frontiers in Climate

TYPE Original Research
PUBLISHED 24 March 2026
DOI 10.3389/fclim.2026.1691030

Improved seasonal precipitation forecasts for the Blue Nile Basin: a deep learning approach

Rebecca Wiegels[1,2]*, Christian Chwala[1], Julius Polz[1,3], Luca Glawion[1], Christof Lorenz[1], Tanja C. Schober[1] and Harald Kunstmann[1,2,4]

[1]Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMKIFU), Karlsruhe Institute of Technology (KIT), Garmisch-Partenkirchen, Germany, [2]Institute of Geography, Augsburg University, Augsburg, Germany, [3]Institute of Meteorology and Climate Research - Atmospheric Trace Gases and Remote Sensing (IMKASF), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [4]Center for Climate Resilience, Augsburg University, Augsburg, Germany

Check for updates

OPEN ACCESS

EDITED BY
Anthony Lupo,
University of Missouri, United States

REVIEWED BY
Neil Hart,
University of Oxford, United Kingdom
Jagabandhu Panda,
National Institute of Technology
Rourkela, India

*CORRESPONDENCE
Rebecca Wiegels
✉ rebecca.wiegels@kit.edu

RECEIVED 22 August 2025
REVISED 06 February 2026
ACCEPTED 20 February 2026
PUBLISHED 24 March 2026

CITATION
Wiegels R, Chwala C, Polz J, Glawion L, Lorenz C, Schober TC and Kunstmann H (2026) Improved seasonal precipitation forecasts for the Blue Nile Basin: a deep learning approach.
Front. Clim. 8:1691030.
doi: 10.3389/fclim.2026.1691030

Seasonal precipitation forecasts are essential for climate-sensitive sectors such as agriculture and water management in East Africa. However, the application of seasonal forecasts at regional scales requires post-processing due to systematic errors and insufficient spatial resolution to capture local characteristics. Yet current statistical methods have remaining limitations in terms of spatial consistency and the representation of extreme events. Here, we propose a deep learning approach, Seasonal AFNOCast, based on an adaptive Fourier Neural Operator architecture, to bias-correct and downscale SEAS5 precipitation forecasts for the Blue Nile Basin, a transboundary catchment in Ethiopia and Sudan. We evaluate Seasonal AFNOCast alongside the established statistical method, Bias Correction and Spatial Disaggregation (BCSD), using forecasts from 2017–2023. Results show that both methods substantially improve precipitation distributions, spatial patterns, and the Continuous Ranked Probability Skill Score (CRPSS) of approx. 0.3 compared to raw SEAS5. Despite only modest improvements over climatology across the entire evaluation period (CRPSS approx. 0.03), both methods show clear skill enhancements during the months March to May (MAM), a highly variable yet operationally critical season for decision-making. While onset predictability remains challenging at a seasonal scale, even after post-processing, this study identifies key differences in the application of the post-processing methods: BCSD performs best at short lead times, whereas Seasonal AFNOCast maintains higher skill at longer leads and indicates an improved representation of high-intensity rainfall and spatial frequency characteristics. Moreover, Seasonal AFNOCast generates forecasts 5–20 times faster than BCSD, making it particularly suitable for operational contexts. Our findings show that deep learning can complement and extend conventional post-processing, improving seasonal forecasts for subsequent applications and supporting hydrological and agricultural decision-making where representation of extreme events and spatial consistency, as well as computational efficiency, are critical.

KEYWORDS
bias-correction, convolutional network, deep learning, downscaling, ensemble prediction, Fourier neural operator network, SEAS5, seasonal forecasts

1 Introduction

Seasonal forecasts are crucial to water management, agriculture and disaster preparedness. The daily scale of seasonal forecasts several months into the future offers insights into dry spells, droughts and floods, enabling proactive decision-making (World Meteorological Organization, 2020). This is particularly important for regions highly dependent on precipitation, such as Ethiopia and Sudan, where rainfed agriculture or dam management rely on timely and accurate predictions. However, global seasonal forecasts suffer from biases and drift (Johnson et al., 2019; Saha et al., 2014), as well as from coarse spatial resolution that limits the representation of regional characteristics. Precipitation remains one of the most difficult variables to forecast due to its strong spatial and temporal variability, which includes extreme events as a particularly challenging aspect for forecasting models (Foufoula-Georgiou et al., 2020; Oliveira et al., 2023).

To address challenges of global forecasting systems, the seasonal forecasts are refined by post-processing either with dynamical downscaling of Regional Climate Models or empirical statistical bias-correction techniques. Dynamical downscaling methods are computationally expensive (Hwang et al., 2011; Lorenz et al., 2021) and oftentimes additionally require post-processing with a statistical bias-correction (Hwang et al., 2011; Manzanas et al., 2018; Nikulin et al., 2018). Conventional techniques for statistical bias-correction are mainly done by means of distribution mapping which ranges from simple linear scaling to more advanced techniques like empirical quantile mapping (Crochemore et al., 2016). The Bias Correction and Spatial Disaggregation (BCSD) method, originally proposed by Wood et al. (2002) for bias-correcting climate forecasts, is an example of such a technique. Lorenz et al. (2021) have applied BCSD for post-processing seasonal forecasts, specifically SEAS5 of the European Centre for Medium-Range Weather Forecasts (ECMWF), in the Blue Nile Basin. Despite its effectiveness in improving the daily seasonal forecasts, BCSD has several limitations: first, spatial context and interaction are not captured, as empirical quantile mapping applies corrections independently at each grid cell, not accounting for neighboring interactions. Second, predicting the likelihood of extremes at sub-seasonal to seasonal scale remains challenging (King et al., 2021), and although approaches like constant correction method (Boé et al., 2007) or generalized extreme value fitting (Trentini et al., 2022) have been proposed, their impact on the predictive skill for extreme events has not yet been comprehensively evaluated. Third, the use of broad temporal windows (around 30 days) to estimate the cumulative distribution function (CDF), required to preserve time correlation (Vannitsem et al., 2021) and produce climatologically robust forecasts, can reduce prediction skill of seasonal transitions, particularly around the onset of the rainy season.

Deep Learning (DL) techniques have demonstrated large potential in weather forecasting (Lam et al., 2023; Kurth et al., 2023; Pathak et al., 2022) showing similar skill to conventional forecasting systems while requiring orders of magnitude less computational resources. Furthermore, the application of DL methods for post-processing weather to climate forecasts have proven successful in both, bias-correction and downscaling tasks. For bias-correction, studies have employed Convolutional Neural Networks (CNNs) (Han et al., 2021) or Generative Adversarial Networks (GANs) (Pan et al., 2021). For downscaling, DL has been effective from climate scales (Harder et al., 2023; Vandal et al., 2017; Prasad et al., 2024; Yang et al., 2024; Glawion et al., 2025) to weather prediction (Leinonen et al., 2023; Koldunov et al., 2024; Han et al., 2021). The Fourier Neural Operator (FNO) architecture, originally developed and proven to successfully solve partial differential equations (Li et al., 2021), has so far seen limited use in downscaling. Recent work by Yang et al. (2024) shows that an FNO-based model can significantly outperform super-resolution CNNs, GANs and Swin Transformers for downscaling tasks. Prasad et al. (2024) further highlight the superior transferability of FNOs and Vision Transformers, with FNOs excelling in cross-variable generalization. Despite these advances, DL applications for post-processing seasonal forecasts have so far been insufficiently explored. While DL methods for quintile prediction were investigated, such as Civitarese et al. (2021) demonstrating the skill of transformer architectures in extreme precipitation forecasts, there is missing research on DL techniques applied for post-processing daily probabilistic seasonal forecasts. To the best of our knowledge, one exception is Yang et al. (2025) who introduce a CycleGAN that can simultaneously downscale and bias–correct daily summer precipitation forecasts over China.

To address the challenges of daily-scale seasonal precipitation post-processing, we implement a hybrid Adapted Fourier Neural Operator–Convolutional (AFNO-Conv) architecture. This choice is motivated by three key advantages. First, AFNO-based models have shown high skill in downscaling tasks, outperforming CNNs, GANs, and transformer-based models in reconstructing high-resolution precipitation patterns (Yang et al., 2024). Second, the AFNO architecture is computationally efficient, making it suitable for large ensemble prediction and operational workflows requiring high-resolution spatial output (Pathak et al., 2022). Third, FNOs offer strong transferability across variables and spatial domains, performing particularly well in generalization tasks compared to other deep learning models (Prasad et al., 2024). In addition, we assume the hybrid AFNO-Conv setup captures both local and non-local spatial and temporal dependencies. Together, these properties make the AFNO-Conv architecture a promising approach for improving seasonal precipitation forecasts to enable subsequent impact modeling over complex regions like the Blue Nile Basin. The transboundary Blue Nile Basin, spanning Ethiopia and Sudan, is a climatically and topographically complex region, where precipitation is highly variable in space and time. Accurate rainfall prediction is critical for rainfed agriculture, hydropower, and water resource management. The basin experiences two distinct rainy periods: the pre-season (March–May, MAM) and the main rainy season (June–September, JJAS). Forecasting the onset and variability of these seasons is essential, as deviations can severely affect agricultural productivity and water availability. While some studies have investigated rainfall onset and variability (Lala et al., 2020; Scheuerer et al., 2024), the skill of post-processed seasonal forecasts, particularly for SEAS5, remains largely unexplored for this region. This highlights the urgent need for improved forecasting methods that enhance both accuracy and spatial representation of seasonal precipitation.

In this study, we evaluate the effectiveness of our new Seasonal AFNOCast deep learning architecture for bias-correcting and downscaling SEAS5 precipitation forecasts in the Blue Nile Basin, with the goal of generating physically plausible precipitation distributions, improving skill, and tackling key limitations of conventional statistical approaches while maintaining computational efficiency for operational applications. In particular, we address the following research questions:

1. To what extent can Seasonal AFNOCast accurately reproduce precipitation distributions and patterns compared to the reference data?

2. How does the skill of Seasonal AFNOCast compare to conventional statistical post-processing methods, with BCSD as benchmark method, for precipitation forecasting in the Blue Nile Basin?

3. What is the added value of the DL approach Seasonal AFNOCast compared to conventional approaches in terms of spatial consistency, prediction of extreme precipitation and onset predictability of the Blue Nile Basin?

2 Materials and methods

2.1 Study region: the Blue Nile Basin

The Blue Nile Basin was selected as study region both for its hydrological significance and its relevance for regional climate services. As part of the mandate of the IGAD Climate Prediction and Application Centre (ICPAC), our collaboration with this institution through trainings and capacity-building activities aims to ensure that the methods developed here can eventually support operational seasonal forecasting.

The Blue Nile Basin is a transboundary basin located in Ethiopia and Sudan (see Figure 1). The Blue Nile originates in the highlands of Ethiopia and meets downstream with the White Nile in Khartoum and finally reaches Egypt. The Nile runoff is dominated by the Blue Nile river with more than a 60% contribution (Yitayew and Melesse, 2011). All three countries heavily depend on the water, i.e., energy production of Grand Ethiopian Renaissance Dam or on irrigation schemes and rainfed agriculture in Sudan and Ethiopia (Ahmed, 2020). However, the region is vulnerable to floods and droughts and the impacts of such disasters can potentially lead to failed crop-seasons, loss of livestock and damage to forestry and fishery (Food and Agriculture Organization of the United Nations (FAO), 2020). Two different climatic zones divide the Blue Nile Basin. First, the humid Ethiopian highlands with an annual mean rainfall of $> 2,000$mm (Ali et al., 2014). This part, the Upper Blue Nile, is dominated by a 3-season regime: the dry season Bega (October–January), the mid-rainy season Belg (February/March–May), and Kiremt (June–September). The main rainy season, Kiremt, accounts for 65%–95% of total annual rainfall in the country (Lala et al., 2020). Second, the semi-arid southeast of Sudan with annual rainfalls $<$ 200 mm in Khartoum (Ali et al., 2014) in which the rainy season lasts from June to September (Ahmed, 2020). Given the varying definitions of rainfall seasons in the literature (Nicholson, 2017), this study adopts the widely used classification for East Africa (Haile et al., 2020; Seregina et al., 2019): March–May (MAM), corresponding to the Belg season in the Upper Blue Nile, and June–September (JJAS), covering both the Kiremt season in the Upper Blue Nile Basin and the rainy season over the north of the Blue Nile Basin.

2.2 Data

2.2.1 ERA5-land

The reanalysis precipitation data from the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5-Land (Muñoz-Sabater et al., 2021) is applied as reference product for the statistical post-processing method, as training target set for the deep learning approach, and as reference for the evaluation. ERA5-Land is based on ERA5, the 5th generation of European Reanalysis. The enhanced product has a resolution of 0.1° (about 11 km in the Blue Nile Basin), and a temporal resolution of 1 h. For this study, the precipitation data was directly downloaded as 24-h accumulations, time-stamped at 00:00 UTC of the following day (ECMWF, 2025). Accordingly, the time stamp corresponds to the end of the accumulation interval. To associate each accumulation with the day during which precipitation occurred, we shifted the time labels backward by one day.

While widely used, ERA5-Land has known limitations in topographically complex regions such as the Blue Nile Basin (Zargar et al., 2025) and known biases in the tropics (Muñoz-Sabater et al., 2021). Given that Seasonal AFNOCast and BCSD are data-driven approaches, the improved predictions inherently learn the biases present in the reference data. Consequently, forecast skill and particularly onset timing, as given by ERA5-Land, may reflect systematic uncertainties embedded in the reanalysis product. Nonetheless, ERA5-Land was chosen for three main reasons: (1) it provides a widely adopted and consistent reference; (2) it ensures methodological consistency with the BCSD baseline approach of Lorenz et al. (2021), which also relies on ERA5-Land to maintain physical coherence across multiple variables (precipitation, temperature and radiation); and (3) its finer spatial resolution provides a more suitable reference for hydrological and impact-model applications, especially in the topographically complex Upper Blue Nile region.

2.2.2 SEAS5

The initial seasonal forecast, the seasonal forecast system 5 (SEAS5) of ECMWF (Johnson et al., 2019) is used as input for the bias-correction and downscaling of precipitation. The global forecast is based on the IFS forecasting system coupled with an ocean and sea-ice model, capable of predicting climate signals such as the El-Nino-Southern-Oscillation. The variable used was total precipitation (tp, variable ID: 228) from the single-level fields, provided on a regular 0.25° grid (approx. 26 km in the region). The product provides a 215 day long global forecast, roughly a 7-month forecast, which is initialized on the first of every month and is available roughly on the fifth of each month. While the hindcasts (retrospective forecasts available from 1981, up to 2016)

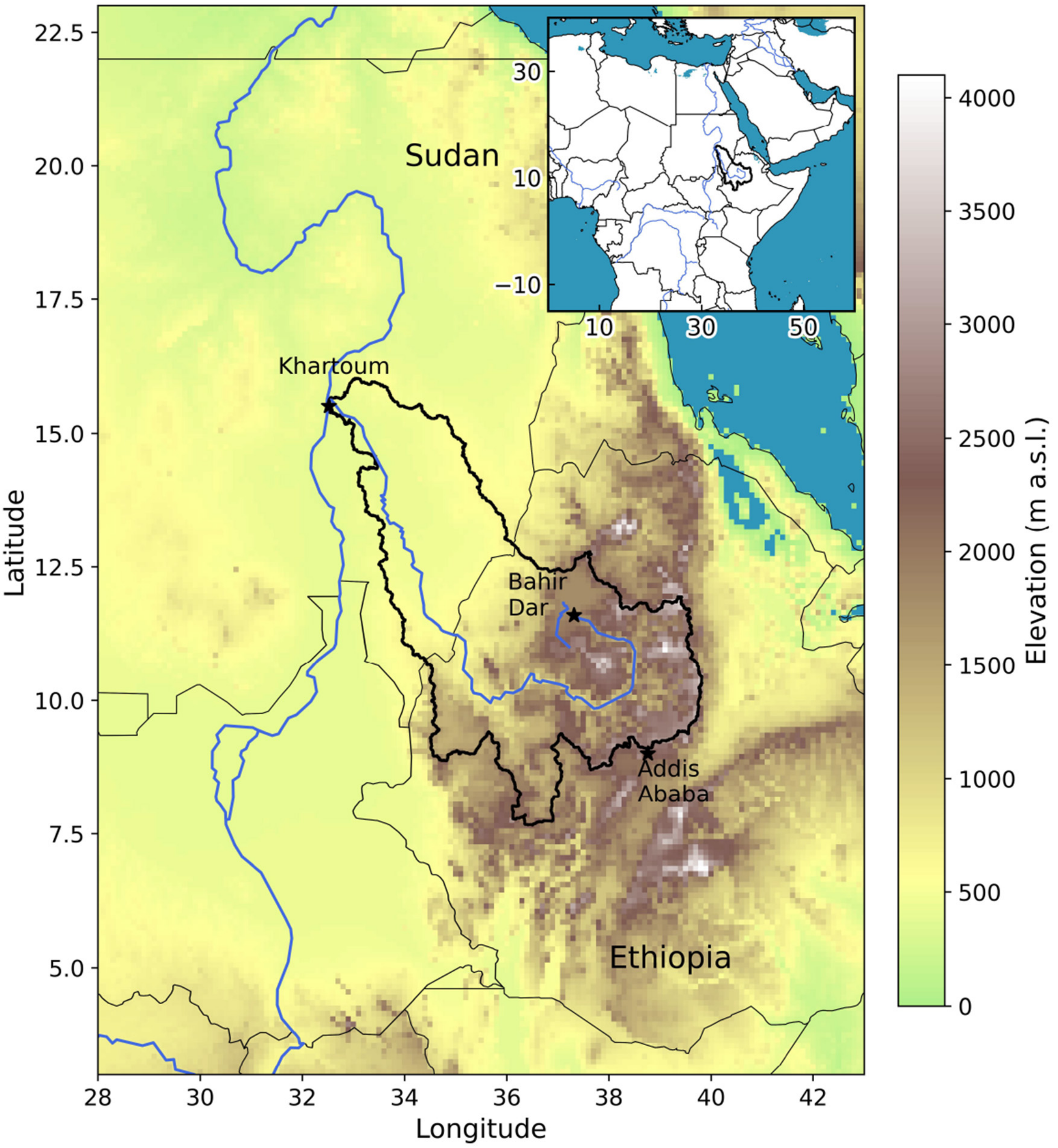

FIGURE 1
Topography and location of the transboundary Blue Nile Basin (thick black line) and the Blue Nile River (Amante and Eakins, 2009).

have an ensemble size of 25 members, the forecasts increased their ensemble size in 2017 to 51 members. Although other studies show the high potential of multi-model forecast systems over Africa (Gebrechorkos et al., 2022), this study focuses on SEAS5 as global forecast input due to its high resolution compared to other global seasonal forecasts.

SEAS5 forecasts are initialized at 00 UTC (ECMWF, 2021). Precipitation is provided as a cumulative total (running accumulation) from the model start time. Thus, the raw accumulated precipitation values do not directly represent daily totals. To derive daily precipitation, we de-accumulated the forecasts by calculating the discrete difference along the forecast step dimension. The first forecast step was retained to preserve dimensional consistency via concatenation. The resulting daily values represent the precipitation accumulated during the 24-h interval receiving the time stamp of the day in which the precipitation accumulated, consistent with the timestamp labeling of ERA5-Land.

2.2.3 Statistical and spatio-temporal characteristics of the precipitation datasets

This study uses ERA5-Land as the reference dataset and raw SEAS5 forecasts as input for the post-processing methods. The post-processing methods are calibrated (trained) using the SEAS5 hindcast with 25 members and the corresponding daily ERA5-Land data from 1981–2016. The evaluation is performed over the forecast period (2017–2023) with 51 members, at both daily and monthly resolutions.

Table 1 summarizes statistical properties of each dataset. The period 1981–2016 corresponds to the hindcast (forecasts initialized through December 2016, valid times extending to June 2017) and 2017–2023 to the forecast period (initializations through December 2023, valid times extending to June 2024). Note that the number of grid cells differs between datasets because SEAS5 includes additional dimensions (initialization date, lead time up to 215 days/7 months, and ensemble members). Furthermore, SEAS5 has a lower spatial resolution, resulting in 324 grid cells over the Blue Nile Basin, compared to 2,568 cells for the higher-resolution ERA5-Land dataset.

Overall the statistical description of the precipitation datasets SEAS5 and ERA5-Land in Table 1 shows similar mean precipitation, but the table suggests differences across statistical metrics and time scales. For the 2017–2023 period, SEAS5 is slightly wetter on average at both monthly and daily scales, although the differences are small. The monthly SEAS5 dataset exhibits slightly higher central values (50–75 percentiles) and more wet days. By contrast, ERA5-Land shows substantially higher monthly maxima and higher standard deviation, indicating that extreme wet months occur only in ERA5-Land and that month-to-month variability is greater. At the daily scale (2017–2023), both datasets have similar mean precipitation, but ERA5-Land again displays stronger extremes and higher variability, consistent with its higher daily maximum and standard deviation. Differently, the wet-day percentages are similar between the two datasets at daily resolution. For the hindcast period (1981–2016), both datasets present a lower precipitation average and slightly reduced standard deviation compared to the forecast period. ERA5-Land is wetter overall, has more wet days, and greater variability, while SEAS5 shows fewer rainy days and lower central tendency but presents a maximum value higher than ERA5-Land.

The temporal evolution of spatially averaged precipitation is shown at the monthly scale for the entire Blue Nile Basin (Figure 2). The basin is characterized by a single dominant rainy season each year, peaking during June to September (JJAS), which is evident as a recurring annual maximum in the time series. The reference curve (black solid line) in Figure 2 highlights several distinct features of the basin's precipitation regime. First, the onset timing of the rainy season varies considerably between years. For instance, 2017 displays an unusually early peak in May, indicating exceptionally high rainfall for that month. Second, the magnitude of rainfall fluctuates markedly across years, with 2023 showing substantially lower precipitation compared to previous seasons. While the raw SEAS5 forecast (green lines) captures the seasonal pattern, the skill is limited in forecasting the year-to-year differences. These interannual differences are particularly relevant for the application of these forecasts, such as water resource management and decision-making in the region.

Figure 3 further illustrates spatial precipitation differences in the Blue Nile Basin during the 2017 rainy season (JJAS) as an example. Columns 2–5 depict statistical variations across SEAS5 ensemble members for lead month 0, corresponding to the forecast initialization month. The maximum and minimum ensemble members are selected based on their spatially averaged precipitation. This spatial visualization confirms substantial ensemble variability in precipitation magnitude, with the highest variability observed in the Upper Blue Nile, as indicated by the dark red areas in the final column representing the standard deviation. Nonetheless, the spatial plots also reveal higher magnitude precipitation events present in the ERA5-Land reference that do not appear in any SEAS5 ensemble member.

Overall, this comparison demonstrates that SEAS5 exhibits greater uniformity, which limits its ability to fully capture spatiotemporal variability and extreme precipitation events observed in ERA5-Land. These systematic deficiencies highlight the need for post-processing to adjust biases and improve representation of both rainfall magnitude and variability.

2.3 Methods

Global seasonal forecasts run into issues for regional applications due to known biases, drifts and the coarse resolution that cannot account for small-scale characteristics. Therefore, post-processing methods are used to regionally enhance seasonal precipitation forecasts. In the following, the statistical method BCSD and the novel DL-based method Seasonal AFNOCast are introduced as methods to post-process SEAS5. The post-processed forecasts are evaluated in depth for the Blue Nile Basin in the 7-year evaluation period from 2017 to 2023.

2.3.1 Bias correction and spatial disaggregation

The statistical post-processing method Bias Correction and Spatial Disaggregation (BCSD) of Lorenz et al. (2021) is applied as the baseline method, as an exemplary statistical approach. The method was first developed by Wood et al. (2002) and consists of two steps: first, a spatial disaggregation utilizing a simple bilinear interpolation toward the reference resolution. Different from the initial implementation, the interpolation is applied on absolute values rather than the climatological anomalies. Second, the bias-correction utilizes empirical quantile mapping. Specifically, BCSD executes the quantile-based correction with the help of pixel-wise and daily CDFs of the reference climatology and hindcasts (1981–2016), each using a sliding ±15-day window. Lorenz et al. (2021) extends the empirical quantile mapping in BCSD by accounting for special cases such as extremes and precipitation intermittency.

For this study, BCSD is implemented in Python (Lorenz et al., 2025), with parallelization handled by Dask to optimize computational efficiency. The full processing pipeline is performed at daily temporal resolution and includes downloading the global seasonal forecasts (covering 215 days), converting them into NetCDF format, spatially truncating them to a bounding box around the Blue Nile Basin, and remapping them to the target resolution. Before bias-correction, the forecasts are rechunked and aligned with the corresponding CDFs constructed from reference and hindcast data. The bias-correction step itself is executed in

TABLE 1 Summary statistics of precipitation datasets for different temporal resolutions (mo., monthly; d., daily).

Dataset	Temp. resolution	Time period	Mean	Std	Min	25%	50%	75%	Max	Wet days (%)
SEAS5	mo.	2017–	3.3	4.45	0	0.02	0.99	5.60	30.93	50
ERA5-Land		2023	53.21	5.27	0	0.03	0.79	4.22	91.94	47
SEAS5	d.	2017–	3.37	6.64	0	0.00	0.00	3.51	355.15	34
ERA5-Land		2023	3.24	7.85	0	0.00	0.08	2.82	381.99	33
SEAS5	d.	1981–	2.47	5.60	0	0.00	0.00	1.59	389.32	28
ERA5-Land		2016	2.67	6.40	0	0.00	0.07	2.30	337.45	32

Unit: mm/day.

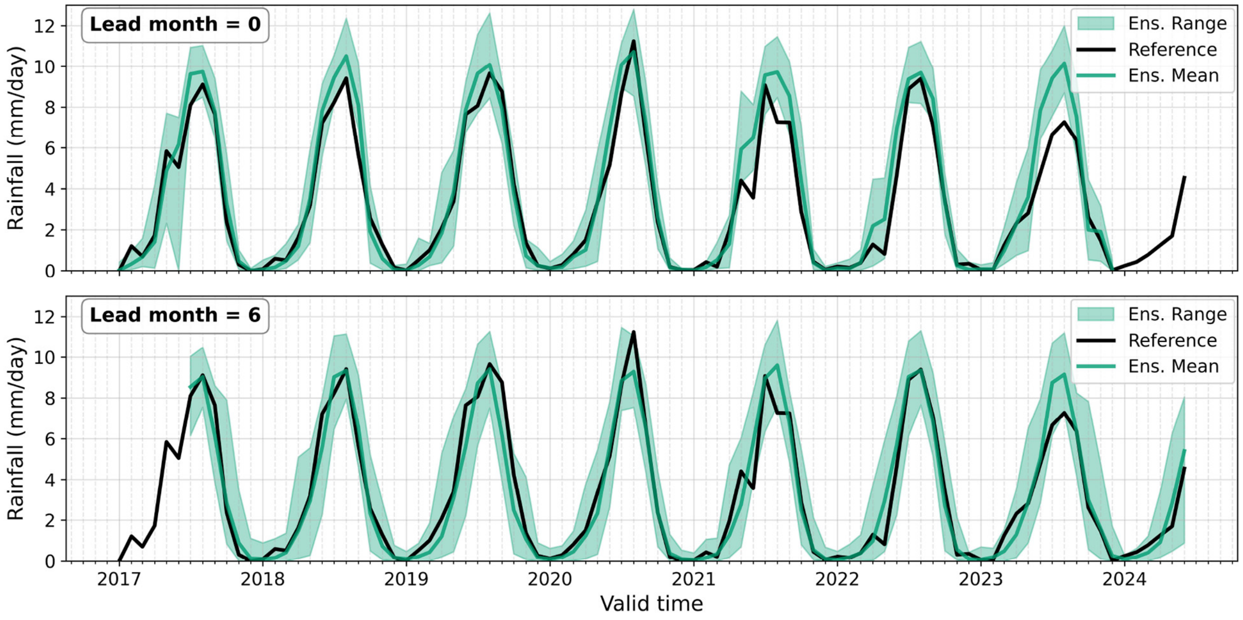

FIGURE 2
Monthly basin-average precipitation from SEAS5 (ensemble mean and range) compared with ERA5-Land for 2017–2023. ERA5-Land is shown as a black line, while the SEAS5 ensemble mean and range are shown by the solid green line and green shading, respectively. The **top** row shows forecasts initialized at the beginning of the valid month (lead time: 0 months), and the **bottom** row shows forecasts initialized 7 months earlier (lead time: 6 months), resulting in a slight temporal shift in the plotted forecast time series.

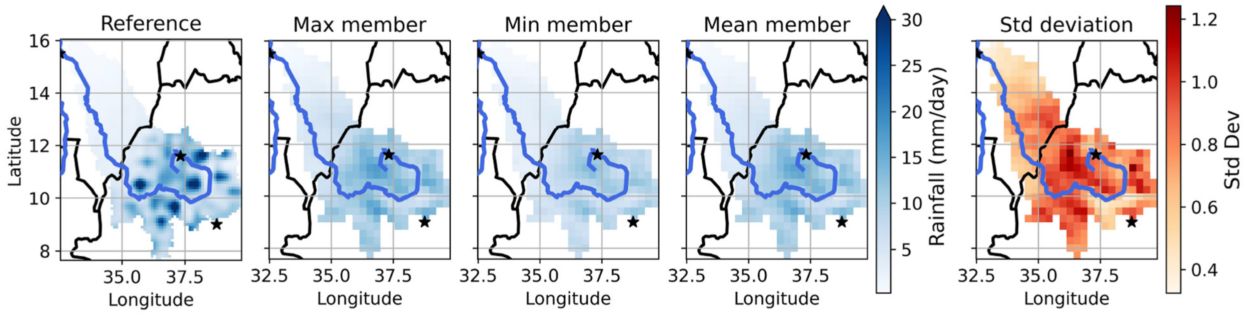

FIGURE 3
Spatial comparison of SEAS5 ensemble precipitation and ERA5-Land over the Blue Nile Basin (JJAS 2017). Column 0 shows the reference, ERA5-Land. For lead month 0, columns 1–4 illustrate SEAS5 ensemble characteristics, including the members with the highest (Max) and lowest (Min) basin-mean precipitation and the member average (Mean). The final column displays ensemble standard deviation.

parallel and requires approximately 30 min using 30 Dask workers on a CPU node.

2.3.2 Seasonal AFNOCast

2.3.2.1 Model architecture

To post–process SEAS5 precipitation forecasts, we developed Seasonal AFNOCast, a custom encoder–decoder convolutional neural network which at its core applies FNO, explicitly the Adapted Fourier Neural Operator (AFNO) (Guibas et al., 2022; Pathak et al., 2022). The architecture evolved from an initial U-Net-like (Ronneberger et al., 2015) CNN design, which, however, exhibited excessive spatial smoothing. Unlike standard CNNs, AFNO works in spectral space, which facilitates learning along different frequencies and thereby captures both short' and long–range dependencies. Pathak et al. (2022) compares AFNO to a global convolution, however, implemented in an efficient way. Including an AFNO block effectively improved sharper precipitation patterns. In this study we extend AFNO to three spatiotemporal dimensions (time window × latitude × longitude), allowing the network to learn frequency interactions simultaneously along time and space.

Figure 4 illustrates the architecture. Two 3–D convolutional "encoder" blocks (kernel = 3 × 3 × 3, ReLU) first raise the feature depth while a bicubic interpolation step upsamples the spatial resolution. A pointwise 1 × 1 × 1 convolution finalizes each block. The resulting feature cube enters a single AFNO3D block, implemented after Guibas et al. (2022) but modified to include the temporal dimension in the Fourier transform. The block inherits transformer-style components such as layer normalization, positional embedding and feed-forward MLP, yet, replaces the conventional attention mechanism with the AFNO. A 3–D dropout layer ($p = 0.1$) follows. Two symmetric "decoder" conv–blocks then reconstruct the output field, with skip connections from the raw input and the preAFNO state (see dashdotted lines in Figure 4) to mitigate information loss, consistent with U-Net–style architectures (Ronneberger et al., 2015). In our experiments, we observed these connections to be essential for preserving member diversity and preventing collapse toward overly similar predictions across the ensemble. The final convolution collapses the time dimension (stride = 4 × 1 × 1) and reduces the channel depth to one, yielding an output tensor of shape (ensemble member, latitude, and longitude). A ReLU activation ensures nonnegative precipitation values. The entire architecture is implemented as a vectorized function using torch.vmap, enabling efficient parallel processing of each ensemble member separately with shared weights. This design allows the model to retain the individual identity of each ensemble member while significantly improving computational efficiency. Importantly, it propagates the intrinsic SEAS5 uncertainty rather than producing new uncertainty in the model. To support this, the ensemble standard deviation (Figure 4) is passed as a second input to each member branch, providing shared uncertainty information across all members without compromising their individual correction.

2.3.2.2 Model training

The model was trained on a H100 Tensor Core GPU with the Adam optimizer, a learning rate of 0.0003, and a batch size of 4. This hyperparameter configuration was selected to balance the constraints of limited GPU memory while achieving the lowest loss. The data set was divided into three subsets: training (1981–2010, 74,820 samples), validation (2011–2016, 18,060 samples), and testing (2017–2023, 15,480 samples), with the testing subset held out of training and used as unseen data for evaluation.

We highlight the loss function used for training. The ensemble output for each day requires a specific loss function to enable training with reanalysis data that provide only one observation per timestep. The network is therefore trained with the continuous ranked probability score (CRPS). This forces the network to optimize the learnable parameters toward a calibrated ensemble with sharper individual ensemble member predictions compared to more common mean squared error (MSE) loss function, with which the model tends to learn the ensemble mean.

As illustrated in Figure 4, one input sample corresponds to a single forecast timestep to be corrected. Each sample comprises the spatial domain, a temporal window of ±5 days centered on the day being corrected (i.e., 11-day window), and the full ensemble of 25 (increasing to 51 from 2017 onward) members $\{y_0, \ldots, y_e\}$. Consequently, the input tensor has dimensions (1, member: 25/51, time: 11, latitude: 38, longitude: 28), and the output tensor has dimensions (1, member: 25/51, time: 1, latitude: 106, longitude: 78). Note that the full ensemble (i.e., 25 members in years 1981–2016 and 51 members in years 2017-2023) input is passed to the model (Figure 4a), however, the architecture processes each ensemble member independently and in parallel (Figure 4b). The corrected members are then concatenated to reconstruct the ensemble (Figure 4c), which is used to compute the CRPS loss with the target field for the corresponding timestep. The target corresponds to the verifying reanalysis, with dimensions (1, member: 1, time: 1, latitude: 106, longitude: 78). The time of the target and prediction coincide with the input timestep, such that the model corrects precipitation for that same day.

Finally, note that training took place separately for each target month, hence, one model was trained per month. Monthly models enable training and learning toward seasonally different rainfall patterns and distributions rather than being hindered by conflicting rainfall characteristics.

2.3.2.3 Model inference

The prediction pipeline is implemented in Python, with data preparation steps comparable to those of the BCSD method. Forecast input data are first downloaded and then spatially truncated to the Blue Nile Basin, remapped, and formatted into structured data arrays with dimensions (samples, ensemble member, time window, latitude, and longitude). Prediction using Seasonal AFNOCast is executed on a GPU and requires approximately 1.75 min per ensemble forecast. When executed on CPU hardware, the same process takes approximately 7 min. These timings include only the inference step, assuming all necessary input data have been preprocessed and prepared in advance.

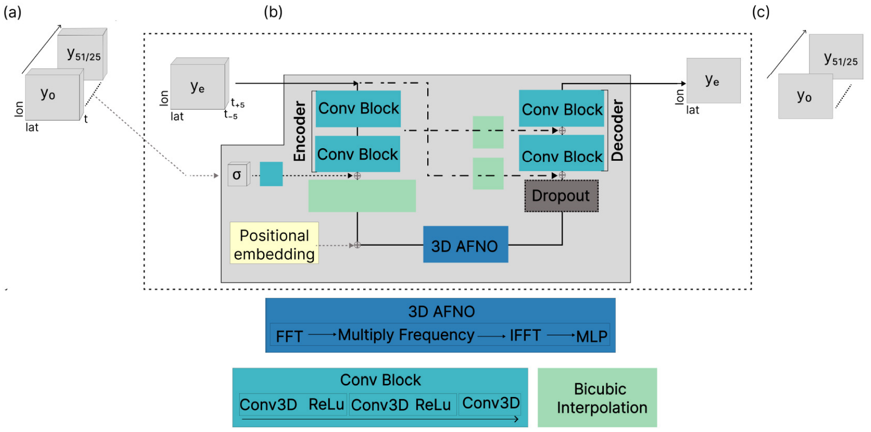

FIGURE 4
Schemativ overview of the Seasonal AFNOCast set-up: **(a)** Ensemble input, consisting of pixel-wise standard deviation and single-member (y_e) inputs in 3-dimensions, **(b)** the architecture of Seasonal AFNOCast, and **(c)** the predicted output, where corrected members (y_e) are concatenated into an ensemble.

2.3.3 Evaluation metrics

As Murphy (1993) already stated, the forecast quality cannot be captured by a single verification metric, but a set of metrics are required to capture the multifaceted forecasts. Deterministic and ensemble metrics were used to quantify forecast skill, spectral metrics to evaluate the spatial signal, and binary scores to assess event occurrence and likelihood, particularly for events relevant to downstream tasks and applications.

The analyses were conducted at both daily and monthly temporal resolutions. Monthly values were computed as averages of daily precipitation over each calendar month. Because SEAS5 forecasts are initialized once per month and extend up to seven lead months, careful alignment between forecast lead times and valid periods is required to ensure consistent evaluation against the ERA5-Land reference. Table 2 summarizes this alignment, showing how valid months (or days) correspond to lead months (or days) for two example forecasts. Each forecast provides 215 lead days, corresponding to a complete forecast horizon of seven months (lead months 0–6). In some cases, however, the final few forecast days extend into an additional, incomplete eighth lead month. These partial months are excluded from the monthly analysis to maintain consistent temporal coverage.

The probabilistic skill scores are evaluated relative to a climatological reference. In this study, the climatology is defined as monthly precipitation averages computed over the 1981–2016 period.

2.3.3.1 Radially averaged power spectrum density

Evaluating the spatial characteristics of rainfall fields is essential, particularly for DL models that often tend to smooth fine-scale features (Tan et al., 2024). The Power Spectral Density (PSD) describes how signal power is distributed across spatial frequencies, providing insight into the representation of variability at different scales. The analysis is based on the Fourier transform of precipitation fields, decomposing the spatial signal into frequency components. The Radially Averaged Power Spectral Density (RAPSD) quantifies the average spectral power as a function of wavelength, radially averaged over the spatial domain.

Following recent studies that apply RAPSD for DL-based precipitation evaluation (Hess et al., 2023; Glawion et al., 2023), we use this approach to compare the spatial pattern fidelity of both the Seasonal AFNOCast forecasts and the baseline method BCSD to the reference dataset ERA5-Land. A key advantage of the RAPSD is its ability to evaluate how well patterns are preserved across different scales (Harris et al., 2001). In log-log space, the spectral slope further indicates the smoothness of the precipitation fields. In this study, RAPSD is computed for each ensemble member and valid timestep (monthly scale) using the implementation provided in the pysteps library (Pulkkinen et al., 2019). Ideally, the RAPSD curve of the post-processed forecasts should closely follow that of ERA5-Land, indicating an accurate spatial reproduction of the underlying precipitation structures across spatial scales.

2.3.3.2 Forecast evaluation: error metrics, ensemble skill scores and categorical metrics

This section introduces the evaluation framework used to assess forecast performance at a monthly temporal resolution. The forecast accuracy between the reference and the raw or post-processed forecast and skill is quantified via deterministic and probabilistic metrics. We compute the Root Mean Squared Error (RMSE), based on the Mean Squared Error (MSE), on a pixel-wise basis. For each grid point (i, j), the forecast value of the ensemble member m, initialization date s, and lead time l is compared to the corresponding reference at the valid time $t = s + l$.

Let $f_{i,j,m,s,l}$ denote the forecast and $y_{i,j,t}$ the corresponding reference value at valid time t, so that the number of samples $n = i * j * s * l$.

$$\mathrm{MSE}_m = \frac{1}{n}\sum^{n}\left(f_{i,j,m,s,l} - y_{i,j,t}\right)^2, \qquad \mathrm{MSE}_m \geq 0, \tag{1}$$

$$\mathrm{RMSE}_m = \sqrt{\mathrm{MSE}_m}, \qquad \mathrm{RMSE}_m \geq 0, \tag{2}$$

TABLE 2 Illustration of monthly averaging and alignment of forecast lead months and valid months exemplary for January and February 2017 initializations.

Valid month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Day count	31	28	31	30	31	30	31	31
Date range	1–31 Jan	1–28 Feb	1–31 Mar	1–30 Apr	1–31 May	1–30 Jun	1–31 Jul	1–31 Aug
Init Jan 2017								
Lead month	0	1	2	3	4	5	6	-
Lead day	0–30	31–58	59–89	90–119	120–150	151–180	181–211	212–215
Init Feb 2017								
Lead month		0	1	2	3	4	5	6
Lead day		0–27	28–58	59–88	89–119	120–149	150–180	181–211

The results of the error metrics per member are summarized using descriptive statistics. We compute the standard deviation, minimum, 25% quantile, median, mean, 75% quantile, maximum and the average of the error over the ensemble members m.

The ensemble forecast skill is determined with the Continuous Ranked Probability Score (CRPS). The CRPS allows one to compare an ensemble forecast with an observation, without limitations to the choice and number of classes (Hersbach, 2000). It describes the difference between the cumulative distribution function of the ensemble and the Heaviside step function of the observation, written as

$$\mathrm{CRPS}(F,x) = \int_{-\infty}^{\infty} \left(F(x) - H(x - x_0)\right)^2 dx, \qquad \mathrm{CRPS}(F,x) \geq 0, \tag{3}$$

where $F(x)$ is the cumulative distribution function (CDF) of the forecast, which in this study is approximated by the empirical CDF. The observed value is x with the Heaviside step function $H(x - x_0)$, defined as

$$H(x - x_0) = \begin{cases} 0, & x < x_0 \\ 1, & x \geq x_0. \end{cases} \tag{4}$$

The Continuous Ranked Probability Skill Score (CRPSS) is given by

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_f}{\mathrm{CRPS}_r}, \qquad \mathrm{CRPSS} \in (-\infty, 1], \tag{5}$$

where $CRPS_f$ is the forecast CRPS and $CRPS_r$ is the CRPS of a reference forecast (e.g., climatology or raw forecast). If $CRPS_r$ refers to the climatology, note, the CDF $F(x)$ used in the CRPS calculation is based on the distribution of historical years, while for an ensemble reference, $F(x)$ is based on the ensemble member distribution. The skill score allows to interpret the skill of the post-processed forecasts in comparison to the climatology or the raw forecast. Here, a value above 0 describes an improvement compared to the reference product (raw SEAS5 or climatology). While 1 would indicate a perfect forecast, well-calibrated ensemble forecasts that preserve a realistic uncertainty spread typically attain maximum CRPSS values of around 0.7–0.8. A negative value describes that the reference product (raw SEAS5 or climatology) has a lower CRPS than the post-processed forecast (BCSD or AFNOCast). Arid areas, grid points with monthly precipitation <0.1 mm are excluded in this calculation. The CRPSS is calculated for monthly averaged predictions and pixel-wise to account for the high differences in the region.

We provide descriptive statistics (minimum, 25% quantile, median, mean, 72% quantile, and maximum) of the probabilistic skill scores (CRPS and CRPSS as boxplots) across time and space. Differently, the spatial maps present the CRPSS computed based on temporally averaged CRPS values.

Finally, we evaluate the categorical forecast performance using the Receiving Operating Characteristic (ROC, Mason, 1982; Fawcett, 2006). Pixel-wise tercile boundaries are defined based on the climatological reference period (1981–2016) to distinguish three classes: Below Normal, Near Normal, and Above Normal. The ROC curve is computed seperately for each tercile class by transforming the reference into binary events (True/False). We determine the correct classifications (True Positives, TP; True Negatives, TN), and incorrect classifications (False Positives, FP; False Negatives, FN) for each possible probability threshold, derived from the fraction of ensemble members falling into the respective class. The ROC curve represents the relationship between the False Positive Rate (FPR) and the True Positive Rate (TPR), each defined as

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{TN + FP}. \tag{6}$$

A perfect classifier corresponds to the point (0,1) in ROC space, whereas random guessing is represented by the diagonal [(0,0), (1,1)]. This interpretation is summarized by the area under the ROC curve (AUC), where random guessing yields an AUC of 0.5, while a perfect classifier yields an AUC of 1. In the context of meteorological forecasting, an AUC $>$ 0.5 indicates positive discrimination skill for the respective tercile class, whereas AUC $\leq$ 0.5 is considered to show no discrimination skill.

2.3.3.3 Brier Score for extreme event detection

We evaluate the skill of probabilistic forecasts in detecting extreme precipitation events at daily temporal resolution using the Brier Score (BS), which quantifies the mean squared difference between forecast probabilities and binary reference outcomes (Brier, 1950). The score is negatively oriented, meaning that lower values indicate better forecast performance, with a perfect score being 0. Seasonal forecasts are not expected to precisely predict the timing and location of an extreme event several months in advance. Instead, we assess their ability to correctly estimate the probability of an extreme event occurring within a specified temporal window across the entire Blue Nile Basin, reflecting the utility of the forecasts for early warning and preparedness applications. An extreme event is defined based on precipitation thresholds derived from the reference dataset (ERA5-Land): The 99th percentile corresponds to a threshold of 40 mm/day, used to define extreme events. The 99.9th percentile, corresponding to 120 mm/day, is used to identify very extreme events. To qualify as an event, a minimum of two consecutive days of spatially overlapping pixels exceeding the defined threshold. A forecast hit is defined when the predicted event occurs within a ±3-day window relative to the start date of the reference extreme event. This temporal tolerance accounts for uncertainty in lead times at seasonal scales. Forecast probabilities are calculated as the fraction of ensemble members that register a hit for a given event. The BS is then computed for each event, observed in the reference or predicted by the post-processed forecasts, and subsequently averaged annually and across the full evaluation period.

2.3.3.4 Onset prediction evaluation using the Brier Score

To complement the evaluation of extreme event prediction, we also assess the forecast skill in predicting the onset of the rainy season at daily temporal resolution, using the BS as the main verification metric. The onset is defined following the criteria of Scheuerer et al. (2024): it is the first occurrence of three consecutive wet days (each with $\geq$1 mm/day) that accumulately exceed a total of 20 mm. To avoid false onset signals caused by early isolated rainfall events, the definition includes an additional constraint as no dry spell should follow within the subsequent 21 days. A dry spell is defined as any sequence of seven or more consecutive days with precipitation <1 mm/day. The onset definition stays a widely differently defined subject (Lala et al., 2020), however, we stick to the onset definition of Scheuerer et al. (2024) given its strong outreach in Eastern Africa through the IGAD Climate Predictions and Applications Centre (ICPAC). The forecasts used to determine the onset date were initialized in April of the respective year. For the forecasts, onset timing is determined individually for each ensemble member, and the spread of predicted onset dates is summarized using boxplots to visualize ensemble dispersion. Forecast skill is then quantified using the BS, calculated under varying temporal tolerances for what constitutes a correct onset prediction i.e., a hit. Specifically, we evaluate hits based on forecasted onset dates falling within ±5, ±10, ±15, and ±20 days of the observed onset date in the reference dataset (ERA5-Land). For each spatial pixel, the forecast probability is computed as the fraction of ensemble members predicting onset within the specified tolerance window. The BS is calculated pixel-wise so that single pixel, e.g., of Bahir Dar, can be evaluated as well as the subsequently spatially averaged score to obtain an overall assessment of onset prediction skill across the basin. In general, forecasts with a BS $\leq$ 0.25 have an accuracy rate higher than 50%, hence show a higher skill than climatology (Coelho et al., 2006; Rodwell and Doblas-Reyes, 2006; da Rocha Júnior et al., 2021).

3 Results

3.1 Physical consistency and spatial patterns

3.1.1 Qualitative analysis of the daily member prediction

A qualitative analysis of the precipitation fields at member and daily level offers insights into the pixel-level corrections introduced by each post-processing method. In Figure 5, the first column shows that both methods produce ensemble mean patterns that closely resemble climatology, suggesting that the ensemble mean of post-processing tends to converge toward climatological signals. However, a closer inspection of individual ensemble members reveals that both methods are capable of capturing characteristic spatial features of the rainy season as seen in the reference dataset (ERA5-Land). Notably, the northwest to southeast precipitation gradient, from Khartoum toward the Ethiopian Highlands, is well reproduced in both the ensemble means and individual members. Furthermore, localized rainfall variability, including rainfall peaks over the Ethiopian Highlands in the reference, as well as dry regions (precipitation <0.1mm/day, displayed in white), are apparent in individual members and preserved by both methods. The main distinction between the two methods lies in how they perform the correction. BCSD, based on quantile mapping, applies a rigid correction, i.e. the same input value at a specific pixel and timestep will always yield the same output. In contrast, Seasonal AFNOCast demonstrates greater flexibility, potentially learning from spatial and temporal context. An example is illustrated in Member 40 of Figure 5, where AFNOCast reconstructs a rainfall structure absent in the original forecast but present in the reference.

3.1.2 Radially averaged power spectrum density

Deep learning models often struggle to reproduce realistic spatial structures, particularly fine-scale features that are frequently smoothed by convolutional architectures. The Radially Averaged Power Spectral Density (RAPSD) assesses how well post-processed forecasts reproduce precipitation structures across spatial scales, with closer agreement to ERA5-Land indicating improved spatial representation. Figure 6 shows that all methods exhibit similar spectral shapes across wavelengths and seasons, but Seasonal AFNOCast aligns more closely with the reference, as highlighted by marginally smaller absolute deviations shown in the second row, where zero indicates perfect agreement. Quantitatively, AFNOCast deviates from ERA5-Land by 10 and 7 dB × $\Delta(\log_{10}$ wavelength) in MAM and JJAS, respectively, compared to larger deviations for BCSD [15 and 10 dB × Δ ($\log_{10}$ wavelength)]. Because the spectra are shown on a logarithmic scale, visually similar deviations correspond to larger absolute power differences at longer wavelengths. We suspect that the anomalous bump in the reference RAPSD between 144 and 72 km is related to structural characteristics of the ERA5-Land dataset, possibly likely due to storing or processing transformation applying spectral methods.

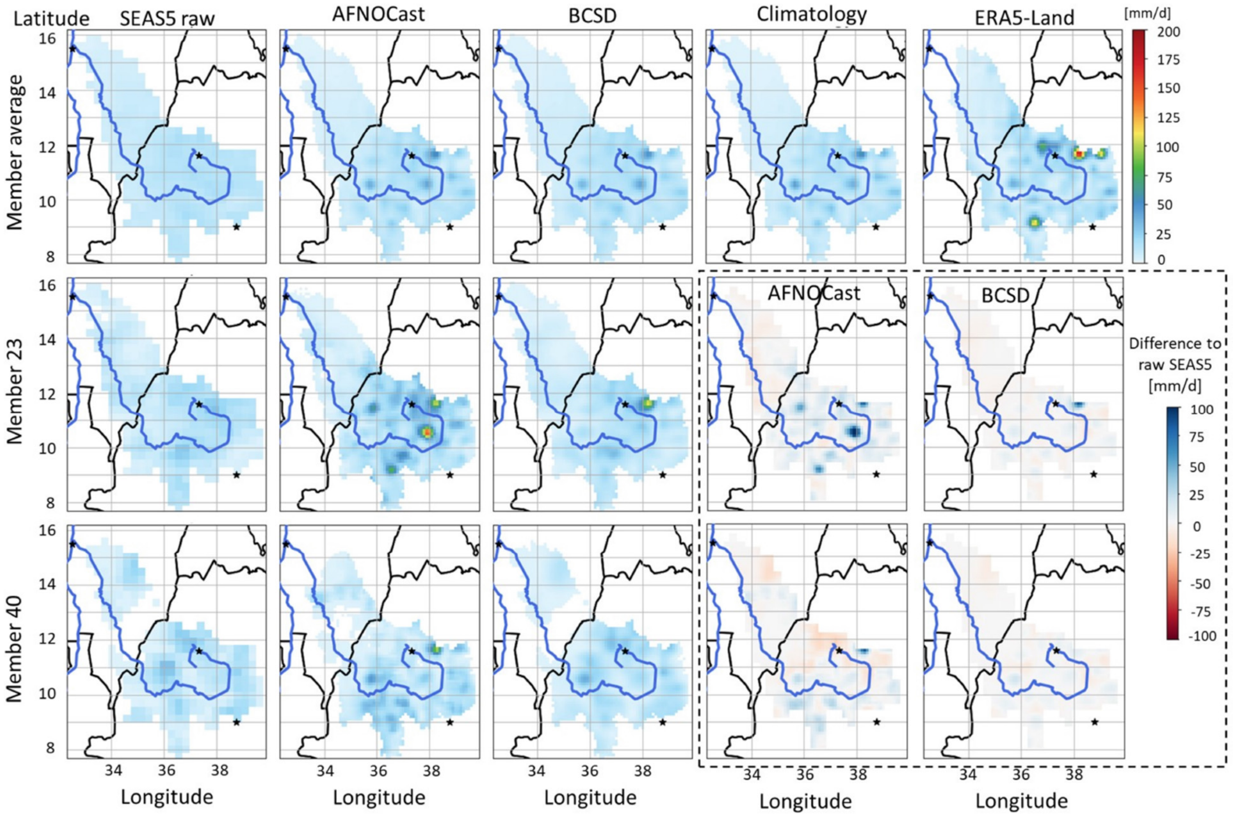

FIGURE 5
Visual and qualitative comparison of the ensemble mean (first row) with a selected single-member correction for an exemplary case (3 August 2018, initialized in May 2018, lead day 94). The panels enclosed by the dashed box show the deviation of each post-processing method from the raw SEAS5 forecast for the corresponding member.

3.1.3 Bias-corrected distributions at daily and monthly scale

Matching the precipitation distribution is an explicit objective of both post-processing approaches, particularly for BCSD which applies quantile mapping. For the DL approach AFNOCast, which is not explicitly constrained to quantile mapping, reproducing the reference distribution serves as a sanity check. Thus, the precipitation distribution are evaluated using histograms, with the goal of matching (i.e., closely aligned histograms) the reference distribution of ERA5-Land in the independent verification period from 2017 onwards, which was not used for training or calibration. To ensure a fair comparison, the reference data are aligned with the forecast initialization dates and their lead time horizon, such that the number of samples is identical across forecasts and reference. This alignment results in a repetition of ERA5-Land values across the seven forecast lead months, which is visible in the histogram curve as a slightly elevated tail (Figure 7).

As expected, the raw SEAS5 forecasts substantially underestimate high-intensity rainfall, with a maximum near 30 mm/day, reflecting the limitations imposed by the model's coarse spatial resolution (Figure 7). This bias is effectively corrected by both post-processing methods. All ensemble members of both post-processed forecasts, BCSD (orange) and AFNOCast (blue), closely follow the reference distribution (black), including for high-intensity rainfall events. A minor deviation is observed in AFNOCast, which slightly overestimates the frequency of precipitation values in the 15–25 mm/day range. Nevertheless, both methods improve the raw forecasts and confirm effective correction of the daily precipitation distribution, with consistent alignment at the monthly scale.

A similar pattern is observed at the daily scale (see Supplementary material). The raw SEAS5 forecasts significantly underestimate high precipitation intensities (≥ 40 mm) in both MAM and JJAS. Both post-processing methods correct this bias and closely reproduce the reference distribution across seasons and lead time. While small seasonal differences between BCSD and AFNOCast are apparent, with BCSD being closer to the reference distribution during JJAS and AFNOCast during MAM, the figure shows that both methods substantially improve the biased daily precipitation distribution of the raw forecasts toward the reference as intended.

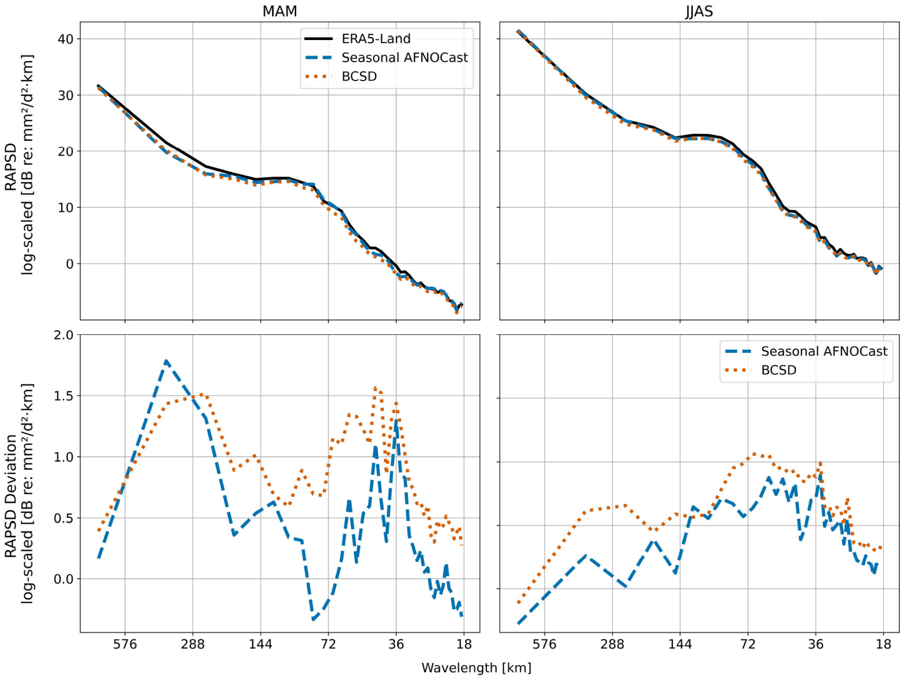

FIGURE 6
Logarithmic RAPSD for MAM and JJAS averaged over the evaluation period. The first row shows the ERA5-Land and both post-processing methods Seasonal AFNOCast and BCSD. The second row shows the deviation of the post-processing methods to ERA5-Land.

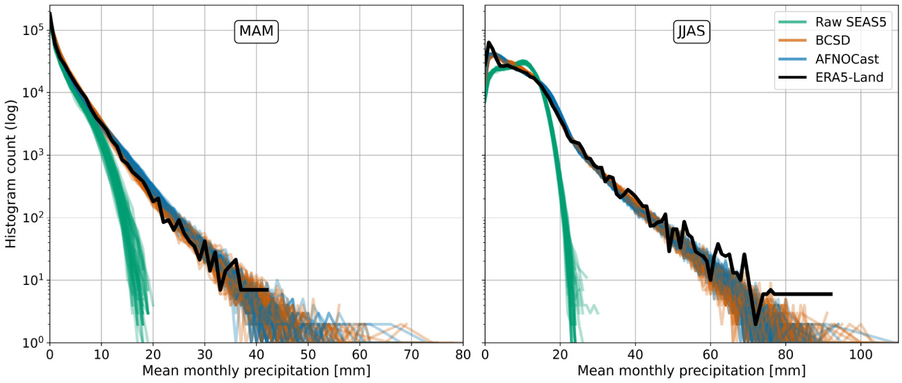

FIGURE 7
Histogram for the raw forecast and the post-processed forecasts compared to the reference for two seasons based on monthly averaged rainfall forecasts.

3.2 Prediction skill of post-processed seasonal forecasts

3.2.1 Overall prediction skill and lead time dependency

This subsection summarizes the overall predictive skill of the post-processed seasonal precipitation forecasts for monthly averaged precipitation and its dependence on lead time. Performance is assessed using deterministic error metrics based on the deviation of the ensemble mean from the reference (RMSE), as well as probabilistic error metrics (CRPS) that evaluate the full ensemble distribution. To quantify relative improvements, we further report the Continuous Ranked Probability Skill Score (CRPSS), which compares the ensemble performance to both the raw SEAS5 forecast and climatology. Positive CRPSS values indicate improvement over the respective baseline, zero or negative values denote no added skill to reduced skill, and the theoretical upper limit is 1. Because climatology is a strong baseline on seasonal timescales, CRPSS values relative to climatology are expected to be smaller than those relative to raw SEAS5. Accordingly, even modest positive values indicate relevant improvements compared to climatology. Finally, we report the categorical forecast skill using ROC curves and the area under the curve (AUC) for tercile

TABLE 3 Descriptive statistics for RMSE over the ensemble (51 members) of monthly averaged precipitation (mm/day).

Metric	Dataset	Mean	Std	Min	25%	50%	75%	Max
RMSE	SEAS5	3.68	0.03	3.63	3.66	3.67	3.69	3.78
	BCSD	2.44	0.05	2.34	2.42	2.44	2.46	2.59
	AFNOCast	**2.36**	0.03	2.31	2.34	2.36	2.38	2.49

The average best-performing product (i.e., lowest mean RMSE) is indicated in bold.

TABLE 4 Statistical description of the probabilistic error metric CRPS for raw SEAS5, post-processed forecasts, and climatology (clim) across space (grid pixels) and time (initializations and lead time).

Metric	Dataset	Mean	Std	Min	25%	50%	75%	Max
CRPS	SEAS5	1.92	3.39	0.00	0.38	0.88	2.27	76.83
	BCSD	1.02	1.35	0.00	0.30	0.60	1.25	43.18
	AFNOCast	1.12	1.53	0.00	0.29	0.62	1.38	46.31
	Clim	0.74	1.18	0.00	0.07	0.38	0.92	36.72

precipitation categories. These metrics evaluate the ability of the forecasts to discriminate between below normal (BN), near normal (NN), and above normal (AN) precipitation conditions, thereby providing insight into the presence of a predictable climate signal in the forecasts.

The results of the deterministic analysis in Table 3 show that both post-processing methods reduce the overall mean RMSE of the raw forecast by more than 30%, but Seasonal AFNOCast is slightly closer to the reference with a RMSE of 2.36 mm/day compared to BCSD with a RMSE of 2.44 mm/day. The standard deviation (Std), lowest for Seasonal AFNOCast and highest for BCSD, indicates a higher variation in BCSDs ensemble.

The probabilistic performance is assessed using the Continuous Ranked Probability Score (CRPS; Table 4), which measures the integrated difference between the forecast ensemble cumulative distribution function and the reference, confirms both post-processed forecasts (BCSD: 1.02; AFNOCast: 1.12) outperform the raw SEAS5 forecast (1.92) by a clear margin, indicating that post-processing significantly improves probabilistic accuracy after post-processing. However, BCSD slightly outperforms AFNOCast across all descriptive statistics (lower mean, median, and 75% percentile). This suggests BCSD may reduce spread and bias more effectively for precipitation in this region, though both methods provide meaningful improvement over SEAS5. While BCSD performs better on average, AFNOCast achieves a lower 25% percentile suggesting that it performs particularly well under favorable conditions. Even after correction, the scores remain comparable to or slightly above climatology (lowest mean CRPS 0.74), indicating that while post-processing effectively reduces systematic errors, inherent predictability limitations remain when averaging over time and space.

The Continuous Ranked Probability Skill Score (CRPSS), contrary to the ensemble error metric CRPS, enables a direct comparison of the post-processed seasonal forecasts toward either raw SEAS5 forecast or the climatology and presents the relative improvement. Overall, across all initializations, lead times and spatial grid cells, both post-processing methods show a clear improvement (CRPSS >0) over the raw forecasts. The median CRPSS across all grid cells is approximately 0.3 for both methods (BCSD: 0.28, AFNOCast: 0.27), indicating a substantial increase in probabilistic forecast skill (Figure 8A). However, a more detailed examination of skill along the lead time horizon reveals notable differences between the methods. At lead month 0, corresponding to the month in which the forecast is initialized and issued, BCSD outperforms AFNOCast, with a median CRPSS close to 0.5, compared to approximately 0.4 for AFNOCast. From lead month 3 onward, the trend reverses: AFNOCast demonstrates a slightly superior skill, with improvements relative to BCSD at longer lead times.

For the application of post-processed seasonal forecasts, performance relative to climatology is particularly relevant, shown with the CRPSS in Figure 8B. The raw SEAS5 forecast exhibits a negative median CRPSS, indicating that climatology outperforms the uncorrected model forecast. While both methods confirm an improvement over the raw SEAS5 forecast, this comparison highlights the greater difficulty of outperforming climatology. The median CRPSS for both post-processing methods are only slightly above zero (BCSD: 0.04, AFNOCast:0.02), indicating limited additional skill. Nevertheless, BCSD shows a median CRPSS of about 0.2 at lead month 0, suggesting meaningful short-term improvements. AFNOCast achieves a median CRPSS near 0.1 for longer lead times, demonstrating skill even when compared to climatology. Note, Figure 8 reveals that the improvement is not uniformly distributed, i.e. a substantial proportion of grid points perform worse than climatology, emphasizing the persistence of spatial and temporal variability in forecast skill despite post-processing.

The spatial and temporal variability of the skill improvements is further examined through additional Figures 9, 10 that illustrate the spatial distribution of the Continuous Ranked Probability Skill Score (CRPSS) relative to climatology, shown for each valid month and all lead months. Note that the maps represent an average across years in the evaluation periods (2017–2023).

Overall, BCSD (Figure 9) shows a clear and consistent improvement over climatology, particularly at lead month 0. The strongest enhancement occurs for valid month March and May, when almost all grid points within the Blue Nile Basin exhibit CRPSS values around 0.4. This positive trend persists

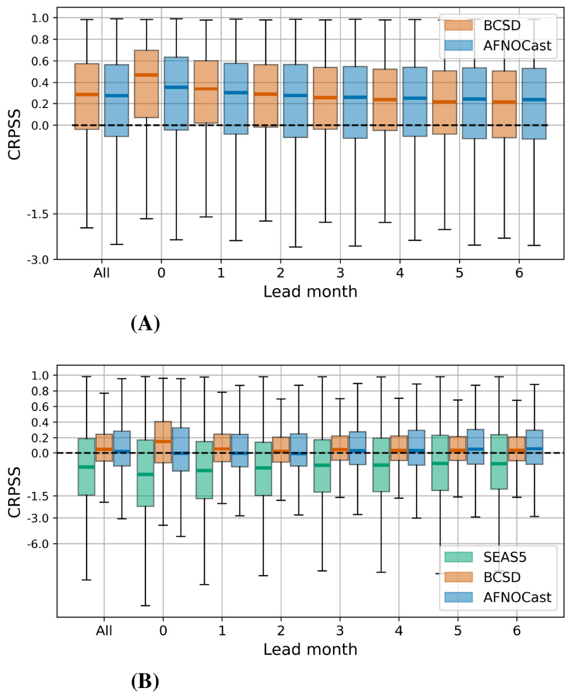

FIGURE 8
The overall Continuous Ranked Probability Skill Score (CRPSS). The black dashed line represents the skill as good as the reference (raw SEAS5/clim), values >0 indicate an improvement relative to the reference. **(A)** CRPSS relative to raw SEAS5. **(B)** CRPSS relative to climatology.

through lead month 6. Similarly, the improvement during March is widespread and sustained across multiple lead times. AFNOCast also demonstrates notable improvements relative to climatology, though its spatial and temporal pattern differs (Figure 10) . Its skill is most evident at longer lead times, confirming the tendencies discussed in the preceding analysis. The months of March–May (MAM) show particularly strong performance: in March, skill is concentrated over Ethiopia, whereas in April and May, improvements shift northward into Sudan, with the exception of lead month 0, where improvements are again strongest over Ethiopia. Among the rainy season months (JJAS), July shows the weakest performance relative to climatology, with only limited areas of improvement visible for both post-processing methods. However, June and August display large areas of improved CRPSS. AFNOCast especially enhances the forecast in September where CRPSS values up to 0.4 are present, even at longer lead times.

In summary, the pronounced improvements in the period March-May (MAM) for both BCSD and AFNOCast indicate that precipitation variability during these transitional months is not well captured by climatology but can be better represented by the post-processed forecasts. During the peak rainy season (JJAS), it becomes more difficult to outperform climatology due to the relatively homogeneous rainfall regime. Nevertheless, BCSD shows consistently strong skill at short lead times while AFNOCast exhibits improved performance primarily over Sudan, with noticeable skill during longer lead months, particularly in MAM, June and September.

The ability of the forecasts to discriminate between above (AN), near (NN) and below normal (BN) conditions is first illustrated in Figure 11 using receiver operating characteristic (ROC) curves. The dashed gray diagonal represents no predictive discrimination for the respective category (AUC = 0.5). Two exemplary target months, April and July, were selected to represent different seasons and to highlight cases in which each post-processing product outperforms the others. For the valid month May, BCSD outperforms both the raw and AFNOCast post-processed forecasts across all tercile categories and lead times. This advantage is more pronounced at shorter lead times (lead month 1, LM1). For all forecast products, discrimination skill decreases markedly with increasing lead time, with AUC values approaching 0.5, indicating limited predictive skill at longer leads. In July, AFNOCast generally outperforms the other two products, especially for the discrimination of BN and AN conditions. Notably, AFNOCast remains useful discrimination skill even at longer lead times (lead month 5, LM5) with AUC values remaining ≥0.6.

A summary of the discrimination skill across valid months is presented in Figure 12. The raw SEAS5 forecast exhibits measurable, albeit weak and seasonally intermittent, skill in discriminating tercile categories, indicating the presence of a large-scale predictive signal. For most months and tercile categories, predictive discrimination decreases with increasing lead time. An exception is observed for the near normal (NN) category, for which AUC values are frequently close to 0.5, particularly in April and from June to September, indicating limited discriminability irrespective of lead time. While both post-processing methods improve the AUC relative to the raw forecast in most cases, there are instances where no improvement is achieved, for example for NN conditions in April and July. Overall, BCSD exhibits superior discrimination skill for most months and lead times, however, AFNOCast achieves the highest AUC values particularly in July and mostly in April.

3.2.2 Precipitation intensity dependent skill

To better understand the source of forecast skill, we analyze the Continuous Ranked Probability Score (CRPS) values across rainfall intensity bins. This approach allows us to identify which precipitation regimes contribute most to the observed improvements in forecast performance. Figure 13 presents CRPS values grouped according to bins of daily rainfall intensity. For lower to moderate intensities, up to approximately 40 mm/day, BCSD demonstrates slightly better skill than AFNOCast with lower CRPS values, indicating that the statistical method is effective in correcting the central portion of the rainfall distribution. However, for high intensity rainfall events exceeding 40 mm/day, the situation reverses. AFNOCast shows consistently lower CRPS values than BCSD, with the gap between the two methods increasing as precipitation intensity rises. This indicates a clear advantage of the DL approach in representing the tail of the distribution. In particular, the lower CRPS for these extremes reflects a novel skill in predicting high-impact events that are typically challenging for traditional statistical methods.

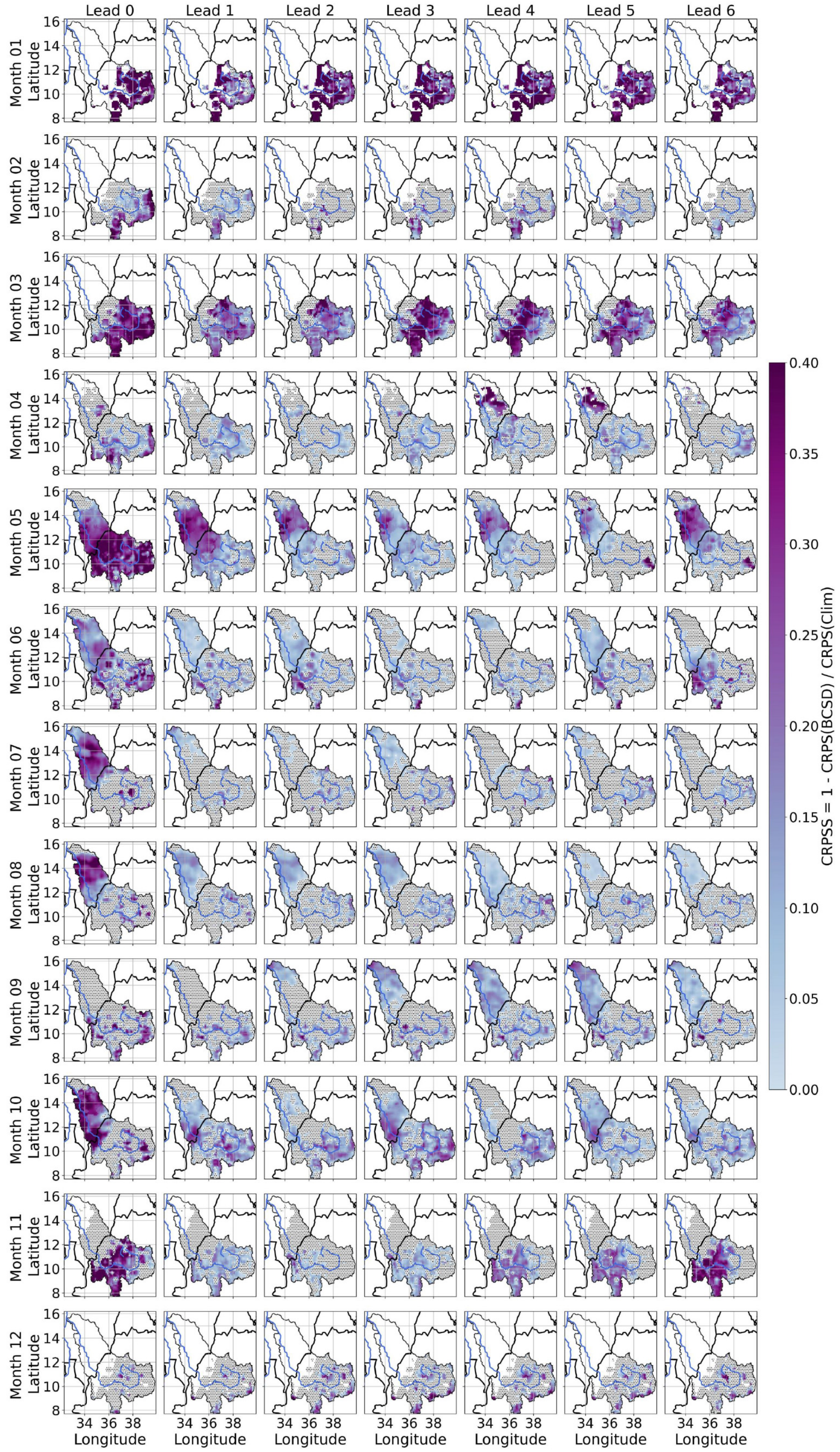

FIGURE 9
Spatial CRPSS of BCSD relative to climatology by lead and calendar month. Each panel shows CRPSS averaged over all forecast initializations (dotted grid points: no improvement over climatology, white grid points in the BN Basin: masked-out pixels with precip <0.1 mm/day).

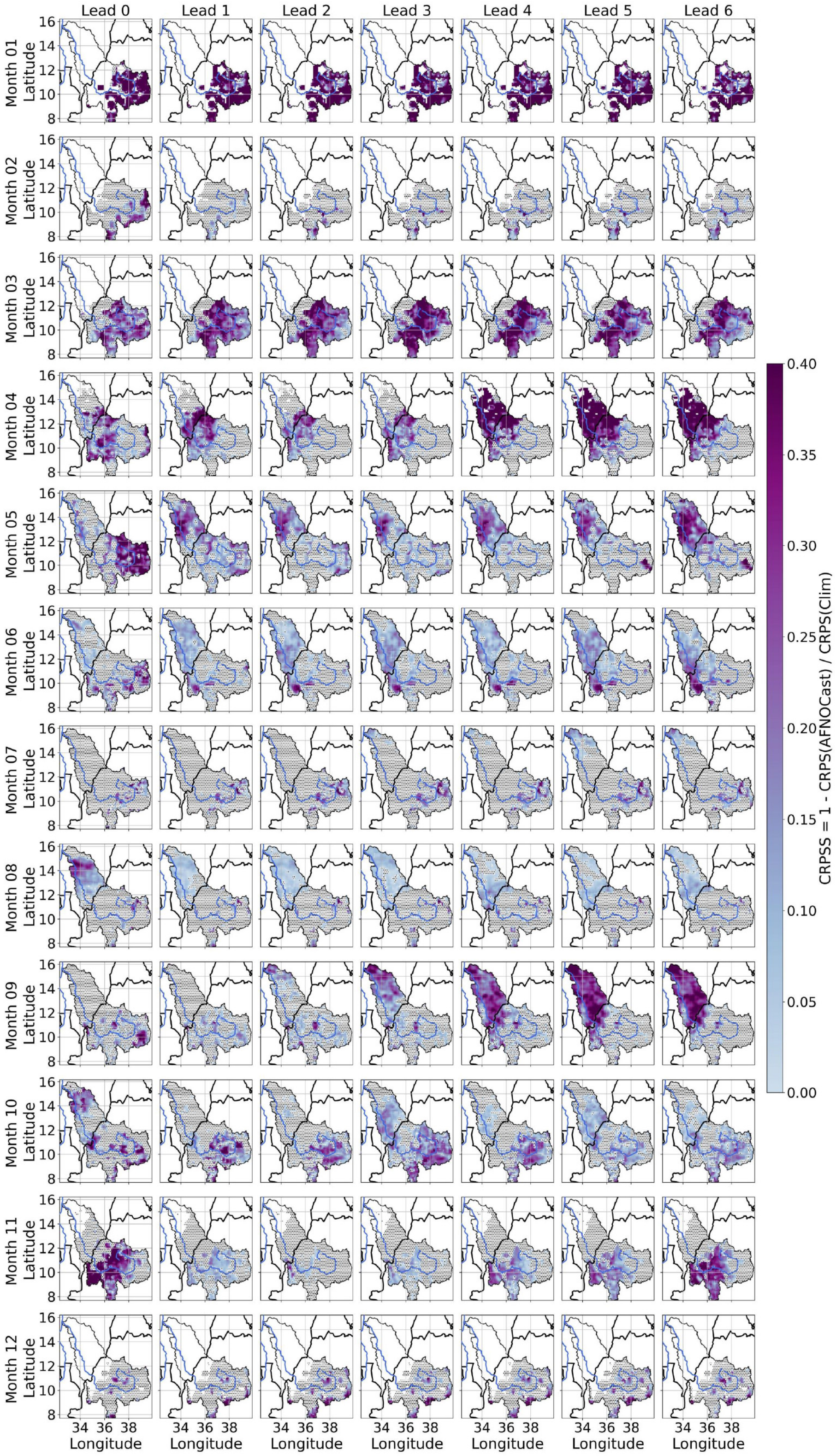

FIGURE 10
Spatial CRPSS of AFNOCast relative to climatology by lead and calendar month. Each panel shows CRPSS averaged over all forecast initializations (dotted grid points: no improvement over climatology, white grid points in the BN Basin: masked-out pixels with precipitation <0.1 mm/day).

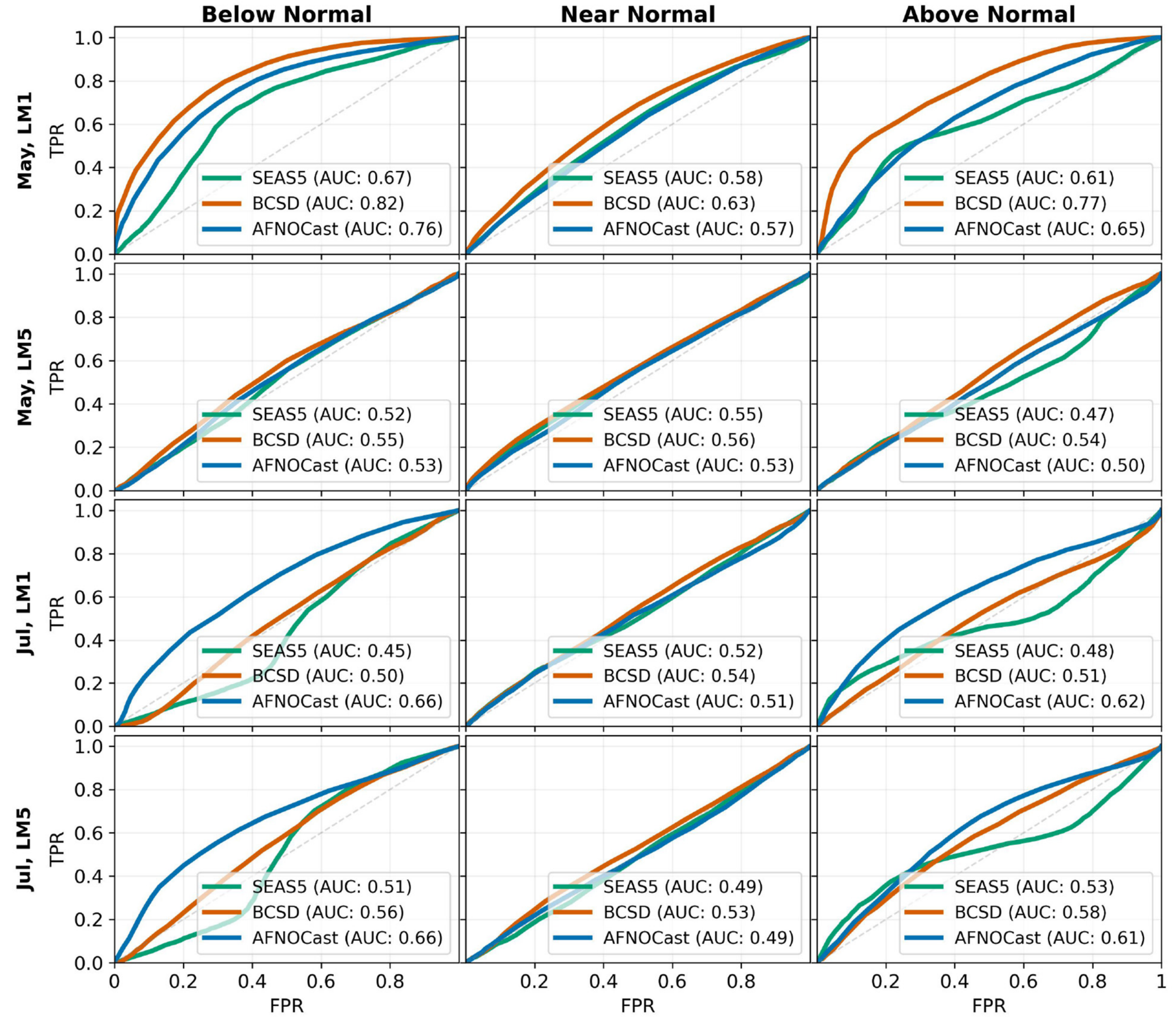

FIGURE 11
The ROC curves for the raw (SEAS5) and the post-processed forecasts (BCSD, AFNOCast) presented for the tercile categories below normal (BN), near normal (NN), and above normal (AN) for two exemplary months (May and July) and two exemplary lead months (LM1 and LM5).

3.3 Extreme event prediction

Capturing interannual variability in precipitation, particularly related to extreme events, remains a significant challenge for seasonal forecasting. While both BCSD and AFNOCast are able to reproduce the general occurrence of extreme and very extreme precipitation events, accurately predicting the timing, duration, and frequency of such events continues to be a limitation. As an example, we illustrate the contrasting cases of the years 2019 and 2022 (Figure 14). In 2022, a year characterized by fewer extreme events in the reference dataset (ERA5-Land), the BCSD ensemble incorrectly indicates a high probability of extreme precipitation in August, despite no such events occurring. In contrast, AFNOCast demonstrates a more conservative probability for August and better alignment with the observed lower event frequency.

The year 2019 presents a case of higher occurrence of such extreme events, particularly early in the season. While BCSD underestimates the likelihood of such events in June, AFNOCast captures a higher probability during the late May to mid-June period, more closely matching the reference. One clear difference between the two post-processing methods is AFNOCast's ability to represent early-season extremes, particularly in the May to mid-June period, whereas BCSD consistently underestimates them each year. From July to October, both post-processing methods exhibit similar behavior across different years, with little interannual variability. This suggests that the driving seasonal forecast model (SEAS5) does not provide substantial predictive information for this period. Notably, during drier years, BCSD occasionally overcorrects by predicting a higher frequency and likelihood of extreme events than observed. Overall, the Brier Score (BS) averaged across years (Table 5) indicates that predicting interannual variability in extremes remains difficult and has a notable impact on forecast skill. For extreme events (>40 mm/day), AFNOCast achieves a slightly lower BS compared to BCSD, and

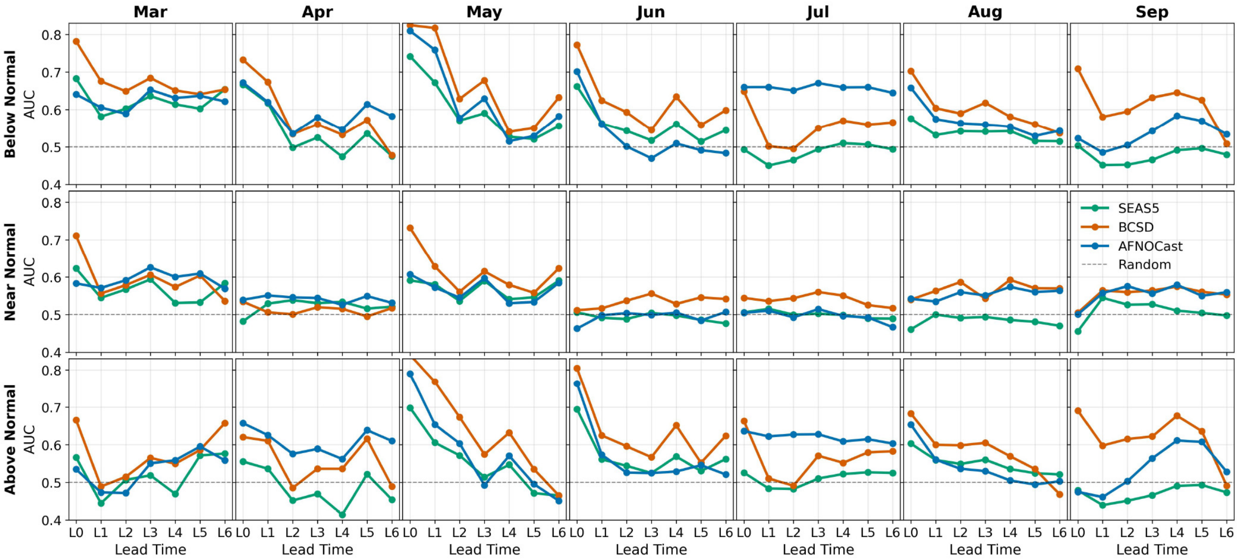

FIGURE 12
The area under the ROC curve (AUC) is a compact metric to represent the discrimination skill of each raw (SEAS5) and post-processed (BCSD, AFNOCast) forecast. The AUC is provided for each lead month (L0-6) and valid months of the relevant seasons (MAM and JJAS) of the region. The dashed gray line represents the skill of a random forecast and AUC values lower mean no skill.

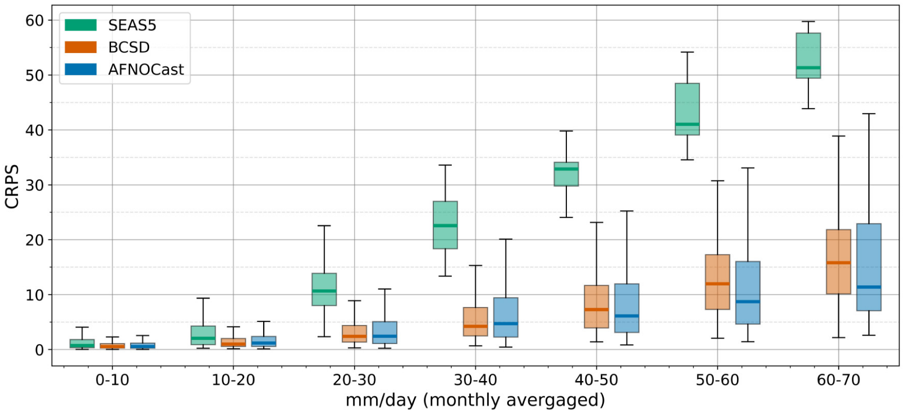

FIGURE 13
The CRPS against rainfall intensity bins for each forecast (raw SEAS5 in green, BCSD in orange, AFNOCast in blue).

for very extreme events (>120 mm/day), it improves the BS by 0.01.

3.4 Onset prediction skill

The onset of the rainy season in the Blue Nile Basin is characterized by substantial interannual variability of up to 50 days, ranging from early May to late June. Figure 15 shows the onset of the rainy season at Bahir Dar (Latitude 11.60°, Longitude 37.32°), located in the Upper Blue Nile Basin near Lake Tana, which similarly exhibits substantial interannual variability, with onset dates in the reference dataset spanning up to 55 days. In Figure 15 the onset predictions from forecasts are based on forecasts initialized in April, focusing on a single grid point rather than a basin average. While the raw SEAS5 forecasts capture the general trend of onset timing, such as later onsets in 2020 and 2022, the median forecast onset displays a reduced variability (maximum deviation of 11 days). Differently, the ensemble spread spans a wide range of up to 80 days, reflecting the high uncertainty in the forecast. Differences introduced by the post-processing methods (BCSD and AFNOCast) are minor and do not significantly shift the median onset. All forecasts tend to remain close to the local climatological mean but still correctly reflect interannual anomalies, suggesting moderate skill in distinguishing between earlier- and later-than-normal onset years, even at a sub-basin scale.

The predictability of onset timing across years is evaluated pixel-wise using the Brier Score (BS), calculated for varying temporal tolerance windows. Specifically, a hit is defined as a

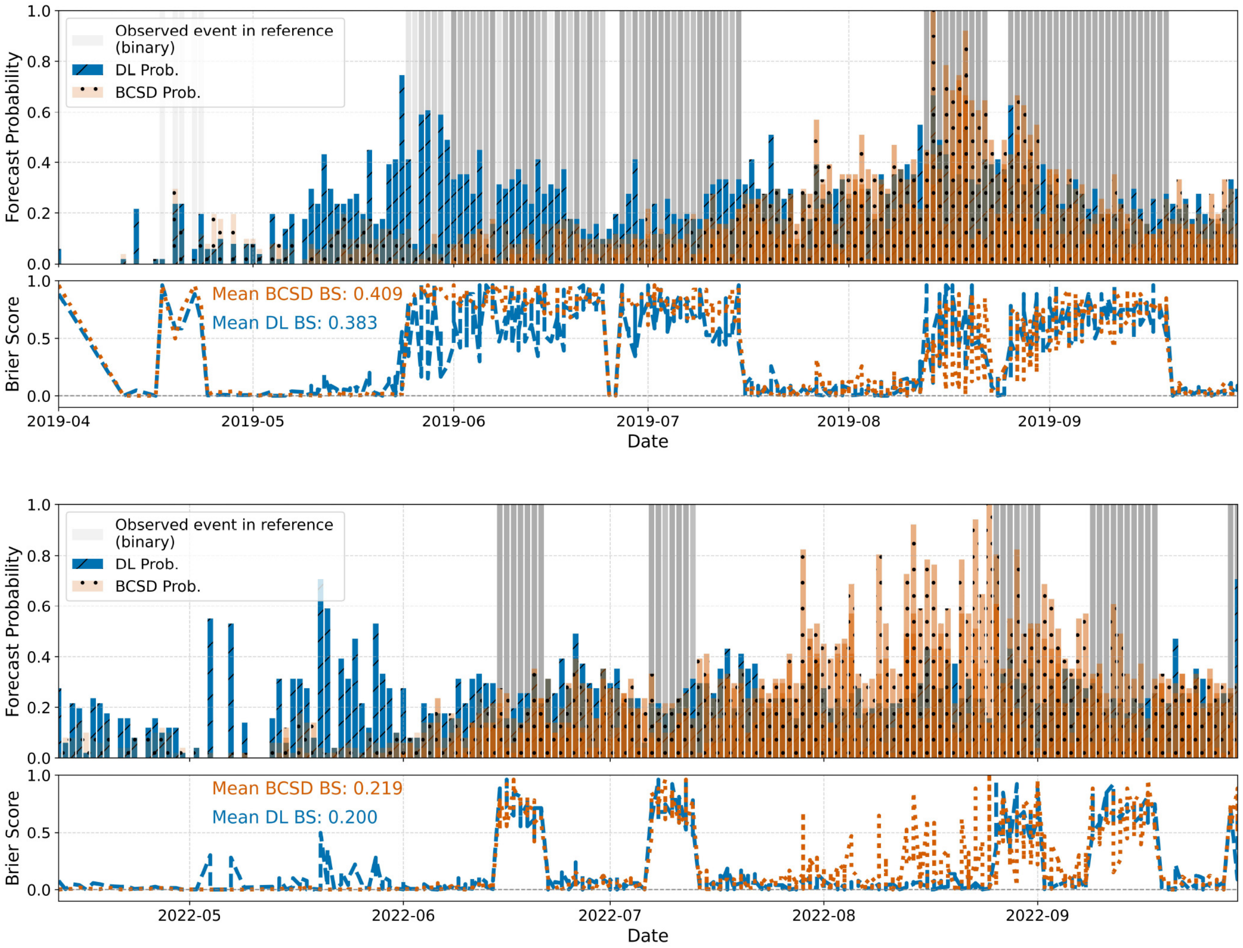

FIGURE 14
The probability of extreme event (>40 mm) detection, the reference extreme event (gray), and the corresponding Brier score in **(top)** 2019 and **(bottom)** 2022.

TABLE 5 Comparison of BS of BCSD and DL for >40mm and >120mm events across years.

Year	>40mm		>120mm	
	BCSD	DL	BCSD	DL
2017	0.326	0.326	0.423	**0.381**
2018	0.382	**0.357**	0.398	**0.367**
2019	0.409	**0.383**	0.412	**0.379**
2020	**0.359**	0.389	0.453	**0.450**
2021	**0.279**	0.294	**0.452**	0.460
2022	0.219	**0.200**	**0.251**	0.259
2023	**0.259**	0.274	**0.338**	0.344
All years	0.317	**0.316**	0.386	**0.376**

The best-performing product in each case (i.e., lowest BS) is indicated in bold.

predicted onset date occurring within ±5, ±10, ±15, or ±20 days of the observed onset in the reference. As expected, the BS decreases (i.e., improves) as the tolerance window increases (Figure 16). The BS analysis reveals only marginal differences in onset prediction skill between the two post-processing methods. Importantly, it suggests that at the current level of forecast quality a temporal window of approximately ±15 days is required to achieve skillful onset predictions exceeding climatology with an BS of around 0.25. This highlights the challenge of accurately predicting the onset of the rainy season at finer temporal resolutions in this region.

4 Discussion

This study presents a DL approach, named Seasonal AFNOCast, for bias-correcting and downscaling seasonal precipitation forecasts. Our results show that this method can reproduce precipitation distributions compared to the reference data at both daily and monthly resolutions across all ensemble members, comparable to the performance of the conventional statistical approach, BCSD. Nonetheless, both approaches show distinct strengths and weaknesses.

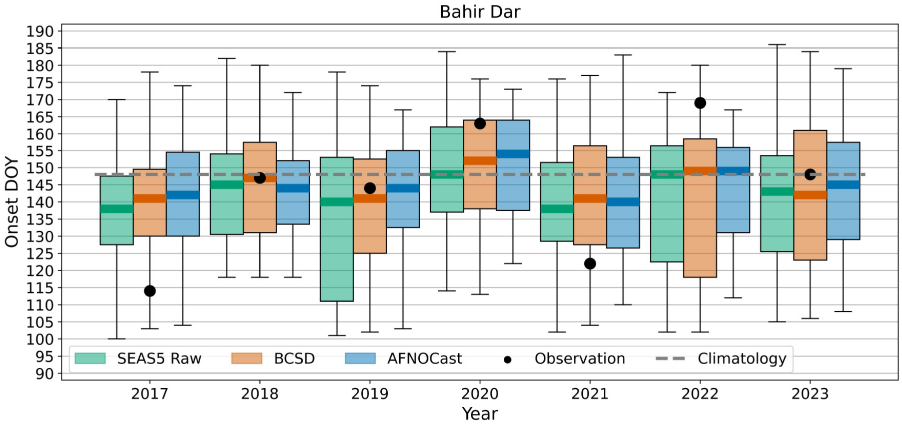

FIGURE 15
Onset prediction (boxplots represent the ensemble), climatological onset average (gray dashed line), and onset observed in the reference for a single pixel: Bahir Dar, Ethiopia.

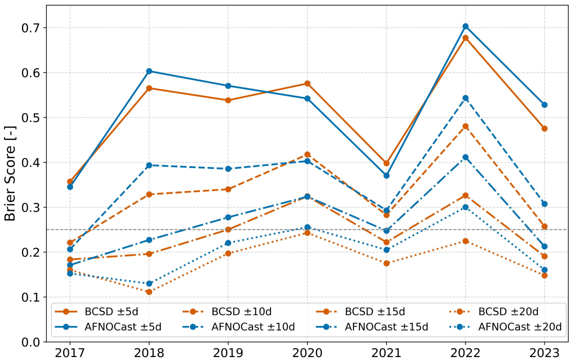

FIGURE 16
The Brier Score of predicting the onset for varying time windows (spatially averaged), based on the April initialized forecasts for each post-processed forecast.

4.1 Reproduction of spatial precipitation patterns

The findings of the qualitative analysis suggest that while both methods improve spatial match with the reference compared to raw forecasts, AFNOCast's capacity to adaptively correct based on context may offer an advantage in capturing the dynamic and heterogeneous nature of rainfall during the East African rainy season. However, although BCSD applies pixel-wise corrections that often reduce spatial coherence, both qualitative analysis and RAPSD results indicate that BCSD still reproduces the spatial patterns of the reference data reasonably well. Consequently, the added value of the DL approach for improving overall spatial precipitation structures appears to be limited. Nevertheless, the RAPSD analysis highlights that the DL model more accurately captures the spectral characteristics of the reference data, particularly at smaller spatial scales. This finding supports the hypothesis that DL methods, with convolutional layers, inherently account for spatial context and can reproduce scale-dependent spectral variability.

4.2 Overall forecast skill across lead times and rainfall intensities

In terms of overall forecast skill, our analysis shows that the DL approach can improve the general predictive ability of SEAS5, similar to BCSD. However, an important distinction emerges when analyzing performance by forecast lead time. While BCSD exhibits particularly high skill in the first lead month, the DL model underperforms at this short lead time. Conversely, for longer lead times, the DL model shows an advantage, outperforming BCSD in reducing model drift and maintaining predictive skill, particularly beyond the third month.

Skill relative to climatology is particularly relevant for forecast applications. While the raw SEAS5 forecast shows no probabilistic skill compared to climatology, positive CRPSS values demonstrate that both post-processing methods add probabilistic skill. In contrast, ROC-based discrimination analysis indicates that some large-scale predictive signal is already present in the raw forecast, albeit weak. This discrimination skill is clearly enhanced by both post-processing approaches. Taken together, these results indicate that SEAS5 contains a usable large-scale signal that due to its bias is unable to outperform climatology probabilistically on its own but can be extracted and amplified through post-processing. Where the raw forecast exhibits clear initial discrimination skill, particularly at short lead times, BCSD tends to provide the strongest improvements, as seen for several months at lead month zero, suggesting that the DL model does not fully exploits the strong initial-condition signal present in SEAS5. In contrast, AFNOCast demonstrates improved discrimination performance in specific seasons, most notably in July, where it outperforms both BCSD and the raw forecast for BN and AN categories. Importantly here, AFNOCast maintains useful discrimination skill at longer lead times, even when the raw forecast signal is weak, which aligns with the findings of the CRPSS analysis. Note, that the limited discrimination skill for near normal conditions, with AUC values frequently close to 0.5 for the raw forecast, and only modest improvements (0.5–0.6) after post-processing is consistent with previous studies [e.g., for Eastern Africa (Walker et al., 2019)].

Skill divergence is also evident when analyzed across rainfall intensities. BCSD performs better in the prediction of moderate rainfall events. In contrast, Seasonal AFNOCast displays a stronger ability to represent extreme precipitation events, especially for daily intensities above 40 mm.

4.3 Impact-oriented evaluation: the added value and limitations of Seasonal AFNOCast

To further evaluate the added value of DL for extremes, we conducted an event-based analysis using defined extreme thresholds and verified the predictions with the BS. The results confirm the findings from the CRPS-based evaluation: although year-to-year variability remains high and improvements are modest, Seasonal AFNOCast shows promising skill in predicting extreme rainfall events. These results are consistent with Wilks (2002), who noted that while individual daily weather events on seasonal time scales are inherently random, the statistics of daily events can nonetheless be predictable through changes in the underlying seasonal statistics. The modest improvements in extreme precipitation event representation are therefore likely driven primarily by improved matching of the reference distribution rather than enhanced predictability, as both post-processing approaches ultimately remain constrained by the limited signal available at daily scale in the raw forecast.

While seasonal forecasts are currently issued at monthly intervals, making runtimes of hours operationally acceptable, computational efficiency remains relevant when considering future developments and resource constraints. As forecast systems evolve toward more frequent sub-seasonal to seasonal updates, and potentially at higher resolution, faster post-processing will become increasingly advantageous. Moreover, once trained, Seasonal AFNOCast requires only a lightweight inference step, reducing both runtime and computational resource demands. In contrast, BCSD requires approximately 30 min per forecast (with parallelization), compared to 1.75 min on GPU or 7 min on CPU for AFNOCast, making the DL approach a promising method for future developments in resource constrained forecasting contexts.

However, the evaluation of rainfall onset did not show improvements using the DL approach. Despite training the model on a smaller temporal window (±5 days), which should, in theory, help differentiate seasonal transitions better than BCSD, the model struggled to outperform the statistical method. One potential reason is the monthly training strategy, which may limit the DL model's ability to incorporate broader temporal context or interannual variability that is critical for reliable onset detection. In addition, Scheuerer et al. (2024) report biases of up to 15 days for the JJAS onset when using SEAS5 forecasts, even after bias-correcting precipitation intensities. As noted by the authors, this indicates that post-processing does not necessarily remove systematic timing biases, which may explain why skill in our analysis only emerges at similarly wide temporal windows. Moreover, the small sample size of seven years used in the evaluation limits the robustness of the findings. The sensitivity of the onset score to both the time window and year further hinder final conclusions between the skill of each method. However, assessing onset prediction skill within a 15-day window offers a useful reference point for future studies aiming to improve temporal precision in seasonal forecasts predicting the onset.

Overall, the limited skill found for both extreme event timing and rainfall onset prediction is consistent with the inherently high local and temporal variability of precipitation, which makes precipitation particularly difficult to predict at daily resolution on seasonal time scales. While the large-scale seasonal signal is captured by both the raw and post-processed forecasts, its translation into precise spatio-temporal information remains limited. Consequently, post-processing primarily improves the statistical representation of precipitation characteristics rather than the deterministic timing of individual events or seasonal transitions.

5 Conclusions

This study demonstrates the potential of post-processing to enhance seasonal precipitation forecasts by systematically correcting biases, refining spatial structure, and improving forecast skill relative to both the raw forecast and climatology. For the Blue Nile Basin, we showed that Seasonal AFNOCast, a DL-based post-processing method, effectively downscales and bias-corrects SEAS5 precipitation forecasts, improving agreement with reference rainfall distributions at both daily and monthly time scales.

Across both deterministic and probabilistic metrics, post-processing leads to clear and consistent skill improvements. The reduction in RMSE confirms that systematic errors in rainfall magnitude are effectively corrected, while improvements in CRPS and CRPSS demonstrate that the full predictive distribution is better aligned with the reference. These improvements are expected outcomes of bias correction, yet their consistent manifestation across multiple metrics is important: reduced RMSE indicates improved accuracy of the ensemble mean, whereas improved CRPS and CRPSS confirm enhanced quality of probabilistic seasonal forecasts. Comparing the two post-processing approaches shows complementary strengths. Deterministically, Seasonal AFNOCast reduces the RMSE of monthly averaged rainfall maps by approximately 0.1 mm/day compared to BCSD, highlighting its enhanced accuracy in reproducing observed rainfall magnitudes. The probabilistic analysis shows that Seasonal AFNOCast performs comparably to BCSD based on the CRPSS, significantly improving the raw SEAS5 forecast. It does not reach BCSD's skill during the first forecast month (lead month 0) but outperforms BCSD for longer lead times (lead month 4 to 6). This improved skill at extended lead times could be particularly valuable for climate sensitive decision-making processes such as in agricultural planning or reservoir operation.

Beyond statistical skill, post-processing substantially improves spatial characteristics of the forecasts. We hypothesized that learning corrections in a spatial and temporal context via DL, rather than applying rigid, pixel-wise distribution adjustments as done by BCSD, would lead to improved spatial pattern representation. Indeed, while BCSD already provides a strong reproduction of large-scale spatial structures, Seasonal AFNOCast slightly enhances the representation of spectral frequency characteristics, particularly through better capture of small-scale features. Both BCSD and

Seasonal AFNOCast increase the effective spatial resolution relative to SEAS5, which is a critical requirement for subsequent impact-oriented applications such as streamflow and sediment transport modeling, where the location and spatial coherence of precipitation are as important as the total amount. The improved agreement of spatial rainfall patterns with the reference further strengthens the applicability of post-processed forecasts for such modeling chains.

This slight improvement of AFNOCast in fine-scale structure is also reflected in its enhanced representation of extreme precipitation events, as evidenced by improved performance in both event detection (Brier Score) and skill of high precipitation values (CRPS). The indicated ability to represent extremes is a key advantage, with significant implications for early warning systems, enabling improved preparedness and potentially reducing the societal impacts of severe events.

A key question for practical application is whether post-processed forecasts provide information beyond climatology. The CRPSS relative to climatology shows that both post-processing methods achieve skill exceeding climatological benchmarks, although the improvement is generally modest. The detailed spatio-temporal CRPSS analysis showed that both post-processing methods substantially improve forecast skill relative to climatology during the MAM season, underlining the added value of bias-correction and downscaling in the pre-rainy season. During the JJAS season, improvements in skill relative to climatology are most pronounced over the northern Blue Nile Basin (Sudan). Here, BCSD provides the main improvement at lead month 0, with skill persisting for longer lead times mainly through August, while Seasonal AFNOCast maintains enhanced skill at longer lead times notably into August and September. This temporal persistence of skill represents a key advantage of the DL approach, especially for applications requiring reliable forecasts later in the rainy season. Discriminative skill analysis using ROC curves further reveals that SEAS5 contains a weak but detectable large-scale predictive signal. Importantly, this signal is preserved and amplified through post-processing, leading to improved discrimination of below- and above-normal precipitation. While AFNOCast is less effective than BCSD at retaining the signal at lead month 0, it substantially outperforms BCSD in highly relevant months such as July. The largest improvements occur when the raw forecast already exhibits skill, highlighting that post-processing acts primarily as a skill amplifier rather than an independent source of predictability. Consequently, the use of skillful raw seasonal forecasts would likely translate directly into further gains in post-processed forecast performance.

From an operational perspective, computational efficiency is a critical advantage of DL-based approaches. Seasonal AFNOCast requires only 5%–23% of the prediction time needed by the statistical method BCSD, ensuring applicability in resource-limited settings and cost-effective deployment as forecast frequencies and resolutions may increase.

Despite these advances, important limitations remain. The skill of both post-processing methods is fundamentally constrained by the large-scale predictive information contained in the SEAS5 input. While post-processing enhances and refines this signal, it cannot compensate for missing large-scale predictability. Although large-scale early and late onset signals are visible in both raw and post-processed forecasts, skill in the timing and location of rainfall onset and extreme daily precipitation remains limited due to the weak coupling between large-scale predictability and daily-scale variability, as well as persistent temporal onset biases even in bias-corrected forecasts. We conclude that enhanced representation of extreme precipitation events of ANFOCast, shown in both event detection (Brier Score) and skill of high precipitation values (CRPS), is a result of the improved fine-scale structures and matching of the reference distribution tail. While the indicated ability to represent extremes can be a key advantage, with significant implications for early warning systems, enabling improved preparedness and potentially reducing the societal impacts of severe events, these findings emphasize that post-processed seasonal forecasts remain dependent on the quality of the underlying dynamical forecast.

Future research should explore the integration of additional predictors, such as large-scale atmospheric patterns, to enhance the information content available to Seasonal AFNOCast and improve its capability to predict complex seasonal features like onset timing. Here, a key advantage of the DL framework becomes evident: additional input variables, such as teleconnection indices or multivariate predictors, can be incorporated without redesigning the overall architecture, enabling the model to learn complex relationships across multiple drivers.

This study highlights the potential of DL-based post-processing to enhance seasonal forecasting in topographically complex regions. By combining competitive skill and enabling efficient operational deployment with improved spatial detail, Seasonal AFNOCast provides a promising foundation for impact-orientated forecasts. Thus, for the next generation of operational climate services aimed at supporting adaptation, preparedness, and sustainable resource management. Given these potential benefits, further research along this path is important to fully realize the contribution of deep learning approaches to operational climate services, disaster risk reduction and climate resilience.

Data availability statement

The reference dataset used for this study is publicly available and can be downloaded from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download. The raw SEAS5 data provided by ECMWF are available for registered users and can be downloaded from MARS https://confluence.ecmwf.int/display/FCST/Implementation+of+Seasonal+Forecast+SEAS5.

Author contributions

RW: Methodology, Data curation, Conceptualization, Validation, Software, Investigation, Writing – review & editing, Writing – original draft, Formal analysis, Visualization. CC: Conceptualization, Supervision, Methodology, Writing – review & editing. JP: Conceptualization, Methodology, Writing – review & editing. LG: Methodology, Conceptualization, Writing – review & editing. CL: Supervision, Writing – review & editing, Conceptualization, Data curation, Funding acquisition, Software. TS: Conceptualization, Writing – review & editing, Supervision. HK: Funding acquisition, Project administration, Writing – review & editing, Conceptualization.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research has been supported by the German Federal Ministry of Research, Technology and Space (Bundesministerium für Forschung, Technologie und Raumfahrt, BMFTR) (grant no. 02WGR1643A-C) through the project “SPS Blue Nile”. The project funding was acquired by CL and HK. We acknowledge the ECMWF for providing ERA5-Land and SEAS5 data.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. ChatGPT (GPT-5) has been used to improve the English language of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fclim.2026.1691030/full#supplementary-material

References

Ahmed, S. M. (2020). Impacts of drought, food security policy and climate change on performance of irrigation schemes in Sub-saharan Africa: the case of Sudan. *Agric. Water Manag.* 232:106064. doi: 10.1016/j.agwat.2020.106064

Ali, Y. S., Crosato, A., Mohamed, Y. A., Abdalla, S. H., and Wright, N. G. (2014). Sediment balances in the Blue Nile River Basin. *Int. J. Sediment Res.* 29, 316–328. doi: 10.1016/S1001-6279(14)60047-0

Amante, C., and Eakins, B. W. (2009). *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis.* NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center; NOAA. doi: 10.7289/V5C8276M

Boé, J., Terray, L., Habets, F., and Martin, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *Int. J. Climatol.* 27, 1643–1655. doi: 10.1002/joc.1602

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78, 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Civitarese, D. S., Szwarcman, D., Zadrozny, B., and Watson, C. (2021). Extreme precipitation seasonal forecast using a transformer neural network. *arXiv preprint arXiv:2107.06846.*

Coelho, C. A. S., Stephenson, D. B., Balmaseda, M., Doblas-Reyes, F. J., and Van Oldenborgh, G. J. (2006). Toward an integrated seasonal forecasting system for South America. *J. Clim.* 19, 3704–3721. doi: 10.1175/JCLI3801.1

Crochemore, L., Ramos, M.-H., and Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.* 20, 3601–3618. doi: 10.5194/hess-20-3601-2016

da Rocha Júnior, R. L., Cavalcante Pinto, D. D., dos Santos Silva, F. D., Gomes, H. B., Barros Gomes, H., Costa, R. L., et al. (2021). An empirical seasonal rainfall forecasting model for the northeast region of Brazil. *Water* 13:1613. doi: 10.3390/w13121613

ECMWF (2021). *Seas5 user guide, version 1.2.* Available online at: https://www.ecmwf.int/en/elibrary/81237-seas5-user-guide (Accessed November 08, 2025).

ECMWF (2025). *Era5-land: data documentation.* Available online at: https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation (Accessed November 08, 2025).

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Food and Agriculture Organization of the United Nations (FAO) (2020). *The Sudan Flood Impact Rapid Assessment – September 2020.* Technical report, Rome.

Foufoula-Georgiou, E., Guilloteau, C., Nguyen, P., Aghakouchak, A., Hsu, K.-L., Busalacchi, A., et al. (2020). Advancing precipitation estimation, prediction, and impact studies. *Bull. Am. Meteorol. Soc.* 101, E1584–E1592. doi: 10.1175/BAMS-D-20-0014.1

Gebrechorkos, S. H., Pan, M., Beck, H. E., and Sheffield, J. (2022). Performance of state-of-theŘart C3S european seasonal climate forecast models for mean and extreme precipitation over Africa. *Water Resour. Res.* 58:e2021WR031480. doi: 10.1029/2021WR031480

Glawion, L., Polz, J., Kunstmann, H., Fersch, B., and Chwala, C. (2023). spateGAN: spatioŘtemporal downscaling of rainfall fields using a cGAN approach. *Earth Space Sci.* 10:e2023EA002906. doi: 10.1029/2023EA002906

Glawion, L., Polz, J., Kunstmann, H., Fersch, B., and Chwala, C. (2025). Global spatio-temporal era5 precipitation downscaling to km and sub-hourly scale using generative AI. *NPJ Clim. Atmosp. Sci.* 8:219. doi: 10.1038/s41612-025-01103-y

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B. (2022). *Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers.* GitHub repository. Available online at: https://github.com/NVlabs/AFNO-transformer (Accessed July 22, 2025).

Haile, G. G., Tang, Q., Hosseini-Moghari, S.-M., Liu, X., Gebremicael, T. G., Leng, G., et al. (2020). Projected impacts of climate change on drought patterns over east Africa. *Earth's Fut.* 8:e2020EF001502. doi: 10.1029/2020EF001502

Han, L., Chen, M., Chen, K., Chen, H., Zhang, Y., Lu, B., et al. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. *Adv. Atmosph. Sci.* 38, 1444–1459. doi: 10.1007/s00376-021-0215-y

Harder, P., Hernandez-Garcia, A., Ramesh, V., Yang, Q., Sattegeri, P., Szwarcman, D., et al. (2023). Hard-constrained deep learning for climate downscaling. *J. Mach. Learn. Res.* 24, 1–40. doi: 10.5194/egusphere-egu23-4350

Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., and Levit, J. J. (2001). Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeorol.* 2, 406–418. doi: 10.1175/1525-7541(2001)002<0406:MSPOAH>2.0.CO;2

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* 15, 559–570. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hess, P., Lange, S., Schtz, C., and Boers, N. (2023). Deep learning for biasŘcorrecting CMIP6Řclass earth system models. *Earth's Fut.* 11:e2023EF004002. doi: 10.1029/2023EF004002

Hwang, S., Graham, W., Hernández, J. L., Martinez, C., Jones, J. W., and Adams, A. (2011). Quantitative spatiotemporal evaluation of dynamically downscaled MM5 precipitation predictions over the Tampa Bay region, Florida. *J. Hydrometeorol.* 12, 1447–1464. doi: 10.1175/2011JHM1309.1

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., et al. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.* 12, 1087–1117. doi: 10.5194/gmd-12-1087-2019

King, J. A., Washington, R., and Engelstaedter, S. (2021). Representation of the Indian Ocean Walker circulation in climate models and links to Kenyan rainfall. *Int. J. Climatol.* 41, E616–E643. doi: 10.1002/joc.6714

Koldunov, N., Rackow, T., Lessig, C., Danilov, S., Cheedela, S. K., Sidorenko, D., et al. (2024). Emerging ai-based weather prediction models as downscaling tools. *arXiv preprint arXiv:2406.17977*.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., et al. (2023). "FourCastNet: accelerating global high-resolution weather forecasting using adaptive Fourier neural operators," in *Proceedings of the Platform for Advanced Scientific Computing Conference, PASC '23* (New York, NY, USA: Association for Computing Machinery), 1–11. doi: 10.1145/3592979.3593412

Lala, J., Tilahun, S., and Block, P. (2020). Predicting rainy season onset in the Ethiopian highlands for agricultural planning. *J. Hydrometeorol.* 21, 1675–1688. doi: 10.1175/JHM-D-20-0058.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science* 382, 1416–1421. doi: 10.1126/science.adi2336

Leinonen, J., Hamann, U., Nerini, D., Germann, U., and Franch, G. (2023). Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., et al. (2021). "Fourier neural operator for parametric partial differential equations," in *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lorenz, C., Portele, T. C., Laux, P., and Kunstmann, H. (2021). "Bias-corrected and spatially disaggregated seasonal forecasts: a long-term reference forecast product for the water sector in semi-arid regions. *Earth Syst. Sci. Data* 13, 2701–2722. doi: 10.5194/essd-13-2701-2021

Lorenz, C., Wiegels, R., Chwala, C., Fersch, B., Weber, J. N., Sawadogo, W., et al. (2025). *Pycast-s2s: a python framework for subseasonal-to- seasonal forecast post-processing.* Zenodo. Version 0.8.0.

Manzanas, R., Gutirrez, J., Fernández, J., Van Meijgaard, E., Calmanti, S., Magariño, M., et al. (2018). Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: added value for user applications. *Clim. Serv.* 9, 44–56. doi: 10.1016/j.cliser.2017.06.004

Mason, I. (1982). A model for assessment of weather forecasts. *Aust. Meteor. Mag* 30, 291–303. doi: 10.1071/ES82036

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* 13, 4349–4383. doi: 10.5194/essd-13-4349-2021

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

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. *Rev. Geophy.* 55, 590–635. doi: 10.1002/2016RG000544

Nikulin, G., Asharaf, S., Magario, M. E., Calmanti, S., Cardoso, R. M., Bhend, J., et al. (2018). Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa. *Clim. Serv.* 9, 72–85. doi: 10.1016/j.cliser.2017.11.003

Oliveira, E. C. L., d., Nogueira Neto, A. V., Santos, A. P. P., d., da Costa, C. P. W., et al. (2023). Precipitation forecasting: from geophysical aspects to machine learning applications. *Front. Clim.* 5:1250201. doi: 10.3389/fclim.2023.1250201

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., Lee, J., et al. (2021). Learning to correct climate projection biases. *J. Adv. Model. Earth Syst.* 13:e2021MS002509. doi: 10.1029/2021MS002509

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). Fourcastnet: a global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Prasad, A., Harder, P., Yang, Q., Sattegeri, P., Szwarcman, D., Watson, C., et al. (2024). "Evaluating the transferability potential of deep learning models for climate downscaling," in *Proceedings of the 41st International Conference on Machine Learning, Machine Learning for Earth System Modeling Workshop (ICML 2024)*.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., et al. (2019). Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0). *Geosci. Model Dev*. 12, 4185–4219. doi: 10.5194/gmd-12-4185-2019

Rodwell, M. J., and Doblas-Reyes, F. J. (2006). Medium-range, monthly, and seasonal prediction for europe and the use of forecast information. *J. Clim.* 19, 6025–6046. doi: 10.1175/JCLI3944.1

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer international publishing), 234–241. doi: 10.1007/978-3-319-24574-4_28

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP Climate Forecast System Version 2. *J. Clim.* 27, 2185–2208. doi: 10.1175/JCLI-D-12-00823.1

Scheuerer, M., Bahaga, T. K., Segele, Z. T., and Thorarinsdottir, T. L. (2024). Probabilistic rainy season onset prediction over the greater horn of Africa based on long-range multi-model ensemble forecasts. *Clim. Dyn.* 62, 3587–3604. doi: 10.1007/s00382-023-07085-y

Seregina, L. S., Fink, A. H., van der Linden, R., Elagib, N. A., and Pinto, J. G. (2019). A new and flexible rainy season definition: validation for the Greater Horn of Africa and application to rainfall trends. *Int. J. Climatol.* 39, 989–1012. doi: 10.1002/joc.5856

Tan, J., Huang, Q., and Chen, S. (2024). Deep learning model based on multi-scale feature fusion for precipitation nowcasting. *Geosci. Model Dev*. 17, 53–69. doi: 10.5194/gmd-17-53-2024

Trentini, L., Dal Gesso, S., Venturini, M., Guerrini, F., Calmanti, S., and Petitta, M. (2022). A novel bias correction method for extreme events. *Climate* 11:3. doi: 10.3390/cli11010003

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R. (2017). "Deepsd: generating high resolution climate change projections through single image super-resolution," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 1663–1672. doi: 10.1145/3097983.3098004

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., et al. (2021). Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world. *Bull. Am. Meteorol. Soc.* 102, E681–E699. doi: 10.1175/BAMS-D-19-0308.1

Walker, D. P., Birch, C. E., Marsham, J. H., Scaife, A. A., Graham, R. J., and Segele, Z. T. (2019). Skill of dynamical and ghacof consensus seasonal forecasts of east african rainfall. *Clim. Dyn.* 53, 4911–4935. doi: 10.1007/s00382-019-04835-9

Wilks, D. S. (2002). Realizations of daily weather in forecast seasonal climate. *J. Hydrometeorol.* 3, 195–207. doi: 10.1175/1525-7541(2002)003<0195:RODWIF>2.0.CO;2

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophy. Res.* 107, ACL 6–1–ACL 6–15. doi: 10.1029/2001JD000659

World Meteorological Organization (2020). *Guidance on operational practices for objective seasonal forecasting (wmo-no. 1246)*. Available online at: https://library.wmo.int (Accessed July 22, 2025).

Yang, Q., Hernandez-Garcia, A., Harder, P., Ramesh, V., Sattigeri, P., Szwarcman, D., et al. (2024). Fourier neural operators for arbitrary resolution climate data downscaling. *J. Mach. Lear. Res.* 25, 1–30. Available online at: http://jmlr.org/papers/v25/23-0597.html (Accessed November 08, 2025).

Yang, S., Ling, F., Luo, J.-J., and Bai, L. (2025). Improving the seasonal forecast of summer precipitation in southeastern China using a CycleGAN-based deep learning bias correction method. *Adv. Atmosph. Sci.* 42, 26–35. doi: 10.1007/s00376-024-4003-3

Yitayew, M., and Melesse, A. M. (2011). "Critical water resources issues in the Nile River Basin," in *Nile River Basin: Hydrology, Climate and Water Use*, ed. A. M. Melesse (Dordrecht: Springer), 401–416. doi: 10.1007/978-94-007-0689-7_20

Zargar, M., Bronstert, A., Francke, T., Zimale, F. A., Worku, K. B., Wiegels, R., et al. (2025). Comparison and hydrological evaluation of different precipitation data for a large tropical region: the Blue Nile Basin in Ethiopia. *Front. Water* 7:1536881. doi: 10.3389/frwa.2025.1536881