SPECIAL
Waves to Weather (W2W)
COLLECTION

Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall in the Tropics

PETER VOGEL,[a,b] PETER KNIPPERTZ,[a] ANDREAS H. FINK,[a] ANDREAS SCHLUETER,[a,c] AND TILMANN GNEITING[d,b]

[a] *Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany*
[b] *Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany*
[c] *Department of Computer Science, Stanford University, Stanford, California*
[d] *Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*

(Manuscript received 26 May 2020, in final form 1 September 2020)

ABSTRACT: Precipitation forecasts are of large societal value in the tropics. Here, we compare 1–5-day ensemble predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF, 2009–17) and the Meteorological Service of Canada (MSC, 2009–16) over 30°S–30°N with an extended probabilistic climatology based on the Tropical Rainfall Measuring Mission 3 B42 gridded dataset. Both models predict rainfall occurrence better than the reference only over about half of all land points, with a better performance by MSC. After applying the postprocessing technique ensemble model output statistics, this fraction increases to 87% (ECMWF) and 82% (MSC). For rainfall amount there is skill in many tropical areas (about 60% of land points), which can be increased by postprocessing to 97% (ECMWF) and 88% (MSC). Forecasts for extremes (>20 mm) are only marginally worse than those of occurrence but do not improve as much through postprocessing, particularly over dry areas. Forecast performance is generally best over arid Australia and worst over oceanic deserts, the Andes and Himalayas, as well as over tropical Africa, where models misrepresent the high degree of convective organization, such that even postprocessed forecasts are hardly better than climatology. Skill of 5-day accumulated forecasts often exceeds that of shorter ranges, as timing errors matter less. An increase in resolution and major model update in 2010 has significantly improved ECMWF predictions. Especially over tropical Africa new techniques such as convection-permitting models or combined statistical-dynamical forecasts may be needed to generate skill beyond the climatological reference.

SIGNIFICANCE STATEMENT: Accurate forecasts of rainfall could support tropical countries to more effectively manage key resources such as water, food, health, and energy. Here we assessed the usefulness of 1–5-day predictions from two leading weather centers against satellite-based rainfall estimates. The forecast models failed to predict the probability of rainfall occurrence better than a climatological reference in many parts of the tropics but showed some value in predicting rainfall amounts and even extremes. Statistical correction methods can significantly improve the raw model output except for high mountain ranges, some coastal areas, and most of tropical Africa. Future studies should refine statistical correction methods, run forecast models at higher spatial resolution, improve model physics, and experiment with statistical forecast techniques.

KEYWORDS: Tropics; Precipitation; Statistical techniques; Ensembles; Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF)

1. Introduction

Numerical weather prediction (NWP) has steadily improved over the last decades, allowing a multitude of socioeconomic benefits to be realized (Bauer et al. 2015; Alley et al. 2019). While progress is unmistakable for 500-hPa geopotential heights and mean sea level pressure in the extratropics, improvements in the predictions of many other parameters are more variable (Navascués et al. 2013). For example, forecasts of European cloud cover have hardly improved over the last more than 10 years (Haiden et al. 2018). A region generally characterized by low forecast skill and high uncertainty is the tropical belt. Haiden et al. (2012) note that 1-day precipitation forecasts at low latitudes have skill similar to 6-day forecasts in the extratropics. Little progress has been made also for free-tropospheric winds in the tropics (Haiden et al. 2018).

For variables with large forecast uncertainty, ensemble prediction is of particular importance, even for short ranges (Leutbecher and Palmer 2008; Zhang and Pu 2010). However, Vogel et al. (2018) find that there is little to no skill in precipitation forecasts from ten global NWP ensemble prediction systems over northern tropical Africa. Their results are robust against temporal and spatial aggregation and point to fundamental problems in predicting precipitation in this region. Similar problems were reported by Medina et al. (2019) for tropical Brazil. Models appear to perform better in areas in the outer tropics and subtropics (Medina et al. 2019; Webster et al. 2011) or during time periods

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

Corresponding author: Peter Knippertz, peter.knippertz@kit.edu

DOI: 10.1175/WAF-D-20-0082.1

when the tropics are strongly influenced by the more predictable extratropical circulation (Davis et al. 2013; van der Linden et al. 2017). The sobering performance of current NWP systems in the tropics have substantial socioeconomic implications, as the majority of developing countries are located in this area. Their populations are especially vulnerable to weather disasters and often underserved by forecasting (Alley et al. 2019).

Why is there so little progress in tropical weather forecasting, although many challenges have been realized for decades (e.g., Smith et al. 2001)? First, initial uncertainties tend to be largest in equatorial regions (Žagar 2017). This is caused by an insufficient observational network, data assimilation algorithms optimized for midlatitude conditions, and large model errors, which also contribute to a fast degradation of forecast quality (Privé and Errico 2013). Conventional observations such as surface stations and weather balloons are scarce at low latitudes, particularly over the vast tropical oceans. Consequently, the observing system is dominated by satellite data, which are heavily skewed toward measuring atmospheric mass variables rather than wind (e.g., Baker et al. 2014). However, data denial experiments for periods with a much enhanced radiosonde network during field campaigns over West Africa have shown a relatively small impact on model performance, illustrating the importance of model errors and the assimilation system (Agustí-Panareda et al. 2010; van der Linden et al. 2020).

Second, the tropics are dominated by convective processes and are therefore particularly sensitive to the representation of deep convection, which is parameterized in all current global NWP models. This can create an erroneous diurnal cycle and impede the mesoscale organization of convection, which in turn can quickly lead to a degradation of, for example, the West African monsoon circulation (Marsham et al. 2013) with impacts on forecasts far beyond Africa (Pante and Knippertz 2019). Despite many improvements, however, forecasts using explicit convection still suffer from biases and other deficits (Kniffka et al. 2020; Peters et al. 2019), particularly in areas where a high degree of convective organization makes forecasts challenging. Estimates of intrinsic predictability using storm-resolving simulations show that in the tropics convection limits the forecast horizon to few days at scales of 100 km (Judt 2020).

Third, small-scale physical processes such as cloud microphysics and radiation can relatively easily affect scales large enough to be of interest to predictions through their effects on the vertical profiles of latent (and radiative) heating and thus divergent wind. For example, convective invigoration by increased cloud condensation nuclei (Rosenfeld et al. 2008) and larger or longer-lived anvils (Fan et al. 2013) affect convective organization and coupling to larger-scale circulations. The most important example of such a coupling on weather time scales are equatorial waves, classically referring to planetary-scale solutions of the shallow water equations for the tropics (Matsuno 1966; Wheeler and Kiladis 1999). The coupling relies on a wave-induced modification of environmental conditions for convection such as convergence, stability, moisture availability, and shear (Schlueter et al. 2019a,b). Although a relatively high level of intrinsic predictability has recently been shown for equatorial waves (Li and Stechmann 2020; Judt 2020), NWP models are known to lose wave energy too quickly and to misrepresent propagation, partly due to precipitation being triggered too easily by convective parameterization schemes (Lin et al. 2008; Frierson et al. 2011; Dias et al. 2018; Bengtsson et al. 2019).

This paper provides a comprehensive assessment of our current ability to predict rainfall in the tropics with global ensemble systems. Predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF, 2009–17) and the Meteorological Service of Canada (MSC, 2009–16 only due to limited data availability) will be compared, since both performed well in past model intercomparisons over West Africa (Vogel et al. 2018) and Ethiopia by Stellingwerf et al. (2020). The analysis will evaluate the whole probability distribution with separate assessments for rainfall occurrence, amount, and extremes. In addition, the potential of statistical postprocessing to correct for systematic biases and dispersion errors (see Vannitsem et al. 2018, for an overview) is tested here systematically for rainfall forecasts for the first time, to our knowledge. The results will be a first step toward potential improvements to be designed and tested specifically for the tropics, which in the long run can inform socioeconomically important decision in areas such as weather warnings (particularly of heavy precipitation and flooding; see, e.g., Engel et al. 2017), water management, energy production, agriculture, and disease prevention.

Section 2 introduces the analyzed ensemble forecasts and the satellite observations used for validation. Section 3 explains the construction of our climatological reference forecast and the methods used for forecast evaluation and to statistically postprocess raw ensemble precipitation forecasts. Section 4 presents the results of our investigations, starting with the assessment of calibration and reliability of raw and postprocessed ensemble forecasts before considering the skill in predicting precipitation occurrence, amount, and extremes. Additionally, the improvement over the investigation period is analyzed. Section 5 summarizes the main outcomes and gives an outlook.

2. Data

The evaluation of precipitation forecasts by the ECMWF and MSC ensembles will be done over the tropical belt between 30°S and 30°N. Both systems will be described in section 2a. Tropical Rainfall Measuring Mission (TRMM) rainfall estimates will be used as an observational reference and are described in section 2b. All datasets are spatially averaged over the same $1° \times 1°$ longitude–latitude boxes resulting in 21 600 data points. Forecast evaluation results will likely depend on spatial resolution but Vogel et al. (2018) showed that their conclusions for northern tropical Africa were fairly robust for latitude–longitude boxes from $0.25° \times 0.25°$ to $2° \times 5°$. We assess forecast quality for accumulation periods between 1 and 5 days. Due to data availability the evaluation will cover the period 2009–17 for ECMWF forecasts and 2009–16 for MSC. Only annual statistics will be presented but particularly for the outer (and often drier) parts of the tropics, a more seasonal perspective would be beneficial.

a. Forecasts

ECMWF is one of the leading providers of ensemble prediction information worldwide. Its ensemble prediction system (EPS) 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 spatial resolution. The ENS members have the same spatial resolution as the CNT run but are started from perturbed initial conditions and are subject to a stochastic representation of model uncertainties (Buizza et al. 1999). Molteni et al. (1996) and Leutbecher and Palmer (2008) describe generation and properties of the ECMWF EPS in detail. For comparison, we consider the EPS of the Meteorological Service of Canada (MSC; 21 members). It is among the best-performing EPSs for accumulated precipitation in northern tropical Africa (Vogel et al. 2018). Both ensemble forecasts are accessible via the TIGGE archive (https://confluence.ecmwf.int/display/TIGGE). Park et al. (2008) and Bougeault et al. (2010) discuss objectives and the setup of TIGGE, including the participating EPSs, while Swinbank et al. (2016) report on recent research and achievements. The data were spatially averaged from the original 0.5° to a 1° grid. For both models, we use forecasts initialized at 0000 UTC. The forecast quality of both EPSs can be monitored in quasi-real time at the World Meteorological Organization (WMO) Lead Centre on Verification of Ensemble Prediction Systems website http://epsv.kishou.go.jp/EPSv.[1] It displays average scores for standard atmospheric variables for the tropical belt between 20°S and 20°N, and the Northern and Southern Hemisphere extratropics.

b. Observations

For a spatially consistent and complete forecast verification, we rely on the TRMM 3B42 gridded dataset. TRMM merges active measurements from a space-borne precipitation radar with passive, radar-calibrated information from infrared as well as microwave measurements (Huffman et al. 2007). Based on monthly accumulations, TRMM estimates are calibrated against nearby gauge observations. The data are available on a 0.25° × 0.25° grid with 3-hourly temporal resolution.

The TRMM 3B42 product is regarded to be one of the best available satellite precipitation estimates (e.g., Maggioni et al. 2016) and has been shown to represent daily and even subdaily rainfall over tropical Africa (e.g., Pfeifroth et al. 2015; Camberlin et al. 2019). There are, however, a number of known deficiencies (Huffman et al. 2007). Over land, TRMM generally underestimates the frequency and amount of rain from warm clouds, typically found over coastal areas with onshore trade or monsoonal winds and in the vicinity of mountains (e.g., Dinku et al. 2018). Another potential problem is an underestimation of extreme values, partly due to beam filling in the microwave bands (Young et al. 2014; Monsieurs et al. 2018). Over ocean, precipitation detection is more challenging than over land and calibration with gauges is not possible. The warm rain/drizzle problem is most severe over dry subtropical regions with extensive marine stratus.

c. Köppen–Geiger climates

For an assessment of forecast quality at a regional level, the tropics are divided into Köppen–Geiger climates by continents. This classification (Köppen 1900; Geiger 1961) uses five main climate zones and subgroups within each zone that are defined by seasonal precipitation patterns. Kottek et al. (2006) provide an updated Köppen–Geiger classification with a resolution of 0.25° × 0.25°, available at http://koeppen-geiger.vu-wien.ac.at/present.htm, that we coarsened to 1° × 1° to match the other datasets. We defined in total ten climatic regions with similar characteristics (see color shadings in Fig. 2a). For the most frequent main climates Equatorial (A) and Arid (B), we added continental labels: Arid North (N) Africa, Tropical Africa, Arid South (S) Africa, Arid Americas (mostly Mexico, eastern Brazil, and areas near the Andes), Tropical Americas, Arid Asia (mostly southwestern Asia and parts of India), Tropical Asia (including the Maritime Continent and northern Australia), and Arid Australia. All areas on different continents belonging to Warm Temperate (C) are labeled ''Temperate climates.'' The main climates Snow (D) and Polar (E) are found in only 6 and 91 grid boxes, respectively, in the high Andes and Himalayas, and are merged under the label ''Mountain climates.'' The number of grid boxes in each region are provided in Table 1.

3. Methods

For probabilistic forecasts, both the correctness of the probabilistic statement and its sharpness need to be evaluated. To measure the skill of a forecast, an adequate reference needs to be defined (section 3a). To measure the calibration of an ensemble system, probability integral transform (PIT) histograms and reliability diagrams will be used (section 3b). The actual evaluation then requires the application of proper scoring rules (section 3c). Finally the employed ensemble postprocessing method will be detailed (section 3d). In the results section, we will then analyze both raw and postprocessed forecasts side-by-side to bring out the benefit of statistical correction, which we deem useful for model developers, forecasters, and users of forecast products.

a. Reference forecasts

For a reference forecast, the concept of an extended probabilistic climatology (EPC) was applied following Vogel et al. (2018). For a given date, EPC generates an observation-based ensemble forecast by using all observations for this calendar date during 1998–2017, but without the considered year. Recently, Lang et al. (2020) compared various climatological methods and showed superiority of the EPC approach. For more robust statistics, we explored adding further days around the day of interest, testing window lengths of up to ±40 days with a step size of ±5 days. In terms of skill measured by the cross-verified continuous ranked probability score (CRPS; see section 3c), different window lengths for individual Köppen–Geiger climate regions do not deviate by more than 0.008 from a reference window length of ±20 days (Fig. 1). There is a

[1] On this web page, the MSC is denoted as Canadian Meteorological Centre (CMC).

TABLE 1. Fraction of grid boxes (%) with positive skill for 1-day precipitation forecasts in each climatic region and all regions combined. Individual columns show values for the three skill measures, BSS for occurrence of precipitation (threshold 0.2 mm), CRPSS, and BSS for extreme events (exceedance of 20 mm), for the two models ECMWF (2009–17) and MSC (2009–16), and for raw and postprocessed forecasts (left and right values, respectively). The numbers in parentheses after the region names give the total number of grid boxes in each region.

Region	BSS 0.2 mm		CRPSS		BSS 20 mm	
	ECMWF	MSC	ECMWF	MSC	ECMWF	MSC
Arid Americas (150)	35|81	45|82	58|91	61|84	34|67	47|60
Arid Asia (394)	58|94	21|95	71|98	46|92	46|59	55|52
Arid Australia (426)	98|100	99|100	100|100	100|100	83|96	87|95
Arid North Africa (1103)	68|91	46|84	46|91	17|65	22|29	20|16
Arid South Africa (244)	82|98	80|99	85|99	91|98	16|72	36|55
Mountain climate (97)	0|39	0|29	1|77	1|51	7|60	6|39
Tropical Africa (793)	24|74	46|60	19|97	46|84	17|82	24|65
Tropical Americas (1034)	44|93	67|90	69|100	87|99	51|99	63|94
Tropical Asia (529)	30|87	57|78	81|100	88|99	80|99	71|95
Temperate climate (817)	34|83	39|82	66|99	75|98	61|96	73|87
Combined (5587)	49|87	53|82	60|97	62|88	44|76	50|67

general tendency for arid climates to perform better for wider windows and for tropical climates to have lowest CRPS for narrower windows, possibly due to effects of seasonal changes such as monsoon onsets. Ultimately, we decided to use the optimal value within the ±5–40-day range for each grid box in order to maximize the skill of EPC over the tropical belt.

b. Calibration

PIT histograms and verification rank histograms are standard tools for the assessment of calibration. Hamill (2001), Gneiting et al. (2007), and Wilks (2019) provide further insights into their use and interpretation. To accommodate ensemble forecasts with different numbers of ensemble members, we use unified PIT (uPIT) histograms as in Vogel et al. (2018). The forecast distribution is divided into 20 bins of equal width such that each bin has a nominal value of 5%. This allows for a unified treatment of PIT and verification rank histograms. Calibrated probabilistic forecasts have uniform PIT histograms, while a U-shape (skew) indicates underdispersion (bias). The evaluation of probability of precipitation (PoP) or accumulation above a given threshold is based on reliability diagrams (e.g., Wilks 2019), where the observed frequency of occurrence is plotted against the forecast probability.

c. Proper scoring rules

To evaluate precipitation forecast quality we use proper scoring rules that assess calibration and sharpness simultaneously (Gneiting and Raftery 2007; Wilks 2019). We evaluate the quality of forecasts for the PoP by means of the Brier score (BS; Brier 1950). For a probabilistic precipitation forecast with cumulative distribution function (CDF) F and verifying observation y, the CRPS (Gneiting et al. 2007) is defined as

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

Here, 1 is an indicator function, equal to 1 if the argument is true and equal to 0 otherwise. The CRPS is negatively oriented, reported in the unit of the observation (here, millimeter). This way higher scores correspond to less skillful forecasts.

For comparative assessments, we rely on skill scores (i.e., the BSS and CRPSS) that indicate skill relative to a reference forecast (here EPC). Thereby a higher (lower) forecast quality of the investigated forecast compared to the reference forecast is indicated by positive (negative) skill. Equal predictive performance of both forecasts yields a skill of zero and the skill of a perfect forecast is one.

d. Statistical postprocessing

Statistical postprocessing corrects systematic deficiencies of NWP model output and allows to assess its true value (Vannitsem et al. 2018). In view of the typically small differences in predictive performance between different statistical postprocessing methods, we rely in the following on the well-established and computationally efficient method of ensemble model output statistics (EMOS; Gneiting et al. 2005) using generalized extreme value (GEV) distributions.

The idea of the EMOS GEV approach by Scheuerer (2014) is to convert an ensemble forecast into a parametric GEV distribution. The three-parameter GEV family of probability distributions allows a point mass for zero precipitation and flexible modeling in positive precipitation accumulations, depending on the specifics of the ensemble forecast at hand. For mathematical details we refer to the original paper by Scheuerer (2014).

Postprocessing techniques rely on statistical parameters that need to be estimated from training data, comprising forecast–observation-pairs from the TRMM pixel at hand and typically from a rolling training period consisting of the n most recent days for which data are available at the initialization time. We use a local neighborhood approach with $n = 500$ training days such that for each TRMM pixel its past 500 forecast–observation-pairs as well as the forecast–observation-pairs of eight adjacent grid boxes are used for training data composition. Near coasts the eight nearest grid boxes that belong to the land–sea class of

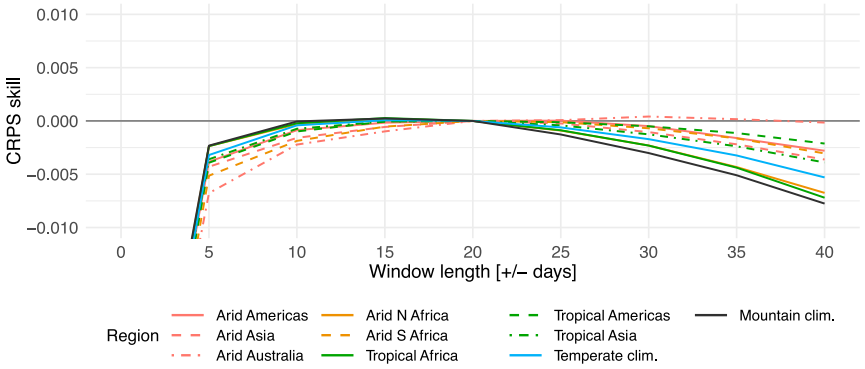

FIG. 1. CRPS skill of 1-day EPC-based precipitation forecasts using different window lengths (in steps of ±5 days) for different climatic regions (see Fig. 2 for the region definition). The reference length used here is 20 days. Results are shown for the entire study region 30°S–30°N and the TRMM data record 1998–2017.

the original point are considered. Parameter estimation is then based on CRPS minimization over the training data. All computations were performed in R (R Core Team 2018) and the details of the EMOS GEV implementation closely follow Vogel et al. (2018). The fact that we evaluate predictions postprocessed using TRMM data relative to TRMM-based EPC forecasts should lead to a somewhat optimistic skill estimate. However, Vogel et al. (2018) could show for northern tropical Africa that conclusions on forecast performance do not depend on whether satellite or station observations are used.

4. Results

The results section is organized as follows: The first three subsections discuss results for ECMWF only in all detail. The first of these concentrates on aspects of calibration and reliability with respect to the occurrence of precipitation on forecast day 1, while the following two address aspects of skill. As a threshold for the occurrence of precipitation, we use 0.2 mm irrespectively of the accumulation period, but our results change minimally under different thresholds up to 1 mm. The latter part will be broken down into aspects of rainfall occurrence, rainfall distribution, and extremes as well as into 1- versus 5-day accumulation periods. Results for raw ensemble output will be compared to postprocessed forecasts throughout. In the fourth subsection corresponding results for the MSC model will be presented and discussed relative to ECMWF. The fifth subsection present regional summary statistics for different skill measures, the two models, and raw and postprocessed forecasts. The final subsection will then address the question to what extent we can see improvements in the forecasting systems over time.

a. Calibration and reliability of the ECMWF ensemble

Concentrating first on land areas, Fig. 2a displays PIT histograms based on 1-day accumulated precipitation forecasts by the ECMWF EPS for the Köppen–Geiger climates and regions as introduced in section 2c. For all ten regions ECMWF raw ensemble forecasts are strongly underdispersive (or overconfident), as indicated by the tendency of the observation to rank lowest or highest compared to all ensemble members. Moreover there is a clear positive skew in the PIT histograms in all regions indicating that the observation frequently lies below the smallest ensemble member. This is mostly caused by a tendency of the model to produce light rain, when no precipitation occurs in reality. The fraction of such situations is indicated by the leftmost bin, which comprises between 30% (Arid Australia) and 48% (Arid Americas) of all forecasts. This value should be compared to the nominal value of 5% for a uniform distribution. There is no clear pattern in how this fraction is distributed geographically.

The miscalibration evident from the PIT histograms can be summarized in a single number, the so-called (scaled) discrepancy measure as defined by Berrocal et al. (2007). It attains values between zero and one, where lower values indicate better calibration. Figure 2b displays the spatial distribution of this measure for the ECMWF raw ensemble forecasts used for Fig. 2a. The results confirm that the ECMWF ensemble is not well calibrated anywhere in the tropics, but that calibration is even worse over many oceanic areas. The highest values are reached over the so-called oceanic deserts to the west of South America and southern Africa. These areas receive very little precipitation in reality and are dominated by persistent stratocumulus decks with occasional drizzle or light rain. TRMM is known to have large uncertainties in rain fraction (see Fig. 4c in Berg et al. 2010) and a comparatively large dry bias (Huffman et al. 2007) in these regions. The often light rain from warm clouds is generally challenging to detect from space (Young et al. 2018). This suggests that the calibration (and skill) of the model is presumably assessed worse in these regions than it actually is—particularly in recent years, as ECMWF has addressed relevant problems in their forecast model (Ahlgrimm and Forbes 2014). In contrast, the oceanic intertropical convergence zone (ITCZ) regions, which are dominated by frequent rainfall from deep convection, have much lower discrepancy values. Interestingly, such clear differences between moister and drier parts of the tropics are not seen over land. Several coastal areas stand out as having particularly low calibration (e.g., eastern Brazil, eastern Madagascar, central Mexico, eastern Kenya, and Tanzania). These are all regions characterized by moist onshore flow that often feeds warm

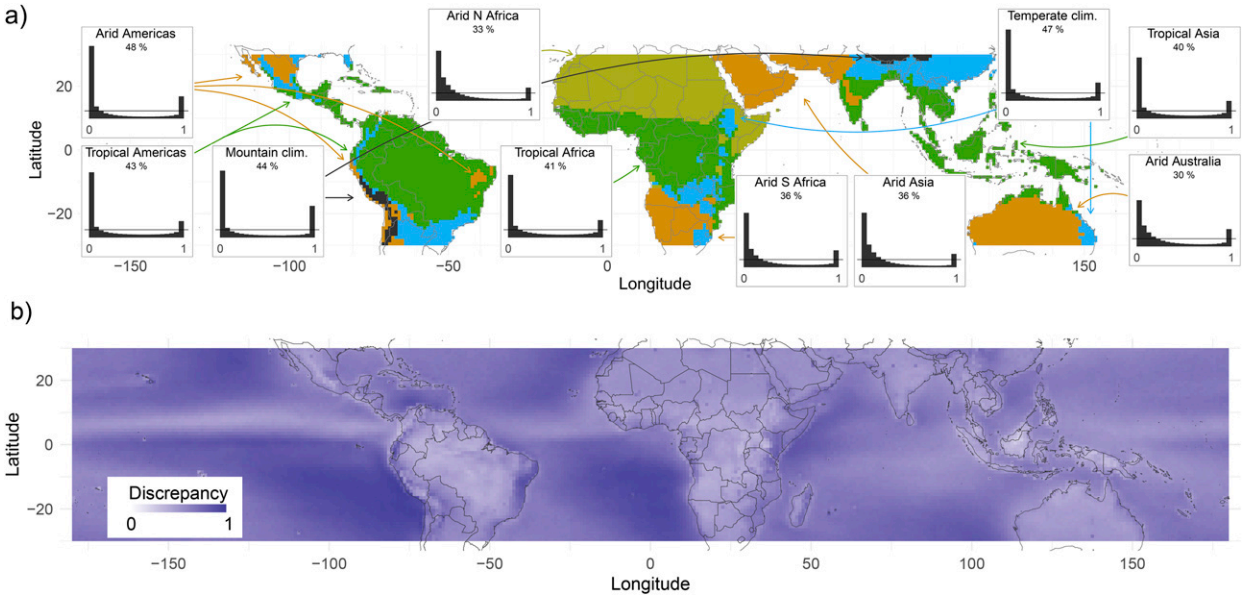

FIG. 2. Calibration of 1-day ECMWF raw ensemble forecasts for precipitation during 2009–17. (a) PIT histograms with 20 bins for the 10 Köppen–Geiger climates indicated with color shading and correspondingly colored arrows. The percentage of uPIT values in the leftmost bin is indicated and should be compared to the nominal value of 5% under calibration. (b) Spatial distribution of the discrepancy between ECMWF raw ensemble and calibrated forecasts as defined by Berrocal et al. (2007). Lower values indicate better calibration.

rain. It is well known that TRMM tends to have a negative bias in such conditions (see section 2b) that could explain the low calibration, while additional model errors can of course not be ruled out. The overall lack of calibration in ECMWF ensemble forecasts is robust across accumulation times from 1 to 5 days (not shown). However, there is a general tendency of wet (dry) areas to become better (worse) calibrated with longer accumulation periods. We speculate that this is related to a relatively high frequency of 5-day dry periods in observations, while the model tends to rain too often creating an even larger mismatch.

Figure 3a displays reliability diagrams for the 1-day occurrence of precipitation as forecast by EPC. As expected, the reliability of such a forecast is close to perfect in all climatic regions. The gray bars at the bottom of the individual diagrams show the climatological distribution of PoP. All arid regions favor low values with some variability between the very dry Sahara and slightly moister Arid S Africa for example. The moister regions either feature a unique mode at high values such as the Tropical Americas or show a bimodal distribution such as Tropical Africa that reflects seasonal shifts (also evident in Mountain climates). The Temperate climate zone stands out as having an almost uniform distribution.

A comparable analysis for forecasts by the ECMWF raw ensemble (Fig. 3b) reveals a clear tendency in all regions to predominantly issue forecasts close to 0% and 100% PoP, while all other categories are sparsely populated. The share between the two extreme categories appears to be related mostly to the overall climatology (i.e., arid regions being dominated by dry forecasts). Such behavior has also been described by Medina et al. (2019) for subregions in Brazil. In Mountain climates many forecasts with low rain probability do have precipitation in reality (about a quarter). This discrepancy indicates potential problems in both observations and model forecasts. Clearly, the model resolution is not sufficient to represent the many orographic effects that can trigger convection (e.g., elevated heating, rotors, mountain waves, barrier winds, see review by Houze 2012). This often results in lower predictive performance as analyzed, for example, by Richard et al. (2007) for the European Alps. On the other hand, it has been shown that TRMM performs relatively poor in the detection of precipitation over mountainous terrain (Barros et al. 2006; Hirpa et al. 2010; Maggioni et al. 2016), prompting more caution in the interpretation of the results for this category. Despite these regional differences in low-probability predictions, forecasts of very high probabilities for rainfall occurrence generally verify in only about 60%–90% of cases depending on the region. This underlines that the forecasts are overall highly overconfident.

After statistical postprocessing ECMWF forecasts are much better calibrated, but with a light tendency to be underconfident leading to small deviations from a uniform distribution in the PIT histogram (not shown). Figure 3c shows the improvement in reliability. In all regions, much larger parts of the probability space reflected by the gray bars are populated, indicating a lower resolution of postprocessed forecasts. The PoP distribution is now much closer to that of EPC (Fig. 3a), although a number of smaller deviations remain. For example in several arid regions, too many low probability forecasts still occur. The highest rainfall probability category is so sparsely populuated in some regions that sampling

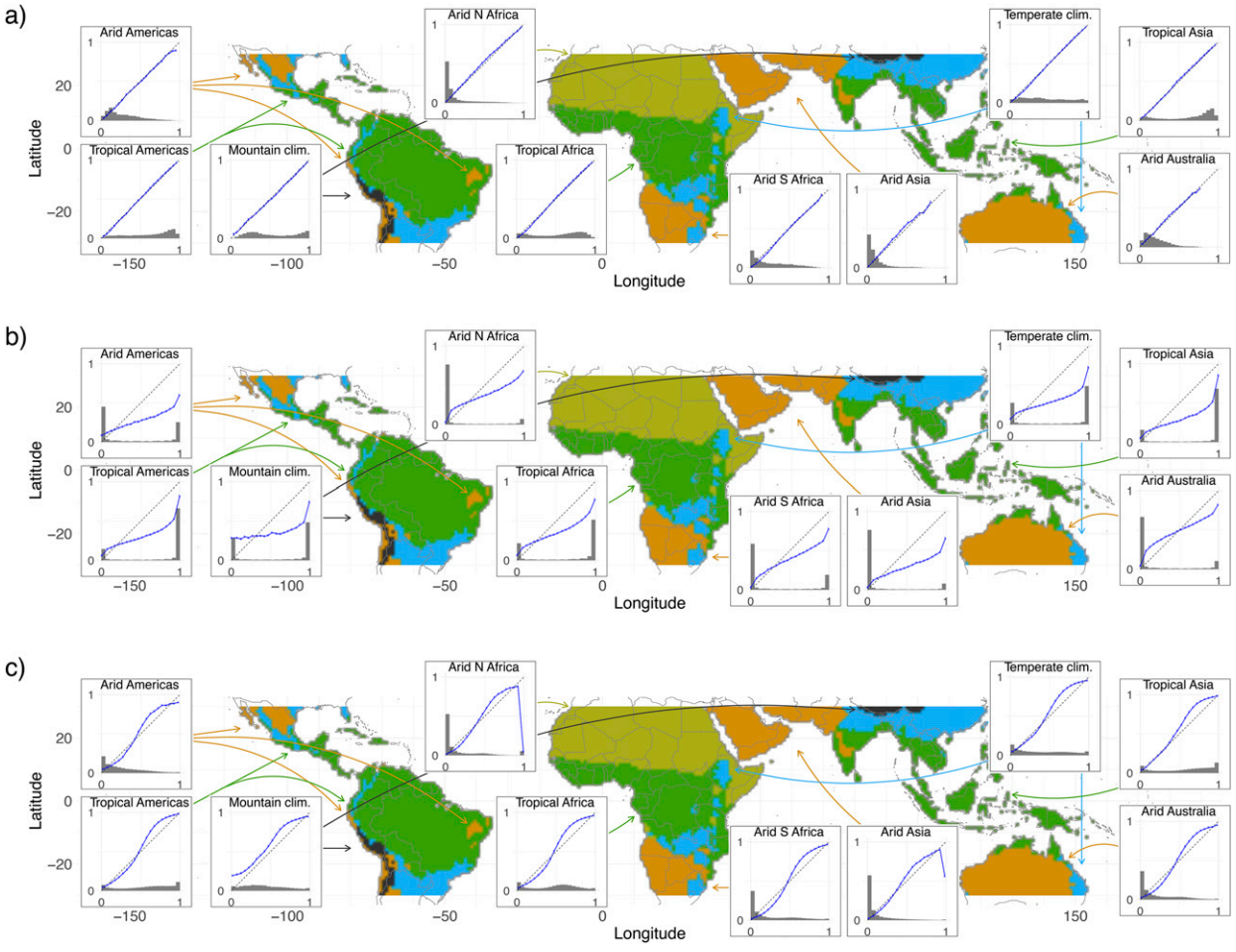

FIG. 3. Reliability diagrams for 1-day (a) EPC and ECMWF, (b) raw, and (c) postprocessed ensemble forecasts for occurrence of precipitation (threshold 0.2 mm) during 2009–17.

problems lead to nonmonotonicities (e.g., Arid North Africa and Asia). The reliability curves are now much closer to the diagonal in all regions despite a general tendency to be underforecasting for low-probability forecasts and overforecasting for high-probability forecasts.

b. Skill of the ECMWF ensemble

Figure 4 displays the BSS of ECMWF raw 1-day ensemble forecasts for the occurrence of precipitation relative to EPC. We see a clear contrast between land areas and oceans. Over the latter, BSS is neutral to negative almost everywhere. Skillful forecasts are only found right next to land areas (e.g., off the coast of northwestern Australia, Persian Gulf). Areas with neutral skill over open oceans are predominantly moist regions such as the ITCZ over the equatorial Pacific and Atlantic Oceans as well as the warm water areas around the Maritime Continent and over the equatorial Indian Ocean. Interestingly, skill is also enhanced over the South Pacific and Atlantic Convergence Zones and just west of the northern African and Central American landmasses. The latter may be related to continental convective complexes moving out to the ocean and to forcing from the extratropics during the transition seasons (e.g., Kiladis and Weickmann 1997; Knippertz 2007). To first order, the pattern over the ocean has some resemblance with the discrepancy measure shown in Fig. 2b, indicating that a lack of calibration explains at least part of the poor skill over the ocean. Forecasts over land are generally more skillful but regional contrasts can be very large. There is a general tendency for higher skill in the relatively drier outer tropics away from largest mountain chains (Australia, southern Brazil, the Sahara, southwestern Africa, southwestern Asia). Skill is often negative in coastal and mountainous regions in the inner tropics (e.g., Andes, Central America, eastern Brazil, western Central Africa, lowlands in East Africa, eastern Madagascar, Himalayas). It appears that the local topographic features responsible for stratiform rainfall generation or the triggering of convection (and possibly its organization) are much better represented by EPC than by the dynamical forecast model in these areas. However, as already pointed out in the previous subsection, skill may be underestimated in some coastal areas with moist onshore flow due to issues of warm rain detection by TRMM.

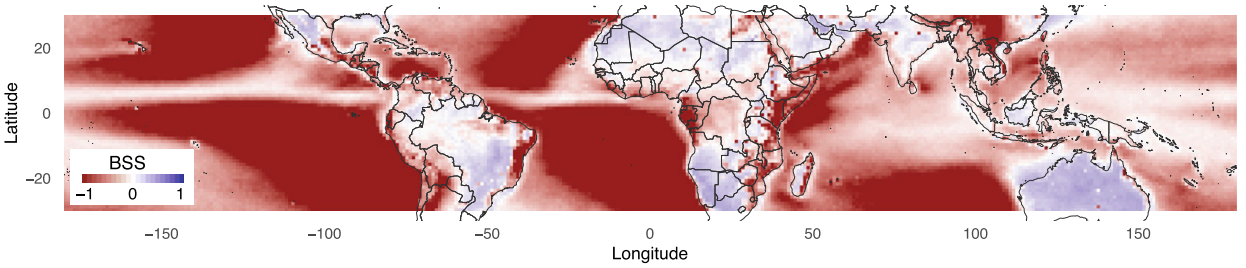

FIG. 4. Brier skill score (BSS) for 1-day ECMWF raw ensemble forecasts for occurrence of precipitation (threshold 0.2 mm) relative to EPC during 2009–17.

For the probability of 5-day accumulated rainfall above 0.2 mm (not shown), the BSS generally increases, most likely as timing errors become less relevant. Over many parts of the oceans the negative skill turns to neutral or only slightly negative values. Over land, regions of negative skill also tend to contract and become mostly confined to complex terrain (South America and Asia) and coastal areas (tropical West Africa). Along the coast of East Africa, however, the region of negative skill expands inland, indicating more fundamental problems beyond timing issues. This includes the dry bias of TRMM in areas with warm rain as discussed in previous sections.

While the BS (and BSS) assesses the probability of occurrence of accumulated precipitation above given thresholds, the CRPS (and CRPSS) allow evaluating forecast quality with respect to the full probabilistic forecast distribution. Figure 5a displays the mean CRPSS of raw ECMWF ensemble forecasts for 1-day accumulated precipitation and the period 2009–17 relative to EPC. The overall pattern has some similarities with the BSS shown in Fig. 4 but overall the skill is considerably higher. This suggests that the ECMWF model struggles particularly to discriminate between rain-no rain situations, while the forecast of the rainfall amount is more reasonable in many parts of the tropics. Areas with consistently poor forecast performance are the oceanic deserts over the southern (and to a much lesser extent the northern) Pacific and Atlantic as well as the Himalayas and Andes. Many other parts of the oceanic and terrestrial tropics show weakly positive skill including densely populated areas such as India, Australasia, and eastern Brazil. The striking exception is tropical Africa, which is characterized by consistently negative skill, apart from highlands in eastern and southern Africa. The affected areas are known to have large contributions from intense convective systems organized at the mesoscale (see Fig. 11 in Nesbitt et al. 2006) and it is known that convective parameterizations struggle to realistically represent this process, leading to forecasts with too much light and too little intense rainfall overall (Stephens et al. 2010; Marsham et al. 2013; Pearson et al. 2014; Birch et al. 2014; Pantillon et al. 2015). A similar conclusion was already drawn by Vogel et al. (2018) for the poor performance of the TIGGE models over northern tropical Africa.

Postprocessing is capable of eliminating areas of negative skill almost entirely (Fig. 5b; the remaining grid points over Egypt have little practical relevance). A large fraction of tropical land and ocean now shows moderately positive skill. Even the highly problematic oceanic deserts and high mountain regions (Andes, Himalayas) reach at least neutral skill after postprocessing. The striking counterexample is tropical Africa. Despite only moderately negative skill in the raw forecasts, postprocessing can only achieve neutral skill here. This suggests that the discrepancy between the more frequent and lighter rain generally produced by convection schemes and the very concentrated, long-lived and intense mesoscale convective systems in reality is too large to be cured by a relatively simple statistical correction. The singularity of tropical Africa is also seen in Fig. 2a by Wheeler et al. (2017) using anomaly correlations applied to the ECMWF ensemble mean. The BSS distribution for postprocessed ECMWF PoP forecasts is almost identical to the CRPSS (Fig. 5b) and is therefore not shown here. This implies that for both rainfall occurrence and amount, EPC is currently the best (and easiest to use) probabilistic forecast information we can provide for large parts of tropical Africa.

How does predictive performance for rainfall amount change for longer accumulation periods? Fig. 5c shows a CRPSS distribution analogous to Fig. 5a but for 5-day accumulated precipitation forecasts. This demonstrates that most of the forecast performance is inherited from errors for 1-day predictions. Postprocessing can again improve forecasts practically everywhere (Fig. 5d) but some regions (oceanic deserts, northern tropical Africa) show a further deterioration compared to 1-day forecasts (Fig. 5b). Overall this indicates that the misrepresentation of local conditions important for rainfall generations dominate forecast behavior, while effects of decreasing predictability for longer lead times or smoothing by longer accumulation periods appear to have comparably little influence.

c. *Skill of the ECMWF ensemble for extreme rainfall events*

An important aspect of precipitation forecasts is their ability to predict extreme events, as this allows for precautionary action to be taken. Exemplarily, Webster et al. (2011) report on extreme rainfall events in Pakistan in 2010, which were embedded in the Indian monsoon during a period of anomalous large-scale flow. These were predicted by the ECMWF model with high probabilities 6–8 days ahead. However, not all extreme precipitation events are connected to well-predictable and large-scale features, and it is unclear if and where models are able to predict extreme precipitation reliably. Sampling

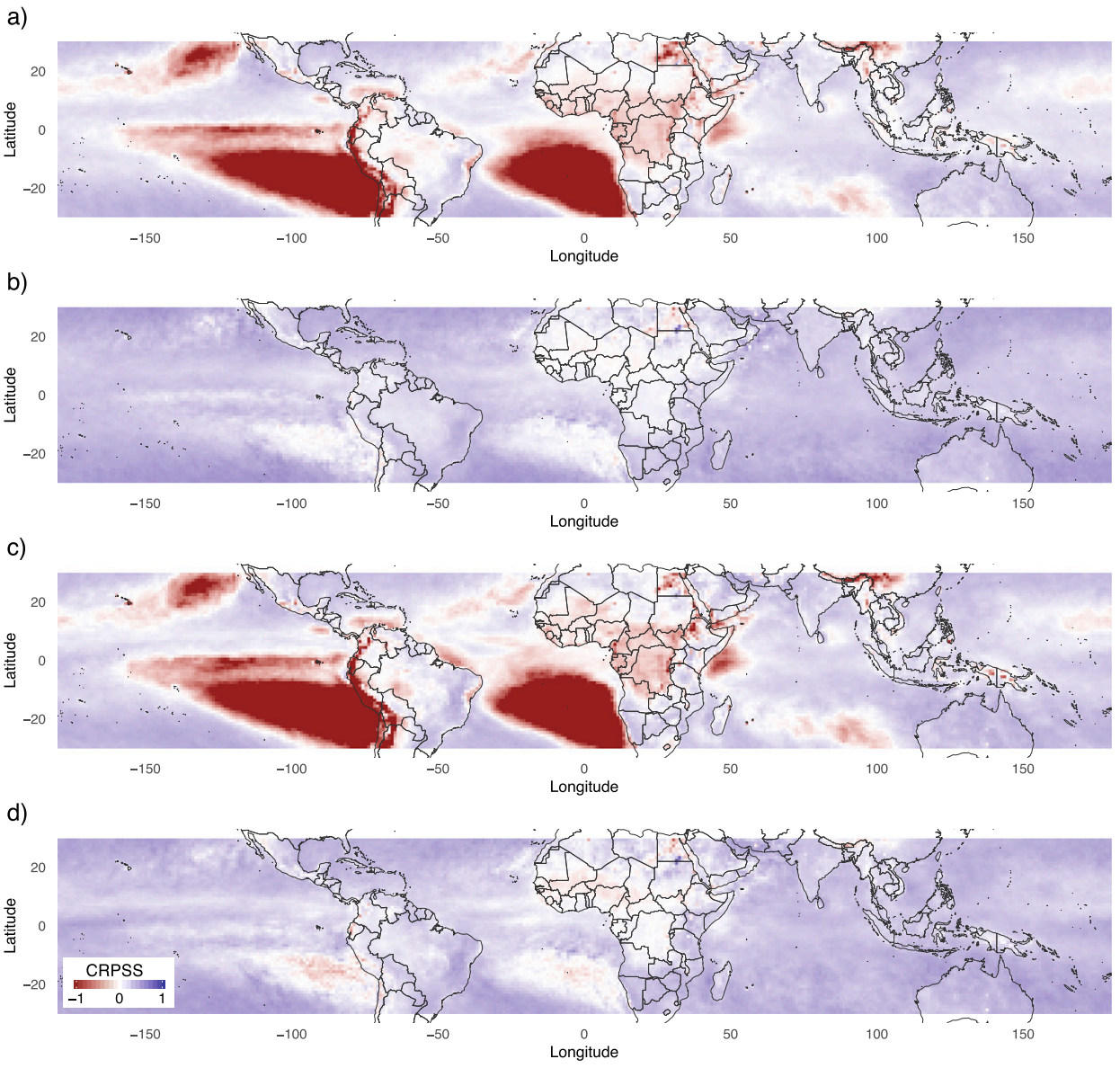

FIG. 5. Continuous ranked probability skill score (CRPSS) for 1-day ECMWF (a) raw and (b) postprocessed forecasts for accumulated precipitation amount relative to EPC during 2009–17. (c),(d) As in (a) and (b), but for 5-day forecasts.

uncertainty typically impedes our ability to analyze forecast skill for the most extreme cases (Lerch et al. 2017). A common compromise is to increase the number of events by using thresholds low enough to give robust statistics for a given time series length. Here, we use 20 mm within 1 day and 50 mm within 5 days as thresholds for the occurrence of extreme events and exclude grid boxes where the considered event occurs with a frequency of less than 1%, or about 33 events in 2009–17. We only display results for continents, where impacts are most important. The evaluation is based on the BSS as in Fig. 4. For 1-day events and raw ECMWF ensemble forecasts (Fig. 6a), positive skill with values of up to 0.3 is found for most of tropical Australasia and with local exceptions over higher ground (Himalayas, Papua New Guinea). Central and South America show a more mixed result with lower skill over mountainous areas (e.g., Andes) and higher skill over eastern Brazil and Mexico. As already found for the CRPSS (Fig. 5a), tropical Africa to the west of the East African highlands stands out as a region of particularly low predictive performance.

Postprocessing improves skill almost everywhere and eliminates areas of negative BSS (Fig. 6b). However, while areas with negative skill in the raw ensemble can be turned to positive skill through postprocessing in the Americas, over the Maritime Continent, and in Asia, tropical Africa only reaches BSS values around zero. This general pattern is

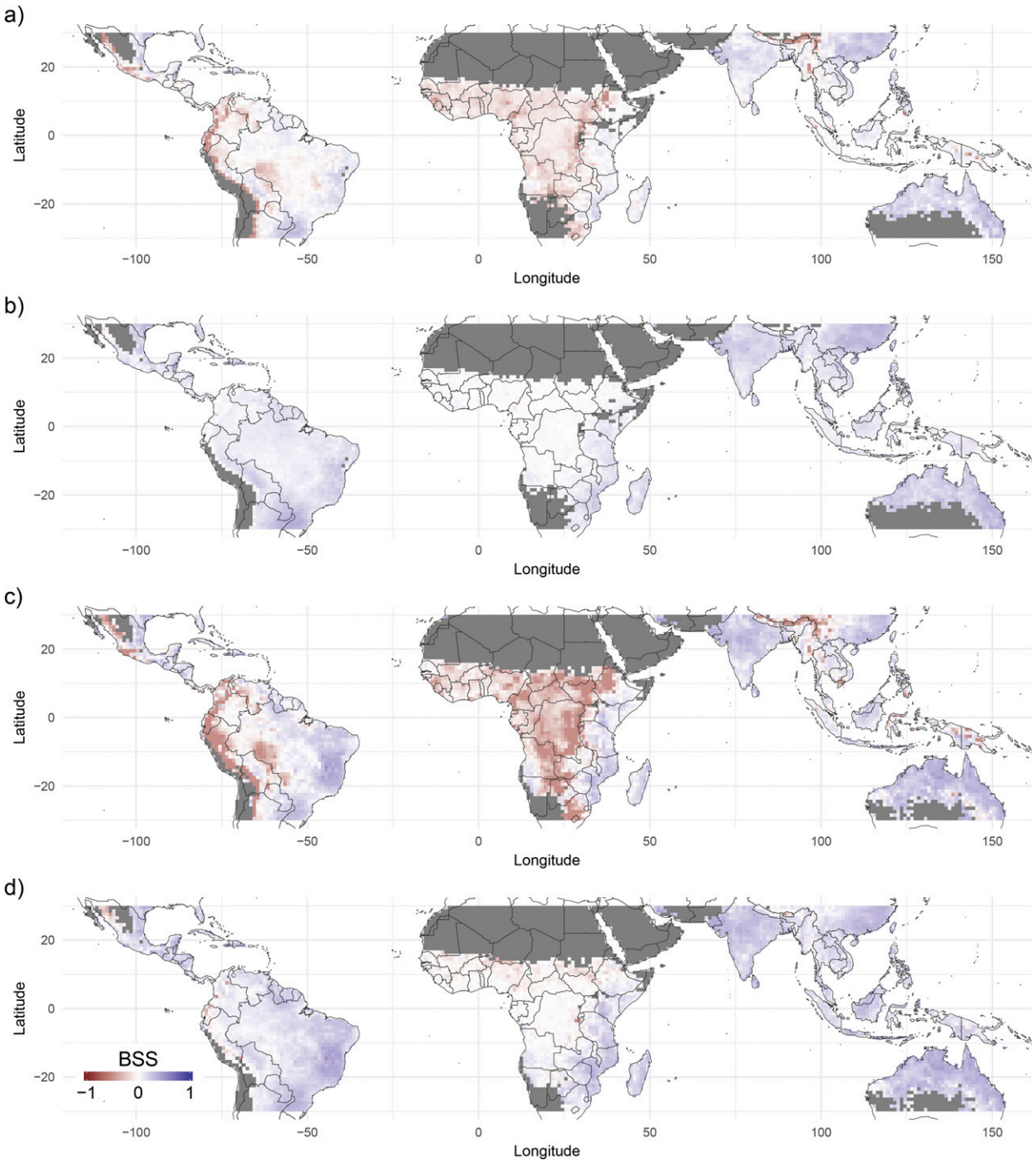

FIG. 6. BSS for ECMWF (a),(c) raw and (b),(d) postprocessed ensemble forecasts for the exceedance of 20 mm within 1 day in (a) and (b) and the exceedance of 50 mm within 5 days in (c) and (d) during 2009–17. Displayed is skill only over land and where the considered event has an occurrence frequency above 1%.

robust for accumulations of 5 days and a threshold of 50 mm (Figs. 6c,d). In the raw forecasts, signals generally tend to amplify over 5 days (i.e., both positive and negative values increase in magnitude) (Fig. 6c). Postprocessing (Fig. 6d) can again correct for the bulk of deficiencies but performance remains slightly worse for 5 days than for 1 in areas with low skill, while areas with good skill are similar or even better for the longer accumulation period. This is largely consistent with the analysis of precipitation amount shown in Fig. 5.

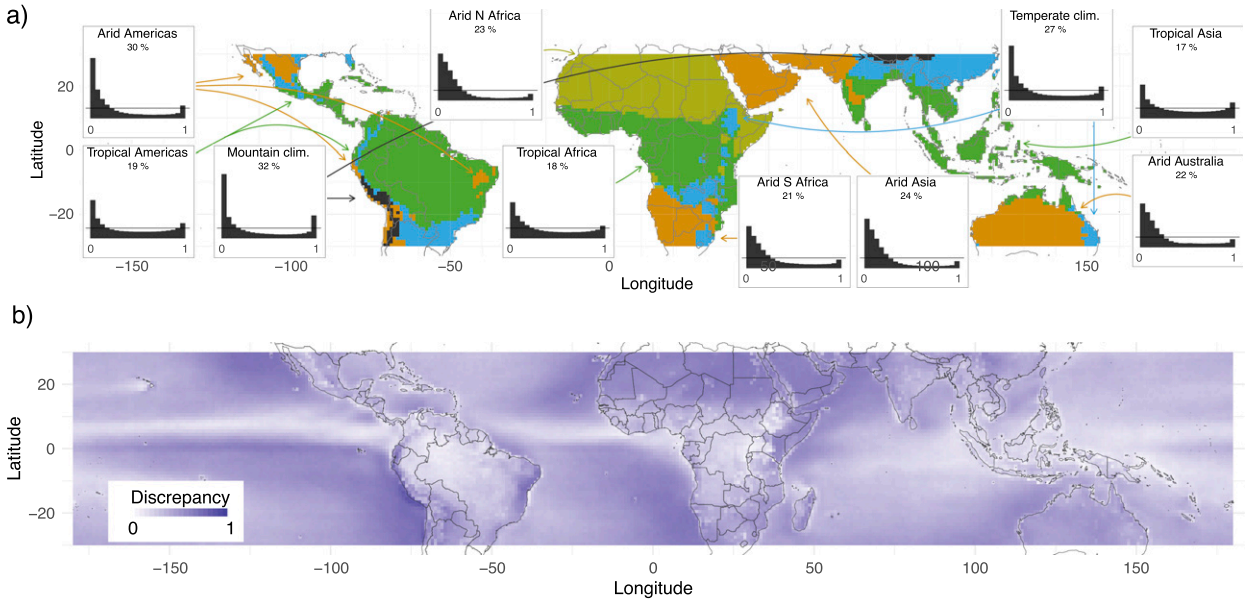

FIG. 7. As in Fig. 2, but for the MSC model and during 2009–16.

d. Comparison to the MSC ensemble

In this section, the raw and postprocessed MSC ensemble forecasts will be analyzed and compared to the ECMWF results discussed in the previous section. As pointed out in section 2, MSC data are only available for 2009–16. When ECMWF results are restricted to this period (not shown), changes are minimal and therefore we conclude that this discrepancy is an unlikely explanation for the relatively large and systematic differences between the two systems. Figure 7 displays calibration of the MSC ensemble in the same way as Fig. 2. Although the MSC raw ensemble is also not well calibrated, it shows a more uniform distribution than ECMWF. PIT histograms are slightly right skewed in all regions, implying observations to rank lowest too often. The leftmost bin contains between 17% (Tropical Asia) and 32% (Mountain climates) of all forecasts.

The geographical distribution of the discrepancy measure for the MSC ensemble in Fig. 7b has structural similarities with that for ECMWF ensemble forecasts (Fig. 2b). It reveals good calibration over large parts of the Indian and western Pacific Oceans as well as along the Atlantic and eastern Pacific ITCZ. The oceanic deserts stand out as areas of large discrepancy but not as much as in ECMWF. As indicated by the PIT histograms, calibration is also better in MSC than in ECMWF over land in tropical Africa and northwestern South America.

The reliability of the MSC raw ensemble is investigated in Fig. 8a. As for the ECMWF ensemble (Fig. 3a), there is a general tendency to frequently forecast rainfall with probabilities of near zero and near one. Apart again from the Mountain climates region, many low-probability forecasts do in fact verify, while forecasts with higher probabilities overestimate the occurrence of rainfall, indicating an overconfident system. The best reliability is found for Arid Australia, Tropical Asia, Tropical Africa, and the Tropical Americas. Overall, the MSC raw ensemble is more reliable than the ECMWF raw ensemble for most regions.

The spatial distribution of the BSS for the MSC raw ensemble is displayed in Fig. 8b. It has many similarities with the corresponding pattern for ECMWF (Fig. 4) but the skill is overall higher. This is particularly true for the moister oceanic regions but also for some land areas such as South America. Lowest skill is again found over oceanic deserts and in mountainous terrain. This demonstrates that the better calibration in MSC forecasts does in fact lead to more reliable forecasts than in the ECMWF EPS, at least when raw model output is considered.

Now focusing on the entire rainfall distribution, Fig. 9a shows horizontal maps of CRPSS for raw MSC ensemble forecasts for 1-day accumulated precipitation. Again, the overall pattern is similar to that of the ECMWF ensemble (Fig. 5a). Regional differences are found for tropical Africa, where the MSC raw ensemble has neutral instead of negative skill, for South America, where negative skill is restricted to the Andes region in the MSC raw ensemble, and for arid northern Africa, where the MSC ensemble performs worse than the ECMWF raw ensemble (and EPC).

Postprocessing increases skill almost everywhere as expected, but parts of northern Africa still have negative skill (Fig. 9b). A direct comparison to the corresponding ECMWF forecasts (Fig. 9c) reveals that the better skill in the MSC raw ensemble does not necessarily yield better postprocessed predictions. Over most parts of the tropics, ECMWF shows superior or equal performance. Exceptions are merely the dry oceanic areas off the coast of Peru and off the coast of Namibia and Angola, as well as over dry northeastern Africa and the Arabian Peninsula. A more detailed look reveals that the postprocessing leads to a calibration and reliability similar to ECMWF, while the resolution is slightly lower in MSC

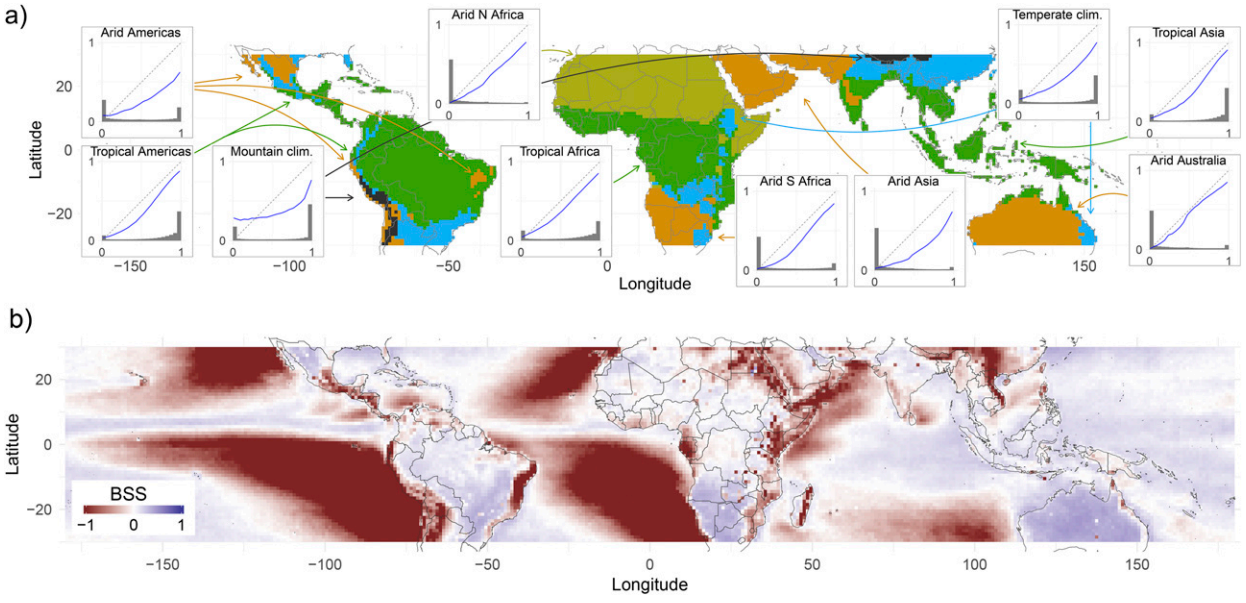

FIG. 8. As in Figs. 3b and 4, respectively, but for the MSC model and during 2009–16.

(not shown). This suggests that while the MSC raw ensemble is better calibrated and more reliable than the ECMWF raw ensemble in many regions, as reflected in higher BSS and CRPSS, it does not necessarily contain superior predictive information. Interestingly, MSC raw and postprocessed ensemble forecasts for extreme rainfall are slightly worse than their ECMWF counterparts (Fig. 6) but with a similar spatial distribution (not shown). Recently, Stellingwerf et al. (2020) evaluated the ECMWF and MSC EPSs specifically over Ethiopia and found ECMWF to be the best individual model after bias correction, while MSC shows the most realistic ensemble spread. Over Ethiopia, MSC also performs best with respect to extremes, which we cannot confirm for the larger tropical area.

e. *Summary statistics*

For a better overview of the results discussed so far, Table 1 presents summary skill statistics for all regions, both models, and both raw and postprocessed forecasts. The numbers given are the percentages of grid boxes with skill larger than zero. The main take-home messages from this analysis are:

- Postprocessing improves forecasts in all cases but for BSS 20 mm in Arid Asia and Arid North Africa, where such extremes are rare, leading to poor statistical robustness. For all regions combined (bottom row in Table 1), fractions of positive skill increase from 44%–62% for raw forecasts to 67%–97% after postprocessing.
- ECMWF performs worse than MSC in 19 out of 30 cases for raw forecasts but only in 3 cases after postprocessing, confirming the previously discussed aspect of worse calibration but better predictive potential. This is also reflected in the statistics for all regions combined.
- For the majority of regions, performance measured in CRPSS is better than for the other two metrics, both before and after postprocessing. Forecasts of extremes are often only marginally worse than forecasts of precipitation occurrence.
- The region with the overall best performance is Arid Australia, where good skill is achieved already for the raw forecasts.
- Mountain climates stand out as the region with poorest performance, where skill remains relatively low even after postprocessing. This region comprises of only 97 grid boxes with extreme conditions that challenge both models and observations as discussed above.
- Another region of note is Tropical Africa, where particularly raw ECMWF forecasts perform poorly in all three skill measures. It is likely that the high degree of convective organization found here contributes considerably to this poor performance, as it leads to a strong concentration of rainfall into few intense events. Given that this issue is rather systematic, postprocessing is capable of curing some of the deficiencies. For example, for ECMWF the positive fraction for CRPSS increases from 19% to 97% but most of this skill is still only marginally above zero (see Fig. 5b). Similar problems are also seen in Arid North Africa, where organized convection occurs during the summer rainy season. Given their sizes of more than 1000 grid boxes, the two African regions contribute significantly to the combined fractions at the bottom of Table 1.

f. *Improvement over time*

In previous subsections, the ability of ECMWF and MSC raw and postprocessed ensemble forecasts to predict rainfall amount, occurrence, and extreme events was assessed with respect to the regional and spatial distribution based on the mean skill. Here we examine whether the model skill has improved over the investigation period (2009–17 for ECMWF and 2009–16 for MSC) due to, for example, higher resolution or better model physics and data assimilation.

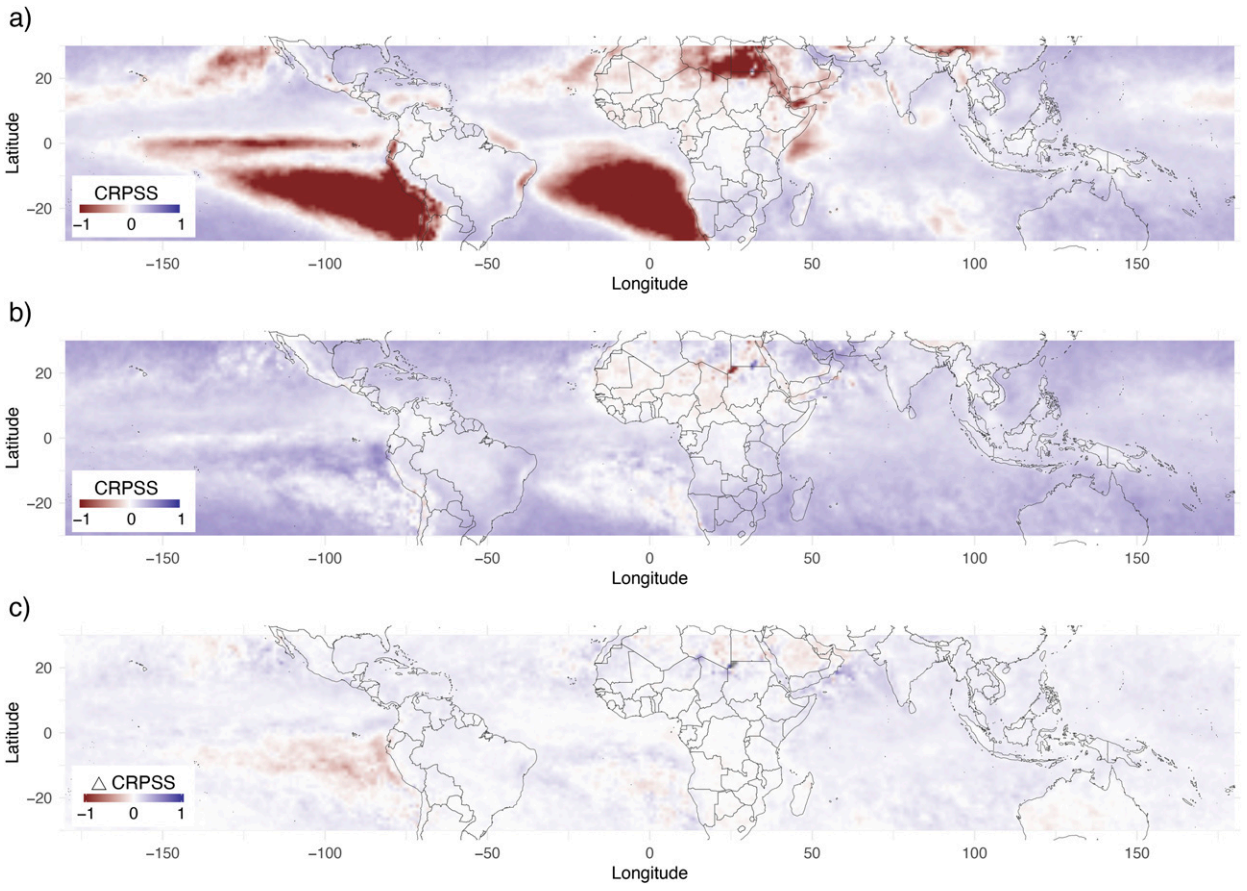

FIG. 9. (a),(b) As in Figs. 5a and 5b, respectively, but for the MSC model and during 2009–16. (c) Differences between postprocessed forecasts from ECMWF (2009–17) and MCS (2009–16) (i.e., between Fig. 5b and Fig. 9b).

Figure 10a displays the temporal evolution of CRPSS for raw ECMWF forecasts for 1-day accumulated precipitation in each Köppen–Geiger climate region (see, e.g., Fig. 2a). In 2009, all regions except for Arid Asia and Arid Australia have negative skill. Performance is worst for Mountain climates, followed by Tropical Africa, Arid Americas, and Arid North Africa. From 2009 to 2010, forecast skill increases markedly in most climates. For some regions, the increase in skill continues until 2011, with most showing positive skill by then. After 2011, no clear change in CRPS skill is detectable anymore. We hypothesize that the improvement early in the time series is related to the increase in horizontal resolution introduced on 26 January 2010, when grid spacing was reduced from 25 to 16 km for the HRES run and from 50 to 32 km for the CNT and ENS runs (Miller et al. 2010), and the introduction of the 5-species prognostic microphysics scheme on 9 November 2010 (Forbes et al. 2011). For the period 2011–17, nonparametric trend tests do not detect change at the 5% level except for tropical Africa (improvement) and Mountain climates (deterioration), suggesting that all changes to the ensemble system introduced after 2010 have little effect on the metrics used here to assess tropical rainfall forecasting.

Figure 10b shows the corresponding time series for postprocessed forecasts. Skill is now positive for all regions and all years varying between 0 and 0.3 with an overall more gradual increase in most regions, indicating that the postprocessing is able to remedy some of the negative impacts of lower resolution before 2010. There is a marked gap between the problematic Mountain climates, Arid North Africa, and Tropical Africa on one hand, and the other climate zones on the other hand. Arid Australia and Americas show a large increase in postprocessed skill with the resolution change from 2009 to 2010 but skill stays roughly constant after that or even deteriorates in the case of Arid Australia. For Tropical Africa and Tropical Americas, in contrast, postprocessed forecast skill improves significantly (on the 5% level) after the resolution increase (i.e., over 2011–17). Over the entire nine years the increase in CRPSS in the three tropical regions is on the order of 0.05. CRPSS for Mountain climates increases by about 0.06, starting at almost zero in 2009, with a significant improvement during 2011–17.

After postprocessing, the Mountain climates region continues to show the worst performance of all Köppen–Geiger climates, while Arid Australia performs best as already seen for the raw forecasts. This indicates that the predictive information contained in the raw forecasts sets limits to what postprocessing can achieve. However, a different behavior is observed for Arid Northern and Tropical Africa, which

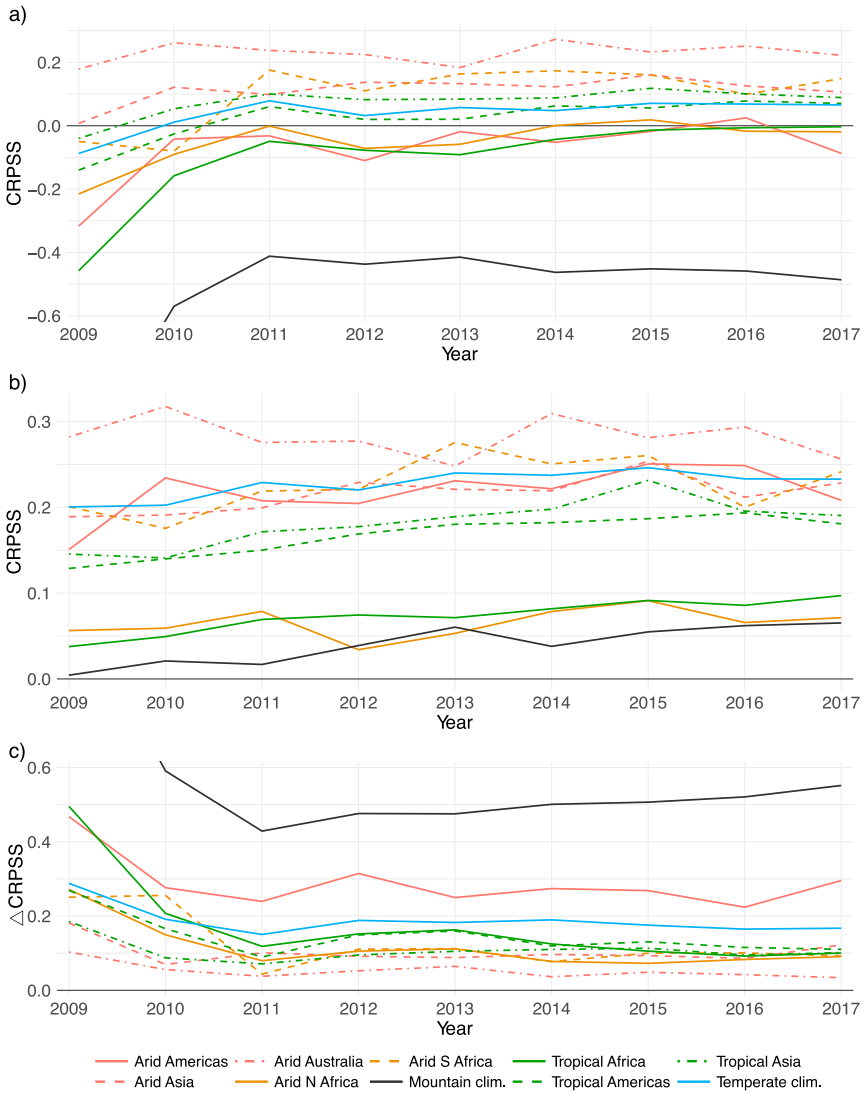

FIG. 10. Temporal evolution of CRPSS for (a) raw and (b) postprocessed ECWMF 1-day ensemble forecasts for accumulated precipitation relative to EPC during 2009–17. (c) The corresponding temporal evolution of the gap in skill between postprocessed and raw forecasts.

reveal a raw ensemble skill similar to Arid Americas but have relatively much smaller skill after postprocessing. This indicates that forecasts for Arid Americas contain more predictive information, although the raw ensemble forecasts for the three regions have similar levels of miscalibration. With less convective organization in Arid Americas (see, e.g., Nesbitt et al. 2006), this is a further indication of deficiencies in the representation of highly organized convective systems.

These regional differences are also evident in the temporal evolution of the skill gap, the difference in CRPSS between raw and postprocessed forecasts (Fig. 10c). It shows a clear narrowing in all regions from 2009 to 2011, when it decreases for the majority of regions from a CRPSS difference between 0.1 and 0.3 down to 0.05–0.15. After 2011, however, the gap in skill remains about constant in most regions except for a significant increase (on the 5% level) for Mountain climates. This behavior is consistent with the rather constant improvement by postprocessing found by Hemri et al. (2014), who verified ECMWF forecasts of temperature and 1-day accumulated precipitation against WMO station observations worldwide.

For the MSC model (2009–16 only), raw ensemble forecasts have neutral or slightly positive CRPSS in all regions and for all years (i.e., already in 2009 in contrast to ECMWF), except for Mountain climates and to a much lesser degree Arid Americas (Fig. 11a). The former and all tropical climates (Asia, Americas, Africa) show a significant positive trend. In agreement with ECMWF the consistently best performance is seen for Arid Australia. After postprocessing, most regions show consistently positive skill but problems are still evident

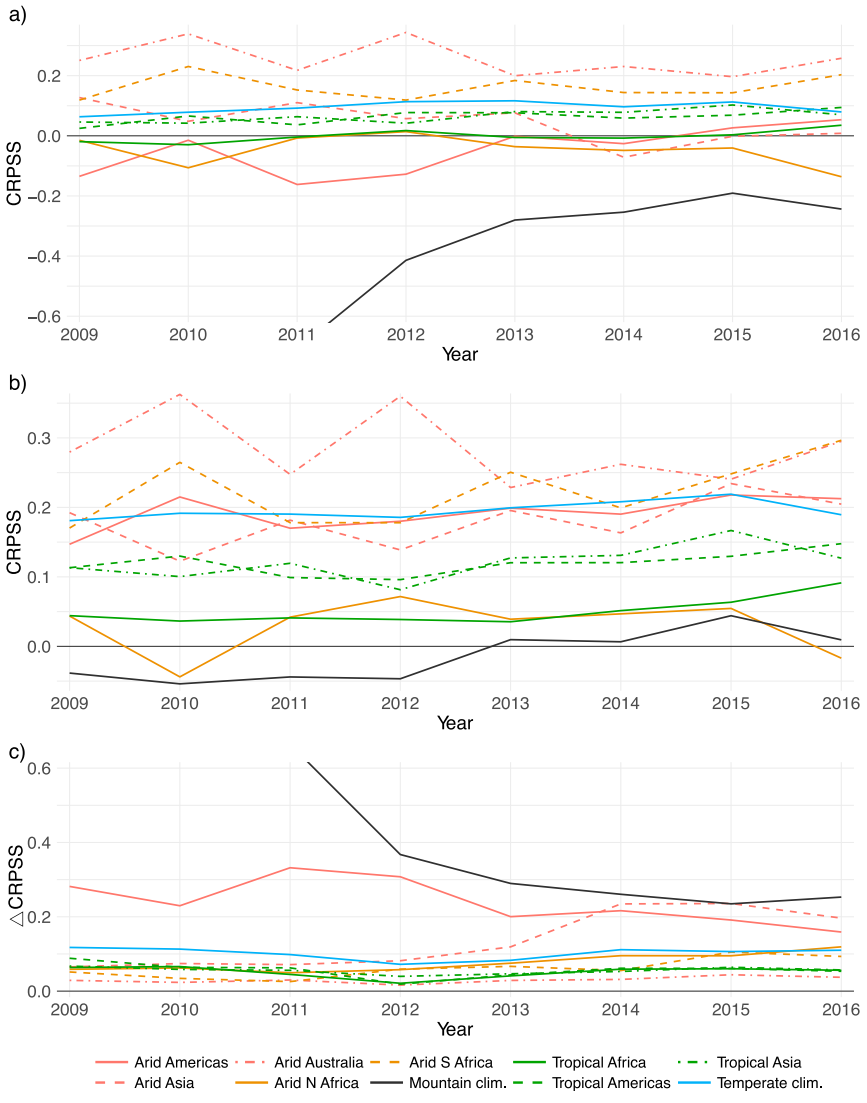

FIG. 11. As in Fig. 10, but for the MSC model and during 2009–16.

in Mountain climates, Arid N and Tropical Africa (Fig. 11b), consistent with the ECMWF results. Significant positive trends are now restricted to Tropical Africa and Mountain climates. As discussed in the previous subsection, raw MSC ensemble forecasts are slightly more skillful than ECMWF in most regions but postprocessing largely reverses this order (cf. Fig. 10 with Fig. 11). Due to the overall better calibration, the skill gap is consistently smaller for the MSC in most regions throughout the entire period (Fig. 11c). As for ECMWF, skill gaps in MSC are largest for Arid Americas and Mountain climates. Somewhat surprisingly, the skill gap grows markedly in 2014 for Arid Asia, leading to an overall significant positive trend.

5. Conclusions

The quality of precipitation forecasts from two leading operational ensemble predictions systems (ECMWF and MSC) was assessed specifically for the tropics between 30°S and 30°N. TRMM satellite estimates were used as an observational reference. Predictions were evaluated for accumulation periods of 1 and 5 days with respect to occurrence and amount of precipitation as well as the occurrence of extreme rainfall relative to a probabilistic reference forecast based on climatology (termed EPC). The potential of statistical ensemble postprocessing to correct for biases and dispersion errors in the ensemble forecasts was investigated. Performance over land is summarized for specific Köppen–Geiger climatic regions.

The main results are as follows:

- Forecasts of precipitation occurrence (using a threshold of 0.2 mm): Both the ECMWF and MSC models do not perform better than the climatological reference over about half of all land points (Table 1) and over relatively dry oceanic regions. This is due to substantial calibration problems and biases,

particularly a strong overconfidence for high-probability forecasts. Postprocessing increases forecast performance significantly, with the fraction of land grid boxes with positive skill rising to 87% (ECMWF) and 82% (MSC). This demonstrates that postprocessing brings out the enhanced predictive information in the ECMWF forecasts despite even larger calibration problems than MSC.

- Forecasts of precipitation amount: There is moderate skill in many parts of the tropics in the raw ensemble forecasts (about 60% of all land points have positive skill, see Table 1). Most problematic regions are the oceanic deserts (particularly over the South Atlantic and Southeast Pacific), high mountain ranges (particularly the Andes and Himalayas), and the west and central lowlands of tropical Africa. Over Tropical Africa as a whole only 19% of grid points in ECMWF (and 46% in MSC) have skill. Postprocessing leads to a considerable improvement almost everywhere, but over the cores of the oceanic deserts and tropical Africa skill remains close to zero (Figs. 5b and 9b). While the former is of little practical relevance and may well stem at least partly from problems with TRMM handling light rain from warm clouds, the latter is worrying given the large population of tropical Africa and socioeconomic importance of rainfall. The most likely reason for this deficit is the inability of convective parameterization schemes to represent the vertically tilted structure crucial for the upscale growth and propagation of mesoscale convective systems (Vogel et al. 2018).
- Forecasts of extreme rainfall events (using a threshold of 20 mm in 1 day): For land points model forecasts of extremes are only marginally worse than those of rainfall occurrence (44% with positive skill in ECMWF and 50% in MSC, see Table 1). This is partly due to some very poor performance in arid regions, where relatively few such cases occur, but also again due to Tropical Africa, where models struggle to represent the intensity of organized convection. Overall, postprocessing does not improve performance as much as for occurrence, likely due to smaller sample sizes.
- 5- versus 1-day accumulation times: Results for a 5-day accumulation period bear many resemblances to the 1-day results, indicating that the predictive performance is dominated by model error (e.g., boundary layer, convection). Particularly in wetter areas, the longer accumulation time even leads to improvement, as timing errors become less relevant. Such behavior appears to be typical of the inner tropics and is in strong contrast to higher latitudes (see Fig. 2a in Wheeler et al. 2017).
- Time evolution: Calibration in the ECMWF raw ensemble improves between 2009 and 2011 but not much afterward. This is likely associated with the increase in model resolution and a major cloud microphysics upgrade in 2010. Skill of postprocessed forecasts increases more gradually in most regions, indicating that the postprocessing is able to remedy some of the earlier model deficits. Mountain climates are generally forecast poorest but at least there is an increase by about 0.06 in CRPSS from 2009 to 2017. Findings for the MSC model broadly agree but the better calibration leads to a smaller skill gap.

The skill of the ensemble forecasts was assessed against the TRMM rainfall product, which has known spatiotemporarily varying limitations. Past research has demonstrated particular issues over mountains and coastal regions with warm rain—despite the gauge calibration of TRMM. In future studies, it would therefore be desirable to use the successor product Integrated Multisatellite Retrievals for GPM (IMERG), potentially other daily products such as Global Precipitation Climatology Project (GPCP) daily and Multi-Source Weighted-Ensemble Precipitation (MSWEB), and as many rainfall stations as possible over tropical and subtropical land areas.

Nevertheless, the results of this study have unveiled a number of considerable deficiencies in our ability to forecast rainfall in the tropics with global ensemble prediction systems. While for example over most of Australia, forecast performance is satisfactory and can only be improved rather little with postprocessing, the raw ensemble model output is hardly useful in many other regions and postprocessing is needed to increase predictive skill. A prominent exception is tropical Africa where forecasts are only little better than the climatological reference even after postprocessing. This shows that the deficits in realistically representing rainfall processes in this region dominated by organized convection impedes benefitting from (presumably useful) predictions of the environmental conditions that influence mesoscale convective systems formation and maintenance.

We propose several lines of research to improve tropical rainfall prediction: The first is to try alternative postprocessing approaches such as those used by Rasp and Lerch (2018), Medina et al. (2019), and Hewson and Pillosu (2020). The second is to improve global NWP models. Increased computing power and its more efficient use, as well as improvements in the understanding and parameterization of relevant processes (e.g., couplings between the boundary layer and shallow and deep convection), will likely help increase precipitation forecast skill. As the explicit simulation of deep convection is not computationally feasible on the global scale at the moment, we advocate experiments with limited-area convection-permitting deterministic and ensemble forecast systems (e.g., Pante and Knippertz 2019). The third idea is to use the coupling between more predictable synoptic-to planetary-scale wave phenomena and convection (Schlueter et al. 2019a,b) to improve forecasts. This would require training adequate statistical models, such as neural network approaches, with past observations and can in principle be done based on observations alone or additionally taking into account current model predictions of environmental factors. In particular for monsoon regions, a differentiation between seasons will be beneficial in any approach that builds heavily on statistics. Finally, efforts are needed to improve initial conditions in the tropics, where uncertainty is particular large in current operational systems and limits practical predictability (e.g., Žagar 2017). The long-term goal of such activities should be to lift the quality of forecasts to a level that is high enough to underpin real-world decision making to create socioeconomic benefit. In the short term, we strongly advocate the further development and operational use of ensemble postprocessing methods to provide essential forecast information to the vulnerable societies in the many

developing countries in the tropics (Webster 2013; Alley et al. 2019).

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 and via the Fellowship Programme at ECMWF. We thank Alexander Jordan and Michael Scheuerer for providing R code. We would also like to acknowledge WMO for initiating TIGGE in the framework of the World Weather Research Programme (WWRP). We also acknowledge valuable contributions from three anonymous reviewers that considerably helped improve a previous version of this manuscript.

Data availability statement. This work is based on TIGGE data. The Interactive Grand Global Ensemble (TIGGE) is an initiative of the World Weather Research Programme (WWRP). The ECMWF and MSC ensemble forecasts used here are freely accessible via the TIGGE archive (https://confluence.ecmwf.int/display/TIGGE). TRMM data were downloaded from https://gpm.nasa.gov/data-access/downloads/trmm (TRMM 2011).

REFERENCES

Agustí-Panareda, A., A. Beljaars, C. Cardinali, I. Genkova, and C. Thorncroft, 2010: Impacts of assimilating AMMA soundings on ECMWF analyses and forecasts. *Wea. Forecasting*, **25**, 1142–1160, https://doi.org/10.1175/2010WAF2222370.1.

Ahlgrimm, M., and R. Forbes, 2014: Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. *Mon. Wea. Rev.*, **142**, 668–685, https://doi.org/10.1175/MWR-D-13-00153.1.

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.

Baker, W. E., and Coauthors, 2014: Lidar-measured wind profiles: The missing link in the global observing system. *Bull. Amer. Meteor. Soc.*, **95**, 543–564, https://doi.org/10.1175/BAMS-D-12-00164.1.

Barros, A. P., S. Chiao, T. J. Lang, D. Burbank, and J. Putkonen, 2006: From weather to climate—Seasonal and interannual variability of storms and implications for erosion processes in the Himalaya. *Tectonics, Climate, and Landscape Evolution*, S. D. Willett et al., Eds., Geological Society of America, 17–38.

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.

Bengtsson, L., and Coauthors, 2019: Convectively coupled equatorial wave simulations using the ECMWF IFS and the NOAA GFS cumulus convection schemes in the NOAA GFS model. *Mon. Wea. Rev.*, **147**, 4005–4025, https://doi.org/10.1175/MWR-D-19-0195.1.

Berg, W., T. L'Ecuyer, and J. M. Haynes, 2010: The distribution of rainfall over oceans from spaceborne radars. *J. Appl. Meteor. Climatol.*, **49**, 535–543, https://doi.org/10.1175/2009JAMC2330.1.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.*, **135**, 1386–1402, https://doi.org/10.1175/MWR3341.1.

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.

Buizza, R., M. Miller, and T. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Camberlin, P., and Coauthors, 2019: Evaluation of remotely sensed rainfall products over central Africa. *Quart. J. Roy. Meteor. Soc.*, **145**, 2115–2138, https://doi.org/10.1002/qj.3547.

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.

Dias, J., M. Gehne, G. N. Kiladis, N. Sakaeda, P. Bechtold, and T. Haiden, 2018: Equatorial waves and the skill of NCEP and ECMWF numerical weather prediction systems. *Mon. Wea. Rev.*, **146**, 1763–1784, https://doi.org/10.1175/MWR-D-17-0362.1.

Dinku, T., C. Funk, P. Peterson, R. Maidment, T. Tadesse, H. Gadain, and P. Ceccato, 2018: Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Quart. J. Roy. Meteor. Soc.*, **144**, 292–312, https://doi.org/10.1002/qj.3244.

Engel, T., A. H. Fink, P. Knippertz, G. Pante, and J. Bliefernicht, 2017: Extreme precipitation in the West African cities of Dakar and Ouagadougou: Atmospheric dynamics and implications for flood risk assessments. *J. Hydrometeor.*, **18**, 2937–2957, https://doi.org/10.1175/JHM-D-16-0218.1.

Fan, J., L. R. Leung, D. Rosenfeld, Q. Chen, J. Zhang, and H. Yan, 2013: Microphysical effects determine macrophysical response for aerosol impacts on deep convective clouds. *Proc. Natl. Acad. Sci. USA*, **110**, E4581–E4590, https://doi.org/10.1073/pnas.1316830110.

Forbes, R. M., A. M. Tompkins, and A. Untch, 2011: A new prognostic bulk microphysics scheme for the IFS. ECMWF Tech. Memo. 649, 28 pp., https://doi.org/10.21957/bf6vjvxk.

Frierson, D. M. W., D. Kim, I.-S. Kang, M.-I. Lee, and J. Lin, 2011: Structure of AGCM-simulated convectively coupled Kelvin waves and sensitivity to convective parameterization. *J. Atmos. Sci.*, **68**, 26–45, https://doi.org/10.1175/2010JAS3356.1.

Geiger, R., 1961: *Überarbeitete Neuausgabe von Geiger, R. Köppen-Geiger/Klima der Erde (Wandkarte 1:16 Mill.).* Klett-Perthes.

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.

——, ——, 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.

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.

——, M. Janousek, J.-R. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart, 2018: Evaluation of ECMWF forecasts, including the 2018 upgrade. ECMWF Tech. Memo. 831, 52 pp., https://doi.org/10.21957/ldw15ckqi.

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.

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.

Hewson, T. D., and F. M. Pillosu, 2020: A new low-cost technique improves weather forecasts across the world. arXiv, 2003.14397, 27 pp., https://arxiv.org/ftp/arxiv/papers/2003/2003.14397.pdf.

Hirpa, F. A., M. Gebremichael, and T. Hopson, 2010: Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. *J. Appl. Meteor. Climatol.*, **49**, 1044–1051, https://doi.org/10.1175/2009JAMC2298.1.

Houze, R. A., 2012: Orographic effects on precipitating clouds. *Rev. Geophys.*, **50**, RG1001, https://doi.org/10.1029/2011RG000365.

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.

Judt, F., 2020: Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *J. Atmos. Sci.*, **77**, 257–276, https://doi.org/10.1175/JAS-D-19-0116.1.

Kiladis, G., and K. Weickmann, 1997: Horizontal structure and seasonality of large-scale circulations associated with submonthly tropical convection. *Mon. Wea. Rev.*, **125**, 1997–2013, https://doi.org/10.1175/1520-0493(1997)125<1997:HSASOL>2.0.CO;2.

Kniffka, A., and Coauthors, 2020: An evaluation of operational and research weather forecasts for southern West Africa using observations from the DACCIWA field campaign in June-July 2016. *Quart. J. Roy. Meteor. Soc.*, **146**, 1121–1148, https://doi.org/10.1002/qj.3729.

Knippertz, P., 2007: Tropical-extratropical interactions related to upper-level troughs at low latitudes. *Dyn. Atmos. Oceans*, **43**, 36–62, https://doi.org/10.1016/j.dynatmoce.2006.06.003.

Köppen, W., 1900: Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt *Geogr. Z.*, **6**, 593–611.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World map of the Köppen-Geiger climate classification updated. *Meteor. Z.*, **15**, 259–263, https://doi.org/10.1127/0941-2948/2006/0130.

Lang, M. N., S. Lerch, G. J. Mayr, T. Simon, R. Stauffer, and A. Zeileis, 2020: Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes Geophys.*, **27**, 23–34, https://doi.org/10.5194/npg-27-23-2020.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, https://doi.org/10.1214/16-STS588.

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.

Li, Y., and S. N. Stechmann, 2020: Predictability of tropical rainfall and waves: Estimates from observational data. *Quart. J. Roy. Meteor. Soc.*, **146**, 1668–1684, https://doi.org/10.1002/qj.3759.

Lin, J.-L., M.-I. Lee, D. Kim, I.-S. Kang, and D. M. W. Frierson, 2008: The impacts of convective parameterization and moisture triggering on AGCM-simulated convectively coupled equatorial waves. *J. Climate*, **21**, 883–909, https://doi.org/10.1175/2007JCLI1790.1.

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.

Matsuno, T., 1966: Quasi-geostrophic motions in the equatorial area. *J. Meteor. Soc. Japan*, **44**, 25–43, https://doi.org/10.2151/jmsj1965.44.1_25.

Medina, H., D. Tian, F. Marin, and G. Chirico, 2019: Comparing GEFS, ECMWF, and postprocessing methods for ensemble precipitation forecasts over Brazil. *J. Hydrometeor.*, **20**, 773–790, https://doi.org/10.1175/JHM-D-18-0125.1.

Miller, M., R. Buizza, J. Haseler, M. Hortal, P. Janssen, and A. Untch, 2010: Increased resolution in the ECMWF deterministic and ensemble prediction systems. *ECMWF Newsletter*, No. 124, ECMWF, Reading, United Kingdom, 10–16.

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.

Monsieurs, E., and Coauthors, 2018: Evaluating TMPA rainfall over the sparsely gauged East African Rift. *J. Hydrometeor.*, **19**, 1507–1528, https://doi.org/10.1175/JHM-D-18-0103.1.

Navascués, B., and Coauthors, 2013: Long-term verification of HIRLAM and ECMWF forecasts over southern Europe: History and perspectives of numerical weather prediction at AEMET. *Atmos. Res.*, **125–126**, 20–33, https://doi.org/10.1016/j.atmosres.2013.01.010.

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.

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.

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.

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.

Peters, K., C. Hohenegger, and D. Klocke, 2019: Different representation of mesoscale convective systems in convection-permitting

and convection-parameterizing NWP models and its implications for large-scale forecast evolution. *Atmosphere*, **10**, 503, https://doi.org/10.3390/atmos10090503.

Pfeifroth, U., J. Trentmann, A. H. Fink, and B. Ahrens, 2015: Evaluating satellite-based diurnal cycles of precipitation in the African Tropics. *J. Appl. Meteor. Climatol.*, **55**, 23–39, https://doi.org/10.1175/JAMC-D-15-0065.1.

Privé, N. C., and R. M. Errico, 2013: The role of model and initial condition error in numerical weather forecasting investigated with an observing system simulation experiment. *Tellus*, **65A**, 21740, https://doi.org/10.3402/tellusa.v65i0.21740.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

R Core Team, 2018: R: A language and environment for statistical computing. R Foundation for Statistical Computing, accessed 1 October 2020, https://www.R-project.org/.

Richard, E., A. Buzzi, and G. Zängl, 2007: Quantitative precipitation forecasting in the Alps: The advances achieved by the Mesoscale Alpine Programme. *Quart. J. Roy. Meteor. Soc.*, **133**, 831–846, https://doi.org/10.1002/qj.65.

Rosenfeld, D., U. Lohmann, G. B. Raga, C. D. O'Dowd, M. Kulmala, S. Fuzzi, A. Reissell, and M. O. Andreae, 2008: Flood or drought: How do aerosols affect precipitation? *Science*, **321**, 1309–1313, https://doi.org/10.1126/science.1160606.

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.

Schlueter, A., A. H. Fink, and P. Knippertz, 2019a: A systematic comparison of tropical waves over northern Africa. Part II: Dynamics and thermodynamics. *J. Climate*, **32**, 2605–2625, https://doi.org/10.1175/JCLI-D-18-0651.1.

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

Smith, R. K., G. Garden, J. Molinari, and B. R. Morton, 2001: Proceedings of an international workshop on the dynamics and forecasting of tropical weather systems. *Bull. Amer. Meteor. Soc.*, **82**, 2825–2829, https://doi.org/10.1175/1520-0477(2001)082<2825:MSPOAI>2.3.CO;2.

Stellingwerf, S., E. Riddle, T. Hopson, J. Kneivel, B. Brown, and M. Gebremichael, 2020: Optimizing precipitation forecasts for hydrological catchments in Ethiopia using statistical bias correction and multi-modeling. *Earth Space Sci.*, in press.

Stephens, G. L., and Coauthors, 2010: Dreary state of precipitation in global models. *J. Geophys. Res.*, **115**, D24211, https://doi.org/10.1029/2010JD014532.

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.

TRMM, 2011: TRMM (TMPA) rainfall estimate L3 3 hour 0.25° × 0.25° V7. Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), accessed 1 October 2020, https://doi.org/10.5067/TRMM/TMPA/3H/7.

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.

——, P. Knippertz, A. H. Fink, I. Ingleby, M. Maranan, and A. Benedetti, 2020: The influence of DACCIWA radiosonde data on the quality of ECMWF analyses and forecasts over southern West Africa. *Quart. J. Roy. Meteor. Soc.*, **146**, 1719–1739, https://doi.org/10.1002/qj.3763.

Vannitsem, S., D. Wilks, and J. Messner, Eds., 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 362 pp., https://doi.org/10.1016/C2016-0-03244-8.

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.

Webster, P. J., 2013: Improve weather forecasts for the developing world. *Nature*, **493**, 17–19, https://doi.org/10.1038/493017a.

——, V. E. Toma, and H.-M. Kim, 2011: Were the 2010 Pakistan floods predictable? *Geophys. Res. Lett.*, **38**, L04806, https://doi.org/10.1029/2010GL046346.

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.

——, H. Zhu, A. H. Sobel, D. Hudson, and F. Vitart, 2017: Seamless precipitation prediction skill comparison between two global models. *Quart. J. Roy. Meteor. Soc.*, **143**, 374–383, https://doi.org/10.1002/qj.2928.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp.

Young, M. P., C. J. R. Williams, J. C. Chiu, R. I. Maidment, and S.-H. Chen, 2014: Investigation of discrepancies in satellite rainfall estimates over Ethiopia. *J. Hydrometeor.*, **15**, 2347–2369, https://doi.org/10.1175/JHM-D-13-0111.1.

——, J. C. Chiu, C. J. R. Williams, T. H. M. Stein, M. Stengel, M. D. Fielding, and E. Black, 2018: Spatio-temporal variability of warm rain events over southern West Africa from geostationary satellite observations for climate monitoring and model evaluation. *Quart. J. Roy. Meteor. Soc.*, **144**, 2311–2330, https://doi.org/10.1002/qj.3372.

Žagar, N., 2017: A global perspective of the limits of prediction skill of NWP models. *Tellus*, **69A**, 1317573, https://doi.org/10.1080/16000870.2017.1317573.

Zhang, H., and Z. Pu, 2010: Beating the uncertainties: Ensemble forecasting and ensemble based data assimilation in modern numerical weather prediction. *Adv. Meteor.*, **2010**, 432160, https://doi.org/10.1155/2010/432160.