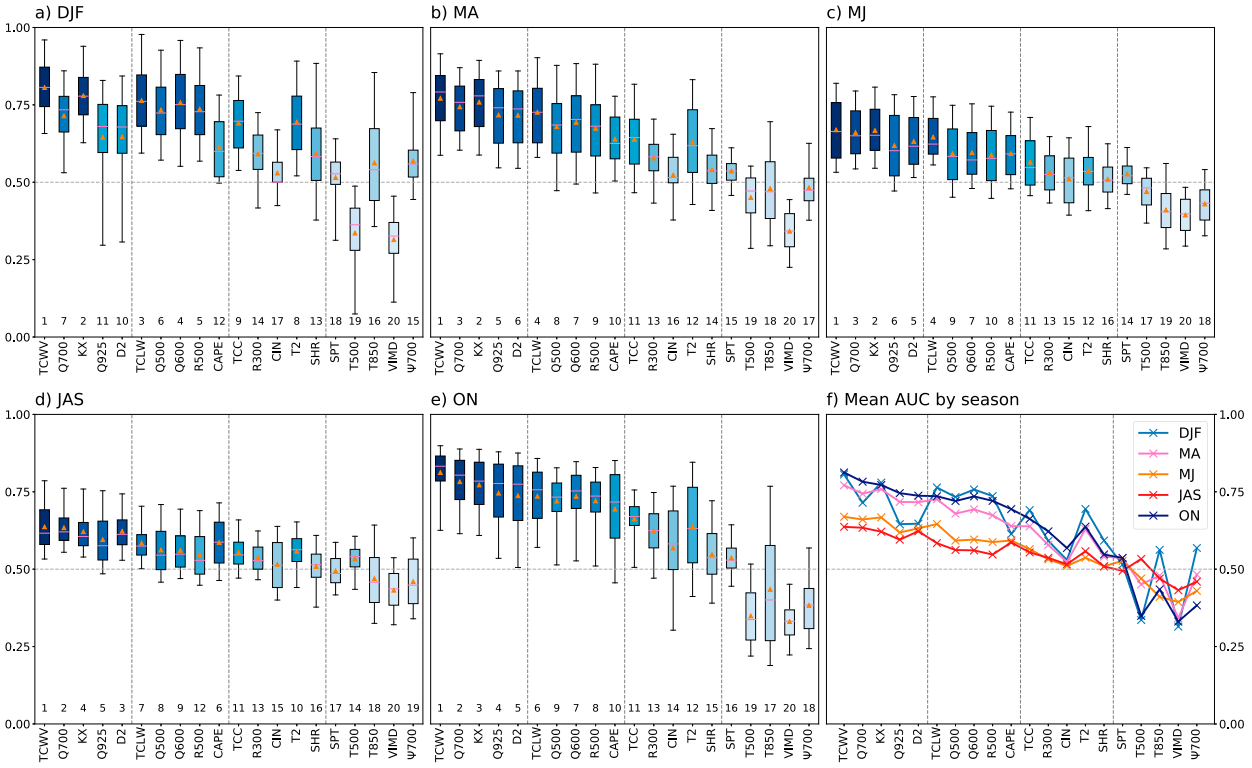

FIG. 2. Boxplots of gridpoint AUC values between ERA5 variables from Table 1 and precipitation occurrence in season (a) DJF, (b) MA, (c) MJ, (d) JAS, and (e) ON. The arrangement of the predictor variables on the horizontal axis is in the order of the spatially averaged CPA value for precipitation accumulation when CPA is computed without splitting into seasons. The orange marks in (a)–(e) and the line plots in (f) indicate the mean AUC value over grid points for the season at hand. The box color from dark to light blue indicates the rank from high to low of the seasonal mean AUC value, as shown beneath the box. In combination, this information allows us to identify differences between yearly vs seasonal perspectives.

smaller and AUC values tend to be closer to 0.5, while their order mostly agrees to that based on annual CPAs. Remarkable differences to MA are less dependence on T2 and clearer relations to T850 and Ψ700 (<0.5). The latter may indicate a dependence of rainfall on the existence of cyclonic perturbations such as AEWs. The general magnitude of AUC values close to 0.5 is likely a reflection of the overall improved conditions for convection, which makes individual storms less dependent on particular circumstances, thereby creating a higher degree of stochasticity [see also discussion in Rasheeda Satheesh et al. (2023)]. This trend continues going into the main monsoon season JAS (Fig. 2d) when most variables show AUC values close to 0.5. Diminished ranges in the boxplots indicate less local variability during a period when rains penetrate deeply into the continent. As expected, in the postmonsoon season ON (Fig. 2e), conditions resemble those discussed for MA (Fig. 2b), even with slightly larger amplitudes. Remarkable differences to MA are that rainfall occurrence depends more on CIN and T2, possibly because, in ON, the solar angle is already flatter and the daytime heating is further dampened by the higher moisture availability after the rainy season. As for DJF, rain depends on cold T500, and the relation to T850 is highly variable and can take both directions, however, with a clear tendency to cooler conditions when rain occurs. ON also shows the clearest relation to cyclonic perturbations as reflected in AUC values below 0.5 for Ψ700. These may grow in importance relative to other mechanisms, as triggering by daytime heating weakens. Finally, Fig. 2f shows a summary plot of mean AUC values for all five seasons. This plot underlines the similar behavior of MJ and JAS (with a consistently higher amplitude for MJ), as well as of MA and ON (with a consistently higher amplitude for ON). DJF often shows the highest magnitude, as rain depends strongly on unusual conditions to occur, but given the many dry days, the overall behavior appears quite noisy.

The corresponding analysis for CPA is shown in Fig. 3. Overall, there are many similarities to Fig. 2, indicating that variables that work as predictors for occurrence also work for amount. This is particularly true for the wet part of the year (MJ, JAS, and ON), where plots look largely identical (Figs. 3c–e). For MA (Fig. 3b), there is still large agreement across all variables, but the magnitude of CPA values is smaller and the boxplots have smaller ranges than for AUC. This indicates that in this somewhat marginal rainfall season, the amount is harder to predict than occurrence. This trend is even more evident for the dry DJF season (Fig. 3a) when some boxplots show a very small range and magnitudes fall underneath those of ON on average, as shown by the summary plot (Fig. 3f).

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

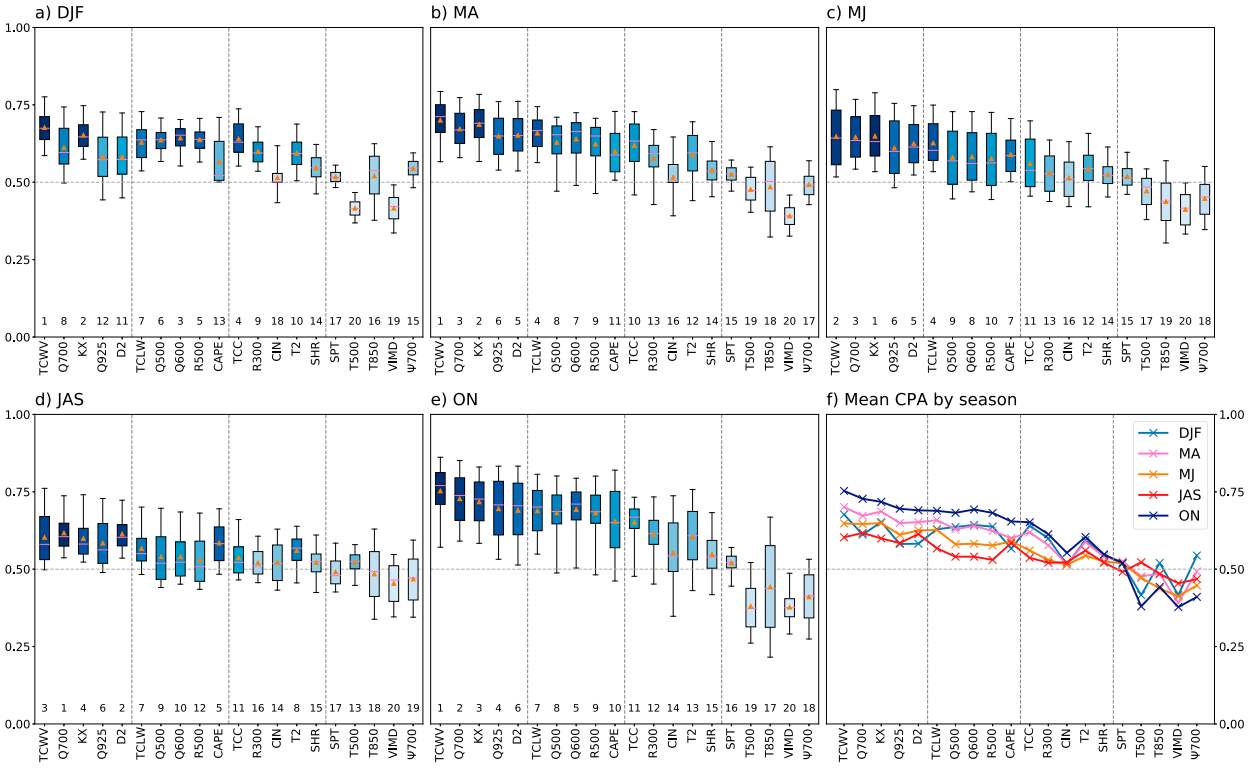

FIG. 3. As in Fig. 2, but for CPA and precipitation amount.

For an improved understanding of the ranges in the boxplots, Fig. 4 shows the spatial pattern of CPA of selected meteorological variables exemplarily for the peak monsoon season JAS. Consistent with the leftmost boxplot in Fig. 3d, CPA values for TCWV are at or above 0.5 almost everywhere in the study region (Fig. 4a) while featuring an interesting three-tier structure. Over northern parts of the domain, where moisture is a general limiting factor, CPA values are high, especially over the dry eastern Sahel. Further south, along the main rain belt and stretching into the Congo basin, CPA values are close to 0.5, indicating limitations through convective triggering or stability rather than moisture availability. To the south of the rain belt, i.e., along the Guinea Coast and over the East African highlands, moisture appears to become a limiting factor again. A very similar pattern but with a smaller range emerges for KX (Fig. 4b). The largest differences to TCWV are found along the Guinea Coast, where conditions are often close to moist neutral requiring a lifting mechanism to produce rain (cf. Fig. 1.31 in Fink et al. 2017). Similar but slightly northward-shifted structures are found for CAPE (Fig. S1c in the online supplemental material).

A much larger range (0.35–0.75) but with a similar three-tier structure is found for R500 and CIN (Figs. 4c,d). One would expect that a moister midtroposphere and less convective inhibition (recall that CIN is negatively oriented) enhances rainfall amounts, so the behavior within the rain belt is somewhat counterintuitive. The most likely explanation is that in areas of abundant moisture and often neutral stratification, large rainfall amounts can most effectively be generated by organized convective systems that require some barrier to accumulate CAPE over the following day and a relatively dry midtroposphere to allow rainfall evaporation and downdrafts, which in turn can trigger new convection through cold pools [cf. Table 11.2 in Lafore et al. (2017b)]. It is interesting to note that CPA for TCC is similarly structured as R500, showing CPA values well below 0.5 in the rain belt (Fig. S1a). Finally, CPA values for T850 and Ψ700 are both characterized by a marked north–south division around 12°N (Figs. S1e,f). The patterns indicate that in the north, high rainfall amounts are accompanied by lower T850, likely indicating a northward progression of the moist and cool monsoon layer, while in the south, warm air at 850 hPa may indicate more instability on the following day. With respect to Ψ700 low values in the north indicate that rainfall is accompanied by more cyclonic conditions, likely due to the trough passage of AEWs, while in the south, weak anticyclonic conditions prevail.

Most meteorological variables from Table 1 show spatial patterns akin to those in Fig. 4 though some feature hard-to-interpret local signals that entail a wider range of CPA values (e.g., T2 and SHR; see Figs. S1b,d). It is also worth mentioning that corresponding spatial structures for AUC largely agree with CPA (not shown). Comparing JAS with the other four seasons, we find a high consistency in the discussed patterns that largely shift northward and southward with the seasonal evolution of the West African monsoon system (not shown).

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

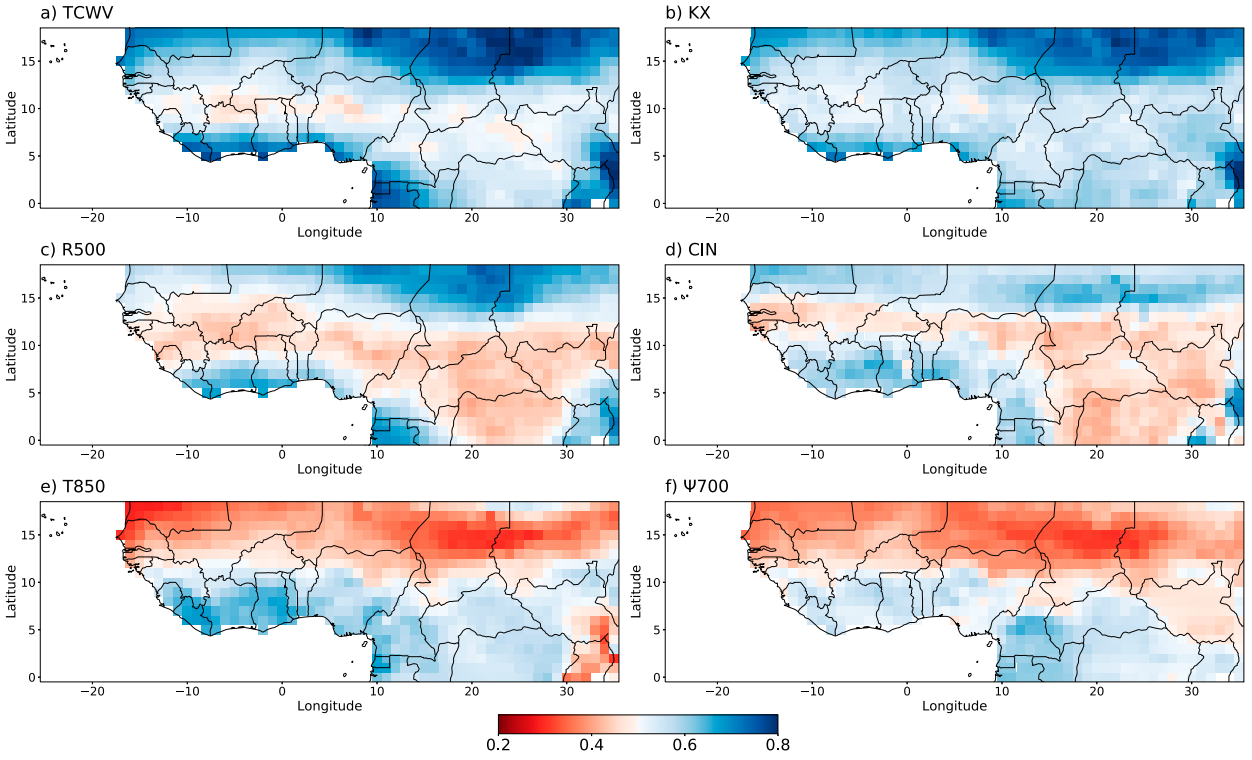

FIG. 4. Spatial pattern of CPA between the ERA5 predictors (a) TCWV, (b) KX, (c) R500, (d) CIN, (e) T850, and (f) Ψ700 from Table 1 and precipitation amount in season JAS.

For the construction of statistical models, correlations between predictor variables matter, as they hinder interpretation and may yield unstable statistical parameters. Figure 5 visualizes Spearman's rank correlation coefficients in JAS for the 20 predictor variables from Table 1. We compute Spearman's coefficient at each grid point and then average over grid points. Note that here, we want the correlation coefficient to be symmetric (in contrast to the asymmetric relation between target and predictor variables). This analysis has been conducted for all five seasons (Fig. S2), but due to the large similarities between them, we discuss the peak monsoon season JAS only. We return to these issues in sections 4c and 5a, where we report on variable selection for our statistical models.

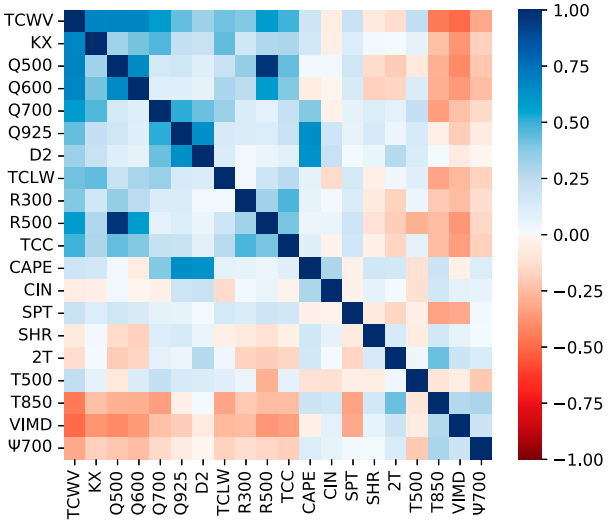

FIG. 5. Spatially averaged Spearman's rank correlation coefficient in season JAS between the ERA5 variables from Table 1.

Not surprisingly, there are generally high correlations between the moisture variables (TCWV, Q500, Q600, Q700, Q925, D2, TCLW, R500, R300, and TCC), and it is noteworthy that R500 is more strongly correlated to Q500 than T500. KX and CAPE show considerably different patterns, with KX being highly correlated with the moisture variables but surprisingly also associated with cold T850, which to some extent counteracts the impact of moister conditions. CAPE is most sensitive to low-level moisture and associated with warm T850, as does CIN but to a smaller degree. A positive SPT is weakly associated with a moister and warmer atmosphere, consistent with the southerly flow behind an AEW trough, where the moister atmosphere suppresses longwave cooling. SHR, T2, and T500 show overall weak and unsystematic correlations, in agreement with the difficult-to-interpret spatial patterns for CPA discussed above. Finally, T850, VIMD, and Ψ700 are consistently negatively correlated with the moisture variables and KX, with the exception of D2. While the relation to VIMD is straightforward, T850 may indicate north–south movements of the monsoon layer, bringing overall moister or drier conditions. The negative correlation between

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

TABLE 2. Overview of probabilistic forecast methods for precipitation occurrence and/or accumulation, including general type, brief description, acronym, and availability of training data. Methods marked with an asterisk (*) yield PoP occurrence forecasts only; for methods marked with "**," we do not present results for PoP forecasts. The final column notes from which year and month onward training data are available and used.

Type	Description	Acronym	Training
Climatological	Monthly probabilistic climatology	MPC	December 2000
Physics-based	ECMWF ensemble prediction system	EPS	—
	Isotonic regression applied to EPS	EPS+ISO*	December 2006
	EMOS applied to EPS	EPS+EMOS**	December 2006
	EasyUQ applied to HRES	HRES+EasyUQ	December 2001
Statistical	Logistic regression, baseline (5 predictors)	Logit-base*	December 2000
	Logistic regression, full model (25 predictors)	Logit-full*	December 2000
	Distributional index model, baseline (5 predictors)	DIM-base**	December 2000
	Distributional index model, full model (25 predictors)	DIM-full**	December 2000
Machine learning	EasyUQ applied to convolutional neural network	CNN+EasyUQ	December 2000
Hybrid	Mixture of HRES+EasyUQ and CNN+EasyUQ	Hybrid	—

moisture variables and Ψ700 reflects the wet conditions associated with cyclonic disturbances, e.g., AEW troughs or vortices.

4. Physics-based and data-driven forecast methods

Forecasts for precipitation occurrence and precipitation amount ought to be probabilistic to account for the chaotic nature of the atmosphere; thus, for the former, they should output a probability of precipitation (PoP), and for the latter, they should output a probability distribution. We investigate forecasts for precipitation occurrence and precipitation amount separately, which allows us to connect our results to Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), where the binary setting was considered only. Furthermore, we can compare between the comparably easy task of producing PoP forecasts and the more challenging task of constructing probabilistic forecasts for precipitation amount. To assess the skill of statistical and machine learning models, it is essential to use baseline models to compare their forecast performance. In the following subsections, different types of forecasting models are presented that are physics-based NWP models, purely data-driven statistical or machine learning techniques, or mixtures of both. Table 2 provides an overview of all considered approaches.

As discussed in section 2, our evaluation period for 24-h forecasts of precipitation amount and precipitation occurrence ranges from 1 December 2010 to 30 November 2019. The DJF season runs across two subsequent calendar years, and we generally assign it to the second year. Reporting a yearly seasonal or overall mean instead of a single mean score over the complete evaluation period allows for a more distinct comparison between forecasting models and provides insights into the temporal evolution of forecast skill.

Except for the ECMWF EPS, all types of forecasting methods require training data and some form of training procedure. The hybrid model combines the CNN+EasyUQ and HRES+EasyUQ forecasts in a way that does not require additional training, but the constituent models do require training. In this study, we use annually growing, expanding training sets that resemble operational settings, where only past data are available. Nonparametric statistical methods such as IDR and especially machine learning approaches benefit from having as much (relevant) training data available as possible. Subject to this caveat, the predictive performance generally does not depend very much on the details of the training scheme. For example, we also implemented the EPS+EMOS technique using a rolling training period of the most recent 730 days and obtained similar results. In our expanding training setting, the initial training period ranges from the first day of the month in the right-most column of Table 2 (hereinafter the start date) to 30 November 2010, and the thus trained methods are used to generate day-ahead 24-h forecasts for the period from 1 December 2010 to 30 November 2011. Then we successively add one more year to the training period, ranging now from the start date through 30 November in the year 2010 + x, and use the thus trained methods to generate forecasts for the 12-month period that begins on 1 December in the year 2010 + x, where $x \in \{1, \ldots, 8\}$. This procedure is followed until training is on data through 30 November 2018, and the thus trained methods are used to generate forecasts for 1 December 2018 through 30 November 2019. Thus, there are nine evaluation folds in total, which we associate with calendar years 2011–2019, respectively.

a. Climatological forecasts

Arguably, the simplest possible type of probabilistic forecast is a climatology constructed from past observations. Here, we use GPM IMERG to construct a monthly probabilistic climatology (MPC). The MPC forecast for a specific valid date is an ensemble constructed by using all past observations from the month at hand. For example, for a test date in January 2014, the MPC forecast is constructed based on data from January 2001 to 2013, which yields an ensemble of size $31 \times 13 = 403$. To obtain the MPC PoP forecast, the relative frequency of ensemble members with rainfall exceeding 0.2 mm is computed.

We also explored persistence as the most basic of all possible forecasts; however, the performance (not shown) was poor in our setting.

b. Physics-based forecasts

Our comparison includes raw and postprocessed probabilistic forecasts from physics-based NWP models run by the

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

ECMWF (section 2c). The postprocessed forecasts require training, for which we use expanding training sets with start dates listed in Table 2 as described above. Training is performed at each grid point individually.

1) Operational ECMWF NWP ensemble

The operational ECMWF EPS comprises 51 NWP runs, namely, a control member and 50 perturbed members. Just as for the climatological MPC approach, the EPS PoP forecast is the relative frequency of members that exceeds 0.2 mm.

2) Statistically postprocessed ECMWF NWP ensemble

Statistical postprocessing is used to correct for systematic biases in raw ensemble forecasts. Here, we use ensemble model output statistics (EMOS), originally developed by Gneiting et al. (2005), to generate full predictive probability distributions by linking ensemble information to distributional parameters. The optimal coefficients are found by optimizing a performance metric on training data.

In the binary case, we recalibrate the EPS PoP by using nonparametric isotonic regression (Zadrozny and Elkan 2002), here referred to as EPS+ISO. For precipitation amount, we apply the EMOS technique proposed by Scheuerer (2014) which models positive rainfall accumulations with generalized extreme value distributions, to generate the EPS+EMOS forecast. While EPS+EMOS induces a PoP forecast, the predictive performance is very similar, though typically slightly inferior, to EPS+ISO. Therefore, we do not report results for the respective PoP forecasts (cf. Table 2).

3) EasyUQ on the HRES model

The HRES model from ECMWF generates a deterministic NWP forecast. We use the EasyUQ technique, introduced in section 2d, to transform this single-valued forecast into a postprocessed predictive distribution, to yield the HRES+EasyUQ forecast.

c. *Statistical forecasts*

Statistical approaches use training data to learn relationships between a target variable and one or more predictor variables. Here, the target variable is the precipitation amount at a given grid point, which in the case of precipitation occurrence is thresholded at 0.2 mm. We use logistic regression to obtain PoP forecasts and DIMs (Henzi et al. 2023) for probabilistic forecasts of precipitation amount, based on predictor variables from section 3. Statistical models require training, and we use annually expanding training sets with a start date in December 2000 (Table 2) as described above. Training is performed at each grid point individually.

The analysis in section 3 provides a thorough understanding of the influence of the selected variables from Table 1 on precipitation occurrence and amount and enables us to link them to typical seasonal weather phenomena. However, overall, the effect of meteorological variables on precipitation is similar across seasons when taking into account the latitudinal shifts associated with the monsoon system. As a consequence, we found little difference in model performance between fitting models on seasonal data versus the whole available training period, as temporal effects such as seasonal changes can be captured by predictor variables that encode the day of the year. Therefore, instead of fitting seasonal models, we train models that apply year-round.

We distinguish baseline models with two predictors that encode the day of the year and three correlated rainfall predictors (section 3a) from full models that additionally use 20 predictor variables from the ERA5 (section 3b). To prevent a statistical model from overfitting, regularization techniques can be applied. However, in this experiment, the performance of the statistical models, which use modest numbers of at most 25 predictor variables only, does not improve when using the regularization techniques we tested. Consequently, we refrain from performing any feature selection beyond the choices made in section 3, which were driven by meteorological expertise and extant literature in atmospheric physics. We note that this strategy is well in line with extant experience in weather prediction, where the use of highly correlated predictor variables typically yields slight only, if any, degradation, of the predictive performance. Tables 3 and 7 in Raftery et al. (2005) provide a striking illustration of this phenomenon. For further analysis and an experiment with fewer, less dependent feature variables, see section 5a.

1) Logistic regression

We use logistic regression (Logit) models to generate statistical PoP forecasts. Specifically, let m be the number of predictor variables, which we denote by $x_1, \ldots, x_m$, and let p be the PoP forecast. The logistic regression model then is of the form

$$\text{logit}(p) = \log\frac{p}{1-p} = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_j, \tag{4}$$

where the statistical coefficients $\alpha_0, \alpha_1, \ldots, \alpha_m$ are estimated from training data. Our baseline model (Logit-base) originates from Vogel et al. (2021) and Rasheeda Satheesh et al. (2023) and uses $m = 5$ predictor variables, namely, three correlated rainfall predictors x_1, x_2, and x_3 at temporal lags of 1, 2, and 3 days, respectively, as described in section 3a, and two variables $x_4 = \sin(2\pi d/365)$ and $x_5 = \cos(2\pi d/365)$ that depend solely on the day of the year d. The full model (Logit-full) extends to $m = 25$ predictor variables in Eq. (4), now including the 20 ERA5 variables from Table 1.

2) Distributional index models

To produce probabilistic forecasts for accumulated precipitation, we use the DIM approach introduced by Henzi et al. (2023), which combines the classical single index model with IDR (Henzi et al. 2021). In a nutshell, an index is learned that represents the conditional mean of the target variable, here log-transformed precipitation accumulation, and then a predictive distribution is estimated nonparametrically under a stochastic ordering constraint. As before, let $x_1, \ldots, x_m$ be

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

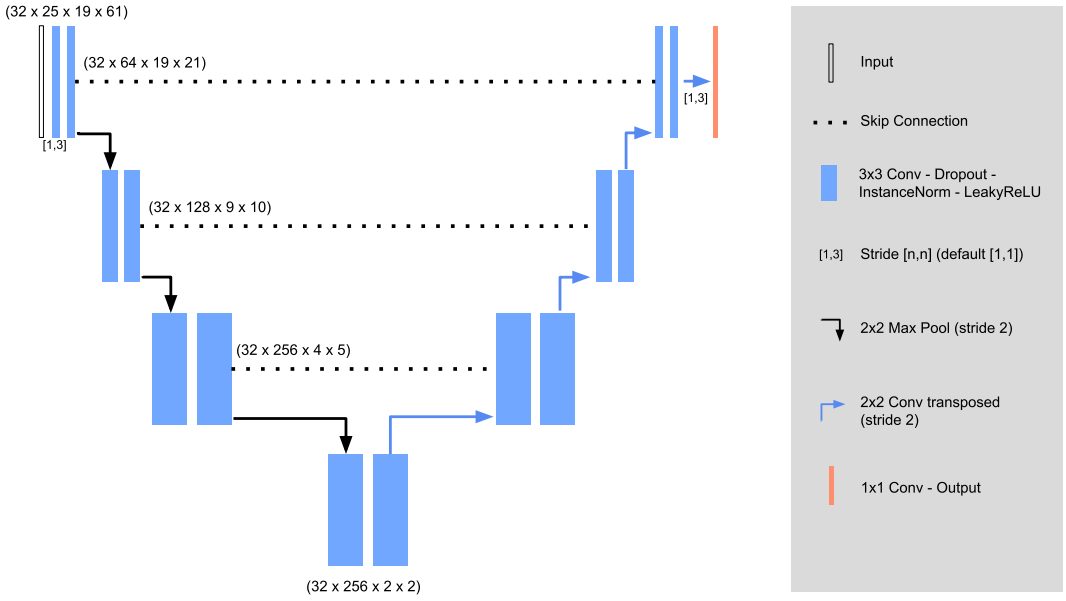

FIG. 6. U-Net architecture of our CNN model with convolution blocks and stages, along with filter sizes and strides. Each blue box corresponds to a multichannel feature map, and the associated descriptor refers to the batch size, the number of channels, and the dimensionality or size of the feature map at each layer. For example, the descriptor ($32 \times 25 \times 19 \times 61$) at upper left indicates a batch size of 32, the number 25 of feature variables, and the size of the 19×61 latitude and longitude grid of our evaluation domain (Fig. 1). The arrows represent operations specified in the legend at right. For definitions of technical terms, we refer to Ronneberger et al. (2015) and Isensee et al. (2021).

predictor variables, and let y be the target, namely, precipitation accumulation. The index model assumes the following relationship:

$$\log\left(y + \frac{1}{100}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j, \tag{5}$$

where the statistical coefficients β_0, β_1, …, β_m are learned from training data. Subsequent to the training of the index model, the nonparametric IDR distributions are estimated on the same training data, augmented with the fitted index values. We distinguish a baseline model (DIM-base, $m = 5$) and an extended model [DIM-full, $m = 25$ in Eq. (5)], for which we use the same sets of predictor variables as in the Logit approach from section 4c(1).

Note that PoP forecasts can be extracted from the DIM-base and DIM-full distributions. These yield similar, though slightly inferior, results than the Logit-base and Logit-full PoP forecasts, respectively (not shown).

d. *Machine learning–based forecasts: CNN+EasyUQ*

The aforementioned statistical models are applied at each grid point individually. Thus, including spatial information has to be done by manually engineering features accordingly, such as the correlated rainfall predictors from section 3a. In contrast, CNN models operate directly on the two-dimensional input space and can learn spatial relations from the data without the need to extract spatial information beforehand. CNN models are most commonly used for image tasks, where the input usually is a two- or three-dimensional array of pixel values. The gridded weather data over our evaluation domain can be envisaged as two-dimensional pseudo images of size 19×61. These dimensions correspond to latitude and longitude, respectively, spanning the study domain (Fig. 1) from 0° to 18°N and from 25°W to 35°E with a grid resolution of $1° \times 1°$. With a suitable architecture, a single CNN model produces a two-dimensional array with forecasts for all grid points at once instead of training models at each grid point individually. Due to their inherent inductive bias toward local neighborhood connectivity, CNNs are well suited for predicting precipitation on the 19×61 grids, as they effectively exploit spatial correlations and structures within a grid, recognizing patterns within local areas that may be indicative of specific weather conditions. For this reason, the three correlated rainfall predictors and the 20 ERA5 variables in the set of predictor variables for the full statistical models from section 4c are replaced by 19×61 grids of IMERG precipitation accumulations (section 2a) at temporal lags of 1, 2, and 3 days and 19×61 grids of the ERA5 variables (section 2b) at 0000 UTC, respectively. The two predictors that encode the day of the year are independent of the location and remain scalar features.

Motivated by their successful application in related meteorological tasks (Ayzel et al. 2020; Weyn et al. 2020; Lagerquist et al. 2021; Chapman et al. 2022; Otero and Horton 2023), we employ a CNN architecture in the form of the U-Net (Ronneberger et al. 2015; Isensee et al. 2021). The network architecture is sketched in Fig. 6 and follows standard choices for convolution blocks and stages. The architecture of the U-Net consists of a contracting (downsampling) path and an expansive (upsampling) path, which are symmetric in terms of individual layer properties, giving it a U-like shape. We make use of max pooling operations for

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

downsampling and transposed convolutions for upsampling layers. A crucial feature of the U-Net is skip connections between layers of the same size in the contracting and expanding paths. Applied to the precipitation data grid, these connections allow the network to use information from multiple resolutions, combining the context from the contracting path with the localization information from the expansive path. The only deviations from standard U-Net architecture choices, such as asymmetric strides and output padding in the upsampling layers, are dictated by the grid structure and allow to model longer spatial range dependencies in the data. To avoid overfitting, we make use of dropout (Srivastava et al. 2014) with a dropout rate of 0.2 throughout the network. With a view toward operational implementations, we refrain from tuning and adopting standard choices from Isensee et al. (2021) for all hyperparameters, thereby showcasing the capabilities of off-the-shelf CNNs in this application.

To transform the deterministic precipitation forecasts of the CNN model into probabilistic forecasts, the EasyUQ technique introduced in section 2d is applied at each grid point individually, subsequent to the training of the index model, and based on the same training data as for the neural network, augmented with the deterministic CNN output. As noted, the resulting CNN+EasyUQ forecast distributions are discrete and have mass exclusively at outcomes observed during training. Code for the implementation of the CNN+EasyUQ approach in Python (Python Software Foundation 2023) is publicly available under https://github.com/evwalz/precipitation. Once more we emphasize that while our usage of EasyUQ in concert with the CNN model is novel, we employ standard choices, such as quadratic loss and 3×3 convolutional kernels, for the neural network architecture and neural network training.

e. *Hybrid approaches*

NWP models represent the physical laws of atmospheric dynamics through a set of differential equations. Statistical or machine learning–based approaches, on the other hand, do not encode physical laws but learn patterns based exclusively on past data. A hybrid model is a combination of both approaches and thus can benefit from both the physical expertise embodied in NWP output and the flexibility of data-driven approaches. In this paper, we base hybrid approaches on the deterministic HRES forecast from section 2c and the deterministic CNN forecast from section 4d. We consider three approaches to obtain probabilistic forecasts from the deterministic HRES and CNN forecasts. First, the NWP forecast can be used as an additional gridded feature in the CNN model, followed by the gridpoint-based application of EasyUQ. Second, we can apply IDR using both deterministic forecasts as input features. Last, a simple approach is to use a weighted or unweighted average of the predictive distributions generated by HRES+EasyUQ and CNN+EasyUQ. We found experimentally that the first two approaches do not improve predictive ability, generally showing similar forecast performance to the CNN+EasyUQ forecast. The last approach in its most basic form of an equal average between the HRES+EasyUQ and CNN+EasyUQ distributions shows slight forecast improvements. In view of the simplistic nature of this latter approach, which does not require any additional training, we adopt it without further, formal model selection and refer to it as the hybrid model.

5. Forecast evaluation

In this section, we report major findings from the forecasting experiment. The discussion concentrates on the peak monsoon season JAS, but we provide results for the other seasons in additional figures in supplemental S1. As described at the start of section 4, our experiment uses expanding training sets to learn the forecasting models, and we frequently report annual results from the evaluation folds for 2011–2019. As evaluation metrics, the mean BS from Eq. (1) and the mean CRPS from Eq. (2) are used.

a. *Effects of variable selection in statistical models*

To better understand the influence of the predictor variables on the forecast performance of the statistical models, namely, the Logit PoP forecast [section 4c(1)] and the DIM forecast for precipitation accumulation [section 4c(2)], a visual analysis is provided in Fig. 7. Starting with the mean score of the base model, which has five predictor variables, one more variable is successively added and the corresponding mean score is shown until the full model with 25 predictor variables is reached. The variables are selected in the order of the distance between 0.5, and the mean AUC or mean CPA is computed without splitting into seasons. An AUC or CPA value of 0.5 suggests a useless feature.

Figure 7a shows that the addition of TCWV to the Logit base model yields an improvement of the BS on the order of 5% in all years. Small further improvements of less than 1% are obtained by adding midlevel humidity (Q700) and static stability (KX). The addition of further variables yields minor improvement only, with the striking exception of 2-m temperature (T2), which leads to an improvement comparable to Q700 and KX, despite AUC values barely above 0.5 (Fig. 2d). We are unable to provide a meteorological interpretation of this effect and encourage follow-up studies. Qualitatively, improvements in CRPS per predictor regarding precipitation amount (Fig. 7b) show similar results, yet the percentage improvements are smaller such that adding variables other than TCWV and Q700 barely improves performance. Generally, the performance difference between years is large, and the ranking of the years differs between the BS, where the lowest values are seen for 2017, and the CRPS, where they are seen for 2013. The mean CRPS covaries with the total rainfall amount (Fig. 9b) and increased amounts of training data facilitate improved performance in later years. Results for seasons other than JAS are qualitatively similar, except that the overall level of the scores varies strongly between seasons (Figs. S3 and S4).

Let us now return to the discussion of correlated predictor variables and variable selection in sections 3c and 4c. In Fig. 8, we report on the same experiment as in Fig. 7, but for using a subset of weakly correlated predictor variables, chosen according to the analysis in Fig. 5 and Fig. S2. The comparison between Figs. 7 and 8, where the full and the reduced model

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

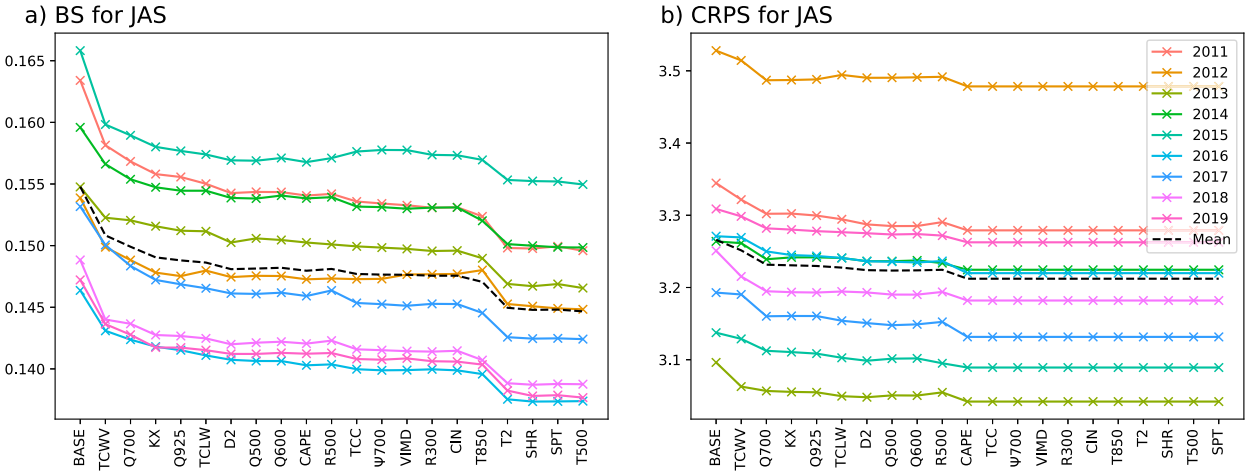

FIG. 7. (a) Mean BS for the Logit PoP forecast under successive addition of the predictor variables displayed on the horizontal axis. The base model includes three correlated rainfall predictors and two time features. The BS is averaged over space and season JAS for evaluation folds from 2011 to 2019. (b) Corresponding display for the DIM forecast of precipitation accumulation and the CRPS.

achieve essentially identical predictive performance, corroborates the insight that generally, the use of highly correlated predictor variables does not degrade performance in this type of setting.

b. *Comparative evaluation of predictive performance*

Figure 9a visualizes the mean BS in season JAS for the PoP forecasting models from Table 2. Similar to the results in Vogel et al. (2021), the ECMWF EPS shows inferior or, in later years, comparable performance to MPC, and both EPS and MPC are outperformed by a simple logistic regression approach based on correlated rainfall predictors only (Logit-base). The inclusion of ERA5 predictors into the logistic regression model (Logit-full) leads to a clear improvement beyond the postprocessed EPS-ISO PoP forecast. Surprisingly, the HRES+EasyUQ forecast shows better performance than the ensemble-based EPS+ISO forecast. The CNN+EasyUQ PoP forecast outperforms all other methods, except for the hybrid forecast, which shows nearly the same performance. In Diebold and Mariano (1995)'s tests based on a time series of spatially aggregated daily scores over the complete evaluation period, the null hypothesis of equal predictive ability in terms of the Brier score gets rejected at the 0.01 level for all 28 pairs of models, with the exceptions of EPS-ISO versus Logit-base, and, notably, hybrid versus CNN+EasyUQ.

The mean CRPS for the forecasting models for precipitation accumulation from Table 2 is displayed in Fig. 9b. Through 2014, EPS clearly shows the lowest forecast skill; thereafter, its skill improves and gets close to the performance of MPC and DIM-base. Unlike the Logit-full PoP forecast, DIM-full does not outperform the postprocessed EPS+EMOS forecast. The fact that EPS+EMOS does not share the inhomogeneous behavior of EPS for the years before and after

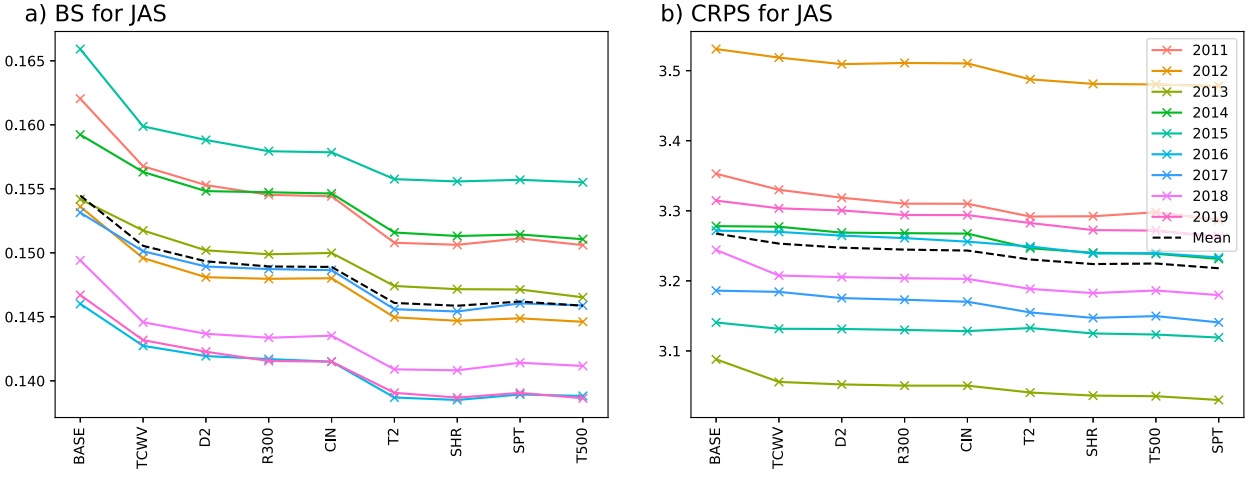

FIG. 8. As in Fig. 7, but for a subset of weakly correlated predictor variables.

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

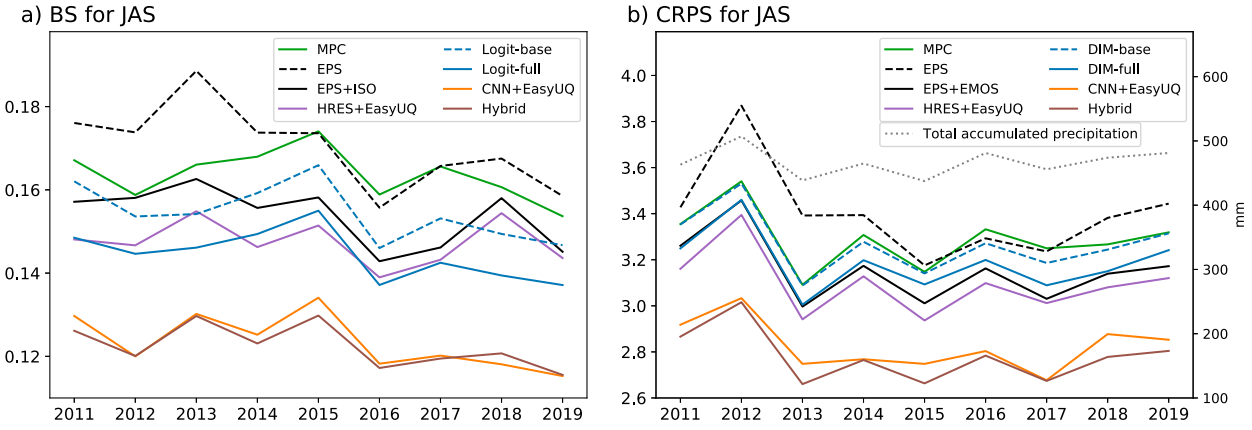

FIG. 9. (a) Mean BS for PoP forecasts from Table 2, averaged over space and season JAS for evaluation folds from 2011 to 2019. (b) Corresponding display for the CRPS and probabilistic forecasts of precipitation amount, along with spatially averaged total accumulated precipitation for JAS, in the unit of millimeters.

2014 indicates that postprocessing can mostly cure the large miscalibration inherent in earlier versions of EPS. The HRES+EasyUQ approach yields better scores than EPS+EMOS, probably due to the flexibility of the EasyUQ forecast distributions. The CNN+EasyUQ approach shows a forecast improvement within the evaluation period, and the hybrid model performs similar or slightly better for some years. As can be seen, by the dotted line giving the JAS area-averaged rainfall, the mean CRPS covaries with the total rainfall amount; thus, the years with the best performance are usually also the driest. In Diebold and Mariano (1995)'s tests, the null hypothesis of equal predictive ability in terms of the CRPS gets rejected at the 0.01 level for all 28 pairs of models, with the sole exception of DIM-base versus MPC.

Figures S5 and S6 show analogous evaluation results for all five seasons. Throughout, the CNN+EasyUQ and hybrid forecasts perform similarly to each other and outperform their competitors by considerable margins.

c. *Spatial structure of predictive performance*

For an understanding of spatial patterns of forecast performance, skill score plots of the forecast approaches considered here with MPC as reference forecast are shown in Fig. 10 for precipitation occurrence and in Fig. 11 for precipitation accumulation, both for the JAS peak monsoon season and across evaluation folds.

With respect to the PoP forecasts for rainfall occurrence, EPS shows negative skill relative to MPC over the southern parts of the study domain, particularly over the relatively dry areas along the Guinea Coast and over Gabon and southern Cameroon, where rainfall tends to be rather localized and short-lived such that precipitation occurrence is hard to predict (Fig. 10a). Senegal/Mauritania and Chad/Sudan are the only areas with considerable positive skill, while the rest of the domain ranges close to 0. Applying statistical postprocessing (EPS+ISO; Fig. 10b) removes the large negative skill along the Guinea Coast but shows remaining issues in a stretch from Nigeria to South Sudan with mostly weakly negative skill. Remarkably, postprocessing deteriorates skill around the highlands in Guinea/Sierra Leone and westernmost Ethiopia. Over the Sahel, in contrast, the postprocessing leads to an overall improvement and consistently positive skill. A possible reason is the stronger influence of predictable features such as AEWs or midlatitude perturbations here in contrast to the more stochastic rains in the south (Rasheeda Satheesh et al. 2023). The comparison between the EPS+ISO and HRES+EasyUQ (Fig. 10c) demonstrates that for forecasts at individual sites, there is no added value in running an NWP ensemble system, even after postprocessing. The structures are fairly consistent (e.g., with problematic regions in Guinea/Sierra Leone, the Central African Republic, South Sudan, and Ethiopia), but the values are consistently more positive for the HRES+EasyUQ technique, which is based on the HRES model alone, as opposed to using an ensemble.

Moving to the data-based approaches (Figs. 10d–g), we see consistent improvement over most areas of the study domain though PoP forecasts for western Ethiopia remain a challenge, possibly related to the rough topography in this area. While in the simpler Logit-base approach (Fig. 10d) some areas of negative skill remain, the inclusion of additional predictors in Logit-full (Fig. 10e) leads to a consistent improvement and thus positive skill almost everywhere in the study region. It is also noteworthy that the Logit models generate overall smoother skill fields compared to the physics-based approaches. Finally, the CNN+EasyUQ and hybrid methods (Figs. 10f,g) outperform all other approaches to a large extent, reaching up to 40% improvement relative to the climatological benchmark MPC. The improvement relative to EPS is particularly impressive over the Guinea coastal region (e.g., Ivory Coast and Ghana), where EPS performs much worse than MPC and illustrates the ability of the CNN to learn complex physical relationships that determine local rainfall probability. The inclusion of NWP information from the HRES model in the hybrid approach yields small improvements in some places but no clear advance relative to CNN+EasyUQ. This demonstrates that knowing the ambient conditions shortly

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

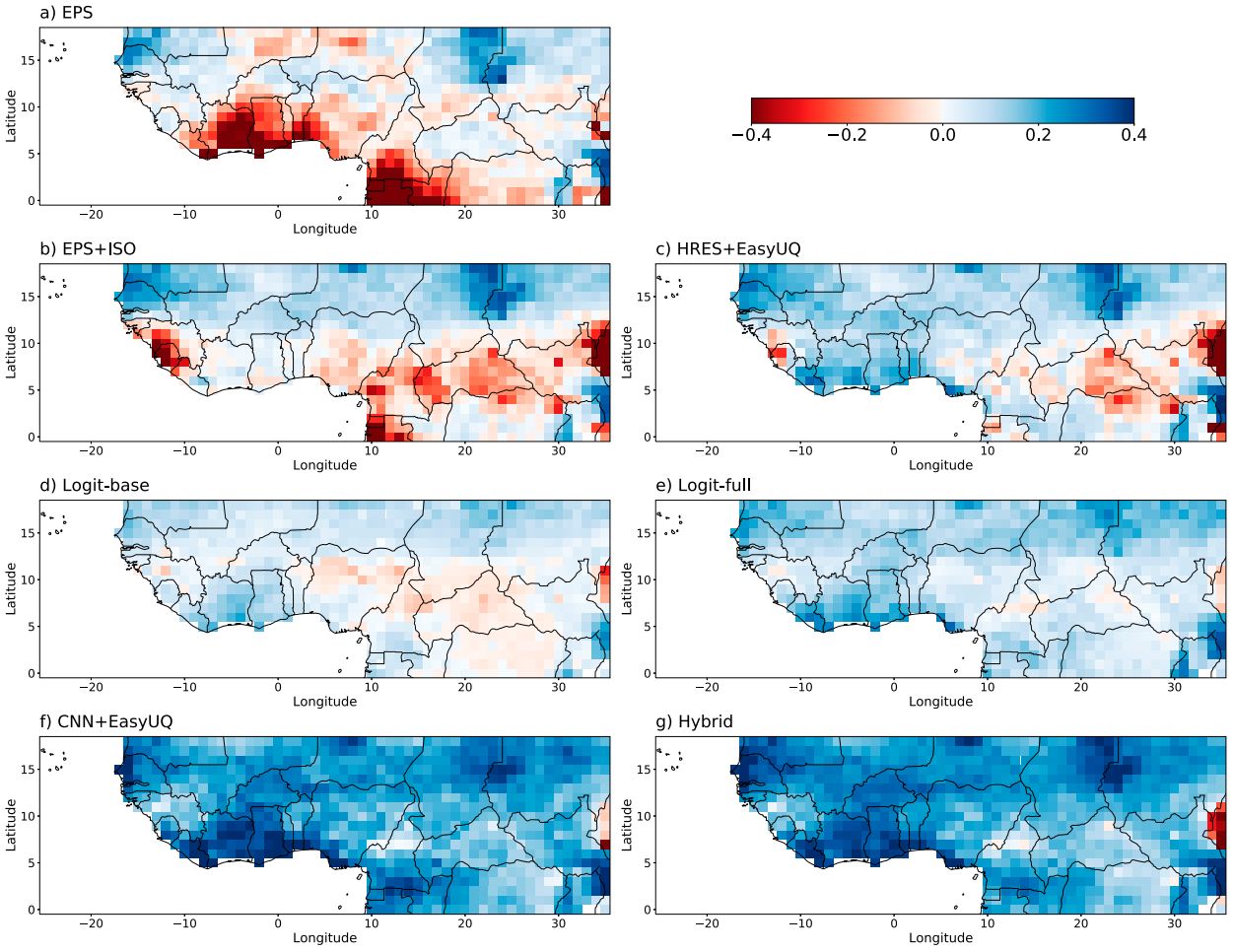

FIG. 10. Spatial structure of the Brier skill score for probability forecasts of precipitation occurrence with (a) EPS, (b) EPS+ISO, (c) HRES+EasyUQ, (d) Logit-base, (e) Logit-full, (f) CNN+EasyUQ, and (g) the hybrid forecast from Table 2, relative to MPC as baseline, for season JAS and combined evaluation folds from 2011 to 2019.

before the beginning of the 24-h forecast period is much more important than knowledge of the forecast evolution during that period.

The corresponding analysis for rainfall amount (Fig. 11) reveals many parallels to rainfall probability. EPS (Fig. 11a) stands out as having many areas of negative CRPS skill, with an overall similar structure to the occurrence analysis (Fig. 10a). Postprocessing (EPS+EMOS; Fig. 11b) cures many issues of EPS, leading to mostly weakly positive skill, but does not perform as well as the computationally much less expensive HRES+EasyUQ technique (Fig. 11c). Here, the skill fields for amount are overall smoother than for occurrence with less contrast between the Sahel and the southern areas. The DIM models (replacing the Logit models for amount) show negligible further advance. The skill of DIM-base (Fig. 11d) is close to 0 everywhere with a negative area in the southeast and positive elsewhere, while the inclusion of additional predictors (DIM-full; Fig. 11e) slightly improves skill over most areas. Finally, as for occurrence, the machine learning–based CNN+EasyUQ and hybrid methods (Figs. 11f,g) outperform all other approaches to a large extent with positive CRPS skill of up to 30%. Here, the hybrid approach leads to a more considerable improvement relative to CNN+EasyUQ, yielding fairly equal skill improvement across the entire, quite heterogeneous domain. These improvements are more prominent in areas where the physics-based HRES model may better represent the time evolution of dynamical features such as AEWs and extratropical influences.

d. Calibration and discrimination ability

We now assess the calibration and discrimination ability of the forecasts. Following Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), reliability diagrams for the PoP forecasts from Table 2 at the grid point closest to Niamey (13°N, 2°E) are presented in Fig. 12. The choice of Niamey reflects typical conditions in the Sahel and allows direct comparison to results in the earlier papers. The panels use the CORP approach of Dimitriadis et al. (2021) and show the decomposition from Eq. (3) of the mean BS into MCB, DSC, and UNC components. Instead of considering each evaluation fold separately, the decomposition is computed once on forecasts in the peak

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

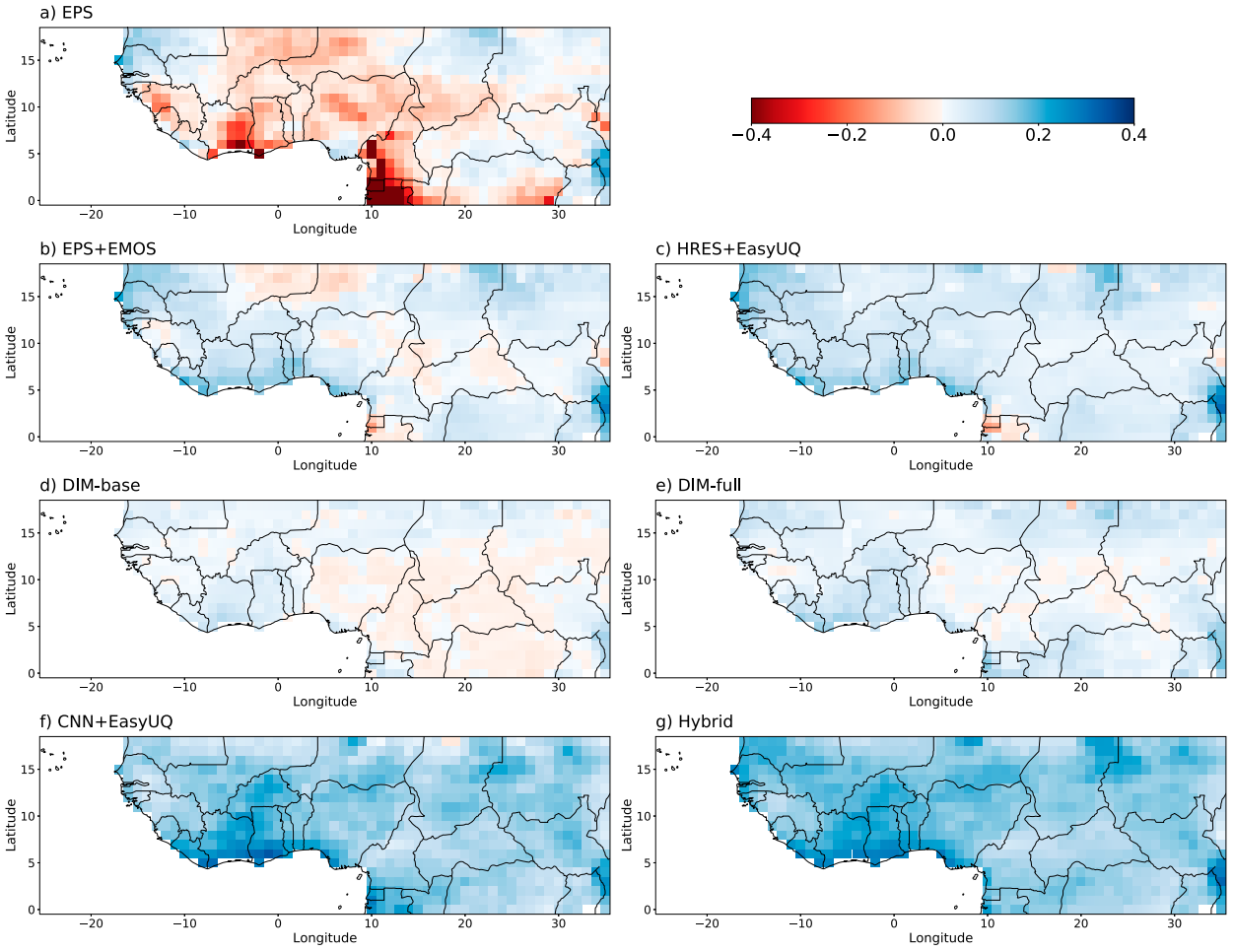

FIG. 11. Spatial structure of the CRPS skill score for probabilistic forecasts of precipitation accumulation with (a) EPS, (b) EPS+EMOS, (c) HRES+EasyUQ, (d) DIM-base, (e) DIM-full, (f) CNN+EasyUQ, and (g) the hybrid forecast from Table 2, relative to MPC as baseline, for season JAS and combined evaluation folds from 2011 to 2019.

monsoon season JAS from all nine evaluation years together. If the reliability curve is close to the diagonal, a PoP forecast is calibrated (reliable). Deviations from the diagonal indicate some type of miscalibration: S-shaped curves indicate underconfidence (PoP too close to the overall observed frequency of rain), inverse S-shaped curves correspond to overconfidence (PoP too close to 0 or 1), and curves that are mostly below the diagonal, or mostly above the diagonal, indicate biased PoP. The climatological MPC PoP forecast has a very limited range of forecast probabilities and lacks discrimination ability but shows excellent calibration. The poor calibration of the raw ENS PoP is corrected by postprocessing (ENS+ISO). In agreement with the findings of Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), the Logit-base PoP forecast is well calibrated and has moderate discrimination ability. In comparison, Logit-full shows a lower BS (more skillful PoP forecasts) reflected in both better calibration and improved discrimination ability. The CNN+EasyUQ and hybrid techniques show superior performance—they are similarly well calibrated as EPS-ISO and Logit-full but show considerably higher discrimination ability.

To assess the calibration of the probabilistic forecasts for accumulated precipitation at the grid point closest to Niamey, Fig. 13 shows probability integral transform (PIT) histograms. For the MPC and EPS ensemble forecast, a universal PIT (uPIT) histogram is shown (Vogel et al. 2018); for the other methods, the randomized version of the PIT is used [Gneiting and Resin 2023, Eq. (1)]. A uniform histogram indicates calibrated forecasts, while a U-shaped (hump shaped) histogram suggests underdispersed (overdispersed) forecasts, meaning that the forecasts are overconfident (underconfident). Skewed histograms indicate biases. The ECMWF ensemble (EPS) is underdispersed, which is corrected for in the EPS+EMOS forecast, though a bias remains. The other forecasts show PIT histograms that are nearly uniform. The associated decomposition [Eq. (3)] of the mean CRPS demonstrates the superior calibration of the climatological MPC forecast and the outstanding discrimination ability and overall predictive performance of the CNN+EasyUQ and hybrid approaches.

Finally, we use the decomposition of the mean BS or mean CRPS into MCB, DSC, and UNC components for a spatially

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

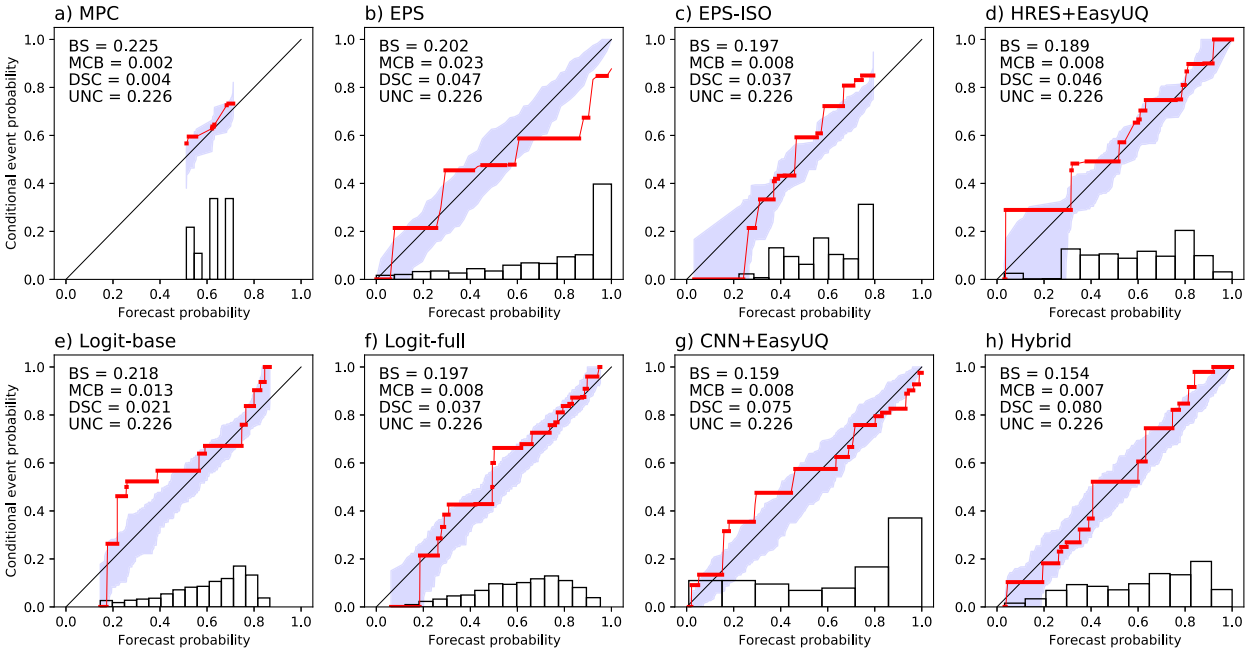

FIG. 12. Reliability diagrams for PoP forecasts at the grid point closest to Niamey (13°N, 2°E) with (a) MPC, (b) EPS, (c) EPS+ISO, (d) HRES+EasyUQ, (e) Logit-base, (f) Logit-full, (g) CNN+EasyUQ, and (h) the hybrid approach from Table 2, for season JAS and combined evaluation folds from 2011 to 2019, with 90% consistency bands under the assumption of calibration (Dimitriadis et al. 2021). The panels also show the mean BS and its MCB, DSC), and UNC components from Eq. (3). The histograms along the horizontal axis show the distribution of the forecast probabilities.

aggregated quantitative assessment. We compute the decomposition [Eq. (3)] at each grid point based on forecasts in the peak monsoon season JAS from all nine evaluation years, and the score components are then averaged across grid points. The MCB–DSC plots for the mean BS (Fig. 14a) and mean CRPS (Fig. 14b) provide a spatially consolidated comparison of the forecast methods. In both cases, the climatological MPC forecast shows the lowest MCB and the lowest DSC component. The ECMWF raw ensemble (EPS) has higher MCB than all other methods, and the miscalibration is taken care of by postprocessing (EPS+ISO and EPS+EMOS). Regarding the statistical forecasts, the inclusion of the ERA5 predictors (Logit-full and DIM-full) in addition to the correlated rainfall predictors (Logit-basic and DIM-basic) improves DSC while MCB remains similar. The superiority of the CNN+EasyUQ forecast stems from its elevated discrimination ability. The hybrid forecast shows slightly improved skill relative to CNN+EasyUQ and trades better calibration for even higher discrimination ability. These findings are stable and apply across all five seasons (Figs. S7 and S8).

6. Conclusions

In this work, the predictability of 1-day-ahead 24-h precipitation occurrence and amount over northern tropical Africa is investigated on the basis of conventional and new data-driven tools. Our study builds on previous papers with a focus on forecasting rainfall occurrence for the summer season JAS, which compared the performance of climatological, raw and postprocessed ECMWF ensemble forecasts, and a simple logistic regression model based on correlated rainfall predictors. This binary forecasting problem is revisited in this paper with major adaptions. Instead of TRMM, GPM IMERG is used as a ground truth data source. Forecasts are produced for the entire year instead of just the summer season (JAS), and ERA5 predictor variables are used to augment the logistic regression model. To this end, an extensive analysis of weather variables from the ERA5 is performed to investigate and understand their relation to and their influence on precipitation. The meteorological interpretation of these dependencies is obtained by combining previously conducted research and results from statistical analysis performed in this work.

A key contribution of our work is that we additionally investigate the more challenging problem of producing probabilistic forecasts for accumulated precipitation. Since the climatology and the NWP model output in this paper are in the form of ensembles, they can be readily used as probabilistic forecasts for precipitation amount. To produce data-driven statistical forecasts, the DIM is introduced, which is simple but very effective and thus can serve as a persuasive baseline. To account for the recent rise of machine learning in weather forecasting, a CNN model is presented which has the additional benefit of inherently exploiting spatial relations. To obtain a probabilistic output, we couple the CNN model with the recently introduced EasyUQ approach, to yield the CNN+EasyUQ technique. These different forecasting approaches provide a detailed forecasting benchmark covering the range of simple to sophisticated models and ideas from

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

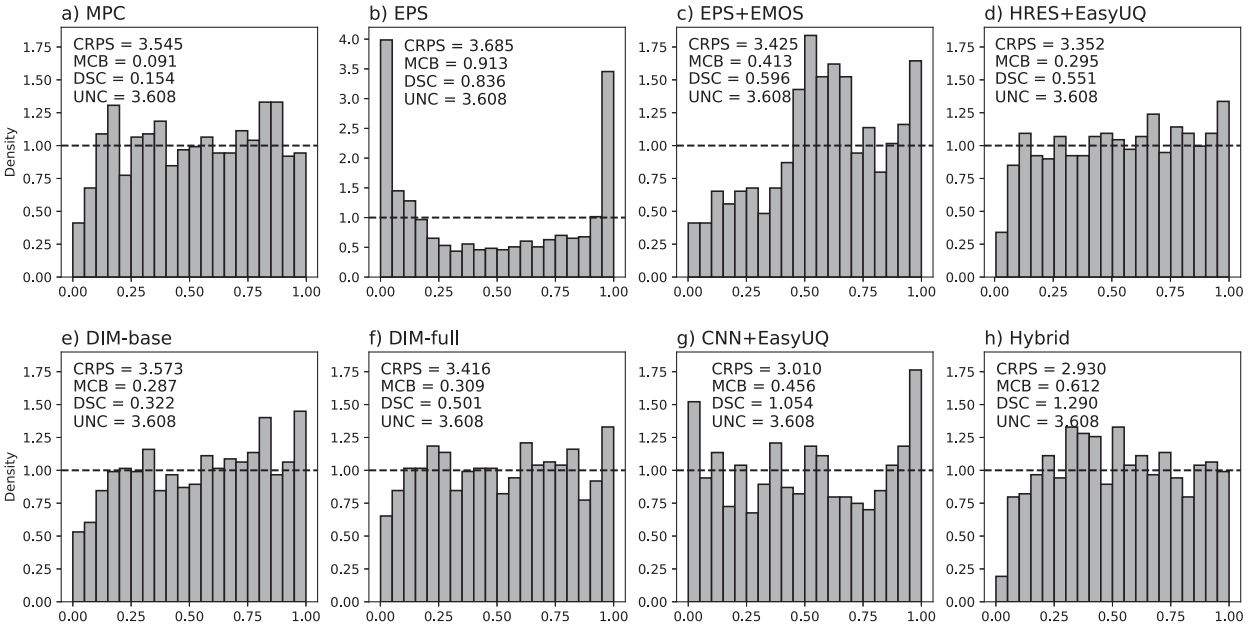

FIG. 13. PIT histograms for probabilistic forecasts of precipitation accumulation at the grid point closest to Niamey (13°N, 2°E) with (a) MPC, (b) EPS, (c) EPS+EMOS, (d) HRES+EasyUQ, (e) DIM-base, (f) DIM-full, (g) CNN+EasyUQ, and (h) the hybrid approach from Table 2, for season JAS and combined evaluation folds from 2011 to 2019. The panels also show the mean CRPS and its MCB, DSC, and UNC components from Eq. (3). The vertical scale of the histograms is shared across forecasts, except for EPS. The horizontal dashed line in each histogram represents the (desired) uniform distribution and aids the interpretation of deviations.

NWP, statistics, and machine learning in an unprecedented way.

The CNN+EasyUQ technique outperforms its competitors by a large margin, except for the hybrid forecast, which is a simple arithmetic average of the HRES+EasyUQ and CNN+EasyUQ forecast distributions that does not require any additional training and yields minor only (if any) further improvement. It is interesting to place our results for 1-day-ahead 24-h forecasts in the context of recent advances in data-based precipitation forecasts. For nowcasts at prediction horizons up to 12 h, progress has been

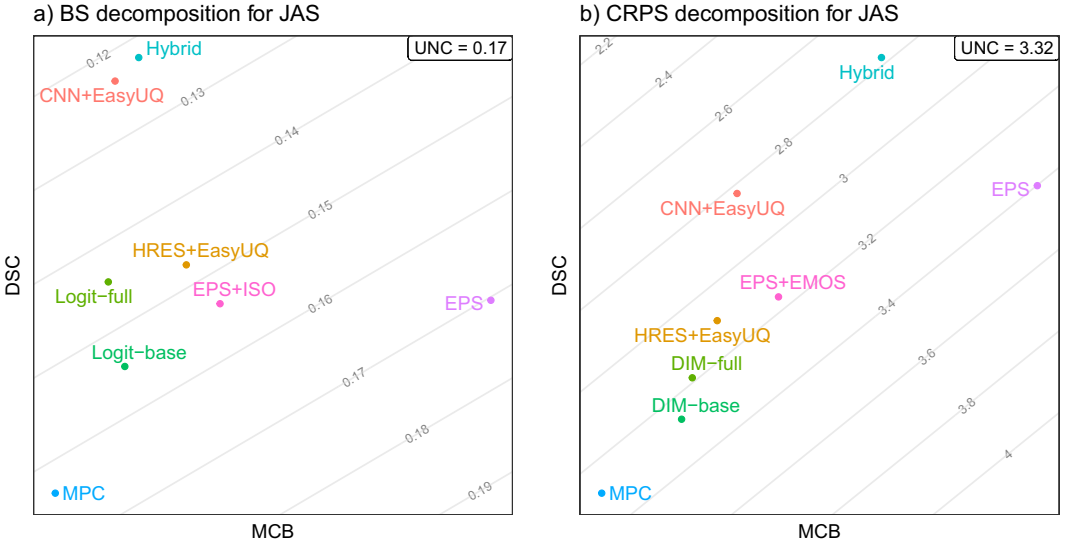

FIG. 14. MCB, DSC, and UNC components of (a) the mean BS for probability forecasts of precipitation occurrence and (b) the mean CRPS for probabilistic forecasts of precipitation accumulation in the unit of millimeters, as described in Table 2. The score decomposition in Eq. (3) is applied at each grid point, based on the combined evaluation folds from 2011 to 2019, and the mean score and score components are then averaged over grid points. Parallel lines correspond to equal mean scores.

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

persuasive (Ayzel et al. 2020; Lagerquist et al. 2021; Ravuri et al. 2021; Espeholt et al. 2022; Zhang et al. 2023). In stark contrast, recent developments in neural network–based weather forecasts at prediction horizons of days ahead have provided sparse attention to rainfall (Bi et al. 2023; Rasp et al. 2024), arguably due to the recognition that "precipitation is sparse and non-Gaussian" (Lam et al. 2023, p. 6). The CNN+EasyUQ technique provides an elegant and computationally highly efficient way of addressing the non-Gaussianity of precipitation accumulation. In very recent work, Andrychowicz et al. (2023) find that the data-driven MetNet-3 approach outperforms the ECMWF and NOAA raw ensembles in terms of CRPS for hourly precipitation accumulation over the continental United States at lead times up to 20 h, but not beyond. However, unlike our study, which compares the CNN+EasyUQ forecast with state-of-the-art competitors, Andrychowicz et al. (2023) do not compare MetNet-3 to postprocessed NWP ensemble forecasts nor to statistical forecasts of the type considered in our paper. Related work by Scheuerer et al. (2020), Ghazvinian et al. (2022), and Horat and Lerch (2024) at longer lead times concerns the use of neural networks for the postprocessing of precipitation forecasts generated by NWP ensembles.

In view of its outstanding performance in this study, the CNN+EasyUQ approach can likely improve operational probabilistic forecasts of day-ahead 24-h rainfall in northern tropical Africa. Our current implementation does not involve hyperparameter tuning in learning the neural network and thus leaves potential for further improvement in predictive performance. However, the superior performance of the off-the shelf neural network shows that no intensive tuning or expert knowledge is required to construct a data-driven forecasting model that outperforms NWP forecasts. This encourages the operational usage of neural network models and facilitates its implementation. To make real-time forecasts feasible, one would need to use the IMERG Early Run (NASA 2024) in lieu of IMERG, an option that remains to be tested. To obtain ensemble forecasts of entire, spatiotemporally coherent precipitation fields, rather than forecasts at individual locations and fixed prediction horizons, the HRES+EasyUQ and CNN+EasyUQ approaches can be coupled with empirical copula techniques (Clark et al. 2004; Schefzik et al. 2013), for which we encourage follow-up studies.

While our study is limited in geographic scope, we feel that data-driven approaches of this type have the potential to revolutionize rainfall forecasts throughout the tropics. Furthermore, the results of comparative studies by Little et al. (2009) for the United Kingdom and Andrychowicz et al. (2023) for the continental United States admit the speculation that the CNN+EasyUQ technique can improve probabilistic forecasts of 24-h precipitation in the extratropics as well. Finally, a very interesting and relevant research question is whether similar advances in predictive performance are feasible at prediction horizons larger than a day ahead.

Acknowledgments. We thank the editor in chief, three anonymous referees, Sebastian Lerch, and Marlon Maranan for their comments and discussion. The work of Eva-Maria Walz was funded by the German Research Foundation (DFG) through Grant 257899354. Tilmann Gneiting is grateful for support from the Klaus Tschira Foundation. The research leading to these results has been accomplished within phase 2 of project C2, "Statistical-dynamical forecasts of tropical rainfall," of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Research Foundation (DFG).

Data availability statement. The reproduction of the results in this paper requires access to GPM IMERG precipitation data (version 6B), predictor variables from ERA5, and ECMWF NWP forecasts. For GPM IMERG (Huffman et al. 2020) and ERA5 (Hersbach et al. 2020), our sources are freely accessible, which makes results for MPC, the statistical approaches (Logit and DIM), and, our key innovation, the CNN+EasyUQ technique, readily reproducible. For the more elaborate CNN+EasyUQ approach and CORP reliability diagrams, code in Python (Python Software Foundation 2023) is available at https://github.com/evwalz/precipitation (Release v1.0.0) and https://github.com/evwalz/corp_reldiag (Release v1.0.0), respectively. As noted, we downloaded the NWP forecasts from ECMWF's Meteorological Archival and Retrieval System (MARS; ECMWF 2018), for which access conditions depend on the user. These forecasts are freely available from the International Grand Global Ensemble (TIGGE) archive (Bougeault et al. 2010).

REFERENCES

Ageet, S., A. H. Fink, M. Maranan, and B. Schulz, 2023: Predictability of rainfall over equatorial East Africa in the ECMWF ensemble reforecasts on short to medium-range time scales. *Wea. Forecasting*, **38**, 2613–2630, https://doi.org/10.1175/WAF-D-23-0093.1.

Andrychowicz, M., L. Espeholt, D. Li, S. Merchant, A. Merose, F. Zyda, S. Agrawal, and N. Kalchbrenner, 2023: Deep learning for day forecasts from sparse observations. arXiv, 2306.06079v3, https://doi.org/10.48550/arXiv.2306.06079.

Arnold, S., E.-M. Walz, J. Ziegel, and T. Gneiting, 2023: Decompositions of the mean continuous ranked probability score. arXiv, 2311.14122v1, https://doi.org/10.48550/arXiv.2311.14122.

Ayzel, G., T. Scheffer, and M. Heistermann, 2020: RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.*, **13**, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Becker, T., P. Bechtold, and I. Sandu, 2021: Characteristics of convective precipitation over tropical Africa in storm-resolving global simulations. *Quart. J. Roy. Meteor. Soc.*, **147**, 4388–4407, https://doi.org/10.1002/qj.4185.

Ben Bouallègue, Z., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine-learning based weather forecasts in an operational-like context. *Bull. Amer. Meteor. Soc.*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bougeault, P., and Coauthors, 2010: The THORPEX interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Chapman, W. E., L. Delle Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

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.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for AI weather prediction. *Nature*, **619**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

ECMWF, 2018: MARS – The ECMWF meteorological archive. ECMWF, https://www.ecmwf.int/en/elibrary/80518-mars-ecmwf-meteorological-archive.

——, 2021: A brief description of reforecasts. ECMWF, https://confluence.ecmwf.int/display/S2S/A+brief+description+of+reforecasts.

——, 2023a: Changes to the forecasting system. ECMWF, https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system.

——, 2023b: IFS cycle upgrades pre 2015. ECMWF, https://confluence.ecmwf.int/display/FCST/IFS+cycle+upgrades+pre+2015.

Espeholt, L., and Coauthors, 2022: Deep learning for twelve hour precipitation forecasts. *Nat. Commun.*, **13**, 5145, https://doi.org/10.1038/s41467-022-32483-x.

Fink, A. H., and Coauthors, 2017: Mean climate and seasonal cycle. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., John Wiley and Sons, 1–39.

Galvin, J. F. P., 2010: Two easterly waves in West Africa in summer 2009. *Weather*, **65**, 219–227, https://doi.org/10.1002/wea.605.

Gebremichael, M., H. Yue, and V. Nourani, 2022: The accuracy of precipitation forecasts at timescales of 1–15 days in the Volta river basin. *Remote Sens.*, **14**, 937, https://doi.org/10.3390/rs14040937.

Ghazvinian, M., Y. Zhang, T. M. Hamill, D.-J. Seo, and N. Fernando, 2022: Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *J. Hydrometeor.*, **23**, 1365–1382, https://doi.org/10.1175/JHM-D-22-0021.1.

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 E.-M. Walz, 2022: Receiver Operating Characteristic (ROC) movies, Universal ROC (UROC) curves, and Coefficient of Predictive Ability (CPA). *Mach. Learn.*, **111**, 2769–2797, https://doi.org/10.1007/s10994-021-06114-3.

——, and J. Resin, 2023: Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electron. J. Stat.*, **17**, 3226–3286, https://doi.org/10.1214/23-EJS2180.

——, A. E. Raftery, A. H. Westveld III, 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 the SEEPS score. *Mon. Wea. Rev.*, **140**, 2720–2733, https://doi.org/10.1175/MWR-D-11-00301.1.

Henzi, A., J. F. Ziegel, and T. Gneiting, 2021: Isotonic distributional regression. *J. Roy. Stat. Soc.*, **83B**, 963–993, https://doi.org/10.1111/rssb.12450.

——, G.-R. Kleger, and J. F. Ziegel, 2023: Distributional (single) index models. *J. Amer. Stat. Assoc.*, **118**, 489–503, https://doi.org/10.1080/01621459.2021.1938582.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Horat, N., and S. Lerch, 2024: Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Mon. Wea. Rev.*, **152**, 667–687, https://doi.org/10.1175/MWR-D-23-0150.1.

Hou, A. Y., and Coauthors, 2014: The global precipitation measurement mission. *Bull. Amer. Meteor. Soc.*, **95**, 701–722, https://doi.org/10.1175/BAMS-D-13-00164.1.

Hubert, H., 1939: Origine Africaine d'un cyclone tropical Atlantique. *Ann. Phys. France d'Outre-Mer*, **6**, 97–115.

Huffman, G. J., and Coauthors, 2020: Integrated multi-satellite retrievals for the Global Precipitation Measurement (GPM) mission (IMERG). *Satellite Precipitation Measurement: Volume 1*, V. Levizzani et al., Eds., Springer, 343–353.

Isensee, F., P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, 2021: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, **18**, 203–211, https://doi.org/10.1038/s41592-020-01008-z.

Kiladis, G. N., C. D. Thorncroft, and N. M. J. Hall, 2006: Three-dimensional structure and dynamics of African easterly waves. Part I: Observations. *J. Atmos. Sci.*, **63**, 2212–2230, https://doi.org/10.1175/JAS3741.1.

Klein, C., F. Nkrumah, C. M. Taylor, and E. A. Adefisan, 2021: Seasonality and trends of drivers of mesoscale convective systems in southern West Africa. *J. Climate*, **34**, 71–87, https://doi.org/10.1175/JCLI-D-20-0194.1.

Lafore, J. P., and Coauthors, 2017a: Deep convection. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., John Wiley and Sons, 90–129.

——, and Coauthors, 2017b: West African synthetic analysis and forecast: WASA/F. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., John Wiley and Sons, 423–451.

Lagerquist, R., J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using deep learning to nowcast the spatial coverage of convection from *Himawari-8* satellite data. *Mon. Wea. Rev.*, **149**, 3897–3921, https://doi.org/10.1175/MWR-D-21-0096.1.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Little, M. A., P. E. McSharry, and J. W. Taylor, 2009: Generalized linear models for site-specific density forecasting of U.K.

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC

daily rainfall. *Mon. Wea. Rev.*, **137**, 1029–1045, https://doi.org/10.1175/2008MWR2614.1.

Maranan, M., A. H. Fink, and P. Knippertz, 2018: Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment. *Quart. J. Roy. Meteor. Soc.*, **144**, 1628–1648, https://doi.org/10.1002/qj.3345.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

NASA, 2024: IMERG – Early run. NASA, accessed 19 March 2020, https://gpm.nasa.gov/taxonomy/term/1357.

Nesbitt, S. W., R. Cifelli, and S. A. Rutledge, 2006: Storm morphology and rainfall characteristics of TRMM precipitation features. *Mon. Wea. Rev.*, **134**, 2702–2721, https://doi.org/10.1175/MWR3200.1.

Nicholls, S. D., and K. I. Mohr, 2010: An analysis of the environments of intense convective systems in West Africa. *Mon. Wea. Rev.*, **138**, 3721–3739, https://doi.org/10.1175/2010MWR3321.1.

Otero, N., and P. Horton, 2023: Intercomparison of deep learning architectures for the prediction of precipitation fields with a focus on extremes. *Water Resour. Res.*, **59**, e2023WR035088, https://doi.org/10.1029/2023WR035088.

Python Software Foundation, 2023: The python language reference. Python Software Foundation, accessed 9 July 2020, https://python.org/.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

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.*, **149**, 3721–3737, https://doi.org/10.1002/qj.4581.

Rasp, S., and Coauthors, 2024: WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.*, **16**, e2023MS004019, https://doi.org/10.1029/2023MS004019.

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

Regula, H., 1936: Druckschwankungen und Tornados an der Westküste von Afrika. *Ann. Hydrogr. Mar. Meteor.*, **64**, 107–111.

Roca, R., J. Aublanc, P. Chambon, T. Fiolleau, and N. Viltard, 2014: Robust observational quantification of the contribution of mesoscale convective systems to rainfall in the Tropics. *J. Climate*, **27**, 4952–4958, https://doi.org/10.1175/JCLI-D-13-00628.1.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th Int. Conf., Part III*, Munich, Germany, Springer, 234–241, https://link.springer.com/book/10.1007/978-3-319-24574-4.

Rotunno, R., J. B. Klemp, and M. L. Weismann, 1988: A theory for strong, long-lived squall lines. *J. Atmos. Sci.*, **45**, 463–485, https://doi.org/10.1175/1520-0469(1988)045<0463:ATFSLL>2.0.CO;2.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, https://doi.org/10.1214/13-STS443.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schlueter, A., A. H. Fink, P. Knippertz, and P. Vogel, 2019: A systematic comparison of tropical waves over Northern Africa. Part I: Influence on rainfall. *J. Climate*, **32**, 1501–1523, https://doi.org/10.1175/JCLI-D-18-0173.1.

Schneider, U., M. Ziese, A. Becker, A. Meyer-Christoffer, and P. Finger, 2015: Global precipitation analysis products of the GPCC. Deutscher Wetterdienst, 14 pp., https://opendata.dwd.de/climate_environment/GPCC/PDF/GPCC_intro_products_v2015.pdf.

Schroeder de Witt, C., C. Tong, V. Zantedeschi, D. De Martini, A. Kalaitzis, M. Chantry, D. Watson-Parris, and P. Bilinski, 2021: RainBench: Towards data-driven global precipitation forecasting from satellite imagery. *Proc. AAAI Conf. Artif. Intell.*, **35**, 14 902–14 910, https://doi.org/10.1609/aaai.v35i17.17749.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Stellingwerf, S., E. Riddle, T. M. Hopson, J. C. Knievel, B. Brown, and M. Gebremichael, 2021: Optimizing precipitation forecasts for hydrological catchments in Ethiopia using statistical bias correction and multi-modeling. *Earth Space Sci.*, **8**, e2019EA000933, https://doi.org/10.1029/2019EA000933.

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.

——, ——, T. Gneiting, A. H. Fink, M. Klar, and A. Schlueter, 2021: Statistical forecasts for the occurrence of precipitation outperform global models over northern tropical Africa. *Geophys. Res. Lett.*, **48**, 2020GL091022, https://doi.org/10.1029/2020GL091022.

Walz, E.-M., A. Henzi, J. Ziegel, and T. Gneiting, 2024: Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output. *SIAM Rev.*, **66**, 91–122, https://doi.org/10.1137/22M1541915.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002109, https://doi.org/10.1029/2020MS002109.

Zadrozny, B., and C. Elkan, 2002: Transforming classifier scores into accurate multiclass probability estimates. *Proc. Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, ACM Digital Library, 694–699, https://doi.org/10.1145/775047.775151.

Zhang, Y., M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang, 2023: Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, **619**, 526–532, https://doi.org/10.1038/s41586-023-06184-4.

Unauthenticated | Downloaded 11/03/25 11:49 AM UTC