Evaluation of the skill of monthly precipitation forecasts from global prediction systems over the Greater Horn of Africa

Weather and Forecasting

Evaluation of the Skill of Monthly Precipitation Forecasts from Global Prediction Systems over the Greater Horn of Africa

--Manuscript Draft--

Manuscript Number:	WAF-D-20-0177
Full Title:	Evaluation of the Skill of Monthly Precipitation Forecasts from Global Prediction Systems over the Greater Horn of Africa
Article Type:	Article
Corresponding Author:	Hussen Seid Endris, Ph.D IGAD Climate Prediction and Applications Centre (ICPAC) Nairobi, KENYA
Corresponding Author's Institution:	IGAD Climate Prediction and Applications Centre (ICPAC)
First Author:	Hussen Seid Endris, PhD
Order of Authors:	Hussen Seid Endris, PhD
	Linda Hirons, PhD
	Zewdu Tessema Segele, PhD
	Masilin Gudoshava, PhD
	Steve Woolnough, PhD
	Guleid A. Artan, PhD
Abstract:	The skill of precipitation forecasts from global prediction systems has a strong regional and seasonal dependence. Quantifying the skill of models for different regions and timescales is important, not only to improve forecast skill, but to enhance the effective uptake of forecast information. The sub-seasonal to seasonal prediction (S2S) database contains near real-time forecasts and re-forecasts from 11 operational centres and provides a great opportunity to evaluate and compare the skill of operational S2S systems. This study evaluates the skill of these state-of-the-art global prediction systems in predicting monthly precipitation over the Greater Horn of Africa. This comprehensive evaluation was performed using deterministic and probabilistic forecast verification metrics. Results from the analysis showed that the prediction skill varies with months and region. Generally, the models show high prediction skill during the start of the rainfall season in March and lower prediction skill during the peak of the rainfall in April. ECCC, ECMWF, KMA, NCEP and UKMO show better prediction skill over the region for most of the months compared with the rest of the models. Conversely, BoM, CMA, HMCR and ISAC show poor prediction skill over the region. Overall, the ECMWF model performs best over the region among the 11 models analyzed. Importantly, this study serves as a baseline skill assessment with the findings helping to inform how a subset of models could be selected to construct an objectively consolidated multi-model ensemble of S2S forecast products for the Greater Horn of Africa region, as recommended by the World Meteorological Organization.

Evaluation of the Skill of Monthly Precipitation Forecasts from Global Prediction Systems over the Greater Horn of Africa

Hussen Seid Endris[1]*, Linda Hirons[2], Zewdu Tessema Segele[1], Masilin Gudoshava[1, 3], Steve Woolnough[2], Guleid A. Artan[1]

[1] IGAD Climate Prediction and Applications Centre (ICPAC), Nairobi, Kenya

[2]National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom

[3]National University of Science and Technology, Bulawayo, Zimbabwe

*Corresponding author address: Hussen Seid Endris

IGAD Climate Prediction and Applications Centre (ICPAC), Nairobi, Kenya.

E-mail: hussen.seid1@gmail.com

Abstract

The skill of precipitation forecasts from global prediction systems has a strong regional and seasonal dependence. Quantifying the skill of models for different regions and timescales is important, not only to improve forecast skill, but to enhance the effective uptake of forecast information. The sub-seasonal to seasonal prediction (S2S) database contains near real-time forecasts and re-forecasts from 11 operational centres and provides a great opportunity to evaluate and compare the skill of operational S2S systems. This study evaluates the skill of these state-of-the-art global prediction systems in predicting monthly precipitation over the Greater Horn of Africa. This comprehensive evaluation was performed using deterministic and probabilistic forecast verification metrics. Results from the analysis showed that the prediction skill varies with months and region. Generally, the models show high prediction skill during the start of the rainfall season in March and lower prediction skill during the peak of the rainfall in April. ECCC, ECMWF, KMA, NCEP and UKMO show better prediction skill over the region for most of the months compared with the rest of the models. Conversely, BoM, CMA, HMCR and ISAC show poor prediction skill over the region. Overall, the ECMWF model performs best over the region among the 11 models analyzed. Importantly, this study serves as a baseline skill assessment with the findings helping to inform how a subset of models could be selected to construct an objectively consolidated multi-model ensemble of S2S forecast products for the Greater Horn of Africa region, as recommended by the World Meteorological Organization.

Key words: Forecast, Re-forecast, S2S, Skill, Precipitation, Models, GHA

1. Introduction

Sub-seasonal predictions, from 2 weeks to a season, are relevant for informing decision making and early warning across a range of sectors in the Greater Horn of Africa (e.g., agriculture, energy, water and disaster risk management). Sub-seasonal forecasts bridge the gap between medium-range weather and seasonal forecasts (Vitart et al. 2012; Robertson et al. 2015; Vitart et al. 2017; White et al. 2017), and have the potential to contribute to early warning and early action for both flooding and drought disasters (Moron et al. 2018).

Given the potential applications of sub-seasonal predictions (White et al. 2017), and the increasing demand for forecast information across sectors in recent years, the World Weather Research Programme (WWRP) and World Climate Research Programme (WCRP) launched a joint research initiative called the sub-seasonal to seasonal (S2S) prediction project and a multi-model database of S2S forecasts and re-forecasts. The database provides an opportunity to make comparisons between the outputs of different prediction models and advance knowledge of S2S prediction **(**Vitart et al. 2017). Since the establishment of the S2S database, some studies have evaluated the skill of S2S models in different regions. Li and Robertson (2015) assessed the weekly prediction skill of three global prediction systems over the globe and indicated the models had very good skill for the first week. Over Africa, de Andrade et al. (2021) evaluated the sub-seasonal forecasts for three global prediction systems and found that although skill was relatively low in week 3 and week 4, compared to weeks 1 and 2, the probabilistic forecasts still had skill in weeks 3-4. de Andrade et al. (2019) compared the performance of sub-seasonal precipitation re-forecasts from 11 S2S models considering lead times up to 4 weeks using deterministic verification metrics and indicated higher skill during the first week and reduced skill as lead time increased. Vigaud et al. (2017) also examined the sub-seasonal rainfall forecast skill over summer monsoon regions of the Northern Hemisphere using sub-monthly lead times and found good skill (reliability) in multi-model forecasts for forecasts beyond 1 week.

Because of different drivers of S2S variability, and the non-linear response to these drivers, the skill at predicting the precipitation varies widely from region to region and timescale to timescale. Evaluating the forecast skill for different regions and timescales is vitally important to identify model errors, improve forecast skill and also promote the uptake and use of forecast information in decision making. In this study, we thoroughly assessed the skill of 11 S2S models over the Greater Horn of Africa (GHA) during the March-April-May (MAM) rainfall season with a focus on monthly timescales.

Past studies have shown that the MAM rainfall commonly known as the long-rains over the GHA is weakly associated with large-scale oceanic and atmospheric features (e.g., Hastenrath et al. 1993; Rowell et al. 1994; Vellinga and Milton 2018) and has low predictability compared to the October-November-December (OND) rainfall known as the short-rains (Camberlin and Philippon 2002). Furthermore, it has been noted that there is an intraseasonal inhomogeneity within the long-rains season. The spatial rainfall anomaly patterns are similar in March and April but quite different in May (Camberlin and Philippon 2002). Other studies (e.g., Rowell et al. 1995; Nicholson and Kim 1997) also found that time series of interannual variability for the months of March, April, and May are different. Nicholson (2015) also indicated that the prevailing atmospheric circulation and controls on interannual variability are clearly different during the three months of the long-rains. As a result of this inhomogeneity within the season, some authors (e.g., Camberlin et al. 2009; Moron et al. 2013; Rowell et al. 1994) have suggested that sub-seasonal analysis is required for the long-rains season to advance the understanding and prediction of precipitation variability.

It is also important to recall that the World Meteorological Organization (WMO) Executive Council at its 69th Session in May 2017 recommended the operational Regional Climate Centres (RCCs) and National Meteorological and Hydrological Services (NMHSs) to access digital forecast and reforecast data from the WMO Lead Centres for long-range forecasts and produce an objectively consolidated sub-seasonal and seasonal forecast product that is traceable and reproducible. In the recommendations, the need to assess the skill of forecasting models for different regions was stressed as well as the selection of a subset of models which have better skill for the region of interest for the construction of the relevant multi-model ensemble. Therefore, the results from this study address these recommendations and provide a crucial baseline for identifying skillful models over GHA on S2S timescale.

2. Data and methods

2.1 Data

2.1.1 Observed data used for verification

The observed data used to verify rainfall re-forecasts is the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) version v2.0 (Funk et al. 2015). This dataset is a blended product of 0.05° resolution satellite imagery and in-situ station data provided by the Climate Hazards Group. CHIRPS dataset is available from 1981 to near-present. Validations of CHIRPS rainfall data has been conducted over the different parts of East Africa by comparing CHIRPS with rain-gauge data and other satellite rainfall products such as African Rainfall Climatology version 2 (ARC2) and the Tropical Applications of Meteorology using Satellite and ground-based observations (TAMSAT) (e.g., Maidment et al. 2017, Dinku et al. 2018). It has been found that CHIRPS performed significantly better than ARC2 and TAMSAT with higher skill, low bias and lower random errors particularly at dekadal (10-days) and monthly time-scales (Dinku et al. 2018) and indicated its suitability for use as a reference rainfall dataset.

The European Centre for Medium-Range Weather Forecasts (ECMWF) fifth generation reanalysis (ERA5, Hersbach (2020)) datasets was used to evaluate the mean circulation features. This global dataset is available from 1979 to near present with a 0.25 resolution. In this study, monthly 850 hPa zonal and meridional winds are utilized for the analysis period. The observed Sea Surface Temperature (SST) data utilized in this study is version 2 of the National Oceanic and Atmospheric Administration (NOAA) Optimum Interpolation SST (NOAA_ OI_SST_V2) analysis, retrieved from https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-optimal-interpolation-oi-sst-analysis-version-2-oisstv2-1x1. The NOAA_OI_SST_V2 integrates both in situ and satellite data and is available from 1982 to present at 1.0^{o} spatial resolution.

2.1.2 Model Data

The S2S database consists of re-forecasts and near real-time forecasts (3 weeks behind) from 11 global prediction centres, which have been made available for scientific research via the data archive portal at the ECMWF and the China Meteorological Administration (CMA)

(Vitart et al. 2016/7). The 11 global prediction centers are Australian Bureau of Meteorology (**BoM**), China Meteorological Administration (**CMA**), Météo-France/Centre National de Recherche Meteorologiques (**CNRM**), Environment and Climate Change Canada (**ECCC**), **ECMWF**, Hydrometeorological Centre of Russia (**HMCR**), the Institute of Atmospheric Sciences and Climate (**ISAC**), Japan Meteorological Agency (**JMA**), Korea Meteorological Administration (**KMA**), National Centers for Environmental Prediction (**NCEP**) and the United Kingdom's Met Office (**UKMO**). Not all 11 models are exactly independent from each other. The UKMO and KMA use the same system and have the same configuration, but different atmospheric initial conditions and ensemble size.

The re-forecasts and forecasts are archived on a common 1.5-degree grid horizontal resolution in the S2S database. The re-forecasts, also known as hindcasts, are a set of forecasts with start and prediction dates in the past, and are used to assess the skill of the model in reproducing the past forecasts and to calibrate real-time forecasts. Re-forecasts are similar in every aspect with the real-time forecasts apart from differences in ensemble size. This study assesses the skill of 11 global prediction systems in predicting the monthly rainfall over GHA.

As the S2S models are developed and run by different prediction centres, they have different configurations. For instance, some models have **fixed** re-forecast configuration, whereas others have **on-the-fly** configuration. Fixed re-forecasts are produced once during the lifetime of a given version of the model (e.g., BoM, CMA, Meteo-France and NCEP). On the other hand, on-the-fly re-forecasts are produced at the same time as the real-time forecasts (e.g., ECMWF, KMA and UKMO). The frequency and initial start date of the re-forecast also varies from model to model. Some models are run in continuous mode on a daily basis (e.g., CMA, NCEP), whereas others run on weekly or sub-weekly basis (e.g., BoM, ECMWF). In addition to that, the re-forecast length and time range varies from model to model. For example, the NCEP has 12 years re-forecasts initialized every day from 1999 to 2010, whereas ECMWF produces re-forecasts on-the-fly covering the past 20 years, initialized 2 days per week (Monday and Thursday) for each model version. The re-forecast ensemble size also varies from model to model. Some models are atmosphere-only models forced by observed SSTs, while others have the atmospheric component coupled to an ocean model and a sea ice model. The general features of the global prediction systems used for this study are summarized in Table 1.

Even if the S2S prediction systems have different configuration or set-up, there are some common features between them to make the model inter-comparisons possible (de Andrade et al. 2019). For instance, all of the prediction systems have re-forecasts covering the period 1999-2010. Each model also has a control re-forecast member using a single unperturbed initial condition and perturbed forecast members produced for sampling uncertainty in the initial conditions. Further, most of the prediction systems produce forecasts and re-forecasts starting on the 1st and middle of each month. Therefore, it is possible to make the model comparisons using the common period 1999-2010.

In this analysis, all re-forecasts (control and perturbed) from one week lead to zero lead have been used. For example, to assess the skill of the models during April, all re-forecasts initialized from 23rd to 31st of March have been analyzed. The rationale for choosing this is: (1) to include the models that have shorter forecast range in the model comparison analysis;

and (2) to get a sufficiently large number of ensemble members for the probabilistic verification as some models, especially the models run on a daily basis, have few ensemble members if we only consider one or two initialization dates. To enable the comparison between all models, the analysis is performed over a common period from 1999 to 2010 (for 12 years). For computational purposes, both CHIRPS and model re-forecasts have been re-gridded to half degree (0.5°) using bilinear interpolation prior to the skill analysis. We have chosen the 0.5 degree as this is the spatial resolution currently used operationally at IGAD Climate Prediction and Applications Centre (ICPAC) the RCC over the GHA, when producing the monthly and seasonal downscaled climate outlooks for the region.

2.2 Verification Methods

It is important to note that forecast quality is multifaceted and there is no single verification metric that captures all aspects of forecast quality (Murphy 1993). It is therefore important to assess the forecast skills using a range of different statistical measures. Currently, there are several methods available to evaluate the skill of weather and climate forecasts - ranging from simple traditional statistics and scores to methods for more detailed and advanced verifications. In the present analysis, the skills of the models have been assessed using three deterministic and three probabilistic forecast verification measures. The deterministic forecast measures include mean error, linear correlation and root mean square error. The probabilistic forecast evaluation metrics include the Ranked Probability Skill Score, Relative Operating Characteristic and Spread-Error Ratio. The deterministic forecast verification assessment is performed between the ensemble mean of all re-forecast members (control plus perturbed members) and the verifying observation, whereas the probabilistic forecast verification analysis is performed using all the individual ensemble members. In addition to the above verification metrics, Taylor and reliability (attribute) diagrams, which provide summary statistical information between the model and reference field are used.

2.2.1 Deterministic Verification Metrics

In this section we summarize the deterministic verifications methods utilized. The mathematical equations for the deterministic metrics are presented in the supplementary materials.

2.2.1.1 Mean Error

The mean error represents the average difference between forecast and verification values. The mean error is primarily a measure of the systematic part of the forecast error. It is important to note that the mean error does not measure the magnitude of the errors. It also does not measure the correspondence between forecast and observation as it is possible to get a perfect score for a bad forecast if there are compensating errors (Kendzierski et al. 2018).

2.2.1.2 Root Mean Square Error (RMSE)

The RMSE represents the square root of the average of the squared differences between forecasts and verification data. It is a measure of the random component of the forecast error and often used for representing the accuracy of forecasts. The RMSE is sensitive to large errors and provides information on the average magnitude of the forecast errors. However, the RMSE does not indicate the direction of the deviations. The RMSE puts greater influence on large errors than smaller errors (Jorgensen 2016) and thus it might be a good indicator of large errors.

2.2.1.3 Linear Correlation

Correlation is one of the most widely used measures for forecast verification, and provides an assessment of the strength of the linear association between forecasts and the verifying observation. It is a good measure of linear association or phase error. Jolliffe and Stephenson 2012 noted that it is possible for a forecast with large errors to still have a good correlation coefficient with the observation.

2.2.1.4 Taylor diagram

A Taylor diagram (Taylor, 2001) summarizes the statistical relationship between model and the observed/reference field. The diagram is useful for evaluating the accuracy of multiple model outputs against a reference data. Further information on the taylor diagram is provided in the supplementary materials.

2.2.2 Probabilistic Verification Metrics

2.2.2.1 Ranked Probability Skill Score (RPSS)

The ranked probability score (RPS) is a measure of the prediction skill of probabilistic forecasts issued for categorical events (i.e., tercile-based categorical forecasts). The RPS is defined as the sum of the squared differences between cumulative forecast probabilities and cumulative observed probabilities (Murphy 1993). The RPS measures both the reliability and resolution of a forecast and is closely related to the Brier score (Tippett, 2008). The RPS is the same as the Brier score in the case of two category forecasts. The discrete expression of the RPS is given as follows:

$$RPS_t = \sum_{n=1}^{N} (F_n^t - O_n^t)^2 \quad \textbf{(1)}$$

Where

F_n^t is the forecast probability at time t, given by P (forecast$_n$ < thresh$_n$)

O_n^t is the observed probability at time t, given by P (observed$_n$ < thresh$_n$)

n is the probability category

The ranked probability skill score (RPSS) is a skill score based on the RPS values. It is computed as the percentage improvement over reference score:

$$RPSS = \left(1 - \frac{RPS}{RPS_{ref}}\right) x100 = \left(1 - \frac{RPS}{RPS_{clim}}\right) x100 \qquad \textbf{(2)}$$

The RPSS compares the RPS of a forecast to some reference forecast, such as a climatology, and the score ranges between negative infinity and 1. An RPSS below 0 indicates that the forecast is less skillful than climatology, and above zero indicates the forecast is more skillful than climatology where 1 is a 'perfect' forecast. Scores equal to zero are equivalent to forecasts given by the climatology. Müller et al. (2005) and Tippett (2008) noted the dependence of the RPSS on ensemble size. It has been indicated that RPSS is negatively biased for ensemble prediction systems with small ensemble sizes. In this analysis, an ensemble size corrected RPSS called Fair RPSS (Ferro, 2014) is used for evaluating and comparing the skill of operational S2S systems. Further information about Fair RPSS score can be found in Ferro (2014).

2.2.2.2 Relative Operating Characteristic (ROC)

ROC measures the ability of a forecast to discriminate between events and non-events, and measures the degree of forecast discrimination (Mason, 1982). Discrimination is the ability to distinguish one categorical outcome from another. The ROC is not sensitive to bias in the forecast, so it does not say anything about reliability. A biased forecast, however, may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration (Jolliffe and Stephenson 2012). The ROC score, which is computed as the area under the ROC curve, is considered as a useful summary measure of forecast skill. A ROC score of 0.5 indicates unskillful forecasts (i.e., the system is no better than climatology). A ROC score above 0.5 indicates positive discrimination skill and a score of 1.0 represents a perfect forecast. More information on the ROC can be found in Mason (1982), and Jolliffe and Stephenson (2003, 2012).

2.2.2.3 Reliability (or Attribute) Diagram

The reliability (also known as attribute) diagram is a graphical method used to evaluate the reliability of probabilistic forecast systems. The diagram presents the observed frequency against the forecast probability. It basically answers the question of how well the predicted probabilities of an event correspond to their observed frequencies. A forecast system is reliable if and only if all the forecast probabilities are reliable (Toth et al. 2003). A reliability diagram displays a range of forecast probabilities for a given event and their corresponding observed frequencies collected over the re-forecast period (Weisheimer and Palmer 2014). Generally, the reliability is high when correspondence between the forecast probabilities and the observed frequencies is good, and it is low when this correspondence is poor. It is expected that all data points will lie on a straight diagonal line in the reliability diagram when the correspondence between the forecast probabilities and the observational frequencies are perfect. A reliability diagram is a form of attribute diagram when the no-resolution (distance from the horizontal or climatological line) and no-skill with respect to the climatology lines are included in the diagram. In the attribute diagram if the curve lies below the line, it indicates overestimation (i.e., the forecast probabilities are too high). On the other hand, if the curve lies above the line, it indicates underestimation (i.e., forecast probabilities are too low).

2.2.2.4 Spread-Error Ratio (SPR)

The SPR is used to assess the relationship between ensemble spread and the deterministic forecast error. It is defined as the square root of the ratio of mean ensemble variance to the mean squared error of the ensemble mean with the verifying observation. The variance is a measure of the forecast member spread of a particular forecast indicating whether the forecast ensemble spread is large or small, while the RMSE is a measure of the forecast error of the ensemble mean forecast. Thus, the SPR evaluates the ability of the ensemble spread (variance) to depict the forecast error of the data expressed as the RMSE of the ensemble means. When the RMSE and spread are equal, the ensemble successfully predicts the forecast error. When the RMSE is superior to the spread meaning that the SPR is less than 1, it is considered as underdispersive (overconfidence). Conversely, SPR greater than 1 indicates overdispersive (underconfidence). For a reliable forecast system, the ensemble forecasts are expected to have the same size of ensemble spread as their RMSE (Leutbecher and Palmer, 2008; Leutbecher, 2009). The SPR is suitable for verification of ensemble forecasts and sensitive to both forecast resolution and reliability (Christensen et al. 2015).

3. Results and discussion

3.1 Rainfall Climatology

We first analyzed the spatial distribution of rainfall climatology for individual months using CHIRPS data. Figure 1 shows the observed rainfall climatology during March, April and May averaged for the period 1981 to 2010. Climatologically, during the month of March the maximum rainfall is located over southern parts of the region mainly in most parts of Tanzania, Burundi and Rwanda. During April and May, the rainfall band moves from the southern to the northern part of GHA following the position of the Inter-tropical convergence zone (ITCZ). In April, a marked increase in rainfall occurs throughout the region. In May, the maximum rainfall is located over western part of Ethiopia, most parts of South Sudan and Uganda. The following sections presents the monthly rainfall skill of S2S models over GHA for the individual months using the verification metrics described above.

3.2 Deterministic Verification Scores

3.2.1 Mean Error

Figures 2a, b and c show the spatial distribution of mean errors of rainfall between the S2S models and CHIRPS over GHA for March, April and May, respectively. During March, CMA, HMCR, ISAC and JMA overestimated, while BoM underestimated the monthly rainfall over most parts of the region. In particular, the overestimation of total monthly precipitation for HMCR and ISAC systems is quite notable. The rest of the models show a mixed signal with variations existing in terms of the location and magnitude of the mean error. Generally, BoM, CMA, CNRM, HMCR and ISAC show large errors, while ECCC, ECMWF, JMA, KMA, NCEP and UKMO show smaller mean errors over the region during the month of March.

In April, most of the models show larger errors (Fig. 2b) compared to March (Fig. 2a). Consistent with the results for March, the magnitudes of errors are smaller for ECCC, ECMWF, JMA, KMA, NCEP and UKMO models. In contrast, CMA, CNRM, HMCR and ISAC largely overestimate the rainfall especially the overestimation in HMCR and ISAC models over the northern part of the region is notable.

During May, the majority of the models overestimate the rainfall mainly over the northern part of the region (Fig. 2c). In contrast, the BoM underestimates the rainfall in most parts of the region. Moreover, some of the models including CMA, JMA, KMA, NCEP and UKMO show a dry bias over the southern part of the region. It is noted that KMA and UKMO models show similar bias patterns in the region. BoM, CMA, CNRM, HMCR and ISAC still show large errors over the region.

In general, the results from the mean error analysis show that the magnitude of mean errors are low during the month of March compared to April and May for all the prediction models. CMA, CNRM, HMCR and ISAC overestimate the monthly rainfall over most part of the region, whereas BoM systematically underestimate the rainfall throughout most of the region. Overall, ECCC, ECMWF, JMA, KMA, NCEP and UKMO show low bias over the region during March, April and May. The spatial distribution of the mean error of rainfall from KMA and UKMO are almost identical in most parts of the region. This might be due to the fact that the two models have exactly the same configurations. As mentioned earlier, the only difference between the two models is the atmospheric initial condition (Noh et al. 2016). The reason for the month-to-month skill difference will be discussed later.

3.2.2 Root Mean Square Error (RMSE)

The spatial distributions of RMSE from the S2S models with reference to CHIRPS are presented from Fig. 3a to c. It can be seen that RMSE are generally higher in April compared to March and May. BoM, CMA, HMCR and ISAC show large errors over the region in all the months with HMCR and ISAC performing worse (with mean RMSE more than 100 mm), which is consistent with the mean error results. On the other hand, ECCC, ECMWF, KMA, NCEP and UKMO exhibit good prediction skills over the region in terms of RMSE. It can also be seen that KMA and UKMO prediction systems exhibit similar RMSE patterns over the region. Generally, the magnitudes of the mean errors are small during March compared with April and May.

3.2.3 Linear Correlation

Figures 4a to 4c illustrate the spatial distribution of correlation coefficients of rainfall between models and CHIRPS for March, April and May, respectively, for the period from 1999 to 2010 over GHA. Cross-hatches indicate regions where the correlation is statistically significant at the 95% confidence level computed using Student's t test. It can be seen that the skill of the model in producing the rainfall forecast varies from month to month. During March, the majority of the models, with the exception of the HMCR model, show high correlation within the 95% confidence level over the equatorial and southern sector of the region and mainly higher towards the coast. Some of the models show low correlation over the northern part of the GHA, mainly over Sudan, South Sudan and northern and western parts of Ethiopia, but it is important to note that March is not the rainfall season over the northern part of the region (Fig. 1). Overall, ECMWF, JMA, KMA, NCEP and UKMO show relatively high and significant correlation over the equatorial sector compared to the rest of the models. During April, the correlation skills are relatively low over the region compared to March with some models showing a negative correlation in parts of the region. Most notably, CMA model shows negative correlation over the eastern part of the region in April (Fig. 4b). Furthermore, CNRM, HMCR, ISAC, JMA and NCEP also exhibit negative correlation over parts of the equatorial East Africa, mainly over parts of Kenya and Somalia. BoM, ECMWF, JMA, KMA and UKMO show relatively improved skill compared to the other models, mainly over the equatorial and southern part of the region. This may be linked with increased predictability in that region associated with the development of low-level Somali Jet and Asian Summer Monsoon system in May as shown by Nicholson 2015. A discussion about the predictability of Jet and monsoon will be discussed later in section 3.4. During May (Fig. 4c) the models generally show better skill than during the month of April. ECCC, ECMWF, KMA, NCEP and UKMO models show relatively higher skill with significant correlation over the region compared to the other models. It is found that HMCR presents the negative correlations over most parts of the region reflecting the fact that the model fails to reproduce the inter-annual variability.

In addition to evaluating the S2S models at monthly timescales, we also analyzed the skill of the models for weeks 1+2 and weeks 3+4 to investigate if the skill for the monthly forecast is coming from weeks 1+2 only or there is skill in weeks 3+4. In March (SFig. 1) for weeks 1+2 the correlation coefficients are statistically significant at 5% level for most models except HMRC showing that the prediction skill is high. In weeks 3+4 (SFig. 2), the skill is lower in comparison to weeks 1+2. However, ECMWF, KMA, NCEP, and UKMO still have prediction skills with correlations greater than 0.5 over most of the southern and equatorial region. In April (SFig. 3), the weeks 1+2 prediction skill is high for most models except for CMA, CNRM and HMRC which in some areas have weak negative correlations. The majority of the models during April have lower skill in weeks 3+4 with most models showing weak negative and positive correlations (SFig. 4). Only ECCC model shows statistically significant correlations in equatorial parts of the region. Since these statistics are calculated over a 12 year period, a larger sample would provide a greater confidence on the skill for weeks3+4 in April. In May (SFig.5-6), most models show high prediction skill (significant correlations) in weeks 1+2 except for the CMA, ECCC and HMRC models. The weeks 3+4 prediction skills in May are generally higher compared to weeks 3+4 in April. During weeks 3+4 of May, CMA, KMA, NCEP and UKMO show higher prediction skill in comparison with the other models. Thus, in general even though the models have lower prediction skills in weeks 3+4, the models do have skill in weeks 3+4. These results are consistent with Vigaud et al. (2018) who found that during the February to April season the ECMWF model had skill up to weeks 3+4. Thus, issuing out the monthly forecasts is likely to aid in tactical decision making over the various sectors in the region that utilize forecast information from the S2S models.

Overall, the results from the correlation analysis show that the correlation skills are highest during March and poor during April. The high prediction skill during March might be linked with high association of March rainfall with tropical sea surface temperatures (SSTs) compared to April and May as indicated by Camberlin et al. 2009 and Moron et al. 2013. On the other hand, the low prediction skill during April might be related with the wind and pressure pattern changes over the Indian Ocean as there is a directional shift in low level winds from northeast (in March) to southwest (in May).

3.2.4 Taylor diagram

Figure 5 shows a Taylor diagram displaying normalized statistical comparison (i.e., correlation, root-mean-square error and amplitude of variation) of monthly total rainfall of the S2S models with CHIRPS during March, April and May, respectively. The rainfall is spatially-averaged for the GHA domain by masking out the regions outside GHA. In March, most models (including CMA, CNRM, ECCC, ECMW, JMA, KMA and UKMO) show high correlation (> 0.6) in comparison with the observation. In particular, ECMWF, KMA and UKMO present relatively high correlation (> 0.8) and low root-mean-square difference and have a variation close to the reference data. On the other hand, BoM, HMCR and NCEP show low correlation (< 0.6) with HMCR showing the lowest correlation (i.e, 0.1) and a variation far from the reference field. During April, correlations are relatively low in comparison to March. Moreover, most of the models underestimate the magnitude of year-to-year variation relative to CHIRPS, while three models (CMA, JMA, and ISAC) overestimate the variation. BoM, ECCC and ECMWF have relatively high correlation (r> 0.6) compared with other models. ISAC shows a variation much higher than CHIRPS, while CMA exhibits the lowest correlation. During May CNRM, ECCC, ISAC, KMA, NECP and UKMO have relatively high correlation (r > 0.6) compared with other S2S models, while JMA and HMCR presents the lowest correlation. It is also noticed that HMCR and JMA indicate extremely high variation compared to CHIRPS.

3.3 Probabilistic Verification Scores

3.3.1 RPSS

The Fair RPSS from the 11 S2S models for March, April and May are presented in Figures 6a, 6b and 6c, respectively. During March most models show positive RPSS (i.e., a forecast better than the climatological forecast values) over most parts of the region, with maximum score over the equatorial sector (Figure 6a). Consistent with other verification metrics, HMCR shows the lowest skill by presenting negative scores over most parts of the region. In April, the skill for most S2S models is relatively low compared to March. More grid points with negative scores are found than for March. ECCC, ECMWF and KMA show relatively better skill in the region. During May, the skills of the forecasts are generally higher than April, but lower than March. While ECCC, ECMWF, KMA, NCEP and UKMO present the highest skill, CMA, HMCR and ISAC show the lowest skill (Figure 6c). Overall, the results of RPSS indicate that the skill of the S2S model forecasts is lower in April than March and May, agreeing with the previous results of mean error and correlations. The RPSS values obtained in this study are relatively higher than those in Vigaud et al. (2018) for seasonal evaluation, highlighting the importance of the monthly updates during the season. It is also noted that most models predict worse than climatology over the northern parts of the region, mainly over Sudan. But it is important to note that the northern part of the GHA is generally dry during this season (Fig. 1).

3.3.2 ROC

Figures 7a, b and c show ROC Skill Scores (ROCSS) for lower-tercile forecasts for March, April and May, respectively. During March, most of the models show a forecast skill better than the climatological forecast (Figure 7a). In particular, CMA, CNRM, ECCC, ECMWF, ISAC, KMA, NCEP and UKMO show good skill over the region. On the other hand, BoM, HMCR and JMA present a forecast worse than a climatological forecast over parts of the region especially over parts of Kenya, Somalia, Ethiopia, South Sudan, Uganda and Tanzania. In April, most of the S2S models show lower skill than in March. ECMWF, KMA and UKMO perform better than other models, with the ECMWF model showing high ROCSS over the region and outperforming the rest of the models. The rest of the models including BoM, CMA, CNRM, ECCC, HMCR, ISAC, JMA and NCEP exhibit skill scores of less than 0.5 over equatorial parts of the region indicating the forecast from those systems is worse than the climatological forecast over the specified region. During May, ECMWF, KMA, NCEP and UKMO prediction systems show good prediction skill over the region compared to the other prediction systems. In contrast, HMCR performs the worst. In general, April forecasts exhibit lower skill than in both March and May. The ROC skill scores for the upper-tercile forecasts have also been analysed and the results are very similar to lower-tercile forecasts (Fig not shown). ROC skill scores for the lower-tercile in weeks 1+2 and weeks 3+4 for each month was also computed (SFig. 7- 12). The results reveal that nonetheless weeks 1+2 have higher skill than weeks 3+4, the weeks 3+4 still have skill especially in March and May. de Andrade et al. (2021) also evaluated the quality of sub-seasonal precipitation forecasts over Africa using reforecasts from three models (ECMWF, UKMO, and NCEP) and indicated that the probabilistic forecasts showed reasonable skill in weeks 3+4.

3.3.3 Reliability (or attribute) diagrams

Figure 8 shows the attribute diagrams of precipitation for the below-normal category over GHA from the 11 S2S models during March, April and May. During March, it can be seen that the majority of the models lie within the grey area particularly for higher probabilities indicating good reliability in the issued re-forecast probabilities. Only three of the S2S models, namely CMA, HMCR and ISAC, lie below the no skill line for forecast probabilities above 0.4. During April, most prediction systems including BOM, CMA, CNRM, ECCC, HMCR, ISAC and NCEP are away from the perfect reliability diagonal (45°) line particularly for higher forecasted probabilities and indicate the lack of reliability and resolution for the issued hindcast probabilities. The rest of the S2S models show good reliability. In particular, the curve for ECMWF, KMA, NCEP and UKMO are much closer to the perfect reliability line, indicating a much better agreement between the forecast probabilities and observed frequencies. In May, the three S2S models (i.e, BoM, HMCR and ISAC) showed the lowest skill by indicating lower resolution and overconfidence. It is also noted that the majority of the models underestimated the low probabilities (below the climatological line). During the three months, it has been found ECMWF shows better prediction skill than the rest of the S2S models. The results for above normal category (Fig not shown) were found to be consistent with the results of below normal category.

3.3.4 SPR

The SPR from the 11 S2S models for March, April and May are presented in Figures 9a, 9b and 9c, respectively. In general, it can be seen that most of the S2S models indicate underdispersion (overconfidence) over wet areas and overdispersion (underconfidence) over the dry areas in the northern parts of the region mainly over Sudan. A recent study by de Andrade et al. (2021) also noted overconfidence in ECMWF, NCEP and UKMO models in all weeks and suggested the need to apply calibration for more reliable predictions. In March (Fig. 9a), most of the models show good performance particularly over the equatorial and southern parts of the region. In the HMCR model, the spread is much smaller than the error. During April (Fig. 9b), most models have an SPR less than 1 indicating underdispersion (overconfidence). ECCC and ECMWF outperform other models by presenting SPR values close to 1. In May (Fig. 9c), similar to April, the majority of the models present an error larger than the spread reflecting underdispersive characteristics, with the exception of the northern parts of the region. ECMWF and ECCC perform better than the rest of the prediction systems, while HMCR performs the worst in terms of spread-error relationship. de Andrade et al. (2021) indicated enhanced skill in ECMWF and associated the forecast skill with correct representations of climate drivers' teleconnections such as El Niño–Southern Oscillation (ENSO), Indian Ocean Dipole (IOD) and Madden Julian Oscillation (MJO).

3.4 SST and atmospheric features

Further to the evaluation of the skill of S2S models in predicting the monthly rainfall, this study assessed the ability of the models in representing some of the important large-scale features. The goal is to provide insight into the connection between the skill of rainfall forecasts and the representation of key processes that drive monthly rainfall variability in the region.

3.4.1 Indian Ocean SST

The Indian Ocean plays an important role in modulating the climate variability of the GHA. Previous studies (e.g., Camberlin and Phillipon, 2002; Vellinga and Milton 2018; Wainwright et al. 2019) have shown the influence of SST anomalies over the tropical Indian Ocean on the East African long-rains. In this study, we assessed the ability of S2S models to reproduce the teleconnections between SSTs over the Indian Ocean and corresponding rainfall over the GHA. This was done by regressing grid-point rainfall over the GHA to SST indices over the Indian Ocean. The specific regions (boxes) used to compute the indices are shown in Fig. 10a. These regions (boxes) were selected in accordance with previous studies and are based on the correlation analysis between spatially averaged observed monthly rainfall over the GHA and concurrent grid point SST shown in Fig. 10a. During March, the SST gradient between the northern (40°E-75°E, 5°S-10°N) and southern (20°E-60°E, 40°S -20°S) Indian Ocean is used following Wainwright et al. (2019), which linked a reduced March rainfall and delayed onset of the long-rains with warm SSTs south of Madagascar. For the May index, average SSTs in the northern Indian Ocean box (5°S–15°N, 50°–90°E) were used, where the correlations with the rainfall are the strongest and statistically significant.

Figure 10b shows SST-rainfall teleconnection patterns obtained by regressing March rainfall against the meridional SST gradient over the Indian Ocean for observations (top left panel) and individual S2S models (all other panels). The observed patterns indicate that the equatorial parts of the region (5°S–10°N) are positively correlated with the index indicating above normal rainfall when the north-south gradient is strong. On the other hand, the southern and southeastern parts of Tanzania are negatively correlated with the index. In this case, warm SSTs over south western Indian Ocean weaken the meridional SST gradient which creates local convective activity (enhanced moisture convergence), and lead to enhanced rainfall in that part of the region. This is consistent with Wainwright et al. 2019, which suggested warmer SSTs to the south delay the northward progression of the rain-band and lead to increased March rainfall in the southern part, but reduced rainfall over the equatorial and northern part of the GHA. The positive coefficients over the eastern horn of Africa are statistically significant at the 95% confidence level. It can be seen that most S2S models reasonably reproduce the observed features (Fig. 10b). This supports the idea that the relatively strong coupling of SST and rainfall in March is well captured by the S2S models, and that this leads to the high monthly skill found for March.

Rainfall teleconnections for May against the SST index over the northern tropical Indian Ocean are shown in Figure 10c. The observations exhibit significant positive coefficients over most of the equatorial and southern parts of the region, and negative coefficients over western parts of Ethiopia and the South Sudan-Sudan border areas. This implies that warm SST anomalies in the northern Indian Ocean bring enhanced rainfall over most parts of Eastern Africa, but reduced rainfall over parts of western Ethiopia, South Sudan and Sudan. Most models poorly represented both the spatial distribution and amplitude of this teleconnection pattern, particularly the positive associations over southern and eastern parts of the region and the negative association over the summer monsoon areas. It can also be seen that there is a linkage between the forecast skill and the teleconnection patterns. For example, ECCC has quite good skill in May over northern Somalia compared to the other models (Fig 4c & 6c) and also has the best representation of the teleconnection in that region (Fig. 10c). Similarly, ECMWF showed good skill over Western Kenya, and has a good representation to the SST teleconnection in that area.

3.4.2 Somali Low-Level Jet (SLLJ)

The SLLJ, a major component of the Asian summer monsoon system, is one of the most important sources of moisture for East Africa, particularly during the summer season. It plays an important role in transporting moisture from the Indian Ocean to the region. Although the jet is most intense during the boreal summer season, the northward cross-equatorial flow of the jet starts in April and the jet becomes active over the Indian Ocean during May. A study by Nicholson (2015) indicated that the surface features of the SLLJ begin to develop over the Indian Ocean in April, and by May a deep and well-developed monsoon low becomes evident.

The climatological pattern of SLLJ during May from ERA5 and mean errors of the jet from S2S models in comparison to the ERA5 are shown in SFig. 13. ERA5 shows the jet is characterized by southeasterly flow south of the equator, meridional flow around the equator along the East African coast and southwesterly monsoonal flow over the Arabian Sea. Generally, models which are able to capture these large-scale features have higher skill. Consistent with precipitation performance, ECCC, ECMWF, JMA, KMA, NCEP and UKMO show smaller errors than the rest of the models. On the other hand, BoM, CMA, HMCR and ISAC show the largest bias.

To examine the ability of S2S models in representing the spatial patterns and magnitude of rainfall teleconnections with the SLLJ, a regression analysis was applied to a scalar index of the Jet. A scalar index of jet intensity was constructed by computing the square root of twice the spatial mean kinetic energy (KE) of 850 hPa horizontal wind over a spatial domain 5^{o}S – 20^{o}N; 50^{o}E –70^{o}E, as in Boos and Emanuel (2009). Figure 11 shows rainfall teleconnections against SLLJ Index estimated by linear regression during May from observation and the S2S models. The teleconnection patterns from ERA5 (Fig. 11 top left) indicates a positive association between the SLLJ Index and rainfall over the summer rainfall region (northwestern parts of the analyzed domain), indicating wet conditions associated with a strong jet, possibly through increased moisture flux to the region. It can be seen that most S2S models fail to capture the pattern and the amplitude of the positive teleconnection over the northern part of the region. In particular, BoM, CMA, HMCR, and NCEP produced signals with opposite signs to those found in ERA5 over those areas. ECMWF and ECCC generally capture the positive relationship between the SLLJ index and rainfall, although ECMWF tends to overestimate the magnitude and spatial extent of the positive teleconnection patterns.

Most areas of the equatorial and southern part of the region have weak and inverse relationships with the strength of the SLLJ (ERA5). This implies that enhancement of the Jet leads to reduced rainfall over the equatorial and southern part of the region. A study by Nicholson (1996) has also indicated that a strengthening of the SLLJ is associated with enhanced frictionally-induced subsidence on the coast of East Africa. The majority of S2S models fairly capture the negative relationship between the strength of the SLLJ and rainfall over the equatorial and southern parts of the region. Analysis of the rainfall teleconnections against SLLJ Index from observations over a longer period (1981-2018) revealed that regression coefficients are statistically significant at the 95% confidence level over most parts of the region (Fig. not shown). This suggests that a large sample is crucial to have a greater confidence on the skill of the models representing the teleconnection patterns.

Overall, our analyses of the important large-scale features revealed that the ability of the models in reproducing the rainfall is partly linked to their ability in representing the important potential oceanic and atmospheric circulation features. However, it is important to note that many other processes contribute to the regional rainfall variability, and thus more in-depth analysis of other relevant atmospheric and oceanic features (such as the MJO, quasi-biennial oscillation (QBO) and Arabian heat low) is crucial to better understand the mechanisms behind the sources of monthly rainfall predictability and elucidate both strengths and deficiencies in the S2S models. For example, Vitart et al. (2017) showed that the ECMWF and UKMO models consistently have higher bivariate correlation for the MJO than the other models, with MJO correlation remaining above 0.6 at several weeks leadtime. The ability of such models to better capture large-scale drivers like the MJO could explain their consistently higher skill throughout the different months.

4. Summary and conclusions

Due to the increasing demand for the availability of S2S forecast products and information from the user community, it is important to assess and document the prediction skill of operational prediction systems for different regions and timescales. This study evaluates and compares the skill of 11 state-of-the-art operational models from the S2S database in predicting the monthly precipitation over the Greater Horn of Africa during the long-rains. The prediction skill of S2S models has been examined using re-forecast/hindcast data by combining forecasts at lead times from one week to zero over the common period of 1999-

2010. The skill has been quantified using different deterministic and probabilistic forecast verification metrics. The deterministic skill assessment is performed using ensemble mean of all re-forecast members, whereas the probabilistic forecast verification analysis is performed using all the ensemble members. It has been found that the skill of the models in predicting rainfall is dependent on both the month and region. The models generally showed good prediction skill during the early stage of the rainy season in March and poor prediction skill during the peak of the rainfall season in April. In addition to the monthly evaluation, analysis for model skill in weeks 1+2 and weeks 3+4 is also conducted. It is shown that although weeks 1+2 have higher skill than weeks 3+4, the weeks 3+4 still exhibit some skill, especially in March and May. The high prediction skill observed during March is likely linked to strong teleconnections between March rainfall and SST over the Indian Ocean, which is well represented by most S2S models. This is in accordance with Camberlin et al. (2009) and Moron et al. (2013) findings, which indicate the March rainfall anomaly patterns are more spatially coherent compared to April and May, and highly associated with tropical SSTs. The low prediction skill during April might be linked with the directional shift in low level winds as there is a progressive directional shift from northeasterly in March to southeasterly in April, where the southeasterlies become stronger and evident in May as highlighted by Nicholson 2015. In May, a diagnostic of SLLJ suggests that the mean error (phase bias) in the position of the jet is a stronger contributor to the quality of the rainfall forecast than its representation of the large-scale teleconnections.

Among the 11 prediction systems, ECCC, ECMWF, KMA, NCEP and UKMO demonstrate noticeably better skill than the other models. In contrast the BoM, CMA, HMCR and ISAC prediction systems tend to yield poor prediction skills over the region. Overall, ECMWF outperforms the rest of the models, in terms of both deterministic and probabilistic verification metrics. The best and worst performing models identified in this study are in agreement with findings of the recent study by de Andrade et al 2019, which assessed the deterministic forecast quality of weekly accumulated precipitation over the globe. This study provides a crucial baseline skill assessment for selecting those models which perform better, thus informing which could be used to construct a multi-model ensemble for producing consolidated forecasts for the GHA region. It is worth noting that in doing so this study directly addresses the WMO recommendation of the need to critically evaluate the skill of forecasting models for different regions and timescale and for selecting a subset of models for producing operational objective S2S forecasts. It has been revealed that the prediction skill of the models in reproducing the regional rainfall was partly linked with the correct representation of some of the important potential atmospheric and oceanic processes and teleconnections such as the SLLJ and SST anomalies over the tropical Indian Ocean. Further diagnostic analysis of other potential drivers is needed to better understand the sources of sub-seasonal predictability and the linkage between the skill of rainfall forecast and representation of key processes. Moreover, this analysis was performed over a relatively short period (12 years) and thus a large sample size is needed to provide greater confidence on the skill of the S2S models in predicting the rainfall as well as representing the teleconnection patterns.

5. Acknowledgments

This work was supported by UK Research and Innovation as part of the Global Challenges Research Fund, African SWIFT programme, grant number NE/P021077/1. Hussen Seid was also supported by Intra-ACP Climate Services and Related Applications (ClimSA) project funded by the 11th EDF (ACP/FED/038-833). Linda Hirons and Steve Woolnough were also supported by the National Centre for Atmospheric Science ODA national capability programme ACREW (NE/R000034/1), which is supported by NERC and the GCRF. Zewdu Segele was supported by the Weather and Climate Information Services (WISER) Support to ICPAC Project (W2-SIP).

6. Data Availability Statement

All the datasets analyzed in this study (S2S hindcasts, observational and reanalysis datasets) are openly available and can be accessed from the following links: **S2S hindcasts:** http://apps.ecmwf.int/datasets/data/s2s, **CHIRPS:** https://data.chc.ucsb.edu/products/CHIRPS-2.0/, **ERA5:** https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels-monthly-means?tab=form, and **NOAA OISSTv2:** https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-optimal-interpolation-oi-sst-analysis-version-2-oisstv2-1x1.

7. Reference

Boos, W. R., & Emanuel, K. A. (2009). Annual intensification of the Somali jet in a quasi-equilibrium framework: Observational composites. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 135(639), 319-335.

Camberlin, P., & Philippon, N. (2002). The East African March–May rainy season: Associated atmospheric dynamics and predictability over the 1968–97 period. *Journal of Climate*, *15*(9), 1002-1019.

Camberlin, P., V. Moron, R. E. Okoola, N. Philippon, and W. Gitau (2009). Components of rainy seasons' variability in equatorial East Africa: Onset, cessation, rainfall frequency and intensity, Theor. Appl. Climatol., 98(3–4), 237–249, doi:10.1007/s00704-009-0113-1.

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. Quarterly Journal of the Royal Meteorological Society, 141(687), 538-549.

de Andrade, F. M., Young, M. P., MacLeod, D., Hirons, L. C., Woolnough, S. J. and Black, E. (2021) Sub-seasonal precipitation prediction for Africa: forecast evaluation and sources of predictability. Weather and Forecasting. ISSN 0882-8156 doi: https://doi.org/10.1175/WAF-D-20-0054.1

de Andrade, F. M., Coelho, C. A., & Cavalcanti, I. F. (2019). Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Climate Dynamics, 52(9-10), 5451-5475.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., & Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall estimates over eastern Africa. Quarterly Journal of the Royal Meteorological Society, 144, 292-312.

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140(683), 1917-1923.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S....Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data, 2, 150066. https://doi.org/10.1038/sdata.2015.66.

Hastenrath, S., Nicklis, A., & Greischar, L. (1993). Atmospheric-hydrospheric mechanisms of climate anomalies in the western equatorial Indian Ocean. Journal of Geophysical Research: Oceans, 98(C11), 20219-20235.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., & Simmons, A. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, *146*(730), pp.1999-2049.

Jolliffe, I. T., & Stephenson, D. B.(2003). Forecast Verification: A Practitioner’s Guide in Atmospheric Science. John Wiley and Sons, 240 pp.

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

Jorgensen, S. E. (Ed.). (2016). Handbook of ecological models used in ecosystem and environmental management (Vol. 3). CRC press.

Kendzierski, S., Czernecki, B., Kolendowicz, L., & Jaczewski, A. (2018). Air temperature forecasts' accuracy of selected short-term and long-term numerical weather prediction models over Poland. *Geofizika*, *35*(1), 67-85.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of computational physics*, *227*(7), 3515-3539.

Leutbecher, M. (2009). Diagnosis of ensemble forecasting systems. In *Seminar on Diagnosis of Forecasting and Data Assimilation Systems* (pp. 235-266).

Li, S., & Robertson, A. W. (2015). Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Monthly Weather Review*, *143*(7), 2871-2889.

Maidment, R. I., Grimes, D., Black, E., Tarnavsky, E., Young, M., Greatrex, H., ... & Alcántara, E. M. U. (2017). A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa. Scientific data, 4, 170063.

Mason, I. (1982). A model for assessment of weather forecasts. *Aust. Meteor. Mag*, *30*(4), 291-303.

Moron, V., Camberlin, P., & Robertson, A. W. (2013). Extracting subseasonal scenarios: an alternative method to analyze seasonal predictability of regional-scale tropical rainfall. Journal of climate, 26(8), 2580-2600.

Moron, V., Robertson, A. W., & Vitart, F. (2018). Sub-seasonal to Seasonal Predictability and Prediction of Monsoon Climates. *Frontiers in Environmental Science*, *6*, 83.

Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J., & Liniger, M. A. (2005). A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. Journal of Climate, 18(10), 1513-1523.

Murphy, A.H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting, 8, 281-293.

Nicholson, S. E. (2015). The predictability of rainfall over the Greater Horn of Africa. Part II: prediction of monthly rainfall during the long rains. Journal of Hydrometeorology, 16(5), 2001-2012.

Nicholson, S. and Kim, J. (1997). The relationship of the el nino-southern oscillation to African rainfall. International Journal of Climatology, 17(2):117–135.

Nicholson, S. E., (1996). A review of climate dynamics and climate variability in Eastern Africa. *The Limnology, Climatology and Paleoclimatology of the East African Lakes,* T. C. Johnson and E. O. Odada, Eds., Gordon and Breach, 25–56.

Noh, Y. C., Sohn, B. J., Kim, Y., Joo, S., & Bell, W. (2016). Evaluation of temperature and humidity profiles of Unified Model and ECMWF analyses using GRUAN radiosonde observations. Atmosphere, 7(7), 94.

Robertson, A. W., Kumar, A., Peña, M., and Vitart, F. (2015). Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.* 96, ES49–ES53. doi: 10.1175/BAMS-D-14-00139.1

Rowell, D. P., C. K. Folland, K. Maskell, and M. N. Ward (1995). Variability of summer rainfall over tropical north Africa (1906–1992): Observations and modelling, Q. J. R. Meteorol. Soc., 121(523), 669–704, doi:10.1002/qj.49712152311.

Rowell, D. P., J. M. Ininda, and M. N. Ward (1994). The impact of global sea surface temperature patterns on seasonal rainfall in East Africa, in Proc. Int. Conf. On Monsoon Variability and Prediction, Trieste, Italy, pp. 666–672, WMO, Geneva, Switzerland.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192.

Tippett, M. K. (2008). Comments on "The discrete Brier and ranked probability skill scores". *Monthly Weather Review*, *136*(9), 3629-3633.

Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). Probability and ensemble forecasts. Forecast verification: A practitioner's guide in atmospheric science, 137-163.

Vigaud, N., Robertson, A. W., Tippett, M. K., & Acharya, N. (2017). Subseasonal predictability of boreal summer monsoon rainfall from ensemble forecasts. *Frontiers in Environmental Science*, *5*, 67.

Vigaud, N., Tippett, M. K., & Robertson, A. W. (2018). Probabilistic skill of subseasonal precipitation forecasts for the East Africa–West Asia sector during September–May. *Weather and Forecasting*, *33*(6), 1513-1532.

Vellinga, M., & Milton, S. F. (2018). Drivers of interannual variability of the East A frican "Long Rains". Quarterly Journal of the Royal Meteorological Society, 144(712), 861-876.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The Subseasonal to Seasonal (S2S) prediction project database. Bull. Amer. Meteor. Soc. 98, 163–173. doi: 10.1175/BAMS-D-16-0017.1

Vitart, F., Robertson, A. W., & Anderson, D. L. (2012). Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. Bulletin of the World Meteorological Organization, 61(2), 23.

Wainwright, C. M., Marsham, J. H., Keane, R. J., Rowell, D. P., Finney, D. L., Black, E., & Allan, R. P. (2019). 'Eastern African Paradox' rainfall decline due to shorter not less intense Long Rains. *npj Climate and Atmospheric Science*, *2*(1), 1-9.

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. Journal of the Royal Society Interface, 11(96), 20131162.

White, B. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., et al. (2017). Potential applications of subseasonal-to-seasonal (S2S) prediction. *Meteol. App*. 24, 315–325. doi: 10.1002/met.1654

Model	Re-forecast configuration	Time range (days)	Re-forecast length	Re-forecast frequency	Re-forecast size	Ocean Coupling
BoM	Fixed	0-62	1981-2013	6/month	33	Yes
CMA	Fixed	0-60	1994-2014	Daily	4	Yes
CNRM	Fixed	0-61	1993-2014	4/monthly	15	Yes
ECCC	On-the-fly	0-32	1998-2017	Weekly	4	No

ECMWF	On-the-fly	0-46	past 20 years	2/week	11	Yes
HMCR	On-the-fly	0-61	1985-2010	weekly	10	No
ISAC	Fixed	0-32	1981-2010	every 5 days	5	No
JMA	Fixed	0-33	1981-2010	3/month	5	No
KMA	On-the-fly	0-60	1991-2010	4/month	3	Yes
NCEP	Fixed	0-44	1999-2010	daily	4	Yes
UKMO	On-the-fly	0-60	1993-2015	4/month	7	Yes

Table 1: Summary of configuration of the global prediction systems (models) used in this analysis. The re-forecast length, time range, frequency and number of ensemble members depend on the modeling center.

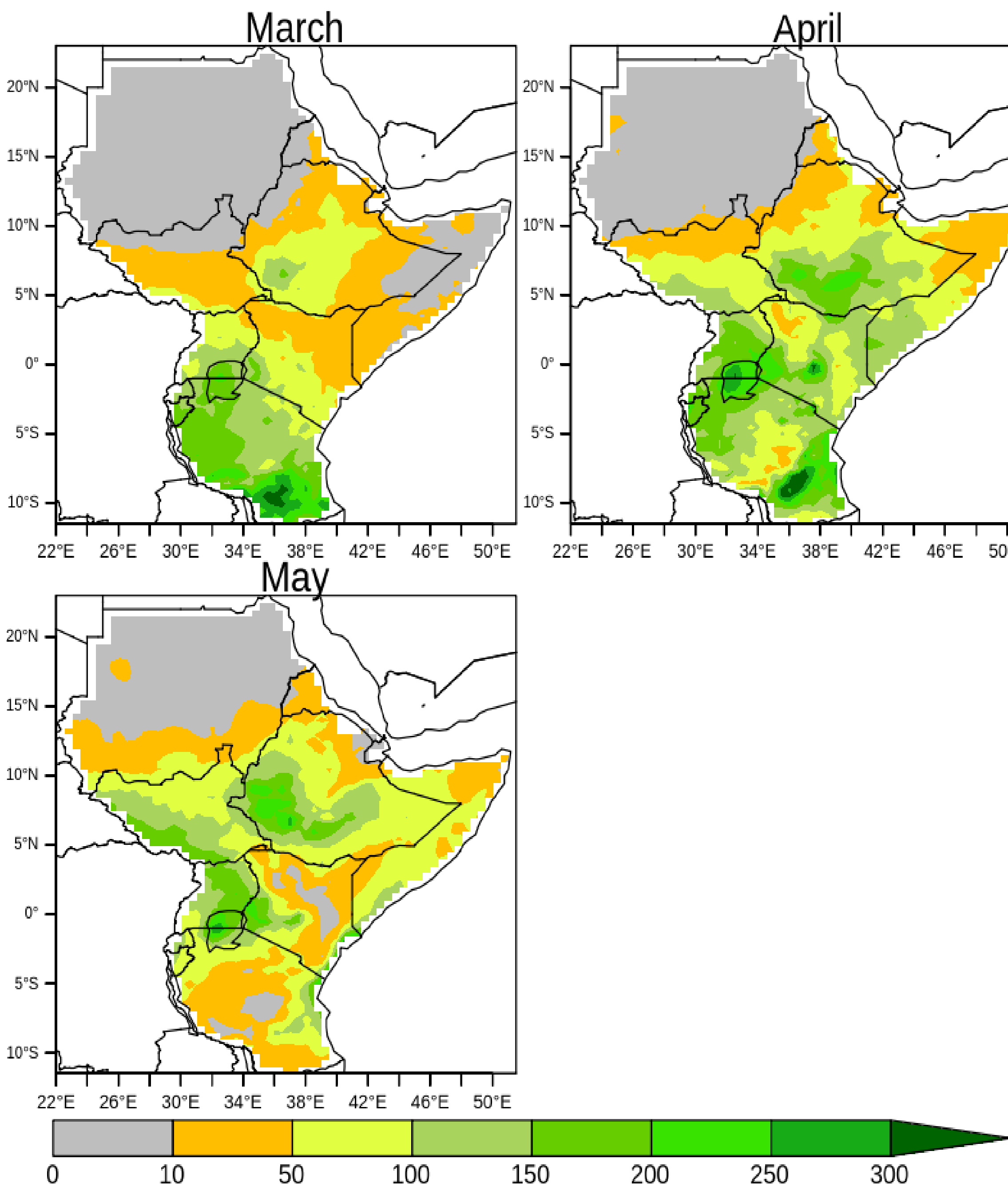

Figure 1. Spatial distribution rainfall climatology during March, April and May over GHA using CHIRPS observed data.

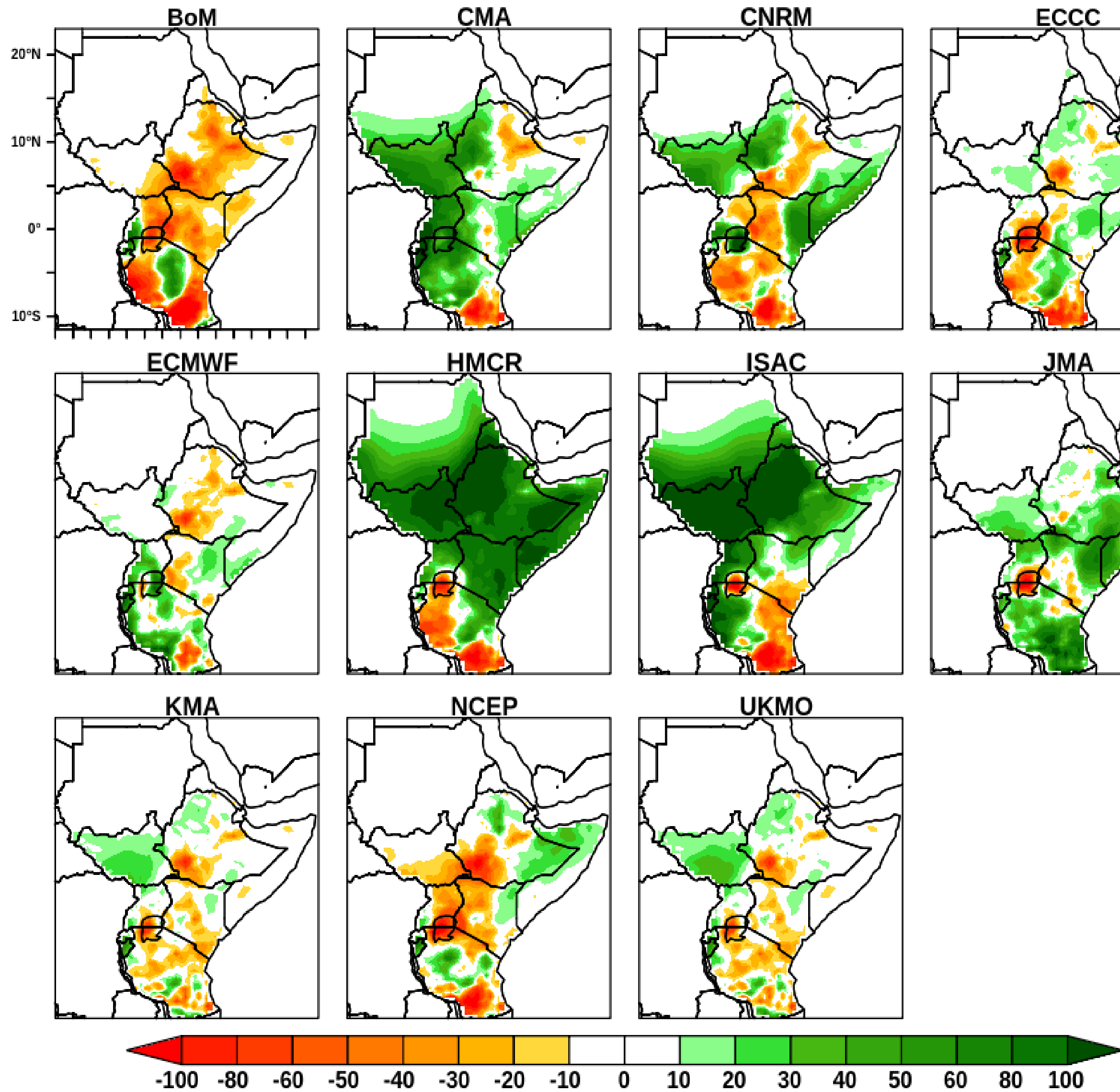

Figure 2a. Spatial distribution of Mean Error of rainfall between models and CHIRPS during March over GHA for the period from 1999 to 2010.

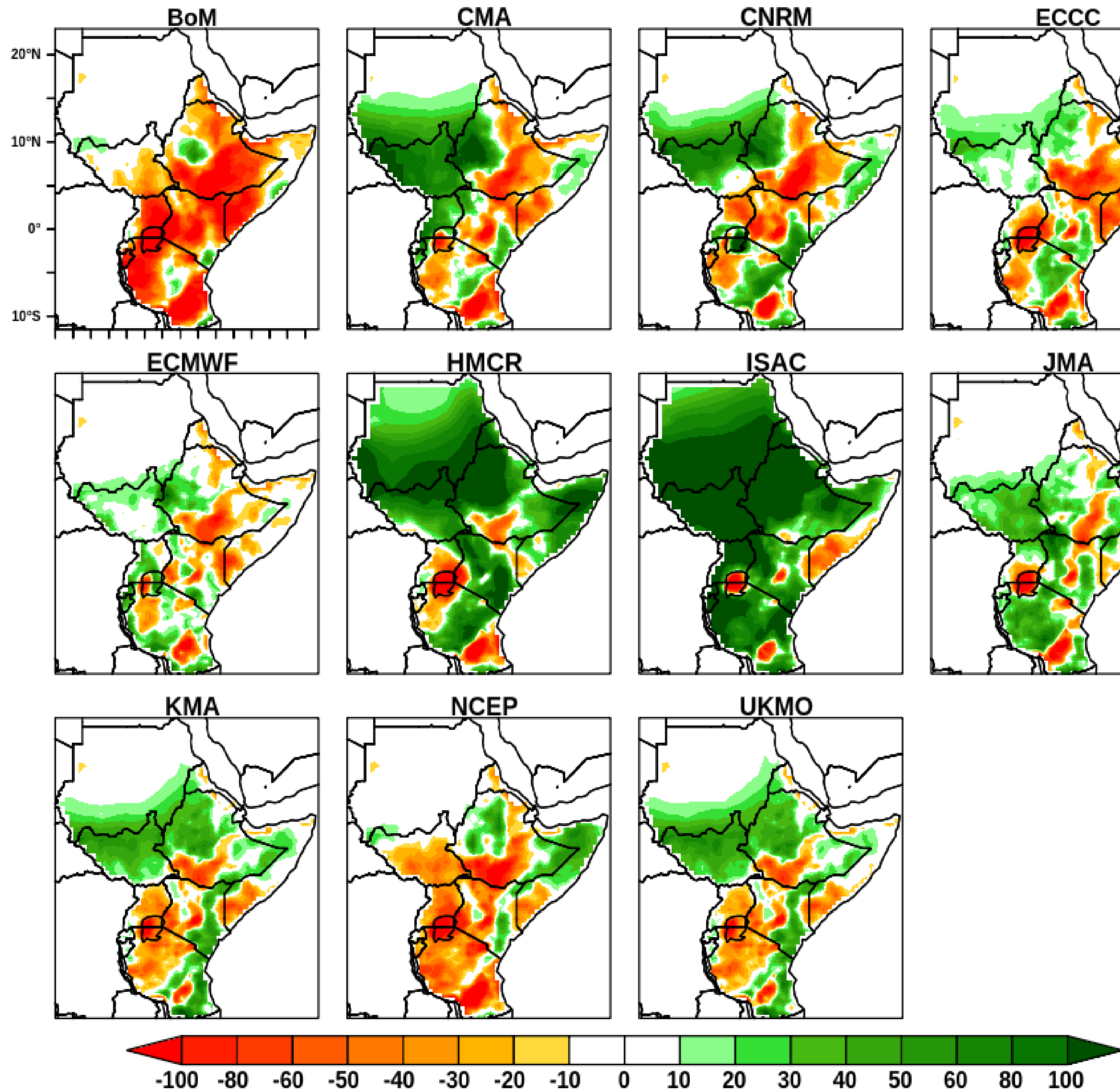

Figure 2b. Spatial distribution of Mean Error of rainfall between models and CHIRPS during April over GHA for the period from 1999 to 2010.

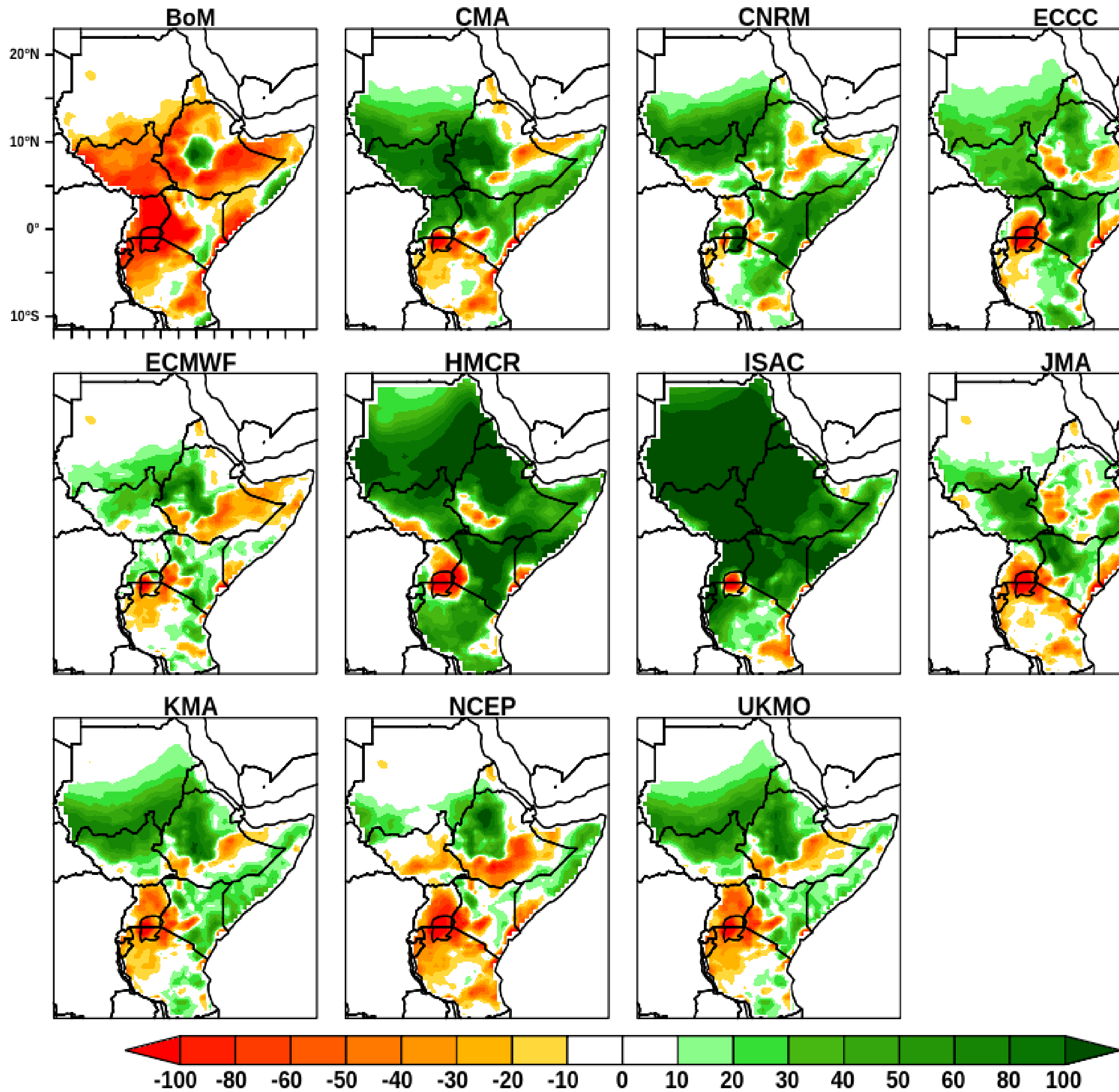

Figure 2c. Spatial distribution of Mean Error of rainfall between models and CHIRPS during May over GHA for the period from 1999 to 2010.

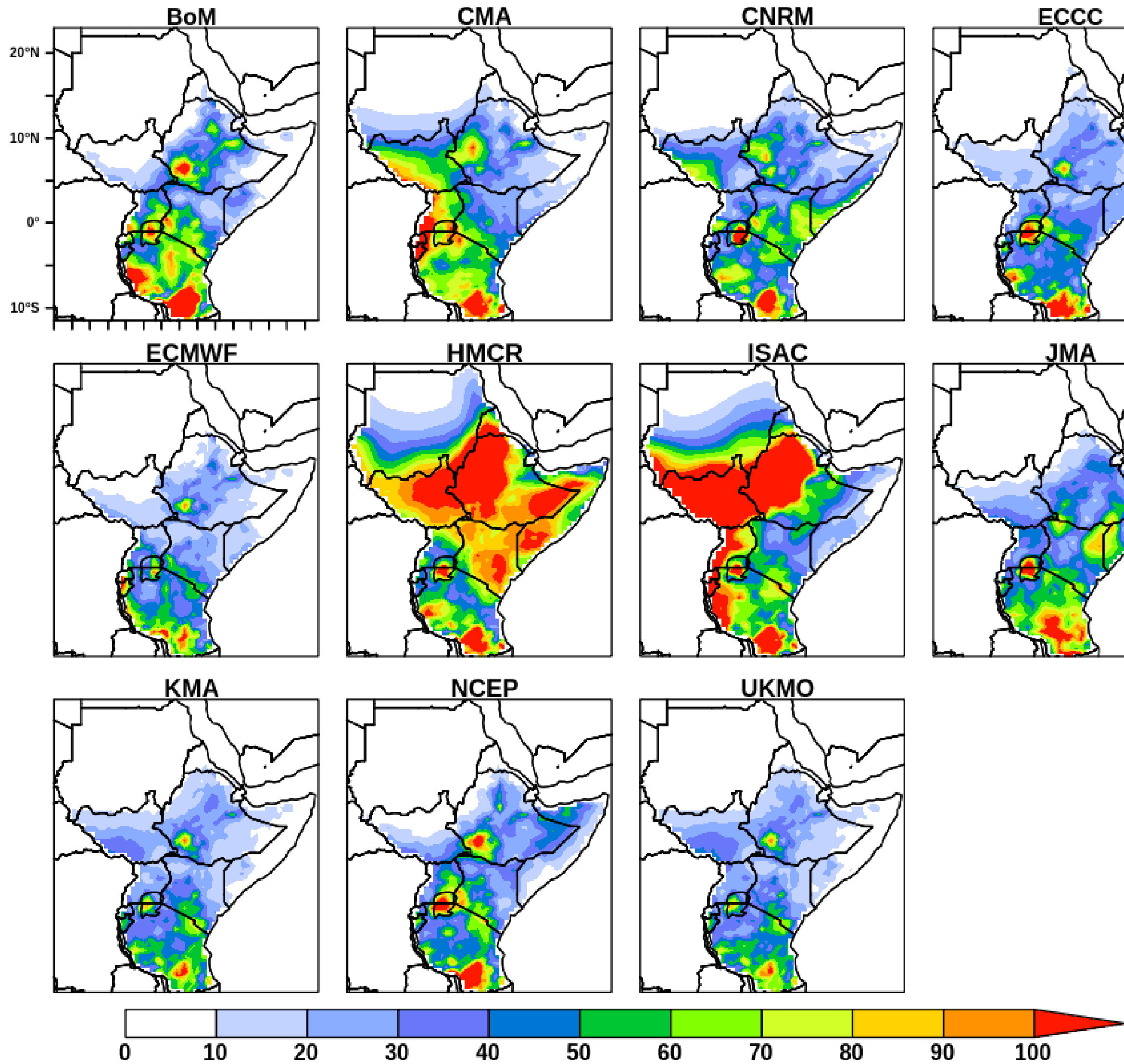

Figure 3a. Spatial distribution of RMSE of rainfall between 11 S2S models and CHIRPS during March over GHA.

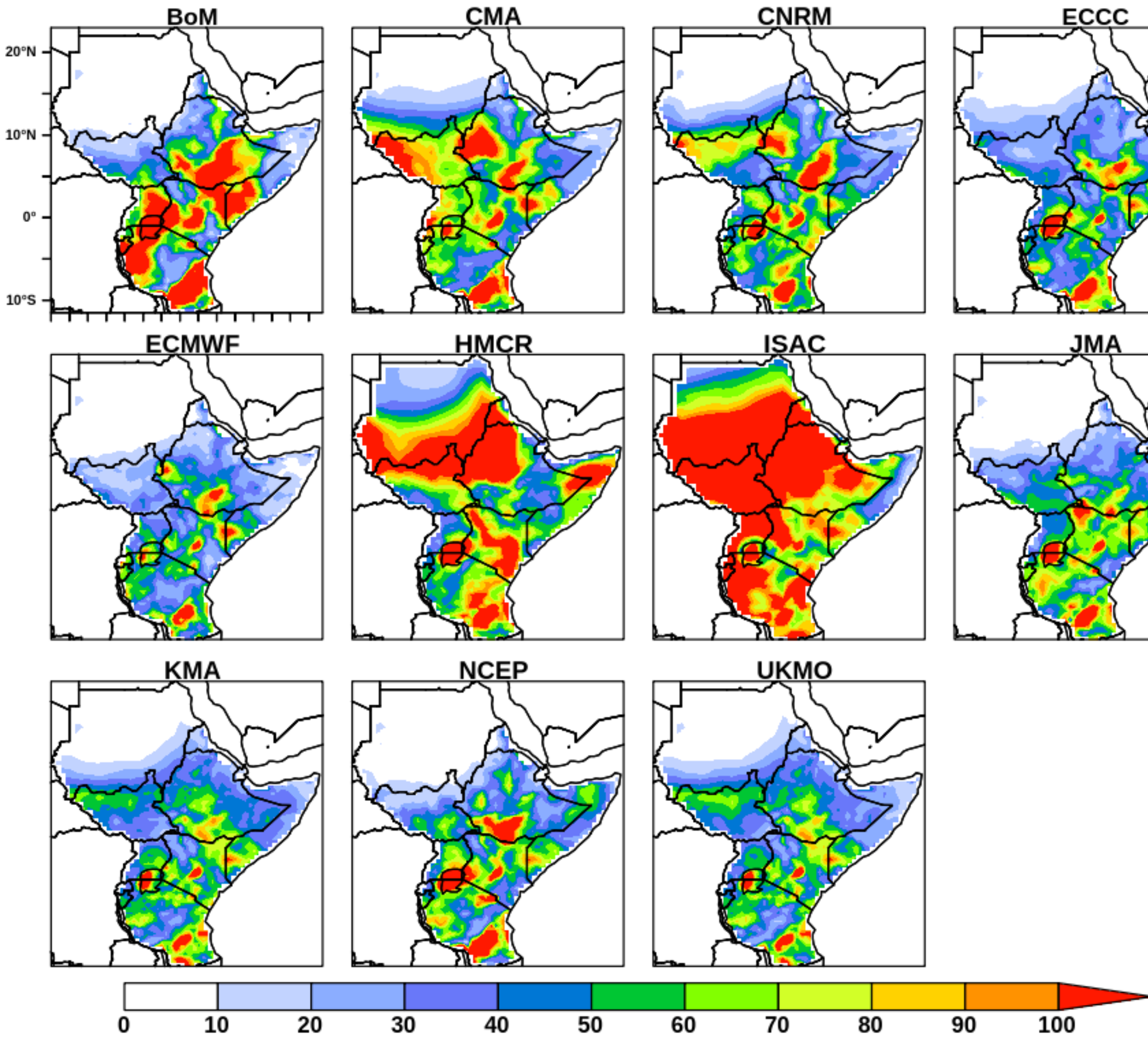

Figure 3b. Spatial distribution of RMSE of rainfall between 11 S2S models and CHIRPS during April over GHA.

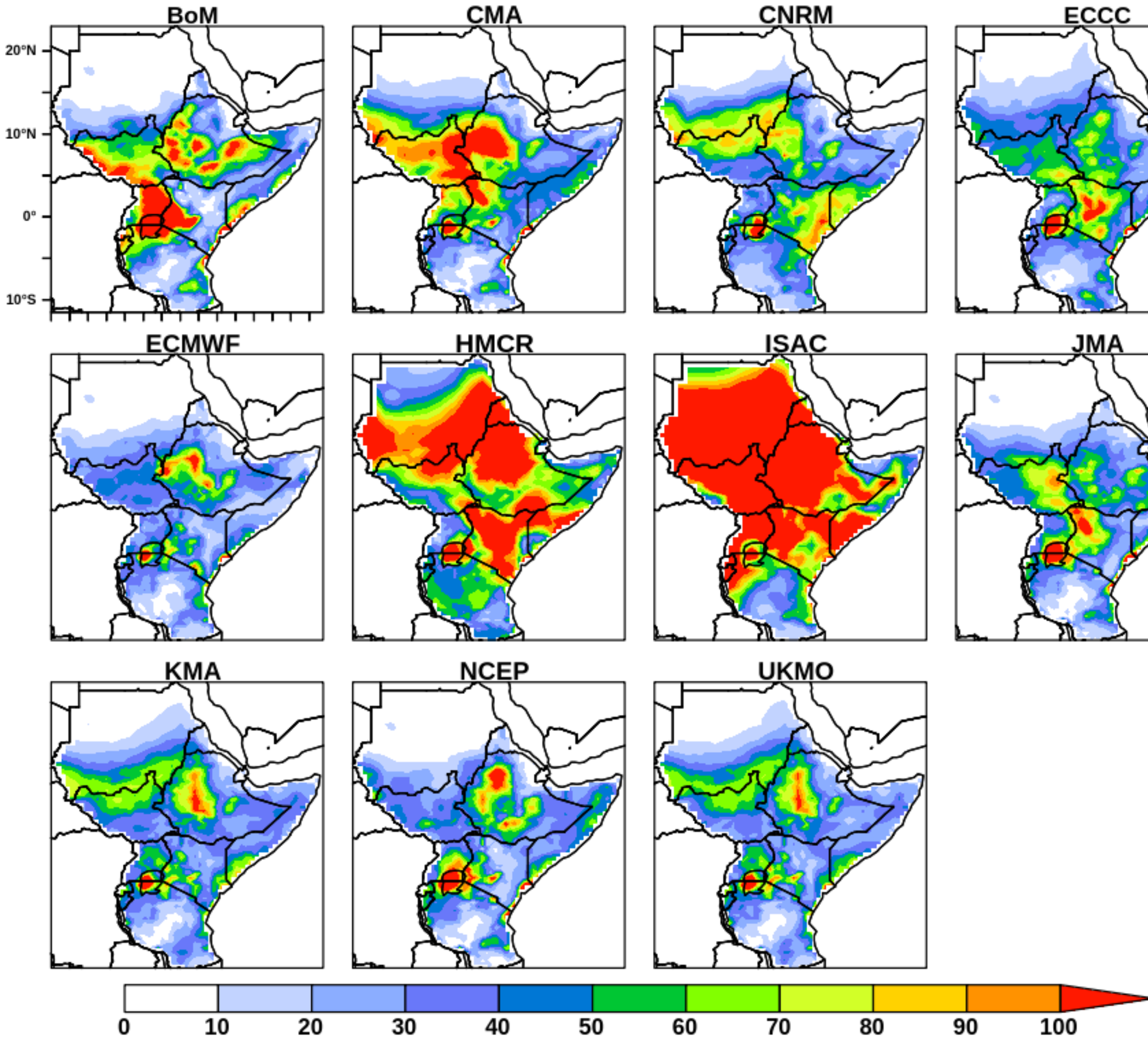

Figure 3c. Spatial distribution of RMSE of rainfall between 11 S2S models and CHIRPS during May over GHA.

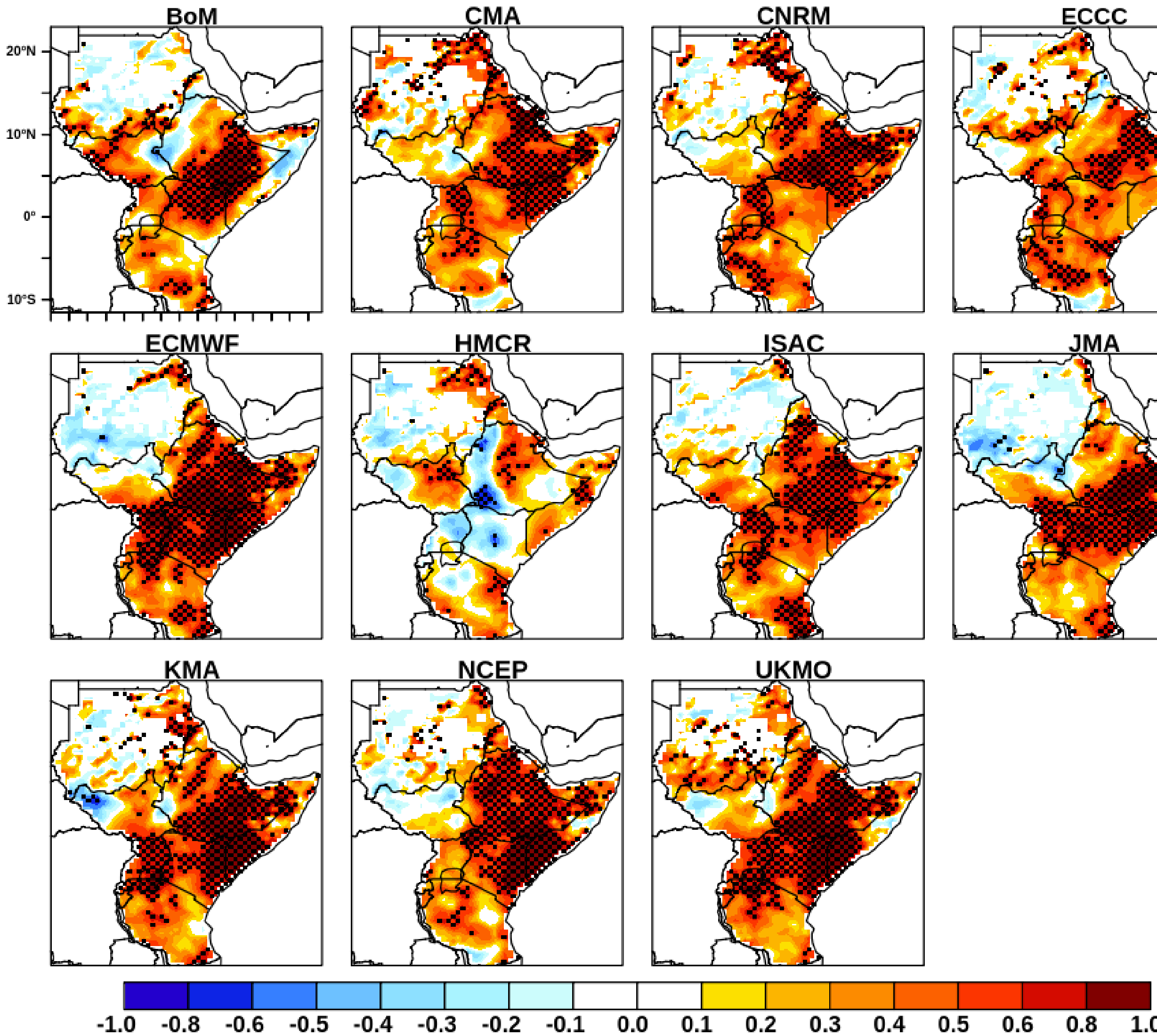

Figure 4a. Spatial distribution of correlation coefficient of rainfall between models and CHIRPS during March for the period from 1999 to 2010. Hatching indicates regions where the correlation is statistically significant at the 95% confidence level.

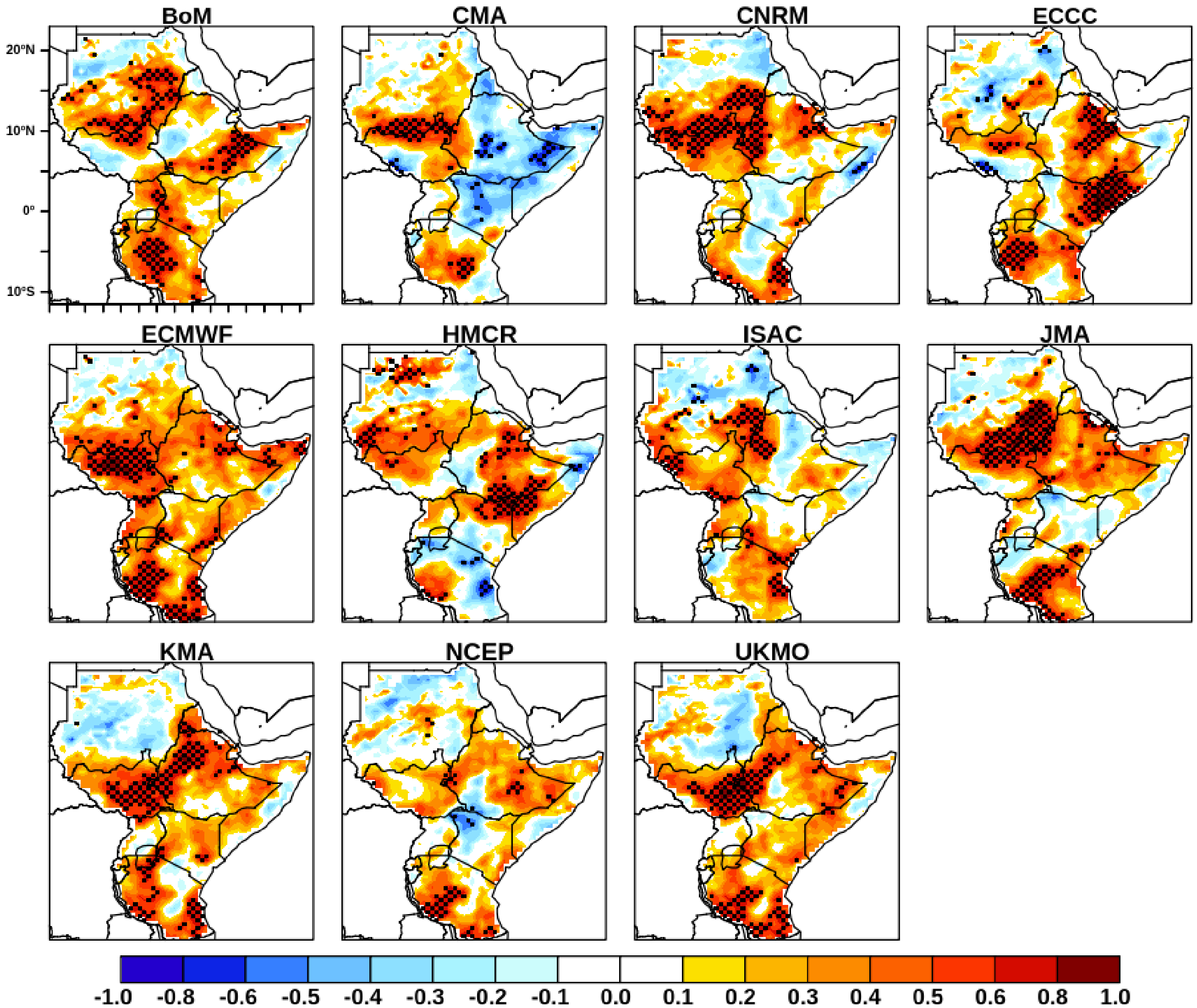

Figure 4b. Spatial distribution of correlation coefficient of rainfall between models and CHIRPS during April for the period from 1999 to 2010. Hatching indicates regions where the correlation is statistically significant at the 95% confidence level.

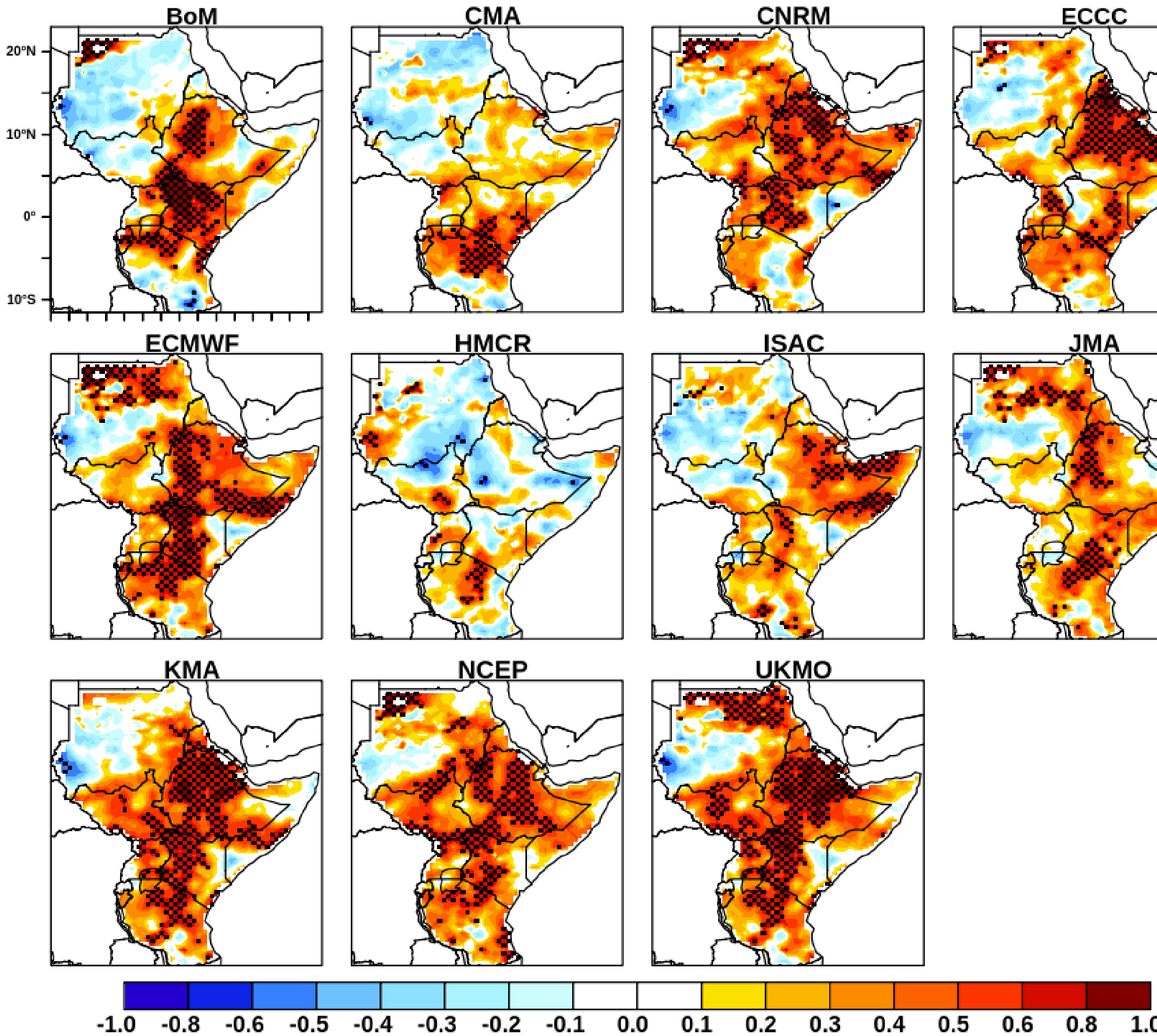

Figure 4c. Spatial distribution of correlation coefficient of rainfall between models and CHIRPS during May for the period from 1999 to 2010. Hatching indicates regions where the correlation is statistically significant at the 95% confidence level.

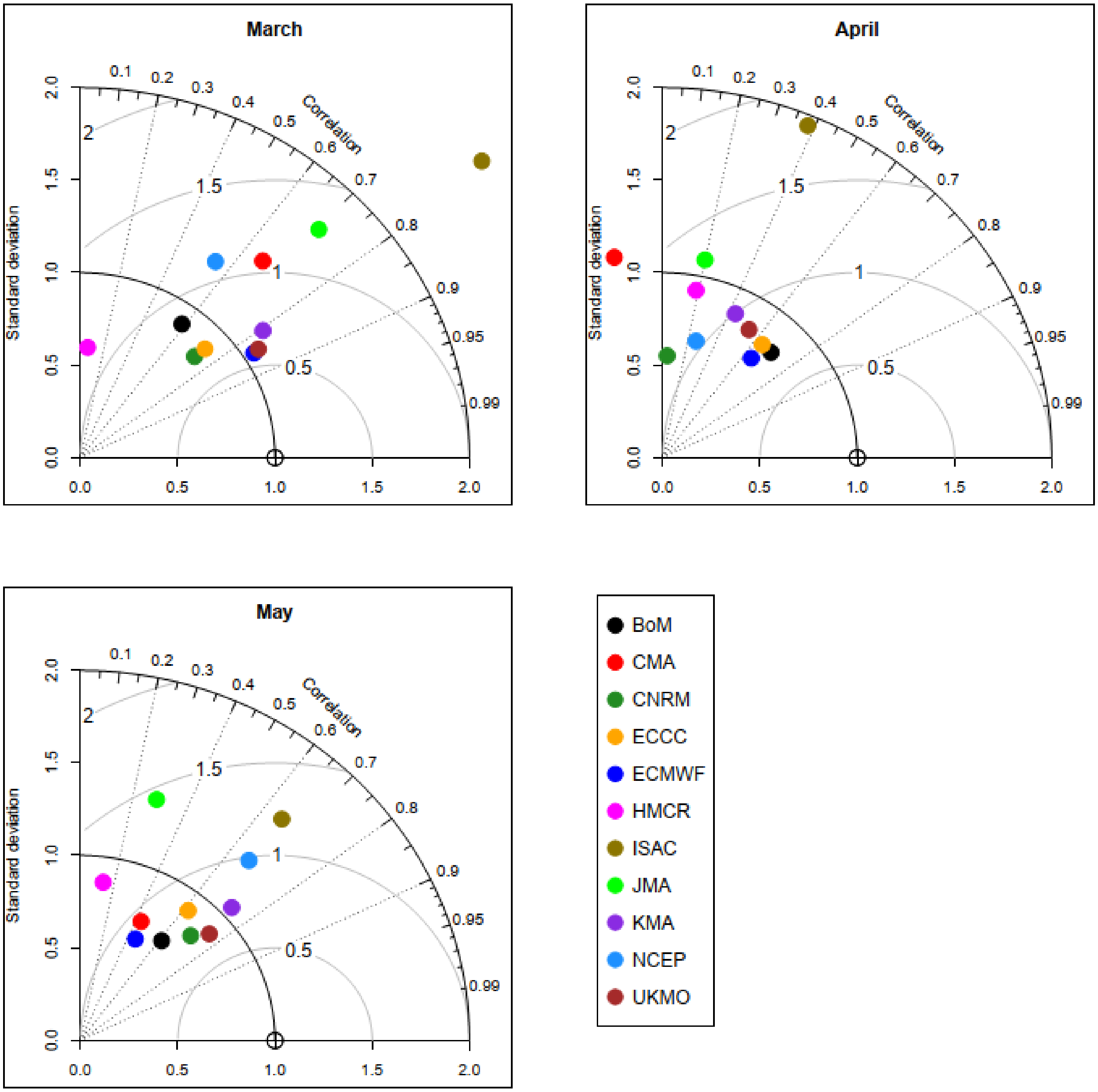

Figure 5. Taylor diagram displaying normalized statistical comparison of monthly total rainfall of the S2S models with CHIRPS during March (top-left), April (top-right), and May (bottom-left).

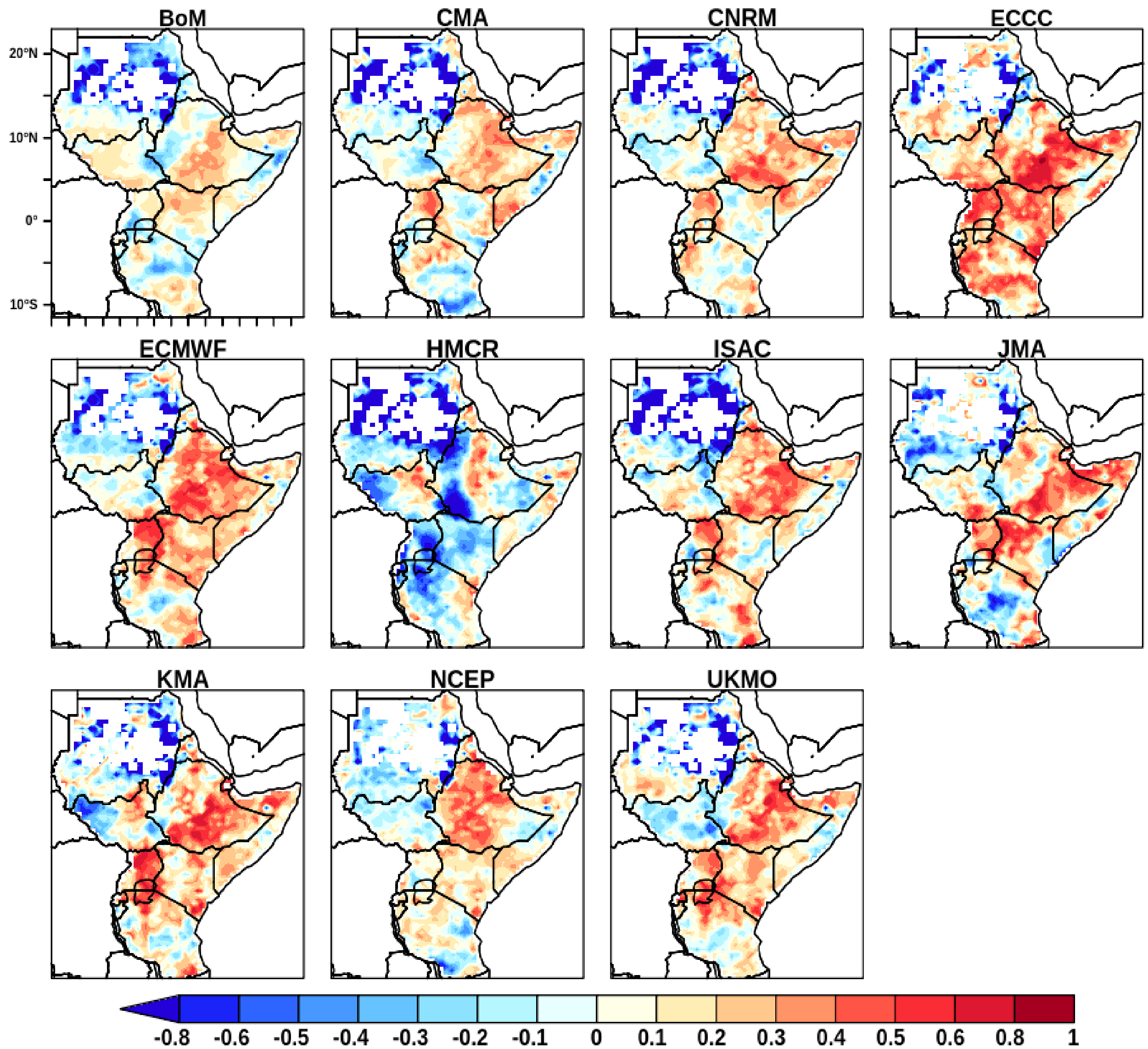

Figure 6a. Ranked Probability Skill Score (RPSS) from 11 S2S models for **March** validated against CHIRPS for the period from 1999 to 2010

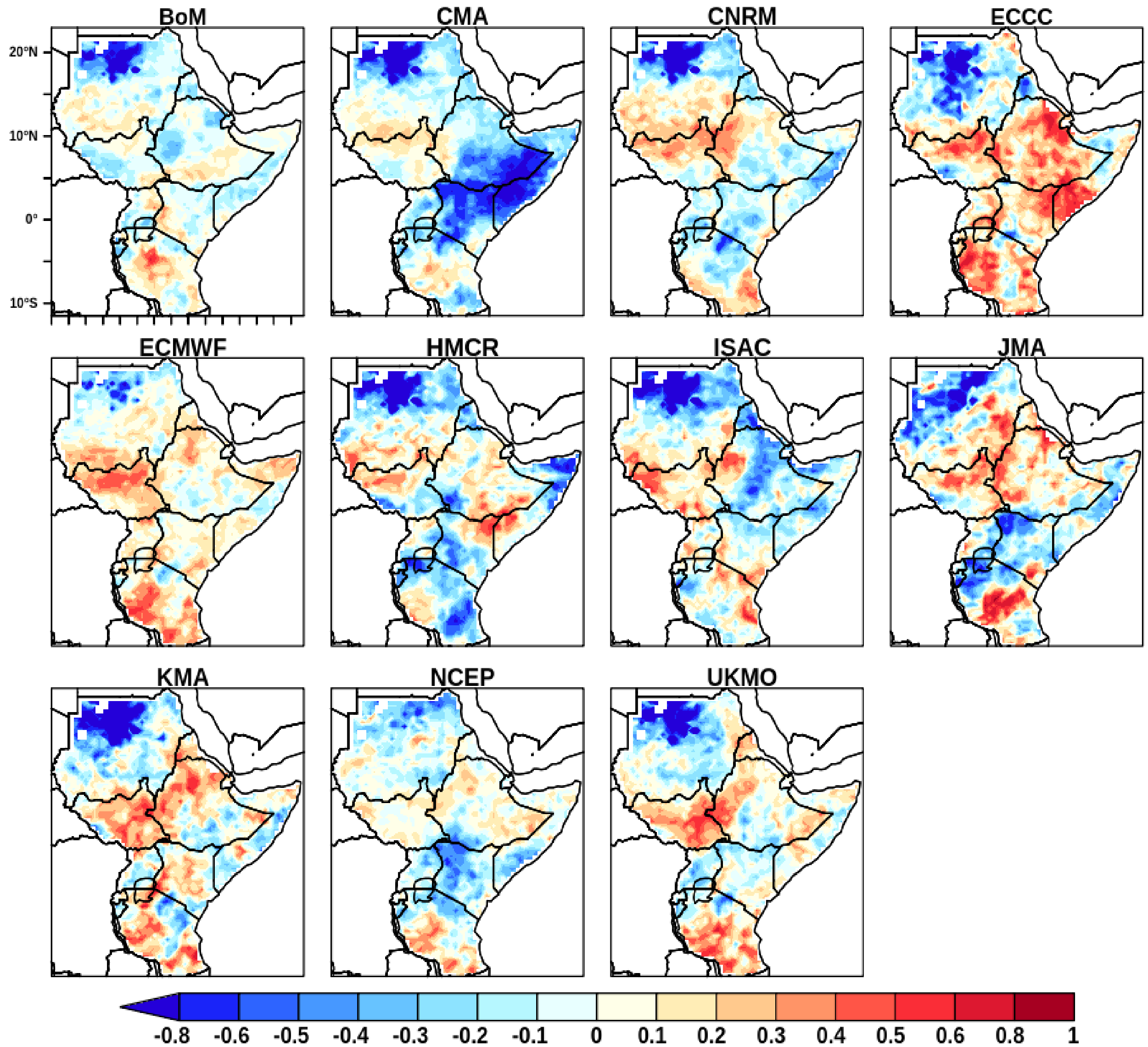

Figure 6b. Ranked Probability Skill Score (RPSS) from 11 S2S models for **April** validated against CHIRPS for the period from 1999 to 2010.

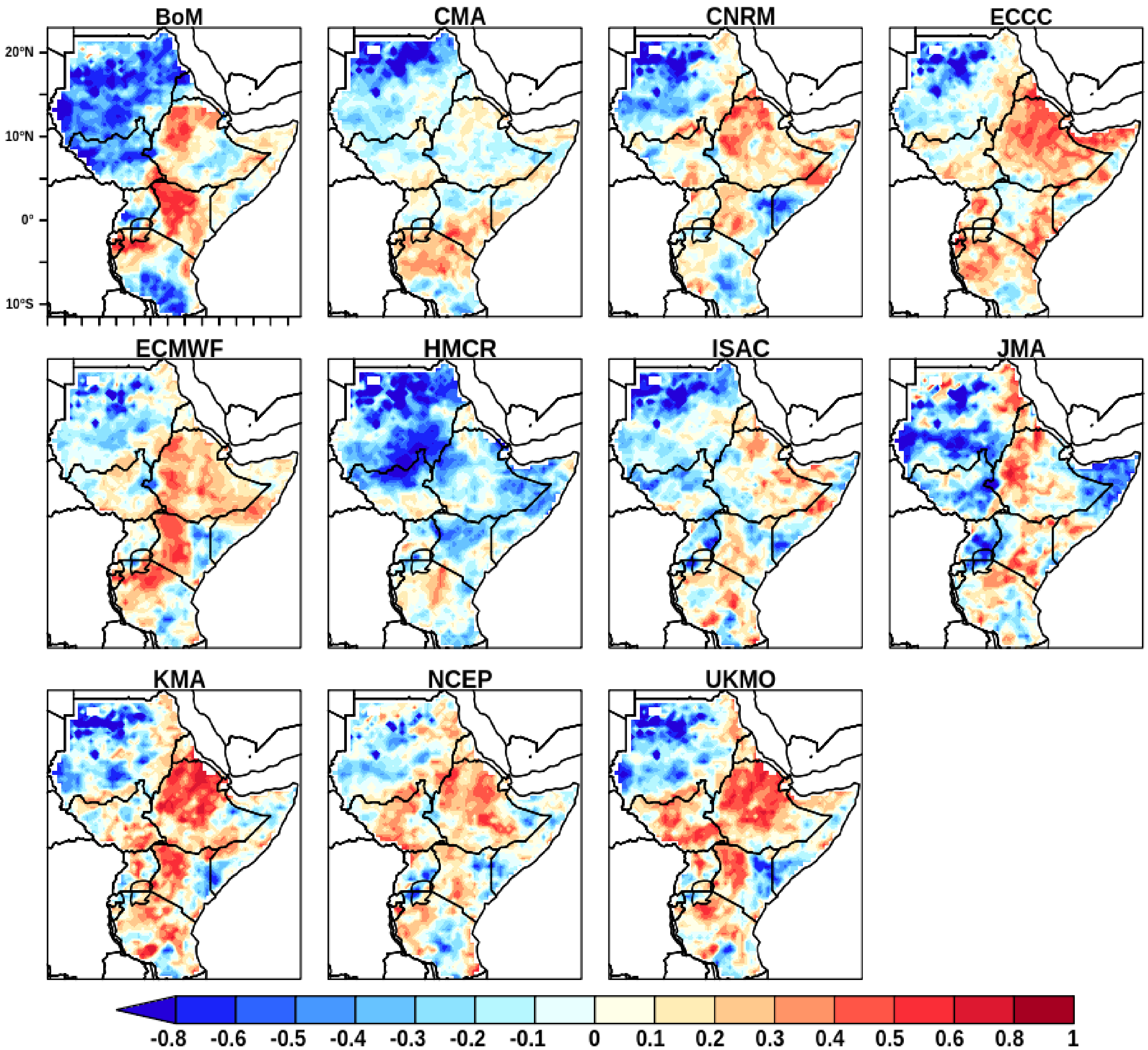

Figure 6c. Ranked Probability Skill Score (RPSS) from 11 S2S models for **May** validated against CHIRPS for the period from 1999 to 2010.

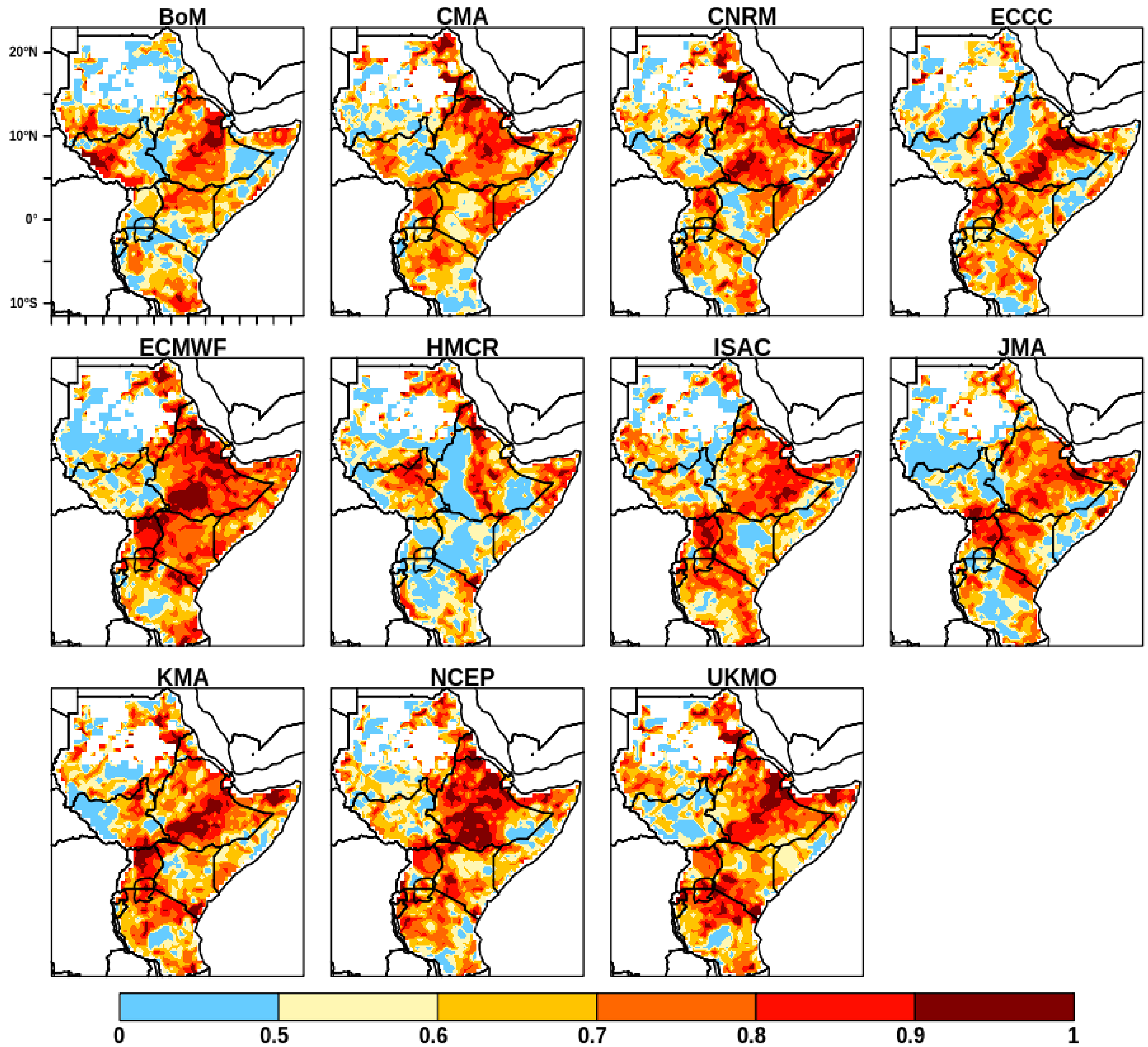

Figure 7a. Relative Operating Characteristic Skill Score **(ROCSS)** for lower tercile during **March** for the period from 1999 to 2010.

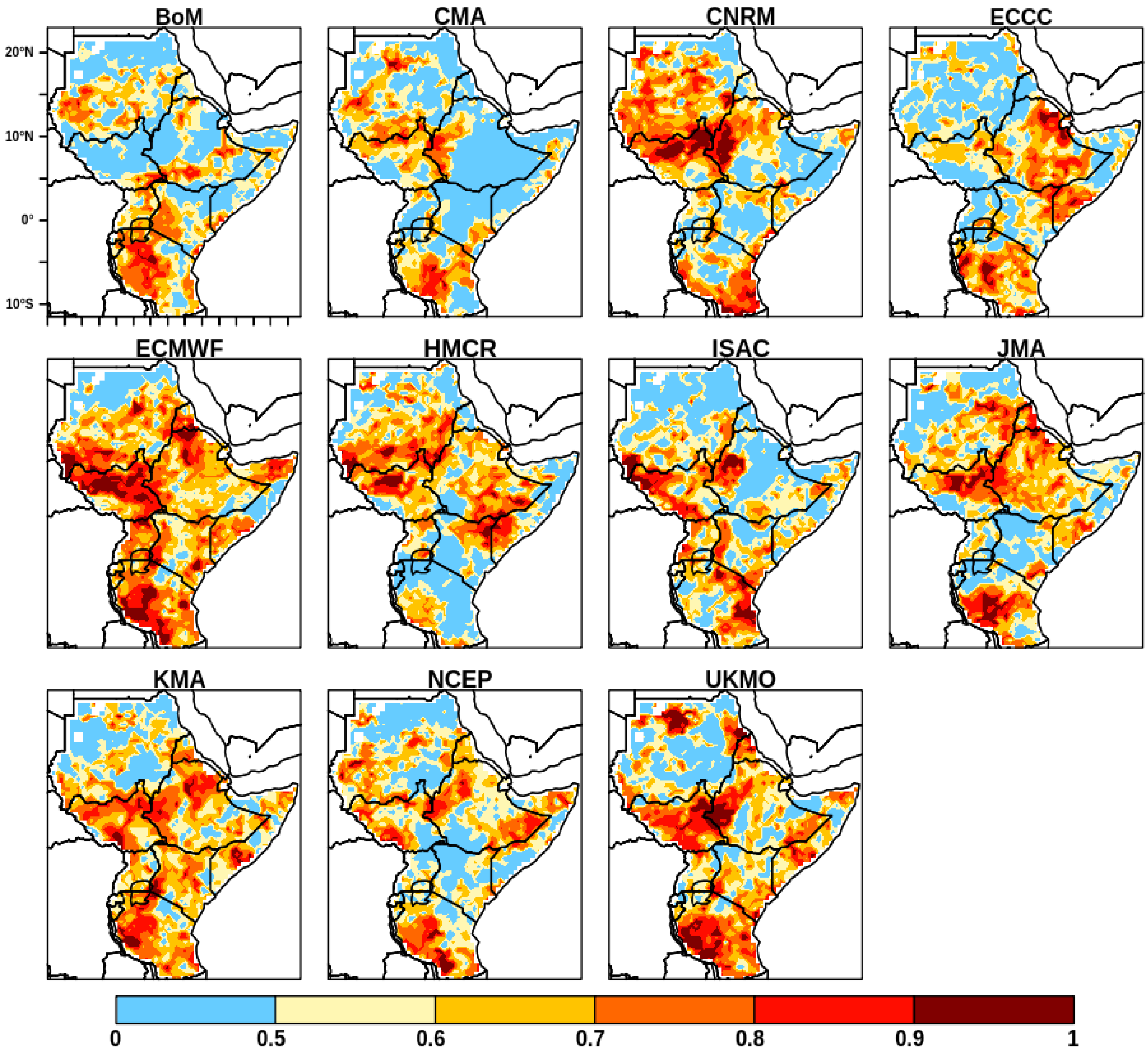

Figure 7b. Relative Operating Characteristic Skill Score **(ROCSS)** for lower tercile during **April** for the period from 1999 to 2010.

.

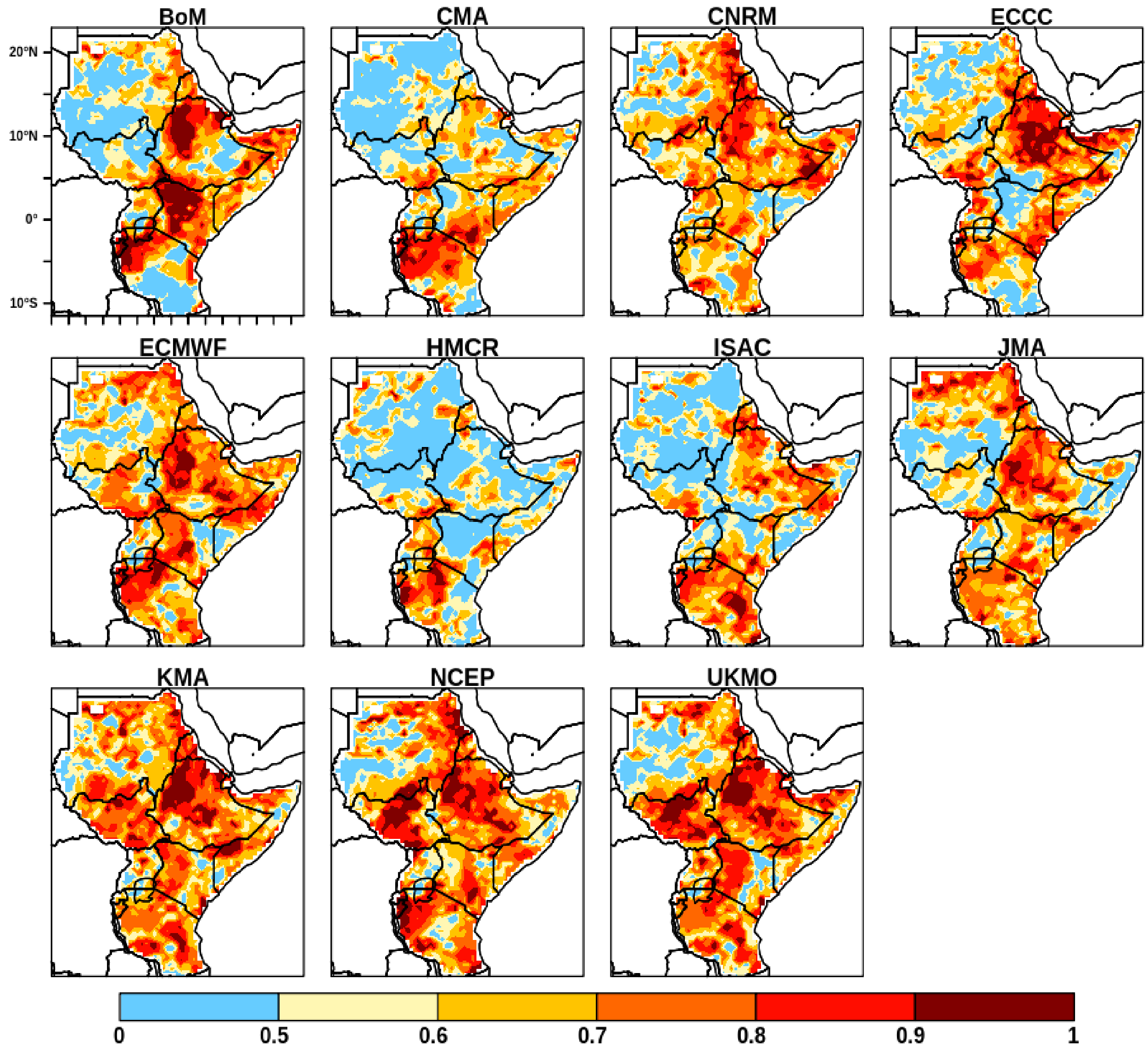

Figure 7c. Relative Operating Characteristic Skill Score **(ROCSS)** for lower tercile during **May** for the period from 1999 to 2010.

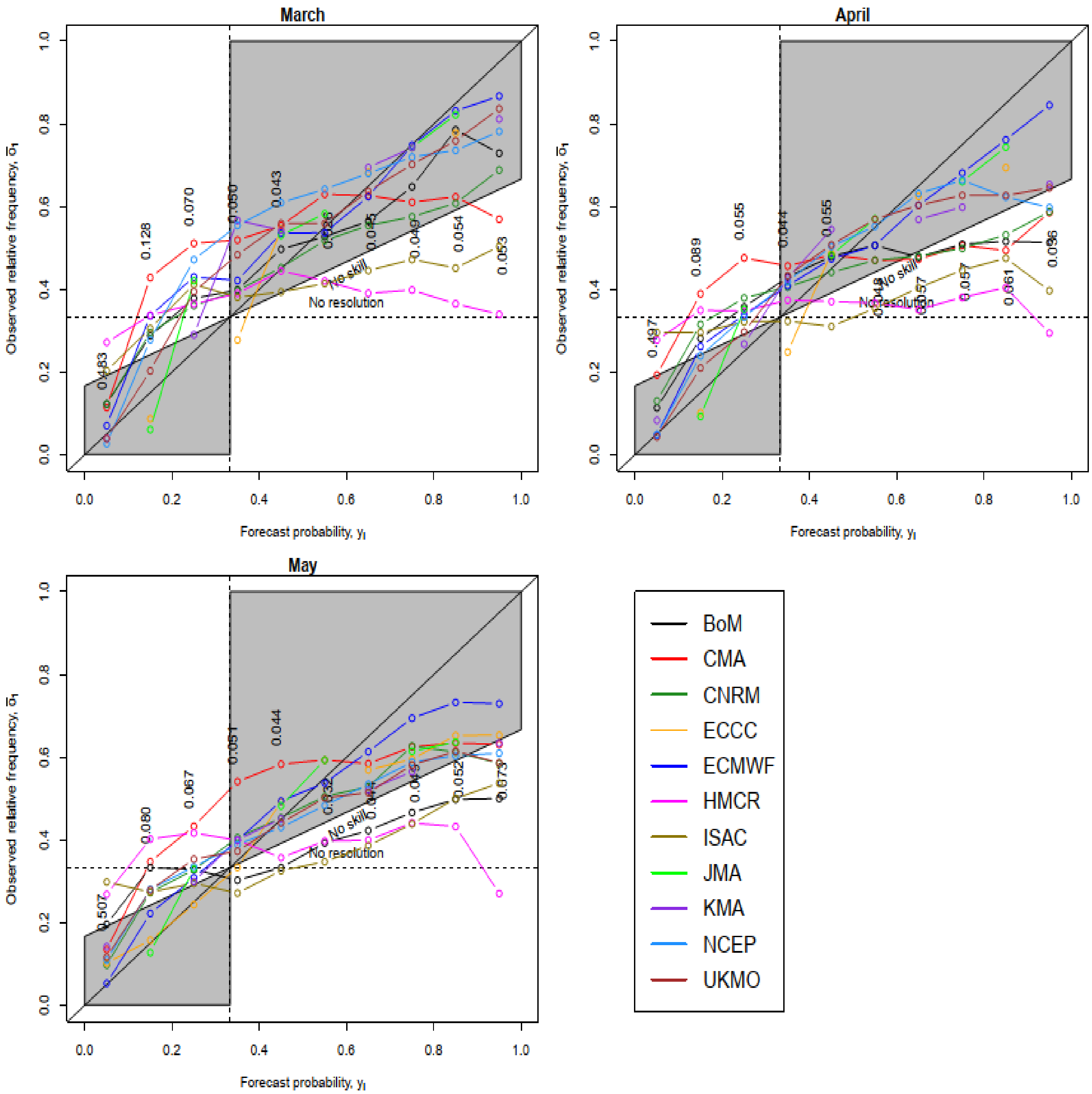

Figure 8. Attribute diagram in predicting the monthly precipitation during March, April and May over GHA for the below-normal category (upper tercile) for the period from 1999 to 2010. In the diagram, the x-axis shows the average forecast probability and the y-axis shows the corresponding observed relative frequency.

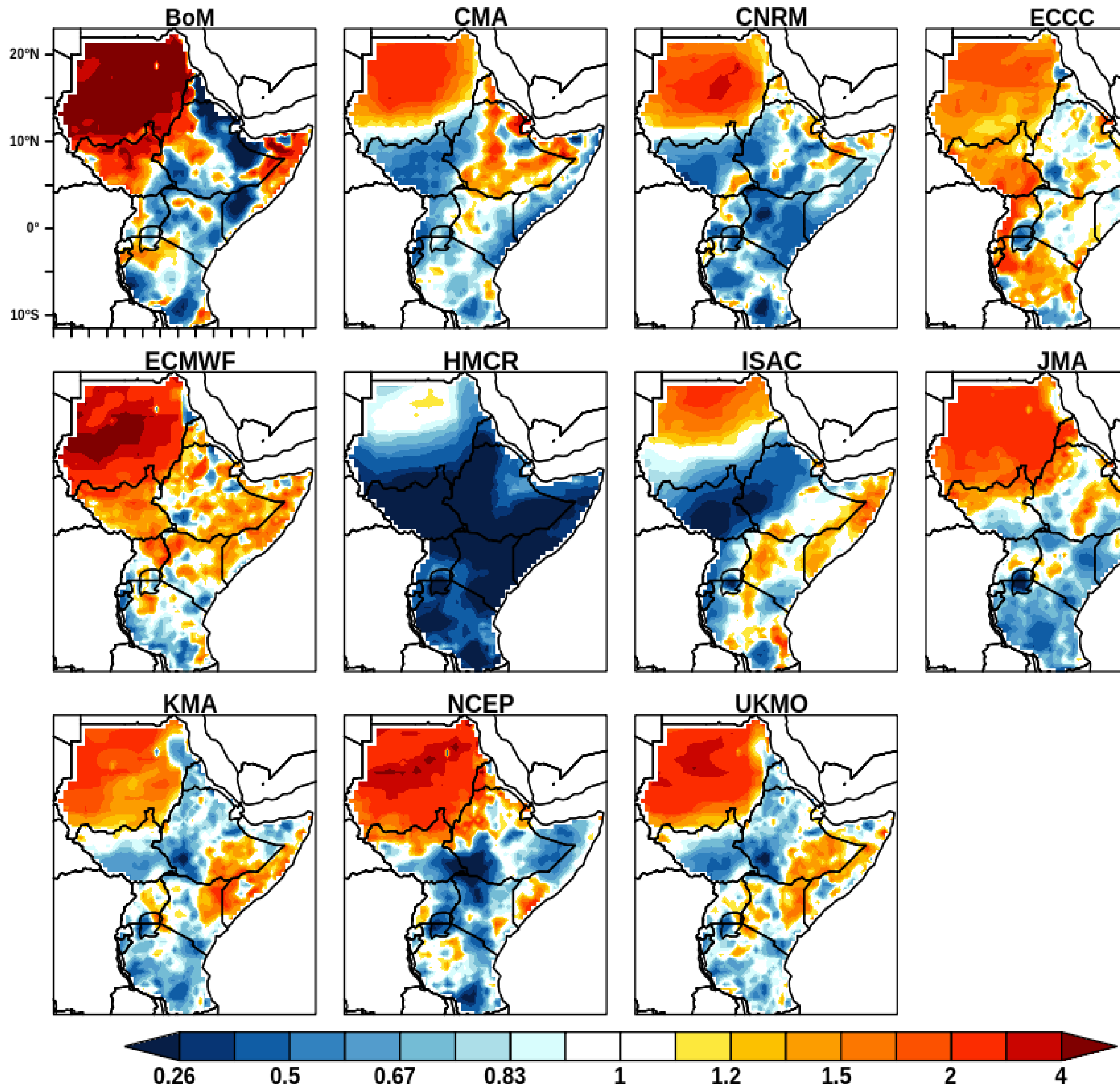

Figure 9a. Spread-error (SPR) ratio for **March** for the period from 1999 to 2010. SPR below 1 indicates underdispersive (overconfidence) and SPR greater than 1 indicates overdispersion (underconfidence).

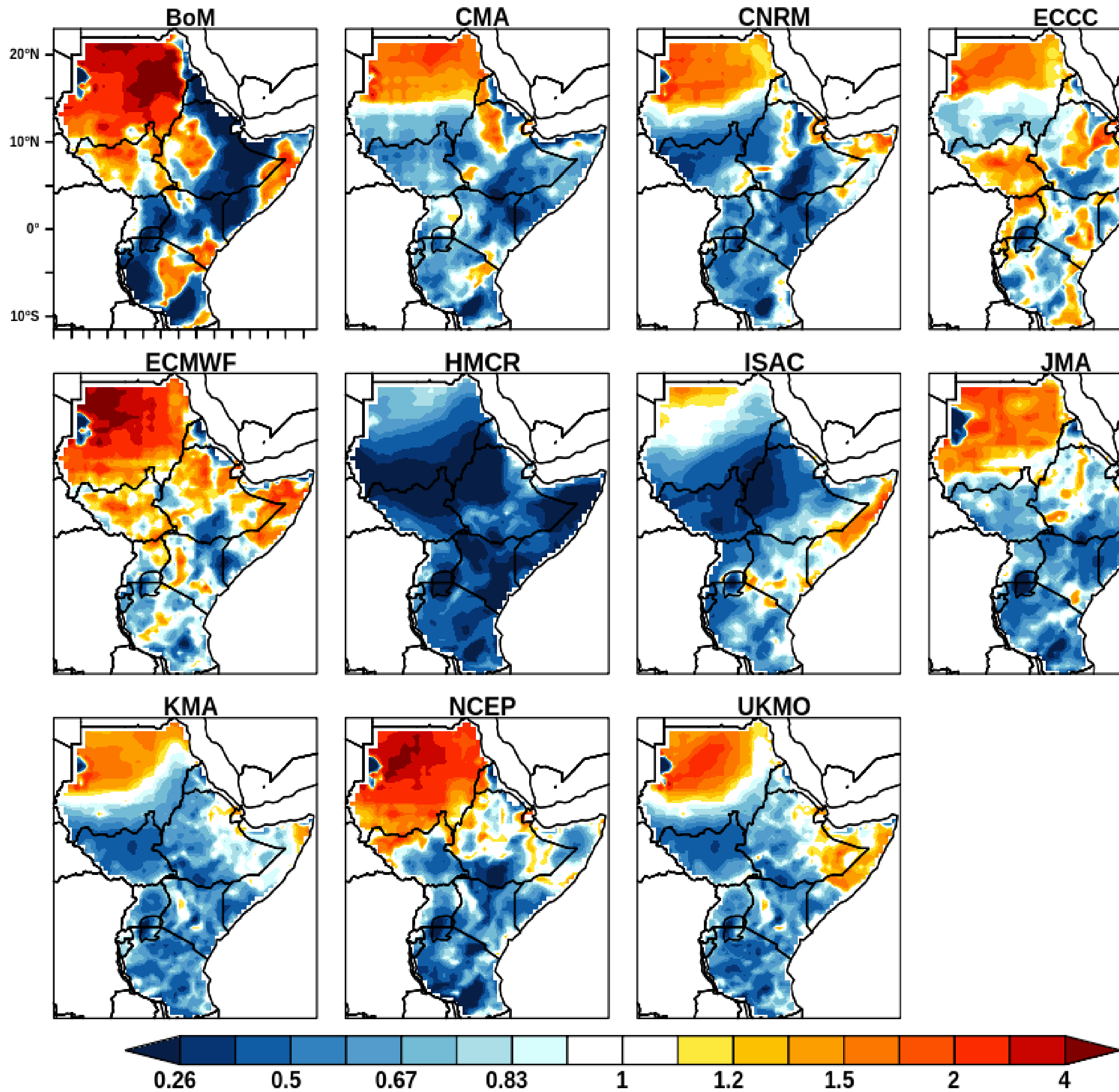

Figure 9b. Spread-error ratio for **April** for the period from 1999 to 2010. SPR below 1 indicates underdispersive (overconfidence) and SPR greater than 1 indicates overdispersion (underconfidence).

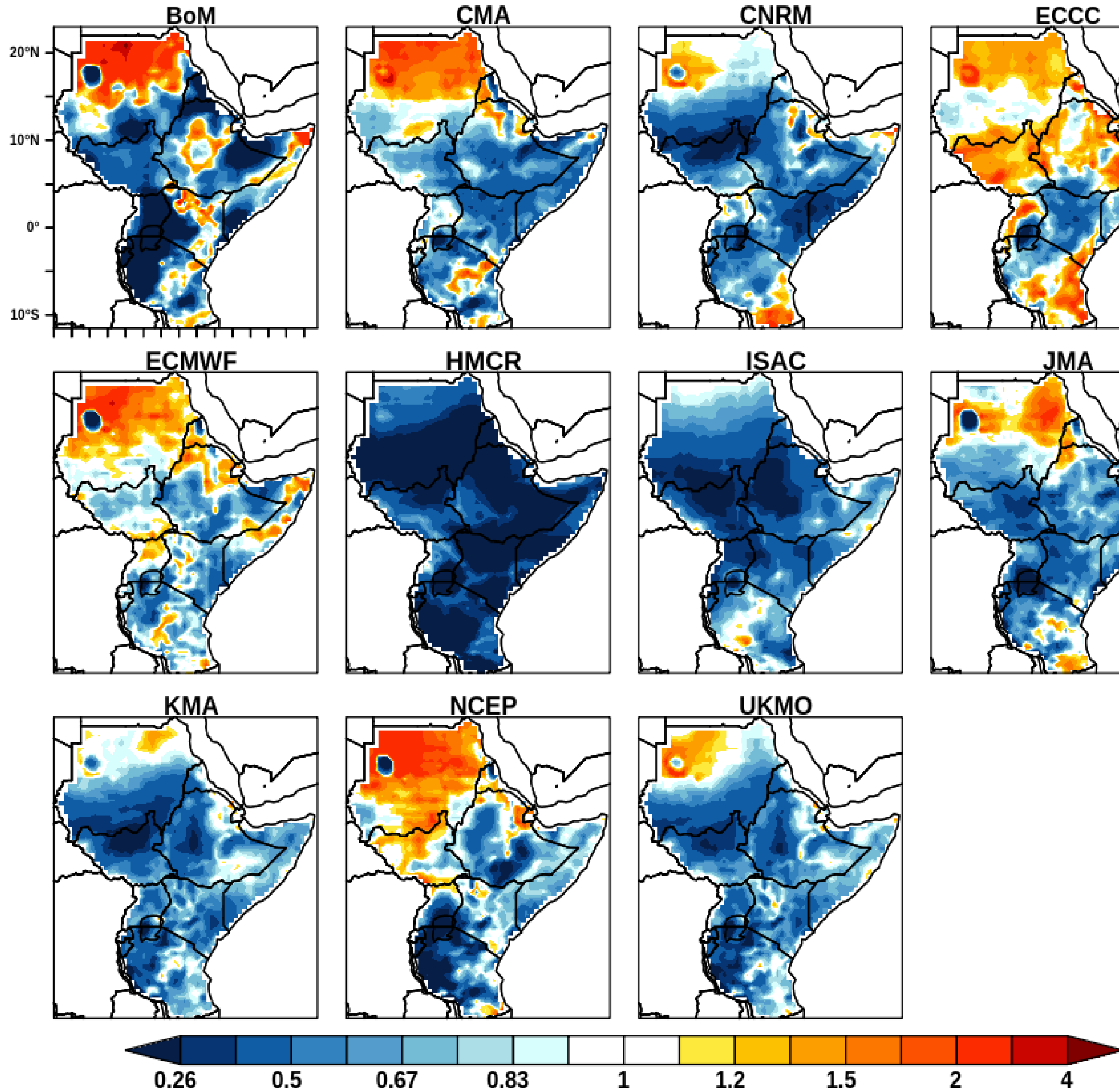

Figure 9c. Spread-error ratio for **May** for the period from 1999 to 2010. SPR below 1 indicates underdispersive (overconfidence) and SPR greater than 1 indicates overdispersion (underconfidence).

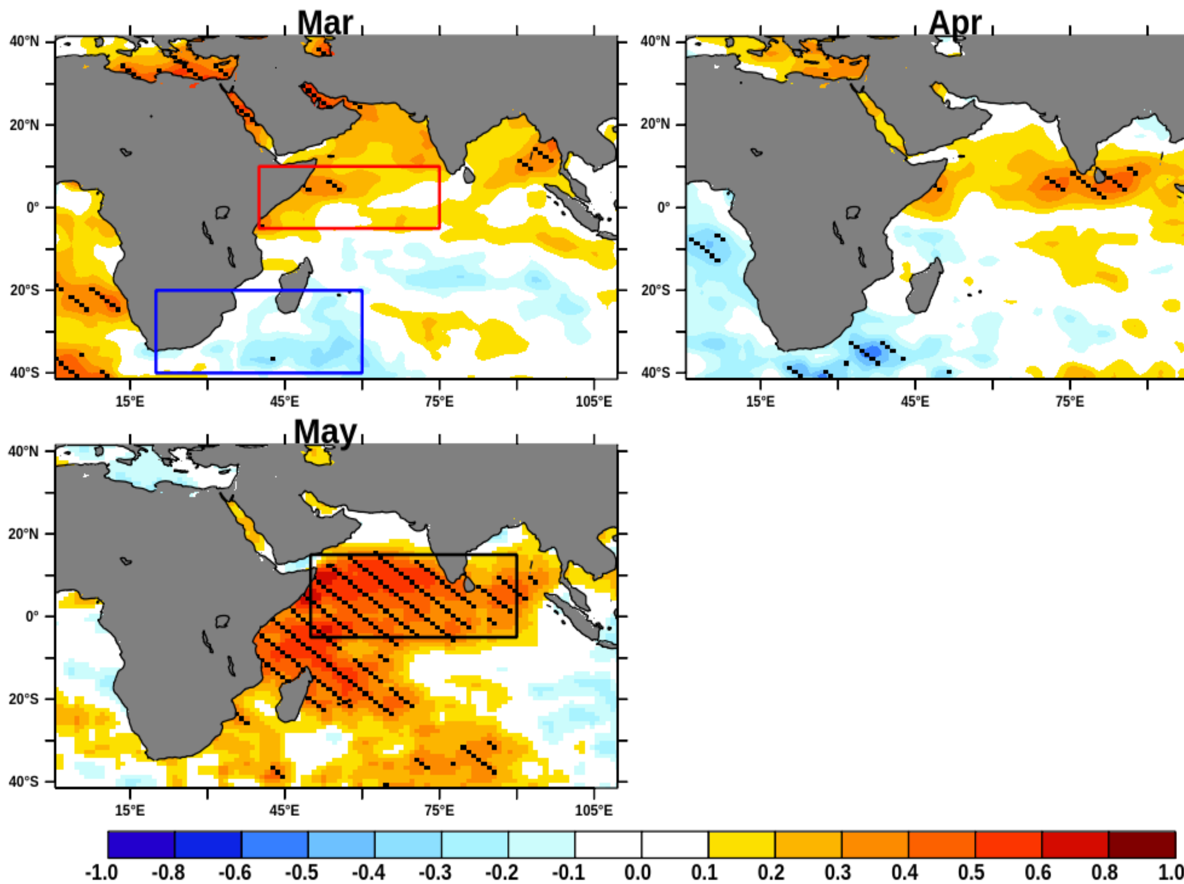

Figure **10a**: Correlations between monthly rainfall (Mar, Apr & May) averaged over GHA region and concurrent grid-point SSTs for the period from 1982-2018 using CHIRPS rainfall and NOAA SST data. Hatching indicates regions where the correlation is statistically significant at the 95% confidence level. The boxes indicate location of SST regions used to compute indices for the regression analysis. For March analyses, a western Indian Ocean meridional index is formed by taking the difference between average SSTs over the northern (red) and southern (blue) boxes shown.

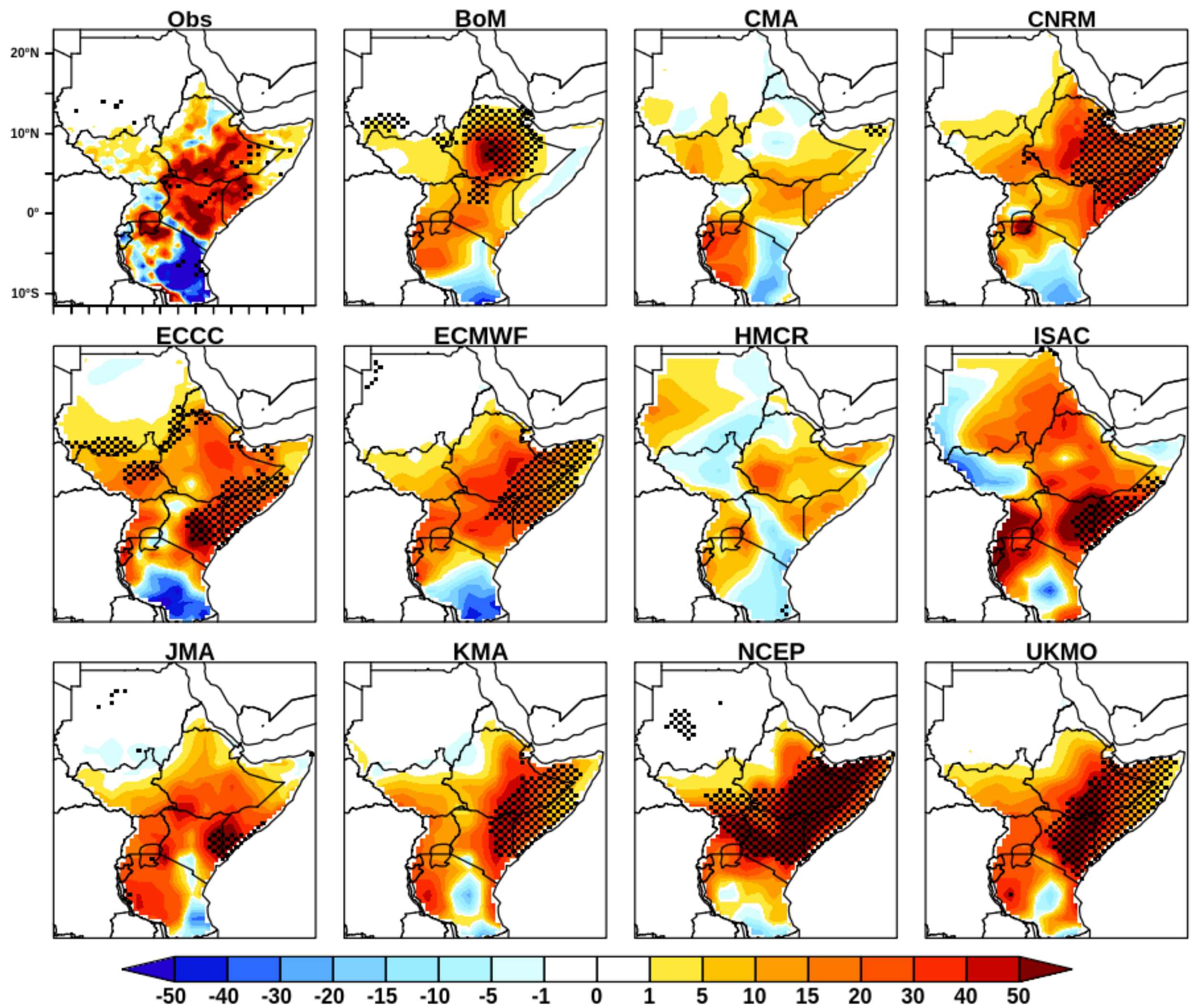

Figure 10b. Linear regression between March rainfall and the SST index (meridional gradient) over the western tropical Indian Ocean for the for the period from 1999–2010. Hatching indicates regions where the regression coefficient is statistically significant at the 95% confidence level. Units are mm/month/°C.

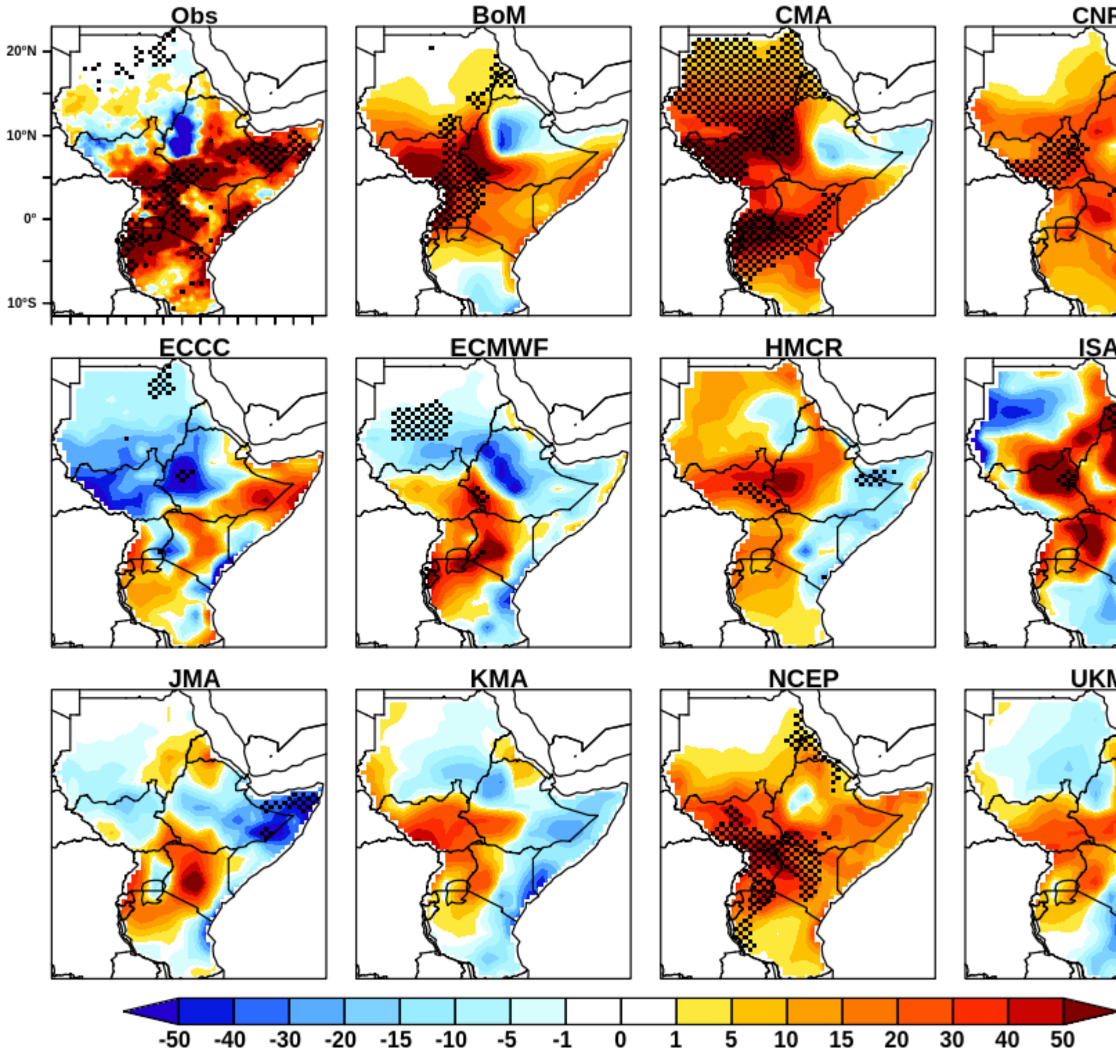

Figure 10c. Linear regression between May rainfall and SST index over the northern tropical Indian Ocean for the period from 1999–2010. Hatching indicates regions where the regression coefficient is statistically significant at the 95% confidence level. Units are mm/month/°C.

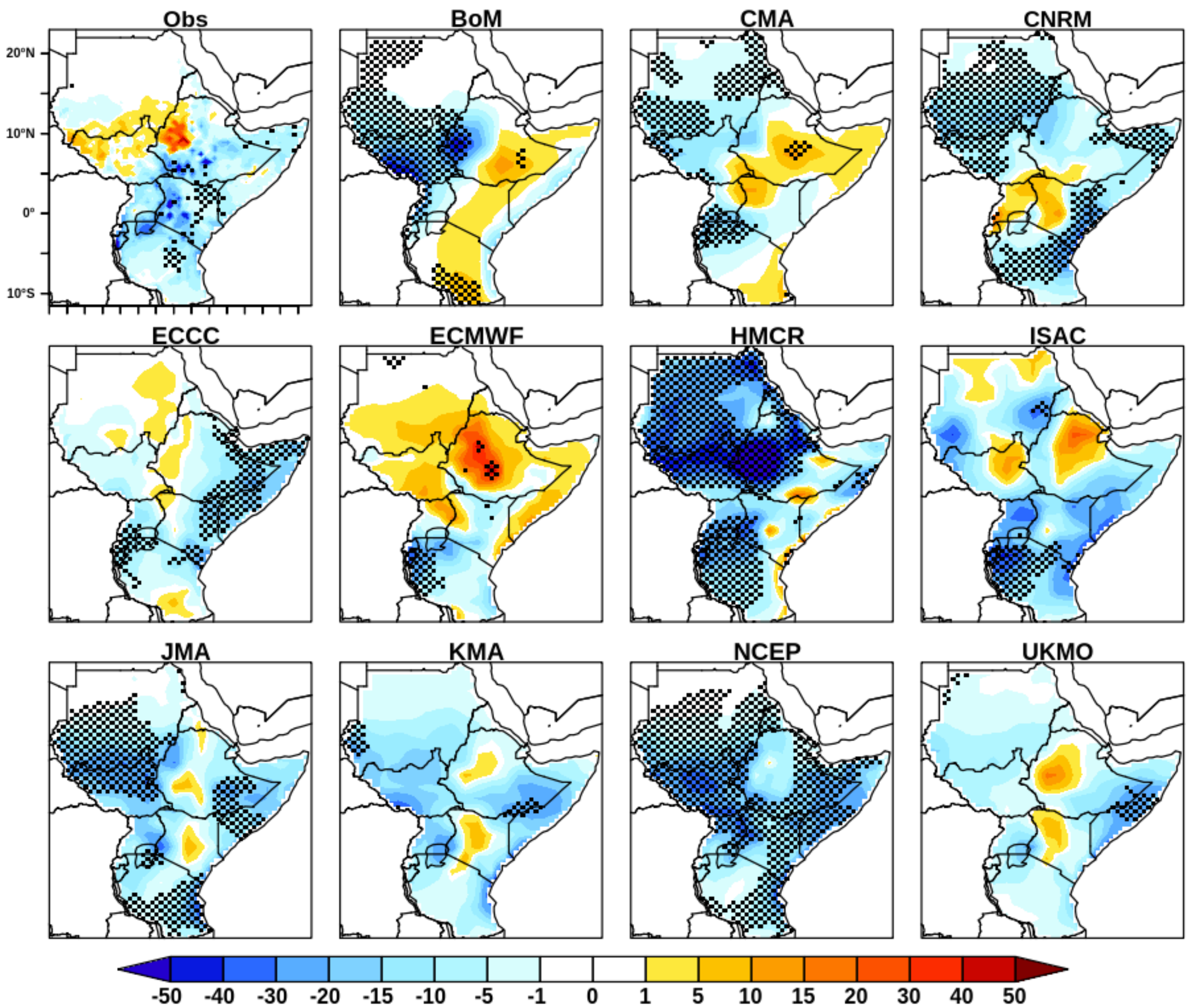

Figure 11. Linear regression between May rainfall and the SLLJ index for the period from 1999–2010. Hatching indicates regions where the regression coefficient is statistically significant at the 95% confidence level. Units are mm/month/m/s.