Optimizing Precipitation Forecasts for Hydrological Catchments in Ethiopia Using Statistical Bias Correction and Multi-Modeling

Sippora Stellingwerf\(^1\), Emily Riddle\(^2\), Thomas M. Hopson\(^3\), Jason C. Knievel\(^2\), Barbara Brown\(^2\), and Mekonnen Gebremichael\(^1\)

\(^1\)Wageningen University and Research, Amersfoort, Netherlands, \(^2\)National Center for Atmospheric Research, Boulder, CO, USA, \(^3\)University of California, Los Angeles, CA, USA

Abstract Accurate rainfall forecasts on timescales ranging from a few hours to several weeks are needed for many hydrological applications. This study examines bias, skill and reliability of four ensemble forecast systems (from Canada, UK, Europe, and the United States) and a multi-model ensemble as applied to Ethiopian catchments. By verifying these forecasts on hydrological catchments, we focus on spatial scales that are relevant to many actual water forecasting applications, such as flood forecasting and reservoir optimization. By most verification metrics tested, the bias corrected European model is the best individual model at predicting daily rainfall variations, while the Canadian model shows the most realistic ensemble spread and thus the most reliable forecast probabilities, including those of extreme events. The skill of the multi-model ensemble outperforms individual models by most metrics, and is skillful up to 9 days ahead. Skill is higher for the 0–5 day model accumulation than for the first 24 h, suggesting that timing errors strongly penalize the skill of forecasts with shorter accumulation periods. Due to seasonality in the model biases, bias correction is best applied to each month individually. Forecasting extreme rainfall is a challenge for Ethiopia, especially over mountainous regions where positive skill is only reached after bias correction. Compared to individual models, the multi-model ensemble has a higher probability of detecting extreme rainfall and a lower false alarm rate, with usable skill at 24 h lead times.

1. Introduction

Timely, reliable forecasts of precipitation are of great importance for a broad range of hydrometeorological applications, such as East Africa's rapidly expanding irrigated agriculture (Laurance et al., 2013), flood warning systems, and hydropower (Conway et al., 2017). Many meteorological centers around the world run global numerical weather prediction (NWP) models that produce quantitative precipitation forecasts for East Africa. These ensemble forecasts are publicly available in real-time or near real-time and can be directly fed into hydrological models or used in real-time to produce alerts and guidance to farmers, reservoir managers, and governments. While a few studies have examined the skill of these models for Africa (Kipkogi et al., 2016; Mutemi et al., 2007), the utility of this suite of model output has not yet been fully assessed for hydrological applications over Ethiopia. In this study, we analyze operational forecasts from four forecast centers (NCEP, UKMET, ECMWF, and Canada) obtained through the THORPEX International Grand Global Ensemble (TIGGE; Bougeault et al., 2010; Swinbank et al., 2016) archive. We examine how a multi-model ensemble mean forecast performs compared with ensemble mean forecasts from individual models. We determine how many days in advance these forecasts are skillful, and assess how increasing the temporal scale of forecasts (daily vs. 5-day rainfall accumulations) influences their skill. Finally, we explore how well the bias corrected ensemble distributions can be used to estimate probabilities of extreme rainfall events. In all of these investigations, we use a variety of standard metrics and visualization tools to thoroughly understand the advantages and disadvantages of different forecast systems and strategies. The remainder of the study is divided into four sections. The background section provides a geographic context for the study, including drivers of precipitation variability over Ethiopia, and previous research on forecast verification in this area. The methods section provides a detailed description of the models, observational data, our bias correction technique, and details about our metrics and assessment tools. Results are presented next followed by a summary of our main conclusions.
2. Background

2.1. Geographical and Meteorological Context

Ethiopia is located in East Africa, bordered by Sudan, South Sudan, Kenya, Somalia, Djibouti, and Eritrea. Major rivers include the Blue Nile and Atbarah river, both draining to the Nile; the Awash river ending in Lake Abbe in Djibouti; the Jubba river, flowing into the Indian Ocean; and the Omo draining into Lake Turkuna on the border with Kenya. In total, Ethiopia has 12 major river basins, all of which originate from within Ethiopia itself (Awulachew et al., 2007) (see Table 1). The elevation ranges from 125 m below sea level in the Danakil Depression up to 4,550 m above sea level in the Semien Mountains (see Figure 1). The complex topography of the region defines distinct climatic areas, with arid desert to the low-lying east and temperate conditions in the mountains (Fazzini et al., 2015).

The seasonal migration of the Intertropical Convergence Zone (ITCZ), which is regionally manifest over central Africa as the Africa Intertropical Front (ITF) (Lélé & Lamb, 2010), is the primary driver of seasonal rainfall variability in Ethiopia (Diro et al., 2008). In southern Ethiopia, precipitation patterns are bimodal with a rainfall peak related to the arrival of the ITCZ in spring and the subsequent retreat in autumn. Here, a primary spring rainy season, called belg, occurs in March, April, and May (MAM), and a shorter, weaker rainy season occurs in October and November (ON) (Riddle & Cook, 2008). In Northern Ethiopia, which is

| ID | Name       | Size [km²] | ID | Name       | Size [km²] |
|----|------------|------------|----|------------|------------|
| 1  | Mereb Gash | 5,965      | 8  | Baro Akobo | 76,203     |
| 3  | Tekeze     | 83,476     | 9  | Ogaden     | 80,009     |
| 4  | Denakil    | 63,853     | 10 | Wabi Shebele | 202,220   |
| 5  | Abbay      | 198,891    | 11 | Omo Gibe   | 78,189     |
| 6  | Awash      | 110,439    | 12 | Rift Valley | 51,989     |
| 7  | Ayisha     | 4,321      | 13 | Genale Dawa | 172,133    |

Table 1. IDs, Names, and Sizes of all 12 Ethiopian Catchments (by Previous Convention, There is No. #2)

Figure 1. A topographic map of Ethiopia. Source: Mengistul et al. (2013).
roughly the northernmost extent of the ITCZ, the rainfall pattern is monomodal with a peak in the summer. This June to September rainy season is referred to as *kiremt*.

Ethiopia's complex topography and its location at the boundary between the Central African ITF and the South Asian monsoon system make it a complicated region to study meteorologically. For example, studies have identified between four and six homogeneous sub-regions, each with different seasonal rainfall cycles and/or different modes of interannual variability (e.g., Diro et al., 2008; Gissila et al., 2004; Riddle & Cook, 2008). During the boreal spring, a low-level cross-equatorial jet, the Somali Jet, begins to form, bringing moist air from the southern Indian Ocean onto the southern flank of the Ethiopian highlands (Riddle & Cook, 2008; Riddle & Wilks, 2013). By June, this moisture source is diverted as the Somali Jet extends across the Arabian Sea toward India, becoming the inflow to the Indian Monsoon. During the boreal summer, low-level flows from the Congo Basin to the southwest, and the Red Sea to the northeast converge over Ethiopian highlands fueling the summer Ethiopian rainy season (Jury, 2011; Jury & Chiao, 2014; Viste & Sorteberg, 2013b). While the Red Sea and Indian Ocean are the largest of these moisture sources (Viste & Sorteberg, 2013b), variability in moisture transport from the west (Congo Basin) has a greater impact on rainfall variability (Nicholson, 2017; Riddle & Wilks, 2013; Viste & Sorteberg, 2013a).

Tropical waves can be a source of predictability in the tropics at timescales ranging from several hours to several weeks. African easterly waves (AEW), caused by disturbances in upper tropospheric jets, can be a source predictability at timescales of a few days, especially over central and western Africa. A few studies have suggested that these westward traveling waves may have precursors over the Arabian Sea and Ethiopia. For example, Mekonnen and Rossow (2011) and Mekonnen and Thornicroft (2016) show that wave activity in the upper-level tropospheric winds can concentrate disorganized convection over the Arabian Sea into larger mesoscale convective systems that can transition into AEWs over Ethiopia and Sudan (Jury, 2018). At longer timescales up to 30 days, the Madden–Julian Oscillation (MJO) influences rainfall over the arid regions of eastern Ethiopia (Schlueter et al., 2019), as well as coastal Kenya and Tanzania (Pohl & Camberlin, 2006a, 2006b) during the transition seasons. An analysis of the relative influence of all types of equatorial waves on African precipitation, however, suggests that the impact of tropical waves over Ethiopia is limited compared with central and western Africa (Schlueter et al., 2019). Together, these studies highlight the meteorological complexity of the region for rainfall forecasting.

2.2. Numerical Weather Prediction for Ethiopia

The global NWP models used in this study are physics-based numerical models that estimate how the atmosphere evolves with time, given certain initial and boundary (surface) conditions. These models are the backbones of most weather forecasts across the globe. However, NWP models face some challenges in the tropics. Recent studies have shown that error growth over the initial 1–2 days of a forecast is much larger in the tropics than in the midlatitudes (e.g., Bechtold, 2019; Žagar, 2017). These errors are due in part to a sparser observing network for model initialization, especially of upper-level winds (Žagar, 2017), which are important for concentrating and propagating convective disturbances. In addition, the models struggle to realistically model the character of tropical convection (e.g., Stephens et al., 2010). They produce too many days with light rain, underestimate the rainfall rates of the highest intensity events (Trenberth et al., 2017), are unable to capture realistic diurnal cycles (e.g., Knievel et al., 2004; Yang & Slingo, 2001), and are often not run at high enough resolution to properly model the interaction of atmospheric flows with complex topography (e.g., Woodhams et al., 2018). Finally, the ensemble systems tend to be overconfident, or under-dispersive, because they fail to properly model the unpredictable, stochastic aspects of rainfall generation (Watson et al., 2017).

Statistical methods can correct some of these model shortcomings. The models used in this study are run in ensemble mode: each model is run multiple times, using different perturbations of the initial conditions to create an ensemble of possible outcomes. This ensemble represents the uncertainty in the forecasts. Averaging together multiple ensemble members serves to cancel out the noise in individual model runs, and enhance the model signal. Averaging the ensembles from several different forecast centers together further reduces noise and tends to improve the forecast, especially if the individual models have opposing errors (Hamill, 2012; Park et al., 2008). Another key aspect of why a multi-modeling approach is often so successful is explained by Hagedorn et al. (2005). Each model has its own strengths and weaknesses.
Which model is “the best” single model often depends on the particular situation. A model that for example performs well for predicting summer rainfall might not be the best model for winter rainfall. The multi-model however performs consistently well, especially when compared to single models instead of the (varying) best model. Even models that seem to perform poorly, might enhance the multi-model by the aforementioned opposing errors. Only when a particular model consistently performs the worst, it indeed degrades the skill of the multi-model. Vice versa, only when one single model consistently performs the best, the multi-model will have lower skill. Next to the multi-modeling approach, known biases in the forecasts can also be corrected using historical records of forecasts and observations.

Only a few previous studies have focused on assessing NWP model skill over eastern Africa and Ethiopia at timescales ranging from 24 h to several weeks. Vogel et al. (2020) provide a comprehensive analysis of the probabilistic skill of the European and Canadian models for the global tropics. Woodhams et al. (2018) compare the results of the UKMET operational global model with a limited-domain high-resolution convection permitting (CP) model over eastern Africa for forecasts up to 48 h. Those authors find significantly better representation of the diurnal cycle and 3-h rainfall rates in the CP model, but only modest improvements in the skill of 24 h averages over the global model. Mutemi et al. (2007) assesses the skill of seven ensemble forecast models over Africa in general, comparing individual model forecasts with a multi-model mean and a simple method for weighting the models and correcting the mean biases in the forecasts. They showed that the multi-model ensemble performed better than the individual best model, but care should be taken in the selection of forecasts to be included in the multi-model: “bad” models degrade the skill of the multi-model. From the NWP models available in the TIGGE archive, we have performed an initial skill analysis and selected four of the best performing state-of-the-art ensemble prediction systems to create a multi-model ensemble.

As in our current study, Kipkogei et al. (2016) use the TIGGE multi-model ensemble to assess skill over the Greater Horn of Africa. They also study four TIGGE models: ECMWF (Europe), NCEP (US), UKMET (UK), and CPTEC (Brazil) and show that the skill, as measured by spatial correlations and the equitable threat score (ETS; also a metric of the spatial pattern of rainfall), is the highest for a weighted bias corrected multi-model average (super-ensemble), followed by the UKMET model, uncorrected for biases. The authors suggest that more work is needed to test other bias correction and multi-modeling methods, and to determine if the skill scores are high enough to be of practical use.

Our study extends this previous work in several ways along the lines of what Kipkogei et al. (2016) suggest. First, we use and test a bias correction approach, called quantile-to-quantile (Q2Q) mapping (Hopsen & Webster, 2010), that corrects the full rainfall distribution instead of just the mean rainfall. We also test whether doing a separate bias correction on each month of data improves the results over performing the bias correction on the aggregated year-round data. Second, our choice of temporal and spatial scales is motivated by real-world water management challenges, and we examine skill separately for individual catchments (see Figure 2) as well as average the results regionally. Compared to Kipkogei et al. (2016) who validate spatial rainfall patterns daily on 625 km² grid boxes, we examine skill on catchments ranging in size between 4,000 and 200,000 km² (see Table 1), for both 24-h and 5-day averages (the latter being the day 0–5 and day 5–10 average rainfall accumulations). Given the challenges that NWP models have in the tropics, it is useful to test whether averaging to these large temporal and spatial scales improves skill. Third, we extend the skill analysis to include many more skill metrics and visualization tools frequently used in forecast verification. This approach gives a more complete picture of the utility of the forecasts for specific catchments, as well as insight into reasons why different models and methods succeed or fail.
3. Methodology

3.1. Data Description

3.1.1. Forecast Data

The gridded forecast data of four ensemble prediction systems (EPSs) are used in this study. Some descriptives of the models are given in Table 2. Model data produced from 2011 to 2016 are used in this study. Mladek (2019) gives a full description of the NWP models in the TIGGE archive.

3.1.1.1. Multi-Model Ensemble

The multi-model ensemble that we use here is a collection of equally weighted ensemble members of all available models mentioned in Table 2. UKMET’s forecasts go out to seven days. Therefore, the multi-model consists of 101 members for lead times up to 7 days, and 90 members for longer lead times. The forecasts are all from 00Z initialization cycles. Any cycle without complete fields out to the final lead time of the respective model are dropped. The TIGGE archive provides the forecasts fields for all models at 0.5° × 0.5°. Given this, it is important to note that within the smallest catchment, effectively only a couple of forecast grids represent the basin, while close to one hundred represent the largest catchment. As such, this has implications for the expected basin-averaged rainfall forecast skill of each catchment, where all things being equal, forecast displacement errors diminish as the basin spatial scale increases.

3.1.2. Observational Data

The observational data set is a merged satellite product from two near real-time satellite products that use measurements of microwave and infrared radiation in their construction: the GSMaP (Global Satellite Mapping of Precipitation (Aonashi et al., 2009)) and NASA’s TRMM (Tropical Rainfall Measuring Mission; Huffman et al., 2007). The GSMaP product is released hourly with a horizontal spacing of 0.1 by 0.1 lat/lon degrees; while the NASA’s TRMM product is released every three hours, with an horizontal spacing of 0.25 by 0.25 lat/lon degrees. Because satellite rainfall estimates from a number of centers have been shown to have significant random and systematic errors in our study region (Hirpa et al., 2010) and lack of country-wide rain gauge and/or radar data availability precludes the application of a geographically extensive corrective error algorithm for these products separately, we have merged these two products together to (presumptively) reduce the errors (random and systematic) from using only one product.

To merge the two products, we first brought the two products to a common spatial and temporal grid, bilinearly interpolating spatially the NASA product to the higher resolution GSMaP 0.1° by 0.1° grid, and temporally accumulating the GSMaP data up to the NASA 3 h time-step. Then, where both are available, the two products were averaged using equal weighting grid-by-grid, time-step-by-time-step; using only the single product where only one product is available. Finally, the 0.1° by 0.1° merged satellite product is spatially averaged over the Ethiopian catchments (using full grid point weights for grids falling fully within the catchment, and fractional grid point weights along the catchment boundaries).

3.2. Bias Correction Using Quantile-to-Quantile Mapping

Bias correction by so-called Q2Q mapping (Hopson, 2005; Hopson & Webster, 2010) is a relatively straightforward technique to remove systemic biases from ensemble forecasts by requiring that the forecasts be replaced with samples from the cumulative distribution function (CDF) of the associated observational historical record. This is done, in short, by replacing the forecast percentile (as determined relative to the model historical CDF) with the observation value associated with the equivalent percentile (as determined relative to the observational historical CDF). This is visually explained in Figure 3. In this study, Q2Q bias correction is done for catchment-integrated predictions. For ensemble forecasts, every ensemble member is adjusted separately to preserve the ordering by lead time (i.e., temporal covariance) of each ensemble.

The first step to bias correct a catchment’s rainfall ensemble forecast value from an ensemble forecast system is to select data to represent a well-sampled model hindcast CDF, one for each lead time. Similarly,
observational data is selected to represent a well-sampled observational CDF (but with no lead time dependence). These CDFs are then stored and used in the following two-step process: for each lead time of a particular ensemble member, the forecast value's percentile is determined from the model CDF; then the equivalent percentile of the observed CDF is extracted, and used in replacement of the forecast. This is done for each lead time for a given ensemble member. Please see the Appendix A for more details.

We apply the Q2Q bias correction separately to each model, each catchment, each accumulation period, and each lead time (while maintaining the lead time ordering of each ensemble member). To test the importance of seasonal variations in the biases, we compare the results of a Q2Q bias correction performed on the aggregated data from the whole year, with results of a Q2Q bias correction performed on individual months, separately. From the whole set of ensemble members, each forecast day we randomly select three ensemble members to create the forecast climatology. This reduces computation time and also limits using the same forecast data for both Q2Q bias correction and validation, while still having a sufficiently large sample size. As this data set consists of 7 years of data, leaving out days with missing data, the sample size that makes up the model climatologies is still over 5,000 (the adequacy of the reduced 3-member climatologies was also confirmed by testing subsets of data, one with three members and the other with the full ensemble set, and found no appreciable difference in results).

Because this Q2Q bias correction is applied to precipitation forecasts, there will be many occasions where the predicted precipitation is zero. Assume some model predicts zero precipitation 20% of the time. This means the lower 20% of the CDF all consist of zero precipitation. In this case we randomly pick a percentile of the model CDF from this 20%. If the cumulative chance for zero precipitation is larger in the observed climatology, say 30%, this will always result in an “adjusted” forecast also producing zero precipitation. If on the other hand the cumulative chance for zero precipitation is smaller in the observed climatology, in some random occasions the zero precipitation forecast will be corrected to producing some (typically light) precipitation. Using the same example where the modeled chance of zero precipitation is 20%, but now the
observed chance is only 10%, the model will be corrected to producing (some) precipitation on average half of the time it predicted zero precipitation.

On the right side of the distribution there are also some considerations to be made. The observational data set in this study only consists of 7 years of data. This means that very rare rainfall events that happened more than 7 years ago have not been captured. Also, the model climatology is three times as large (3 random ensemble members per day) as the observational climatology (1 value per day). Because of this, the highest (fractional) percentile present in the forecast climatology is higher than the (fractional) highest percentile in the observed climatology (consider, as an example, an observational data set with a sample size of 100 and a forecast data set with a sample size of 300. Within the observational data set, the sample with the highest rainfall amount corresponds to a percentile of $(100 - 1)/(100) = 0.990$, while in the case of the forecast data set the highest rainfall amount corresponds to a percentile of $(300 - 1)/(300) = 0.997$). When the percentile of an uncorrected ensemble member exceeds the highest percentile of the observed climatology, it is constrained to the maximum observed rainfall. More sophisticated methods such as the GEV method (Katz et al., 2002) could be applied, but are beyond the scope of this research.

### 3.3. Statistical Validation Techniques

For statistically validating the original and bias corrected forecasts against data from satellite observations, we will make use of several verification metrics. These include summary measures, difference and correlation metrics, so-called skill scores, and accuracy measures for binary forecasts. An overview of metrics used in this study are given in Table 3. A theoretical background of these metrics is given in the Appendix A.

We assess the performance of the ensemble mean precipitation using the deterministic measures of difference, correlation, and skill. The probabilistic accuracy is assessed using the brier skill score as well as plots illustrating the reliability and resolution. Finally, we evaluate the performance of binary forecasts for high-intensity rainfall.

To assess the uncertainty and related statistical significance of differences between models and accuracy changes after applying bias correction, we apply a technique called bootstrapping (Efron, 1979). In short, by randomly sampling the forecast data many times, different outcomes of accuracy measures are simulated. Suppose we have 100 days of rainfall predictions and observations; we pick 100 random days to create a sample. This is done with replacement, so that it can occur that one day is picked more than once, or not at all. We repeat this procedure many times, in our study 1,000 times, to simulate 1,000 different possible

| Forecast type | Metric type | Metric name | Abbreviation |
|---------------|-------------|-------------|--------------|
| Deterministic Difference Mean absolute error MAE | Root mean squared error RMSE | Bias bias | Index of agreement $d$ |
| Deterministic Correlation Coefficient of determination $R^2$ | Anomaly correlation AC | Skill | Skill score of the root mean squared error SS$_{\text{RMSE}}$ |
| Deterministic | binary | Accuracy of binary forecasts | |
| Deterministic | Skill | Brier skill score BSS | |
| Deterministic | Skill | Threat score $T$ | |
| Deterministic | Skill | Probability of detection POD | |
| Deterministic | Skill | False alarm rate $F$ | |
| Deterministic | Skill | False alarm ratio FAR | |
| Deterministic | Skill | Success ratio SR | |
| Deterministic | Skill | Bias $B$ | |
values for each metric. The distribution of these 1,000 values can then be used to infer the uncertainty around the mean value of the metric.

Which model performs “best” or “worst” can depend on the metric chosen (Jolliffe & Stephenson, 2012). The broad range of verification metrics used in this study allows for a thorough evaluation of different aspects of model performance. Special focus is given to how model performance varies with the changing of seasons, how model performance changes with lead time and temporal integration time (accumulation time), and finally how accurate models predict heavy rainfall events. Whereas accurate rainfall predictions with longer lead times are more important for water management purposes, accurate rainfall predictions for shorter lead times are essential for flood forecasting.

4. Results and Discussion

This chapter demonstrates the need for bias correction in Section 4.1. In Sections 4.2 to 4.4 we assess the dependency of skill on the time of the year, the temporal characteristics of the forecast, and finally the total rainfall amount.

4.1. Exploratory Data Analysis

Figure 4 shows the climatology of both the observed and forecast precipitation for all 12 Ethiopian catchments. The seasonality of the rainfall varies by region. Catchments 9, 10, and 13, in the southeastern region, show a strongly bimodal rainfall pattern, with rainfall in the MAM and ON months. Catchments 4, 6, and 7, mostly located in the Rift Valley in the northeast, show a different bimodal rainfall pattern, with the second rainfall peak occurring in summer. Catchments 1, 3, and 5 (the northern catchments) show a clear summer rainfall peak, with most rainfall in July and August. Catchments 8, 11, and 12, the southwestern catchments, have a dominant spring (belg) rainfall season with a precipitation maximum usually around April/May. In general, these models tend to systematically overpredict the amount of precipitation during the rainy seasons. Regardless of the bias, skill can still be present as long as the patterns match. Figure 5 shows scatterplots of predicted versus observed rainfall. While the ensemble mean generally produces too much rainfall when observed values are low, extreme rainfall events tend to be underpredicted by the ensemble mean. As we are averaging over multiple models, the multi-model ensemble mean will also tend to be less extreme than the ensemble mean of individual models. As opposed to ensemble means, the individual ensembles are trying to simulate actual realizations (and the variability) of the atmosphere, and are able to produce a more realistic distribution of daily rainfall rates (Figure 6). UKMET is the only model that underestimates the frequency of heavy rainfall events.

4.1.1. Validation Statistics of the Ensemble Means

Table 4 provides validation statistics averaged over the Ethiopian catchments. The table compares the raw output of the NWP model (Raw) to two versions of Q2Q bias corrected models: bias correcting on all model
data at once (All) or bias correcting conditioned on the month (Mon). These statistics are based on ensemble means, not on individual ensemble members. The model data are all 24-h forecasts from initialization (zero day lead time).

All raw model predictions tend to overestimate the average rainfall. Q2Q bias correction successfully removes the bias, significantly decreasing the mean absolute error and the RMSE. The $R^2$ and the index of agreement ($d$) increase slightly, because these metrics are almost insensitive to bias. The AC and SS$_{RMSE}$ both clearly increase after Q2Q bias correction. We have assessed the uncertainty of performance changes for all metrics by applying bootstrapping. The scores of all metrics and all models change significantly ($p < 0.05$) after Q2Q bias correction, except for the SS$_{RMSE}$ for UKMET and the $R^2$ for ECMWF, UKMET, and NCEP. This is denoted with a single asterisk in Table 3. Figure 7 lays focus on the uncertainty of the increase of the SS$_{RMSE}$, as this metric evaluates overall model performance. As systemic biases change over the seasons, monthly Q2Q bias correction is consistently most effective in lowering the error as well as increasing correlation. In the remainder of this chapter bias corrected results are obtained from monthly bias corrected forecasts.

Overall, when averaging over all Ethiopian catchments, the multi-model EPS performs best. This difference in performance of the multi-model EPS is significant ($p < 0.05$) for all models and all metrics, except for the bias (all models have almost zero bias) and the $d$ and AC of ECMWF. This is shown by the double asterisk in Table 3. That the multi-model EPS significantly outperforms other models, is shown in Figure 8. This figure compares the SS$_{RMSE}$ of the NWP models after applying monthly Q2Q bias correction, including its uncertainty. It is also worthwhile to assess the spatial variability of the best scoring models. Figure 9 shows the number of catchments for which each model's ensemble mean scores best based on error and correlation metrics.
Before Q2Q bias correction, the multi-model is most frequently the best when looking at the RMSE, $R^2$, $d$, and AC. Judging a model based on its bias or MAE, there is no clear “winner.” After Q2Q bias correction the multi-model is clearly the best choice for most catchments. The bias is almost negligible after correction (Table 4) and therefore the distribution of best models is more or less random.

The summary statistics and the error and correlation metrics give a good indication of overall model performance, when using the ensemble mean as the forecast. Rank histograms give more information about how reliable the whole set of ensemble members are. Rank histograms bin observations according to their values relative to the sorted values predicted by ensemble members. An under-dispersive EPS will have a U-shaped histogram, with observed values often falling outside the rainfall range predicted by the ensemble members. Vice versa, an over-dispersive EPS will have a large population in the middle ranks. EPSs with a dry bias will have a large population in the highest rank, and vice versa. Figure 10 shows the rank histograms. Models tend to have a wet bias before Q2Q bias correction, which turns into a (smaller) dry bias after the correction. The EPSs are generally under-dispersive before and after Q2Q bias correction, but less so after.

In the remainder of this chapter we will more elaborately explore the effect of Q2Q bias correction on the variability of forecast skill.

### 4.2. Seasonal Variability of Skill

The forecast skill, expressed through the anomaly correlation and the skill score of the RMSE, is temporally variable (Figure 11). The dry season is characterized by the lowest skill scores before applying Q2Q bias correction, influenced by the systemic overestimation of low-intensity rainfall. After removing the systemic
biases, the skill score improves significantly. Skill is highest during the spring (MAM) and autumn (ON) rainy seasons. The predictability of the summer rainy season is lowest.

During the rainy seasons, the multi-model ensemble has the highest skill, quickly followed by ECMWF. After Q2Q bias correction, NCEP performs the worst. There is quite a lot of variability in skill from catchment to catchment (triangles in Figure 11). This is likely due to the regional differences in the topography, as well as (related or non-related) differences of rainfall drivers. These results highlight the complexity of the region both spatially and temporally, with skill scores varying both by catchment and from one season to the next. Further work is needed to better understand the relationship between spatial and temporal variations in these skill scores and the underlying mechanisms that cause daily rainfall variability across Ethiopia.

### 4.3. Skill Versus Lead Time and Integration Time

Skill decreases with lead time. This becomes apparent through both the decrease of the anomaly correlation and the SS_{RMSE} (Figure 12).

The anomaly correlations for predicting 5-day accumulation averages, both the day 0–5 and day 5–10 accumulation averages (shown as crosses in Figure 12) are larger than the skill for predicting 24-h accumulations. Note that the average lead time for day 0–5 (day 5–10) accumulations correspond to two (seven) days, because the accumulation starts on day zero (five). One explanation for this higher skill is that increasing the accumulation time smooths the rainfall peaks. After all, it is more likely for any individual day to have extreme (and harder to predict) rainfall than five consecutive, extremely wet days. Another explanation is

---

Table 4
Validation Statistics Over the Ethiopian Catchments

| Model | Bias | MAE | RMSE | d   | R²  | SS | AC  |
|-------|------|-----|------|-----|-----|----|-----|
| cmc   | Raw  | 0.65| 1.66 | 2.83| 0.72| 0.37| -0.01| 0.38|
|       | All  | 0.16| 1.30 | 2.62| 0.69| 0.38| 0.07| 0.40|
|       | Mon  | -0.01**| 1.27| 2.52| 0.74| 0.42| 0.11| 0.45|
| ecmwf | Raw  | 0.47| 1.52 | 2.82| 0.75| 0.45| 0.00| 0.48|
|       | All  | -0.02| 1.26| 2.58| 0.78| 0.44| 0.09| 0.51|
|       | Mon  | -0.01**| 1.21| 2.49| 0.80**| 0.47*| 0.12| 0.54**|
| ukmet | Raw  | 0.82| 1.71 | 2.92| 0.74| 0.41| -0.04| 0.42|
|       | All  | 0.08| 1.37 | 2.85| 0.75| 0.38| -0.01| 0.45|
|       | Mon  | 0.03**| 1.31| 2.73| 0.77| 0.41*| 0.03**| 0.49|
| ncep  | Raw  | 0.92| 1.80 | 3.17| 0.71| 0.37| -0.13| 0.39|
|       | All  | -0.08| 1.39| 2.93| 0.71| 0.33| -0.03| 0.41|
|       | Mon  | -0.01**| 1.34| 2.81| 0.75| 0.38*| 0.01| 0.44|
| mm    | Raw  | 0.63| 1.51 | 2.63| 0.77| 0.48| 0.06| 0.50|
|       | All  | -0.05| 1.19| 2.41| 0.79| 0.48| 0.15| 0.54|
|       | Mon  | -0.01| 1.15| 2.33| 0.81| 0.51| 0.18| 0.57|

Note. For each model the uncorrected forecasts (Raw), forecasts using quantile-to-quantile (Q2Q) bias correction on all model data at once (All), forecasts using Q2Q bias correction conditioned on the month (Mon) are evaluated. All forecast metrics are based on the ensemble means of +24 h forecasts. The sample size is 1,729 daily values for all models, the average observed rainfall 1.96 mm/day. The sequential color scale and indicates the score per metric for bias corrected models, with the best scores in green, mediocre scores in yellow and the worst in red. The asterisks indicate that:

* difference between score of monthly bias correction versus no bias correction is NOT significant ($p > 0.05$).

** difference between score of monthly bias corrected multi-model versus other bias corrected models is NOT significant ($p > 0.05$).
that temporal phase errors are relatively less important for longer accumulation periods. Suppose that a tropical wave and related precipitation is crossing a catchment on day 1, day 2–5 being dry. Now suppose a model perfectly forecasts the amount of rainfall, but is one day too late with the arrival of the tropical wave. The model produces rainfall only on day 2, when it is observed to be dry. This will result in zero skill for day 1 and day 2, but a perfect skill for a 5-day forecast (when we suppose the model had perfectly forecast no rain for day 3–5). This temporal phase error is directly related to a spatial phase error. Suppose we have five catchments in the east-west direction. Suppose a hypothetical disturbance arrives from the east and it starts to rain on day 1 in the most easterly catchments. On day 2 the system has moved to the adjacent catchment, and so on until it rains in the most westerly catchment on day 5. The model is again consistently 1 day late, but is a perfect forecast otherwise. For five consecutive days the skill of 24-h rainfall accumulations is zero in all catchments. However, for the 5-day accumulation the skill is perfect (except for the last catchment, as rain is forecast on day 6, outside of the 5-day accumulation period).
In general, then, there is a tendency for skill to be higher for longer temporal averaging periods since a longer averaging window smooths over errors in the phase of tropical wave propagation, and cancels out errors in stochastic variability at hard to predict temporal and spatial scales. In contrast, skill generally decreases with lead time as forecast model errors continue to accumulate. These two effects work in opposition when comparing 0–24 h and 0–5 day forecasts (or a 120–144 h and 5–10 day forecast). Which of the two is higher in skill will depend on many factors, including the characteristic timescales of rainfall producing mechanisms, the inherent difficulty of predicting rainfall in the region (the so-called “predictability limit”), the quality and resolution of the model initialization, as well as other factors. Our research suggests that model

Figure 9. Bar charts that show the number of catchments for which each model’s +24 h ensemble mean forecast has the best score, per verification metric. A model is depicted as “the best” when it has the lowest bias (in absolute sense), the lowest MAE, the lowest RMSE or the highest $R^2$, $d$, or AC.

Figure 10. Rank histograms for +24 h ensemble forecasts before (red) and after (blue) applying Q2Q bias correction. A rank histogram is created by sorting ensemble members from low to high precipitation amounts and then binning observations according to their magnitude relative to the individual ensemble members. The blue, horizontal line represents the shape a perfectly reliable ensemble prediction system would have.
errors (due to phasing, or other factors) strongly penalize the skill of forecasts with shorter accumulation periods, so that the advantage of being near the model initialization is overshadowed by the advantage of a longer averaging period.

For the SS\textsubscript{RMSE}, this skill score is only higher for 5-day accumulations after Q2Q bias correction. This is likely because biases accumulate with increasing integration periods, worsening the SS\textsubscript{RMSE}.

The multi-model ensemble is the most skillful, though the difference with ECMWF is small and diminishes with lead time. We can also see that skill based on anomaly correlation is not completely analogous to skill based on the SS\textsubscript{RMSE}. For example, CMC, together with NCEP, performs worst considering anomaly correlation (after Q2Q bias correction). However, considering the SS\textsubscript{RMSE}, CMC is the third best. This apparent inconsistency can be explained as follows: Murphy and Epstein (1988) showed that the skill score of the mean squared error can be decomposed the following way:

\[ \text{SS}_{\text{clim}} \approx r_{\tilde{P}'\tilde{O}'}^2 - \left( r_{\tilde{P}'\tilde{O}'} - \frac{s_{\tilde{P}'} }{s_{\tilde{O}'}} \right)^2 - \left( \frac{\bar{P} - \bar{O}}{s_{\tilde{O}'} } \right)^2 \]

The three terms on the right side are the anomaly correlation, conditional, and unconditional biases, respectively. When the bias terms are negligible, the skill score and the square (!) of the anomaly correlation are

Figure 11. The seasonal variability of AC (upper) and skill score of the root mean squared error SS\textsubscript{RMSE} (lower) before (left) and after (right) Q2Q bias correction. The metrics are calculated using the ensemble means of +24 h forecasts and subsequently averaged over all Ethiopian catchments for the whole temporal range of the data set. In addition to the skill metrics the figures also show the monthly average of relative frequency of rainy days. The whiskers represent spatial variability and extend to the 10th and 90th percentile. The lines are a spatial average of the skill metrics over all catchments while the triangles identify, per month, the highest and lowest values associated with all of the catchments.
equivalent. Thus, the square of the anomaly correlation can be seen as a measure of potential skill, whereas the skill score shows the actual skill (Murphy & Epstein, 1988; Wilks, 2011). This means that even though the anomaly correlations of CMC and NCEP are similar, NCEP has more conditional bias (as the unconditional biases are practically zero for both models, see Table 4).

Applying Q2Q bias correction significantly increases skill for all lead times considered. Before Q2Q bias correction the two best models, the multi-model ensemble and ECMWF, were somewhat skillful for lead times up to three days. After correction, these models show skill up to a lead time of at least 9 days. These two EPSs perform best throughout the forecast lead times considered here.

4.4. Skill Versus Rainfall Amount

Rainfall data often follow a Gamma distribution (Wilks, 2011): low rainfall intensities are common while heavy rains are more rare (Figure 6). Figure 13 shows conditional quantile (CQ: Murphy et al., 1989) plots for bias corrected forecasts. Given a certain forecast, different possible outcomes (observations) can occur. The CQ plot shows the distribution (through selected quantiles) of observed rainfall for the forecast of individual ensemble members (Wilks, 2011), along with a forecast of climatology (average rainfall for each day of the year; i.e., it is the forecast if you would use the average rainfall on a specific date as a prediction). The 1:1 line represents a perfect forecast. At the low end of the rainfall distribution, ensemble members are unbiased, with observed rainfall not being over- or under-forecast. However, this is not the case for the high
end of the rainfall distribution. Individual (bias corrected) ensemble members that forecast a large daily rainfall usually are still too wet, with observed rainfall being much lower in most cases. This latter behavior is typically expected for uncertain forecasts of extreme events. However, keep in mind these conditional quantile plots show the joint distribution of individual ensemble members and observations, not the ensemble mean (as in, e.g., Figures 5 and 9).

Related to this, a second aspect that can be inferred from the CQ plot is the resolution. The resolution refers to the ability of models to distinguish among different observed events for different forecasts, or put another way, the ability of the forecast model to differentiate between the conditional averages of the observations for different values of the forecast. Suppose that there is hardly any difference in the observed rainfall for forecasts of rainfall of 10 and 20 mm/day; the forecasts would have low resolution. The resolution can be inferred from the angle of the quantile lines of the observations. Generally, the quantile lines become more horizontal with forecasts of increasing rainfall (i.e. the same distribution of the observations occurs for forecasts of different values), indicating a decrease of resolution. No correction or calibration method can correct for a lack of resolution. Comparing between models, CMC shows the lowest resolution. In the end, the CQ plots suggest that individual ensemble members, especially when forecasting heavy rainfall, do not provide enough resolution for certain kinds of decision-making, and motivate the need to utilize the whole ensemble set, which we will focus on in the remainder of this study.

Heavy rainfall events are most relevant in the context of flood forecasting. We chose to compare the skill of forecasts for rainfall exceeding the observed (catchment-specific) 90th percentile. Figure 14 shows the reliability diagrams of each EPS model, as well as the frequencies of forecast probabilities exceeding the 90th percentile. Exceedance probabilities are calculated from the raw and bias corrected ensemble distributions, neither of which are explicitly calibrated to account for overconfident or over-dispersive ensembles. Models are reliable when, on average, the events being forecast occur with a relative frequency that is close to the
The reliability diagrams clearly show that all models tend to overestimate the probability of heavy rainfall events. This wet bias is mostly fixed by the Q2Q bias correction. Another aspect that can be inferred from the reliability diagram is the model's resolution. Models have a good resolution when different forecast probabilities are reflected by different observed frequencies. UKMET has the poorest resolution, with the weakest relation between forecast probability and observed relative frequency. Interestingly, CMC performs well looking at resolution of forecast probabilities while its individual ensemble members showed the least resolution (Figure 13). This is likely due to the large (not shown), more accurate ensemble spread that characterizes CMC (see Figure 10). This large spread might degrade the resolution of individual ensemble members, while increasing resolution for forecast probabilities. If the model forecasts an extreme event with a high probability, which does not occur often for the CMC model (as seen by the right hand side of the sharpness diagram inset), this is more often reflected by more accurate frequencies of occurrence. The multi-model, with per definition the largest ensemble spread, also has a high-resolution. For all models, resolution is increased after bias correction. Finally, the inset boxes show the model's sharpness, which is a property of the forecast alone and does not depend on the observations. Models are sharp when forecast probabilities are often different from the climatological average frequency. UKMET and ECMWF have a similar sharpness, however ECMWF is much more reliable, with higher probabilities also reflected in a higher relative observed frequency. Bias correction using Q2Q mapping decreases sharpness but increases reliability and resolution. The improved reliability and resolution are reflected by an increased Brier skill score.

The Brier skill score shows the accuracy of probability forecasts for binary events. In this case, the event of interest is exceedance of the 90th percentile of the observed precipitation, and the reference forecast is the daily average climatology. Figure 15 shows the spatial distribution of the BSS. The lowest skill scores are found in the mountainous regions in the west. A similar spatial pattern in skill was found for probabilistic
skill metrics by Vogel et al. (2020) Q2Q bias correction increases the BSS for all models and for all catchments. The multi-model ensemble has the largest BSS. CMC and ECMWF have similar performance to each other, and UKMET and NCEP are the least skillful.

We have seen that the multi-model ensemble has the highest Brier skill score for extreme rainfall. Also for less extreme rainfall the multi-model ensemble usually outperforms the other models (Figure 16). Generally, the BSS increases with increasing rainfall percentile, and most models are skillful for rainfall predictions exceeding the 40th percentile. Because these rainfalls are per definition increasingly rare, using climatology as a predictor will give a lower and lower BS_{clim}. The skill scores of all models increase after Q2Q bias correction. Following the multi-model ensemble, CMC is usually the best when looking at the median BSS, closely followed by ECMWF. NCEP and UKMET are generally the least skillful.
4.5. Threshold-Based Decision-Making

For the purpose of flood forecasting, thresholds need to be set for deciding when to let the “alarm bells” go off and take preventive action. Because we are dealing with ensemble forecasts we could for example start to issue flood warnings when the chance of exceeding a threshold rainfall (e.g., the 90th percentile) exceeds a certain threshold probability. In choosing this probability threshold, there is always a trade-off between the probability of detection, which can be increased by decreasing the threshold probability, and the false alarm rate which inevitably increases along with it. The ROC diagram (Mason, 1982), shown in Figure 17, shows this trade-off. A perfect forecast has a POD of one and an F of zero, and would display a graph along the left and upper boundaries of the ROC diagram. For a forecast using the climatological probability of the event (no skill), the POD and H are always equal, no matter the threshold probability. Skillful forecasts are located to the upper left of the 1:1 line, which represents no skill. As is to be expected for extreme events, that per definition deviate strongly from climatology, all models are skillful. Again, the multi-model ensemble is the most skillful.

Decision-makers should carefully consider the trade-off between the probability of detection and the false alarm rate in choosing a threshold probability for taking action (e.g., issuing a warning). Figure 18a shows how choosing different threshold probabilities influences the POD, the success ratio (SR, which is 1-FAR), the bias (straight lines) and the TS (curved lines). The best possible forecast has a POD, SR, B, and TS of one. Increasing the threshold probability increases the SR, but also decreases the POD and vice versa. A common solution in finding the “best” threshold probability is to find the one that maximizes the TS (Wilks, 2011). The TS is the ratio of correct “yes” forecasts to the total amount of times the event was forecast and/or observed. The TS is similar to the H but does not reward correct “no” forecasts, which would boost the score when applied to rare events. In the case of our example and data set, the maximum TS is obtained using a threshold probability of 0.3, or 30%. So, if at least 30% of the ensemble members exceed the 90th rainfall percentile, a warning is issued (a “yes” forecast). If the probability is lower than that, no warning is issued (a “no” forecast). Choosing to maximize the TS does

Figure 16. Brier skill scores for precipitation exceeding a range of percentiles (x-axis) for all models. The box plots show the spatial variability of the BSS for calibrated forecasts, while the colored diamonds show the non-corrected spatial median of the BSS. +24 h ensemble forecasts were used to create these plots.

Figure 17. ROC curves showing the trade-off between false alarm rate (F) and probability of detection (POD) for precipitation exceeding the 90th percentile. The curve is created using a series of incremental probability thresholds, shown for the multi-model ensemble. The 1:1 line represents no skill. +24 h bias corrected ensemble forecasts were used to create these plots.
come at a price. On average, the $B$ is around 1.2. This means extreme rainfall is forecast around 20% more often than it is observed. For a threshold probability of 30%, this bias is relatively constant with increasing lead time.

Figure 18b shows in more detail how the TS, the POD, the FAR and $B$ develop with increasing lead time, using a threshold probability of 30%. Obviously, the scores are best for shorter lead times. The multi-model ensemble starts off with a TS of almost 40%. This is the highest score among the models, and this highest ranking persists throughout the lead times considered. At a 1-day lead time the TS of the multi-model ensemble is as good as ECMWF’s 0-day lead time forecast. For the other models the difference is even larger: the TS at 4-day lead of the multi-model is higher than the 1-day lead forecast of NCEP, UKMET, and CMC. Also
considering the POD, the multi-model ensemble consistently outperforms all other models. The probability of correct “yes” forecasts, given the rainfall occurred, exceeds 60% for lead times up to 1 day when using the multi-model ensemble. ECMWF is a close second and is followed by UKMET, NCEP, and CMC in order of descending TS and POD. Because in this example we already put out a warning when the probability of exceeding the 90th percentile rainfall amount is 30%, there will be days with false alarms, decreasing the SR. Only for the multi-model ensemble, for 0-day lead time forecasts, the SR is higher than 50%. Again, ECMWF is a close second, followed by NCEP and UKMET, and finally CMC with the lowest SR. Using a threshold probability of 30%, all models overestimate the frequency of rainfall exceeding the 90th percentile (bias).

5. Conclusion and Recommendations

This study explores one method to optimize the skill of large-scale precipitation forecasts from operational NWP systems. We applied Q2Q bias correction to the ensemble forecasts of four EPSs and a multi-model ensemble that combines those individual EPSs. We validated these models using satellite observations of precipitation in 12 catchments in Ethiopia. We used a variety of validation metrics to compare models and to explore the variability of model skill.

We found that applying a simple and straightforward bias correction technique, Q2Q mapping, significantly increases model skill. Especially in mountainous areas, biases are large and should be corrected. Biases change over the seasons and vary spatially. Therefore, the best model improvement is achieved by applying Q2Q bias correction per catchment and per month.

When looking at a spatial average of all Ethiopian catchments, the ensemble mean of the multi-model ensemble performs best when assessed over a broad range of error and correlation metrics. A close second is the ensemble mean of ECMWF. The bias corrected multi-model ensemble provides the best performing predictions for a large majority of catchments. For some individual catchments other models score better, judging based on the error and correlation metrics. Forecasters might therefore select the best model for each catchment individually. Not only is the multi-model ensemble the best for a majority of catchments, it is also best for most months of the year, especially during wet seasons. With increasing lead time, the multi-model ensemble remains the best performer. However, the additional skill over ECMWF diminishes with lead time as other models lose skill more rapidly than ECMWF. The ranking of CMC, UKMET, and NCEP is both spatially and temporally variable and depends on the verification metric used.

Increasing the temporal scale of a forecast increases its skill, albeit at the cost of the forecast's granularity. The anomaly correlation and (after Q2Q bias correction) the skill score of the root mean squared error are larger for 5-day rainfall averages than for 24-h averages. A possible explanation for this result is that the relative importance of temporal and related spatial phase errors decrease when increasing the accumulation period. For future research it might be worth investigating the effect phase error has on skill.

If the goal is to forecast extreme rainfall, CMC as well as the multi-model ensemble are the two most reliable. However, both models underestimate the frequency of high-intensity rainfall. The Brier skill score shows the accuracy of probabilistic forecasts and is highest for the multi-model ensemble. When making decisions related to flood forecasting, probability thresholds could be used to decide whether or not to act on the risk. In choosing the threshold probability, there is always a trade-off between the probability of detection and the false alarm rate and ratio. This trade-off, visualized by the ROC curve and the performance diagram, is most optimal using the multi-model ensemble. When using a threshold of 30%, the multi-model ensemble outperforms all others based on the threat score, the probability of detection and the false alarm ratio. ECMWF is a close second while CMC performs worst.

All results shown in this study were based on validating using satellite observations of rainfall. These observations themselves may very well be biased. Measures that are independent of bias, such as the anomaly correlation, can be seen as the more fair metric to compare the performances of models. Even considering just the anomaly correlation, the multi-model ensemble and ECMWF would still come out as “the best.” It would be worthwhile to further investigate the biases of different rainfall products, to make a quantitative estimate of the uncertainties in Ethiopia.
It is worthwhile to invest in improvements of NWP models so as to prevent precipitation biases from occurring in the first place. In the meantime, statistical post-processing techniques, such as bias correction or calibration, are necessary for providing skillful and reliable forecasts. Therefore, it is also worthwhile to invest in further improvements of these techniques. Perhaps applying an extreme value theory approach would result in better predictions of extreme rainfall.

A very straightforward and simple approach for creating a multi-model ensemble, just using all available ensemble members, significantly improves model skill and reliability. Perhaps, a more sophisticated selection of ensemble members would improve its skill relative to the best individual model even more. For example, one could apply methods to select only those members that perform best in certain regions.

**Appendix: Elaboration of Bias Correction and Validation Metrics**

In this section we expand on the Q2Q mapping bias correction technique and further explain the validation metrics that were introduced in the Section Methods.

### Quantile-to-Quantile Mapping Bias Correction

Expanding somewhat more formally on the explanation of this technique in Section 3.2, the process can be represented as follows. The model $C_{i,f}^m(\cdot)$ and observational $C_{i,f}^o(\cdot)$ climatological CDFs are given by

\[
C_{i,f}^m(y) = P[Y \leq y]
\]  

(A1)

and

\[
C_{i,f}^o(x) = P[X \leq x]
\]  

(A2)

where $i$ represents each catchment, $f$ each forecast lead time, $P[\alpha]$ denotes the probability of the event $\alpha$, and $x(y)$ a specific value of the generic random variable $X(Y)$ representing the observed (forecast) weather variable of interest. Using these CDFs, a direct percentile-by-percentile bias correction is made to each catchment $i$, lead time $f$, and precipitation ensemble member $fcst$ of the forecast value $y$ by simply setting Equation 1 equal to Equation A1, and determining the value of $x$ that satisfies the equation. In this manner, the “observation-space” percentile is matched with the “forecast model-space” percentile as shown in Figure 3, and $x$ is used in replace of $y$ in the hydrometeorological application. Note that this technique ensures that the forecasts produce the same climatological number of “no rain” events as observed. It also preserves the temporal (ranked) covariances of the weather variable forecast fields (as generated by the numerical weather prediction model).

For a specific example, a predicted rainfall amount of, say, 25 mm corresponds to the 75th percentile of model climatology. However, the 75th percentile of observed rainfall is, say, lower at 20 mm. Bias correction is then applied by adjusting the original rainfall prediction of 25–20 mm. This technique is visually explained in Figure 3. For ensemble forecasts every ensemble member is in this way adjusted.

### Difference Measures

In the following, $O$ refers to the observed variate, $P$ refers to the model-predicted variate, and $N$ refers to the sample size. Except for the Brier skill score and the metrics for binary forecasts, the model-predicted rainfall $P$ is the ensemble mean of each individual EPS.

The statistical measures to show the difference, or error, between the model and observations are the mean absolute error (MAE), root mean squared error (RMSE), the bias error, and the index of agreement $d$, as recommended by Willmott (1981). These statistical values are calculated using Equations A3–A6.

The MAE describes the average absolute difference between an observed and model-predicted variate and is calculated using:
The RMSE is related to the MAE, but gives more weight to larger errors due to its squaring of errors. Both measures indicate overall model performance. The RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{N^{-1} \sum_{i=1}^{N} (P_i - O_i)^2}$$

(A4)

Overall bias (also known as mean error) measures the average bias across all forecasts. The bias gives a good indication of the systemic over- or under-prediction of forecast precipitation. This bias is unconditional, as it does not depend on the forecast value. The bias is calculated using:

$$\text{bias} = N^{-1} \sum_{i=1}^{N} (P_i - O_i)$$

(A5)

The index of agreement ($d$, Willmott, 1981) can be applied to compare between models and is both a relative and bounded measure (as opposed to the RMSE). The $d$ is calculated as follows:

$$d = 1 - \left[ \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (P_i^* - O_i^*)^2} \right] \cdot 0 \leq d \leq 1$$

(A6)

wherein $P_i^* = P_i - \bar{O}$ and $O_i^* = O_i - \bar{O}$.

**Correlation Metrics**

The coefficient of determination, $R^2$, shows how much of the variation of the predictand, $P$, is accounted for by the regression, in this case a least squares linear regression $\hat{P} = a + b \cdot O$. $\hat{P}$ is the predicted value of $P$. $R^2$ is then calculated using:

$$R^2 = \frac{\text{SSR}}{\text{SSR} + \text{SSE}} \cdot 0 \leq R^2 \leq 1$$

(A7)

with SSR being the regression sum of squares:

$$\text{SSR} = \sum_{i=1}^{N} (\hat{P}(O_i) - \bar{\hat{P}})^2$$

(A8)

SSE is the sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{N} (P_i - \hat{P}(O_i))^2$$

(A9)

The disadvantage of the $R^2$ is that you can obtain a high score by just getting the seasonality right. This is not the case with the anomaly correlation. An anomaly is the difference between the (observed or forecast) variate, in this case precipitation, and the climatological mean of this variate. Simply said, it shows whether a day is wetter or dryer than usual. The anomaly correlation is a measure that shows the correlation of the anomalies of the forecast precipitation with the actual, observed precipitation anomaly. If the patterns of forecast and observed precipitation are more alike, the AC will be higher, regardless of any systemic bias that may exist. The (uncentered) AC is calculated using Equation A10:

$$AC = \frac{\sum_{i=1}^{N} (P_i - C_i)(O_i - C_i)}{\sqrt{\sum_{i=1}^{N} (P_i - C_i)^2 \sum_{i=1}^{N} (O_i - C_i)^2}} = -1 \leq AC \leq 1$$

(A10)
with $C_i$ being the climatological precipitation. Because the observational data set only consists of 7 years of data, just using the average rainfall for each day will create a rather erratic and even somewhat skillful climatology. To mimic a climatology of a longer time-span, we smoothed the 7-year climatology by computing a running mean over the 7-year climatology with a 20-day averaging window.

**Skill Scores**

Accuracy measures by themselves may be ambiguous. Therefore, we will use several skill scores. Skill scores relate the accuracy of a prediction to a reference forecast using a particular accuracy measure. In this case, we use the climatology ($C_i$) as a reference forecast. Positive skill scores mean that the forecast is more accurate than when simply using the climatological mean value and vice versa. We use two skill scores: the skill score of the RMSE ($\text{SS}_{\text{RMSE}}$) and the Brier skill score ($\text{BSS}$). These are calculated using Equations A11–A13.

The $\text{SS}_{\text{RMSE}}$ is calculated using:

$$\text{SS}_{\text{RMSE}} = 1 - \frac{\text{RMSE}}{\text{RMSE}_{\text{clim}}}, -\infty < \text{SS}_{\text{RMSE}} \leq 1$$  \hspace{1cm} (A11)

with RMSE being the model RMSE and $\text{RMSE}_{\text{clim}}$ being the RMSE using the climatological mean. The BSS is calculated the following:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{clim}}}, -\infty < \text{BSS} \leq 1$$  \hspace{1cm} (A12)

with BS being the model Brier score and $\text{BS}_{\text{clim}}$ being the Brier score using the climatological mean. The Brier score measures the accuracy of forecast probabilities of dichotomous events and is calculated using the following equation:

$$\text{BS} = N^{-1} \sum_{k=1}^{N} (P_k - O_k)^2, 0 \leq \text{BS} \leq 1$$  \hspace{1cm} (A13)

In this case $P_k$ and $O_k$ denote the paired forecast and observed probability of the dichotomous event. Because probabilities of exceeding a certain amount of rainfall vary over the year, using just one value for the $O_k$ could create false skill (Hamill & Juras, 2006). Especially in Ethiopia, where the rainfall pattern is strongly seasonal, positive skill could be achieved by only predicting the overall variations across seasons. Therefore, we calculate monthly values of $O_k$ for computing the BSS. For example, the climatological probability of any dichotomous event, $O_k$, is the same for all days of all Januaries, simply the average relative frequency across all 7 years.

**Accuracy Measures for Binary Forecasts**

Probabilistic forecasts may have to be converted to binary forecasts for the purpose of decision-making. For example, if the probability of exceeding a certain rainfall amount is larger than a certain threshold probability, a flood warning may be issued. In essence, probabilistic forecasts are then converted to a “yes” or

| Table A1 | Contingency Table for Binary Forecasts |
|----------|---------------------------------------|
|          |             | Observed |
|          | Yes | No |
| Forecast | Yes | a | b |
|          | No  | c | d |

Because probabilities of exceeding a certain amount of rainfall vary over the year, using just one value for the $O_k$ could create false skill (Hamill & Juras, 2006). Especially in Ethiopia, where the rainfall pattern is strongly seasonal, positive skill could be achieved by only predicting the overall variations across seasons. Therefore, we calculate monthly values of $O_k$ for computing the BSS. For example, the climatological probability of any dichotomous event, $O_k$, is the same for all days of all Januaries, simply the average relative frequency across all 7 years.
“no” forecast. A contingency table is created for a set of forecasts by applying the same threshold across a set of forecast/observation pairs and categorizing each pair as in Table A1. The accuracy of binary forecasts can then be calculated using a variety of scores: the hit rate \( H \), threat score \( TS \), probability of detection \( POD \), false alarm ratio \( FAR \) and the bias \( B \), Equations A14–A19.

The hit rate is simply the fraction of correct “yes” and correct “no” forecasts. \( H \) is calculated using:

\[
H = \frac{a + d}{a + b + c + d}
\]  

(A14)

For events that seldom occur the hit rate tends to be high. This is caused by the large amount of correct “no” forecasts. The threat score does not take into account correct “no” forecasts, and is therefore a better accuracy measure for rare events. The TS is calculated as follows:

\[
TS = \frac{a}{a + b + c}
\]  

(A15)

The probability of detection is the fraction of correct “yes” forecasts given the event occurred, and is calculated using:

\[
POD = \frac{a}{a + c}
\]  

(A16)

The false alarm ratio is the fraction of incorrect “yes” forecasts over the total number of “yes” forecasts and is calculated using:

\[
FAR = \frac{b}{a + b}
\]  

(A17)

The false alarm rate, not to be confused with the false alarm ratio, is the fraction of incorrect “yes” forecasts over the total number of nonevents and is calculated using:

\[
F = \frac{b}{b + d}
\]  

(A18)

The bias is the ratio of the number of “yes” forecasts to the number of “yes” observations and is calculated as follows:

\[
B = \frac{a + b}{a + c}
\]  

(A19)

**Data Availability Statement**

This work is based on TIGGE data. TIGGE (The Interactive Grand Global Ensemble) is an initiative of the World Weather Research Programme (WWRP). As such, the authors would like to thank ECMWF for continuing to provide open access to this archive, and we are also indebted to Steve Worley and Doug Schuster for earlier access to the NCAR TIGGE archive. Besides the ECMWF TIGGE portal for raw forecasts, processed forecasts used in this study can be found here: http://ral.ucardu/~hopson/ESS/.

**References**

Aonashi, K., Awaka, J., Hirose, M., Kozu, T., Kubota, T., Liu, G., et al. (2009). GSMaP passive microwave precipitation retrieval algorithm: Algorithm description and validation. *Journal of the Meteorological Society of Japan*, 87A, 119–136. [https://doi.org/10.2151/jmsj.87a.119](https://doi.org/10.2151/jmsj.87a.119)

Awalchew, S., Yilma, A., Loulseged, M., Loisandl, W., Ayana, M., & Alamirew, T. (2007). *Water resources and irrigation development in Ethiopia*. International Water Management Institute. (Working Paper 123).
Bechtold, P. (2019). Challenges in tropical numerical weather prediction at ECMWF. In D. A. Randall, J. Srinivasan, R. S. Nanjundiah, P. Mukhopadhyay (Eds.), Current trends in the representation of physical processes in weather and climate models (pp. 29–50). Singapore: Springer Singapore.

Bougeault, P., Toth, Z., Bichop, C., Brown, B., Burridge, D., Chen, D. H., et al. (2010). The THORPEX interactive grand global ensemble. Bulletin of the American Meteorological Society, 91(8), 1059–1072. https://doi.org/10.1175/2010bams2853.1

Conway, D., Dalin, C., Landman, W., & Osborn, T. (2017). Hydropower plans in eastern and southern Africa increase risk of concurrent climate-related electricity supply disruption. Nature Energy, 2(12), 946–953. https://doi.org/10.1038/s41560-017-0037-4

Diro, G. T., Black, E., & Grimes, D. I. F. (2008). Seasonal forecasting of Ethiopian spring rains. Meteorological Applications, 15(1), 73–83. https://doi.org/10.1002/met.63

Elison, E. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fazzini, M., Bisci, C., & Billi, P. (2015). The climate of Ethiopia. In P. Billi (Ed.), Landscapes and landforms of Ethiopia. Dordrecht: Springer. chapter 3.

Gissila, T., Black, E., Grimes, D. I. F., & Slingo, J. M. (2004). Seasonal forecasting of the Ethiopian summer rains. International Journal of Climatology, 24(11), 1345–1358. https://doi.org/10.1002/joc.1078

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A, 57(3), 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. Monthly Weather Review, 140(7), 2232–2252. https://doi.org/10.1175/mwr-d-11-00201.0

Hamill, T. M., & Juras, J. (2006). Measuring forecast skill: Is it real skill or is it the varying climatology? Quarterly Journal of the Royal Meteorological Society, 132(621C), 2905–2923. https://doi.org/10.1256/qj.06.25

Hirpa, F. A., Gehremichael, M., & Hopson, T. (2010). Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. Journal of Applied Meteorology and Climatology, 49(5), 1044–1051. https://doi.org/10.1175/2009jamc2298.1

Hopson, T. (2005). Operational flood forecasting for Bangladesh (PhD thesis). University of Colorado.

Hopson, T. M., & Webster, P. J. (2010). A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. Journal of Hydrometeorology, 11(3), 618–641. https://doi.org/10.1175/2010jhm1068.1

Huffman, G. J., Bolvin, D. T., Nelkin, E. I., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. Journal of Hydrometeorology, 8(1), 38–55. https://doi.org/10.1175/jhm560.1

Jolliffe, I. T., & Stephenson, D. B. (2012). Forecast verification: A practitioner’s guide in atmospheric science. John Wiley & Sons.

Jury, M. R. (2010). Representation of Ethiopian wet spells in global and nested models. Advances in Meteorology, 2014(4), 1–12. https://doi.org/10.1155/2014/237374

Katz, R. W., Parlane, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. Advances in Water Resources, 25(8–12), 1287-1304. https://doi.org/10.1016/s0309-1708(02)00056-9

Kipkogei, O., Bhardwaj, A., Kumar, Y., Ogalo, L. A., Opijah, F. J., Mutemi, J. N., & Krishnamurti, T. N. (2016). Improving multimodel medium range forecasts over the Greater Horn of Africa using the FSU superensemble. Meteorology and Atmospheric Physics, 128(4), 441–451. https://doi.org/10.1007/s00703-015-0430-0

Knevel, J. C., Ahijevych, D. A., & Manning, K. W. (2004). Using temporal modes of rainfall to evaluate the performance of a numerical weather prediction model. Monthly Weather Review, 132(12), 2995–3009. https://doi.org/10.1175/mwr2882.1

Laurance, W., Sayer, J., & Cassman, K. (2013). Agricultural expansion and its impacts on tropical nature. Trends in Ecology & Evolution, 29(2), 107–116. https://doi.org/10.1016/j.tree.2013.12.001

Léle, I. M., & Lamb, P. J. (2010). Variability of the intertropical front (ITF) and rainfall over the West African Sudan-Sahel zone. Journal of Climate, 23(14), 3984–4004. https://doi.org/10.1175/2010jcli3277.1

Mason, I. (1982). A model for assessment of weather forecasts. Australian Meteorological Magazine, 30(4), 291–303.

Mekonnen, A., & Rosows, W. B. (2011). The interaction between deep convection and easterly waves over tropical North Africa: A weather state perspective. Journal of Climate, 24(16), 4276–4294. https://doi.org/10.1175/2011jcli3900.1

Mekonnen, A., & Thornicroft, C. D. (2016). On mechanisms that determine synoptic time scale convection over East Africa. International Journal of Climatology, 36(12), 4045–4057. https://doi.org/10.1002/joc.4614

Mengistul, A. A., Nagel, P., Getahun, A., Saber, S. A., & Loader, S. P. (2013). Updated review of amphibian diversity, distribution and conservation in Ethiopia. Ethiopian Journal of Biological Sciences, 12(1), 81–116. Retrieved From https://www.ajol.info/index.php/ejbs/article/view/98861

Mladen, R. (2019). TIGGE archive model description. ECMWF. https://confluence.ecmwf.int/display/TIGGE/Models

Murphy, A. H., Brown, B. G., & Chen, T. Y. (1989). Diagnostic verification of temperature forecasts. Weather and Forecasting, 4(4), 485–501. https://doi.org/10.1175/1520-0434(1989)004<0485:dvotf>2.0.co;2

Murphy, A. H., & Epstein, E. S. (1988). Skill scores and correlation coefficients in model verification. Monthly Weather Review, 117(3), 572–582.

Mutemi, J. N., Ogalo, L. A., Krishnamurti, T. N., Mishra, A. K., & Vijaya, T. S. (2007). Multimodel based superensemble forecasts for short and medium range NWP over various regions of Africa. Meteorology and Atmospheric Physics, 95(1–2), 87–113. https://doi.org/10.1007/s00703-006-0187-6

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. Reviews of Geophysics, 55(3), 590–635. https://doi.org/10.1002/2016rg000544

Park, Y.-Y., Buzzi, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. Quarterly Journal of the Royal Meteorological Society, 134(637), 2029–2050. https://doi.org/10.1002/qj.334

Pohl, B., & Camberlin, P. (2006a). Influence of the Madden–Julian Oscillation on East African rainfall. I: Intraseasonal variability and regional dependency. Quarterly Journal of the Royal Meteorological Society, 132(621), 2521–2539. https://doi.org/10.1256/qj.05.104

Pohl, B., & Camberlin, P. (2006b). Influence of the Madden–Julian Oscillation on East African rainfall. II: March–May season extremes and intraseasonal variability. Quarterly Journal of the Royal Meteorological Society, 132(621), 2541–2558. https://doi.org/10.1256/qj.05.223

Riddle, E. E., & Cook, K. H. (2008). Abrupt rainfall transitions over the Greater Horn of Africa: Observations and regional model simulations. Journal of Geophysical Research, 113(D15), 55. https://doi.org/10.1029/2007jd009202
Riddle, E. E., & Wilks, D. S. (2013). Statistical indices of the northward rainfall progression over eastern Africa. *International Journal of Climatology, 33*(2), 356–375. https://doi.org/10.1002/joc.3429

Schlueter, A., Fink, A. H., Knippertz, P., & Vogel, P. (2019). A systematic comparison of tropical waves over northern Africa. Part I: Influence on rainfall. *Journal of Climate, 32*(5), 1501–1523. https://doi.org/10.1175/jcli-d-18-0173.1

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelman, A., Golaz, J.-C., Bodas-Salcedo, A., et al. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research, 115*(D24), 1147. https://doi.org/10.1029/2010jd014532

Swinbank, R., Kyouda, M., Buchanam, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society, 97*(1), 49–67. https://doi.org/10.1175/bams-d-13-00191.1

Trenberth, K. E., Zhang, Y., & Gehne, M. (2017). Intermittency in precipitation: Duration, frequency, intensity, and amounts using hourly data. *Journal of Hydrometeorology, 18*(5), 1393–1412. https://doi.org/10.1175/jhm-d-16-0263.1

Viste, E., & Sorteberg, A. (2013a). The effect of moisture transport variability on Ethiopian summer precipitation. *International Journal of Climatology, 33*(15), 3106–3123. https://doi.org/10.1002/joc.3566

Viste, E., & Sorteberg, A. (2013b). Moisture transport into the Ethiopian highlands. *International Journal of Climatology, 33*(1), 249–263. https://doi.org/10.1002/joc.3409

Vogel, P., Knippertz, P., Fink, A., Schleuter, A., & Gneiting, T. (2020). Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics. *Weather and Forecasting, 35*(6), 2367–2385. https://doi.org/10.1175/waf-d-20-0082.1

Watson, P. A. G., Berner, J., Corti, S., Davini, P., Hardenberg, J., Sanchez, C., et al. (2017). The impact of stochastic physics on tropical rainfall variability in global climate models on daily to weekly time scales. *Journal of Geophysical Research: Atmospheres, 122*(11), 5738–5762. https://doi.org/10.1002/2016JD026386

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic Press.

Willmott, C. J. (1981). On the validation of models. *Physical Geography, 2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Woodhams, B. J., Birch, C. E., Marsham, J. H., Bain, C. L., Roberts, N. M., & Boyd, D. F. A. (2018). What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Monthly Weather Review, 146*(9), 2757–2780. https://doi.org/10.1175/mwr-d-17-0396.1

Yang, G.-Y., & Slingo, J. (2001). The diurnal cycle in the tropics. *Monthly Weather Review, 129*(4), 784–801. https://doi.org/10.1175/1520-0493(2001)129<0784:tdecitt>2.0.co;2

Zagar, N. (2017). A global perspective of the limits of prediction skill of NWP models. *Tellus A: Dynamic Meteorology and Oceanography, 69*(1), 1317573. https://doi.org/10.1080/16000870.2017.1317573