

Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa

PETER VOGEL

Institute of Meteorology and Climate Research, and Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany

PETER KNIPPERTZ, ANDREAS H. FINK, AND ANDREAS SCHLUETER

Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

TILMANN GNEITING

Heidelberg Institute for Theoretical Studies, Heidelberg, and Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany

(Manuscript received 25 August 2017, in final form 18 December 2017)


ABSTRACT


Accumulated precipitation forecasts are of high socioeconomic importance for agriculturally dominated societies in northern tropical Africa. In this study, the performance of nine operational global ensemble prediction systems (EPSs) is analyzed relative to climatology-based forecasts for 1–5-day accumulated precipitation based on the monsoon seasons during 2007–14 for three regions within northern tropical Africa. To assess the full potential of raw ensemble forecasts across spatial scales, state-of-the-art statistical post-processing methods were applied in the form of Bayesian model averaging (BMA) and ensemble model output statistics (EMOS), and results were verified against station and spatially aggregated, satellite-based gridded observations. Raw ensemble forecasts are uncalibrated and unreliable, and often underperform relative to climatology, independently of region, accumulation time, monsoon season, and ensemble. The differences between raw ensemble and climatological forecasts are large and partly stem from poor prediction for low precipitation amounts. BMA and EMOS postprocessed forecasts are calibrated, reliable, and strongly improve on the raw ensembles but, somewhat disappointingly, typically do not outperform climatology. Most EPSs exhibit slight improvements over the period 2007–14, but overall they have little added value compared to climatology. The suspicion is that parameterization of convection is a potential cause for the sobering lack of ensemble forecast skill in a region dominated by mesoscale convective systems.

1. Introduction

The bulk of precipitation in the tropics is related to moist convection, in contrast to the frontal-dominated extratropics. Because of the small-scale processes involved in

the triggering and growth of convective systems, quantitative precipitation forecasts are known to have overall poorer levels of skill in tropical latitudes (Haiden et al. 2012). This can be monitored in quasi-real time via the World Meteorological Organization (WMO) Lead Centre on Verification of Ensemble Prediction System website (<http://epsv.kishou.go.jp/EPSv>) by comparing deterministic and probabilistic skill scores for 24-h precipitation forecasts for the 20°N–20°S tropical belt with those for the Northern and Southern Hemisphere extratropics. There are hints that precipitation and cloudiness forecasts in the tropics show enhanced skill during regimes of stronger synoptic-scale forcing (Söhne et al. 2008; Davis et al. 2013; Van der Linden et al. 2017) or in regions of orographic forcing (Lafore et al. 2017), but large parts of

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

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-17-0127.s1>.

Corresponding author: Peter Vogel, p.vogel@kit.edu

DOI: 10.1175/WAF-D-17-0127.1

© 2018 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

the tropical landmasses are dominated by convection that initiates from small-scale surface and boundary layer processes and sometimes is organized into mesoscale convective systems (MCSs). The latter depends mostly on the thermodynamic profile and vertical wind shear.

Within this context, northern tropical Africa, particularly the semiarid Sahel, can be considered a region where precipitation forecasting is particularly challenging. The area consists of vast flatlands, where MCSs during boreal summer provide the bulk of the annual rainfall (Mathon et al. 2002; Fink et al. 2006; Houze et al. 2015) and convergence lines in the boundary layer or soil moisture gradients at the kilometer scale can act as triggers for MCSs (Lafore et al. 2017). Sahelian MCSs often take the form of meridionally elongated squall lines with sharp leading edges characterized by heavy rainfall. Synoptic-scale African easterly waves are known to be linked to squall-line occurrence in the western Sahel (Fink and Reiner 2003) and lead to enhanced skill in cloudiness forecasts over West Africa (Söhne et al. 2008).

However, numerical weather prediction (NWP) models are known to have an overall poor ability to predict rainfall systems over northern Africa. For example, the gain in skill by improved initial conditions due to an enhanced upper-air observational network during the 2006 African Monsoon Multidisciplinary Analysis (AMMA) campaign (Parker et al. 2008) was lost in NWP models after 24 h of forecast time, potentially because of the models' inability to predict the genesis and evolution of convective systems (Fink et al. 2011).

Given the substantial challenges involved in forecasting rainfall in northern Africa, one might hope that ensemble prediction systems (EPSs) provide an accurate assessment of uncertainties and a more useful forecast overall. An ensemble is a set of deterministic forecasts, created by changes in the initial conditions and/or the numerical representation of the atmosphere (Palmer 2002). With clear advantages of ensembles over single deterministic forecasts, EPSs are now run at all major NWP centers, which led to the creation of the TIGGE multimodel ensemble database (Bougeault et al. 2010; Swinbank et al. 2016). TIGGE contains forecasts from up to 10 global EPSs, with the ensemble of the European Centre for Medium-Range Weather Forecasts (ECMWF) being the most prominent and most important contributor (Hagedorn et al. 2012). To our knowledge, this present study is the first to rigorously and systematically assess the quality of ensemble forecasts for precipitation over northern tropical Africa. This is partly related to the fact that for this region ground verification data from rain gauge observations are infrequent on the Global Telecommunication System (GTS), the standard verification data source for NWP centers.

Despite many advances in the generation of EPSs, ensembles share structural deficiencies such as dispersion errors and biases. Statistical postprocessing addresses these deficiencies and realizes the full potential of ensemble forecasts (Gneiting and Raftery 2005). Additionally, it performs implicit downscaling from the model grid resolution to finer resolutions or station locations. The correction of systematic forecast errors is based on (distributional) regression techniques and, depending on the need of the user, several approaches are at hand (Scheffzik et al. 2013; Gneiting 2014). Hamill et al. (2004) and Wilks (2009) proposed and extended logistic regression techniques, which yield probabilistic forecasts for the exceedance of thresholds. Here, we will for the first time explore whether established methods such as Bayesian model averaging (BMA; Raftery et al. 2005) and ensemble model output statistics (EMOS; Gneiting et al. 2005), which provide complete probabilistic quantitative precipitation forecasts, can improve precipitation forecasts for Africa.

The ultimate goal of this paper is to provide an exhaustive assessment of our current ability to predict rainfall over northern tropical Africa, considering the skill of raw and postprocessed forecasts from TIGGE. Any skill, if existing, would be expected to come from resolved large-scale forcing processes as mentioned above. We examine accumulation periods of 1–5 days for the monsoon seasons 2007–14 and verify against about 21 000 daily rainfall observations from 132 rain gauge stations and satellite-based gridded precipitation observations. Section 2 introduces the TIGGE ensemble, as well as the station and satellite-based observations used for verification. Section 3 describes our benchmark climatological forecast and methods for the evaluation of probabilistic forecasts and explains EMOS and BMA in detail. Results are presented in section 4, where we verify 1-day accumulated ECMWF precipitation forecasts against station observations. This analysis is performed in particular depth and serves as a fundamental exemplar. We also evaluate ECMWF ensemble forecasts at longer accumulation times and for spatial aggregations, before turning to the analysis of all TIGGE subensembles. Implications of our findings and possible alternative methods for forecasting precipitation over northern tropical Africa are discussed in section 5.

2. Data

a. Forecasts

The TIGGE multimodel ensemble was set up as part of the THORPEX program in order to “accelerate improvements in the accuracy of 1-day to 2-week

TABLE 1. TIGGE subensembles used in this study, with years of availability, number of ensemble members (number of perturbed members + control run + any high-resolution run), initialization time, and native grid(s) used during the period of 2007–14.

Source	Acronym	Availability	Members	Initialization time (UTC)	Native grid(s)
China Meteorological Administration	CMA	2008–13	14 + 1	0000	TL213/T639
Centro de Previsão Tempo e Estudos Climáticos	CPTEC	2008–14	14 + 1	0000	T126
European Centre for Medium-Range Weather Forecasts	ECMWF	2007–14	50 + 1 + 1	0000	T399/T639
Japan Meteorological Agency	JMA	2007–13/14	50/26 + 1	1200	TL159/TL319/TL479
Korea Meteorological Administration	KMA	2011–14	16 + 1	0000	N320
Météo-France	MF	2010–14	34 + 1	0600	TL798
Meteorological Service of Canada	MSC	2008–14	20 + 1	0000	0.45° uniform
National Centers for Environmental Prediction	NCEP	2008–14	20 + 1	0000	T126
Met Office	UKMO	2007–13	23 + 1	0000	N144/N216/N400

high-impact weather forecasts for the benefit of humanity” (Bougeault et al. 2010, p. 1060). Since its start in October 2006, up to 10 global NWP centers have provided their operational ensemble forecasts, which are accessible on a common $0.5^\circ \times 0.5^\circ$ grid. Park et al. (2008) and Bougeault et al. (2010) discuss objectives and the setup of TIGGE, including the participating EPSs, in great detail. They also note early results using the TIGGE ensemble, while Swinbank et al. (2016) report on achievements accomplished over the last decade. Hagedorn et al. (2012) find that a multimodel ensemble composed of the four best participating TIGGE EPSs, which include the ECMWF ensemble, outperforms reforecast-calibrated ECMWF forecasts. For the evaluation of NWP precipitation forecast quality, TIGGE is the most complete and best available data source for the period 2007–14. Table 1 gives an overview of the nine participating TIGGE EPSs that provide accumulated precipitation forecasts.

In addition to the separate evaluation of each participating TIGGE subensemble, we construct a reduced multimodel (RMM) ensemble. For each of the seven subensembles available for the period 2008–13, the RMM ensemble uses the mean of the perturbed members, and the control run, and in the case of the ECMWF EPS, furthermore, the high-resolution run, as individual contributors. The RMM ensemble therefore consists of 15 members and, as postprocessing performs an implicit weighting of all contributions, a manual selection of subensembles as performed by Hagedorn et al. (2012) is not necessary.

Arguably, the ECMWF EPS is the leading example among the TIGGE subensembles (Buizza et al. 2005; Hagedorn et al. 2012; Haiden et al. 2012). It consists of a high-resolution (HRES) run, a control (CNT) run, and 50 perturbed ensemble (ENS) members. The HRES and CNT runs are started from unperturbed initial conditions and differ only in their resolution. The ENS members are started from perturbed initial conditions and have the same resolution as the CNT run. Molteni et al. (1996) and

Leutbecher and Palmer (2008) describe the generation and properties of the ECMWF system in detail.

b. Observations

Despite multiple advances in satellite rainfall estimation, station observations of accumulated precipitation remain a reliable and necessary source of information. However, the meteorological station network in tropical Africa is sparse and clustered, and observations of many stations are not distributed through the GTS. The Karlsruhe African Surface Station Database (KASS-D) contains precipitation observations from a variety of networks and sources. Manned stations operated by African national weather services provide the bulk of the 24-h precipitation data. Due to long-standing collaborations with these services and African researchers, KASS-D contains many observations not available in standard, GTS-fed station databases. Within KASS-D, 960 stations have daily accumulated (usually 0600–0600 UTC) precipitation observations.

After excluding stations outside the study domain, and removing sites with less than 80% available observations in any of the monsoon seasons, the remaining 132 stations were subject to quality control, as described in the appendix, and passed these tests. Based on their rainfall climate (e.g., Fink et al. 2017) and geographic clustering, the stations were assigned to three regions, as indicated in Fig. 1, referred to in this paper as West Sahel, East Sahel, and Guinea Coast.

As NWP forecasts are issued for grid cells, the comparison of station observations against gridded forecasts is fraught with problems. To allow for an additional assessment of forecast quality without a gauge-to-gridbox comparison and for areas without station observations, we use satellite-based, gridded precipitation estimates. Based on recent studies, version 7 (and also version 6) of the Tropical Rainfall Measuring Mission (TRMM) 3B42 gridded dataset is regarded the best available satellite precipitation product, despite a small dry bias (Roca et al. 2010; Maggioni et al. 2016; Engel et al. 2017).

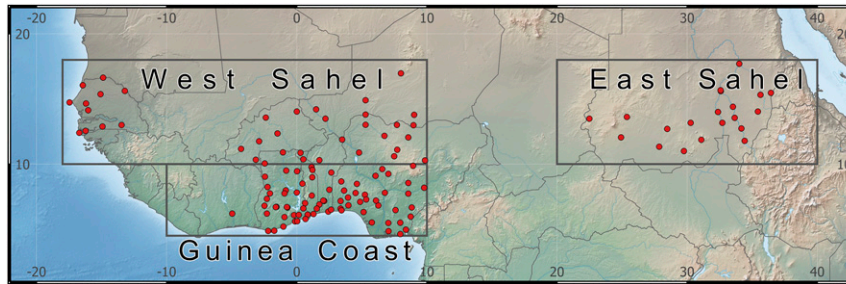


FIG. 1. Geographical overview of the study domain, with the locations of the observation stations (dots) within the three considered regions.

TRMM merges active measurements from the precipitation radar with passive, radar-calibrated information from infrared as well as microwave measurements (Huffman et al. 2007). Based on monthly accumulation sums, TRMM estimates are calibrated against nearby gauge observations. TRMM 3B42-V7 data are available on a $0.25^\circ \times 0.25^\circ$ grid with 3-hourly temporal resolution.

c. Data preprocessing

Based on 1-day accumulated station observations, we derive 2–5-day accumulated precipitation observations by summing over consecutive 1-day observations. As these cover the period from 0600 UTC of the previous day to 0600 UTC of the considered day and as all TIGGE subensembles, except Météo-France (MF), have initialization times different from 0600 UTC, we use the most recent run available at that time and adapt accordingly. Specifically, for the subensembles initialized at 0000 UTC, we use the difference between the 30-h accumulated and the 6-h accumulated precipitation forecasts. For initialization at 1200 UTC, we use the difference between the 42-h accumulated and the 18-h accumulated precipitation forecasts, and for longer accumulation times, we extend this process correspondingly.

To obtain forecasts for a specific station location from gridded NWP forecasts, both bilinear interpolation as well as a nearest-neighbor approach are possible. We use the latter, implying that the forecast for the station is the same as the forecast for the grid cell containing the station. Especially for large gridbox sizes, bilinear interpolation may not be physically persuasive, and the nearest-neighbor approach is more compelling.

TRMM observations are temporally aggregated to the same periods as the station observations. As they do not cover the exact same periods, the first and last 3-h TRMM observations are weighted by 0.5. For evaluation on different spatial scales, NWP forecasts and TRMM observations are aggregated to longitude–latitude boxes of $0.25^\circ \times 0.25^\circ$, $1^\circ \times 1^\circ$, and $5^\circ \times 2^\circ$. As propagation of precipitation systems is a potential error source and in an

environment with predominantly westward movement of these systems, the largest box is tailored to assess NWP forecast quality without this potential source of error.

d. Consistency between TRMM and station observations

In light of the dry bias of the TRMM observations, we evaluate the consistency of TRMM and station observations in our datasets. Specifically, we pair each station observation with the TRMM observation for the $0.25^\circ \times 0.25^\circ$ box that contains the station location. Figure 2 shows contingency tables of TRMM and station observations above and below 0.2 mm, respectively, and two-dimensional frequency plots for TRMM and station observations above 0.2 mm, which is our threshold for the distinction between rain and no rain throughout the paper, as discussed in section 3b. For all regions the prevailing case is the one with both TRMM and the station reporting precipitation amounts below 0.2 mm. Among the disagreeing cases, the one with TRMM observing more than 0.2 mm and the station less than 0.2 mm is more frequent, coinciding with the intuition that a station is more likely to miss a precipitation event reported by TRMM than vice versa. The least squares regression lines in the two-dimensional frequency plots illustrate the dry bias of TRMM relative to station observations when both report rain. Overall, the agreement between the station and TRMM observations is fair. Disagreements of the magnitude and type seen here arise for reasons of differing coverage, spatial variability, and retrieval problems, among other concerns, and are compatible with the extant literature (see, e.g., Roca et al. 2010; Engel et al. 2017).

3. Methods

Probabilistic forecasts are meant to provide calibrated information about future events. To be of use, they should satisfy two properties. First, they should convey correct probabilistic statements, in that observations behave like random draws from the forecast distributions. This property

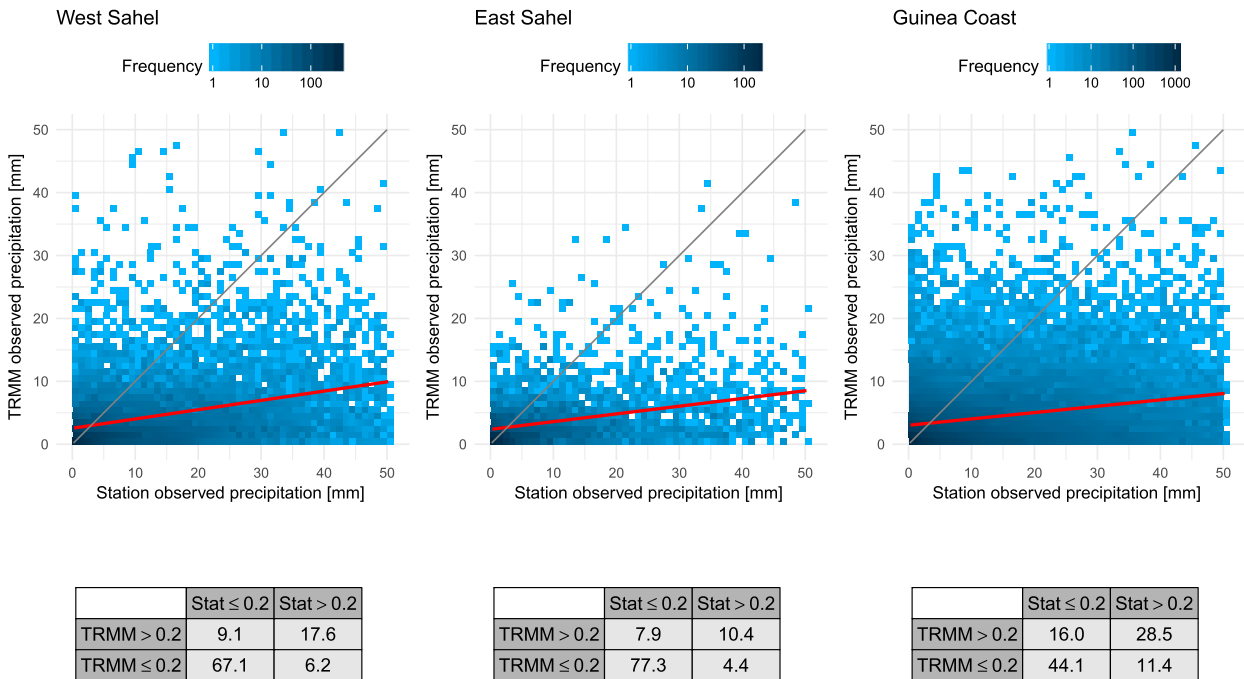


FIG. 2. Comparison of 1-day accumulated station and TRMM observations of precipitation during the monsoon seasons of 2007–14. The contingency tables contain the frequencies of TRMM and station observations below and above 0.2 mm, respectively. The two-dimensional frequency plots show the joint distribution of TRMM and station observations above 0.2 mm, with the linear least squares line overlaid. Observations above 50 mm exist, but are very infrequent.

is called calibration. Second, under all calibrated forecasts, sharper ones with lesser uncertainty are preferred.

a. Reference forecasts

For the assessment of raw and postprocessed ensemble forecast skill, the availability of a benchmark forecast is essential. Here, we introduce the concept of a probabilistic climatology that consists of the observations during the 30 years prior to the considered year at the considered day of the year and location. This can be understood as a 30-member observation-based ensemble forecast that represents the climatological distribution of rainfall at a given location and date, but does not incorporate dynamic information about the state of the atmosphere. We extend the probabilistic climatology by including observations in a ± 2 -day window around the considered day and refer to this as the extended probabilistic climatology (EPC). Our findings generally are insensitive to the range of the window being chosen from ± 2 to ± 20 days as shown in Fig. S1 in the online supplemental material.

Hamill and Juras (2006) note that pooling can lead to a deterioration when performed across data with differing climatologies, leading to a perceived, but incorrect improvement of assessed model forecast skill. In our case, however, neighboring daily climatologies are

very similar, and the pooling is performed over a range of ± 2 days only. EPC has better forecast quality than standard probabilistic climatologies (Fig. S1) and is used as benchmark in the following. As TRMM observations are available for the period 1998–2014 only, the TRMM-based EPC relies on this period without the considered verification year.

b. Assessing calibration: Unified probability integral transform histograms

Verification rank histograms and probability integral transform (PIT) histograms are standard tools for the assessment of calibration, and we refer the reader to Hamill (2001), Gneiting et al. (2007), and Wilks (2011) for in-depth discussions of their use and interpretation. In a nutshell, for calibrated probabilistic forecasts, rank and PIT histograms are uniform, U-shaped histograms indicate underdispersion, and skewed histograms mark biases.

For an ensemble forecast, the verification rank is the rank of the observation when it is pooled with the m ensemble members; clearly, this is an integer between 1 and $m + 1$. If k members predict no precipitation, and no precipitation is observed, the rank is randomly drawn between 1 and $k + 1$. For a probabilistic forecast in the form of a cumulative distribution function (CDF) F

and a verifying precipitation accumulation $y > 0$, the PIT is the value $F(y)$ of the forecast CDF evaluated at the observation. In the case of no precipitation, a value is randomly drawn between 0 and the forecast probability of no precipitation (Sloughter et al. 2007).

In the present study, we compare raw ensemble forecasts to postprocessed forecasts in the form of CDFs, and the TIGGE subensembles have varying numbers of members. We use the term probabilistic quantitative precipitation forecast (PQPF) to denote all these types of forecasts. To allow a compelling visual assessment of calibration in this setting, we introduce the notion of a unified PIT (uPIT). For a forecast in the form of a CDF, the uPIT is simply the PIT. For an ensemble forecast with m members, if the observation has rank i and this rank is unique, the uPIT is a random number from a uniform distribution between $(i-1)/(m+1)$ and $i/(m+1)$. If k members predict no precipitation, and no precipitation is observed, the uPIT is a random number between 0 and $(k+1)/(m+1)$. It is readily seen that for a calibrated PQPF the uPIT is uniformly distributed. Hereinafter, we use 20 equally spaced bins to plot uPIT histograms.

Our uPIT histograms focus on calibration regarding the forecasted precipitation amount. However, any PQPF induces a probability of precipitation (PoP) forecast for the binary event of rainfall occurrence at any given threshold value. We use a threshold of 0.2 mm to define rainfall occurrence irrespectively of the temporal and spatial aggregation at hand, with the results reported on hereinafter being insensitive to this choice.¹ Reliability, the equivalent of calibration for probability forecasts of binary events, means that events declared to have probability p occur a proportion p of the time. This can be checked empirically in reliability diagrams, where the observed frequency of occurrence is plotted versus the forecast probability (e.g., Wilks 2011).

c. Proper scoring rules

For the comparative evaluation of predictive skill, we use proper scoring rules that assess calibration and sharpness simultaneously (Gneiting and Raftery 2007; Wilks 2011). Specifically, the continuous ranked probability score (CRPS) for a PQPF with CDF F and a verifying observation y is defined as

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(x) - 1(x \geq y)]^2 dx,$$

¹ Specifically, we checked thresholds from 0.0 to 1.0 mm, with minimal differences in findings. Exemplary results are available from the authors upon request.

where 1 is an indicator function, equal to 1 if the argument is true and equal to 0 otherwise. From every PQPF, we can extract a deterministic forecast and compute its absolute error (AE). If the deterministic forecast is chosen to be the median of the forecast distribution, the AE can be interpreted as a proper scoring rule (Gneiting 2011; Pinson and Hagedorn 2012).² Both the AE and the CRPS are negatively oriented, and they are reported in the unit of the observation (here, millimeters) and so can be compared directly. In fact, if the forecast distribution is a deterministic forecast, the CRPS reduces to the AE (Gneiting and Raftery 2007).

With the PoP being an essential component of a PQPF, the evaluation of PoP forecast quality by proper scoring rules is desirable and can be accomplished by means of the Brier score (BS; Brier 1950). For a probability forecast p for a binary event to occur, the negatively oriented BS is $(1-p)^2$ if the event occurs and p^2 if it does not occur.

It is well known that not only the BS, but many proper scoring rules for probability forecasts of binary events exist and that forecast rankings can depend on the choice of the proper scoring rule. However, every proper scoring rule admits a representation as a weighted average over so-called elementary scores or losses S_θ , which can be interpreted economically. Specifically, suppose that we are given a probability forecast p for a binary event and need to make a deterministic forecast of whether or not it will happen. If correct decisions do not incur any costs, a false alarm carries cost θ , and a missed event has cost $1-\theta$ for some $\theta \in (0, 1)$, an optimal strategy is to predict that the event will happen when $p > \theta$ and to predict that it will not happen when $p < \theta$.³ The elementary score S_θ is the loss incurred by this strategy. Ehm et al. (2016) advocate the use of so-called Murphy diagrams, which display, for each forecast considered, the mean elementary score as a function of $\theta \in (0, 1)$. If a forecast receives a lower elementary score than another for every θ , then it is preferable for any decision-maker and receives lower scores under just any proper scoring rule (Ehm et al. 2016). Interestingly, the area under a forecast's graph in a Murphy diagram equals half its mean BS, and the height of the graph at $\theta = 1/2$ equals half the misclassification rate when false alarms and misses incur equal costs.

² For this desirable interpretation to be valid, the deterministic forecast needs to be chosen as the median of the forecast distribution. For the mathematical argument and technical details see the review article by Gneiting (2011) and the references therein.

³ When $p = \theta$, either action can be taken. These results are elementary and well known; see Ehm et al. (2016) and the references therein.

A popular graphical tool for the assessment of discrimination ability in binary prediction problems is the receiver operating characteristic (ROC) diagram; for details of which we refer the reader to section 8.4.7 of Wilks (2011). In a nutshell, for any given probability forecast, the ROC curve is a plot of the hit rate versus the false alarm rate as a function of the cutoff value for the binary decision. The area under the ROC curve (AUC) is commonly used as a measure of resolution and discrimination skill, with higher values being preferable. In contrast to Murphy diagrams, which consider both reliability and discrimination and assess the actual value of a forecast in decision-making, ROC curves and AUC values are insensitive to (any lack of) reliability and, therefore, reflect potential skill and value only (Wilks 2011, p. 346).

d. Statistical postprocessing

Statistical postprocessing addresses structural deficiencies of NWP model output. Here, we use the well-established methods of EMOS (Gneiting et al. 2005; Scheuerer 2014) and BMA (Raftery et al. 2005; Sloughter et al. 2007) to correct for systematic errors in ensemble forecasts of precipitation accumulation.

In this section, we review these methods with a focus on the 52-member ECMWF EPS, and we denote the values of its HRES, CNT, and ENS members by x_{HRES} , x_{CNT} , and x_1, \dots, x_{50} , respectively. We write \bar{x}_{ENS} for the mean of the ENS members, \bar{p} for the fraction (out) of (all 52) members that predict no precipitation, and denote the observed precipitation accumulation by y . Adaptations of the postprocessing schemes to the other TIGGE subensembles and the RMM ensemble are straightforward.

1) ENSEMBLE MODEL OUTPUT STATISTICS

The idea of the EMOS approach is to convert an ensemble forecast into a parametric distribution, based on the ensemble forecast at hand (Gneiting et al. 2005). Scheuerer (2014) introduced an EMOS approach for precipitation accumulation that relies on the three-parameter family of left-censored generalized extreme value (GEV) distributions. The left-censoring allows for a point mass at zero and the shape parameter for flexible skewness in positive precipitation accumulations.

Briefly, the EMOS predictive distribution based on the ECMWF ensemble is a left-censored GEV distribution. The location parameter of this distribution is a linear function of x_{HRES} , x_{CNT} , \bar{x}_{ENS} , and \bar{p} , and its scale parameter is a linear function of the ensemble mean difference, which is a more robust measure of ensemble spread than the standard deviation. While all parameters are estimated from training data, the shape parameter does not link to the ensemble values (Scheuerer 2014).

For illustration, Fig. 3a shows an EMOS postprocessed forecast distribution for 5-day accumulated precipitation at Ouagadougou, Burkina Faso. The 52 raw ECMWF ensemble members are represented by blue marks; they include 11 values in excess of 200 mm, with the CNT member being close to 500 mm. The ensemble forecast at hand informs the statistical parameters of the EMOS postprocessed forecast distribution, which includes a tiny point mass at zero, and a censored GEV density for positive precipitation accumulations, with the 90th percentile being at 174 mm.

2) BAYESIAN MODEL AVERAGING

A BMA predictive distribution is a weighted sum of component distributions, each of which depends on a single ensemble member. For the ECMWF ensemble, the BMA method for precipitation accumulation proposed and studied by Sloughter et al. (2007) and Fraley et al. (2010) implies a statistical model of the form

$$y|x_{\text{HRES}}, x_{\text{CNT}}, x_1, \dots, x_{50} \sim w_{\text{HRES}} g_{\text{HRES}}(y|x_{\text{HRES}}) + w_{\text{CNT}} g_{\text{CNT}}(y|x_{\text{CNT}}) + \frac{w_{\text{ENS}}}{50} \sum_{i=1}^{50} g_{\text{ENS}}(y|x_i),$$

with nonnegative weights w_{HRES} , w_{CNT} , and w_{ENS} that sum to 1, and reflects the members' performance during the training period.⁴ Each of the component distributions, g_{HRES} , g_{CNT} , and g_{ENS} , contains a point mass at zero and a density for positive accumulations. The point mass at zero specifies the probability of no precipitation and is estimated in a logistic regression model, where the cube root of the member forecast and a binary indicator of the member forecast being zero are used as predictor variables. The specification for positive amounts is based on a gamma density for the cube-root-transformed precipitation amount, with a mean that is a linear function of the cube-root-transformed member forecast and a variance that is a linear function of the member forecast. While the statistical coefficients for the mean of the gamma model are estimated for g_{HRES} , g_{CNT} , and g_{ENS} separately, the coefficients for the variance of the gamma model are shared. To obtain the BMA predictive distribution for the linear precipitation accumulation in millimeters, rather than the cube root thereof, a backtransformation is applied as described by Sloughter et al. (2007).

Figure 3b shows such a BMA postprocessed forecast distribution for the aforementioned forecast

⁴ Within this context, we take the chance to correct a typographical error in Fraley et al. (2010), where the factor $1/m_i$ is missing in between the summation signs in their Eq. (5).

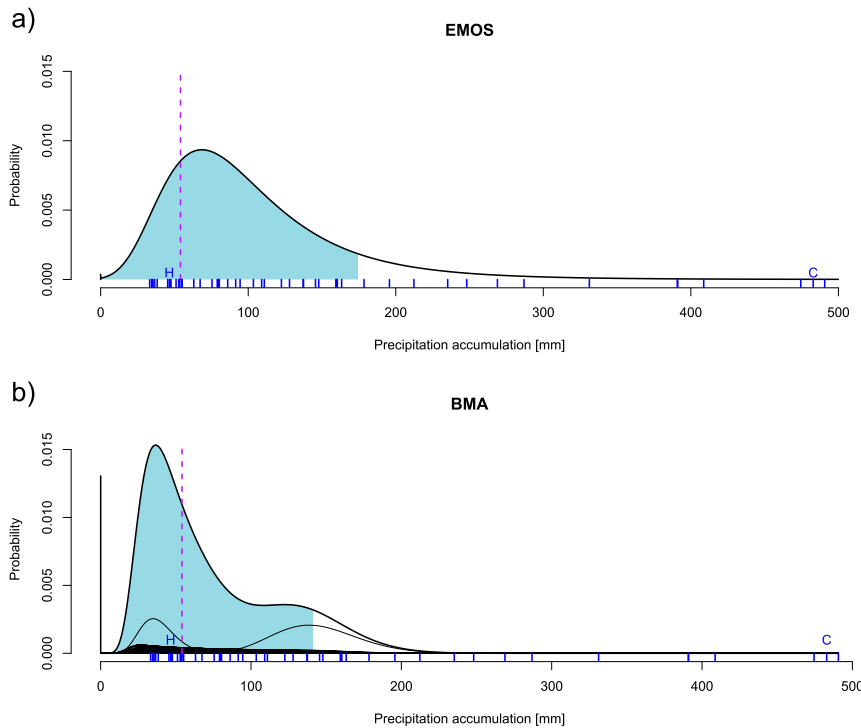


FIG. 3. EMOS and BMA postprocessed ECMWF ensemble forecasts for 5-day accumulated precipitation at Ouagadougou, valid 3–8 Aug 2007. The blue marks at bottom represent the 52 raw ECMWF ensemble members, including the HRES (H) run, the CNT (C) run, and the 50 perturbed ENS members. (a) The EMOS postprocessed forecast includes a tiny point mass at zero and a censored GEV density for positive accumulations. (b) The BMA postprocessed forecast includes a point mass at zero, which is represented by the solid bar, and a mixture of power-transformed Gamma densities for positive accumulations. The 52 component densities are represented by the thin black curves, with the HRES and CNT components standing out. The lower 90% prediction interval is indicated in light blue, and the dashed bar represents the verifying precipitation accumulation.

case at Ouagadougou. The postprocessed distribution involves a point mass of about 0.01 at zero, and a mixture of power-transformed gamma densities for positive accumulations, with the 90th percentile being at 141 mm. In this example, the BMA and EMOS postprocessed distributions are sharper than the raw ECMWF ensemble, and nevertheless the verifying accumulation is well captured.

Adaptations to the other ensembles considered in this paper are straightforward, as described by Fraley et al. (2010). For example, in the case of the RMM ensemble each of the 15 contributors receives its own component distribution, BMA weight, logistic regression coefficients for the probability of no precipitation, and statistical parameters for the gamma mean model, whereas the coefficients for the gamma variance model are shared.

3) ESTIMATION OF STATISTICAL PARAMETERS

Postprocessing techniques such as EMOS and BMA rely on statistical parameters that need to be estimated

from training data, comprising forecast–observation-pairs either from the station or TRMM pixel at hand, or from all stations or applicable TRMM pixels within the considered region, and typically from a rolling training period consisting of the n most recent days for which data are available at the initialization time. We employ the regional approach with a rolling training period of $n = 20$ days, which yields superior results, consistent with the literature (e.g., Thorarinsdottir and Gneiting 2010). As shown in Figs. S2–S5 in the supplementary material, our findings are insensitive to the choice of n when using training periods between 20 and 50 days. The local approach requires longer training periods and (in experiments not shown here) yields very similar results then.

For EMOS, parameter estimation is based on CRPS minimization over the training data, which is computationally efficient, as closed expressions for the CRPS under GEV distributions are available (Scheuerer 2014).

For BMA, we employ maximum likelihood estimation, implemented via the expectation-maximization (EM) algorithm developed by Sloughter et al. (2007). All computations were performed in R (R Development Core Team 2017) based on the ensembleBMA package (Fraley et al. 2011) and code supplied by M. Scheuerer.

4. Results

Our annual evaluation period ranges from 1 May to 15 October, covering the wet period of the West African monsoon. The assessment of ECMWF ensemble forecasts is based on monsoon seasons 2007–14, and for the other TIGGE subensembles we restrict our evaluation according to availability as indicated in Table 1.

For verification against station observations, this yields more than 3000, 6000, and 12 000 forecast–observations pairs per monsoon season in East Sahel, West Sahel, and Guinea Coast. For verification against TRMM observations, we use 30 randomly chosen, non-overlapping boxes per region at $0.25^\circ \times 0.25^\circ$ and $1^\circ \times 1^\circ$ aggregation and eight sites per region for $5^\circ \times 2^\circ$ longitude–latitude boxes. This covers substantial parts of the study region and results in about 5000 forecast–observation pairs per monsoon season at the smaller aggregation levels and well over 1000 pairs at our highest level.

In section 4a, we study the skill of 1-day accumulated ECMWF raw and postprocessed ensemble precipitation forecasts in detail. Sections 4b and 4c present results and highlight differences for longer accumulation times and spatially aggregated forecasts. Section 4d turns to results for all TIGGE subensembles, and we investigate the gain in predictability through intermodel variability using the RMM ensemble. In our uPIT histograms and reliability diagrams, we show results for the last available monsoon season only (2014), given that operational systems continue to be improving (Hemri et al. 2014).

a. 1-day accumulated ECMWF forecasts

Figure 4 shows uPIT histograms for 1-day accumulated raw and postprocessed ECMWF ensemble and EPC forecasts over West Sahel, East Sahel, and Guinea Coast. The histograms for the raw ensemble indicate strong underdispersion as well as a wet bias (Figs. 4a–c). At Guinea Coast, about 56% of the observations are smaller than the smallest ensemble member, a result that is robust across monsoon seasons. EMOS and BMA postprocessed forecasts generally are calibrated (Figs. 4g–i), as is EPC (Figs. 4d–f), except that the tails of the EMOS predictive distributions are too light. Statistical postprocessing also corrects for the systematically too-high PoP values issued by the raw ECMWF

ensemble. As shown in Fig. 5, EMOS and BMA postprocessed PoP forecasts are reliable, but are hardly ever higher than 0.70. Generally, the postprocessed PoP forecasts have reliability and resolution similar to EPC.

Table 2 shows the mean BS, mean CRPS, and mean absolute error (MAE) for the various forecasts and regions, with the scores being averaged across monsoon seasons 2007–14. We use a simple procedure to check whether differences in skill are stable across seasons. If a method has a higher (worse) mean score than EPC during all eight seasons, we mark the score with $^-^-$; if it is judged to be worse during seven seasons, we use a $^-$. Similarly, if a method has a smaller (better) mean score than EPC during all seasons, we mark the score as $^+^+$; if it performs better during seven seasons, we label it as $^+$ in Table 2. Viewed as a (one sided) statistical test of the hypothesis of predictive skill equal to EPC, the associated tail probabilities or p values are $1/2^8 = 0.0039\dots$ and $(1+8)/2^8 = 0.035\dots$, respectively. Clearly, the raw ECMWF ensemble underperforms relative to EPC, with the $^-^-$ designations used throughout, and the EMOS and BMA postprocessed forecasts perform at about the same level as EPC. For the BS, the similar performance of postprocessed and EPC forecasts stems from the fact that not only do postprocessed and EPC forecasts show similar reliability but also similar resolution, as seen from the inset histograms in Figs. 5d–l.

The Murphy diagrams in the top row of Fig. 6 corroborate these findings. For 1-day precipitation occurrence, decision-makers will mostly prefer the climatological reference EPC over the raw ECMWF ensemble, and only some decision-makers will have a slight preference for EMOS or BMA postprocessed forecasts, as compared to EPC. Further light on these issues is shed by the ROC diagrams in the bottom row of Fig. 6. EMOS and BMA PoP forecasts can be interpreted as recalibrated raw ensemble probabilities, and so it is not surprising that for West Sahel and East Sahel, raw and postprocessed forecasts show essentially the same level of discrimination skill, at a level that is slightly superior to EPC. For Guinea Coast, EMOS and BMA have considerably higher AUC values than the raw ensemble, due to the extreme concentration of the raw ensemble probabilities at very high levels, as illustrated in Fig. 5c. In contrast, the Murphy curves are sensitive to calibration and show marked differences between raw and postprocessed forecasts. Overall, these are sobering results, as they suggest that over northern tropical Africa the ECMWF 1-day accumulated precipitation forecasts are hardly of practical use.

What could be possible reasons for the poor performance of the raw forecasts? A number of recent studies

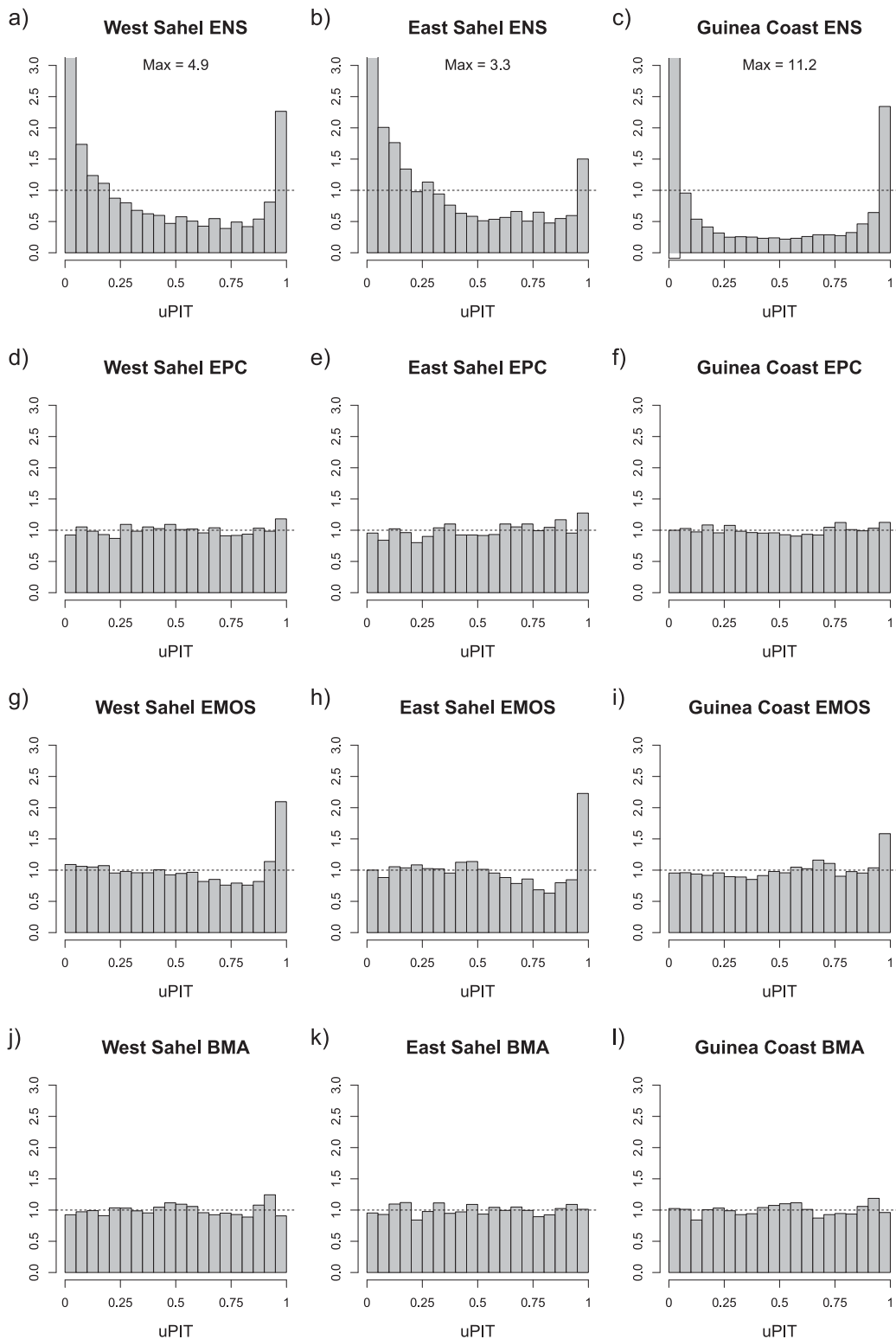


FIG. 4. The uPIT histograms for (a)–(c) raw ECMWF ensemble, (d)–(f) EPC, and (g)–(i) EMOS and (j)–(l) BMA postprocessed forecasts of 1-day accumulated precipitation during the monsoon season of 2014, verified against station observations. Histograms are cut at a height of 3, with the respective maximal height noted. The dashed line indicates the uniform distribution that corresponds to a calibrated forecast.

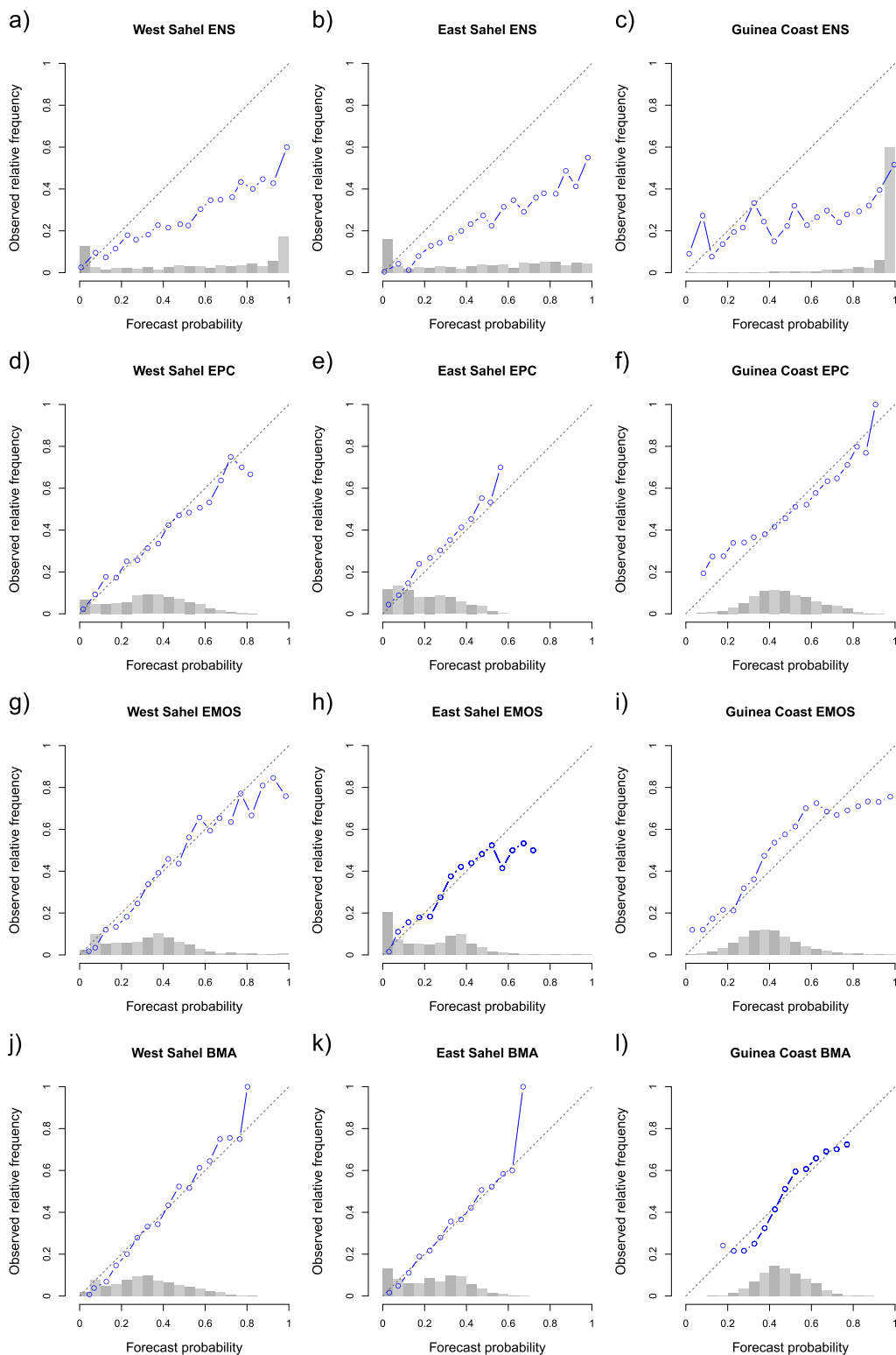


FIG. 5. Reliability diagrams for (a)–(c) raw ECMWF ensemble, (d)–(f) EPC, and (g)–(i) EMOS and (j)–(l) BMA postprocessed forecasts of 1-day accumulated precipitation during the monsoon season of 2014, verified against station observations. The diagonal indicates perfect reliability, and the histograms show the relative frequencies of the PoP forecast values.

TABLE 2. Mean BS at a threshold of 0.2 mm, mean CRPS, and MAE for raw ECMWF ensemble, EPC, and EMOS and BMA post-processed forecasts of 1-day accumulated precipitation during the monsoon seasons of 2007–14, verified against station observations. If a method has a higher (worse) or lower (better) mean score than EPC during all eight seasons, the score is marked with a $^-$ or $^+$, respectively; if it performs worse or better than EPC during seven seasons, the score is marked with a $^-$ or $^+$.

	BS			CRPS			MAE		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	$^-$ 0.32	$^-$ 0.32	$^-$ 0.48	$^-$ 4.50	$^-$ 2.63	$^-$ 6.99	$^-$ 5.36	$^-$ 3.13	$^-$ 8.39
EPC	0.19	0.15	0.23	3.75	2.08	5.28	4.60	2.38	6.57
EMOS	0.19	0.15	0.23	3.75	2.15	$^+$ 5.25	4.65	$^-$ 2.45	6.60
BMA	$^+$ 0.18	0.15	$^+$ 0.22	3.71	2.07	$^+$ 5.20	4.58	2.38	6.53

have shown that the use of convective parameterization is a first-order error source for realistically representing precipitation, cloudiness, wind, and even the regional-scale monsoon circulation in West Africa together with their respective diurnal cycles (e.g., Pearson et al. 2014; Marsham et al. 2013; Birch et al. 2014; Pantillon et al.

2015). Based on these results, and given that all of the models we investigate use convective schemes, we suspect this aspect to be a major cause of the poor performance we find. A visual comparison of 1-day accumulated precipitation forecasts from ECMWF HRES and TRMM shows that rainfall structures in

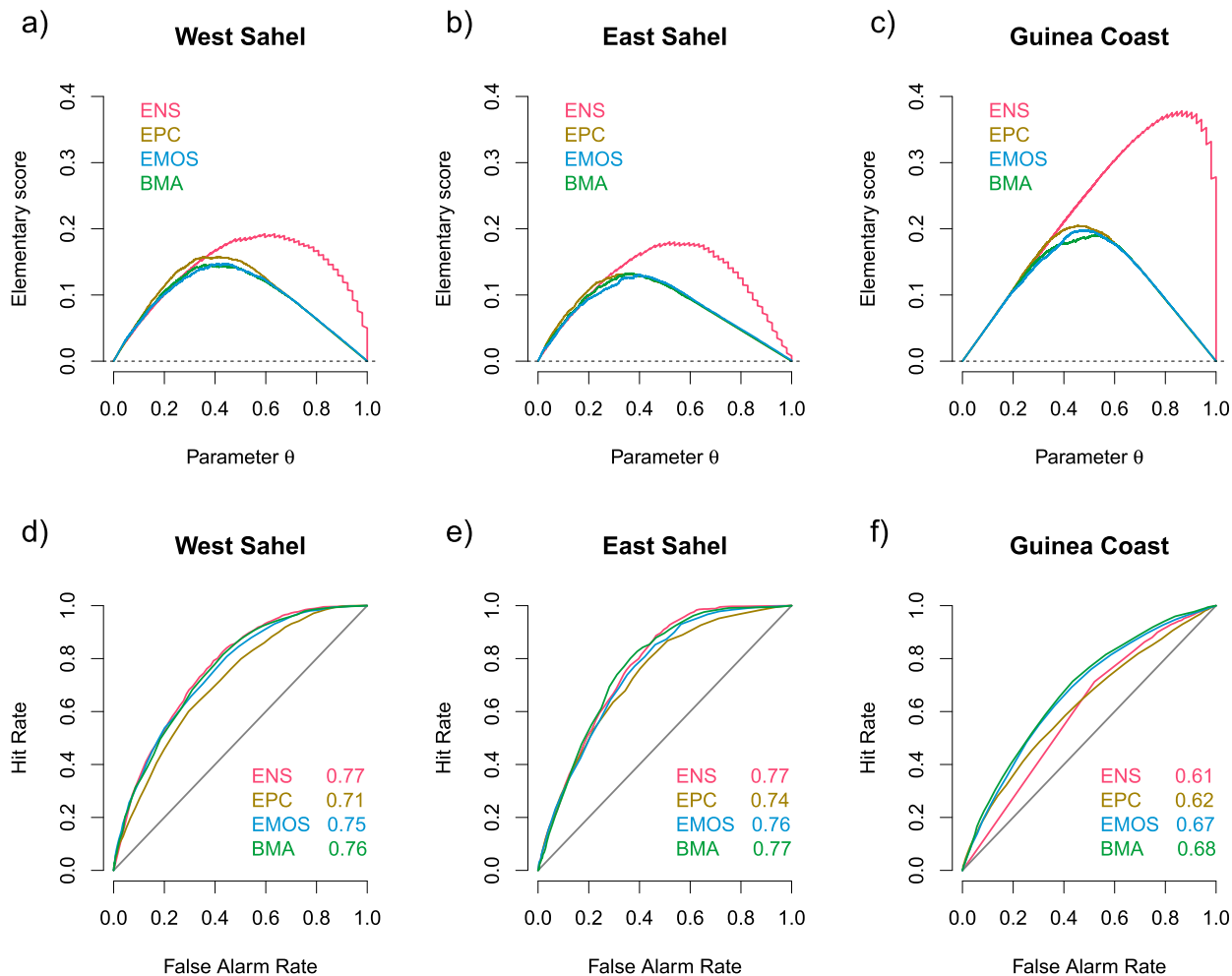


FIG. 6. (a)–(c) Murphy diagrams and (d)–(f) ROC curves (with respective AUC values) for ENS, EPC, and EMOS and BMA postprocessed 1-day accumulated PoP forecasts during the monsoon season of 2014, verified against station observations.

the model tend to be too widespread and too light, lacking signs of mesoscale organization (see Fig. S6 for an example). Inspection of raw ensemble data suggests that, for both station and TRMM observations, agreement between the forecasts and observations is modest at best. Many observed precipitation events are either not predicted at all, are strongly underpredicted, or are predicted by (almost) all ensemble members (with varying amounts of precipitation), yet are not observed (see Fig. S7 for an illustrative example). In particular, the second point is an indication of a misrepresentation of real-world squall-line systems by the model.

b. Longer accumulation times

One might expect NWP precipitation forecasts to improve relative to EPC at longer accumulation times, as the main focus in forecasting shifts from determining time and location of initiation and subsequent propagation of convection toward determining regions with enhanced or reduced activity, based on large-scale conditions. Longer lead times might also lead to growth in differences between perturbed members and, thus, reduce the raw ensemble underdispersion.

The uPIT histogram in Fig. 7a indicates only slight, if any, improvement in calibration for raw ECMWF 5-day accumulated precipitation forecasts over West Sahel, and the results for the other regions are similar (not shown). Raw ensemble reliability improves at longer accumulation times, verified against either station observations in Fig. 7b, or $5^\circ \times 2^\circ$ TRMM observations in Figs. 7c and 7d, though at a loss of resolution.

Table 3 uses the same settings as in Table 2, but the scores are now for 5-day accumulated precipitation. The raw ECMWF ensemble still underperforms relative to EPC. The EMOS and BMA postprocessed forecasts outperform EPC only slightly, with the differences in scores being small and typically not being stable across monsoon seasons. Despite the change in the underlying forecast problem, even postprocessed ECMWF ensemble forecasts are generally not superior to EPC.

c. Spatially aggregated observations

For the assessment of forecast skill at larger spatial scales, we focus on ECMWF raw and BMA postprocessed ensemble forecasts over West Sahel, evaluated by the Brier score and CRPS. This is due to the similarities in CRPS and MAE results, better performance of BMA compared to EMOS in many instances, and results for West Sahel that are as good for BMA postprocessed forecasts as for East Sahel, and better than for Guinea Coast.

The use of spatially aggregated TRMM observations avoids problems of point-to-pixel comparisons, and at

higher aggregation we can assess the forecast quality with minimal propagation error. The dry bias of TRMM disadvantages the raw ensemble compared to EPC and postprocessed forecasts, but does not hinder assessments regarding systematic forecast errors. As illustrated in Fig. 7c, 1-day PoP forecasts from the raw ECMWF ensemble remain unreliable even at the $5^\circ \times 2^\circ$ gridbox scale. It is only under large scales and longer accumulation times simultaneously, when precipitation occurs almost invariably, that raw ensemble PoP forecasts become reliable (Fig. 7d).

Table 4 shows mean Brier and CRPS scores at various spatial aggregations for 1-day precipitation accumulation, verified against TRMM observations. The raw ECMWF ensemble forecast is inferior to EPC at all resolutions and in every single region and season. BMA postprocessed forecasts outperform EPC across aggregation scales, and in every single region and season, but the improvement relative to EPC remains small.

d. TIGGE subensembles and RMM ensemble

In addition to the ECMWF EPS, which we have studied thus far, the TIGGE database contains several more operational subensembles, as listed in Table 1. Figure 8 shows uPIT histograms for the various subensembles and the RMM ensemble for 1-day accumulated precipitation forecasts over West Sahel. All TIGGE subensembles exhibit underdispersion and wet biases, though to strongly varying degrees.

Figure 9 displays Brier and CRPS skill scores relative to EPC for raw and BMA postprocessed TIGGE subensemble and RMM ensemble forecasts during 2007–14, verified against station observations. All raw ensembles underperform relative to EPC, in part drastically so. For most subensembles, a temporal improvement in skill is visible, with the monsoon seasons of 2011–14 revealing higher skill than those during 2007–10. Post-processing by BMA increases forecast quality. The ECMWF, Korea Meteorological Administration (KMA), NCEP, and UKMO ensembles yield the best postprocessed forecasts, exhibiting small positive skill relative to EPC for most monsoon periods. The BMA postprocessed RMM ensemble outperforms all subensembles as well as EPC, but the improvement is small. As shown in Fig. 10, the mean perturbed forecasts from the ECMWF, UKMO, and NCEP ensembles are the top three contributors to the BMA postprocessed RMM forecast.

In further experiments, we have studied raw and postprocessed TIGGE subensemble and RMM ensemble forecasts at accumulation times of up to 5 days and spatial aggregations of up to $5^\circ \times 2^\circ$ grid boxes in TRMM. Our findings generally remain unchanged. The raw ensemble forecasts never reach the quality of the

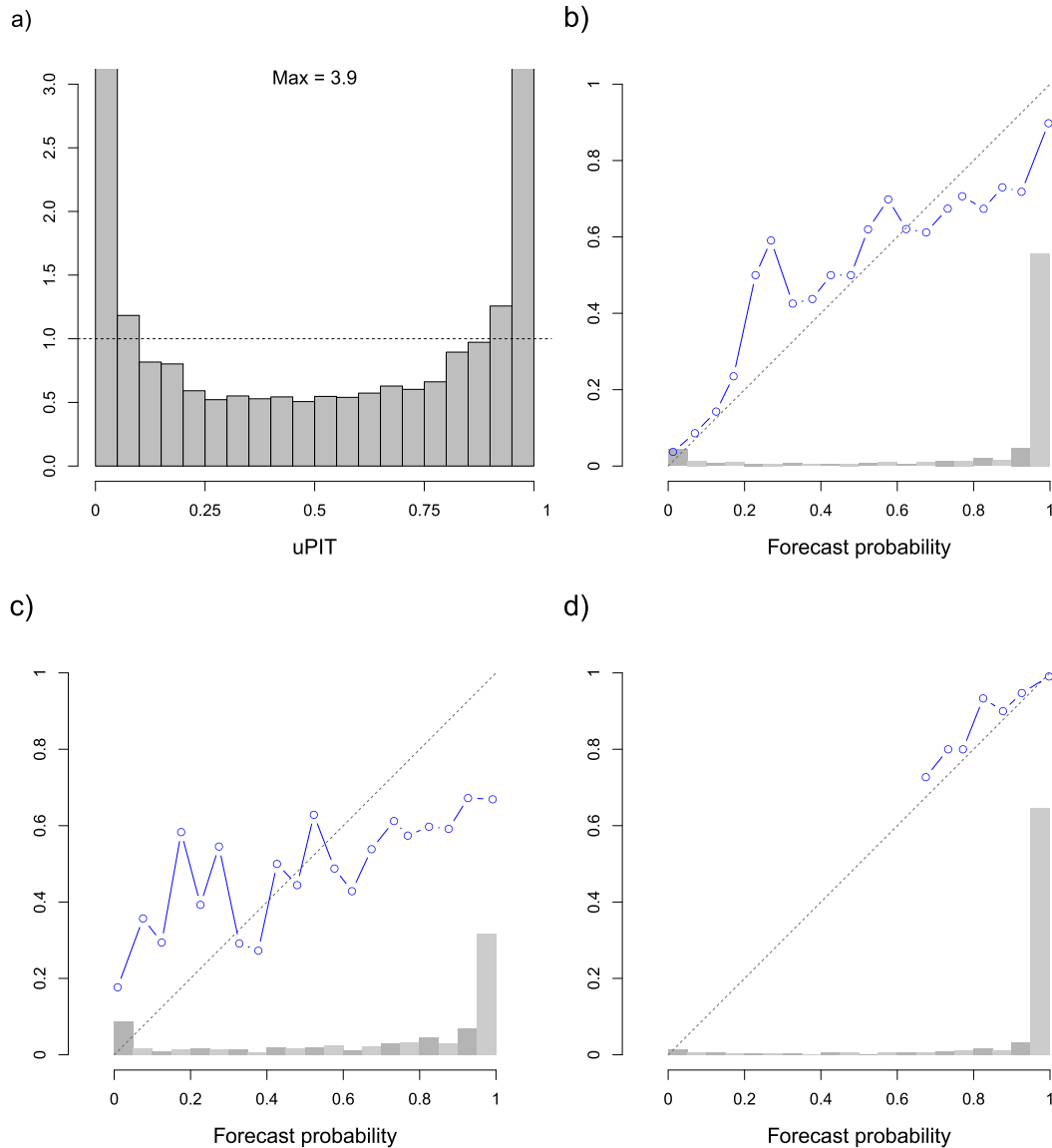


FIG. 7. Calibration and reliability of raw ECMWF ensemble forecasts over West Sahel during the monsoon season of 2014 at 1- and 5-day accumulations. The (a) uPIT histogram and (b) reliability diagram for 5-day accumulated precipitation, verified against station observations. (c),(d) Reliability diagrams for 1- and 5-day accumulated precipitation, verified again $5^{\circ} \times 2^{\circ}$ aggregated TRMM observations. Same setup as in Figs. 4 and 5.

climatological reference EPC. After postprocessing with BMA, the ECMWF ensemble typically becomes the best-performing TIGGE subensemble, showing

slightly better scores than EPC when verified against TRMM observations, at all spatial aggregations. The BMA postprocessed RMM forecast depends heavily on

TABLE 3. Mean BS, mean CRPS, and MAE for raw ECMWF ensemble, EPC, and EMOS and BMA postprocessed forecasts of 5-day accumulated precipitation during the monsoon seasons of 2007–14, verified against station observations. The setup is as in Table 2.

	BS			CRPS			MAE		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	0.14	-0.25	-0.10	-12.80	-8.42	-19.69	16.23	-10.76	-24.41
EPC	0.12	0.16	0.08	11.63	7.07	16.54	16.15	9.56	22.98
EMOS	0.13	0.16	-0.08	11.62	7.34	16.44	+15.99	9.96	22.74
BMA	+0.11	+0.15	0.08	+11.47	+6.94	16.33	+16.07	+9.45	22.92

TABLE 4. Performance of spatially aggregated raw ECMWF ensemble, EPC, and BMA postprocessed forecasts of 1-day accumulated precipitation during the monsoon seasons of 2007–14, verified against TRMM gridbox observations. The setup is as in Table 2.

	TRMM 0.25° × 0.25°/1 day						TRMM 1° × 1°/1 day			TRMM 5° × 2°/1 day		
	BS			CRPS			CRPS			CRPS		
	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast	West Sahel	East Sahel	Guinea Coast
ENS	-0.30	-0.23	-0.48	-2.29	-1.44	-4.03	-2.24	-1.56	-4.43	-1.95	-1.53	-4.22
EPC	0.19	0.14	0.23	1.07	0.57	1.35	0.94	0.58	1.36	0.81	0.49	1.07
BMA	+0.17	+0.13	+0.21	+1.03	+0.55	+1.29	+0.89	+0.55	+1.28	+0.76	+0.45	+0.95

the ECMWF mean perturbed forecast and is superior to both EPC and the BMA postprocessed subensemble.

5. Discussion

In a first-ever thorough verification study, the quality of operational ensemble precipitation forecasts from different NWP centers was assessed over northern tropical Africa for several years, accumulation periods, and for station and spatially aggregated satellite observations. All raw ensembles exhibit calibration problems in the form of underdispersion and biases and are unreliable at high PoP forecast values. They have lower skill than the climatological reference EPC for the prediction of occurrence and amount of precipitation, with the underperformance being stable across monsoon seasons.

After correcting for systematic errors in the raw ensemble through statistical postprocessing, the ensemble forecasts become reliable and calibrated, but only a few are slightly superior to EPC. While ramifications and developments of both EMOS and BMA might be feasible (see, e.g., Fortin et al. 2006; Scheuerer and Hamill 2015), and training sets could be augmented by using reforecast data (e.g., Di Giuseppe et al. 2013), the respective benefits are likely to be incremental at this time, though as the raw ensemble performance improves, they might become substantial. Not surprisingly, forecast skill tends to be highest for long accumulation times and large spatial aggregations. Overall, raw ensemble forecasts are of no use for the prediction of precipitation over northern tropical Africa, and even EMOS and BMA postprocessed forecasts have little added value compared to EPC.

What are the reasons for this rather disappointing level of performance for the state-of-the-art global EPSs? For 1-day accumulated precipitation forecasts, the ability of an NWP model to resolve the details of convective organization is essential. As all global EPSs use parameterized convection, this clearly limits the forecast skill. The fact that even postprocessed 1-day accumulated ensemble forecasts exhibit no skill relative

to EPC implies that ensembles cannot translate information on the current atmospheric state (e.g., tropical waves or influences from the extratropics) into meaningful impacts regarding the occurrence or amount of precipitation. This is robust for verification against station as well as satellite observations and cannot, therefore, be explained by propagation errors.

For longer accumulation times and larger spatial aggregations, the large-scale circulation has a much stronger impact on convective activity, which should weaken the limitations through convective parameterization. The skill of 5-day accumulated precipitation forecasts, however, increases only slightly, if at all, compared to 1-day accumulated forecasts. The most likely reason for this is that squall lines have feedbacks on the large-scale circulation, which are not realistically represented in global NWP models either. Marsham et al. (2013) find that the large-scale monsoon state in (more realistic) simulations with explicit convection differs quite markedly from runs with parameterized convection, even when using the same resolution of 12 km. In the explicit-convection simulation, greater latent and radiative heating to the north weakens the monsoon flow, delays the diurnal cycle, and convective cold pools provide an essential component to the monsoon flux. We suspect that some or all of these effects are misrepresented in global EPS forecasts.

The fact that EPS precipitation forecasts are so poor over northern tropical Africa is a strong demonstration of the complexity of the underlying forecast problem. An interesting question within this context is whether poor predictability in the tropics is unique to northern Africa with its strongly organized, weakly synoptically forced rainfall systems.

Furthermore, the lack of skill motivates complementary approaches to predicting precipitation over this region. Little et al. (2009) compare operational NCEP ensemble, climatological, and statistical forecasts for stations in the Thames Valley, United Kingdom. They note that NCEP forecasts outperform climatological forecasts, but demonstrate that statistical forecasts,

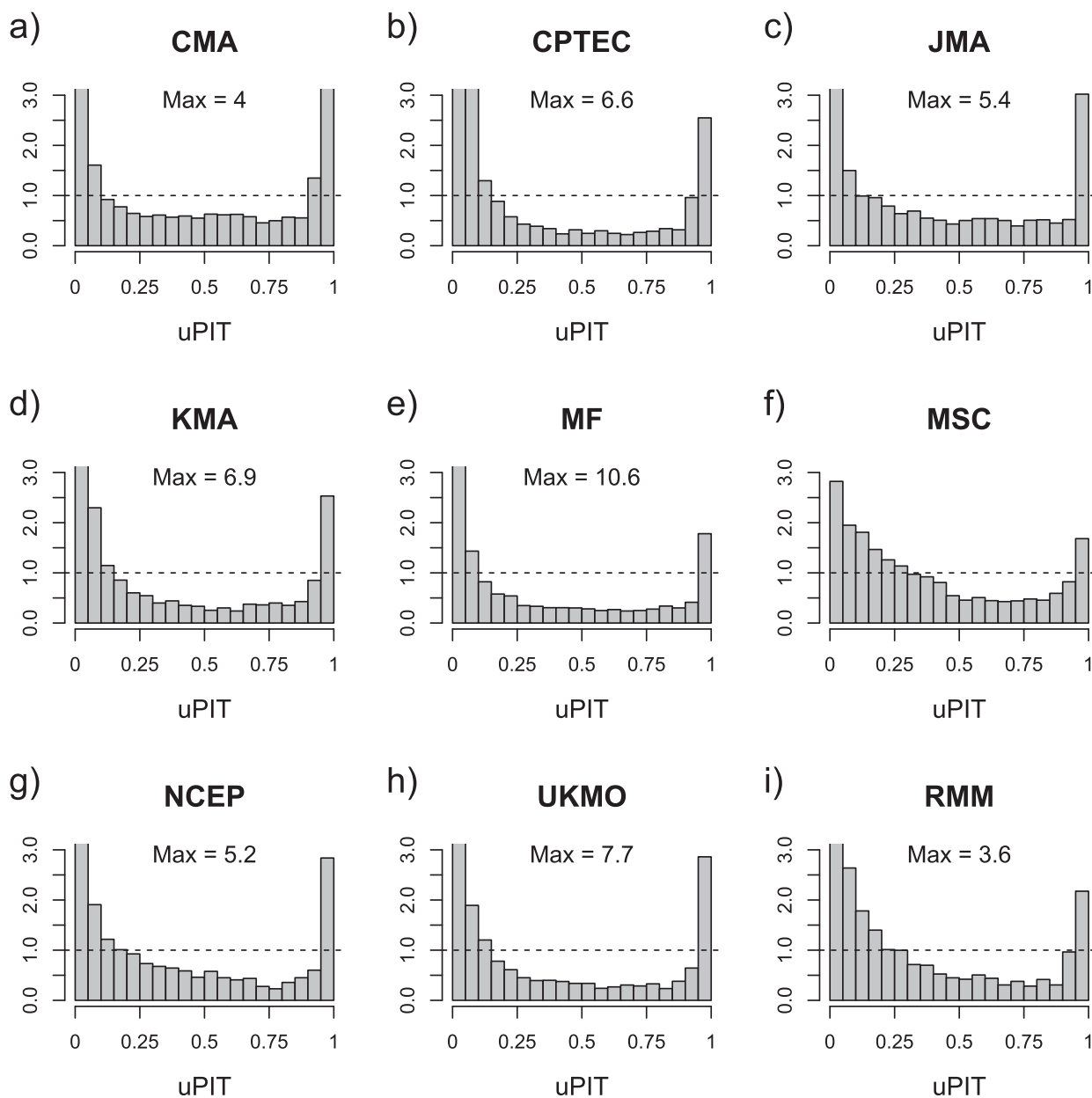


FIG. 8. The uPIT histograms for raw TIGGE subensemble and raw RMM ensemble forecasts of 1-day accumulated precipitation over West Sahel during the monsoon season of 2013, verified against station observations. Same setup as in Fig. 4.

solely based on past observations, can outperform NCEP forecasts by exploiting spatiotemporal dependencies. These also exist over northern tropical Africa and some additional predictability may stem from large-scale drivers such as convectively coupled waves. Fink and Reiner (2003) note a coupling of the initiation of squall lines to African easterly waves and Wheeler and Kiladis (1999) the influence of large-scale tropical waves, such as Kelvin and equatorial Rossby waves or the Madden–Julian oscillation, on convective activity.

Pohl et al. (2009) confirm the relation between the Madden–Julian oscillation and rainfall over West Africa, and Vizy and Cook (2014) demonstrate an impact of potential extratropical wave trains on Sahelian rainfall. Statistical models based on spatiotemporal characteristics of rainfall and extended by such large-scale predictors seem a promising approach for improving precipitation forecasts over our study region, and we expect such forecasts to outperform climatology. This approach will be explored in future work.

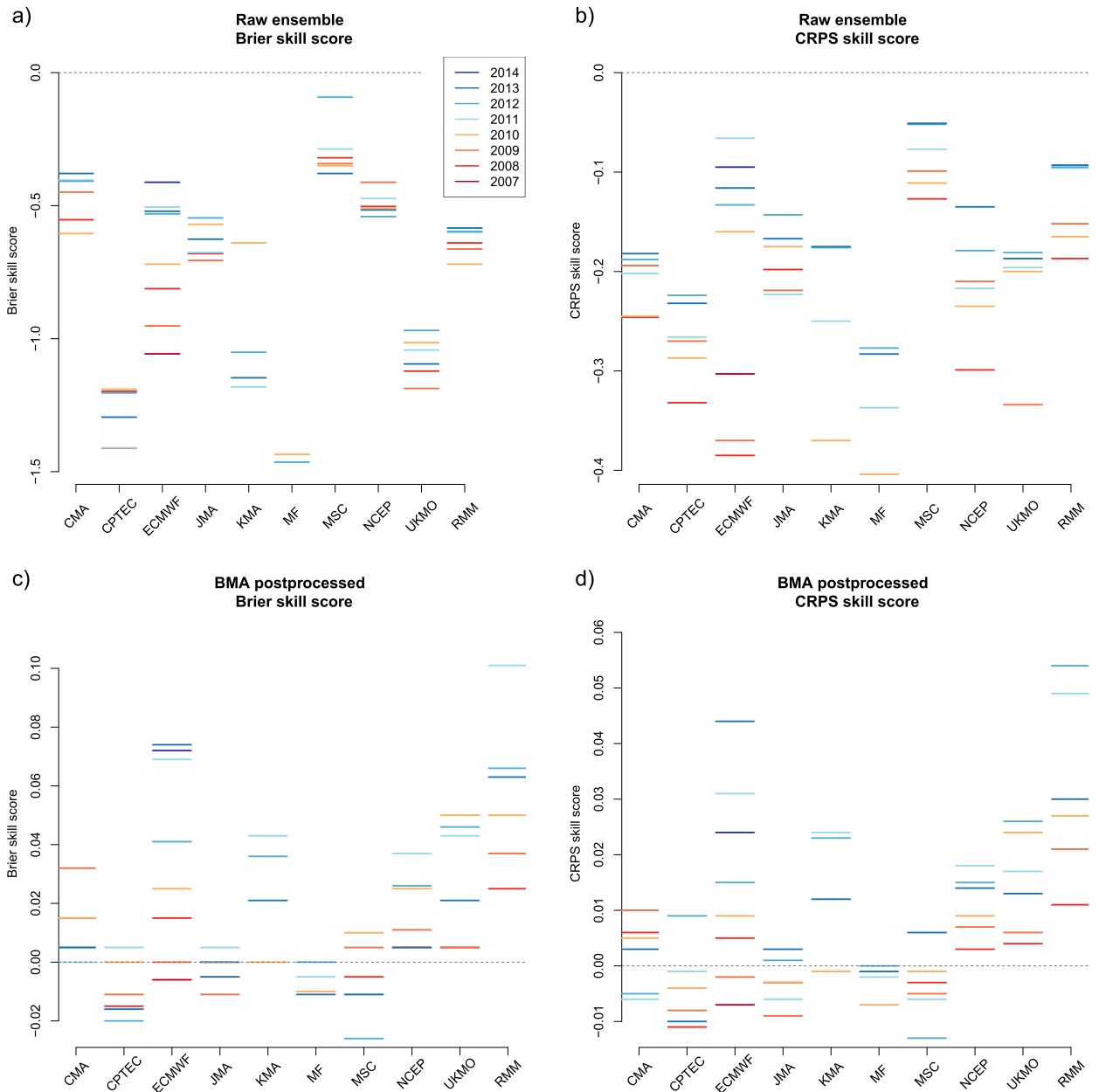


FIG. 9. (a),(c) Brier and (b),(d) CRPS skill scores for raw and BMA postprocessed TIGGE subensemble forecasts of 1-day accumulated precipitation over West Sahel during the monsoon seasons of 2007–14, verified against station observations. Skill equal to EPC is indicated by the dashed line.

As discussed in section 4a, we suspect convective parameterization to be a major cause of the low quality of the model-based forecasts here. Therefore, it would be interesting to test ensembles of convection-permitting NWP model runs, ideally in combination with ensemble data assimilation, but the computational costs are high, and it will take time until a multiyear database will become available for validation studies. Alternatively, it

could be tested whether systematic improvements to convection schemes (e.g., Bechtold et al. 2014) do in fact positively impact ensemble forecast quality. Given the growing socioeconomic impact of rainfall in northern tropical Africa with its rain-fed agriculture, statistical and dynamical approaches should be fostered in parallel in order to improve the predictability of rainfall in this region.

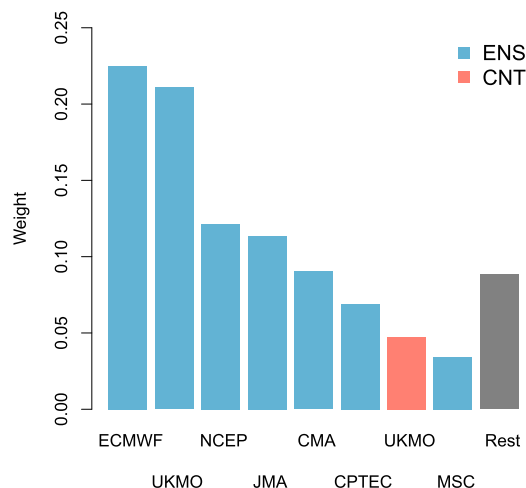


FIG. 10. BMA weights of RMM components for 1-day accumulated precipitation forecasts over West Sahel trained against station observations, averaged over the monsoon seasons of 2008–13. ENS forecasts and CNTs are distinguished by the color of the respective bar.

Acknowledgments. The research leading to these results has been accomplished within project C2 “Prediction of wet and dry periods of the West African Monsoon” of the Transregional Collaborative Research Center SFB/TRR 165 “Waves to Weather” funded by the German Science Foundation (DFG). TG is grateful for support by the Klaus Tschira Foundation. The authors also thank various colleagues and weather services that have over the years contributed to the enrichment of the KASS-D database; special thanks go to Robert Redl for creating the underlying software. We thank Sebastian Lerch for helpful discussions and Alexander Jordan and Michael Scheuerer for providing R code, and we are grateful to Tom Hamill, Ken Mylne, and an anonymous referee for constructive suggestions.

APPENDIX

Quality Control for Precipitation Observations within KASS-D

Rainfall exhibits extremely high spatial and temporal variability, which hinders automated quality checks applicable to other meteorological variables such as temperature or pressure. For precipitation, Fiebrich and Crawford (2001) note only a range and a step test. The global range of station-observed 1-day accumulated precipitation is from 0 to 1825 mm. All KASS-D observations passed this test. The step test checks if the difference of neighboring 5-min accumulated precipitation is smaller than 25 mm. For 1-day accumulated precipitation, tests of this type are not meaningful, nor are the

persistence tests used by Pinson and Hagedorn (2012) for wind speed.

However, the site-specific climatological distributions of precipitation accumulation should be right skewed (i.e., the median should be smaller than the mean), and in the tropics they should have a point mass at zero (Rodwell et al. 2010). As noted, we only consider stations with more than 80% available observations in any of the monsoon seasons, and all 132 stations thus selected passed these tests.

REFERENCES

- Bechtold, P., N. Semane, P. Lopez, J.-P. Chaboureaud, A. Beljaars, and N. Bormann, 2014: Representing equilibrium and non-equilibrium convection in large-scale models. *J. Atmos. Sci.*, **71**, 734–753, <https://doi.org/10.1175/JAS-D-13-0163.1>.
- Birch, C. E., D. J. Parker, J. H. Marsham, D. Copsey, and L. Garcia-Carreras, 2014: A seamless assessment of the role of convection in the water cycle of the West African monsoon. *J. Geophys. Res. Atmos.*, **119**, 2890–2912, <https://doi.org/10.1002/2013JD020887>.
- 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>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, <https://doi.org/10.1175/MWR2905.1>.
- Davis, J., P. Knippertz, and A. H. Fink, 2013: The predictability of precipitation episodes during the West African dry season. *Quart. J. Roy. Meteor. Soc.*, **139**, 1047–1058, <https://doi.org/10.1002/qj.2014>.
- Di Giuseppe, F., F. Molteni, and A. M. Tompkins, 2013: A rainfall calibration methodology for impacts modelling based on spatial mapping. *Quart. J. Roy. Meteor. Soc.*, **139**, 1389–1401, <https://doi.org/10.1002/qj.2019>.
- Ehm, W., T. Gneiting, A. Jordan, and F. Krüger, 2016: Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. Roy. Stat. Soc.*, **78B**, 505–562, <https://doi.org/10.1111/rssb.12154>.
- Engel, T., A. H. Fink, P. Knippertz, G. Pante, and J. Bliedernicht, 2017: Extreme precipitation in the West African cities of Dakar and Ouagadougou: Atmospheric dynamics and implications for flood risk assessments. *J. Hydrometeorol.*, **18**, 2937–2957, <https://doi.org/10.1175/JHM-D-16-0218.1>.
- Fiebrich, C. A., and K. C. Crawford, 2001: The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bull. Amer. Meteor. Soc.*, **82**, 2173–2187, [https://doi.org/10.1175/1520-0477\(2001\)082<2173:TIOUMP>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2173:TIOUMP>2.3.CO;2).
- Fink, A. H., and A. Reiner, 2003: Spatiotemporal variability of the relation between African easterly waves and West African squall lines in 1998 and 1999. *J. Geophys. Res.*, **108**, 4332, <https://doi.org/10.1029/2002JD002816>.
- , D. G. Vincent, and V. Ermert, 2006: Rainfall types in the West African Sudanian zone during the summer monsoon 2002. *Mon. Wea. Rev.*, **134**, 2143–2164, <https://doi.org/10.1175/MWR3182.1>.
- , and Coauthors, 2011: Operational meteorology in West Africa: Observational networks, weather analysis and forecasting. *Atmos. Sci. Lett.*, **12**, 135–141, <https://doi.org/10.1002/asl.324>.

- , and Coauthors, 2017: Mean climate and seasonal cycle. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., Wiley-Blackwell, 1–39.
- Fortin, V., A.-C. Favre, and M. Saïd, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.*, **132**, 1349–1369, <https://doi.org/10.1256/qj.05.167>.
- Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, <https://doi.org/10.1175/2009MWR3046.1>.
- , —, —, J. M. Sloughter, and V. Berrocal, 2011: Probabilistic weather forecasting in R. *R J.*, **3** (1), 55–63.
- Gneiting, T., 2011: Making and evaluating point forecasts. *J. Amer. Stat. Assoc.*, **106**, 746–762, <https://doi.org/10.1198/jasa.2011.r10138>.
- , 2014: Calibration of medium-range weather forecasts. ECMWF Tech. Memo. 719, 28 pp., <https://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf>.
- , and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249, <https://doi.org/10.1126/science.1115255>.
- , and —, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , —, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, <https://doi.org/10.1002/qj.1895>.
- 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>.
- Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2).
- , and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, <https://doi.org/10.1256/qj.06.25>.
- , J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, [https://doi.org/10.1175/1520-0493\(2004\)132<1434:ERIMFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2).
- Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, <https://doi.org/10.1002/2014GL062472>.
- Houze, R. A., Jr., K. L. Rasmussen, M. D. Zuluaga, and S. R. Brodzik, 2015: The variable nature of convection in the tropics and subtropics: A legacy of 16 years of the Tropical Rainfall Measuring Mission satellite. *Rev. Geophys.*, **53**, 994–1021, <https://doi.org/10.1002/2015RG000488>.
- Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.*, **8**, 38–55, <https://doi.org/10.1175/JHM560.1>.
- Lafore, J. P., and Coauthors, 2017: Deep convection. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., Wiley-Blackwell, 90–129.
- Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, <https://doi.org/10.1016/j.jcp.2007.02.014>.
- Little, M. A., P. E. McSharry, and J. W. Taylor, 2009: Generalized linear models for site-specific density forecasting of U.K. daily rainfall. *Mon. Wea. Rev.*, **137**, 1029–1045, <https://doi.org/10.1175/2008MWR2614.1>.
- Maggioni, V., P. C. Meyers, and M. D. Robinson, 2016: A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era. *J. Hydrometeor.*, **17**, 1101–1117, <https://doi.org/10.1175/JHM-D-15-0190.1>.
- Marsham, J. H., N. S. Dixon, L. Garcia-Carreras, G. M. S. Lister, D. J. Parker, P. Knippertz, and C. E. Birch, 2013: The role of moist convection in the West African monsoon system—Insights from continental-scale convection-permitting simulations. *Geophys. Res. Lett.*, **40**, 1843–1849, <https://doi.org/10.1002/grl.50347>.
- Mathon, V., H. Laurent, and T. Lebel, 2002: Mesoscale convective system rainfall in the Sahel. *J. Appl. Meteor.*, **41**, 1081–1092, [https://doi.org/10.1175/1520-0450\(2002\)041<1081:MCSRIT>2.0.CO;2](https://doi.org/10.1175/1520-0450(2002)041<1081:MCSRIT>2.0.CO;2).
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroligias, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, <https://doi.org/10.1002/qj.49712252905>.
- Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774, <https://doi.org/10.1256/0035900021643593>.
- Pantillon, F., P. Knippertz, J. H. Marsham, and C. E. Birch, 2015: A parameterization of convective dust storms for models with mass-flux convection schemes. *J. Atmos. Sci.*, **72**, 2545–2561, <https://doi.org/10.1175/JAS-D-14-0341.1>.
- Park, Y. Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, **134**, 2029–2050, <https://doi.org/10.1002/qj.334>.
- Parker, D. J., and Coauthors, 2008: The AMMA radiosonde program and its implications for the future of atmospheric monitoring over Africa. *Bull. Amer. Meteor. Soc.*, **89**, 1015–1027, <https://doi.org/10.1175/2008BAMS2436.1>.
- Pearson, K. J., G. M. S. Lister, C. E. Birch, R. P. Allan, R. J. Hogan, and S. J. Woolnough, 2014: Modelling the diurnal cycle of tropical convection across the ‘grey zone.’ *Quart. J. Roy. Meteor. Soc.*, **140**, 491–499, <https://doi.org/10.1002/qj.2145>.
- Pinson, P., and R. Hagedorn, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteor. Appl.*, **19**, 484–500, <https://doi.org/10.1002/met.283>.
- Pohl, B., S. Janicot, B. Fontaine, and R. Marteau, 2009: Implication of the Madden-Julian oscillation in the 40-day variability of the West African monsoon. *J. Climate*, **22**, 3769–3785, <https://doi.org/10.1175/2009JCLI2805.1>.
- 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>.
- R Development Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, <http://www.R-project.org>.
- Roca, R., P. Chambon, I. Jobard, P. E. Kirstetter, M. Gosset, and J. C. Bergés, 2010: Comparing satellite and surface rainfall products over West Africa at meteorologically relevant scales during the AMMA campaign using error estimates. *J. Appl. Meteor. Climatol.*, **49**, 715–731, <https://doi.org/10.1175/2009JAMC2318.1>.
- Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **136**, 1344–1363, <https://doi.org/10.1002/qj.656>.
- 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>.
- , and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, <https://doi.org/10.1175/MWR3441.1>.
- Söhne, N., J.-P. Chaboureau, and F. Guichard, 2008: Verification of cloud cover forecast with satellite observation over West Africa. *Mon. Wea. Rev.*, **136**, 4421–4434, <https://doi.org/10.1175/2008MWR2432.1>.
- Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, <https://doi.org/10.1175/BAMS-D-13-00191.1>.
- Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, <https://doi.org/10.1111/j.1467-985X.2009.00616.x>.
- Van der Linden, R., A. H. Fink, J. G. Pinto, and T. Phan-Van, 2017: The dynamics of an extreme precipitation event in northeastern Vietnam in 2015 and its predictability in the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **32**, 1041–1056, <https://doi.org/10.1175/WAF-D-16-0142.1>.
- Vizy, E. K., and K. H. Cook, 2014: Impact of cold air surges on rainfall variability in the Sahel and wet African tropics: A multi-scale analysis. *Climate Dyn.*, **43**, 1057–1081, <https://doi.org/10.1007/s00382-013-1953-z>.
- Wheeler, M., and G. N. Kiladis, 1999: Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *J. Atmos. Sci.*, **56**, 374–399, [https://doi.org/10.1175/1520-0469\(1999\)056<0374:CCEWAO>2.0.CO;2](https://doi.org/10.1175/1520-0469(1999)056<0374:CCEWAO>2.0.CO;2).
- Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- , 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.