Subseasonal precipitation prediction for Africa: forecast evaluation and sources of predictability

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1175/WAF-D-20-0054.1

Subseasonal Precipitation Prediction for Africa: Forecast Evaluation and Sources of Predictability

FELIPE M. DE ANDRADE,[a] MATTHEW P. YOUNG,[a] DAVID MACLEOD,[b] LINDA C. HIRONS,[a]
STEVEN J. WOOLNOUGH,[a] AND EMILY BLACK[a]

[a] *National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom*
[b] *School of Geographical Sciences, University of Bristol, Bristol, United Kingdom*

ABSTRACT: This paper evaluates subseasonal precipitation forecasts for Africa using hindcasts from three models (ECMWF, UKMO, and NCEP) participating in the Subseasonal to Seasonal (S2S) prediction project. A variety of verification metrics are employed to assess weekly precipitation forecast quality at lead times of one to four weeks ahead (weeks 1–4) during different seasons. Overall, forecast evaluation indicates more skillful predictions for ECMWF over other models and for East Africa over other regions. Deterministic forecasts show substantial skill reduction in weeks 3–4 linked to lower association and larger underestimation of predicted variance compared to weeks 1–2. Tercile-based probabilistic forecasts reveal similar characteristics for extreme categories and low quality in the near-normal category. Although discrimination is low in weeks 3–4, probabilistic forecasts still have reasonable skill, especially in wet regions during particular rainy seasons. Forecasts are found to be overconfident for all weeks, indicating the need to apply calibration for more reliable predictions. Forecast quality within the ECMWF model is also linked to the strength of climate drivers' teleconnections, namely, El Niño–Southern Oscillation, Indian Ocean dipole, and the Madden–Julian oscillation. The impact of removing all driver-related precipitation regression patterns from observations and hindcasts shows reduction of forecast quality compared to including all drivers' signals, with more robust effects in regions where the driver strongly relates to precipitation variability. Calibrating forecasts by adding observed regression patterns to hindcasts provides improved forecast associations particularly linked to the Madden–Julian oscillation. Results from this study can be used to guide decision-makers and forecasters in disseminating valuable forecasting information for different societal activities in Africa.

KEYWORDS: ENSO; Madden-Julian oscillation; Precipitation; Forecast verification/skill; Hindcasts; Probability forecasts/models/distribution

1. Introduction

Delivering useful subseasonal forecasts (between 2 weeks and 2 months ahead) remains a great challenge for operational forecasting centers, as this time scale is too long to retain much of the influence of the atmospheric initial conditions and sufficiently short to be dominated by the forced boundary conditions. The lack of subseasonal precipitation forecast quality over many regions worldwide has been identified by evaluating near real-time forecasts and hindcasts made available by the Subseasonal to Seasonal (S2S) prediction project (Vitart et al. 2017). The target goal of the S2S project is to address the predictability gap between medium-range weather predictions and seasonal climate predictions to improve forecast quality on subseasonal time scales for a range of applications, for

instance, agriculture, water resource management, and other socioeconomic activities.

The S2S database has been used to evaluate subseasonal precipitation forecasts on a weekly basis (Vigaud et al. 2017a,b; Coelho et al. 2018; de Andrade et al. 2019; among others). For example, de Andrade et al. (2019) evaluated weekly precipitation hindcasts from all models participating in the S2S project, finding best agreement with precipitation observations during the first two weeks lead and worst quality in subsequent weeks, especially over extratropical regions. Weekly precipitation predictions were also verified over summer monsoon regions of the Northern Hemisphere and the East Africa–West Asia sector (Vigaud et al. 2017b, 2018), both showing worst quality for longer lead times (i.e., beyond two weeks lead).

Despite the fact that there is poorer precipitation forecast quality within S2S models after the first two weeks lead, recent studies have analyzed the role played by particular sources of subseasonal predictability, such as El Niño–Southern Oscillation (ENSO) and the Madden–Julian oscillation (MJO), in modulating the quality of precipitation forecasts. Li and Robertson (2015) evaluated weekly precipitation forecasts as a function of ENSO and MJO metrics supporting the concept that particular

climate drivers' conditions can promote better subseasonal predictions. Moreover, de Andrade et al. (2019) found a reduction in forecast quality after removing ENSO- and MJO-related precipitation patterns from weekly forecasts. Other drivers could also affect the quality of subseasonal precipitation predictions, for instance, tropical–extratropical interactions (Vigaud et al. 2019) and stratosphere–troposphere coupling (Domeisen et al. 2020). For a more comprehensive description of relevant drivers of subseasonal predictability, the reader is referred to Mariotti et al. (2020).

Among the many efforts to improve understanding of subseasonal forecast skill over the past years, one important aspect is forecast verification (Coelho et al. 2019). Several studies have identified areas where precipitation forecast quality of S2S models could be refined (e.g., Vigaud et al. 2018; de Andrade et al. 2019). However, only few of those studies have employed detailed verifications to analyze the attributes of forecast quality defined in Murphy (1993). Since a single verification score is unable to evaluate different attributes of forecast quality, an assessment of a set of metrics is required to help obtain a fully comprehensive overview of S2S models' ability to predict subseasonal precipitation (Coelho et al. 2018). Here, weekly precipitation forecast quality from three S2S models is investigated over the African continent assessing the attributes of deterministic and probabilistic forecast quality using a variety of metrics. Such a comprehensive exploration of subseasonal forecast quality not only has the potential to advance the scientific understanding, but also provide support to forecasters and decision-makers in different sectors of society, improving early warning systems and lives and livelihoods of millions of people in Africa. Furthermore, an evaluation of how well models capture the relationships of important climate drivers with African precipitation and its contribution to the quality of forecasts also deserves investigation to deepen our knowledge of the sources of subseasonal predictability. Thus, this study provides an unprecedented weekly precipitation forecast evaluation for Africa, examining different ensemble prediction systems and key drivers modulating high-impact weather events.

Section 2 outlines the datasets and methods employed to evaluate the attributes of forecast quality. Section 2 also provides a description of the methodology used to analyze particular sources of subseasonal predictability and their links to African precipitation forecast quality. The results of deterministic and probabilistic forecast verification are presented in section 3, followed by an assessment of key driver-dependent forecast quality in section 4. A summary and conclusions are given in section 5.

2. Data and methods

a. S2S hindcasts

Precipitation hindcasts from the S2S database were evaluated for the European Centre for Medium-Range Weather Forecasts (ECMWF), the Met Office (UKMO), and the National Centers for Environmental Prediction (NCEP) models. These hindcasts have different configurations such as forecast length, spatial resolution, frequency, period, ensemble size, and coupling effects (Table 1); see Vitart et al. (2017) for further details. Moreover, ECMWF and UKMO hindcasts are produced gradually by updating their model versions according to near real-time forecasts, whereas in the NCEP model hindcasts have a fixed date for a given model version. We analyzed ECMWF and UKMO hindcasts corresponding to model version dates of the year 2018.

Four start dates per month were chosen based on UKMO initializations (the 1st, 9th, 17th, and 25th). We selected the closest start date for certain nonmatching ECMWF initializations. This discrepancy regarding models' initialization restricted a multimodel evaluation. To have a fair intercomparison among models, three perturbed members, extracted from 1-day lag after initializations, were added to the NCEP ensemble size. This procedure allowed all models having at least seven ensemble members. Since the subseasonal time scale is beyond the weather prediction limit, a weekly time frame was employed for more adequately representing the subseasonal forecast range. Weekly precipitation was obtained considering four accumulation lead times: days 5–11 (week 1), 12–18 (week 2), 19–25 (week 3), and 26–32 (week 4).

b. Observational dataset

Hindcasts were verified using data from the Global Precipitation Climatology Project (GPCP), version 1.2 (Huffman et al. 2001). Daily GPCP precipitation is produced by the National Aeronautics and Space Administration (NASA), by blending precipitation estimates from gauge stations and satellite measurements, and sourced from the National Center for Atmospheric Research (NCAR). GPCP data were linearly interpolated to the 1.5° spatial resolution to match models regridded resolution made available in the S2S database and used to calculate accumulated precipitation for the weekly periods defined in section 2a.

c. Forecast verification framework

Forecast verification is a process to evaluate the robustness of an ensemble prediction system, providing a guide for identifying its strengths and weaknesses when examining the joint distribution of forecasts and observations. A common forecast verification practice consists of assessing the attributes of deterministic and probabilistic forecast quality by computing metrics depending on forecast type (Coelho et al. 2019). Deterministic forecast verification metrics compare quantitative forecasts to observations (e.g., rainfall amounts in millimeters). The evaluation of deterministic forecasts is most often conducted by analyzing the ensemble mean to verify the value of using a set of perturbed initial conditions rather than a single unperturbed forecast. Probabilistic forecast verification metrics compare forecast probabilities to observations (e.g., probability of above-normal rainfall). Probabilities are usually examined in different categories and obtained by taking the proportion of the ensemble members falling in ranges defined by certain predefined thresholds (e.g., 33rd or 67th percentiles). Specifically, binary observations are used to assess probabilistic forecasts of dichotomous variables with two possible outcomes (e.g., rain or no rain events).

TABLE 1. The main features of the three S2S operational models and their hindcasts.

Model	Forecast length	Spatial resolution	Hindcast frequency	Hindcast period	Ensemble size	Ocean coupled	Sea ice coupled
ECMWF	46 days	Tco639/319L91	Two per week	Past 20 years	11	Yes	No
UKMO	60 days	N216 L85	Four per month	1993–2016	7	Yes	Yes
NCEP	44 days	T126 L64	Daily	1999–2010	$4 + 3^a$	Yes	Yes

[a] Three more perturbed members, extracted from 1-day lag after initializations, were added to the NCEP ensemble size.

A variety of deterministic and probabilistic forecast verification metrics were used to evaluate the attributes of forecast quality defined in Murphy (1993). Attributes and metrics are summarized below, with a more detailed description in Coelho et al. (2019):

- Bias is the mean difference between the deterministic forecasts and observations. Bias can indicate a model's overestimation (bias > 0) or underestimation (bias < 0), but it does not provide any information on the magnitude of the absolute error. Bias is assessed by the mean error [ME; (1)]:

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

where N denotes the sample size, F_i the forecast totals, and O_i the observation totals.

- Association describes the linear relationship between deterministic forecasts and observations. Forecasts with good association are highly positively correlated with observations. The Pearson's correlation coefficient [R; (2)] is a common metric of association indicating the direction of deviations [R close to 1 (-1) indicates strong positive (negative) association]:

$$R = \frac{\sum_{i=1}^{N} F_i' O_i'}{\sqrt{\sum_{i=1}^{N} F_i'^2}\sqrt{\sum_{i=1}^{N} O_i'^2}}, \quad (2)$$

where F_i' denotes the forecast anomalies and O_i' the observed anomalies.

- Accuracy is the difference between forecasts and observations, providing the magnitude of forecast errors. Thus, the lower the difference, the better the accuracy. The mean square error [MSE; (3)] assesses deterministic errors, whereas the ranked probability score [RPS; (4)] evaluates probabilistic errors for more than two probability categories:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (F_i' - O_i')^2, \quad (3)$$

$$\text{RPS} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{K-1}\sum_{j=1}^{K}\left[\left(\sum_{z=1}^{j} P_{iz}\right) - \left(\sum_{z=1}^{j} O_{iz}\right)\right]^2\right\}_i, \quad (4)$$

where K is the number of categories, P_{iz} is the cumulative forecast probability and O_{iz} is the cumulative binary observation for occurrence ($O_{iz} = 1$) and nonoccurrence ($O_{iz} = 0$) of an event. RPS is a generalized version of the Brier score (BS) for two categories.

- Skill evaluates the accuracy of forecasts relative to some reference forecast, such as observed climatology. The skill score [SS; (5)] indicates forecasts more (less) skillful than the reference when is positive (negative):

$$\text{SS} = 1 - \frac{S_f}{S_r}, \quad (5)$$

where S_f is the score for forecasts and S_r the score for the reference forecast. A perfect SS would be equal to 1. The SS assesses deterministic and probabilistic skill using the MSE (3) and the RPS (4), resulting in the mean square skill score [MSSS; (6)] and ranked probability skill score [RPSS; (7)], respectively:

$$\text{MSSS} = 1 - \frac{\text{MSE}_f}{\text{MSE}_r}, \quad (6)$$

$$\text{RPSS} = 1 - \frac{\text{RPS}_f}{\text{RPS}_r}, \quad (7)$$

where MSE_f and MSE_r are the MSEs for forecasts (3) and for a reference forecast, respectively. Here, RPS_f is the RPS for forecasts (4) and RPS_r the RPS for a reference forecast. RPSS is sensitive to ensemble size and a negative bias is introduced for small ensemble sizes (Müller et al. 2005). To overcome this sensitivity, we use a debiased (discrete) RPSS [RPSS_D; (8)] derived for any ensemble size and probability category by adding a bias correction term on the reference forecast (Weigel et al. 2007) rather than including a correction term by randomly resampling from climatology (Müller et al. 2005):

$$\text{RPSS}_D = 1 - \frac{\text{RPS}_f}{\text{RPS}_r + D}. \quad (8)$$

For equiprobable K categories, the correction term D is defined as $D = (1/M)[(K^2 - 1)/6K]$, where M is the ensemble size.

- Discrimination is the ability of forecasts at discerning between different observed outcomes. For dichotomous forecasts, it is the ability of a forecast at distinguishing between occurrence and nonoccurrence of events, for instance precipitation falling in a tercile category. The relative operating characteristic (ROC) diagram and the area under the curve (AUC) are metrics adopted for assessing discrimination and providing useful information for decision-makers. The ROC

diagram for a given event is obtained by plotting the hit rate against the false alarm rate computed at different probability thresholds. The hit rate is the ratio between the number of correct forecasts of the event and the total number of occurrences of the event, whereas the false alarm rate is the ratio between the number of noncorrect forecasts of the event and the total number of nonoccurrences of the event. The diagonal line in the ROC diagram is where the hit rate equals the false alarm rate and indicates no discrimination. Better discrimination is found when the ROC curve is above the diagonal line and close to the upper-left corner, indicating the hit rate exceeds the false alarm rate. The AUC is computed from the ROC diagram joining the points associated with each threshold to form a series of trapezoids and adding their areas. The AUC is interpreted as a score, indicating no (perfect) discrimination when equal to 0.5 (1) (Kharin and Zwiers 2003a).

- Reliability measures the conditional bias in forecast probabilities, indicating the extent of their over or underconfidence. A reliable forecasting system is identified for all probability thresholds when the probabilistic outcomes are equal to the observed frequencies. For example, if a system is reliable, we should expect an event to occur 60% of the times the system issues a 60% probability of occurrence. Resolution assesses the degree of variability in the observed frequencies at different forecast probabilities. Sharpness evaluates the ability of forecasts to predict extreme probabilities. The attributes diagram (AD) is a useful way to verify probabilistic forecasts by summarizing the ability of ensemble prediction systems to represent the attributes of reliability, resolution, and sharpness. The AD is constructed by plotting the observed frequency for different forecast probabilities. Stratification is done by binning data into different probability thresholds. The diagonal line in the AD indicates perfect reliability in which the forecast probabilities are equal to the observed frequency. The horizontal line represents the observed climatological frequency, indicating no resolution. The line of no-skill can be found at the midpoint between the perfect reliability and observed frequency climatology. Probabilities falling into the area between the no-skill line and the vertical line replicating the horizontal line contribute to increase skill as demonstrated by the decomposition of the RPS/BS (Murphy 1972, 1973). Histograms provide information on the frequency of forecasts in each bin and the degree of sharpness.

Evaluation was performed over the African continent and adjacent regions to explore forecasting quality not only over land, but also oceanic areas where important atmospheric systems, such as the intertropical convergence zone (ITCZ), are located. To analyze the regional performance of the models, verification metrics were computed over four geographically selected African regions (Fig. 1), referred to as West African Monsoon (WAM), Equatorial West Africa (EWA), Equatorial East Africa (EEA), and Southern Africa (SA). These locations were chosen to represent different climate regions with particular rainy seasons (Zaitchik 2017).
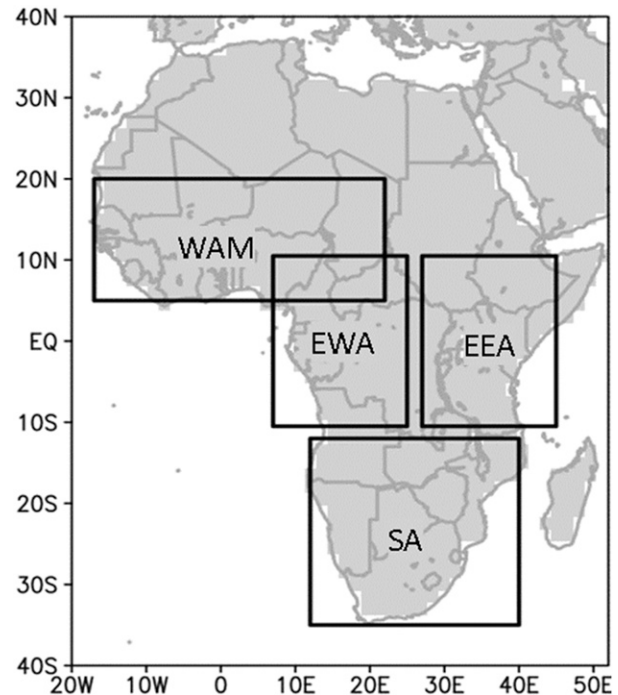

FIG. 1. African regions analyzed in the present study. Black boxes approximately denote the African regions reviewed by Zaitchik (2017): West African Monsoon region (WAM; 4.5°–19.5°N, 16.5°W–21°E), Equatorial West Africa (EWA; 10.5°S–10.5°N, 7.5°–25.5°E), Equatorial East Africa (EEA; 10.5°S–10.5°N, 27°–45°E), and Southern Africa (SA; 34.5°–12°S, 12°–40.5°E).

For deterministic forecasts, ensemble mean anomalies were obtained after subtracting the ensemble mean climatology computed through a leave-one-out cross-validation method without considering the verified year. Such an approach has been applied to ensure that no information from a given forecast is used in the verification procedure of the same forecast (e.g., Vitart 2017). This should provide independence between forecasts and the verification subset to avoid unfair evaluation and minimize potential skill overestimation (Wilks 2006). For probabilistic forecasts, tercile categories (below-normal, near-normal, and above-normal) were analyzed as they are frequently used in forecasting and provide a useful way to assess the model's ability to distinguish between dry, normal, and wet weeks. Tercile categories were defined using precipitation totals for each model ensemble member and employing a cross-validation method leaving one year out. The lower and upper terciles were estimated after pooling all model ensemble members together. Probabilities were obtained by computing the fraction of ensemble members in each tercile category. Ensemble mean anomalies and tercile probabilities were calculated depending on the start date and lead time. Observed anomalies and binaries were calculated in the same way.

Verification metrics were calculated for each model and lead time using forecasts where the start date falls within the following seasons over the common period of 1999–2010:

December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON). While these seasons may differ slightly from localized rainy seasons, they represent the main wet seasons found across Africa and are suitable for an overall evaluation. For each model, 144 forecasts (12 starts per season over 12 years) were examined using the available ensemble members shown in Table 1. Statistical significance of the correlations different from zero was analyzed using a two-sided Student's t test (Wilks 2006) with 95% significance level. The effective sample size was calculated based on lag-1 autocorrelation (Livezey and Chen 1983).

d. Sources of subseasonal predictability

ENSO, the Indian Ocean dipole (IOD), and the MJO are important modes of S2S variability influencing African precipitation (e.g., Behera et al. 2005; Ratnam et al. 2014; Sossa et al. 2017). Thus, their contribution to forecast quality was also evaluated. For the sake of brevity, we have analyzed the driver-dependent forecast quality using the ECMWF 11-member ensemble mean only. We have considered a longer period (1997–2014) and all available model version dates of the year 2017. Using more initializations provides larger sample sizes, enhancing the statistical robustness. The datasets and methodologies are described below.

1) DRIVERS' INDICES

ENSO and IOD indices were obtained, respectively, by averaging sea surface temperature (SST) anomalies in the Niño-3.4 region (5°S–5°N, 120°–170°W) (Bamston et al. 1997) and computing the dipole mode index (DMI) as the difference of area-averaged SST anomalies between the west (10°S–10°N, 50°–70°E) and southeastern (10°S–0°, 90°–110°E) tropical Indian Ocean (Saji et al. 1999). The daily optimum interpolation SST version 2 (OISST.v2) of the National Oceanic and Atmospheric Administration [NOAA; Reynolds et al. (2007)] was used as observational reference and SST hindcasts from S2S database as predicting fields. Observed and forecasted weekly SST was obtained by averaging daily values over the four weeks defined in section 2a. SST anomalies were computed by removing the climatology from the total field considering a cross-validation approach. Weekly ENSO and IOD indices were normalized by the corresponding standard deviation.

The real-time multivariate MJO [RMM; Wheeler and Hendon (2004)] index was calculated as in Gottschalck et al. (2010) and Vitart (2017), which follows the same approach employed for obtaining this index made available in the S2S database. The RMM index components (RMM1 and RMM2) were computed by projecting latitudinally averaged daily anomalies of zonal wind (850 and 200 hPa) and outgoing longwave radiation (OLR) at the top of the atmosphere onto the two dominant observed eigenvectors associated with the MJO. Zonal wind at 0000 UTC from ERA-Interim reanalysis (Dee et al. 2011) and daily interpolated OLR from NOAA (Liebmann and Smith 1996) were used for calculating the observed index. Zonal wind and OLR from S2S hindcasts were selected as corresponding forecasts. Reanalysis and

hindcasts were linearly interpolated from a horizontal resolution of 1.5°–2.5°, matching the same 144 longitudinal grid points of observed OLR and eigenvectors. Zonal wind and OLR anomalies were calculated by subtracting the climatology from the total field considering a cross-validation approach. Low-frequency signals within verifying datasets were filtered by removing the 120-day mean of the previous 120 days from each day. The 120-day mean was subtracted from forecasts using a combination of observations and hindcasts, filling with observed data the missing days preceding model's initializations. Then, anomalies were normalized by its respective observed normalization factor as in Gottschalck et al. (2010). Last, anomalies were projected onto the two leading eigenvectors and divided by the corresponding observed standard deviation calculated by Wheeler and Hendon (2004), generating RMM1 and RMM2 time series. Observed and forecasted weekly RMM components were computed following a similar approach applied for obtaining weekly SST.

2) QUALITY OF FORECASTS RELATIVE TO DRIVERS' SIGNAL

To explore the ability of forecasts to capture the relationship between precipitation variability and different drivers, a simple linear regression analysis between weekly precipitation and drivers' indices was performed using observations and hindcasts in weeks 1–4 for initializations within DJF, MAM, JJA, and SON. Over 18 years, 450 forecasts were used in DJF (25 starts), 468 in MAM/SON (26 starts), and 486 in JJA (27 starts). Modeled (observed) regression coefficients were obtained by regressing out hindcast (GPCP) precipitation anomalies with forecasted (observed) drivers' indices. Precipitation anomalies were computed as in section 2c. Since significant associations can exist between ENSO and IOD (e.g., Zhang et al. 2015), a multiple linear regression approach was also employed to examine ENSO- and IOD-related rainfall variability simultaneously. Regression coefficients were scaled to one standard deviation of the index following Lo and Hendon (2000). A two-sided Student's t test (Allen 1997) was applied with 95% significance level for evaluating statistical significance of regression slopes different from zero. Effective sample size was determined as in section 2c.

Forecast quality was initially analyzed through the absolute difference between forecasted and observed regression coefficients to determine model's ability in representing drivers' teleconnections to African rainfall. Next, observed and modeled rainfall variations linearly dependent on drivers were, respectively, removed from observed and predicted fields to evaluate the association between observations and hindcasts after subtracting ENSO-, IOD-, and MJO-related rainfall patterns. After removing the modeled precipitation variability associated with the drivers from hindcasts, the effect of adding observed regression patterns, i.e., obtained by regressing GPCP precipitation anomalies with observed drivers' indices, to the hindcasts was also examined to verify the quality of calibrated forecasts. The regional average of the absolute difference and correlation between observations and forecasts
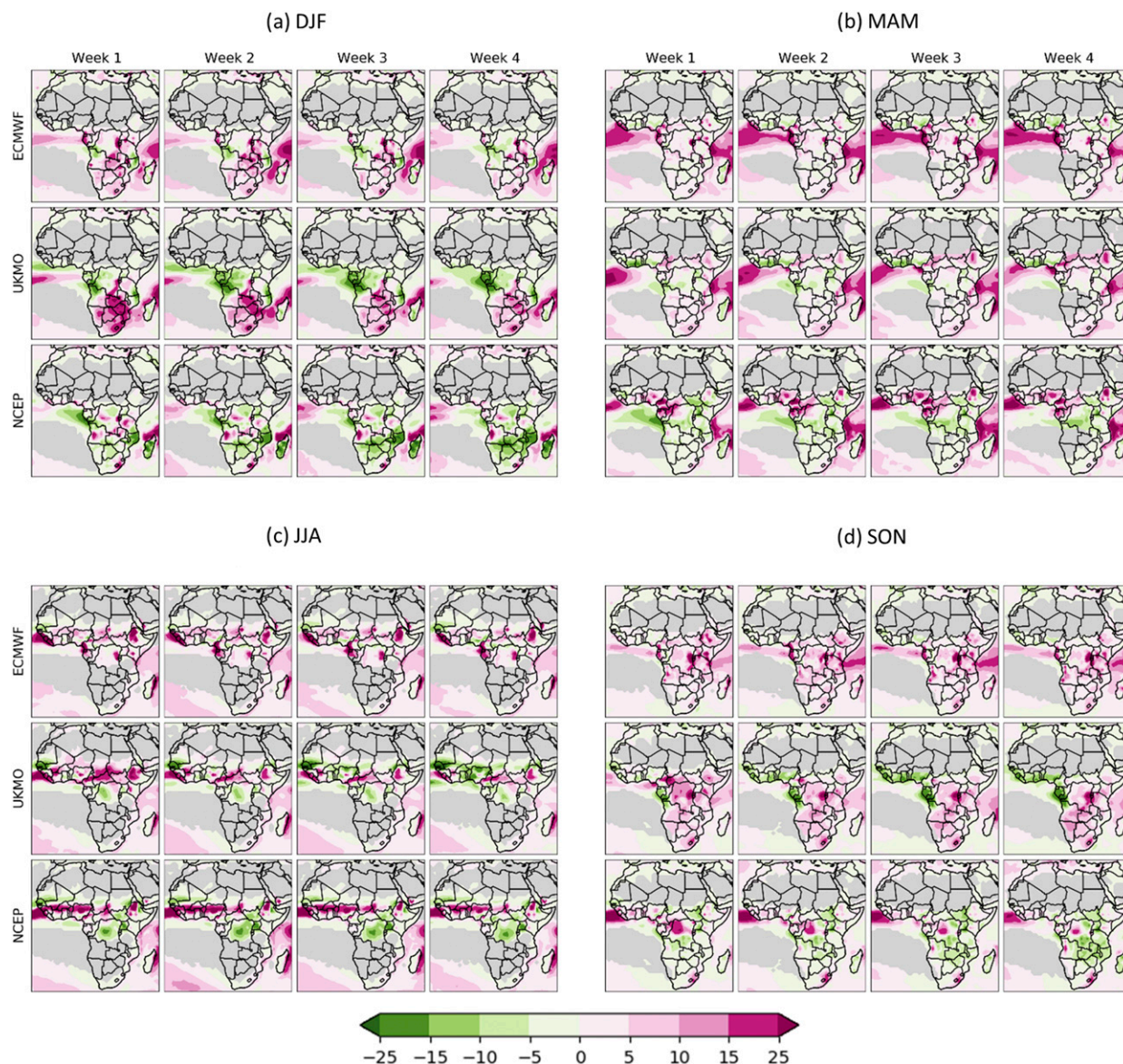
FIG. 2. Mean error (ME) between the hindcast ensemble mean and observed precipitation totals for ECMWF, UKMO, and NCEP models in weeks 1–4 for initializations during (a) DJF, (b) MAM, (c) JJA, and (d) SON over the 1999–2010 period. Units are accumulated millimeters per week. Gray shading denotes a dry mask applied over regions where the observed weekly precipitation climatology is less than 1 mm for more than 50% of start dates within a season.

was analyzed over the four regions shown in Fig. 1. Significant correlations were obtained as in section 2c.

3. Forecast quality assessment

In this section, a subseasonal African precipitation forecast quality assessment for three S2S models (ECMWF, UKMO, and NCEP) is conducted for lead times from one to four weeks ahead considering start dates in DJF, MAM, JJA, and SON during 1999–2010. For consistency between models, only results using seven ensemble members of each model are shown as findings indicated a slight improvement when examining the full ensemble size of ECMWF. Although the first seven ensemble members of

ECMWF have been selected for evaluation, results are similar if chosen at random. The below-normal category assessment overall shows similar performance to the above-normal category, whereas the assessment for the near-normal category indicates unskillful forecasts. For this reason, probabilistic evaluation is focused on results for the above-normal category, with results for the other categories mentioned when necessary and made available in the online supplemental material.

a. Deterministic verification

Figure 2 shows the mean error between hindcast and observed precipitation totals. Biases differ among seasons and in
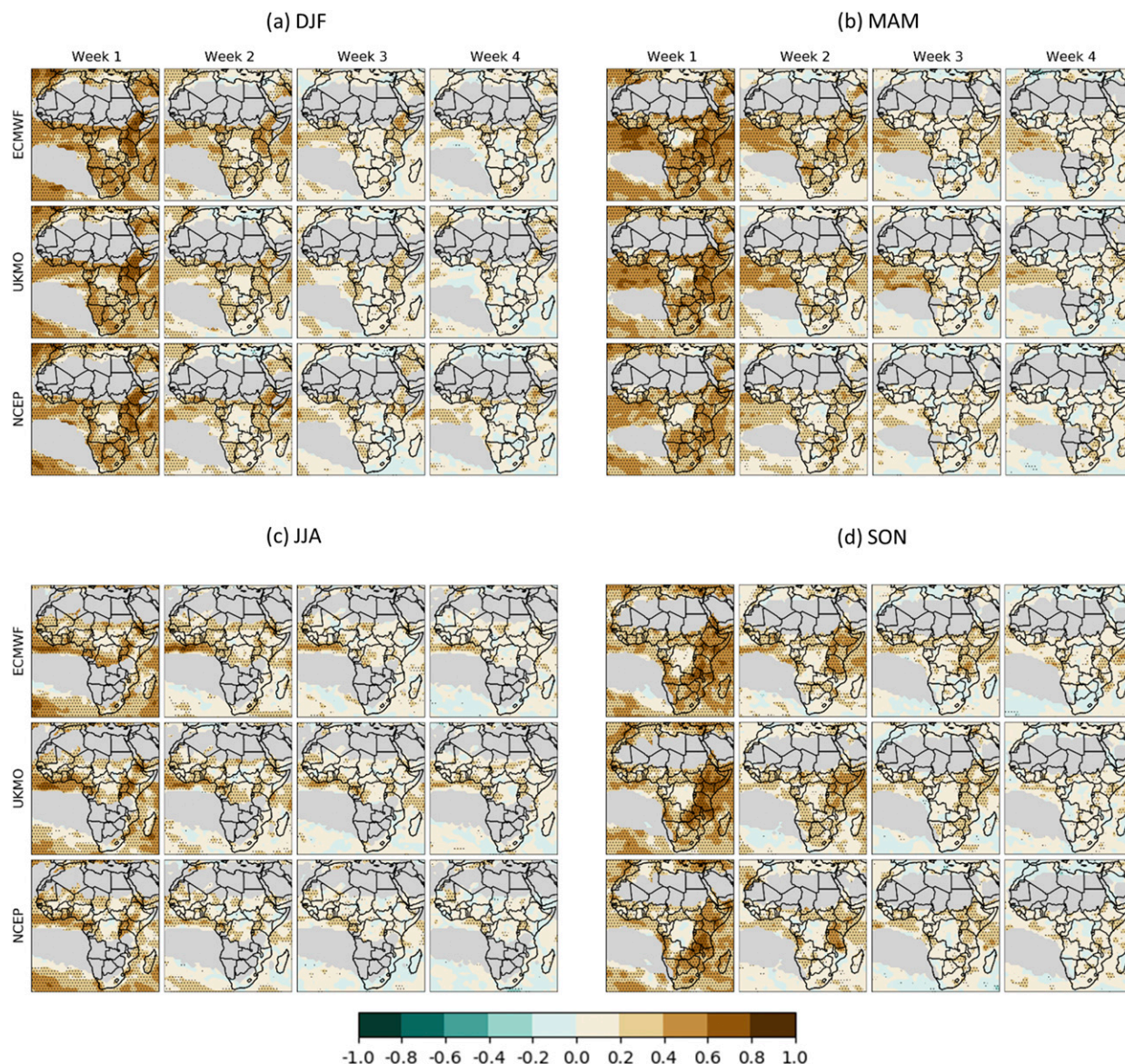
FIG. 3. As in Fig. 2, but for the Pearson's correlation coefficient (R) between the hindcast ensemble mean and observed precipitation anomalies. Stipples indicate correlations statistically significant at the 95% level.

ECMWF and NCEP are approximately constant throughout the weeks over most regions; however, UKMO shows a drying trend with lead time, particularly in DJF and JJA. In general, ECMWF has the lowest biases over land compared to other models, with roughly similar spatial patterns to UKMO, except in DJF and SON when strong negative biases develop over EWA coastal regions in UKMO (Figs. 2a,d). NCEP generally has the opposite sign to ECMWF and UKMO over East and south-southeastern Africa, with overestimation (underestimation) for ECMWF and UKMO (NCEP) in these regions notable during their wet seasons (SON and DJF, respectively). Models have deficiencies in representing precipitation near Mozambique and Madagascar in DJF, which could affect subseasonal prediction of tropical cyclones

across the region (Kolstad 2019). Large positive biases seen on the equatorial Atlantic and Indian Oceans in MAM (Fig. 2b) are likely related to shortcomings in predicting the seasonal migration of the ITCZ (e.g., Shonk et al. 2019). All models show similar biases at weeks 1–2 over the Sahel in JJA (Fig. 2c), with some evidence of a meridional tripole structure, which is particularly zonally uniform in NCEP. The drying trend in UKMO leads to strong negative biases in the core of the WAM by weeks 3–4.

Linear correlation is used to evaluate association between hindcasts and observed precipitation anomalies (Fig. 3). Positive correlations are strongest for week 1 and reduce with increasing lead time, with significant correlations mainly concentrated near the equator after two weeks lead, corroborating
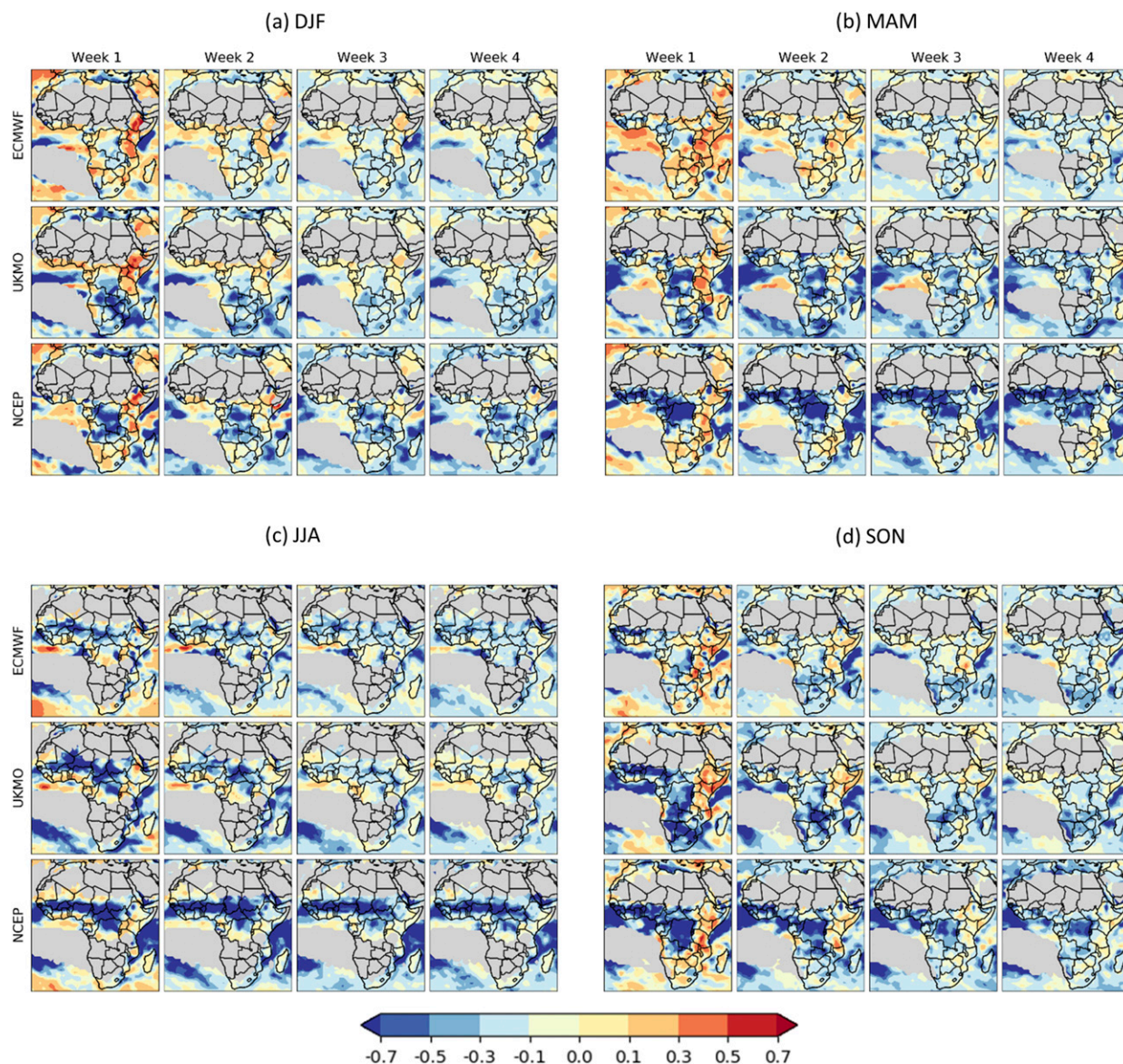
FIG. 4. As in Fig. 2, but for the mean square skill score (MSSS) between the hindcast ensemble mean and observed precipitation anomalies. A zero anomaly forecast was adopted for representing the reference forecast.

with previous assessments (e.g., Li and Robertson 2015). Weak associations are highlighted in weeks 3–4 over SA and adjacent oceans, probably due to the natural unpredictability of the extratropical fluctuations as suggested in de Andrade et al. (2019). Low correlation over Central Africa, near Democratic Republic of the Congo (DRC), suggests models' failure in representing variations on the meridional migration of tropical convection throughout the year. However, the observations may also be uncertain here due to low numbers of rain gauges (Washington et al. 2013). Significant correlations are found over East Africa up to week 4 for starts in DJF, MAM, and SON (Figs. 3a,b,d), particularly for ECMWF. In JJA, high association is shown over West Africa near the Gulf of Guinea (GoG) for all models, with significant correlations up to week 4 (Fig. 3c).

Maps of MSSS obtained by relating the MSE between hindcasts and observed precipitation anomalies to the reference MSE are shown in Fig. 4. Skill substantially decreases over most regions from week 1 to subsequent weeks. Skill is more pronounced over East Africa in DJF, MAM, and SON (Figs. 4a,b,d), showing, for example, positive scores up to week 4 during DJF for ECMWF. Skill in JJA is restricted to a region of West Africa near the equatorial Atlantic (Fig. 4c). Despite presenting large areas of positive correlation (Fig. 3), models show negative MSSS as a remarkable characteristic in all seasons, suggesting large errors at predicting precipitation anomalies, especially UKMO and NCEP. This can be investigated by decomposing the MSSS into three squared components (Murphy 1988).

The first term is R (2), the second is the conditional bias providing the forecast amplitude errors, and the third is the unconditional bias (1), which is zero when considering anomalies. The conditional bias is computed as $[R - (S_f/S_o)]^2$, where S_f and S_o are the standard deviations of the forecasts and observations, respectively. By expanding the conditional bias, the MSSS can be evaluated from $[2R(S_f/S_o)] - (S_f/S_o)^2$. When either the correlation or the ratio of the standard deviations is null, there is no skill improvement compared to the reference forecast. This also holds true for negative correlations. The ratio of the standard deviations indicates that models have stronger underestimation ($S_f/S_o < 1$) for longer leads compared to weeks 1–2 (Fig. 5b). Since the MSSS measures both the linear association and the relationship between the magnitude of the forecasted and observed anomalies, weak positive correlations in many regions (Fig. 5a) and/or underestimation of the magnitude of the anomalies leads to large negative MSSS.

b. Probabilistic verification

Probabilistic skill is assessed through the RPSS$_D$ displayed in Fig. 6. When not accounting for the small ensemble size, RPSS is negative in most regions (Fig. S1 in the online supplemental material). Negative RPSS$_D$ over climatologically dry regions indicates that tercile distribution can be skewed when the lower boundary is not well defined. RPSS$_D$ is positive in most regions and lead times for ECMWF, whereas the other models have a more mixed signal, with particularly strong negative RPSS$_D$ over some regions. While ECMWF shows positive skill over wide regions at all lead times, UKMO and NCEP generally have limited skill beyond week 1, but UKMO maintains relatively high skill over East Africa for starts in MAM and SON (Figs. 6b,d) and NCEP over West Africa in JJA (Fig. 6c). Counterintuitively, UKMO has poorest skill near DRC during weeks 1–2 compared to subsequent weeks, which is also evident to some extent in the MSSS assessment (Fig. 4), and deserves additional investigation.

The ability of probabilistic forecasts to discriminate heavier precipitation events is shown in Fig. 7 through ROC diagrams for the above-normal category using grid points over different African regions. Good discrimination is found when there is high hit rate combined with low false alarm rate. For example, when above-normal rainfall over EEA in SON is forecast with 60% of probability (square marker in Fig. 7), forecasts of above-normal rainfall for week 1 result in a 40%, 53%, and 48% hit rate against a 16%, 20%, and 23% false alarm rate for ECMWF, UKMO, and NCEP, respectively. In contrast, little differences between hit and false alarm rates for the same threshold indicate forecasts with limited value in week 4. Thus, ROC diagrams can provide support to forecast users to trigger advisory action in the decision-making process.

The reduction in discrimination from weeks 1–2 to the following weeks (Fig. 7) is consistent with the reduction of forecast quality seen in other metrics. Discrimination is slightly better for ECMWF/UKMO over NCEP and EEA over other regions, particularly in weeks 1–2, with AUC showing scores around 0.7 in week 1. A ROC score of 0.7, for instance, indicates

that 70% of forecasts have higher probabilities of falling in the above-normal category when above-normal precipitation occurs compared to when it does not occur. ROC scores near 0.5 indicate the model cannot adequately distinguish between different outcomes. This provides worthless random classifications after two weeks lead for most regions.

Figure 8 shows the AD for the above-normal category using grid points over the same regions analyzed in Fig. 7. Models have better reliability and resolution in weeks 1–2 than weeks 3–4, as shown by colored lines closer to the solid diagonal line and farther from the horizontal line. Indeed, only ECMWF forecasts for week 1 fall into the zone of enhanced skill, notably in EEA and SA. In general, ECMWF has slightly better reliability and resolution than UKMO and NCEP in the two highest bins at weeks 1–2. For weeks 3–4, models show roughly similar features, though such comparable results are less apparent in the below-normal category for EEA and EWA (Fig. S4). Probabilistic forecasts can be marginally useful up to week 3 for most regions and even week 4 in the below-normal category. Such forecasts may have usefulness for decision-making as they are close to the no-skill line and the slope of the colored lines is still positive (Weisheimer and Palmer 2014).

Overconfidence is a noticeable feature in all weeks confining higher (lower) probability below (above) the perfect reliability, which means that an event conditioned on a forecast probability of 70% is verified only about 50% of the time, but an event conditioned on a forecast probability of 10% is verified about 20% of the time, for example. Mean forecast probabilities are slightly higher than the mean observed frequency for extreme tercile categories in the ECMWF and UKMO models (not shown). This difference is more pronounced for NCEP and it corroborates the lowest reliability found among models (Fig. 8, Fig. S4). The reliability could be improved by employing calibration, especially after week 2 when the reliability is reduced. For the near-normal category, all models have lower mean forecast probabilities than the mean observed frequency (not shown), and have no resolution (Fig. S5). The histograms present sharper forecasts in weeks 1–2 compared to longer leads, with some cases showing a U-shaped pattern concentrating high frequencies close to the highest and lowest bins. Sharpness drops with increasing lead time with maximum frequencies appearing around climatological frequency (0.2–0.4).

4. Drivers modulation of forecast quality

The previous two sections show that models have best subseasonal forecasting performance for 1–2 weeks ahead, with ECMWF overall more skillful than UKMO and NCEP. Here, the link between weekly African precipitation forecast quality and important large-scale drivers, such as ENSO, IOD, and the MJO is investigated using ECMWF hindcasts during 1997–2014. The observed characteristics of weekly African precipitation variability linearly associated with those drivers are illustrated by regressing out observed rainfall anomalies with weekly mean of observed drivers' indices. Only regression coefficients for week 1 based on starts in DJF,
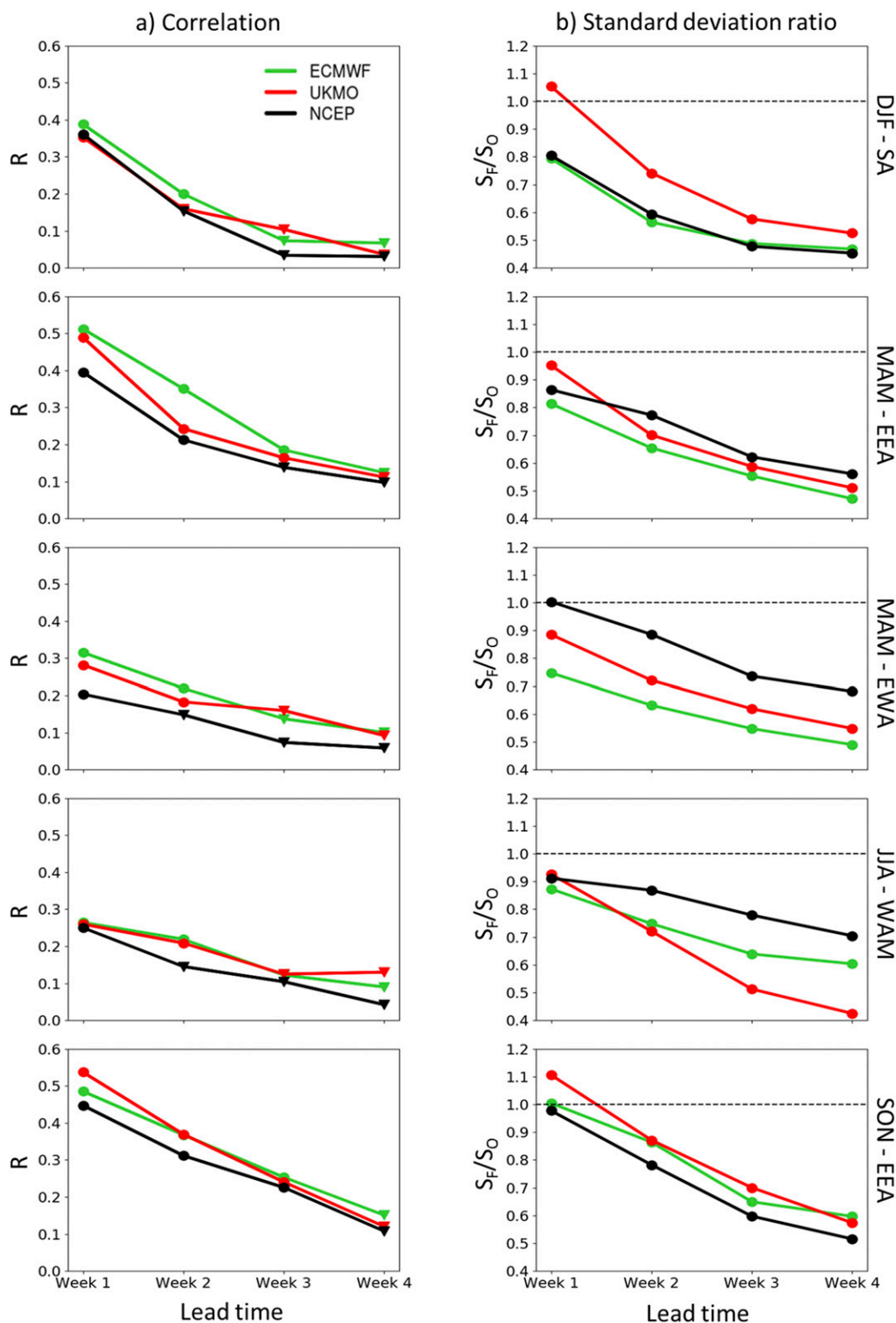
FIG. 5. Regional average of the (a) correlation (R) between the hindcast ensemble mean and observed precipitation anomalies and (b) ratio of the standard deviations of the hindcast ensemble mean and observations (S_F/S_O) over African regions (Fig. 1) for ECMWF (green line), UKMO (red line), and NCEP (black line) models in weeks 1–4 for initializations during DJF, MAM, JJA, and SON over the 1999–2010 period. Circle markers in (a) denote correlation coefficients statistically significant at the 95% level.
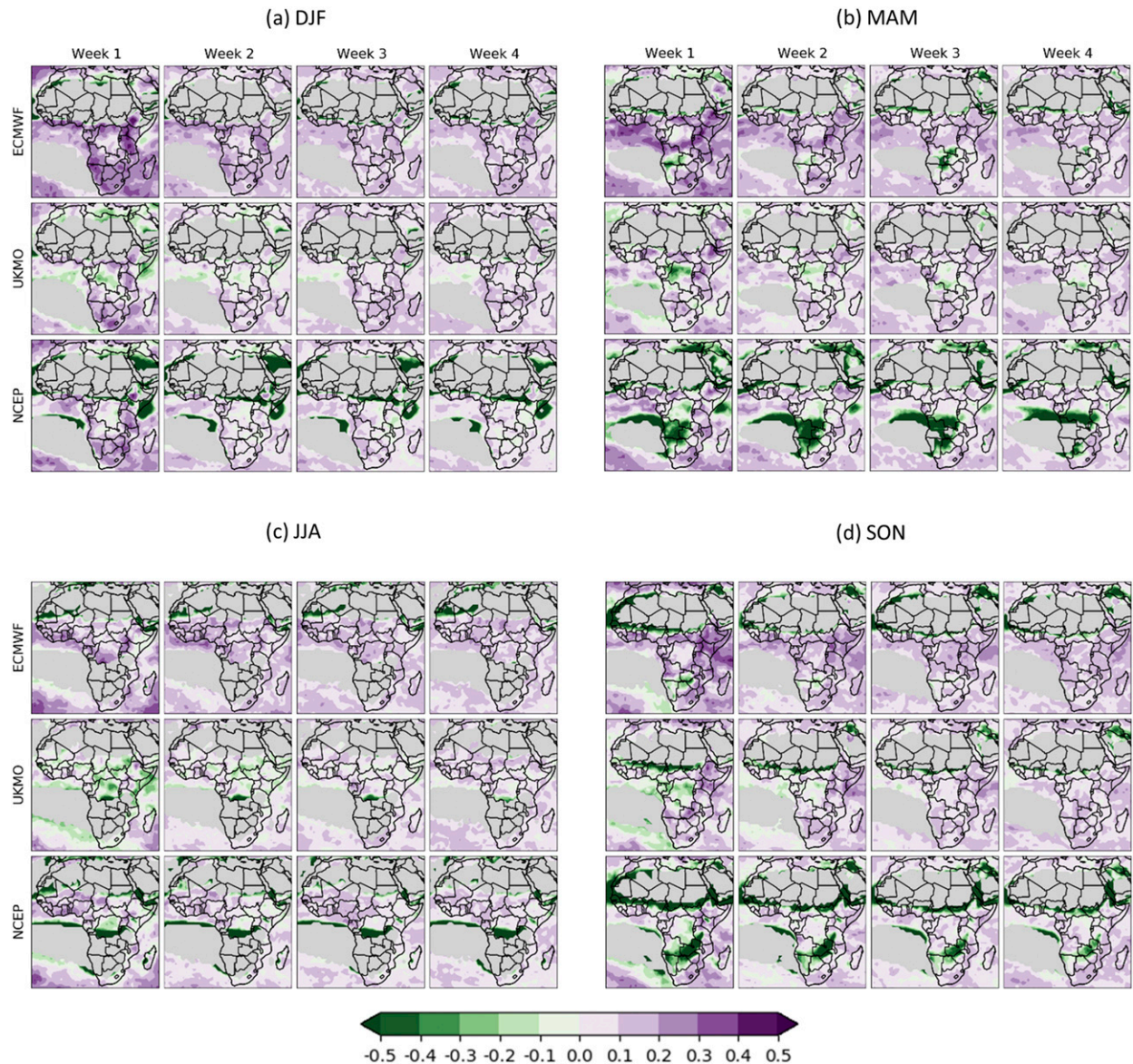
FIG. 6. Discrete ranked probability skill score (RPSS$_D$) between hindcast probabilities and binary observation obtained from precipitation totals in the tercile categories for ECMWF, UKMO, and NCEP models in weeks 1–4 for initializations during (a) DJF, (b) MAM, (c) JJA, and (d) SON over the 1999–2010 period. Gray shading as described in Fig. 2. A climatological probability of 1/3 was used as the reference forecast.

MAM, JJA, and SON are shown in Fig. 9, as they are roughly similar in subsequent weeks.

Weekly ENSO-related rainfall variability is more pronounced over East/Southeastern Africa in DJF compared to other seasons (Fig. 9), with positive (negative) anomalies over East (Southeastern) Africa. Additionally, negative (positive) anomalies in West Africa/Sahel (East Africa) are associated with El Niño influence during JJA (SON). IOD generally starts developing in JJA and reaches its maturity in SON before dissipating around December (Cai et al. 2018). Therefore, the weak regression coefficients in DJF and MAM are likely not related to this driver. A positive relationship is verified between

IOD and rainfall over Sahel in JJA and East Africa in SON. The latter shows the most striking relations between IOD and rainfall, with increasing (decreasing) East African precipitation during positive (negative) phases of the driver. Because there is significant correlation between ENSO and IOD indices (Table 2), it is likely that the regression patterns for these drivers include some signal from the other driver, and when accounting for their combined effects in the subsequent analysis, a multiple linear regression is used. Differences between simple and multiple regression patterns are most noticeable in SON, with the latter showing in particular a weaker positive precipitation signal associated with ENSO over East Africa (not shown).
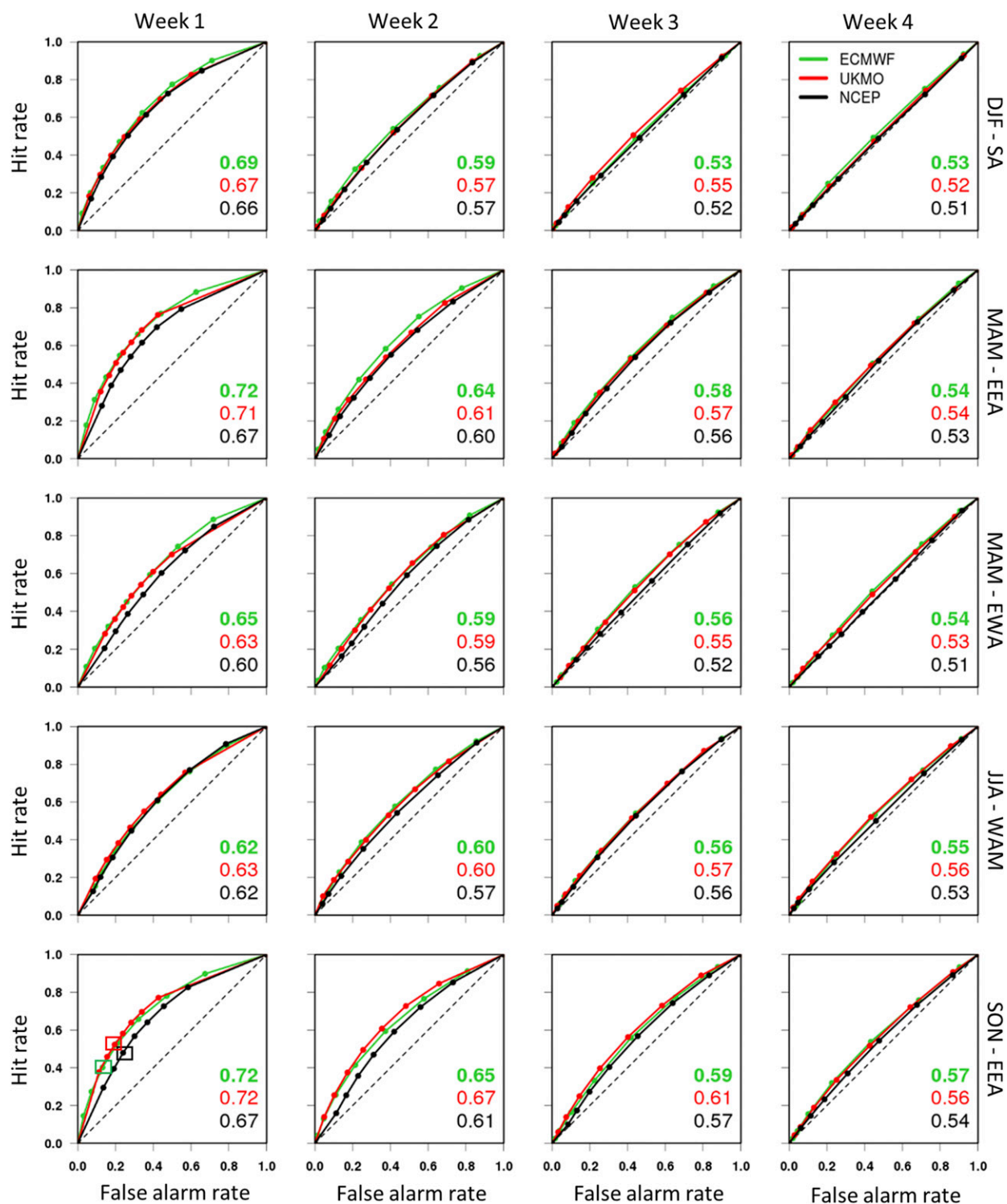
FIG. 7. Relative operating characteristic (ROC) diagram between hit and false alarm rates computed using hindcast probabilities and binary observation obtained from precipitation totals in the above-normal category over African regions (Fig. 1) for ECMWF (green line), UKMO (red line), and NCEP (black line) models in weeks 1–4 for initializations during DJF, MAM, JJA, and SON over the 1999–2010 period. The diagonal line is the line of no discrimination. Circle markers indicate the probability thresholds (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, and 0.1) varying from higher values at bottom to lower values at top. Square markers for SON-EEA in week 1 represent the corresponding 0.6 threshold for each model (see text for details). Colored numbers denote the area under the curve (AUC) for each model.
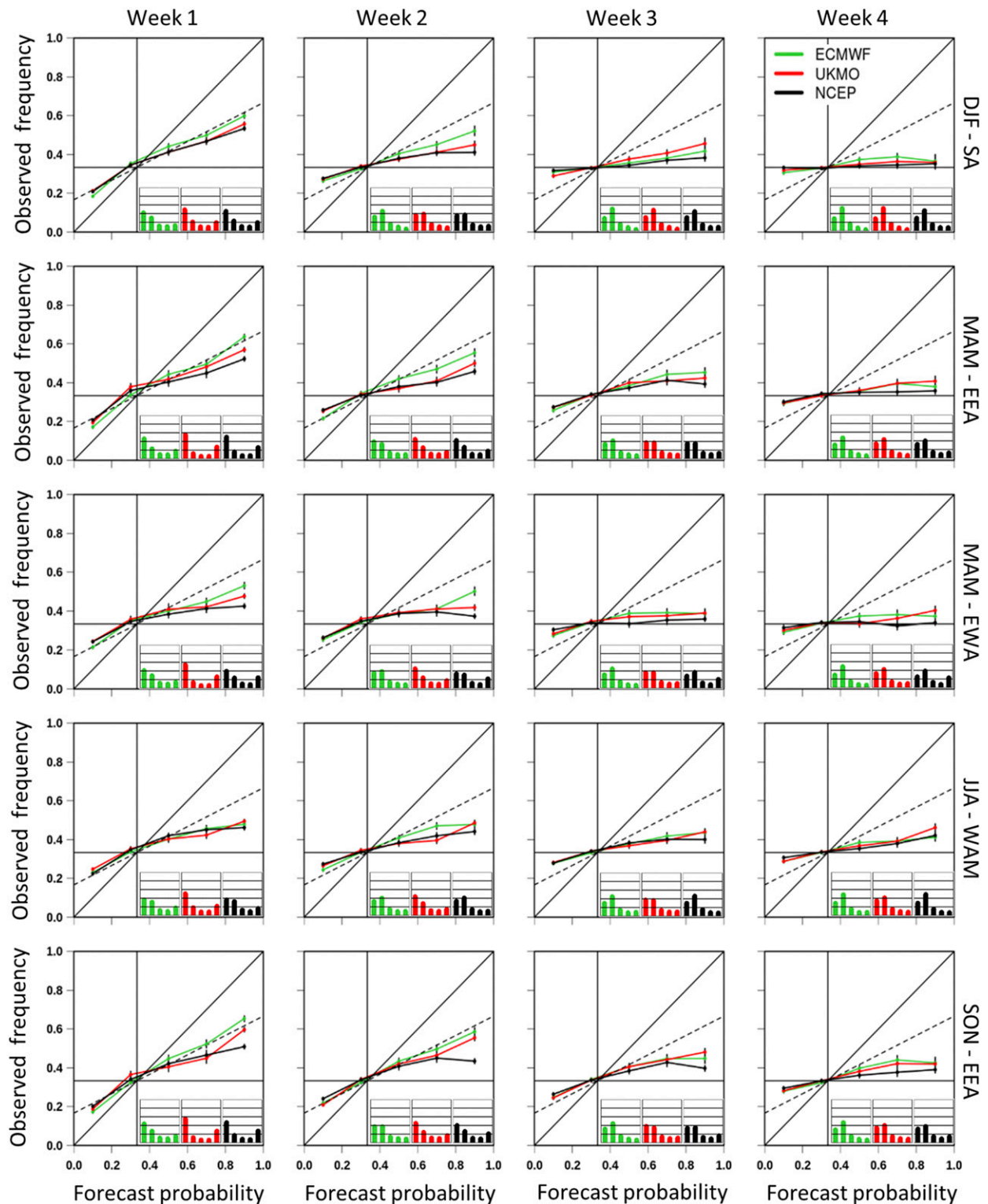
FIG. 8. Attributes diagram (AD) between forecast probability and observed frequency computed using hindcast probabilities and binary observation of precipitation totals in the above-normal category over African regions (Fig. 1) for ECMWF (green line), UKMO (red line), and NCEP (black line) models in weeks 1–4 for initializations during DJF, MAM, JJA, and SON over the 1999–2010 period. The diagonal (horizontal) line indicates perfect reliability (no resolution) and the dashed line is the no-skill line. The vertical line replicates the horizontal line. Histograms are divided into five probability bins (0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1) plotted in ascending order from the left to the right side against the frequency that each forecast probability was issued [ordinate; interval (horizontal grid lines) is the same as for the observed frequency]. Error bars denote the 95% confidence intervals estimated from 1000 bootstrap samples obtained from the available samples.
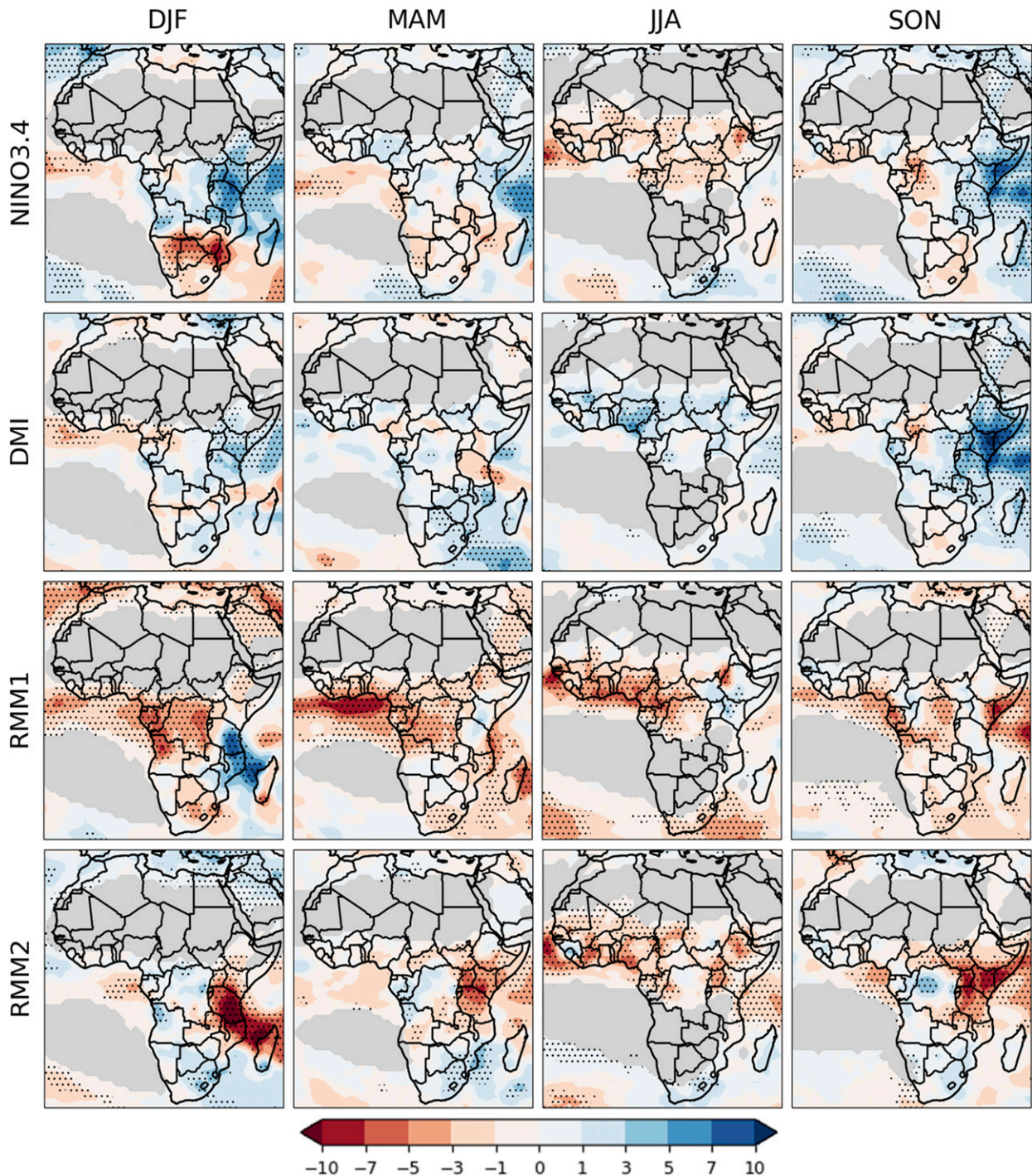
FIG. 9. Simple linear regression between observed weekly precipitation anomalies and weekly mean of Niño-3.4, DMI, and RMM (RMM1 and RMM2 components) observed indices in week 1 for start dates in DJF, MAM, JJA, and SON over the 1997–2014 period. Regression coefficients statistically significant at the 95% level are stippled. Units are accumulated millimeters per week. Indices are normalized by their corresponding standard deviations. Gray shading as described in Fig. 2.

In DJF, RMM1/RMM2 are related to large precipitation variations over Southeastern Africa (Fig. 9). In MAM, large regression coefficients are slightly displaced to the north compared to DJF, showing significant associations between rainfall and different MJO phases in EWA (RMM1) and during the East African long rains (RMM2). The boreal summer (JJA) is characterized by the MJO influence on WAM, highlighting strong rainfall variability near the GoG and on

TABLE 2. Correlations between the observed ENSO, IOD, and MJO indices (Niño-3.4, DMI, RMM1, and RMM2) in week 1 for start dates in DJF, MAM, JJA, and SON over the 1997–2014 period. Correlations are roughly similar to the ones found in weeks 2–4. Correlation coefficients statistically significant determined from a two-sided Student's *t* test at the 95% level are shown in bold. Effective sample size was estimated as in section 2c.

	DJF	MAM	JJA	SON
Niño-3.4 × DMI	**0.17**	−0.06	**0.23**	**0.60**
Niño-3.4 × RMM1	−0.03	−0.02	−0.02	**−0.41**
Niño-3.4 × RMM2	−0.06	**−0.20**	0.22	**−0.35**
DMI × RMM1	0.08	−0.03	−0.11	**−0.38**
DMI × RMM2	−0.03	0.06	−0.12	**−0.22**
RMM1 × RMM2	−0.11	−0.12	**0.15**	**0.25**

westernmost countries, particularly for RMM1. The MJO-related rainfall variations on the East African short rains are verified over central-southern (easternmost) region when regressing RMM2 (RMM1) with precipitation in SON.

Figure 10 shows the regional average of the absolute difference between regression coefficients of hindcast precipitation anomalies with forecasted drivers' indices and the corresponding observed regression coefficients over African regions (Fig. 1) during weeks 1–4. Largest differences are linked to the RMM2 for most regions, except over SA in DJF. These differences are more pronounced in EEA and WAM,

where precipitation variability is more closely related to RMM2 compared to other regions (Fig. 9). Larger discrepancies are also verified either for RMM1 or DMI over the same regions compared to Niño-3.4. Notwithstanding, rainfall anomalies over EEA in MAM are only weakly associated with RMM1 (Fig. 9) and IOD is usually inactive. Overall, ENSO signal is not well captured over SA compared to other relevant drivers in DJF, especially after week 1. Moreover, absolute differences increase with lead time in EEA during SON, which may affect subseasonal short-rains predictions. When considering errors relative to the observed regression patterns, i.e., dividing the absolute differences by the corresponding observed rainfall response to the driver, the errors are more balanced, except for DMI over EEA in DJF and RMM2 over EWA in MAM (not shown). Forecasted regression patterns suggest that largest errors in Fig. 10 are related to model's shortcomings in representing both the location and amplitude of particular driver-related rainfall anomalies (Figs. S6 and S7).

The modulation of subseasonal African precipitation forecast quality by the strength of the drivers' teleconnections within the ECMWF model is investigated in Fig. 11. This association is assessed by the regional average of the correlation between observations and forecasts after removing the observed and modeled regression patterns, computed between the corresponding precipitation anomalies and drivers' indices, from observed and predicted fields, respectively.
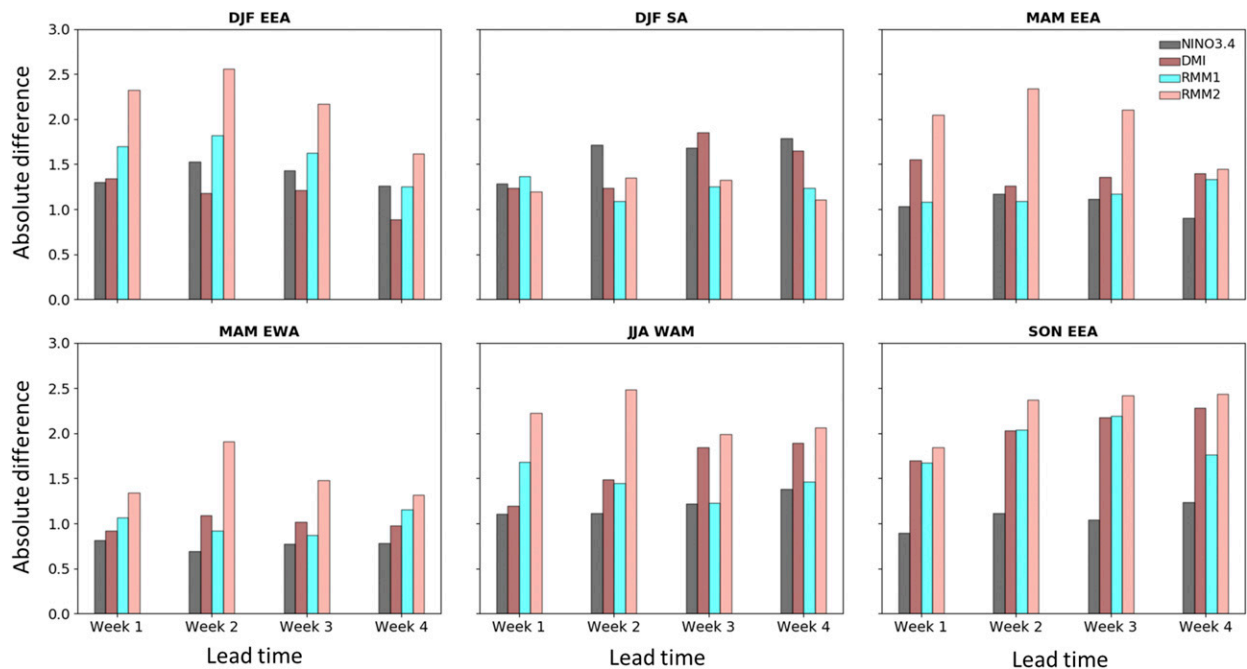

FIG. 10. Regional average of the absolute difference between the linear regression of the ECMWF hindcast ensemble mean precipitation anomalies with forecasted drivers' indices and the corresponding observed regression coefficients over African regions (Fig. 1) during weeks 1–4 for initializations in DJF, MAM, JJA, and SON over the 1997–2014 period. Observed (forecasted) regression coefficients were calculated using observed (forecasted) indices and observed (forecasted) precipitation anomalies. A multiple linear regression approach was employed to assess ENSO and IOD signal on rainfall simultaneously (see text for details). Indices were normalized by their corresponding standard deviations. Units are accumulated millimeters per week.
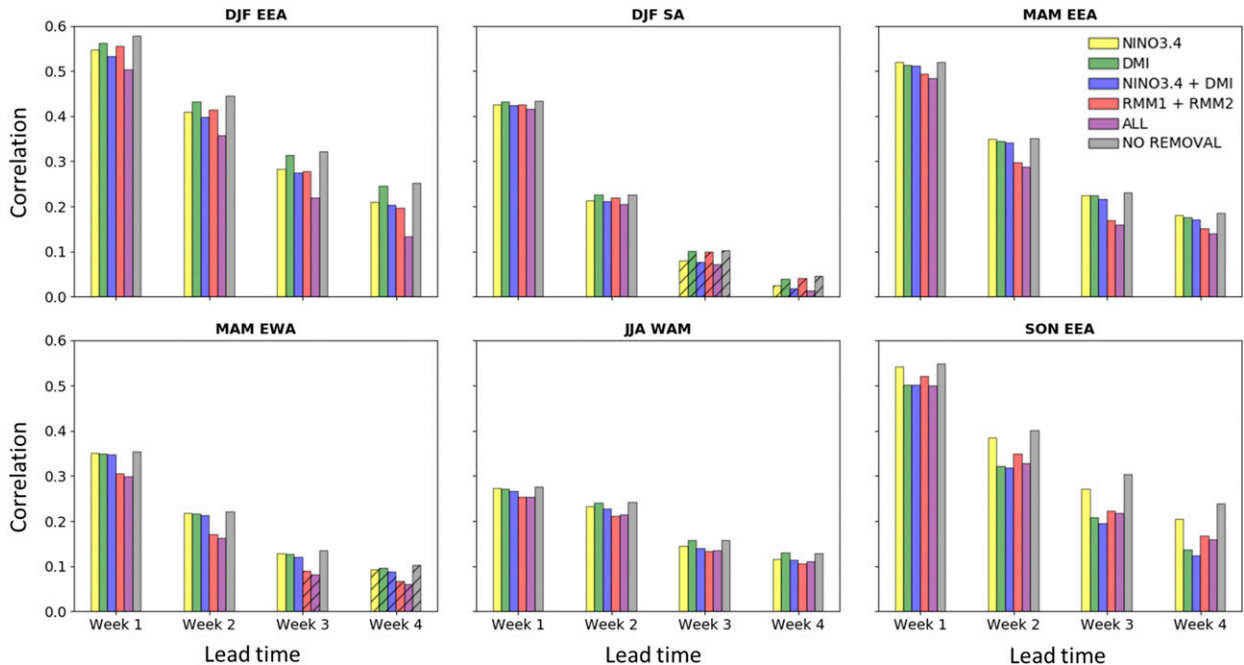
FIG. 11. Regional average of the correlation between the ECMWF hindcast ensemble mean and observed precipitation anomalies over African regions (Fig. 1) during weeks 1–4 for initializations in DJF, MAM, JJA, and SON over the 1997–2014 period. Correlations were obtained after removing particular observed and forecasted regression patterns (colored bars), calculated between the corresponding precipitation anomalies and drivers' indices, from observations and hindcasts, respectively. A multiple linear regression approach was employed to assess ENSO and IOD signal on rainfall simultaneously (see text for details). "ALL" denotes that the correlation was computed after subtracting sequentially all drivers' regression patterns. "NO REMOVAL" indicates that the correlation was obtained without removing any regression pattern. Hatches over the bars denote correlation coefficients are not statistically significant at the 95% level.

Lowest correlations are verified when all driver-related regression patterns are sequentially subtracted compared to when no removal is considered. This difference is more pronounced for longer leads and particular regions, such as EEA in DJF/SON. The impact of removing both ENSO and IOD signals is more noticeable in DJF and SON over EEA than in other seasons and regions, with larger ENSO (IOD)-related effects in the former (latter) season. Intriguingly, the impact of removing the IOD (or IOD+ENSO) signal in SON affects forecast association more than the impact of removing all drivers. This could be related to the fact that all indices are significantly correlated during the period under consideration (Table 2), which means may be removing the same signal in different ways. It is well known that ENSO and IOD are interannual modes of variability and their patterns project onto the RMM index, in particular for ENSO (Wheeler and Hendon 2004). When computing the RMM index, the methodology subtracting the previous 120 days is supposed to account for removing low-frequency variations (see section 2d). However, such approach will not necessarily be effective in the drivers' developing and decaying phases as the last 120 days may not include their signatures. Thus, the correlations between ENSO/IOD indices and the MJO index are not physically easy to interpretate and it may not be fair to perform a multiple linear regression including all indices without underlying physical understanding. Additionally, the correlations between RMM

components are nonzero likely because the two eigenvectors were calculated over all seasons and the correlations in individual seasons might not be null.

Forecast quality in MAM/JJA is more affected by subtracting the MJO-related rainfall variability than other drivers (Fig. 11). This would be expected since ENSO and IOD are usually weak or inactive during those seasons. When assessing the impacts of removing the MJO-related rainfall variability individually, lowest correlations are found for all weeks after subtracting RMM2 signal over EEA in MAM and WAM in JJA (not shown). In contrast, RMM1-related rainfall variability has a more significant association with forecast quality in the first two weeks over EWA in MAM (not shown). Although a different period has been analyzed in section 3a (1999–2010), correlations for ECMWF in Fig. 3 could be linked to the sources of subseasonal predictability examined here, with large associations in regions where those drivers have strong linear relationships with precipitation (Fig. 9).

To further explore drivers' signals on forecast quality, Fig. 12 displays the regional average of the correlation between observations and forecasts after adding the corresponding observed regression patterns to hindcasts, that is replacing the modeled linear response to the driver with the observed response to the driver. The general picture is that a clear improvement in forecast quality is shown if all observed driver-related regression patterns are added to hindcasts and compared to the "NO ADDITION" case (i.e., using
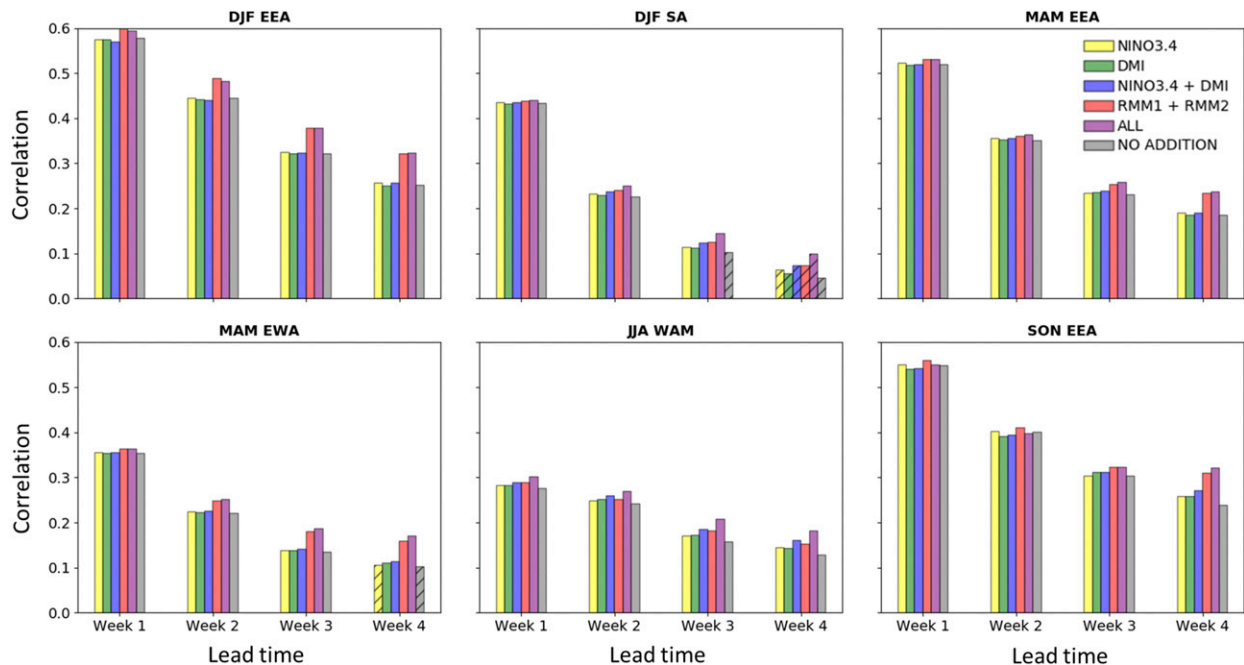
FIG. 12. As in Fig. 11, but for correlations obtained after adding particular observed regression patterns (colored bars) to hindcasts analyzed in Fig. 11. ''ALL'' denotes that the correlation was computed after adding sequentially all drivers' regression patterns. Gray bars are equivalent to those in Fig. 11, with no removal or addition of any regression pattern to observations and hindcasts.

uncalibrated forecasts), especially in weeks 3–4. These enhanced associations mostly respond to the MJO-related rainfall variability, in particular owing to RMM2 signals (not shown), but there is also an improvement in association in response to ENSO and IOD over EEA in SON. This may indicate that better subseasonal predictions of specific MJO phases and its teleconnections could help improve the quality of weekly rainfall forecasts over most regions analyzed here.

5. Summary and conclusions

This study has conducted an evaluation of the quality of subseasonal precipitation forecasts over Africa and examined its relationships with particular climate drivers. A comprehensive assessment of forecasts depends on how well models represent the attributes of forecast quality defined in Murphy (1993). We initially investigated weekly accumulated African precipitation forecast quality using hindcasts provided by three S2S models (ECMWF, UKMO, and NCEP) and precipitation from the GPCP dataset. Start dates within DJF, MAM, JJA, and SON were selected to assess forecasts from one to four weeks ahead during 1999–2010. Deterministic and probabilistic forecasts were evaluated employing a variety of metrics to provide a more detailed assessment. Then, weekly precipitation forecast quality was linked to key drivers (ENSO, IOD, and the MJO) by exploring the ECMWF model's ability in representing drivers' signals on African precipitation and their contribution to the quality of forecasts during 1997–2014.

The deterministic evaluation indicated significant correlations greater than 0.4 between hindcasts and observations for

all models in weeks 1–2 over East Africa in DJF/MAM/SON and near GoG in JJA. This corroborates best MSSS findings, in which skill in weeks 1 and 2 was improved up to 70% and 50% relative to the reference forecast, respectively. Further investigation of this correspondence was provided by decomposing the MSSS, revealing unskillful predictions linked to low forecast association and/or large underestimation of predicted variance. Analysis of bias indicated a large overestimation (underestimation) in wet regions during particular rainy seasons for ECMWF and UKMO (NCEP), though over WAM in JJA models showed a similar bias pattern with a meridional tripole structure.

The evaluation of probabilistic forecasts showed large deficiencies in the near-normal category. Low forecast quality in this category has been related to the fact that such forecasts deviate very little from tercile-based climatological probability (Kharin and Zwiers 2003b). The consequences of issuing poor forecast quality in the near-normal category can be very harmful for forecasters and users. Leading, for example, to increased uncertainty in any tercile-based forecast information and reduced effectiveness of such information in decision-making. Thus, some operational forecasting centers assign the climatological probability to the near-normal category and issue outlooks for the most likely outer tercile category (Peng et al. 2012). Erroneous forecasts in the near-normal category indicate the need to review the scientific knowledge and develop improved methods of estimating probabilities (Kharin and Zwiers 2003b).

One the other hand, more skillful forecasts with roughly similar characteristics were identified in the outer tercile

categories. These forecasts showed better discrimination over EEA compared to other regions, particularly in weeks 1–2. AUC could quantitatively summarize models' performance to discriminate extreme events. For example, ECMWF correctly predicted around 70% (65%) of above-normal forecasts when above-normal rainfall occurred in EEA during the first (second) week of forecasts. Nevertheless, this agreement reduced to less than 60% of forecasts in subsequent weeks, indicating forecasts with limited value to forecasters and decision-makers. Despite having found better reliability, resolution, and sharpness in weeks 1–2, with slightly enhanced skill for ECMWF over EEA/SA in week 1, overconfidence was verified in all weeks, showing probabilities closer to the climatological distribution for longer lead times. Since models' probabilistic skill can be associated with other attributes, such as reliability and resolution, it is suggested that overconfidence has increased forecasting errors, inducing more unskilled forecasts, especially beyond two weeks lead, as verified in the $RPSS_D$ assessments.

One aspect of the forecast verification we have not addressed is the relation between metrics analyzed and its practical implications for forecasting routines. In terms of deterministic forecasts, forecasters would judge how skillful forecasts are by relating the MSSS to correlation and the ratio of the forecasted to observed variances. Forecast quality would be determined by assessing the overall balance between those metrics, with high correlation and small variance errors indicating more skillful forecasts. For probabilistic forecasts, skillful outcomes could be identified by relating $RPSS_D$ to the overall balance between reliability and resolution. Forecasters would identify more skillful forecasts when model's accuracy is large owing to more reliable forecasts and improved resolution. AUC would provide similar qualitative information as the resolution assessment (Toth et al. 2003).

When assessing the ability of the ECMWF model in representing particular climate drivers' signals on regional African rainfall, it was found larger errors in capturing rainfall variations linearly related to the MJO-RMM2 index over most regions compared to other indices (RMM1, DMI, Niño-3.4). This suggests that the model does not reproduce the local impacts of the MJO properly and in particular those phases associated with RMM2 (i.e., 2 and 3; 6 and 7). Shortcomings in simulating driver-related rainfall variability could affect subseasonal predictions of important weather systems influencing African rainfall, as, for instance, ITCZ and tropical cyclones.

To analyze weekly forecast quality linked to the strength of drivers' teleconnections, regional correlations between observations and hindcasts were calculated after removing the corresponding driver-related rainfall regression patterns from observations and forecasts. When removing all drivers' signals sequentially, results showed significant reduction in association compared to when no subtraction was considered. The removal of regression patterns individually indicated that the MJO contribution to forecast quality was more dominant during seasons when ENSO and IOD are usually inactive, such as MAM/JJA. Although ENSO is expected to be correlated with EEA rainfall during the short rains season (e.g., Hoell et al. 2014), enhanced associations were particularly linked to IOD. A multiple linear regression analysis revealed that a large

portion of the ENSO signal on EEA rainfall during SON can be attributed to the IOD.

It is worth noting that even verifying forecast quality closely related to ENSO and IOD during DJF and SON, respectively, the effect of calibrating forecasts by adding observed regression patterns to hindcast revealed improved forecast associations especially linked to the MJO. Despite identifying significant associations, the drivers analyzed could not account completely for the overall forecast quality. The fact the significant correlations remained after removing ENSO and IOD effects indicates that the quality of forecasts does not depend solely on these interannual modes of variability. Furthermore, while as significant contributor to the forecast quality the MJO is not the only important source of subseasonal predictability. This suggests there is a need for assessing other drivers, including but not limited to, SST variability over the GoG and soil moisture initializations. However, our results still support that forecast quality of weekly rainfall over Africa is regime-dependent, i.e., related to the major tropical sources of S2S predictability. Moreover, it is clear that improving the representation of these drivers—and their regional impacts—within the ECMWF model has the potential to deliver better subseasonal predictions for Africa.

This paper investigated single models when developing a weekly precipitation forecast verification framework for Africa. Combining forecasts from a multimodel perspective may help to improve resolution and discrimination of predictions (e.g., Vigaud et al. 2018). Different calibration methods, such as model output statistic (e.g., Doss-Gollin et al. 2018), should be explored to identify which is the best one practice to be employed for delivering more reliable forecasts to operational centers and applications communities. Although these aforementioned techniques have not been adopted here, this comprehensive verification guide for forecasting weekly rainfall across Africa provides a valuable tool for forecasters and decision-makers to better understand the African regions and seasons with useful subseasonal skill. Furthermore, the results linking known S2S drivers with forecast quality in this study have huge potential to assist forecasters to better interpret regime-dependent skill, which, if successfully communicated, can increase confidence in its appropriate use in decision-making across a range of sectors and societal applications.

REFERENCES

Allen, M. P., Ed., 1997: The t test for the simple regression coefficient. *Understanding Regression Analysis*, Springer, 66–70, https://doi.org/10.1007/978-0-585-25657-3_14.

Bamston, A., M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific: Research note. *Atmos.–Ocean*, **35**, 367–383, https://doi.org/10.1080/07055900.1997.9649597.

Behera, S. K., J. J. Luo, S. Masson, P. Delecluse, S. Gualdi, A. Navarra, and T. Yamagata, 2005: Paramount impact of the Indian Ocean dipole on the East African short rains: A CGCM study. *J. Climate*, **18**, 4514–4530, https://doi.org/10.1175/JCLI3541.1.

Cai, W., and Coauthors, 2018: Stabilised frequency of extreme positive Indian Ocean Dipole under 1.5°C warming. *Nat. Commun.*, **9**, 1419, https://doi.org/10.1038/s41467-018-03789-6.

Coelho, C. A. S., M. A. F. Firpo, and F. M. de Andrade, 2018: A verification framework for South American sub-seasonal precipitation predictions. *Meteor. Z.*, **27**, 503–520, https://doi.org/10.1127/metz/2018/0898.

——, B. Brown, L. Wilson, M. Mittermaier, and B. Casati, 2019: Forecast verification for S2S timescales. *Sub-Seasonal to Seasonal Prediction: The Gap between Weather and Climate Forecasting*, F. Vitart and A. Robertson, Eds., Elsevier, 337–361.

de Andrade, F. M., C. A. S. Coelho, and I. F. A. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, https://doi.org/10.1007/s00382-018-4457-z.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Domeisen, D. I., and Coauthors, 2020: The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. *J. Geophys. Res. Atmos.*, **125**, e2019JD030923, https://doi.org/10.1029/2019JD030923.

Doss-Gollin, J., Á. G. Muñoz, S. J. Mason, and M. Pastén, 2018: Heavy rainfall in Paraguay during the 2015/16 austral summer: Causes and subseasonal-to-seasonal predictive skill. *J. Climate*, **31**, 6669–6685, https://doi.org/10.1175/JCLI-D-17-0805.1.

Gottschalck, J., and Coauthors, 2010: A framework for assessing operational Madden–Julian Oscillation forecasts: A CLIVAR MJO working group project. *Bull. Amer. Meteor. Soc.*, **91**, 1247–1258, https://doi.org/10.1175/2010BAMS2816.1.

Hoell, A., C. Funk, and M. Barlow, 2014: La Niña diversity and northwest Indian Ocean rim teleconnections. *Climate Dyn.*, **43**, 2707–2724, https://doi.org/10.1007/s00382-014-2083-y.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.*, **2**, 36–50, https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

Kharin, V. V., and F. W. Zwiers, 2003a: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

——, and ——, 2003b: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701, https://doi.org/10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2.

Kolstad, E. W., 2019: Subseasonal prediction of Idai and other tropical cyclones and storms in the Mozambique channel. *ESSOAr*, https://doi.org/10.1002/essoar.10501336.1, in press.

Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889, https://doi.org/10.1175/MWR-D-14-00277.1.

Liebmann, B., and C. A. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.*, **77**, 1275–1277.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

Lo, F., and H. H. Hendon, 2000: Empirical extended-range prediction of the Madden–Julian oscillation. *Mon. Wea. Rev.*, **128**, 2528–2543, https://doi.org/10.1175/1520-0493(2000)128<2528:EERPOT>2.0.CO;2.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**, 1513–1523, https://doi.org/10.1175/JCLI3361.1.

Murphy, A. H., 1972: Scalar and vector partitions of the ranked probability score. *Mon. Wea. Rev.*, **100**, 701–708, https://doi.org/10.1175/1520-0493(1972)100<0701:SAVPOT>2.3.CO;2.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor. Climatol.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Peng, P., A. Kumar, M. S. Halpert, and A. G. Barnston, 2012: An analysis of CPC's operational 0.5-month lead seasonal outlooks. *Wea. Forecasting*, **27**, 898–917, https://doi.org/10.1175/WAF-D-11-00143.1.

Ratnam, J. V., S. K. Behera, Y. Masumoto, and T. Yamagata, 2014: Remote effects of El Niño and Modoki events on the austral summer precipitation of southern Africa. *J. Climate*, **27**, 3802–3815, https://doi.org/10.1175/JCLI-D-13-00431.1.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high-resolution-blended analyses for sea surface temperature. *J. Climate*, **20**, 5473–5496, https://doi.org/10.1175/2007JCLI1824.1.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata, 1999: A dipole mode in the tropical Indian Ocean. *Nature*, **401**, 360–363, https://doi.org/10.1038/43854.

Shonk, J. K., T. D. Demissie, and T. Toniazzo, 2019: A double ITCZ phenomenology of wind errors in the equatorial Atlantic in seasonal forecasts with ECMWF models. *Atmos. Chem. Phys.*, **19**, 11 383–11 399, https://doi.org/10.5194/acp-19-11383-2019.

Sossa, A., B. Liebmann, I. Bladé, D. Allured, H. H. Hendon, P. Peterson, and A. Hoell, 2017: Statistical connection between the Madden–Julian oscillation and large daily precipitation events in West Africa. *J. Climate*, **30**, 1999–2010, https://doi.org/10.1175/JCLI-D-16-0144.1.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds.,Wiley, 137–163.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017a: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

——, ——, ——, and N. Acharya, 2017b: Subseasonal predictability of boreal summer monsoon rainfall from ensemble forecasts. *Front. Environ. Sci.*, **5**, 67, https://doi.org/10.3389/fenvs.2017.00067.

——, M. K. Tippett, and A. W. Robertson, 2018: Probabilistic skill of subseasonal precipitation forecasts for the East Africa–West Asia sector during September–May. *Wea. Forecasting*, **33**, 1513–1532, https://doi.org/10.1175/WAF-D-18-0074.1.

——, ——, and ——, 2019: Deterministic skill of subseasonal precipitation forecasts for the East Africa–West Asia sector from September to May. *J. Geophys. Res. Atmos.*, **124**, 11 887–11 896, https://doi.org/10.1029/2019JD030747.

Vitart, F., 2017: Madden–Julian oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220, https://doi.org/10.1002/qj.3079.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Washington, R., R. James, H. Pearce, W. M. Pokam, and W. Moufouma-Okia, 2013: Congo Basin rainfall climatology: Can we believe the climate models? *Philos. Trans. Roy. Soc. London*, **368B**, 20120296, https://doi.org/10.1098/rstb.2012.0296.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Zaitchik, B. F., 2017: Madden-Julian Oscillation impacts on tropical African precipitation. *Atmos. Res.*, **184**, 88–102, https://doi.org/10.1016/j.atmosres.2016.10.002.

Zhang, W., Y. Wang, F.-F. Jin, M. F. Stuecker, and A. G. Turner, 2015: Impact of different El Niño types on the El Niño/IOD relationship. *Geophys. Res. Lett.*, **42**, 8570–8576, https://doi.org/10.1002/2015GL065703.