



An IMERG-Based Optimal Extended Probabilistic Climatology (EPC) as a Benchmark Ensemble Forecast for Precipitation in the Tropics and Subtropics

EVA-MARIA WALZ,^{a,b} MARLON MARANAN,^c RODERICK VAN DER LINDEN,^c ANDREAS H. FINK,^c AND PETER KNIPPERTZ^c

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

^b *Computational Statistics, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*

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

(Manuscript received 15 December 2020, in final form 28 April 2021)

ABSTRACT: Current numerical weather prediction models show limited skill in predicting low-latitude precipitation. To aid future improvements, be it with better dynamical or statistical models, we propose a well-defined benchmark forecast. We use the arguably best available high-resolution, gauge-calibrated, gridded precipitation product, the Integrated Multisatellite Retrievals for GPM (IMERG) “final run” in a ± 15 -day window around the date of interest to build an empirical climatological ensemble forecast. This window size is an optimal compromise between statistical robustness and flexibility to represent seasonal changes. We refer to this benchmark as extended probabilistic climatology (EPC) and compute it on a $0.1^\circ \times 0.1^\circ$ grid for 40°S – 40°N and the period 2001–19. To reduce and standardize information, a mixed Bernoulli–Gamma distribution is fitted to the empirical EPC, which hardly affects predictive performance. The EPC is then compared to 1-day ensemble predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF) using standard verification scores. With respect to rainfall amount, ECMWF performs only slightly better than EPS over most of the low latitudes and worse over high-mountain and dry oceanic areas as well as over tropical Africa, where the lack of skill is also evident in independent station data. For rainfall occurrence, EPC is superior over most oceanic, coastal, and mountain regions, although the better potential predictive ability of ECMWF indicates that this is mostly due to calibration problems. To encourage the use of the new benchmark, we provide the data, scripts, and an interactive web tool to the scientific community.

SIGNIFICANCE STATEMENT: Precise precipitation forecasts in the tropics and subtropics are relevant for a large and growing population. To gauge the success of improvements, an adequate baseline is needed. Here we use satellite-based rainfall estimates from 2001 to 2019 to define a climatological reference forecast that we call extended probabilistic climatology (EPC), as it combines rainfall observations from a window of ± 15 days around the date of interest. We show that this simple approach outperforms current weather forecast models in some areas and forecast aspects but is inferior in others. To foster the use of this new benchmark in the scientific and forecasting communities, we provide the EPC data, scripts, and an interactive web tool to display EPC forecasts for selected locations.


KEYWORDS: Africa; Subtropics; Tropics; Precipitation; Satellite observations; Ensembles; Forecasting

1. Introduction

Over the last more than 60 years, scientific and technical advances have tremendously improved numerical weather prediction (NWP) worldwide (Bauer et al. 2015; Alley et al. 2019). The quasi-exponential growth in computing power enabled the implementation of ensemble prediction systems (EPSs) in the 1990s, where each member is started from slightly different initial conditions to allow quantifying forecast

uncertainty (Molteni et al. 1996). EPSs are well in line with recent developments in many research areas in that they foster the transition from deterministic to probabilistic forecasts (Gneiting and Katzfuss 2014).

Despite the overall triumph of NWP, quantitative precipitation forecasts in the tropics remain a great challenge. For example, Haiden et al. (2012) showed that in 2010/11 a deterministic forecast of tropical rainfall with a 1-day lead time was as skillful as a forecast in the extratropics for 6-day lead time. More recently, Vogel et al. (2020) compared 1–5-day ensemble predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Meteorological Service of Canada (MSC) over the tropical belt from 30°S and 30°N with Tropical Rainfall Measuring Mission (TRMM) 3B42 precipitation estimates and found that both models predict

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

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

DOI: 10.1175/WAF-D-20-0233.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

rainfall better than the reference only over about 50% (occurrence) and 60% (amount) of all land points. Forecast performance is best over arid Australia and worst over oceanic deserts, mountains, and large parts of tropical Africa. Specifically for the summer monsoon season in northern tropical Africa, Vogel et al. (2018) analyzed nine global EPSs and found that all individual models and the multimodel ensemble are uncalibrated and unreliable, and have no skill in predicting the occurrence and amount of precipitation when compared to a climatological forecast. This disappointing result is robust for different subregions, accumulation periods, grid spacings, and verification datasets. A possible reason for this is the exceptionally high degree of convective organization observed over tropical Africa (Nesbitt et al. 2006; Roca et al. 2014), a process that is difficult to capture with the convective parameterizations used in global NWP models (Vogel et al. 2018). Kniffka et al. (2020) confirm the overall low skill in predicting local rainfall in Africa but show a positive effect of propagating synoptic-scale disturbances on forecasts of regional precipitation. The overall poor performance of current operational systems with respect to tropical rainfall calls for alternative approaches reaching from convection-permitting resolution (Pante and Knippertz 2019) to methods from statistics and machine learning (Shi et al. 2015; Rasp et al. 2020; Vogel et al. 2021).

Before developing and evaluating new models and approaches, it is essential to establish benchmark forecasts in order to systematically assess forecast improvement. Rasp et al. (2020) recently proposed WeatherBench as a standard for data-driven weather forecasting, but the information provided is deterministic only and precipitation data are taken from reanalysis, which is not well suited for the question at hand, since it is still fundamentally based on numerical modeling. In the context of evaluating ensemble forecasts, a commonly used reference is the probabilistic climatology, which is based on all available past observations for a particular day of the year (Jaun and Ahrens 2009; Pagano et al. 2013; Pinson and Hagedorn 2012; Pappenberger et al. 2015). Closely related to this concept, Vogel et al. (2018) and Vogel et al. (2020) introduced the idea of an extended probabilistic climatology (EPC), which is based on the consideration that climatologies of neighboring days are very similar and can together better capture the rainfall distribution for a given (sub)season. This approach is particularly well suited for predicting precipitation in the tropics, as many tropical regions show large seasonal shifts with strongly varying numbers of dry and wet days. An open question in this approach is the optimal number of days around the considered date in the computation of the EPC. This number should be large enough to give robust seasonal statistics, particularly in more arid parts of the tropics, but small enough to capture the sometimes rather sudden onsets of rainy seasons, particularly in monsoon regions. For northern tropical Africa for example, Vogel et al. (2018) used a window of ± 2 days, while Vogel et al. (2020) varied this number across the tropical belt to account for local differences.

The aim of the present paper is to establish a widely usable and optimized probabilistic benchmark for the specific task of

predicting low-latitude (here 40°S–40°N) rainfall. To achieve this, we use the arguably best currently available, high-resolution, gauge-calibrated, gridded precipitation product Integrated Multisatellite Retrievals for GPM (IMERG) “final run” (referred to as IMERG-F hereafter) (Huffman et al. 2015). We employ the EPC approach proposed by Vogel et al. (2018) and estimate an optimal window length for the computation. A mixed Bernoulli–gamma (MBG) distribution is then fitted to the resulting distributions at each grid point to reduce and standardize the amount of information contained in the benchmark, which is made freely available to the scientific community.

The paper is structured as follows. All relevant datasets and statistical tools used in this study are described in sections 2 and 3, respectively. Section 4 provides an analysis of the optimal number of neighboring days in the construction of the EPC, comparisons of forecast performance to the operational ECMWF EPS and to rain gauge observations over tropical Africa, one of the most problematic forecast regions worldwide (Vogel et al. 2020), as well as the fit of the empirical EPC to the full MBG probability distribution and its impact on forecast performance. Conclusions are given in section 5.

2. Data

a. IMERG-F rainfall estimates

The computation of the EPC is performed for the low-latitude belt from 40°N to 40°S for the years 2001–19 using daily IMERG-F rainfall estimates (Hou et al. 2014; Huffman et al. 2015, 2019a). The data are provided on a $0.1^\circ \times 0.1^\circ$ grid and can be downloaded from (<https://gpm.nasa.gov/data/directory>). The dataset contains blended precipitation estimations based on passive microwave (PMW) and infrared (IR) retrievals at a native temporal resolution of 30 min. All PMW estimates from the TRMM/GPM international constellation of satellites are calibrated toward rainfall estimates of the GPM/TRMM Combined Radar-Radiometer (CORRA) product. Initially available since the start of GPM in mid-2014, the current version V06B (as of 6 August 2020) provides precipitation data dating back to June 2000 (Huffman et al. 2019b), thus covering a major part of the preceding TRMM era (Kummerow et al. 1998; Huffman et al. 2007). Before 2014, radar information was available for 40°N–40°S only, motivating the restriction to this belt. Unlike in the near-real-time runs “early” and “late,” the “final run” data considered are calibrated with rain gauge measurements provided by the Global Precipitation Climatology Centre (GPCC; Schneider et al. 2016) on a monthly basis. Rainfall estimates for shorter time scales are rescaled such that they match the monthly sum. The degree to which the original estimates are adjusted by the gauge calibration process within a given region is generally determined by the number of available rain gauges, which is highly variable across the tropical continents. The analyses in sections 4a, 4b, and 4d are based on daily data from 0000 to 0000 UTC, while section 4c considers data from 0600 to 0600 UTC to better match with reporting practices for weather stations, although this will likely have a rather negligible impact on the

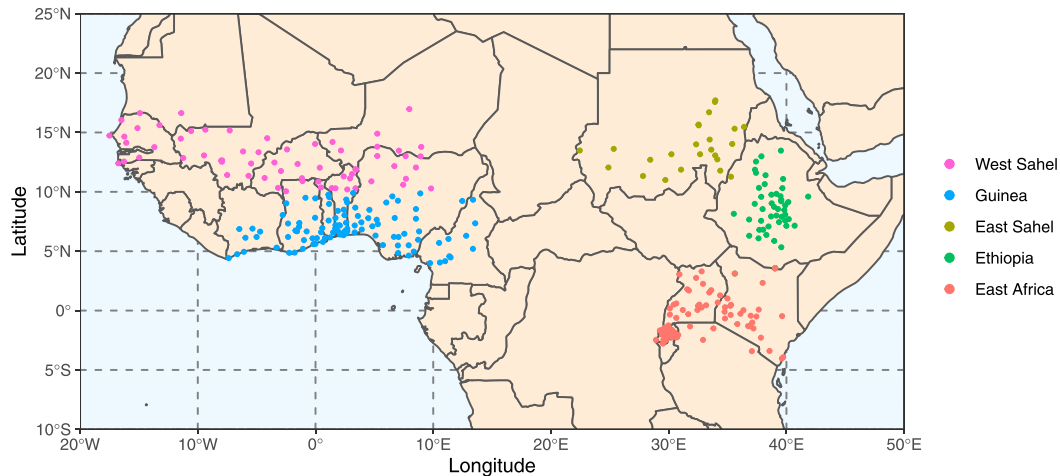


FIG. 1. Geographical overview of the study domain with all considered station locations (dots) colored according to the five regions given in the legend.

climatologies. While sections 4a and 4d use the full resolution of $0.1^\circ \times 0.1^\circ$, IMERG-F data are regridded to $0.25^\circ \times 0.25^\circ$ by applying first-order conservative remapping for the comparison with the lower-resolution ECMWF ensemble forecasts in sections 4b and 4c.

b. ECMWF ensemble forecasts

Based on the years 2012–19, EPC is compared to forecasts from the operational ECMWF EPS, which consists of 50 perturbed members and a single control run at a grid resolution of $0.25^\circ \times 0.25^\circ$. For the sake of simplicity, we only consider 24-h accumulations initialized at 0000 UTC but the results by Vogel et al. (2020) suggest that many forecast errors in ECMWF are systematic and depend only little on lead time. A verification with station observations for the years 2012–16 in section 4c requires an accumulation from 0600 to 0600 UTC, which is obtained by subtracting the accumulation at lead time +6 h from that at lead time +30 h. The data are downloaded from ECMWF’s Meteorological Archival and Retrieval System (MARS) (<https://www.ecmwf.int/en/forecasts>).

c. Rain gauge observations

To provide an unbiased performance comparison for the IMERG-F-based EPC and ECMWF ensemble forecasts, rain gauge measurements from standard weather stations are used. We concentrate here on the largest tropical landmass Africa, where model forecasts have been shown to be particularly challenging (Vogel et al. 2020). Unfortunately, the network of meteorological ground stations in tropical Africa is relatively sparse and station records often show gaps in time. Here we use data from the Karlsruhe African Surface Station Database (KASS-D), which brings together precipitation observations from a wide variety of networks and sources, including not freely available data from research projects and national weather services. KASS-D data were also used in Vogel et al. (2018). KASS-D provides 24-h accumulated precipitation measured between 0600 and 0600 UTC on the following day.

We concentrate on 2012–16, which is characterized by a relatively good spatial coverage and degree of completeness of station records when compared to other recent time periods available in KASS-D. Only stations with more than 80% of available observations in every year of 2012–16 were selected. Based on geographic position and rainfall climate (see Nicholson et al. 2018), the stations are assigned to one of the five regions West Sahel, Guinea Coast, East Sahel, Ethiopia, and East Africa as illustrated in Fig. 1.

3. Forecast evaluation and statistical tests

To evaluate probabilistic forecasts, proper scoring rules provide an appropriate choice of evaluation metrics. To assess forecast performance for the prediction of occurrence and amount of precipitation popular choices of proper scoring rules are the Brier score (BS) (Brier 1950) and the continuous rank probability score (CRPS) (Matheson and Winkler 1976; Gneiting and Raftery 2007). Since BS and CRPS are negatively oriented, smaller values indicate superior performance. To quantify discrimination ability or potential predictive ability of probabilistic forecasts for precipitation occurrence the area under the receiver operating characteristic curve (AUC) (Fawcett 2006) is used. The AUC measure attains values between 0 and 1 with the interpretation that larger values are better.

To assess statistical significant differences of forecast performance, a Diebold Mariano (DM) test is applied. This procedure tests the hypothesis that two methods have equal predictive performance in the sense that the expectation of the score difference vanishes (Diebold and Mariano 1995; Gneiting and Katzfuss 2014). Let F and G be two competing forecasts and $S_n^F = (1/n)\sum_{i=1}^n S(F_i, y_i)$ and $S_n^G = (1/n)\sum_{i=1}^n S(G_i, y_i)$ are the corresponding mean scores. The DM test is based on the following statistic:

$$t_n = \sqrt{n} \frac{S_n^F - S_n^G}{\hat{\sigma}_n}, \quad (1)$$

where $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^n [S(F_i, y_i) - S(G_i, y_i)]^2$ is an estimate of the variance of the score differential. Under standard regularity conditions, t_n is asymptotically standard normal under the null hypothesis of equal predictive performance of F and G . If the null hypothesis is rejected, negative values of t_n indicate that F is superior whereas positive values of t_n indicate that G is superior. In section 4, DM tests are applied at each grid point by considering the CRPS values of competing forecasts over the time period 2018/19. As pointed out by Wilks (2016), summarizing and interpreting local test results requires an adjustment for the effect of test multiplicity on the overall result. Therefore, a Benjamini and Hochberg (1995) procedure to control the false discovery rate at level α is applied. Let m be the number of considered grid points and p_1, \dots, p_m be the p values of the corresponding DM test at each grid point. To apply a Bonferroni-type multiple-testing procedure, let $p_{(1)} \leq \dots \leq p_{(m)}$ be the ordered p values and k be the largest i for which $p_{(i)} \leq (i/m)\alpha$, then reject all null hypothesis corresponding to the p values $p_{(1)}, \dots, p_{(k)}$.

To test forecast equivalence, i.e., whether two forecasts show no significant difference in their performance, we recommend using a specifically tailored equivalence test instead of drawing conclusions based on nonsignificant DM test results. Equivalence tests have been mostly used in pharmaceutical (e.g., Anderson and Hauck 1983) and economic research (e.g., Johnston and Duke 2008). The most prominent example is the “two one-sided t test” (TOST) (Schuirmann 1987). Following the approach of the TOST procedure the two one-sided DM (TODM) test is derived. Therefore, a predefined upper θ_1 and lower θ_2 equivalence bound is required and two composite null hypotheses, H_0^1 and H_0^2 of one-sided DM tests are considered:

$$\begin{aligned} H_0^1 : t_n &\leq \theta_1, \\ H_0^2 : t_n &\geq \theta_2. \end{aligned} \quad (2)$$

Rejection of both one-sided DM tests implies that the observed difference falls within the predefined equivalence margin $[\theta_1, \theta_2]$ and is close enough to zero to be practically equivalent. As shown by Berger and Hsu (1996) rejection of the two individual tests on level α constitutes a rejection at level α of joint null hypothesis. In section 4, we apply TODM equivalence test at each grid point by considering CRPS values of two forecasting methods over the time period 2018/19. A critical requirement in this procedure is to predefine an interval $[\theta_1, \theta_2]$ that represents the range of score differences that indicate equal forecast performance. We choose the equivalence margin to be symmetric around zero, i.e., the interval $[-\theta, \theta]$. The actual value of θ should be defined based on expertise but this choice is not straightforward in the given use case and test results strongly depend on it. Therefore, here we perform the TODM test over a range of θ values and plot the percentage of significant grid points against θ . Like for the classical DM test discussed above, an adjustment for the effects of test multiplicity is required. To achieve this, the Benjamini and Hochberg procedure is applied to both one-sided DM subtests of TODM.

TABLE 1. Performance of IMERG-F-based EPC forecasts with different window lengths for the 40°S–40°N belt and the period 2001–19. IMERG-F data accumulated from 0000 to 0000 UTC with $0.1^\circ \times 0.1^\circ$ grid resolution are used. The time and space averaged CRPS is shown in the top row. Percentage of grid points where EPC x (with x being window length) is statistically superior to EPC0 based on a two-sided DM test with Benjamini–Hochberg correction and significance level $\alpha = 0.05$ is shown in the bottom row. EPC0 is never viewed as superior.

Window (days)	0	2	5	10	15	20
CRPS (mm day ⁻¹)	2.78	2.68	2.66	2.65	2.65	2.64
Percentage (%)	—	96.18	96.54	96.44	96.28	96.05

4. EPC optimization and validation

In this section, we introduce the IMERG-F-based EPC forecast benchmark and compare it to other rainfall information. A first important step is the identification of an optimal window length to construct the EPC (section 4a). Probabilistic forecasts using this optimal window length are then compared to ensemble predictions from the ECMWF EPS (section 4b) and verified against and compared with surface station data (section 4c). Finally a MBG distribution is fitted to the empirical EPC at each grid point to reduce data volume and the impact on forecast performance is assessed (section 4d).

a. Optimal window length

An EPC forecast with a window size of $\pm x$ days is denoted as EPC x . Constructing the EPC x forecast based on N past years thus results in a $N(2x + 1)$ -member ensemble. In the following analysis window sizes of $x \in [0, 2, 5, 10, 15, 20]$ are investigated based on IMERG-F data from 0000 to 0000 UTC with $0.1^\circ \times 0.1^\circ$ grid resolution. We use a cross-validation approach in that EPC x forecasts are successively constructed from 18 out of the 19 years of available IMERG-F data and evaluated on the omitted year in the computation of the mean CRPS values. A time and space average of CRPS (top row in Table 1) indicates that forecast performance generally improves for increasing x but that the CRPS values for 10, 15, and 20 days are almost identical. The annual mean evolution of this parameter (colored solid lines in Fig. 2) indicates spatial mean values around 2.7 mm day⁻¹ with a considerable interannual variability of up to 8%. As expected, the year-to-year variations are highly correlated with area-mean precipitation (gray dashed line in Fig. 2). The relative behavior of the different choices of x , however, is very robust with consistently best performance of forecasts with windows of 15 and 20 days. Based on the evaluation period 2018/19 and by constructing EPC exclusively from past years of IMERG-F data (2001–17 for 2018 and 2001–18 for 2019) we test whether the extended concept, i.e., $x > 0$, results in a more skillful forecast than EPC0. A two-sided DM test applied to all days during 2018/19 and the entire 40°S–40°N belt indicates that longer windows do in fact perform significantly better in more than 96% of the grid points (bottom row in Table 1). The highest fraction is found for $x = 5$ and the lowest for $x = 20$ but differences are overall small. EPC0 is nowhere viewed superior to the nonzero windows at the chosen

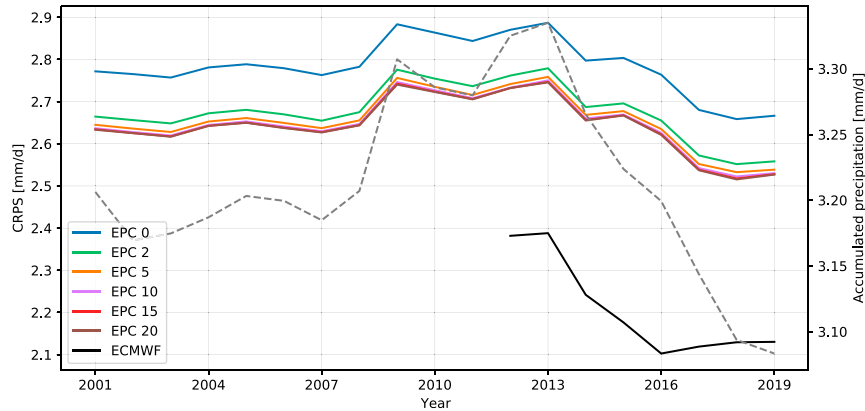


FIG. 2. Annual mean CRPS (left ordinate) for IMERG-F-based EPC forecasts based on windows of size 0 (blue), 2 (green), 5 (orange), 10 (purple), 15 (red), and 20 days (brown) for the study period 2001–19 and averaged spatially over the 40°S – 40°N belt. The gray dashed line shows spatially averaged yearly accumulated precipitation also based on IMERG-F (right ordinate). Rainfall is accumulated from 0000 to 0000 UTC on a $0.1^{\circ} \times 0.1^{\circ}$ grid. For comparison, the CRPS of ECMWF ensemble forecasts on a $0.25^{\circ} \times 0.25^{\circ}$ grid for the years 2012–19 is shown in black.

significance level $\alpha = 0.05$. Applying a TODM equivalence test to EPC15 and EPC20 (Fig. 3) reveals that the two are almost indistinguishable in their performance. Even for an equivalence margin θ as small as 0.03 mm day^{-1} 98% of grid points show significant forecast equivalence. Ultimately, we decided to use a window length of ± 15 days to account for the higher percentage of significant points in Table 1. This choice is a reasonable compromise between robust statistics and the ability to capture sudden seasonal changes in monsoon regions. Additionally, the total window length of 31 days is very close to a month, such that over land gauges and satellite estimates should be fairly consistent as a result of the monthly calibration.

b. Comparison of ECMWF ensemble forecasts with EPC15

A next logical step is to compare the performance of EPC15 with that of a dynamical model-based forecast, here from the ECMWF EPS. Recall that this requires a remapping to $0.25^{\circ} \times 0.25^{\circ}$ before the EPC computation to allow a fair comparison with the coarser model data. Evaluation is done here against IMERG-F rainfall, which creates a small advantage for EPC, as both will contain the same systematic errors. However, we will show in the next section that—at least for Africa—a validation with independent station observations leads to similar conclusions. Evaluation is performed over the period 2012–19. EPC15 forecasts are constructed by successively considering 18 of the 19 available years and using the omitted year for evaluation. The area-mean CRPS for these eight years is 2.59 mm day^{-1} for EPC15 and thus considerably less than the spatially averaged rainfall for 2012–19 of around 3.21 mm day^{-1} (Fig. 2). This value is only slightly lower than the 2.65 mm day^{-1} obtained for the whole period and finer spatial resolution given in Table 1. One would expect that the coarse-graining averages out some local errors to reduce the CRPS.

Corresponding CRPS computations for 1-day forecasts by the ECMWF ensemble yield a much lower value of 2.21 mm day^{-1} , and is in fact lower in every single year of the considered period (black line in Fig. 2). Using EPC15 as a reference forecast, a skill score can be defined for CRPS (CRPSS). The spatial distribution of this parameter is shown in Fig. 4a and reveals positive skill over most parts of the low latitudes, particularly in the subtropics. Negative values occur over the dry oceanic regions over the eastern South Pacific and South Atlantic, while corresponding areas in the Indian Ocean and Northern Hemisphere show neutral skill. Better performance of EPC15 is also found over high mountain regions such as the Andes and Himalayas. Tropical Africa stands out as an area with neutral to negative skill. Compared to the study by

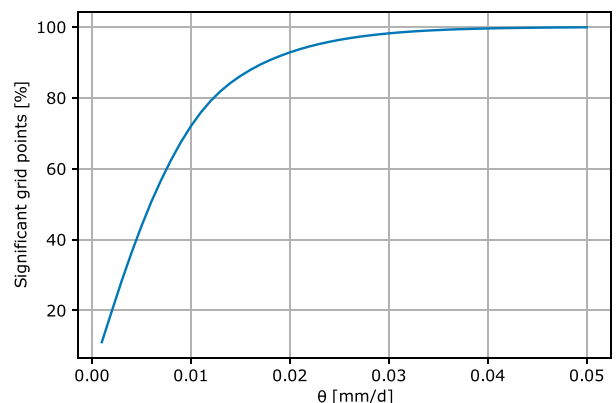


FIG. 3. Comparison of EPC15 and EPC20 based on IMERG-F data (0000–0000 UTC and $0.1^{\circ} \times 0.1^{\circ}$ grid resolution). Shown is the percentage of significant grid points according to a TODM equivalence test with Benjamini–Hochberg correction and $\alpha = 0.05$ plotted against different values of the margin θ (mm day^{-1}).

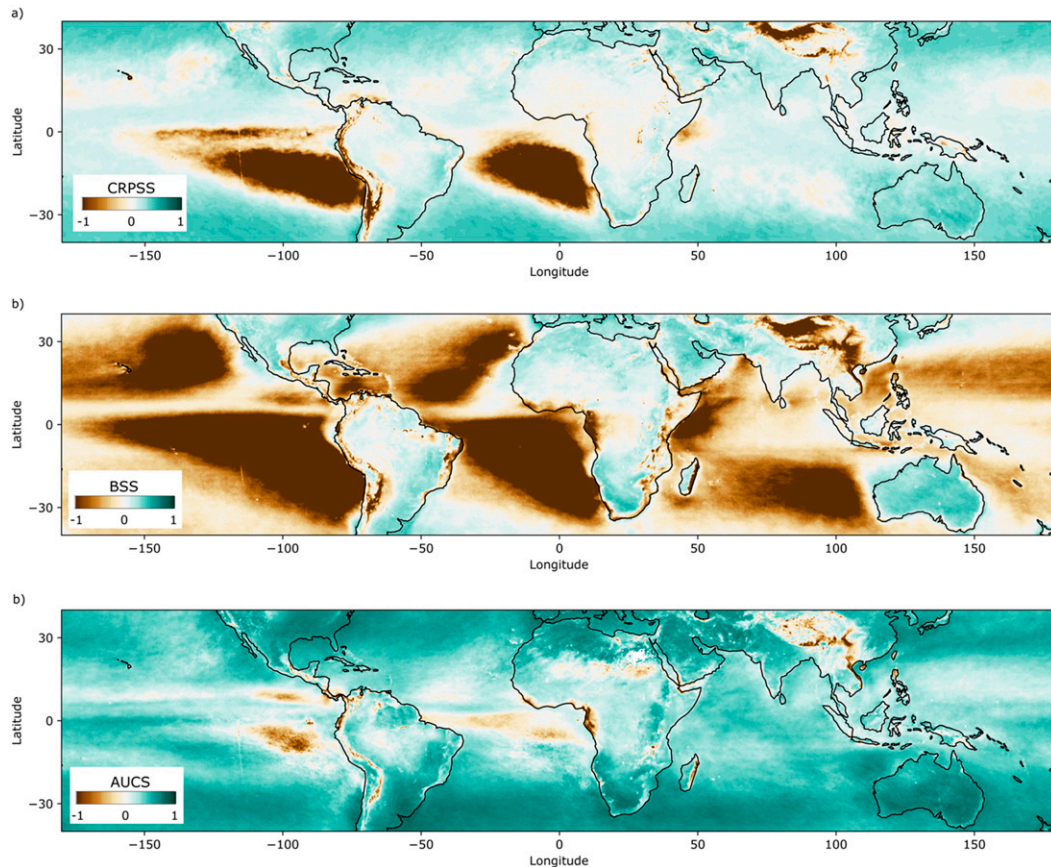


FIG. 4. (a) CRPS skill, (b) BS skill, and (c) AUC skill for ECMWF ensemble predictions for 1-day accumulated precipitation and for occurrence of precipitation obtained by thresholding 1-day accumulated precipitation at 0.2 mm. Skill measures are averaged over 2012–19 and relative to EPC15, which is constructed by successively selecting 18 years from 2001 to 2019 and omitting the year used for evaluation. Rainfall is accumulated from 0000 to 0000 UTC on a $0.25^\circ \times 0.25^\circ$ grid.

Vogel et al. (2020, their Fig. 5a), which uses a coarser resolution, an earlier time period, and TRMM instead of IMERG-F, there are considerable structural similarities but a tendency toward higher CRPS of ECMWF.

In addition to evaluating the full probabilistic forecasts using the CRPS, the skill of predicting occurrence of precipitation is investigated using the BS with a threshold of 0.2 mm. Using EPC15 as reference forecast, the BS skill (BSS) of the ECMWF ensemble is displayed in Fig. 4b. Here, a stark land-sea contrast is evident. Over land, the ECMWF ensemble is skillful over most areas except for some high mountain and coastal regions. Land areas in the inner tropics with frequent rainfall (Amazon basin, tropical Africa, southern India, Southeast Asia, Maritime Continent) show neutral skill. Over the ocean, skill is mostly negative or neutral except for near-continental areas in the subtropics such as the Mediterranean Sea. The drier parts of the oceans have strongly negative BSS, while the intertropical convergence zone region shows neutral skill. The exact reasons for these results are unclear but may lie both in issues with IMERG-F to detect warm rain or drizzle over the sea (Khan and Maggioni 2019) and with the ECMWF

model to realistically represent rainfall triggering over the homogeneous ocean surface with a weak diurnal cycle. Again there are large structural similarities with the TRMM-based analysis presented by Vogel et al. (2020, their Fig. 4).

Finally, to quantify discrimination ability, the AUC skill of ECMWF with EPC15 as the reference is visualized in Fig. 4c. Values are positive almost everywhere, indicating that the performance in CRPS and BS is negatively impacted by miscalibration (see discussion in Vogel et al. 2020). Areas with neutral or even negative discrimination skill are restricted to mountainous areas, parts of tropical Africa, and the inner tropical Atlantic and eastern Pacific Oceans.

Given the overall better quality of IMERG-F data over land and the larger socioeconomic relevance of forecasts, we also present a more detailed performance comparison for five selected tropical and subtropical regions. Figure 5 shows a map of 2001–19 averaged annual precipitation based on IMERG-F. The regions chosen are (i) the rainfall maximum in tropical South America ($82.8^\circ\text{--}52.5^\circ\text{W}$, $12.5^\circ\text{S--}10.5^\circ\text{N}$), (ii) central Africa ($17^\circ\text{W--}41^\circ\text{E}$; $5^\circ\text{S--}17^\circ\text{N}$), (iii) the Indian subcontinent and adjacent waters ($69^\circ\text{--}88^\circ\text{E}$, $7^\circ\text{--}30^\circ\text{N}$), (iv)

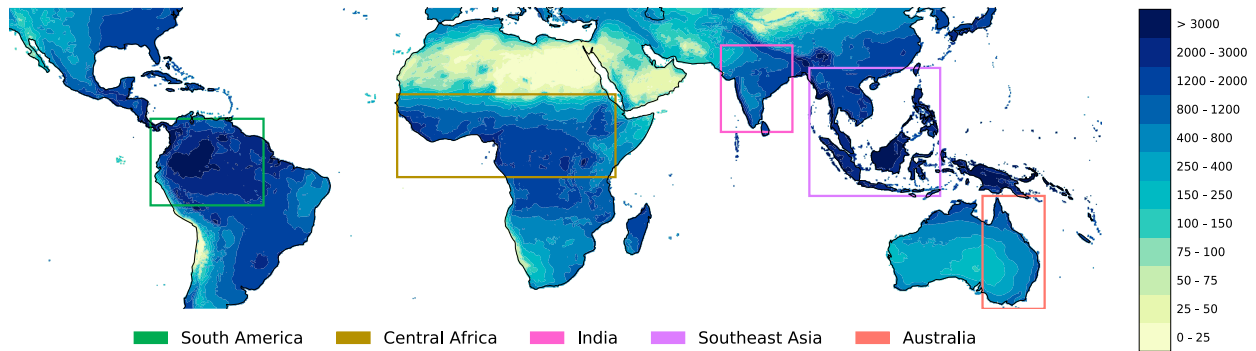


FIG. 5. Annual precipitation amount (mm) averaged over the period 2001–19 based on IMERG-F. The colored boxes define specific areas of interest used in the comparison of EPC15 and ECMWF ensemble forecasts.

Southeast Asia and the very wet Maritime Continent (92.5° – 127.25° E, 10° S– 24° N), and (v) eastern Australia (138.5° – 155° E, 40° – 10° S). For each region the mean values of CRPS, BS, and AUC are given in Table 2. In terms of CRPS, EPC15 outperforms ECMWF slightly over the wet South American region and more clearly over drier central Africa, where the reduction amounts to 4.6%. Over Southeast Asia, India, and relatively dry eastern Australia (see Fig. 5), ECMWF is superior, in the latter region with a reduction by 23%. The BS of EPC15 is always smaller or equal to the BS of ECMWF with the largest difference in central Africa. As discussed above, this is likely a result of miscalibration of the ECMWF model with respect to rainfall frequency. The higher AUC values for ECMWF for all five regions indicate better discrimination ability compared to EPC15, even for central Africa, where both CRPS and BS indicate worse performance. Largest differences are again found for Australia in agreement to the CRPS results.

c. Comparison with African station observations

The fact that the EPC15 forecasts are both constructed from and verified against the IMERG-F dataset could bias results in favor of EPC15. An independent comparison is obtained by using ground observations from KASS-D as described in section 2c. To match typical station reporting practices in Africa we use accumulation periods from 0600 to 0600 UTC for IMERG-F data and ECMWF ensemble forecasts in contrast to sections 4a and 4b. Just like in section 4b, IMERG-F data are regridded to $0.25^{\circ} \times 0.25^{\circ}$ by applying first-order conservative

remapping to allow for a fair comparison with ECMWF forecasts. For each station the nearest grid point is identified using the Haversine formula. As an evaluation period we chose 2012–16, as our station record is relatively good for this period. This implies that the ECMWF system evaluated here differs somewhat from that considered in section 4b but, as shown in Vogel et al. (2020), performance differences over time are small during the post-2011 period. The IMERG-F-based EPC15 forecasts are produced in a quasi-operational way by only using data from preceding years (e.g., 2001–11 for a forecast in 2012). Comparing point-to-area estimates of rainfall is generally problematic due to the high spatial and temporal variability of this quantity. In our case, the long sampling period of 31 days implies a correlation distance for station averages large enough to make them comparable to gridded data at 0.25° (Bell and Kundu 2003). This, however, does not hold for the full daily distribution, as point measurements can better represent extremes.

Table 3 shows mean CRPS values and results from a two-sided DM test for the five regions shown in Fig. 1. All show superiority of EPC15 over ECMWF forecasts. Reductions in CRPS range from 7%–8% in East Africa and West Sahel to 15% in the wet Guinea Coast region. According to the DM test, EPC15 is viewed superior over large fractions of stations reaching from 99% in Guinea Coast to 82% in East Africa, where ECMWF is superior over almost 17% of all stations. The results are largely consistent with those presented in section 4b in that areas over tropical Africa are

TABLE 2. Comparison of forecast performance between EPC15 and the ECMWF ensemble for each box defined in Fig. 5. The first column shows mean CRPS values (best score for each region in bold), and the second and third columns show mean BS and AUC at a threshold of 0.2 mm. Evaluation is performed for the years 2012–19. Both IMERG-F and ECMWF rainfall is accumulated from 0000 to 0000 UTC on a $0.25^{\circ} \times 0.25^{\circ}$ grid.

Area	CRPS (mm day^{-1})		BS		AUC	
	ECMWF	EPC15	ECMWF	EPC15	ECMWF	EPC15
South America	4.04	3.99	0.17	0.15	0.77	0.73
Central Africa	2.36	2.25	0.21	0.15	0.82	0.79
India	2.34	2.76	0.17	0.16	0.87	0.81
Australia	1.54	2	0.19	0.19	0.85	0.63
Southeast Asia	4.35	4.78	0.23	0.19	0.78	0.72

TABLE 3. Verification of IMERG-F-based EPC15 and ECMWF ensemble forecasts against the African station observations displayed and grouped in Fig. 1 for the period 2012–16. The number of available stations is given in the first column. Following columns show mean CRPS values (best score for each region in bold) and the percentages of stations, where EPC15 and ECMWF forecasts are significantly superior according to a two-sided DM test results with Benjamini–Hochberg correction and $\alpha = 0.05$. Rainfall is accumulated from 0600 to 0600 UTC on a $0.25^\circ \times 0.25^\circ$ grid.

Area	No. of stations	CRPS (mm day ⁻¹)		DM test (%)	
		ECMWF	EPC 15	ECMWF	EPC 15
West Sahel	65	2.07	1.93	6.15	92.31
Guinea Coast	103	3.88	3.31	0	99.03
East Sahel	24	1.11	1.02	4.16	91.67
Ethiopia	48	2.38	2.08	10.42	87.50
East Africa	84	2.89	2.70	16.67	82.14

particular difficult to forecast for ECMWF in contrast to other tropical and subtropical regions. The fact that CRPS reductions are comparable to Table 2 and Fig. 4a indicates that the superiority of EPC15 is not strongly influenced by the choice of the verification dataset.

To further investigate this issue, we also compute EPC15 from station observations. For an optimal comparison with the IMERG-F-based EPC15 we select 76 stations from KASS-D with 100% data coverage during the period 2001–16, 19 in West Sahel, 50 in Guinea Coast, and 7 in East Africa. Using station rainfall for verification, we can systematically compare the forecast performance of the ECMWF ensemble with the IMERG-F-based and the station-based EPC15. As expected, ECMWF shows largest CRPS in all three regions followed by IMERG-F EPC15 and then station EPC15, but the difference between the former two is always larger than between the latter two (Table 4). For the West Sahel and Guinea Coast regions the difference in CRPS between the two EPCs is 1%–2%, while it reaches 6% in East Africa, where, however, only seven stations are available. For all three regions, the CRPS increases relative to the results shown in Table 3, which is caused by the smaller numbers of stations and the different time period used here. Interestingly, however, the gap in CRPS between EPC15 and ECMWF remains almost constant, which underlines the robustness of the EPC concept against time and space resolution, dataset used, and time period considered.

d. Mixed Bernoulli–gamma fit

The previous sections have demonstrated that the EPC concept produces forecasts of comparable skill to state-of-the-art NWP models and thus can serve as a benchmark for

future forecast developments. To make the concept easy to access and use, we will test in the following whether the empirical EPC can be replaced by a smooth fitted theoretical distribution without losing much predictive ability. Given that an EPC15 forecast based on 19 years of IMERG-F data consists of 589 rainfall values, a fitted three-parameter distribution reduces the data volume by a factor of almost 200. Fitting a probability distribution is also more convenient to work with (Bröcker and Smith 2008) and allows to derive more consistent quantiles and probabilities, particular for more extreme events (Wilks 2002).

To jointly represent occurrence and amount of precipitation, we follow Williams (1998) and propose using a MBG distribution (Sloughter et al. 2007; Cannon 2008) with parameter p for the probability of a nonzero event, shape parameter α , and scale parameter θ of the gamma distribution (see appendix). To allow for a performance comparison with the raw EPC15, the CRPS for the MBG distribution is derived following the steps from Scheuerer and Möller (2015) (see the appendix) resulting in

$$\begin{aligned} \text{CRPS}(F_{\Gamma_{\alpha,\beta}}, y) = & p y [2\Gamma_{\alpha,\beta}(y)] - p \frac{\alpha}{\beta} [2\Gamma_{\alpha+1,\beta}(y)] \\ & - p^2 \frac{\alpha}{\beta\pi} B\left(\alpha + \frac{1}{2}, \frac{1}{2}\right) + y(1-2p) + p^2 \frac{\alpha}{\beta}, \end{aligned} \quad (3)$$

where B denotes the beta function, $\Gamma_{\alpha,\beta}$ is the gamma distribution, and $\beta = 1/\theta$ is an inverse scale parameter, called rate parameter. The parameters of the gamma distribution are fitted using maximum likelihood estimation.

The following calculations are based on IMERG-F data from 0000 to 0000 UTC and $0.1^\circ \times 0.1^\circ$ grid resolution as in

TABLE 4. Comparison between the mean 2012–16 CRPS (mm day⁻¹) for ECMWF ensemble forecasts, EPC15 based on IMERG-F, and EPC15 based on surface stations with 100% data availability in the period 2001–16 (best score for each region in bold). The number of stations in each region is given in the first column. Verification is done against station observations. Accumulation is from 0600 to 0600 UTC, and the grid resolution is $0.25^\circ \times 0.25^\circ$.

Area	No. of stations	ECMWF	IMERG-F EPC15	Station EPC15
West Sahel	19	2.32	2.18	2.15
Guinea Coast	50	3.97	3.43	3.37
East Africa	7	3.14	2.95	2.77

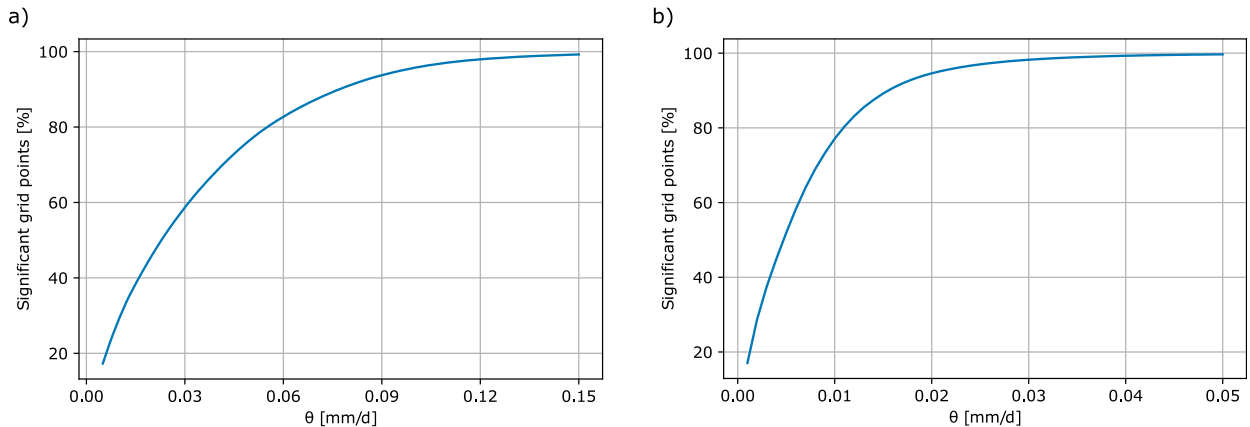


FIG. 6. Equivalence of EPC15 and MBG forecasts for the years 2018 and 2019 based on IMERG-F data (0000–0000 UTC accumulation and $0.1^\circ \times 0.1^\circ$ grid resolution) during 2011–17 and 2011–18, respectively, according to a TODM equivalence test with Benjamini–Hochberg correction and $\alpha = 0.05$. Shown are the percentage of significant grid points for different values of the equivalence margin θ (mm day^{-1}) for (a) all points and (b) land points only (i.e., proportion of water surface $< 50\%$ according to the GPM IMERG land–sea mask; Olson et al. 2019). Note the different θ ranges in the two plots.

section 4a. EPC15 is constructed from the years 2001–17 for a 2018 forecast and from 2001 to 2018 for a 2019 forecast as in section 4b. Comparing EPC15 and MBG for these two years results in mean CRPS values of 2.52 and 2.54 mm day^{-1} , respectively, corresponding to a loss of less than 1%. The former value is slightly lower than the one given in section 4b due to the differences in resolution and time period. Concentrating on land points, i.e., with a proportion of water surface $< 50\%$ according to the GPM IMERG land–sea mask (Olson et al. 2019), the mean CRPS reduces to 1.94 mm day^{-1} for both forecasts. A TODM equivalence test shows that already for a relatively minor margin θ of 0.03 mm day^{-1} ($\sim 1.2\%$ of the CRPS) 58.7% of all grid points are equivalent (Fig. 6a). Results of the two forecasts are practically indistinguishable for θ greater 0.15 mm day^{-1} . The curve for land points only is much steeper, yielding 98.2% equivalence already for a θ of 0.03 mm day^{-1} (Fig. 6b), indicating that MBG is a very good alternative to EPC15 over land.

A horizontal distribution of the CRPSS of MBG relative to EPC15 (Fig. 7) shows that most land points have values close to zero with the exception of arid Australia, where a predominance of dry days may make the MBG fit difficult. Such a behavior, however, is not found over some other continental deserts such as the Sahara and Arabian Desert. Interestingly, some continental areas such as South America and Africa even show a slightly superior forecast by MBG. As evident from the average numbers stated above, performance over the ocean is much less consistent. In particular over the drier regions in the outer tropics MBG has a worse performance. Here, many dry and drizzly days make the fit of the positively skewed MBG distribution difficult, likely leading to too much weight for higher values. The negative bias of IMERG-F in these areas (Khan and Maggioni 2019) may exacerbate this problem. As these areas are largely uninhabited and contribute only small amounts to the global totals, larger errors there are relatively acceptable. We therefore conclude that

MBG should be used as the benchmark forecast instead of the raw EPC15 due to the much smaller data volume and the advantages of having a full probability distribution.

5. Conclusions

Rainfall forecasts, even at short lead times of only a day, still constitutes a large challenge for tropical and subtropical latitudes. Current NWP systems generally have low skill over these areas (Haiden et al. 2012; Vogel et al. 2020), calling for enhanced efforts for improvement through postprocessing, model development, higher resolution, statistical, or hybrid approaches. To systematically and consistently gauge progress in such developments, the definition of an adequate benchmark forecast is needed.

Here we used satellite-based rainfall estimates from IMERG-F version V06B for the period 2001–19 to define such a benchmark for the 40°S – 40°N belt. Given the large stochasticity in tropical and subtropical rainfall, we concentrated on probabilistic forecasts. The concept we use is based on past observations in a window around the date of interest to construct an empirical-climatological probability forecast and is termed Extended Probabilistic Climatology (EPC). Applying the EPC concept to IMERG-F rainfall data and comparing the results to ECMWF 1-day ensemble forecasts and station data over tropical Africa, led to the following main conclusions:

- A length of ± 15 days is identified as an optimal window length for the EPC construction (termed EPC15), as it provides a good compromise between statistical robustness and enough flexibility to account for sudden seasonal onsets of rainfall (e.g., in monsoon regions).
- The area- and time-mean CRPS for EPC15 at the highest spatial resolution of 0.1° is 2.65 mm day^{-1} . Year-to-year variability of this value is closely correlated with area

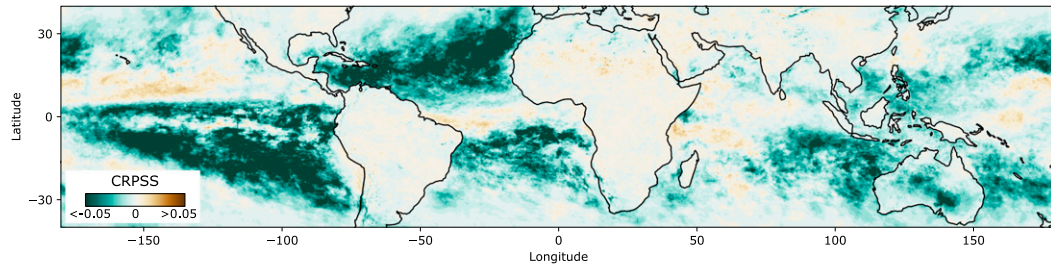


FIG. 7. Average CRPS skill of MBG forecasts relative to EPC15 forecasts for the years 2018 and 2019 based on IMERG-F data (0000–0000 UTC and $0.1^\circ \times 0.1^\circ$ grid resolution) during 2001–17 and 2001–18, respectively.

mean annual rainfall. Coarse-graining the IMERG-F data to 0.25° before EPC computation reduces the CRPS only slightly.

- Raw ECMWF ensemble predictions are superior to EPC15 over most parts of the tropics and subtropics in terms of forecasting the distribution of rainfall with the exception of oceanic deserts, high mountain areas, and parts of tropical Africa, where the difference in CRPS can exceed 10% in subregions.
- With respect to precipitation occurrence, the skill of ECMWF is neutral or slightly positive over land and strongly negative over large parts of the low-latitude oceans. The fact that potential predictive ability is positive almost everywhere, points to considerable miscalibration problems.
- Focusing specifically on tropical Africa, the superiority of EPC15 against ECMWF forecasts is robust against an EPC15 construction and evaluation with surface stations instead of IMERG-F data.
- The empirical EPC15 can be replaced by the three parameters of a fitted MBG distribution without much loss in predictive quality. Over dry oceanic areas, where some issues with IMERG-F data quality have been documented, the MBG fit is difficult, but over many land areas the MBG-based forecasts are even superior.

Based on these findings, we advocate the IMERG-F-based MBG-fitted EPC15 as an adequate and powerful benchmark ensemble forecast for future rainfall prediction studies focusing on low latitudes. To make the results as accessible as possible for a wider community of researchers and operational weather services we have (i) set up a website (<http://www.epc.kit-weather.de>), where EPC15 forecasts are interactively displayed based on a clickable map ($0.1^\circ \times 0.1^\circ$ grid) and lists of significant cities per country; (ii) provided the code to replicate the results of this paper on Github (<https://github.com/ewwalz/epc/>); and (iii) made available parameters of the fitted MBG distribution for the standard benchmark EPC15 ($0.1^\circ \times 0.1^\circ$ grid) for each day of the year under (<https://doi.org/10.5445/IR/1000127274>).

We strongly encourage other scientists to actively use these resources, in particular to study in detail other low-latitude regions based on their station records or to extend the analysis presented in this paper to longer lead times. The new benchmark should also be used as a reference to further develop postprocessing procedures that, as shown in Vogel et al. (2018, 2020), for ECMWF ensemble predictions, can lead to significant

improvements relative to the raw model output evaluated here. Hopefully future generations of NWP models or postprocessing, statistical (e.g., Vogel et al. 2021) or hybrid models will be able to outperform EPC15 and show truly skillful forecasts over tropical Africa. This would be of enormous socioeconomic relevance for the large and growing population, mostly in developing countries.

Acknowledgments. The research leading to these results has been accomplished within project C2 “Statistical-dynamical forecasts of tropical rainfall” of the Transregional Collaborative Research Center SFB/TRR 165 “Waves to Weather” funded by the German Science Foundation (DFG). The authors thank Sebastian Lerch, Alexander Jordan, and Tilmann Gneiting for advice and discussion and for providing preprocessed data. We are grateful to three anonymous reviewers whose constructive comments helped to improve an earlier version of this paper.

Data availability statement. IMERG “final run” data are publicly available and can be downloaded from the NASA data store under <https://gpm.nasa.gov/data/directory>. There, daily data were accessed through the Precipitation Processing System (PPS) data servers, which requires registration. The ECMWF ensemble forecast is provided through the ECMWF Meteorological Archival and Retrieval System (MARS) catalogue under <https://www.ecmwf.int/en/forecasts>.

APPENDIX

Mixed Bernoulli–Gamma Distribution

Let p be the probability of a nonzero event and denote by $g_{\alpha,\beta}$ the probability density function (PDF) of a gamma distribution $\Gamma_{\alpha,\beta}$ with shape parameter α and inverse scale parameter β , called rate parameter, then the mixed Bernoulli–gamma (MBG) PDF is

$$f_{p,\alpha,\beta}(y) = \begin{cases} pg_{\alpha,\beta}(y), & y > 0 \\ 1 - p, & \text{else} \end{cases},$$

and the MBG cumulative distribution function (CDF) is

$$F_{p,\alpha,\beta}(y) = \begin{cases} (1 - p) + p\Gamma_{\alpha,\beta}(y), & y > 0 \\ 1 - p, & \text{else} \end{cases}. \quad (\text{A1})$$

The quantile function for the MBG distribution is

$$Q_{p,\alpha,\beta}(q) = \begin{cases} \Gamma_{\alpha,\beta}^{-1}\left(\frac{q-1+p}{p}\right), & q > 1-p \\ 0, & \text{else} \end{cases}$$

Let X have CDF in Eq. (A1), then

$$\mathbb{E}(X) = p \frac{\alpha}{\beta},$$

$$\text{Var}(X) = \frac{p[\alpha(1+\alpha)]}{(\beta^2) - p^2\alpha^2/\beta^2}.$$

The CRPS for a CDF F and the corresponding observation y is defined as

$$\text{CRPS}(F, y) = \int_{-\infty}^{+\infty} [F(t) - 1_{\{y \leq t\}}]^2 dt$$

$$= \mathbb{E}_F |Y - y| - \frac{1}{2} \mathbb{E}_F |Y - Y'|,$$

where Y and Y' are independent random variables with distribution function F and finite first moment (Gneiting and Raftery 2007). The derivation of the CRPS for the MBG distribution is based on considerations from appendix A in Scheuerer and Möller (2015):

$$\begin{aligned} \mathbb{E}|Y - y| &= \int_0^y (y-t)f_{p,\alpha,\beta}(t) dt - \int_y^\infty (y-t)f_{p,\alpha,\beta}(t) dt + (1-p)y \\ &= py[2\Gamma_{\alpha,\beta}(y) - 1] - \int_0^y tf_{p,\alpha,\beta}(t) dt + \int_y^\infty tf_{p,\alpha,\beta}(t) dt + (1-p)y \\ &= py[2\Gamma_{\alpha,\beta}(y) - 1] - p \frac{\alpha}{\beta} \int_0^y g_{\alpha+1,\beta}(t) dt + p \frac{\alpha}{\beta} \int_y^\infty g_{\alpha+1,\beta}(t) dt + (1-p)y \\ &= py[2\Gamma_{\alpha,\beta}(y) - 1] - p \frac{\alpha}{\beta} [2\Gamma_{\alpha+1,\beta}(y) - 1] + (1-p)y \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}|Y - Y'| &= \int_0^\infty \int_0^\infty |x - x'| f_{p,\alpha,\beta}(x) f_{p,\alpha,\beta}(x') dx dx' \\ &+ (1-p) \int_0^\infty x f_{p,\alpha,\beta}(x) dx + (1-p) \int_0^\infty x' f_{p,\alpha,\beta}(x') dx' \end{aligned}$$

$$\begin{aligned} &= \int_0^\infty \int_0^\infty |x - x'| f_{p,\alpha,\beta}(x) f_{p,\alpha,\beta}(x') dx dx' + 2p(1-p) \frac{\alpha}{\beta} \\ &= 2p(1-p) \frac{\alpha}{\beta} + 2p^2 \frac{\alpha}{\beta\pi} B\left(\alpha + \frac{1}{2}, \frac{1}{2}\right), \end{aligned}$$

where B is the beta function. Putting both results together yields the following:

$$\begin{aligned} \text{CRPS}(F, y) &= py[2\Gamma_{\alpha,\beta}(y) - 1] - p \frac{\alpha}{\beta} [2\Gamma_{\alpha+1,\beta}(y) - 1] + (1-p)y - 2p(1-p) \frac{\alpha}{\beta} - p^2 \frac{\alpha}{\beta\pi} B\left(\alpha + \frac{1}{2}, \frac{1}{2}\right) \\ &= py[2\Gamma_{\alpha,\beta}(y) - 1] - p \frac{\alpha}{\beta} [2\Gamma_{\alpha+1,\beta}(y)] + (1-p)y + 2p^2 \frac{\alpha}{\beta} - p \frac{\alpha}{\beta} - p^2 \frac{\alpha}{\beta\pi} B\left(\alpha + \frac{1}{2}, \frac{1}{2}\right) \\ &= py[2\Gamma_{\alpha,\beta}(y)] - p \frac{\alpha}{\beta} [2\Gamma_{\alpha+1,\beta}(y)] - p^2 \frac{\alpha}{\beta\pi} B\left(\alpha + \frac{1}{2}, \frac{1}{2}\right) + y(1-2p) + p^2 \frac{\alpha}{\beta}. \end{aligned}$$

REFERENCES

Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, <https://doi.org/10.1126/science.aav7274>.

Anderson, S., and W. W. Hauck, 1983: A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun. Stat. Theory Methods*, **12**, 2663–2692, <https://doi.org/10.1080/03610928308828634>.

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>.

Bell, T. L., and P. K. Kundu, 2003: Comparing satellite rainfall estimates with rain gauge data: Optimal strategies suggested by a spectral model. *J. Geophys. Res.*, **108**, 4121, <https://doi.org/10.1029/2002JD002641>.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.

Berger, R. L., and J. C. Hsu, 1996: Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.*, **11**, 283–319, <https://doi.org/10.1214/ss/1032280304>.

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).

Bröcker, J., and L. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus*, **60A**, 663–678, <https://doi.org/10.3402/tellusa.v60i4.15387>.

Cannon, A. J., 2008: Probabilistic multisite precipitation downscaling by an expanded Bernoulli–gamma density

- network. *J. Hydrometeorol.*, **9**, 1284–1300, <https://doi.org/10.1175/2008JHM960.1>.
- Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.
- Fawcett, T., 2006: Introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874, <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- 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>.
- 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>.
- Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.*, **8**, 38–55, <https://doi.org/10.1175/JHM560.1>.
- , D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, C. Kidd, E. J. Nelkin, and P. Xie, 2015: NASA Global Precipitation Measurement (GPM) Integrated Multi-Satellite Retrievals for GPM (IMERG). Algorithm Theoretical Basis Doc., version 4.5, 30 pp., https://gpm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V4.5.pdf.
- , E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan, 2019a: GPM IMERG final precipitation L3 1 day 0.1 degree \times 0.1 degree V06. Goddard Earth Sciences Data and Information Services Center (GES DISC), accessed 12 November 2019, <https://doi.org/10.5067/GPM/IMERGDF/DAY/06>.
- , —, —, —, and —, 2019b: V06 IMERG Release Notes. NASA, 6 pp., https://gpm.nasa.gov/sites/default/files/2020-02/IMERG_V06_release_notes_190503.pdf.
- Jaun, S., and B. Ahrens, 2009: Evaluation of a probabilistic hydrometeorological forecast system. *Hydrol. Earth Syst. Sci.*, **13**, 1031–1043, <https://doi.org/10.5194/hess-13-1031-2009>.
- Johnston, R. J., and J. M. Duke, 2008: Benefit transfer equivalence tests with non-normal distributions. *Environ. Resour. Econ.*, **41**, 1–23, <https://doi.org/10.1007/s10640-007-9172-x>.
- Khan, S., and V. Maggioni, 2019: Assessment of level-3 gridded Global Precipitation Mission (GPM) products over oceans. *Remote Sens.*, **11**, 255, <https://doi.org/10.3390/rs11030255>.
- Kniffka, A., and Coauthors, 2020: An evaluation of operational and research weather forecasts for southern West Africa using observations from the DACCWA field campaign in June–July 2016. *Quart. J. Roy. Meteor. Soc.*, **146**, 1121–1148, <https://doi.org/10.1002/qj.3729>.
- Kummerow, C., W. Barnes, T. Kozu, J. Shiue, and J. Simpson, 1998: The Tropical Rainfall Measuring Mission (TRMM) sensor package. *J. Atmos. Oceanic Technol.*, **15**, 809–817, [https://doi.org/10.1175/1520-0426\(1998\)015<0809:TTRMMT>2.0.CO;2](https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2).
- Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, <https://doi.org/10.1287/mnsc.22.10.1087>.
- 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>.
- 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>.
- Nicholson, S. E., C. Funk, and A. H. Fink, 2018: Rainfall over the African continent from the 19th through the 21st century. *Global Planet. Change*, **165**, 114–127, <https://doi.org/10.1016/j.gloplacha.2017.12.014>.
- Olson, B., D. Bolvin, and G. Huffman, 2019: Land/Sea static mask relevant to IMERG precipitation 0.1 \times 0.1 degree V2 (GPM_ IMERG_LandSeaMask). Goddard Earth Sciences Data and Information Services Center (GES DISC), accessed 19 March 2020, <https://doi.org/10.5067/6P5EM1HPR3VD>.
- Pagano, T. C., D. L. Shrestha, Q. J. Wang, D. Robertson, and P. Hapuarachchi, 2013: Ensemble dressing for hydrological applications. *Hydrol. Processes*, **27**, 106–116, <https://doi.org/10.1002/hyp.9313>.
- Pante, G., and P. Knippertz, 2019: Resolving Sahelian thunderstorms improves mid-latitude weather forecasts. *Nat. Commun.*, **10**, 3487, <https://doi.org/10.1038/s41467-019-11081-4>.
- Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.*, **522**, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>.
- 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>.
- Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Moutadid, and N. Thuerey, 2020: WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002203, <https://doi.org/10.1029/2020MS002203>.
- 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>.
- Scheuerer, M., and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9**, 1328–1349, <https://doi.org/10.1214/15-AOAS843>.
- Schneider, U., M. Ziese, A. Meyer-Christoffer, P. Finger, E. Rustemeier, and A. Becker, 2016: The new portfolio of global precipitation data products of the Global Precipitation Climatology Centre suitable to assess and quantify the global water cycle and resources. *Proc. Int. Assoc. Hydrol. Sci.*, **374**, 29–34.
- Schuurmann, D. J., 1987: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.*, **15**, 657–680, <https://doi.org/10.1007/BF01068419>.
- Shi, X., Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, 2015: Convolutional LSTM Network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., MIT Press, 802–810.
- Sloughter, J., 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>.
- 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>.
- , —, —, —, and —, 2021: Statistical forecasts for the occurrence of precipitation outperform global models over northern tropical Africa. *Geophys. Res. Lett.*, **48**, e2020GL091022, <https://doi.org/10.1029/2020GL091022>.
- Wilks, D. S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836, <https://doi.org/10.1256/qj.01.215>.
- , 2016: “The stippling shows statistically significant grid points”: How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, <https://doi.org/10.1175/BAMS-D-15-00267.1>.
- Williams, P. M., 1998: Modelling seasonality and trends in daily rainfall data. *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., MIT Press, 985–991.