Do Convection-Permitting Ensembles Lead to More Skillful Short-Range Probabilistic Rainfall Forecasts over Tropical East Africa?

CARLO CAFARO,[a] BETH J. WOODHAMS,[b] THORWALD H. M. STEIN,[a] CATHRYN E. BIRCH,[b] STUART WEBSTER,[c] CAROLINE L. BAIN,[c] ANDREW HARTLEY,[c] SAMANTHA CLARKE,[b] SAMANTHA FERRETT,[a] AND PETER HILL[a]

[a] *Department of Meteorology, University of Reading, Reading, United Kingdom*
[b] *University of Leeds, Leeds, United Kingdom*
[c] *Met Office, Exeter, United Kingdom*

(Manuscript received 22 September 2020, in final form 2 February 2021)

ABSTRACT: Convection-permitting ensemble prediction systems (CP-ENS) have been implemented in the midlatitudes for weather forecasting time scales over the past decade, enabled by the increase in computational resources. Recently, efforts are being made to study the benefits of CP-ENS for tropical regions. This study examines CP-ENS forecasts produced by the Met Office over tropical East Africa, for 24 cases in the period April–May 2019. The CP-ENS, an ensemble with parameterized convection (Glob-ENS), and their deterministic counterparts are evaluated against rainfall estimates derived from satellite observations (GPM-IMERG). The CP configurations have the best representation of the diurnal cycle, although heavy rainfall amounts are overestimated compared to observations. Pairwise comparisons between the different configurations reveal that the CP-ENS is generally the most skillful forecast for both 3- and 24-h accumulations of heavy rainfall (97th percentile), followed by the CP deterministic forecast. More precisely, probabilistic forecasts of heavy rainfall, verified using a neighborhood approach, show that the CP-ENS is skillful at scales greater than 100 km, significantly better than the Glob-ENS, although not as good as found in the midlatitudes. Skill decreases with lead time and varies diurnally, especially for CP forecasts. The CP-ENS is underspread both in terms of forecasting the locations of heavy rainfall and in terms of domain-averaged rainfall. This study demonstrates potential benefits in using CP-ENS for operational forecasting of heavy rainfall over tropical Africa and gives specific suggestions for further research and development, including probabilistic forecast guidance.

SIGNIFICANCE STATEMENT: Forecasting the location and timing of precipitation is challenging, especially in the tropics where most rainfall comes from convective systems. In the midlatitudes, convection-permitting ensembles (CP-EPS) have been shown to be beneficial to operational forecasting of precipitation, but only few studies have considered CP-EPS in the tropics. In this study of 24 forecasts over tropical East Africa, we find that CP-EPS have skill and are more skillful than deterministic CP forecasts and global ensembles in predicting the rainfall location and discriminating between events and nonevents. However, skill scores are lower than those found for CP-EPS in the midlatitudes. Further work should focus on improving ensemble spread, including for the global ensemble.

KEYWORDS: Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Ensembles; Model comparison

1. Introduction

In tropical Africa, unlike midlatitude locations, the main contribution to daily rainfall comes from deep convective systems (Fink et al. 2017). Dezfuli et al. (2017b) found, for instance, that convective events contribute to nearly three quarters of the total seasonal precipitation, even if they are rare. The dominance of convection makes rainfall forecasting in this region particularly challenging. The global models that are usually available to local forecasters rely on parameterization schemes to generate convection and are typically unable to reproduce the two main characteristics of precipitation, namely intensity and diurnal timing. Such parameterized convection models produce light rain too frequently, typically miss the heaviest rainfall events (e.g., Holloway et al. 2012) and tend to predict the afternoon peak of the convective rainfall too early (Bechtold et al. 2004). More recently, Vogel et al. (2018) suggested the parameterization of convection as the potential cause of the low skill by nine operation global ensemble prediction systems with respect to climatological forecasts for rainfall prediction in West Africa.

Increasing model resolution to achieve a 4-km horizontal grid spacing or less has proven to be beneficial for forecasting

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

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-20-0172.s1.

Corresponding author: Carlo Cafaro, c.cafaro@reading.ac.uk

DOI: 10.1175/WAF-D-20-0172.1

rainfall for two reasons: 1) convective clouds and their updrafts start to be explicitly resolved and 2) local topographic features (e.g., orography, coastlines, land surface properties) are represented in finer details, allowing better representation of their associated feedbacks on convection (Clark et al. 2016).

In West Africa, Pearson et al. (2014), Marsham et al. (2013), and Birch et al. (2014) showed that running limited area convection-permitting (CP) models with grid spacing between 12 and 1.5 km improved the initiation, propagation, and diurnal cycle of convective systems within the West African monsoon.

In 2011, to support the World Meteorological Organization (WMO) community of meteorological services in Africa, the Met Office began running operationally a 4.4-km CP deterministic model over tropical East Africa. Chamberlain et al. (2014) found that for a 2-month forecasting period for Lake Victoria, this CP model had better skill than the Met Office global model at predicting rainfall. More recently Woodhams et al. (2018) found that over a 2-year period the CP model outperformed the Met Office global parameterized convection model for rainfall prediction throughout East Africa, especially on subdaily time scales and for storms over land. In March 2019, the CP model was expanded to include the whole of West and East Africa, now referred to as the "Tropical Africa Model" (Hanley et al. 2021). Information from this model is freely available to meteorological services covered by the domain.

Due to lack of predictability at small spatiotemporal scales (Lorenz 1969; Hohenegger and Schar 2007), many forecasting centers in the midlatitudes use CP ensembles prediction systems for operational and research purposes (Gebhardt et al. 2011; Schwartz et al. 2015; Raynaud and Bouttier 2016; Hagelin et al. 2017). Several verification studies have shown the benefits of these CP ensembles, mainly for midlatitude regions including the United States (Snook et al. 2019; Schwartz 2019; Gowan et al. 2018; Schwartz and Sobash 2017), United Kingdom (Cafaro et al. 2019; Mittermaier and Csima 2017), northern and continental Europe (Frogner et al. 2019; Klasa et al. 2018; Pantillon et al. 2018; Schellander-Gorgas et al. 2017), and eastern China (Li et al. 2019).

Only a few studies have dealt with short-range CP ensemble over tropical regions and in particular Africa. Torn (2010) experimented with a CP model (4-km horizontal grid spacing) nested inside a mesoscale ensemble (36-km grid spacing) and found that forecasts of African easterly waves from the two ensembles had similar sensitivities to initial conditions, which included various perturbation and initialization times. Maurer et al. (2017) evaluated a CP-ENS using COSMO (2.8-km grid spacing) with land surface and atmosphere perturbations over West Africa. Their single-model setup [using analyses from the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble] was shown to have higher skill (reliability and sharpness) than the multimodel setup (using analyses from three different models) in predicting precipitation, although more underdispersive.

While previous studies have focused on case studies to test the benefits of CP ensemble in forecasting for tropical Africa, the use of CP ensemble in an operational set up has not yet been considered. As part of the African Science for Weather Information and Forecasting Techniques (SWIFT) project (https://africanswift.org/), the Met Office ran a CP ensemble prediction system for the first time in East Africa to support the forecasting testbed hosted by the Kenya Meteorological Department during April–May 2019. The aim of the testbed was to fill the gap between research and forecasting activities (e.g., Ralph et al. 2013). For instance, the Kenya Meteorological Department currently issue heavy rain warnings based on 24-h accumulations determined from parameterized convection forecasts, by using plots from the global Met Office, Global Forecast System (GFS) and ECMWF through the WMO Severe Weather Forecast Project (SWFP) (e.g., http://www.meteo.go.ke/pdf/Heavy\%20Rainfall\%20Alert\%2023rd\%20Jan-2020.pdf). CP deterministic and ensemble forecasts could allow for warnings with more spatial and temporal specificity.

In this paper, we compare and evaluate the CP and global ensemble forecasts over East Africa and consider the implications for operational use of CP ensemble when forecasting precipitation in Tropical Africa. The overarching question of this study is as follows: are CP ensemble forecasts more skillful than global ensemble and deterministic forecasts (both of which are less expensive to run and already operational for East Africa)? To address this question, a neighborhood based approach is applied to both ensembles and deterministic forecasts, after applying a threshold to the rainfall field. This approach allows us to evaluate the added skill in the CP ensemble due to the additional degree of smoothing provided by averaging across all the ensemble members compared to just applying the spatial averaging to the deterministic forecasts Using this approach for the United States, Schwartz et al. (2017) found that their 3-km CP ensemble outperformed the 1-km individual members and they attributed this to ensemble averaging filtering out noise from unpredictable scales. Such an evaluation has not previously been performed in a tropical region.

The paper is structured as follows: section 2 describes the forecasts and the observational data used for the analysis along with the methodology including the neighborhood approach. General characteristics of the forecasts (diurnal cycle, spread) are described in section 3. In section 4 we present the probabilistic verification, including the comparison of the CP ensemble against the deterministic and global forecasts. Additional spatial verification of the CP ensemble is provided in section 5, considering different skill metrics and ensemble postprocessing options to support future operational use. Conclusions and directions for future work are offered in section 6.

2. Data and methodology

a. Data

1) FORECASTS

The simulations supporting the SWIFT forecasting testbed were run from 19 April to 12 May 2019, giving a total number of 24 days. The Met Office Unified Model (MetUM) Tropical East Africa CP ensemble model (hereafter CP-ENS) was run as a downscaler of the of the global ensemble, similar to the

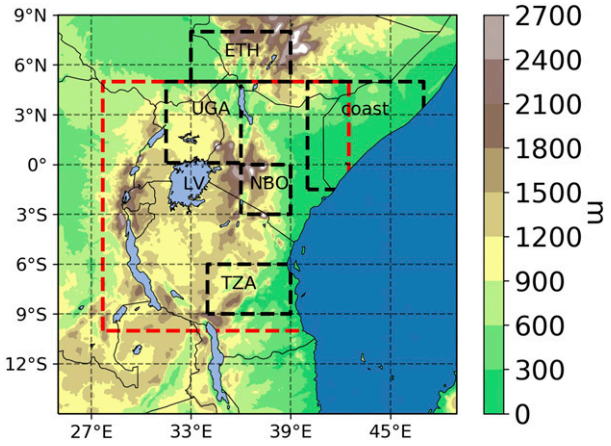

FIG. 1. A map showing the elevation for the domain spanned by the convection-permitting ensemble model for tropical East Africa. Black dashed boxes enclose the different subregions considered for regional differences in rainfall characteristics, including the Lake Victoria basin (LV). The red dashed box encloses the region used for calculating the fractions skill score (FSS). Ocean points are not considered.

set-up used by the Met Office CP model (MOGREPS-UK) up to March 2016 (Hagelin et al. 2017) and for the CP model over Singapore (Porson et al. 2019). Here, the initial and boundary conditions for each CP-ENS member are taken from the MetUM global ensemble (MOGREPS-G, Bowler et al. (2009), running with a horizontal grid spacing of 0.28° with 18 members.

The CP-ENS was run with 80 vertical levels with model lid at 38.5 km and at a horizontal grid spacing of 0.04° (~4.4 km) for a domain size of 600 × 600 grid points spanning East Africa (see Fig. 1). It consisted of 18 members, initialized four times a day (at 0000, 0600, 1200, and 1800 UTC) and ran out to 72 h. The science configuration of the dynamics and physics schemes of the atmosphere and land used for the tropical regions, denoted with "RAL1-T," are documented in Bush et al. (2019) and is the same used in Porson et al. (2019). In particular the tropical configuration differs from the midlatitude configuration used for MOGREPS-UK for these reasons: a different set of vertical levels (more levels in the upper troposphere to allow for a higher tropopause), the presence of boundary layer stochastic perturbations in the midlatitude configuration (useful to initiate convection earlier) and not in the tropical configuration, as well as the use of a prognostic cloud scheme (PC2) in the tropical configuration.

The version of MOGREPS-G run operationally did not provide the diagnostics required for the forecasting testbed, so a limited-area model with global model configuration, including the convective parameterization scheme (Walters et al. 2017) was also nested within MOGREPS-G. It is this limited area version (hereafter Glob-ENS), with the same horizontal grid spacing of MOGREPS-G, that will be used for comparison against the CP-ENS in this paper. Apart from its limited-area setup, the Glob-ENS only differs from the MOGREPS-G configuration by not having stochastic physics activated. The stochastic physics perturbations used in MOGREPS-G were technically difficult to replicate in the Glob-ENS limited-area setup and were therefore switched off. The impact of the stochastic physics on the spread of MOGREPS-G is much smaller than the impact of initial condition perturbations. For the purpose of this paper and the SWIFT testbed, rather than running a separate deterministic configuration, the control members of each respective ensemble (CP-ENS and Glob-ENS) were selected to represent the deterministic forecasts (CP-DET and Glob-DET).

2) OBSERVATIONS

The sparsity of ground observations in tropical regions of Africa makes model verification more challenging than in midlatitude regions. Therefore, precipitation forecasts were compared to gridded satellite observations derived from the Global Precipitation Measurement (GPM) mission (Hou et al. 2014), specifically the Integrated Multisatellite Retrievals for GPM (IMERG) Final Precipitation, version 6 (V06), level 3 product (Huffman et al. 2018; Tan et al. 2019), which we will refer to as GPM-IMERG. GPM-IMERG was preferred over other satellite derived products due to its high temporal and spatial resolution (half-hourly and 0.1°), which is essential to demonstrate the CP model capabilities on subdiurnal time scales. GPM-IMERG has been used extensively for model verification in the tropics, including Africa (Kniffka et al. 2020; Woodhams et al. 2018; Stein et al. 2019). The use of GPM-IMERG comes with some caveats: over southern West Africa, Maranan et al. (2020) found that GPM-IMERG overestimated the frequency and intensity of weak precipitating systems, while it underestimated the intensity of heavier rainfall events. For specific case study days with heavy rainfall events over South Africa, Stein et al. (2019) found that GPM-IMERG matched the radar-observed spatial pattern of rainfall well although not necessarily the amounts. However, in comparison against rain gauges, Dezfuli et al. (2017a) found that GPM-IMERG captured well the annual cycle and the diurnal cycle during the March–April–May "short rains" season over East Africa, which is the focus period of this study.

3) SPATIOTEMPORAL MATCHING

Both the CP-ENS and Glob-ENS rainfall fields were regridded to match the GPM-IMERG grid using the conservative method of the Climate and Forecast (cf) package (https://ncas-cms.github.io/cf-python/introduction.html). Analysis will focus on the 3-h accumulated precipitation since, following Woodhams et al. (2018), the benefit of CP models compared to global models is potentially best demonstrated on subdaily scales. The spatial domains used in this analysis are shown in Fig. 1. To illustrate regional variability in rainfall characteristics, such as the diurnal cycle, different subregions were selected (black dashed boxes in Fig. 1). These subregions correspond to the wettest locations, both in terms of rainfall amount and number of days with daily accumulation equal or exceeding 10 mm day^{-1} (Fig. 2) and can be characterized by the presence of lakes, mountains, and coastlines, which induce local circulations affecting the phase and amplitude of the diurnal cycle.

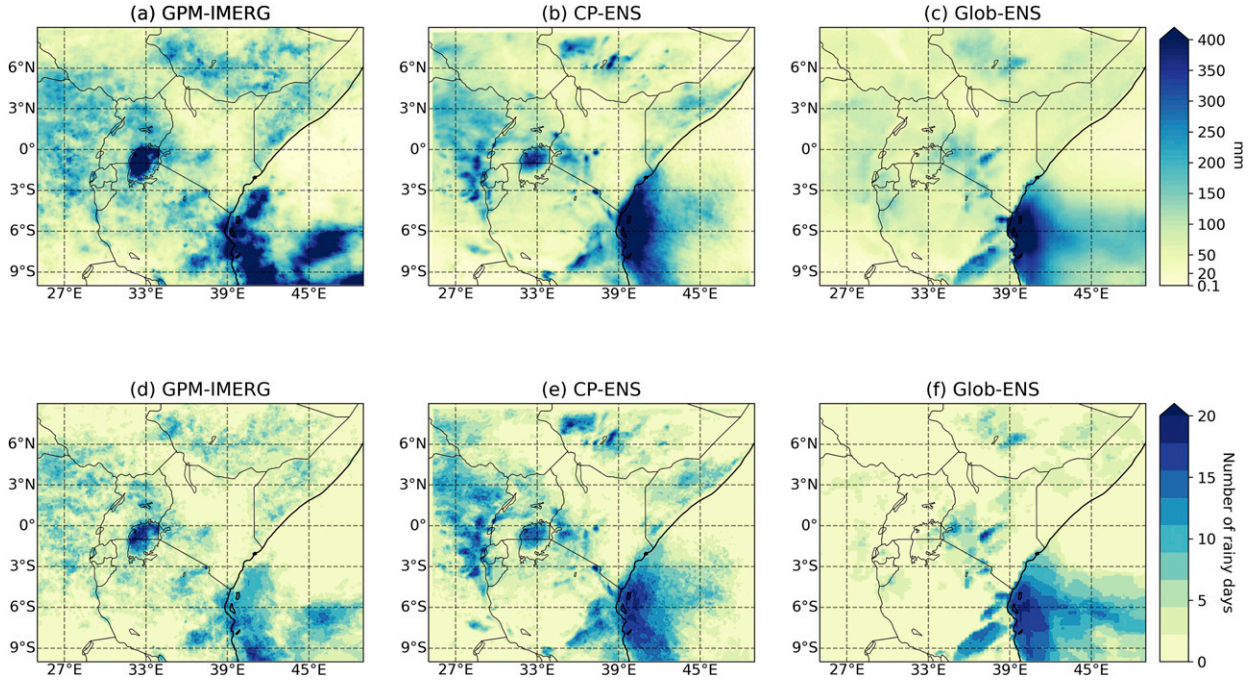

FIG. 2. (a)–(c) Total accumulated precipitation (mm) between 0000 UTC 20 Apr 2019 and 0000 UTC 13 May 2019 and (d)–(f) number of rainy days (defined by exceeding a daily accumulation of 10 mm) for GPM-IMERG and the ensemble mean of the model rainfall. The forecast precipitation is from the $T + 12$ to $T+ 36$ h accumulation, initialized at 1200 UTC for each day of the period.

b. *Forecast spatial verification methods*

Despite greater physical realism provided by CP models compared to global models, they are not expected to match perfectly with observations on a gridpoint scale. Therefore, traditional gridpoint verification methods have given way to neighborhood (or "fuzzy") verification methods (Ebert 2008; Gilleland et al. 2009). In addition to their use in verification, neighborhood methods have also been used to generate probabilities from deterministic forecasts (Theis et al. 2005), by taking the mean of the number of grid points exceeding a particular threshold within each neighborhood (hereafter the neighborhood probability, NP). Schwartz et al. (2010) extended this methodology to ensemble forecasts by further averaging the spatial mean over all the members, a technique which Ben Bouallègue and Theis (2014) referred to as *smoothing*. Schwartz and Sobash (2017) subsequently named it the *neighborhood ensemble probability*, which is how we will refer to it in this paper.

Here, probabilistic forecasts generated using the "neighborhood ensemble probability" (NEP) are compared to probabilistic forecasts generated with the NP method from the deterministic forecasts. The two methods can be described mathematically as follows:

- First, a common step in generating probabilities either from ensembles or deterministic forecasts is to convert the rainfall accumulation field f_{ij} into a binary field by applying a threshold q_j, for each grid point i and ensemble member j:

$$\mathrm{BP}_{ij} = \begin{cases} 1, & \text{if} \quad f_{ij} \geq q_j, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

- Next, in the case of ensembles, the ensemble mean of the binary field is calculated:

$$\mathrm{EP}_i = \frac{1}{N-1} \sum_{j=1}^{N-1} \mathrm{BP}_{ij}, \tag{2}$$

- Finally, for each grid point i, the spatial mean over each square neighborhood S_i, consisting of N_b grid points, is calculated:

$$\mathrm{NEP}_i := \frac{1}{N_b} \sum_{k=1}^{N_b} \mathrm{EP}_k, \quad k \in S_i, \tag{3}$$

$$\mathrm{NP}_i := \frac{1}{N_b} \sum_{k=1}^{N_b} \mathrm{DP}_k, \quad k \in S_i. \tag{4}$$

Thus, NP is only a spatial average, whereas NEP is an ensemble average *as well as* a spatial average (see also Schwartz and Sobash 2017). By comparing NEP and NP, we therefore assess whether the ensemble adds skill to simple neighborhood averaging provided by NP. For spatial verification, we process the observations as a binary field [Eq. (1)] when using the relative operating characteristic (ROC) and fractional [Eq. (4)] when using the fractions skill score (see section 4 for specifics).

where q_j is the percentile threshold calculated for each member separately and BP_{ij} refers to the binary probability.

where EP_i refers to the ensemble mean probability and N is the number of ensemble members. The sum starts from 1, because the control member (member 0, unperturbed) is

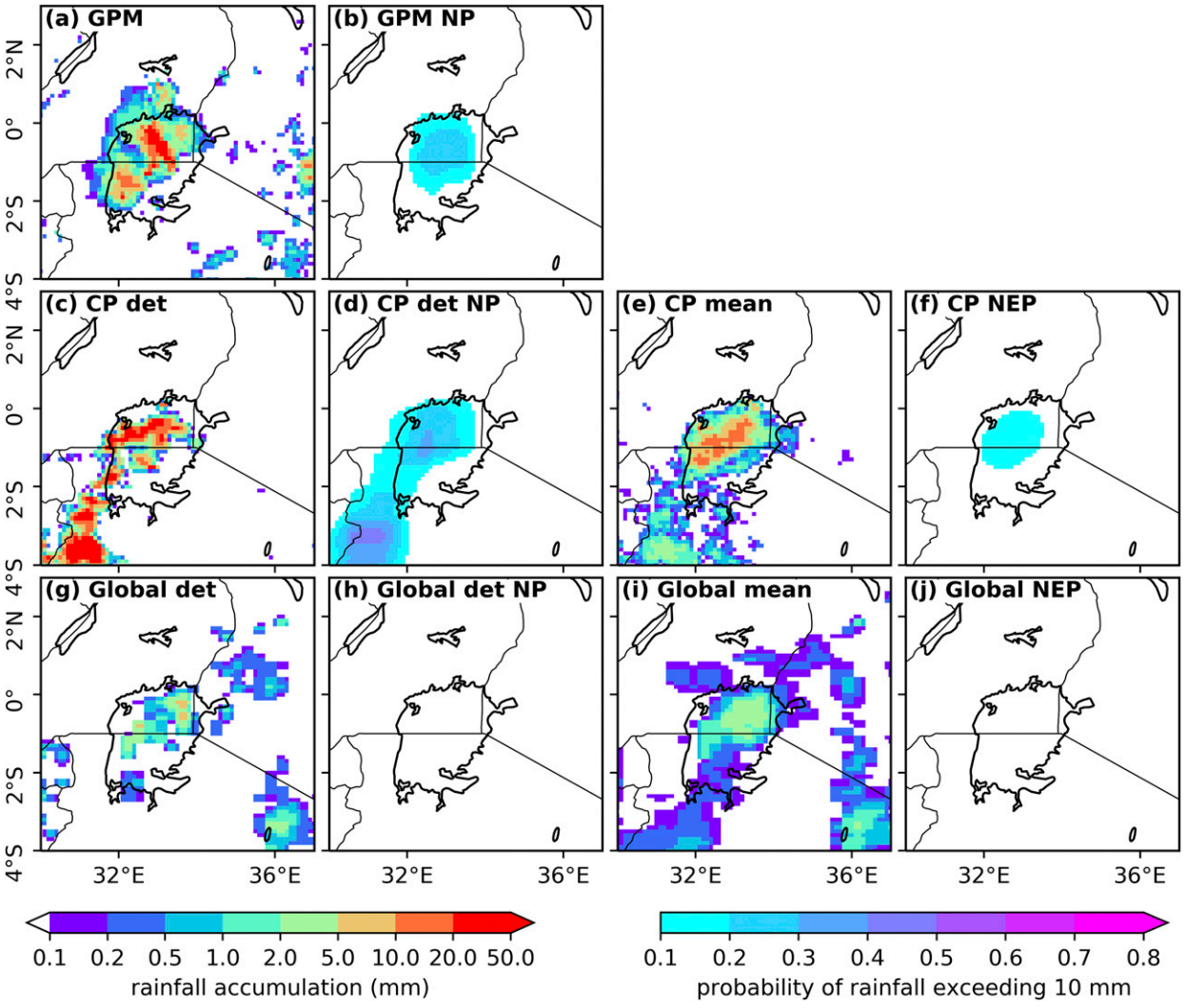

FIG. 3. Observational or forecast data for the 3-h rainfall accumulation between 0300 and 0600 UTC 29 Apr 2019. Forecast data are taken from the 1200 UTC initialization on 27 Apr 2019 ($T - 39$–42). Probabilities are for 3-h rainfall accumulation exceeding 10 mm within an $n = 15$ (~165 km) neighborhood: (top) GPM observations, (middle) CP-ENS forecast, and (bottom) Glob-ENS. (a),(c),(g) Rainfall accumulation and (b),(d),(h) neighborhood probability [NP, Eq. (4)] of threshold exceedance from observations (in the top panels) or the control member of the ensemble (in the middle panels). (e),(i) Ensemble mean rainfall accumulation and (f),(j) neighborhood ensemble probability [NEP, Eq. (3)].

excluded from the calculation of the probabilistic forecast and has been selected to represent the deterministic forecast, see section 2. In the case of deterministic forecasts (i.e., the control member), we define the deterministic probability $DP_i := BP_{i0}$.

An example of NEP and NP probabilistic products is shown in Fig. 3. Figure 3a shows the observed rainfall accumulation for 29 April 2019 between 0300 and –0600 UTC, and Fig. 3b shows the observations as a neighborhood probability (NP) of exceeding 10 mm. The accumulations predicted by the control member of the CP and global ensembles are shown in Figs. 3c and 3g, respectively, with the corresponding NPs in Figs. 3d and 3h. Figures 3e and 3i show the ensemble mean accumulations for the CP and global ensembles, respectively, and the NEPs are shown in Figs. 3f and 3j. Note that the NEP is not the same as the NP of the ensemble mean; rather, the NEP is the average of the NPs across all ensemble members. It is also worth to notice that probabilities from the global model (Figs. 3h,j) are below 0.1, lower than the corresponding probabilities from the CP model (Figs. 3d,f). In general, NEP will be lower than NP because the probability field has undergone more smoothing, as discussed previously.

3. Rainfall characteristics: Intensity and diurnal cycle

In this section, an analysis of the characteristics of rainfall intensity and timing is performed to provide a qualitative assessment of the CP versus global-configuration simulations

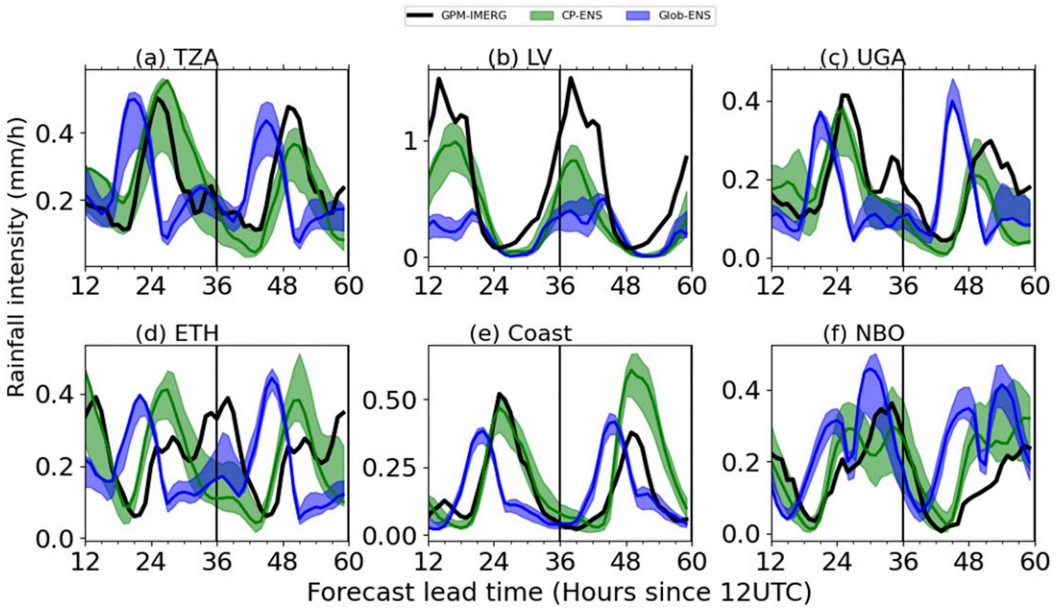

FIG. 4. Mean hourly rainfall for the models and observations with the panels showing averages for the different subregions (as in Fig. 1). Green and blue shadings represent the envelopes of the 18 ensemble members comprising the CP-ENS and Glob Ens, respectively, with solid lines indicating the control members. The black solid line represents the GPM-IMERG satellite observations. Values on the *x* axis represent starting forecast hours of the 3-h accumulation periods (e.g., an *x*-axis value of 24 is for 3 h accumulated between 24 and 27 h). The black vertical solid line indicates the valid time of midnight (in UTC).

against satellite-derived rainfall observations. First, Fig. 2 shows maps of the total accumulated rainfall over all the forecasts for GPM-IMERG and the ensemble mean for each of CP-ENS and Glob-ENS. The CP-ENS agrees better with the observed patterns of precipitation, but in places, such as southern Ethopia, the CP-ENS overestimates the rainfall accumulation with respect to GPM-IMERG. The Glob-ENS accumulations are lower with respect to observations almost everywhere (Figs. 2b,c), which could be due to the underestimation of heavy rainfall rates by the global model (Woodhams et al. 2018). Off the Tanzania coast, we assume that the improved performance of Glob-ENS with respect to other regions is related to large-scale and slow-varying signals, such as the intertropical convergence zone (ITCZ) position, and perhaps an indirect consequence of the Tropical Cyclone Kenneth affecting the region between Madagascar and Mozambique during this period.

a. Diurnal cycles

To investigate the diurnal cycle of rainfall in the different subregions (cf. Fig. 1), hourly rainfall fields are spatially averaged over each subregion for each day and ensemble member and then averaged over the different forecasts. Results for the 1200 UTC initialization are shown in Fig. 4, for lead times of 12–60 h. While we note slight differences with the other initialization times (not shown), the qualitative behavior is as follows:

- *Timing*: In agreement with previous MetUM studies for tropical Africa (Pearson et al. 2014; Birch et al. 2014; Woodhams et al. 2018), the CP-ENS shows a better representation of the diurnal cycle than the Glob-ENS when compared to GPM-IMERG observations. The daytime peaks of observed rainfall are generally well predicted by the CP-ENS, especially over the Somali coast, where the sea breeze was probably the driver of the rainfall systems (Camberlin et al. 2018). Nighttime peaks are missed over Tanzania, Uganda, and southern Ethiopia (Figs. 4a,c,d) by the CP-ENS. In regard to the Glob-ENS, it tends to predict an earlier peak than observed in all the regions, except over Lake Victoria, where the Glob-ENS peaks at the same time of CP-ENS and GPM-IMERG observations. This is in agreement with Woodhams et al. (2018) who found that, over the Lake Victoria basin, the parameterizing convection model reproduced well the timing of the rainfall peak, although underestimating the intensity.
- *Intensity*: Rainfall intensity of the peak is generally well estimated by the CP-ENS up to 36 h, especially over Tanzania (Fig. 4a), Uganda (except for the nighttime peak, Fig. 4b) and Nairobi area (Fig. 4f). In other regions the peak of rainfall is either overestimated (south Ethiopia at $T + 24$ h, Fig. 4d), underestimated (as for Lake Victoria, Fig. 4b) or missed (as for south Ethopia and Uganda at about $T + 36$ h). For day-2 forecasts (from 36 h up to 60 h), the CP-ENS performance deteriorates over the coast and Nairobi area, where it overestimates the observed peak (Figs. 4e,f, respectively). As time progresses, rainfall increases in the Glob-ENS for all subregions except for the Nairobi area and decreases for the CP-ENS over all subregions except for the

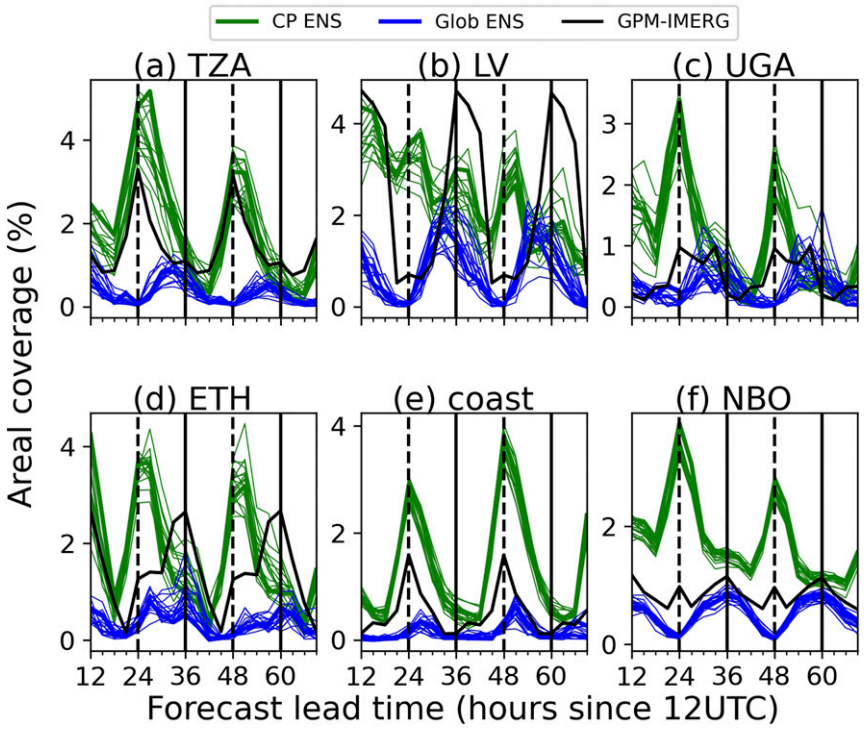

FIG. 5. (a)–(f) Fractions of grid points exceeding the accumulation of 10 mm $(3\,h)^{-1}$ for each panel corresponding to the different subregions. Values on the x axis represent starting forecast hours of the 3-h accumulation periods (e.g., an x-axis value of 24 is for 3 h accumulated between 24 and 27 h). Solid (dashed) black vertical lines refer to the valid time of midnight (midday) in UTC.

Somali coast. The gradual decrease in rainfall in CP configurations of the MetUM was also observed over Southeast Asia by Dipankar et al. (2020), who noted that this behavior depended on the driving model.

- *Spread*: The envelopes of rainfall intensity vary between the different subregions: smaller over the coastal regions and greater over the surroundings of Nairobi and Lake Victoria, with the CP-ENS generally showing greater envelopes than the Glob-ENS.

b. Areal coverage

Aggregate areal coverage of the 3-h rainfall accumulation exceeding defined thresholds provides complementary information to the mean diurnal cycle. Different rainfall thresholds were selected [1, 2.5, 5, 10, 25, 50 mm $(3\,h)^{-1}$]. Here results relative are presented in Fig. 5 for the 10 mm $(3\,h)^{-1}$ threshold. Figures relative to other thresholds are included in the supplemental material. For thresholds up to 5 mm $(3\,h)^{-1}$ CP-ENS has areal coverage less than or equal to GPM-IMERG, whereas Glob-ENS has greater areal coverage than both GPM-IMERG and CP-ENS for thresholds up to 2.5 mm $(3\,h)^{-1}$. This demonstrates that Glob-ENS predicts lighter and more widespread rainfall than both observations and the convection-permitting model, in line with previous findings with the MetU and also other studies over the United States using the Weather Research and Forecasting (WRF) Model (e.g., Schwartz and Liu (2014). For the 10 mm $(3\,h)^{-1}$ accumulation shown here and for greater thresholds (not shown), CP-ENS has greater areal coverage than both the observations and Glob-ENS in all the regions, except for Lake Victoria at $T + 24$ and $T + 48$ h.

Although the diurnal cycle is represented better by CP-ENS than by Glob-ENS, the former predicts too little light rainfall and too much heavy rainfall with respect to GPM-IMERG. The latter finding helps explain the overestimate of the rainfall amplitude seen in Fig. 4. Overestimation of the rainfall amount by CP models was also found by Marsham et al. (2013); Dipankar et al. (2020), among others. Also, in agreement with Fig. 4, areal coverage in Glob-ENS peaks earlier than observed, apart from over the coast.

c. Ensemble characteristics

To assess the spread–error relationship, the root-mean-square error (RMSE) of the domain averaged rainfall over each subregion is computed and compared to the ensemble spread, calculated as the square root of average ensemble variance as in Fortin et al. (2014).

For a perfect ensemble, the spread resembles the RMSE of the ensemble mean (Leutbecher and Palmer 2008; Fortin et al. 2014). In Fig. 6, we show these quantities for the 3-h rainfall accumulation averaged over the different subregions and for both ensembles. For all subregions and for most of the times,

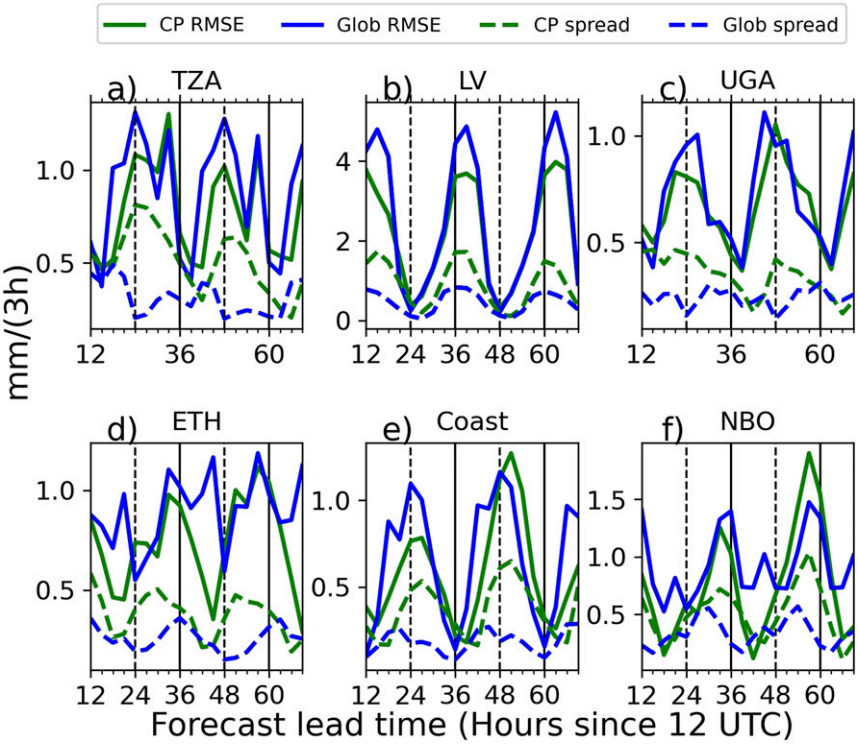

FIG. 6. Ensemble spread (dashed lines) and RMSE of the ensemble mean (solid lines) of the 3-h rainfall accumulation averaged over the different subregions. The values on the x axis represents the starting forecast hours for each accumulation period (e.g., 12 is the 3-h accumulation period between 12 and 15 h). Solid (dashed) black vertical lines refer to the valid time of midnight (midday) in UTC.

both ensembles are underdispersive, i.e., the ensemble spread is lower than the RMSE. Underdisperion is a well-known issue for convection-permitting ensembles (Porson et al. 2020; Loken et al. 2019; Romine et al. 2014), but the Glob-ENS is generally more underdispersive than the CP-ENS, with a higher RMSE and a lower spread. The spread–error relationship—and thus the level of underdispersion—varies across the different subregions and with time (cf. Figs. 4 and 5). For instance, a larger offset in the timing of the peak in rainfall leads to a broad peak in RMSE for Glob-ENS in most of the subregions. The worse initiation of the peak of coastal rainfall by the CP-ENS on day 2 also leads to a greater RMSE compared to day 1 (Fig. 6e). Similar to RMSE, spread follows the diurnal cycle, peaking when the rainfall intensity is largest. The spread–error relationship will also be evaluated spatially in section 5.

4. Probabilistic forecast verification

In this section, probabilistic forecasts from the CP-ENS will be verified and compared against the global and deterministic configurations using two metrics: the fractions skill score (FSS) and the area under the receiver operating characteristic (ROC) curve. They measure two different attributes of a forecasting system, namely the spatial displacement of rainfall patterns and the discriminating ability between events and no-events, respectively.

a. *Fractions skill score*

The FSS (Roberts and Lean 2008) was originally designed for deterministic forecasts to account for the uncertainty in forecasting the location of rainfall and mitigate for the double penalty when rainfall is displaced. With the FSS, the fractions of values above a given threshold within a given neighborhood, are evaluated, leading to values ranging from 0 (no skill) to 1 (perfect forecast). Roberts and Lean (2008) also introduced the *useful scale* as the neighborhood size where FSS $= 0.5 + f_0/2$ (or FSS $= 0.5$ if $f_0 < 0.2$ Skok and Roberts 2016), where f_0 is the observed rainfall frequency, i.e., the fraction of observed points exceeding a threshold. Following Mittermaier et al. (2013), percentile thresholds will be used in order to focus only on the spatial error of the predicted rainfall pattern and to avoid incorporating a frequency bias (see Fig. 5). As described in section 2, for the ensembles we will use NEP (CP NEP and Glob NEP) and for the deterministic forecasts NP (CP Det NP and Glob Det NP). GPM-IMERG observations have also been processed into NP for each given neighborhood size (see Fig. 3 for one example).

The choice of percentile requires a balance between a low enough percentile that gives meaningful statistics, so enough

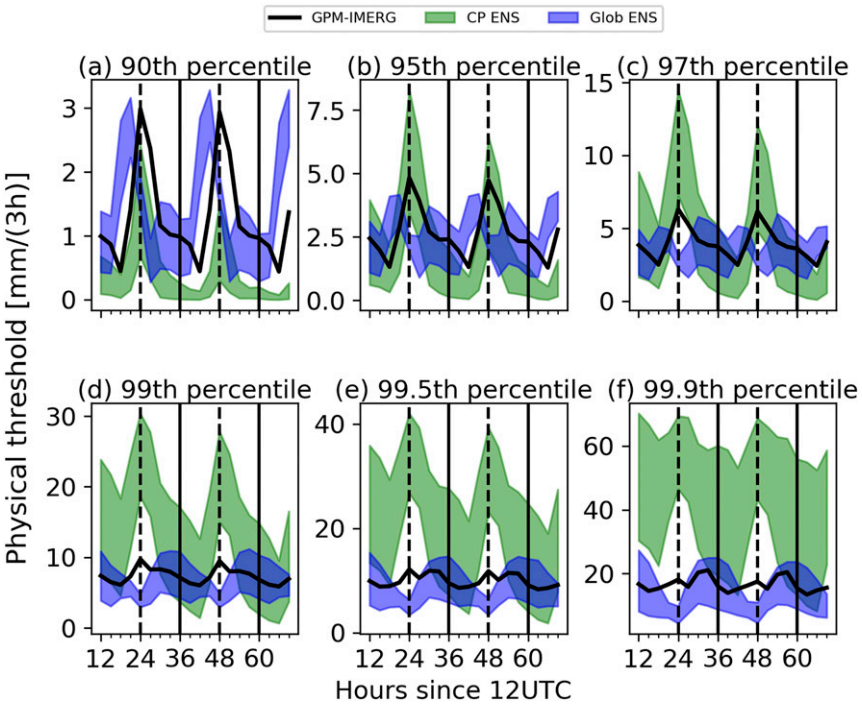

FIG. 7. Average physical thresholds [mm $(3\,\mathrm{h})^{-1}$] over all the forecasts corresponding to (a) 90th, (b) 95th, (c), 97th, (d) 99th, (e) 99.5th, and (f) 99.9th percentile threshold as a function of the forecast hour. The physical thresholds were computed for the large domain (red dashed box in Fig. 1) for each day and for each 3-h period separately. The green and blue shadings encompass the CP and global ensembles distributions, respectively. Values on the x axis represent starting forecast hours of the 3-h accumulation periods (e.g., an x-axis value of 24 is for 3 h accumulated between 24 and 27 h).

spatial coverage (events are not too rare), and a high enough percentile that is related to meaningful (heavy) rainfall values relevant for forecasting in the tropics. In Fig. 7, we show the ensemble spread in average physical thresholds corresponding to different percentiles at different times of the day for the large domain (red dashed box in Fig. 1). As expected from areal coverages relative to other thresholds, included as supplemental material, the Glob-ENS has higher physical values than the CP-ENS for the 90th percentile (Fig. 7a), comparable values for the 95th percentile (Fig. 7b) and lower physical values for the 99th percentile and above (Figs. 7d–f). Biases in the timing of convection described in the previous section can also be identified from Fig. 7. Note, however, that the FSS is calculated for each 3-h period separately on the domain enclosed by the red dashed line as in Fig. 1, with the relevant percentile threshold calculated for each period separately as well, so that any frequency bias due to the timing of the diurnal cycle will not influence the skill. Finally, in order to get a summary score, FSS is then averaged over the different cases, using equation S30 of the supporting information document by Skok and Roberts (2016). Several factors could affect model performance in terms of FSS: neighborhood size N_b, rainfall percentile, accumulation period, but also valid time and initialization time. In Fig. 8, we show FSS as a function of forecast time for different percentile thresholds, considering only the 1200 UTC initialized forecasts and using a neighborhood scale of $n = 23$ grid points (255 km). CP NEP has the highest FSS for all the different percentiles and at nearly all times. The CP Det NP is generally more skillful than Global NEP, while Global NEP is more skillful than the Global Det NP. We see that FSS decreases with forecast lead time—particularly when comparing day 1 and day 2—and as the percentile increases. For the 99th percentile and above (associated with rainfall accumulations greater than 30 mm $(3\,\mathrm{h})^{-1}$ for the CP-ENS and 10 mm $(3\,\mathrm{h})^{-1}$ for GPM-IMERG and Glob-ENS), all configurations mostly have FSS below 0.5, the useful skill value, although the Global NEP and Global Det NP struggle attaining useful skill already at the 97th percentile. This is likely due to the most intense events being localized in nature and therefore more difficult to forecast. Compared to Fig. 8 in Schwartz (2019), who performed a similar analysis over United States, FSS remains low for all percentiles, despite our use of a larger neighborhood. Our values are comparable, however, to those found over a small domain centered on Singapore by Sun et al. (2020).

The FSS shows a diurnal cycle, with the strongest amplitude generated from the CP model: it peaks at around $T + 24$ h and $T + 48$ h for percentiles up to the 97th, which coincides with the timing of maximum rainfall (see Fig. 7). For percentiles equal or greater than the 99th, FSS shows additional peaks at

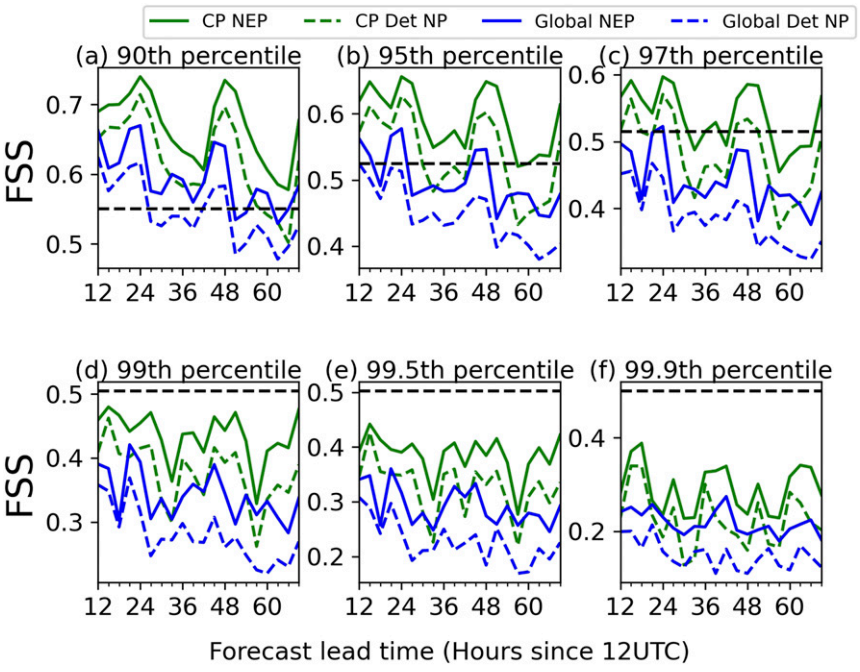

FIG. 8. Mean FSS corresponding to the (a) 90th, (b) 95th, (c) 97th, (d) 99th, (e) 99.5, and (f) 99.9 percentile as a function of the forecast hour on a fixed neighborhood size of $n = 23$ grid points (~255 km), calculated over the large domain. Values of the FSS useful scale are represented by dashed horizontal lines. Values on the x axis represent starting forecast hours of the 3-h accumulation periods (e.g., an x-axis value of 24 is for 3 h accumulated between 24 and 27 h).

$T + 12$ h and $T + 36$ h, corresponding to the nighttime storms, depicted also by Fig. 7. Since the selection of percentiles removes the frequency bias from the FSS, the diurnal cycle is somewhat unexpected, although diurnal signals in the FSS have previously been reported (e.g., Schwartz 2019). A possible explanation concerns the spatial organization of rainfall. At the grid scale (i.e., neighborhood size $n = 1$), the FSS reaches its maximum around $T + 12$ h and $T + 36$ h (not shown) when convection appears more organized, and as the neighborhood size increases FSS peaks at $T + 24$ (as in Fig. 8), when convection appears more scattered. The neighborhood approach thus appears to have a greater impact on scattered patterns, which are likely better captured by the CP model, thus allowing it to more significantly outperform the Global model at those hours. This varying behavior of FSS with the pattern of convection was also noted for U.K. convection by Flack et al. (2018) and related to differences in large-scale forcing for their case studies.

In our remaining analysis with the FSS, we will only present results for the 97th percentile, at which all models but Glob Det NP have useful skill at some times at the scale shown in Fig. 8 or larger, and for which the physical value [~6 mm $(3\,\text{h})^{-1}$ for GPM-IMERG, up to 15 mm $(3\,\text{h})^{-1}$ for CP-ENS] can be considered high enough to be related to intense rainfall (Fig. 7). In Fig. 9, we show the FSS as a function of neighborhood size as well as forecast time. FSS increases with neighborhood size (as expected and by construction, see Roberts and Lean 2008) and decreases with time, the latter consistent with Fig. 8. Also shown in Fig. 9 are the mean and median useful scale, calculated over the useful scales determined for each of the 24 forecasts, where the maximum length of the domain is used if FSS ≤ 0.5 for all neighborhood sizes (Sun et al. 2020). The mean useful scale is always greater than the median, as found by Sun et al. (2020), but this difference is greatest in the global forecasts, implying that these have greater outliers in useful scale than the CP forecasts. As with the FSS, the mean useful scale increases with forecast lead time and has a diurnal cycle which is most evident in the CP forecasts. Generally, the CP-ENS has the highest FSS and therefore the smallest useful scale (~100 km). Also, the Glob-ENS performs better than the Glob DET. We consider the useful scale as a metric in more detail for the different forecast pairs in Fig. 10. A bootstrap technique was employed to characterize uncertainty: 24 samples were drawn randomly (with replacement) from the 24 forecasts, after which the mean useful scale was calculated for each forecast model, including the difference between forecast model pairs. Following Schwartz and Liu (2014), this process was repeated 5000 times allowing estimation of the 95% confidence interval. The largest bootstrapped mean difference in useful scale between CP NEP and CP Det NP is about 100 km (Fig. 10a) and between CP NEP and Global NEP about 150 km (Fig. 10b), but these differences do not occur at the same time in the forecast run. CP NEP is more skillful than CP Det NP in terms of the bootstrapped mean useful scale, with the smallest

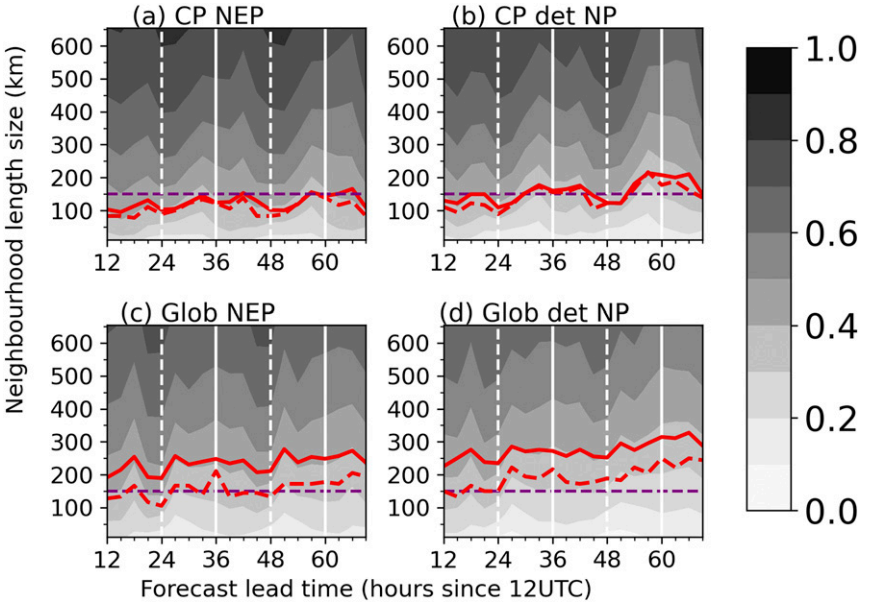

FIG. 9. Mean FSS as function of forecast hour and neighborhood length side (calculated over the large domain). The solid (dashed) red line indicates the mean (median) scale at which FSS = 0.5 over all the different forecasts. The dashed horizontal purple line indicates the 150-km scale. Values on the x axis represent starting forecast hours of the 3-h accumulation periods (e.g., an x-axis value of 24 is for 3 h accumulated between 24 and 27 h). Vertical solid (dashed) white lines indicate midnight (midday) UTC.

differences (which are not significantly different from zero) between around 1200 UTC ($T + 24$ and $T + 48$). This suggests that explicitly resolving convection would be sufficient for predicting the location of intense rainfall over East Africa domain at the time of peak rainfall (as confirmed by the CP Det NP and Global NEP comparison in Fig. 10c), whereas the CP-ENS has additional skill at other times and at longer lead times. Compared to the global model, both CP NEP and CP Det NP have better useful scale (significantly different from zero) at the time of peak rainfall, but their superiority is no longer evident after day 1. The Glob-ENS is generally more skillful than Glob Det NP in terms of useful scale (Fig. 10d), although again this difference is not significantly different from zero. The influence of initialization times on the FSS as a function of valid time was also investigated. Considering only the CP-derived forecasts, none of the initialization times clearly outperforms the others (not shown). For the global forecast, the 1800 UTC was found to have higher skill for the ensembles at all valid times, followed by the 0600 UTC run.

FSS was also calculated for 24-h accumulations exceeding the 97th percentile for comparison with Woodhams et al. (2018). This is the accumulation period mostly used by African weather agencies, partners of the SWIFT project. Figure 11 shows that for 24-h accumulations, the CP-based forecasts are more skillful than Glob det NP and Glob ENS for both the periods, with useful scale at about 150 km. Note that the improvement in useful scale from CP Det to CP-ENS is fairly small, at around 10 km for 24-h accumulations, similar to the grid scale. However, this difference in the useful scale is smaller than the one for the 3-h accumulation (cf. Fig. 9). The improvement from global to CP is more pronounced than found by Woodhams et al. (2018), but we note that the latter had a longer dataset, which included dry spells, compared to our 2-week wet period.

b. *Areas under the ROC curve*

Areas under the ROC curve (AUC; Mason and Graham 2002) were computed for different neighborhood sizes, rainfall thresholds and initialization times for the 3-h rainfall accumulation NEP and NP probabilistic forecast. Physical fixed thresholds were used rather than percentiles, because we want to have a unique definition for events and nonevents across models and observations. The use of physical thresholds is justified because the ROC curve and derived scores are insensitive to any lack of reliability by probabilistic forecasts or forecast biases (Kharin and Zwiers 2003; Vogel et al. 2018). A threshold of 10 mm $(3\,\text{h})^{-1}$ was chosen for relevance to intense events in all three datasets, roughly the 97th percentile for CP and 99th percentile for global and observations (see Fig. 7). ROC statistics have been aggregated on each of the subregions (Fig. 1) with contingency tables populated following the methodology described by Schwartz and Sobash (2017) and Vogel et al. (2018). Specifically, at each grid point, observations are treated as binary [BP, see Eq. (1)] whereas the forecasts are treated as NEP (ensembles) or NP (deterministic).

Figure 12 shows AUC values for the NEP and NP forecasts exceeding the 10-mm accumulation in 3 h for the 1200 UTC initialization on a neighborhood size of $n = 23$

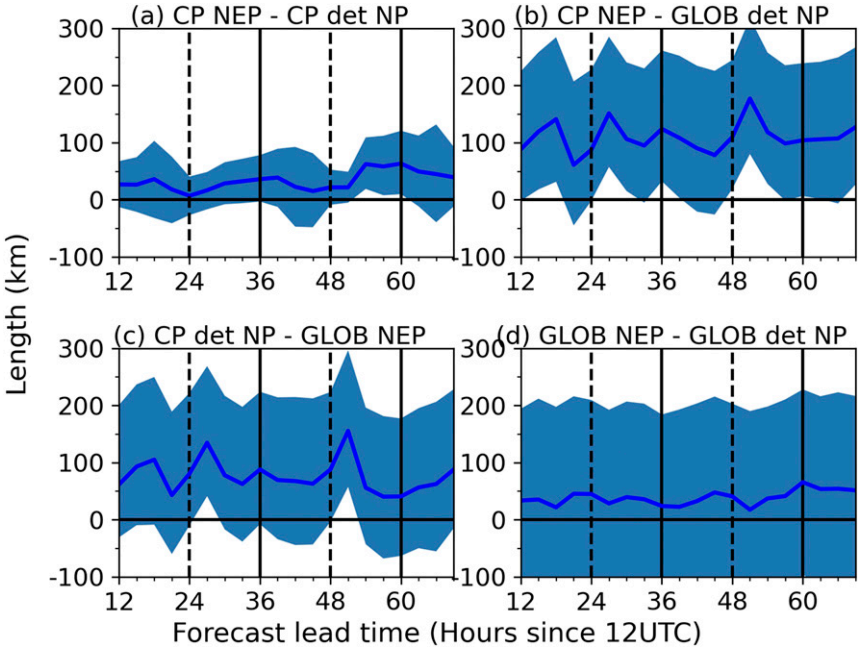

FIG. 10. Differences (in km) of the mean useful scales as a function of forecast hour for the different model pairs (calculated over the large domain), represented by the solid line. The shading represents the 95% confidence interval calculated using a bootstrap resampling with replacement. Black solid (dashed) lines indicate the 0000 UTC (1200 UTC) valid time. The values on the *x* axis represent the starting forecast hours for each accumulation period (e.g., 12 is the 3-h accumulation period between 12 and 15 h). Positive values indicate that the first forecast for each pair is more skillful and vice versa.

grid point (~ 255 km), where AUC > 0.5 indicates ability to discriminate between events and no-events. For most of the times and subregions, CP NEP (solid green) has higher AUC values than the other forecasts, with the highest values at times when convection peaks and lowest values during the diurnal minimum. The time of maximum AUC varies across the subregions, similar to the diurnal cycles shown in Fig. 5: when convection is most active, there are more events to be detected, potentially leading to higher hit rates (and higher false alarm rate) implying higher AUC values and vice versa. At the times of peak convective activity, AUC CP NEP reaches values greater than 0.7, which is above the threshold of usefulness for probabilistic predictions (Buizza et al. 1999). None of the other forecasts reach above this threshold value for significant periods of time, apart from the Glob NEP for the coastal region. For all subregions, similar conclusions can be drawn from AUC analyses using different rainfall thresholds and neighborhood sizes (not shown), with CP NEP retaining AUC above 0.7 and AUC differences between CP NEP and other forecasts increasing for higher thresholds. AUC for the large domain is shown as supplemental material.

5. Spatial spread–error relationship for CP-ENS

In general, the CP-ENS has been shown to be the most skillful model for predicting rainfall over East Africa. Given the novelty of CP-ENS in this region, it is vital to understand how the ensemble data may be processed to provide the best forecast guidance. Using a variety of FSS scores to represent the different guidance, this section will explore which is the most skillful diagnostic rainfall forecast that can be derived from the CP-ENS, and therefore offer the greatest potential to local forecasters. So far, the FSS has been computed for the neighborhood ensemble probability (NEP), thereby assessing the ability of the ensemble to predict the probability of exceedance of a threshold rainfall accumulation. However, rainfall accumulation predictions from ensembles may also be presented as the ensemble mean, or as a collection of the individual ensemble members (e.g., as postage stamp plots). To assess the predictive skill using these different methods, corresponding variations of the FSS are computed. FSS_{ens_mean} is the FSS computed using the neighborhood probability (NP) of the ensemble mean (i.e., essentially treating the ensemble mean as a deterministic forecast). Although taking the mean of all ensemble members unrealistically smooths out the intense regions of precipitation and lowers rainfall rates, this FSS analysis uses percentile thresholds, such that it is only the placement of the rainfall—not the amount—which is evaluated. FSS is also computed using the neighborhood probability for each individual ensemble member (FSS_{em}), with FSS_{det} distinguishing the control member. More details about these different scores are provided in Table 1.

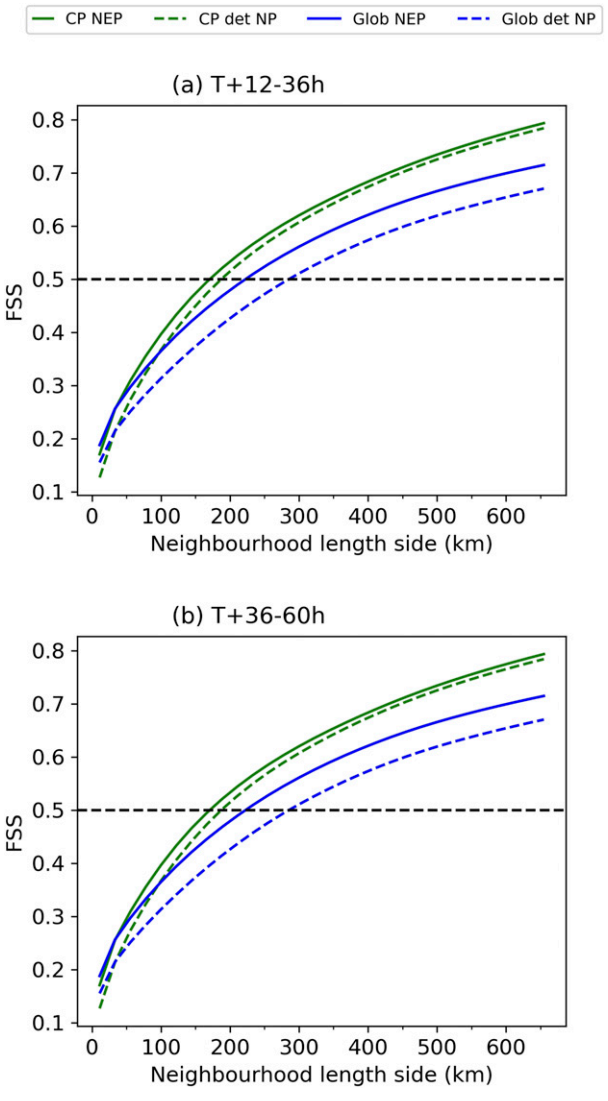

FIG. 11. Mean FSS for the 24-h accumulation period as a function of the neighborhood size for the period (a) from $T + 12$ to $T + 36$ h and (b) from $T + 36$ to $T + 60$ h of the 1200 UTC initialization (calculated over the large domain). The black dashed line refers to the value of FSS = 0.5, the useful scale as in the main text.

Figure 13 shows the different FSS metrics as a function of lead time using the 97th percentile threshold and a neighborhood size of ~250 km. Skill scores for forecasts with less than 12-h lead time are shown for interest, but it should be considered that these forecasts are still within the spinup period. In agreement with findings by Woodhams et al. (2018) for a CP deterministic model over East Africa, 0900–1800 UTC is the most skillful time of day according to FSS_{NEP}, FSS_{det}, and FSSs of the individual ensemble members, especially at lead times exceeding 36 h (Figs. 13d–f). FSS_{ens_mean} shows the most skillful times to be 0300–0900 UTC at lead times shorter than 36 h. For all metrics, 2100–0000 UTC shows the lowest skill, suggesting the model may be unable to capture storms which persist overnight. Between 2100–0000 and 0000–0900 UTC, all metrics show that skill is greatest closer to the valid time (short lead times). For other valid times, skill remains fairly constant or slightly reduces with decreasing lead time.

FSS_{NEP} is almost always the highest score, suggesting that the best way to display information from the CP-ENS is as a probability of threshold exceedance (as was done in section 4). Similar results were obtained by Schwartz et al. (2014), who demonstrated that the best ensemble guidance was realized by applying the neighborhood approach to the gridscale probabilistic forecasts. FSS_{ens_mean} is greater than FSS_{det} when convective activity is low (0000–0900 UTC, Figs. 13a–c), suggesting that the ensemble mean adds value to the deterministic model for the prediction of rainfall location during this time. However, during the period of convective activity (0900–1800 UTC, Figs. 13d–f), the deterministic model is more skillful than the ensemble mean out to a lead time of $T - 24$ h. The deterministic model is often at the upper end of the envelope of skill of the individual members, especially at lead times shorter than 54 h (cf. FSS_{det} and FSS_{em} range), suggesting that the ensemble perturbations may lead to a deterioration in skill.

In section 3 the spread–error relationship for rainfall *intensity* was discussed. FSS can be used to show the spread–error relationship for the *location* of rainfall by comparing the mean FSS between observations and each ensemble member $eFSS_{mean}$ and the mean FSS between each ensemble member–member pair $dFSS_{mean}$ (Dey et al. 2014). For example, high $dFSS_{mean}$ indicates that ensemble members are predicting rainfall in similar locations, therefore the spatial spread is low. The standard deviation of the FSS between each ensemble member–member pair $dFSS_{std}$ is a measure of the range of dFSS values, where a high $dFSS_{std}$ suggests that there are some outlier members with particularly high or low dFSS (Dey et al. 2014). Table 1 gives more details about these measures. Figure 14 shows $eFSS_{mean}$, $dFSS_{mean}$ and $dFSS_{std}$ for (Fig. 14a) 0000–0300 UTC and (Fig. 14b) 1200–1500 UTC rainfall accumulations as a function of lead time. These two times were chosen to be representative of outside (Fig. 14a) and during (Fig. 14b) the main convective period. The $dFSS_{mean}$ is greater than $eFSS_{mean}$ for both times of day and all forecast lead times, showing that the uncertainty in spatial location of the rainfall is not fully captured by the ensemble. This is true for all times of day (not shown). Spatial spread is lower during the convective period (1200–1500 UTC) but fairly constant throughout the forecast (initialization dependence aside). For the 0000–0300 UTC accumulation period, the spread increases ($dFSS_{mean}$ decreases) as forecast lead times increases. The 0600 and 1800 UTC initializations (circles and diamonds) have a greater $dFSS_{mean}$ (i.e., lower spread) than the 0000 and 1200 UTC initializations, possibly related to the data assimilation cycle. The $dFSS_{std}$ is lower for the 1200–1500 UTC accumulation compared to the 0000–0300 UTC accumulation, indicating that there are fewer major outlier ensemble members during the convective period. Few outliers during the main rainfall period suggests that the ensemble perturbations are too small to affect major rain locations. This is consistent with the findings from section 4 that CP NEP and CP Det NP had similar FSS and similar useful scales during the main convective period (corresponding to from $T + 24$ to $T + 30$ h for the 1200 UTC initialization, cf. Figure 10). Overall, the

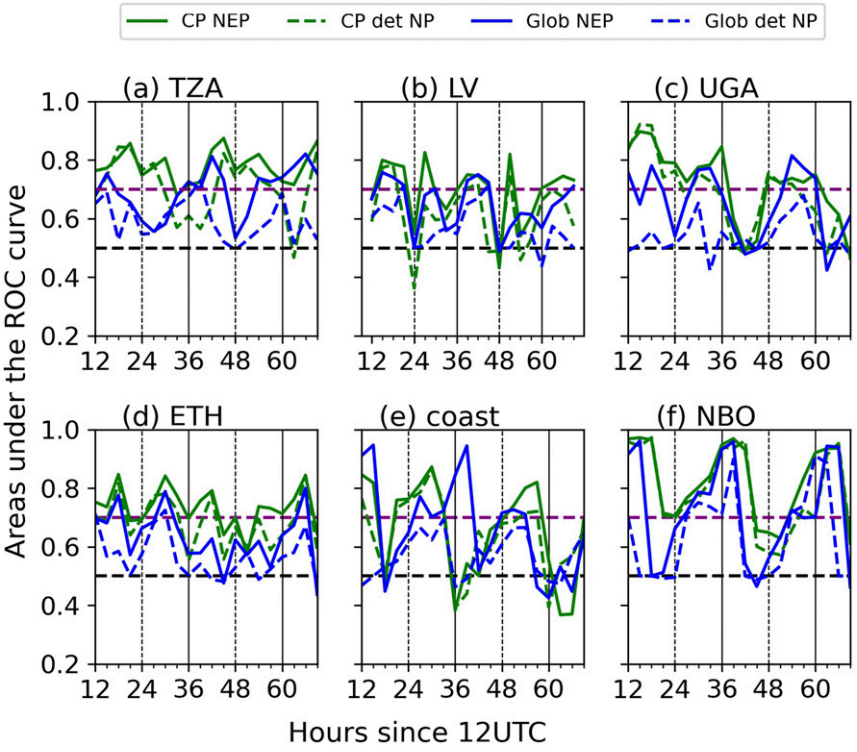

FIG. 12. Areas under the ROC for the probabilistic forecasts of 3-h rainfall accumulation exceeding 10 mm on a neighborhood size of $n = 23$ grid points (approximately 255 km) aggregated over the different subregions and generated either from ensembles (NEP) or control members (NP). Values on the x axis represent starting forecast hours of the 3-h accumulation periods (e.g., an x-axis value of 24 is for 3 h accumulated between 24 and 27 h).

high values of $dFSS_{mean}$ and low values of $dFSS_{std}$ throughout the diurnal cycle imply that the ensemble members are not very spatially different from one another. The lack of spatial spread also explains why FSS_{ens_mean} is often fairly similar to FSS_{det} and often within the envelope of FSS of individual members (Fig. 13).

6. Summary and conclusions

In an operational forecasting testbed environment that occurred during April–May 2019, convection-permitting ensemble forecasts were produced by the Met Office for tropical East Africa for the first time. In this paper, potential benefits of the CP ensemble were assessed compared to the driving global ensemble, first in terms of rainfall characteristics (intensity and diurnal cycle) and then by verifying probabilistic forecasts calculated using a neighborhood approach. The ensemble forecast results were compared with deterministic forecasts (assembled from the ensemble control member). Probabilities for the deterministic forecasts were computed for comparison with the ensemble probabilities, by computing the fractions of grid points exceeding a threshold within a given neighborhood. To assess whether the CP ensemble forecasts added any skill with respect to the global and deterministic forecasts, the FSS was used to discern skill in the location of rainfall and the area under the ROC curve (AUC) to assess the ability to discriminate between events and nonevents. The results of this analysis can be summarized as follows:

1) *Convection-permitting versus parameterized convection*: The CP ensemble model improves the representation of the diurnal cycle with respect to the global ensemble over most of the subregions. The global ensemble tends to peak earlier than GPM-IMERG and CP ensemble, especially for the afternoon rainfall peak, in agreement with previous studies for tropical Africa (Birch et al. 2014; Pearson et al. 2014; Woodhams et al. 2018). However, in some subregions (Uganda and southern Ethiopia) CP ensemble is shown to miss the overnight/early morning peak in rainfall. Further analysis is required to investigate the reasons why the CP ensemble misses convective events in these regions at these times of day. The CP ensemble generally produces more rainfall with respect to GPM-IMERG and the global ensemble, especially for higher rainfall thresholds, also in agreement with other studies (Kendon et al. 2012; Birch et al. 2014; Woodhams et al. 2018; Dipankar et al. 2020).
2) *Spread–error relationship*: Ensemble spread was assessed both in terms of the rainfall amount, compared to the RMSE of the ensemble mean of the two ensembles for the different subregions and in terms of the spatial agreement

TABLE 1. Description of different FSS values as plotted in Figs. 13 and 14.

FSS	Description	Interpretation
FSS^{det}	Traditional FSS from Roberts and Lean (2008) computed applying Eq. (4) to the fractions taken from the CP control member (CP Det NP as in the previous section)	Used as a comparison to find the added value of running an ensemble
FSS^{NEP}	FSS computed applying Eq. (3) to the fractions of CP-ENS members exceeding a threshold (CP NEP as in the previous section)	Ability of the ensemble to predict the probability of threshold exceedance
FSS_{ens_mean}	The deterministic FSS is computed on the ensemble mean of the rainfall field using Eq. (4)	Ability of the ensemble mean to predict threshold exceedance
$eFSS^{mean}$	Mean of the deterministic FSS computed for each individual ensemble member, defined in Dey et al. (2014)	Average ability of each ensemble member to predict threshold exceedance
$dFSS_{mean}$	Mean of the deterministic FSS computed between all ensemble member-member pairs, defined in Dey et al. (2014)	Measure of spread of ensemble members; high $dFSS_{mean}$ shows low spread as all members are similar; ideally would be equal to $eFSS_{mean}$
$dFSS^{std}$	Standard deviation of the deterministic FSS computed between all ensemble member–member pairs, defined in (Dey et al. 2014)	Measure of the range of dFSS values; for a fixed $dFSS_{mean}$, small $dFSS_{std}$ suggests rainfall occurs in slightly offset locations between all members, whereas large $dFSS_{std}$ suggests that most ensemble members produce rainfall in the same location but with a few outlier members

between ensembles for the CP ensemble. Both of the analyses lead to the conclusion that CP ensemble is underdispersive, i.e., not able to capture the expected error associated with either the rainfall amount or the rainfall spatial patterns. In particular, spatial spread was shown to be lower for 0600 and 1800 UTC initializations. Also, the RMSE-spread comparison showed that global ensemble is more underdispersive than CP ensemble.

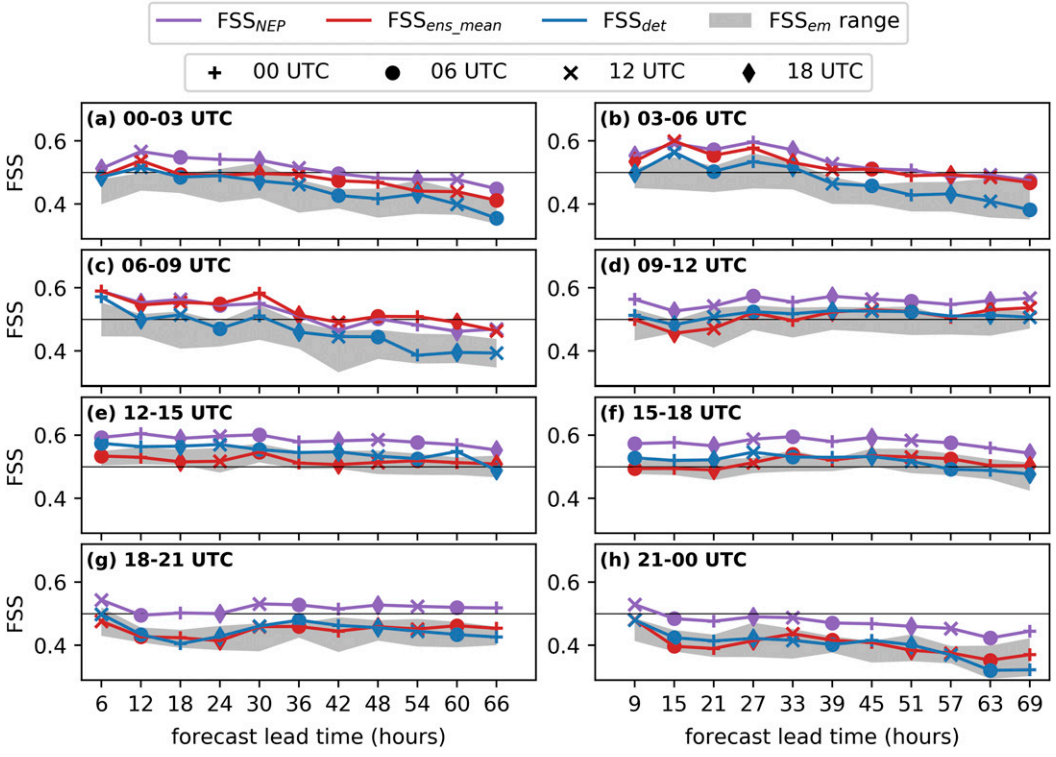

FIG. 13. FSS scores (defined in Table 1) as a function of forecast lead time for 3-h accumulation periods. The gray shading shows the range of FSS scores for individual ensemble members. FSS is computed for a neighborhood of $n = 23$ (~250 km) for rainfall exceeding the 97th percentile. Different markers correspond to different model initialization times.

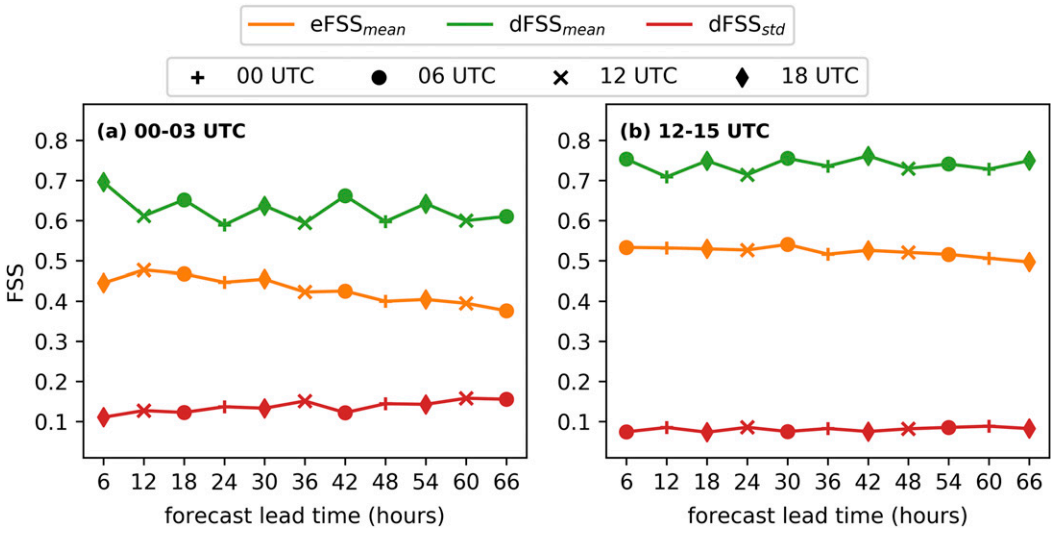

FIG. 14. The eFSS$_{mean}$, dFSS$_{mean}$, and dFSS$_{std}$ (defined in Table 1) as a function of forecast lead time for 3-h accumulation periods (a) 0000–0300 UTC and (b) 1200–1500 UTC. FSS is computed for a neighborhood of $n = 23$ (~250 km) for rainfall exceeding the 97th percentile. Different markers correspond to different model initialization times.

3) *Forecast skill*: Neighborhood probabilistic forecasts based on CP ensemble were shown to be generally more skillful than those generated from either the global ensemble or the respective (deterministic) control members. FSS was found to decrease with percentile threshold and forecast hour (although showing diurnal signals). However, FSS values remain quite low compared to similar analysis performed in the midlatitudes (e.g., Schwartz 2019). CP ensemble forecasts were proven to be more skillful than global forecasts also for the 24-h accumulation, which is the accumulation current weather warnings in East Africa are based on, although the ensembles were only marginally better than the control members for 24-h accumulations. In terms of useful scale, the ensembles were better than their respective control members, though this improvement was generally not statistically significant. The CP ensemble has a useful scale 100 km smaller than global ensemble, which is statistically significant, although a similar improvement was found when comparing the CP Det to global ensemble. ROC areas revealed generally greater discriminating skill by the CP ensemble forecasts, with higher differences for greater thresholds (not shown).
4) *Probabilistic guidance*: The FSS of the deterministic CP model often exceeded that of the ensemble mean and the mean FSS of the individual ensemble members (corresponding to postage stamps). However, the probability of threshold exceedance (CP NEP) was shown to be the most skillful forecast product, highlighting the value of the probabilistic information provided by the CP ensemble. Therefore, this is the product that local forecasters should look at.

A decomposition of the RMSE (see the appendix) indicates that the RMSE is dominated by the forecast variance for the CP ensemble, rather than the bias. Therefore, a bias correction alone may not be sufficient in leading to a more skillful forecast and future efforts should therefore focus on the lack of dispersion. Underdispersiveness is a well-known issue in the meteorological community and research is ongoing to improve this, by understanding the impact of initial, boundary and physical perturbations, as well as postprocessing techniques (e.g., time-lagging, Porson et al. (2020) and references therein). Initial conditions perturbations have a bigger impact in terms of spread and forecast quality in the first hours of forecast integration, whereas boundary conditions dominate for longer lead times over small domains (Vié et al. 2011; Kühnlein et al. 2014; Porson et al. 2020; Dipankar et al. 2020). Boundary conditions perturbations are provided by the global driving model. In regard to perturbations of the initial conditions, there are different ways to generate them (Tennant 2015). Here, a downscaling approach is used: Kühnlein et al. (2014) and Tennant (2015) showed a good performance of the downscaled convective-scale ensemble, especially under conditions of relatively weak synoptic forcing (i.e., convective rainfall). Arguably, the most appealing way to improve CP ensemble spread is way to improve the CP ensemble spread is to improve the spread of the initial conditions of the parent driving ensemble. Porson et al. (2019), for instance, showed that that perturbing the sea surface temperatures (SSTs) in the initial conditions of the parent model generates a higher spread also in the driven CP ensemble than just having fixed SSTs. Another way to enhance the ensemble spread is through the representation of model error in the physics scheme (Bouttier et al. 2012). Whether changes in the physical perturbations have a greater impact than changes in the driving model would depend also on the synoptic forcing (A. Porson 2020, personal communication).

While this study has focused on the CP ensemble and potential ways to improve its performance, it has also demonstrated the

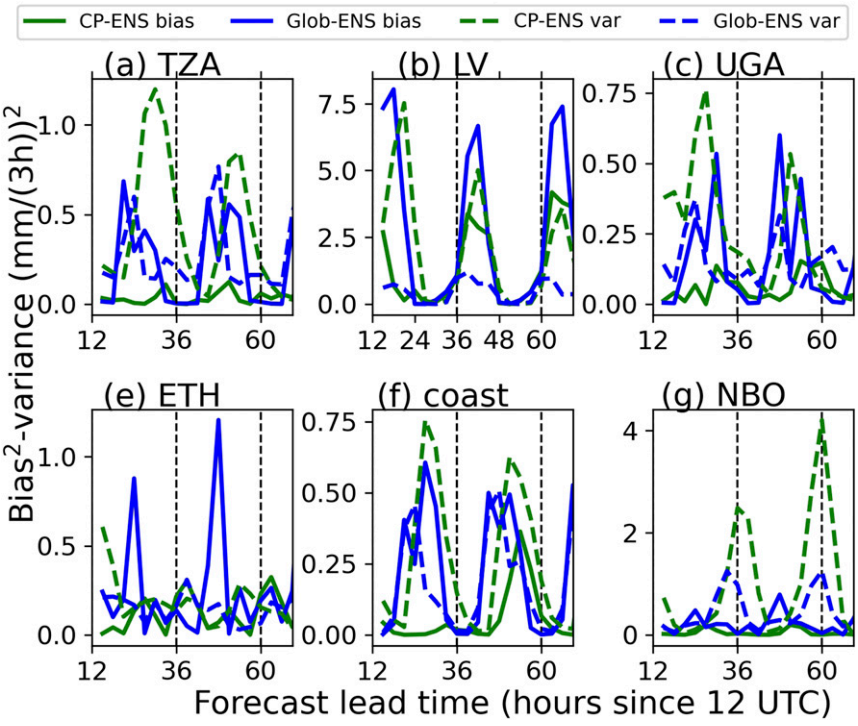

FIG. A1. Bias (solid lines) and variance (dashed lines) decomposition of the mean square error (MSE) of the ensemble mean of the 3-h rainfall accumulation averaged over the different subregions. The values on the *x* axis represent the starting forecast hours for each accumulation period (e.g., 12 is the 3-h accumulation period between 12 and 15 h). Solid (dashed) black vertical lines refer to the valid time of 0000 UTC (1200 UTC).

value of CP deterministic forecasts, which outperform the global ensemble in many ways. Continued evaluation and improvement of CP determinstic forecasts will clearly play an essential role to the forecasting system in East Africa. As advocated by Woodhams et al. (2018), there continues to be a need for more in situ observations (ground and upper air), whose assimilation could increase the CP forecast skill further, especially in the first hours of integration.

Convection-permitting ensemble simulations are only recently being explored for operational forecasting in the tropics. While limited to a brief period of only 24 cases, the findings of this study, should therefore stimulate further investigations in other tropical regions. Future work should involve verification over a longer period or larger set of cases to corroborate the added value by CP ensemble in the tropics. In parallel to the model development point of view presented above, more detailed probabilistic forecast guidance and advice to forecasters is essential for successful adoption of CP ensemble for operational forecasting in the tropics. A future testbed is being planned in the African SWIFT project to investigate how best to exploit information from the CP ensemble for operational forecasting.

Acknowledgments. This work was supported by U.K. Research and Innovation as part of the Global Challenges Research Fund, Grant NE/P021077/1 (GCRF African SWIFT). The authors thank all the people involved with the organization of the SWIFT testbed, in particular the Kenya Meteorological Department for providing the facilities. Finally, three anonymous reviewers are acknowledged for their helpful suggestions to improve this manuscript.

Data availability statement. Model data used for this study are available on the Met Office Managed Archive Storage System (MASS) with the following path: moose:/devfc/u-be957. More information on how to get access to the data can be found at https://www.ceda.ac.uk/blog/access-to-the-met-office-mass-archive-on-jasmin-goes-live/. The GPM-IMERG data were provided by the NASA Goddard Space Flight Center's Precipitation Measurement Missions Science Team and Precipitation Processing System, which develop and compute the GPM IMERG as a contribution to GPM, including data archiving at the NASA GES DISC.

APPENDIX

Bias-Variance Decomposition of the RMSE

In the main text we have calculated the RMSE of the ensemble mean. Here we apply the bias-variance decomposition of the mean squared error (MSE) (Kohavi and Wolpert 1996):

$$\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{var}. \quad \text{(A1)}$$

Results are shown in Fig. A1. First, biases for Glob-ENS are higher than for the CP-ENS. Also, it can be seen that the largest contribution to the MSE comes from the bias for the global ensemble and from the variance for the CP-ENS for most of the regions.

REFERENCES

Bechtold, P., J.-P. Chaboureau, A. Beljaars, A. K. Betts, M. Köhler, M. Miller, and J.-L. Redelsperger, 2004: The simulation of the diurnal cycle of convective precipitation over land in a global model. *Quart. J. Roy. Meteor. Soc.*, **130**, 3119–3137, https://doi.org/10.1256/qj.03.103.

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, https://doi.org/10.1002/met.1435.

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.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776, https://doi.org/10.1002/qj.394.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–189, https://doi.org/10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Bush, M., and Coauthors, 2019: The first Met Office Unified Model/JULES regional atmosphere and land configuration, RAL1. *Geosci. Model Develop.*, **13**, 1999–2029, https://doi.org/10.5194/gmd-2019-130.

Cafaro, C., T. H. A. Frame, J. Methven, N. Roberts, and J. Bröcker, 2019: The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble. *Quart. J. Roy. Meteor. Soc.*, **145**, 1780–1798, https://doi.org/10.1002/qj.3531.

Camberlin, P., W. Gitau, O. Planchon, V. Dubreuil, B. M. Funatsu, and N. Philippon, 2018: Major role of water bodies on diurnal precipitation regimes in Eastern Africa. *Int. J. Climatol.*, **38**, 613–629, https://doi.org/10.1002/joc.5197.

Chamberlain, J. M., C. L. Bain, D. F. A. Boyd, K. McCourt, T. Butcher, and S. Palmer, 2014: Forecasting storms over Lake Victoria using a high resolution model. *Meteor. Appl.*, **21**, 419–430, https://doi.org/10.1002/met.1403.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Dey, S. R., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Dezfuli, A. K., C. M. Ichoku, G. J. Huffman, K. I. Mohr, J. S. Selker, N. van de Giesen, R. Hochreutener, and F. O. Annor, 2017a: Validation of IMERG precipitation in Africa. *J. Hydrometeor.*, **18**, 2817–2825, https://doi.org/10.1175/JHM-D-17-0139.1.

——, ——, K. I. Mohr, and G. J. Huffman, 2017b: Precipitation characteristics in West and East Africa from satellite and in situ observations. *J. Hydrometeor.*, **18**, 1799–1805, https://doi.org/10.1175/JHM-D-17-0068.1.

Dipankar, A., and Coauthors, 2020: SINGV: A convective-scale weather forecast model for Singapore. *Quart. J. Roy. Meteor. Soc.*, **146**, 4131–4146, https://doi.org/10.1002/qj.3895.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

Fink, A. H., and Coauthors, 2017: Mean climate and seasonal cycle. *Meteorology of Tropical West Africa: The Forecasters' Handbook*, D. J. Parker and M. Diop-Kane, Eds., John Wiley & Sons, 1–39, https://doi.org/10.1002/9781118391297.ch1.

Flack, D. L. A., S. L. Gray, R. S. Plant, H. W. Lean, and G. C. Craig, 2018: Convective-scale perturbation growth across the spectrum of convective regimes. *Mon. Wea. Rev.*, **146**, 387–405, https://doi.org/10.1175/MWR-D-17-0024.1.1

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, https://doi.org/10.1175/JHM-D-14-0008.1.

Frogner, I.-L., A. T. Singleton, M. Ø. Køltzow, and U. Andrae, 2019: Convection-permitting ensembles: Challenges related to their design and use. *Quart. J. Roy. Meteor. Soc.*, **145**, 90–106, https://doi.org/10.1002/qj.3525.

Gebhardt, C., S. Theis, M. Paulat, and Z. B. Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

Hanley, K. E., J. S. R. Pirret, C. L. Bain, A. J. Hartley, H. W. Lean, S. Webster, and B. J. Woodhams, 2021: Assessment of convection-permitting versions of the Unified Model over the Lake Victoria basin region. *Quart. J. Roy. Meteor. Soc.*, https://doi.org/10.1002/QJ.3988, in press

Hohenegger, C., and C. Schar, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1794, https://doi.org/10.1175/BAMS-88-11-1783.

Holloway, C. E., S. J. Woolnough, and G. M. S. Lister, 2012: Precipitation distributions for explicit versus parametrized convection in a large-domain high-resolution tropical case study. *Quart. J. Roy. Meteor. Soc.*, **138**, 1692–1708, https://doi.org/10.1002/qj.1903.

Hou, A. Y., and Coauthors, 2014: The Global Precipitation Measurement Mission. *Bull. Amer. Meteor. Soc.*, **95**, 701–722, https://doi.org/10.1175/BAMS-D-13-00164.1.

Huffman, G., and Coauthors, 2018: NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). Algorithm Theoretical Basis Document (ATBD) version 06, Global Precipitation Measurement (GPM) National Aeronautics and Space Administration (NASA), Tech. Rep., NASA, 38 pp., https://storm.pps.eosdis.nasa.gov/storm//IMERG_ATBD_V06.pdf.

Kendon, E. J., N. M. Roberts, C. A. Senior, and M. J. Roberts, 2012: Realism of rainfall in a very high-resolution regional climate model. *J. Climate*, **25**, 5791–5806, https://doi.org/10.1175/JCLI-D-11-00562.1.

Kharin, V. V., and F. W. Zwiers, 2003: 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.

Klasa, C., M. Arpagaus, A. Walser, and H. Wernli, 2018: An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quart. J. Roy. Meteor. Soc.*, **144**, 744–764, https://doi.org/10.1002/qj.3245.

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.

Kohavi, R., and D. Wolpert, 1996: Bias plus variance decomposition for zero-one loss functions. *ICML'96 Proc. 13th Int. Conf. on Machine Learning*, San Francisco, CA, Morgan Kaufmann Publishers Inc., 275–283.

Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt, 2014: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1552–1562, https://doi.org/10.1002/qj.2238.

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, P., Z. Guo, K. Furtado, H. Chen, J. Li, S. Milton, P. R. Field, and T. Zhou, 2019: Prediction of heavy precipitation in the eastern China flooding events of 2016: Added value of convection-permitting simulations. *Quart. J. Roy. Meteor. Soc.*, **145**, 3300–3319, https://doi.org/10.1002/qj.3621.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

Lorenz, E. N., 1969: Predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Maranan, M., A. H. Fink, P. Knippertz, L. K. Amekudzi, W. A. Atiah, and M. Stengel, 2020: A process-based validation of GPM IMERG and its sources using a mesoscale rain gauge network in the West African forest zone. *J. Hydrometeor.*, **21**, 729–749, https://doi.org/10.1175/JHM-D-19-0257.1.

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.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

Maurer, V., N. Kalthoff, and L. Gantner, 2017: Predictability of convective precipitation for West Africa: Verification of convection-permitting and global ensemble simulations. *Meteor. Z.*, **26**, 93–110, https://doi.org/10.1127/metz/2016/0728.

Mittermaier, M., and G. Csima, 2017: Ensemble versus deterministic performance at the kilometer scale. *Wea. Forecasting*, **32**, 1697–1709, https://doi.org/10.1175/WAF-D-16-0164.1.

——, N. Roberts, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteor. Appl.*, **20**, 176–186, https://doi.org/10.1002/met.296.

Pantillon, F., S. Lerch, P. Knippertz, and U. Corsmeier, 2018: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Quart. J. Roy. Meteor. Soc.*, **144**, 1864–1881, https://doi.org/10.1002/qj.3380.

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.

Porson, A. N., S. Hagelin, D. F. Boyd, N. M. Roberts, R. North, S. Webster, and J. C. Lo, 2019: Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Quart. J. Roy. Meteor. Soc.*, **145**, 3004–3022, https://doi.org/10.1002/qj.3601.

——, and Coauthors, 2020: Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Quart. J. Roy. Meteor. Soc.*, **146**, 3245–3265, https://doi.org/10.1002/qj.3844.

Ralph, F. M., and Coauthors, 2013: The emergence of weather-related testbeds linking research and forecasting operations. *Bull. Amer. Meteor. Soc.*, **94**, 1187–1211, https://doi.org/10.1175/BAMS-D-12-00080.1.

Raynaud, L., and F. Bouttier, 2016: Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quart. J. Roy. Meteor. Soc.*, **142**, 854–866, https://doi.org/10.1002/qj.2686.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann, 2017: On the forecast skill of a convection-permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, https://doi.org/10.5194/gmd-10-35-2017.

Schwartz, C. S., 2019: Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Mon. Wea. Rev.*, **147**, 2997–3023, https://doi.org/10.1175/MWR-D-18-0452.1.

——, and Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVAR, ensemble Kalman filter, and "Hybrid" variational–ensemble data assimilation systems. *Mon. Wea. Rev.*, **142**, 716–738, https://doi.org/10.1175/MWR-D-13-00100.1.

——, and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood

approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

——, ——, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

Skok, G., and N. Roberts, 2016: Analysis of fractions skill score properties for random precipitation fields and ECMWF forecasts. *Quart. J. Roy. Meteor. Soc.*, **142**, 2599–2610, https://doi.org/10.1002/qj.2849.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

Stein, T. H. M., and Coauthors, 2019: An evaluation of clouds and precipitation in convection-permitting forecasts for South Africa. *Wea. Forecasting*, **34**, 233–254, https://doi.org/10.1175/WAF-D-18-0080.1.

Sun, X., and Coauthors, 2020: A subjective and objective evaluation of model forecasts of Sumatra squall events. *Wea. Forecasting*, **35**, 489–506, https://doi.org/10.1175/WAF-D-19-0187.1.

Tan, J., G. J. Huffman, D. T. Bolvin, and E. J. Nelkin, 2019: IMERG V06: Changes to the Morphing Algorithm. *J. Atmos. Oceanic Technol.*, **36**, 2471–2482, https://doi.org/10.1175/JTECH-D-19-0114.1.

Tennant, W., 2015: Improving initial condition perturbations for MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **141**, 2324–2336, https://doi.org/10.1002/qj.2524.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, https://doi.org/10.1017/S1350482705001763.

Torn, R. D., 2010: Ensemble-based sensitivity analysis applied to African easterly waves. *Wea. Forecasting*, **25**, 61–78, https://doi.org/10.1175/2009WAF2222255.1.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, https://doi.org/10.1175/2010MWR3487.1.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting, 2018: Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Wea. Forecasting*, **33**, 369–388, do, https://doi.org/10.1175/WAF-D-17-0127.1.

Walters, D., and Coauthors, 2017: The Met Office Unified Model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geosci. Model Dev.*, **10**, 1487–1520, https://doi.org/10.5194/gmd-10-1487-2017.

Woodhams, B. J., C. E. Birch, J. H. Marsham, C. L. Bain, N. M. Roberts, and D. F. A. Boyd, 2018: What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Mon. Wea. Rev.*, **146**, 2757–2780, https://doi.org/10.1175/MWR-D-17-0396.1.