Impact-based Skill Evaluation of Seasonal Precipitation Forecasts

Zahir Nikraftar[1], Rendani Mbuvha[1,2], Mojtaba Sadegh[3,4], Willem A. Landman[5]

[1] Machine Intelligence and Decision Systems (MInDS) Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London (QMUL), London, UK, E1 4NS.
[2]School of Statistics and Actuarial Science, University of Witwatersrand, Johannesburg, South Africa
[3] Department of Civil Engineering, Boise State University, Boise, ID, USA.
[4] United Nations University Institute for Water, Environment and Health, Hamilton, ON, Canada.
[5] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa.

Corresponding Author: Rendani Mbuvha (r.mbuvha@qmul.ac.uk)

Abstract

We introduce an impact-based framework to evaluate seasonal forecast model skill in capturing extreme weather and climate events over regions prone to natural disasters such as floods and wildfires. Forecasting hydroclimatic extremes holds significant importance in an era of increasing hazards such as wildfires, floods, and droughts. We evaluate the performance of five Copernicus Climate Change Service (C3S) seasonal forecast models (CMCC, DWD, ECCC, UK-Met, and Météo-France) in predicting extreme precipitation events from 1993 to 2016 using 14 indices reflecting timing and intensity (using absolute and locally-defined thresholds) of precipitation at a seasonal timescale. Performance metrics, including Percent Bias, Kendall Tau Rank Correlation Score, and model discrimination capacity, are used for skill evaluation. Our findings indicate that the performance of models varies markedly across regions and seasons. While the models generally show good skill in the tropical regions, their skill in extra-tropical regions is markedly lower. Elevated precipitation thresholds (i.e., higher intensity indices) correlate with heightened model biases, indicating deficiencies in modelling severe precipitation events. Our analysis using an impact-based framework highlights the superior predictive capabilities of the UK-Met and Météo-France models in capturing the underlying processes that drive precipitation events, or lack thereof, across many regions and seasons. Other models exhibit strong performance in specific regions and seasons but not globally. These results advance our understanding of an impact-based framework for capturing a broad spectrum of extreme weather and climate events and further inform the strategic fusion of diverse models across different regions and seasons, thereby offering insights for disaster management and risk analysis.

Key Points

- C3S models demonstrate notable skill in forecasting seasonal variability of precipitation, particularly in the tropical and sub-tropical regions.
- UK-Met and Météo-France consistently outperform other models, demonstrating superior accuracy and reliability in various regions and seasons
- Our Impact-Based Framework offers valuable insights for targeted risk assessments, particularly in wildfire- and flood-prone regions using seasonal forecast models.

Plain Language Summary

We introduce a novel model assessment framework by investigating the impact of extreme events in areas prone to disasters such as floods and wildfires. We investigate how well five seasonal forecast models predict extreme precipitation, and related events such as floods and droughts. We assess the performance of models through 14 indices that assess the timing and intensity of precipitation during different seasons and in various regions globally. Our results show that some models outperform others in certain regions and seasons. We find that two models, UK-Met and Météo-France, are particularly skillful at predicting extreme events in various regions and seasons. This information is important for improved management and understanding of the risks of natural disasters.

1. Introduction

Precipitation plays a crucial role in momentum flux exchange at the ocean– atmosphere-land interface (Xue et al., 2020), and as such, is one of the primary outputs of weather and climate models (Tapiador et al., 2019). Numerous international initiatives such as the North American Multi-Model Ensemble (NMME, https://www.cpc.ncep.no aa.gov/products /NMME/) and Copernicus Climate Change Service (C3S, https://cds.climate .copernicus.eu/) multi-system seasonal forecast models forecast precipitation, and other Météorological factors, at various spatiotemporal scales. Such forecasts are used for a variety of purposes, including extreme event early warning. Forecast models rely on the sources of atmospheric predictability, such as modes of variability including El Niño–Southern Oscillation (ENSO), Madden–Julian oscillation (MJO), Quasi-Biennial Oscillation (QBO), and Indian Ocean Dipole (IOD). Other sources of predictability include anomalies in the initial state of an Earth system component with a persistence time that aligns with the projected forecast duration (i.e., large-scale anomalies in upper ocean heat content, sea ice, snowpack, soil moisture), and external forcing (Assessment of Intraseasonal to Interannual Climate Prediction and Predictability, 2010; Baldwin et al., 2003; Committee on Developing a U.S. Research Agenda to Advance Subseasonal to Seasonal Forecasting et al., 2016; Lau & Waliser, 2012; Shukla et al., 2000; Zhang et al., 1997). Despite its significance, seasonal forecast models face difficulties in predicting precipitation spatial patterns, timing and intensity (Tapiador et al., 2019; Mallakpour et al. 2022). This is because the predictive capabilities of seasonal forecasting models are constrained by the uncertainty in initial and boundary conditions, climate change-induced modifications of teleconnection patterns, imperfect parameterization schemes, the variability in parameters, quality of observation systems, model biases, and inherent properties of the climate system (Villarini et al., 2011; Xu et al., 2021; Kumar & Zhu, 2018). Also, seasonal forecast performance can vary across regions due to the complex and region-specific interactions between climate drivers and local environmental factors (Hao et al., 2018).

Accurate precipitation predictions are of great importance in the formulation of mitigation and adaptation measures for weather and climate extreme events as well as minimizing impacts from their cascading hazards such as flood, drought, and wildfire (Gebrechorkos et al., 2022; Frédéric Vitart & Robertson, 2018). Recently several environmental and Climate Forecasting Systems such as the hydrological forecasting system, the Canadian Forest Fire Weather Index System (http://cwfis.cfs.nrcan.gc.ca/en_CA/background/summary/fdr), and the global drought forecasting system (http://iridl. ldeo.columbia.edu/maproom/Global/Drought/Global/CPC_

GOB/MME_pt_Persist.html) have been developed, using seasonal forecasts as input with the purpose of risk assessment and response (Alfieri et al., 2013; Arheimer et al., 2020; Samaniego et al., 2019; Thielen et al., 2009). The accuracy and trustworthiness of such systems is highly dependent on the process representation and parametric accuracy of the seasonal forecast models (Gebrechorkos et al., 2022; Wanders & Wood, 2016). It is essential to assess the underlying performance of different forecast models across diverse global regions, and specific to the impacts that forecast errors may induce to identify the most effective and reliable sector- and region-specific models.

While studies have evaluated the skill of seasonal forecast models in predicting total precipitation at the sub seasonal to seasonal scales (Becker et al., 2014; Gebrechorkos et al., 2022; Nobakht et al., 2021; Roy et al., 2020), there is a need for an impact-based assessment of forecast models that can inform their applicability for target extremes (e.g., flood, wildfire). Traditional forecast model performance assessments conduct a top-down hazard information approach by mainly investigating the model's skill in capturing weather patterns in comparison to the reference datasets (De Andrade et al., 2019; Moron & Robertson, 2020; Frédéric Vitart & Robertson, 2018). The shift towards impact-based assessment framework reflects the evolving landscape of climate science and its increasing relevance in the face of a changing climate (Rad et al. 2022). It emphasizes the importance of moving beyond traditional evaluation methods and towards a comprehensive understanding of how weather forecasts directly influence society, ecosystems, and infrastructure resilience (AghaKouchak et al., 2018; Khorshidi et al., 2020; Mallakpour et al., 2022; Modaresi Rad et al., 2022; Sadegh et al., 2018). Such a framework considers the vulnerability of the local environment to specific weather events and warns of the associated impacts. An instance of such an influence might involve a chain reaction of hazards, like flooding due to back-to-back heavy rainfall events (Sadegh et al. 2018) or wildfires resulting from consecutive weeks of no precipitation and increased temperature, which can create conditions conducive to ignition and wildfire growth (Khorshidi et al. 2020). Impact-based assessment of seasonal precipitation forecasts involve assessing the effectiveness of models by considering the impact of extreme precipitation on various sectors and systems. It goes beyond assessing the mere accuracy of forecasted precipitation and aims to understand how well the forecasts translate into meaningful information for decision-making and risk management (AghaKouchak et al., 2023).

Our main objective in this study is to evaluate the skill of the five state-of-the-art seasonal prediction systems from the Copernicus Climate Change Service (C3S) multi-model at a global scale in predicting particular features of extreme events which could lead to hazards. We note that the primary goal in this study is to identify the models that are most representative of the underlying processes that drive weather and climate events at various spatial and temporal scales. While the importance of post-processing techniques in improving the reliability of the seasonal forecast models is undeniable, we evaluate raw forecasts to gain insights into their inherent model skills and deficiencies without the influence of statistical post-processing. This enables us to detect essential similarities and differences between forecast models and to assess their inherent skill in capturing extreme events. We utilize a selection of climate extreme indices designed to encompass diverse aspects of extreme events, including their frequency, timing, and intensity. These indices are defined by the Expert Team on Climate Change Detection and Indices, and have been investigated in other studies (Chervenkov et al., 2019; Chervenkov & Slavov, 2019). These indices are useful in diagnosing the variability of precipitation at various timescales posing them as proper metrics for impact-based assessment. We refine these indices to capture weather patterns that could cause hazards such as flood and wildfire. We conduct an evaluation of forecast models to assess their capability in discerning situations that result in the occurrence of a specific event from those that lead to its non-occurrence. As a related task, we perform a targeted analysis of model performance in regions with high risk of wildfires and floods. The following questions are answered in this study: Do seasonal forecast models have the capability to represent the variability of precipitation throughout the season? Can specific models lend themselves as superior alternatives for multi-model fusion, and thereby enhance targeted extreme event prediction in specific regions and seasons?

2. Methodology

This study examines the effectiveness of precipitation forecasts of five seasonal forecast models from the C3S project including the Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC: version 35), Deutscher Wetterdienst (DWD: version 21), Environment and Climate Change Canada (ECCC: version 3), Météo France (Météo-France: version 8), and UK Met Office (UK-Met: version 601) models in accurately predicting extreme precipitation indices in a three month lead time during the hindcast period spanning 1993 to 2016 (refer to Table S1). Validation was carried out using the fifth generation ECMWF reanalysis (ERA5) precipitation product which we refer to as "reference data" hereafter.

2.1 Data Preparation

We employed eight types of distinct climate extreme indices, following Expert Team on Climate Change Detection and Indices (https://www.wcrp-climate.org/data-etccdi) definitions, to encompass different aspects of precipitation extremes, such as event duration, intensity, and frequency (Dunn et al., 2022). We used 1mm and 10mm precipitation thresholds (representing wet days, and heavy precipitation days respectively) for the calculation of climate extreme indices. In addition to these absolute thresholds, we incorporated the 75th percentile of daily historical data in each grid from the ERA5 (reference dataset) as a secondary criterion in calculating certain indices. This approach aims to ensure the indices are reflective of local climate. For such indices, both the absolute threshold and the percentile-based criteria should be met. Combination of metrics and thresholds resulted in a comprehensive set of 14 climate extreme indices (Table 1). The assessment of seasonal skill for the models was carried out using forecast initializations on the first day of February, May, August, and November with a 3-month lead time, i.e., March-May (MAM), June-August (JJA), September-November (SON), and December-February (DJF) seasons respectively.

Table 1. Extreme indices used to evaluate seasonal forecast models.

Abbreviation	Index	Abbreviation	Index
cdd1	Maximum consecutive dry days 1mm	nwd10q75	Number of wet days 10mm and 75th percentile
cdd10	Maximum consecutive dry days 10mm	hpd	Heavy precipitation days (days with precipitation above 10mm)
cwd1	Maximum consecutive wet days 1mm	vhpd	Very heavy precipitation days (days with precipitation above 20mm)
cwd10	Maximum consecutive wet days 10mm	h1dp	Highest 1-day precipitation amount
int1	Daily pr intensity 1mm	h5dp	Highest 5-day precipitation amount

int10	Daily pr intensity 10mm	propd1	Proportion of days with precipitation at or above 1mm
nwd1q75	Number of wet days 1mm and 75th percentile	propd10	Proportion of days with precipitation at or above 10mm

The analysis was conducted on the ensemble mean for each model. All forecast models and reference data were re-gridded to a consistent one-degree resolution. For spatial aggregation, we conducted our analysis over the sixth-version of Intergovernmental Panel on Climate Change (IPCC) regions shown in Table S2 (Iturbide et al., 2020). The IPCC divides the world into major regions, each of which includes a group of countries or territories that share similar climate characteristics, geographic features, and socio-economic factors. The IPCC regions, also known as the "IPCC Regional Reporting", are a set of geographical regions used by the IPCC as a framework for understanding how climate change affects different parts of the world and to facilitate the assessment of climate change impact, vulnerability, and adaptation strategies at the regional level.

2.2 Model Evaluation

For evaluation and comparison of forecast models against reference data, we employed Percent Bias (Rudisill et al., 2024; Spies et al., 2015). Here, zero signifies a perfect alignment between model predictions and reference data, and positive-bias/negative-bias signifies over-estimation/under-estimation (Eq.1).

$$PBIAS = 100\frac{\sum_{i=1}^{N}(M_i - R_i)}{\sum_{i=1}^{N} R_i} \qquad \text{Eq.1}$$

where M_i is the model simulation and R_i is the value of the reference data at time i.

We also conducted a discriminant analysis to determine if the forecast skill varied in different sections of the precipitation distribution. To achieve this, we categorized the reference datasets in each grid into three distinct groups using upper and lower terciles: below-normal conditions (lower tercile), above-normal conditions (upper tercile), and normal conditions (middle tercile). Subsequently, we employed a random forest algorithm to perform the classification task on the forecast models in each grid. The random forest classifier was trained on 75% of the forecast data. The performance of the classifier was evaluated using the remaining 25% of the forecast data (i.e., test set). The classifier predicted the probabilities for each precipitation category (below-normal, above-normal, or normal) for each sample in the test set. These probabilities were then used to calculate the area under the Receiver Operating Characteristic (ROC) curve (ROC AUC), which measures the forecast discrimination. The ROC AUC is built based on the true positive rate (sensitivity) and false positive rate (1-specificity) obtained at various threshold values for a multi-class classification. It takes values between 0 to 1, in which a higher ROC AUC score indicates better discriminatory power of the forecasts. Here we categorized the ROC score to level of discrimination (i.e., in our case ability to discriminate extreme events from non-extreme events). The values of ROC score between 0.0 to 0.6 is considered no discrimination, 0.6 to 0.7 is considered satisfactory discrimination, 0.7 to 0.8 is considered good discrimination, 0.8 to 0.9 is considered very good discrimination, and 0.9 to 1.0 is considered excellent discrimination (Mandrekar, 2010). Kendall's Tau rank correlation analysis is used to measure the strength of the relationship between climatic indices extracted from forecast models and reference data aggregated over each IPCC regions (Sen, 1968).

2.3 Impact-based framework

The IPCC regions which were vulnerable to wildfire and flooding were identified by categorizing them based on the proportion of farmland (Friedl & Sulla-Menashe, 2022), proportion of burned areas (Chuvieco et al., 2018; Lizundia-Loiola et al., 2020), percentage of flood-affected zones (Tellman et al., 2021), proportion of built-up regions (Gong et al., 2020), and population density (Schiavina et al., 2023) (see Figure S1). We specifically focused on regions that not only had an elevated risk of wildfire (flood) exposure, but also had a significant built-up area and population density. These selected regions were earmarked for additional analysis.

We carefully selected relevant extreme indices pertinent to the corresponding climate-induced hazard in each region. For the regions with wildfire as a prominent natural hazard, we selected maximum consecutive dry days 1mm (cdd1), and proportion of days with precipitation at or above 1mm (propd1) indices as the relevant indicators of a weather condition conducive to or preventive of wildfire. For regions with a high risk of flooding, we selected number of wet days with 10mm precipitation and 75th percentile of the reference data (nwd10q75) and heavy precipitation days (hpd) indices as the relevant indicators. We also determined the season with the highest occurrence rate for each specific hazard and region. In every region, we identified the top-performing model in terms of predictive accuracy using the following selection criteria. Initially, we prioritized models with a combination of higher correlation (statistically significant) and lower bias. If multiple models demonstrated similar high performance, we utilized Taylor diagrams to select models that aligned more closely with the reference data across various performance metrics. Our evaluation then focused on assessing the models' forecast skill with respect to relevant indices in the identified hazard-prone seasons for each region. By adopting this impact-based approach, we aimed to pinpoint the most suitable models and indices for each climate-induced hazard, enabling more effective and tailored climate risk assessments.

3. Results

3.1 Global Analysis of Model Bias

The analysis of Percent Bias at a global scale across the four seasons reveals a consistent tendency of underestimation in all forecast models for most of the extreme wet precipitation indices (Figure 1 and Figure S2), but the performance for other indices is rather mixed. The cdd1 index, which measures the maximum consecutive dry days at 1mm threshold, shows negative bias for most but not all regions, while cwd1 index, which measures the maximum consecutive wet days at 1mm threshold, shows positive bias for all regions implying that the models predict more precipitation days compared to reference dataset (Figure 1). This is while cdd10 shows positive bias values and cwd10 shows negative bias values, implying that while models are capturing more wet days, they mostly capture light precipitation, and they do not properly capture extreme events with consecutive days of heavy precipitation. This is aligned with the propd1 index showing positive biases (higher number of precipitation days) while propd10 index showing negative biases (Figure 1 and Figure S2). This pattern is consistent with the findings of regional C3S assessment studies, for example in Africa (Gebrechorkos et al., 2022). These observations underscore the limitations in forecast capabilities for accurately modelling persistent wet and dry periods. Introduction of secondary constraints (i.e., 75th of the reference data) to the indices increases the bias, signifying the models' limitation in capturing severe extreme events.

The bias values for int1 index – daily precipitation more than 1mm – are higher than those of int10 across all seasons and models. This is also because of the models' tendency to capture more wet days with at least 1mm precipitation. Focusing on h1dp and h5dp indices, forecasts are able to model the extreme precipitation events that happen over 5 consecutive days more skillfully than those extending one day (Figure S2). Figure 1 and Figure S2 show that the CMCC, DWD, and ECCC models demonstrate relatively lower ability to capture extreme rainfall events within the extratropical IPCC regions compared to the UK-Met and Météo-France models across the four seasons. However, in the tropical and northern subtropical regions, all models (especially UK-Met and Météo-France) exhibit superior performance (lower bias) in capturing extreme events, compared to extratropical regions. This is attributed to the model's predictive skill in grasping large-scale teleconnection patterns (Giuntoli et al., 2022). Refer to section S1 for additional information about global precipitation anomalies patterns across the five forecast models.

For indices measured in terms of the number of days, we observed larger bias compared to those representing total rainfall, indicating limitations of the models in accurately replicating the variation of precipitation throughout the season (Figure 1 and Figure S2). Indices that represent magnitude and intensity of precipitation (i.e., precipitation intensity, highest 1-day precipitation amount, and highest 5-day precipitation amount) exhibit lower biases, suggesting the model's skill in simulating total seasonal precipitation. The UK-Met and Météo-France models exhibit higher skill in capturing extreme events, demonstrating favorable performance across various regions when considering 75th percentile threshold levels.

We opted to utilize the percentiles of reference data climatology as a consistent benchmark across all models, rather than relying on the percentile of forecast models. This choice ensures intercomparability in models' ability to depict climate patterns. Figure S4 illustrates the difference in Percent Bias when using percentiles derived from the reference data climatology versus those from the forecast model, specifically for the UK-Met and Meteo-France models. Further comparison reveals that the Meteo-France model exhibits larger biases than the UK-Met model across most regions globally, indicating its weaker representation of climate patterns in terms of Percent Bias.

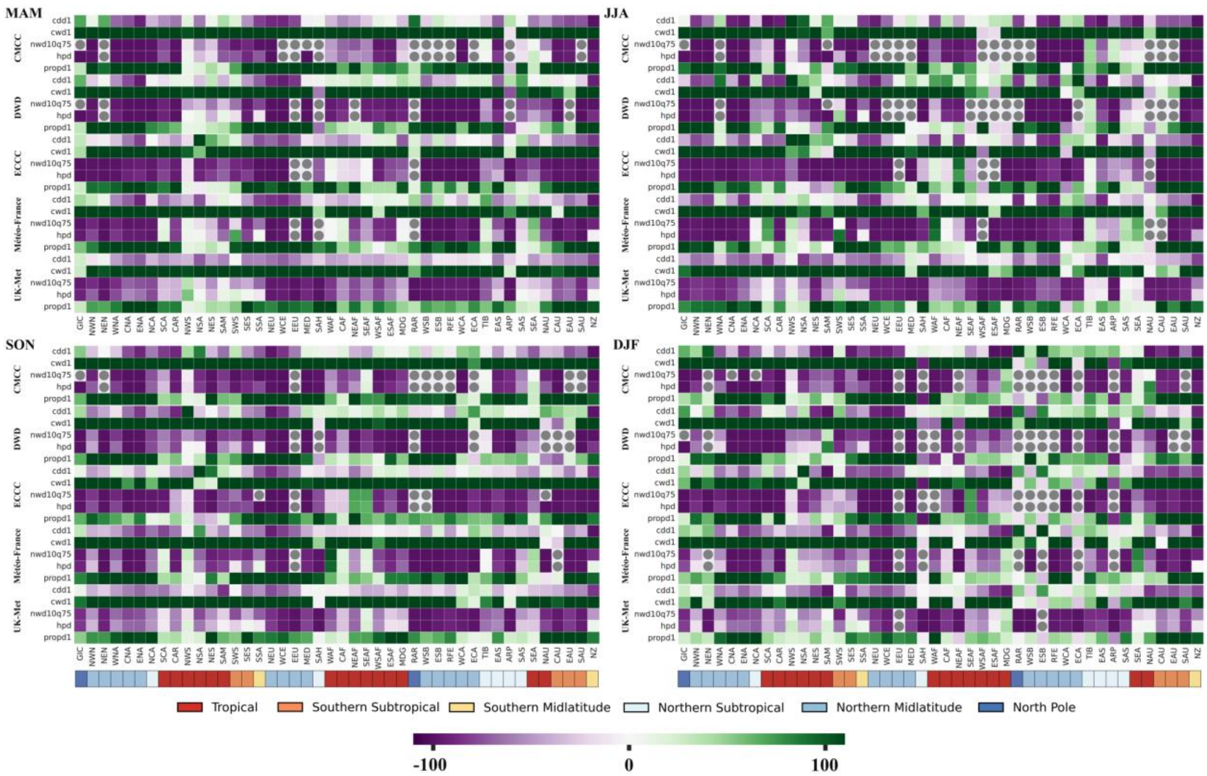

Figure 1. Percent Bias of the five C3S models for a) MAM, b) JJA, c) SON, and d) DJF seasons. Grids with grey circles shows the regions where reference data indicated the existence of extreme events while the forecast models could not capture any events satisfying the index requirements. Grids with no values reflect regions that both reference data and forecasts did not capture any events satisfying the index requirements. The colour bar is constrained to the range of -100 to 100 for visualization purposes. This range was chosen to enhance visibility of variations in regions with small bias. Actual values occasionally exceed this range. IPCC regions are color-coded based on their location within respect to latitudinal zones. See Table S2 for names of IPCC regions.

3.2 Global Analysis of ROC Scores

We use discrimination analysis to assess how well the year-to-year variations in the forecasted precipitation match those in the observations. Measurements of ROC score in Figure 2 and Figures S5 show superior performance of forecast models in the tropical and subtropical regions located in Atlantic, Indian ocean, and west Pacific regions (Guimarães et al., 2021; Jie et al., 2017). The skill level, however, varies across different models and seasons over Africa. Notably, the Météo-France and UK-Met models exhibit superior performance during the SON and MAM seasons (i.e., indices with satisfactory, good, very good, and excellent discriminations are more frequent). The exceptional performance of the Météo-France model in African regions has been the subject of discussion in prior studies (Gebrechorkos et al., 2022). Furthermore, the UK-Met model demonstrates a higher level of skill compared to the other four models in predicting extreme events in several Australian regions. This elevated skill of the UK-Met model is particularly pronounced during the MAM season whereas in JJA, SON, and DJF the skill drops dramatically. The lower performance of ACCESS-S1 model (which is the same model used in UK-Met but with different ensemble generation scheme, ensemble size and the configuration of the system for operational forecasting) over Australia during southern hemisphere summer (DJF) is also concluded in other studies (King et al., 2020).

The prevalence of grids with no discrimination ROC categories is more pronounced in extratropical regions, possibly due to the inherent lower predictability of extratropical variations, and model limitations in representing interactions between tropical and extratropical regions, as well as land surface processes (De Andrade et al., 2019). Notably, the CMCC, DWD, and ECCC models fail to detect any extreme event in many extratropical regions, as indicated by the absence of discrimination categories in Figure S5. This disparity is particularly conspicuous when compared to the UK-Met and Météo-France models.

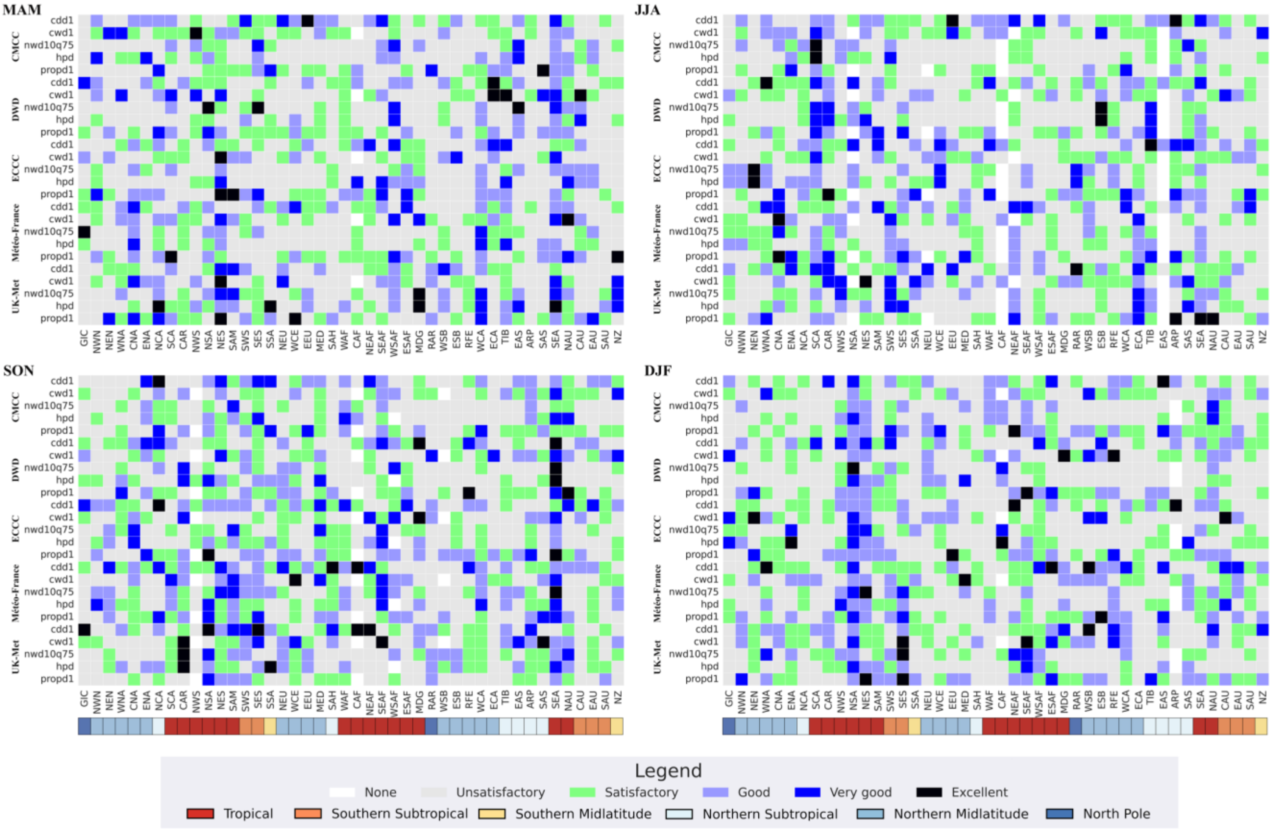

Figure 2. Discrimination levels using categorized ROC score of the five models for a) MAM, b) JJA, c) SON, and d) DJF seasons. Grids that are shaded in white represent regions that either or both reference data and model did not capture any events satisfying the index requirements. IPCC regions are color-coded based on their location within respect to latitudinal zones.

3.3 Wildfire-prone Regions: Targeted Forecast Performance Analysis

Many scientific investigations have underscored the notable influence of climatic patterns on the initiation of wildfires (Sharma et al., 2022; Turco et al., 2023). Extended periods of elevated temperatures devoid of precipitation events establish an environment conducive to fire ignition and propagation, intensifying the combustibility of vegetated areas (Alizadeh et al., 2021, 2023). As the duration of consecutive dry days (days without rainfall or with rainfall below a specific threshold) extends, the moisture content of fuel diminishes, increasing its susceptibility to ignition (Abatzoglou & Williams, 2016). Refer to section S2 for detailed analysis of wildfire-related indices on a global scale, while in the following we will focus on specific regions with wildfire as a prominent natural hazard.

We now focus on the four regions where the wildfire is a prominent natural hazard: Northern Australia (NAU), South-Eastern Africa (SEAF), Western North America (WNA), and Northern South America (NSA) regions. Within each region, a particular season characterized by an elevated likelihood of wildfire incidence is designated for subsequent analysis. In Northern Australia, the peak period for wildfire aligns with the dry SON season. From August to December, many regions of Southern Africa experience the onset of their wildfire season therefore we selected the SON season for further analysis. In the United States, wildfire activity is a year-round concern, but the most severe wildfires arise during the summer months (JJA season), particularly in the western regions. In Latin America, the fire season typically commences at the end of January and extends through April (DJF season).

Focusing on the NAU region and using the maximum number of consecutive dry days index, all models except for CMCC demonstrate a notable correlation with the reference data (Figure 3a). Furthermore, Météo-France and ECCC display lower bias compared to other models (Figure 3b). However, in the Taylor diagram presented in Figure 3c, the ECCC model establishes its supremacy over Météo-France by exhibiting a standard deviation that is more closely aligned with the reference data. The overestimation of precipitation (and consequently underestimation of dry days) in CMCC and UK-Met models over NAU region is visible in Figure 4 where they exceed the 1mm threshold earlier and with steeper slope compared to other models resulting in the underestimation of cdd1 index.

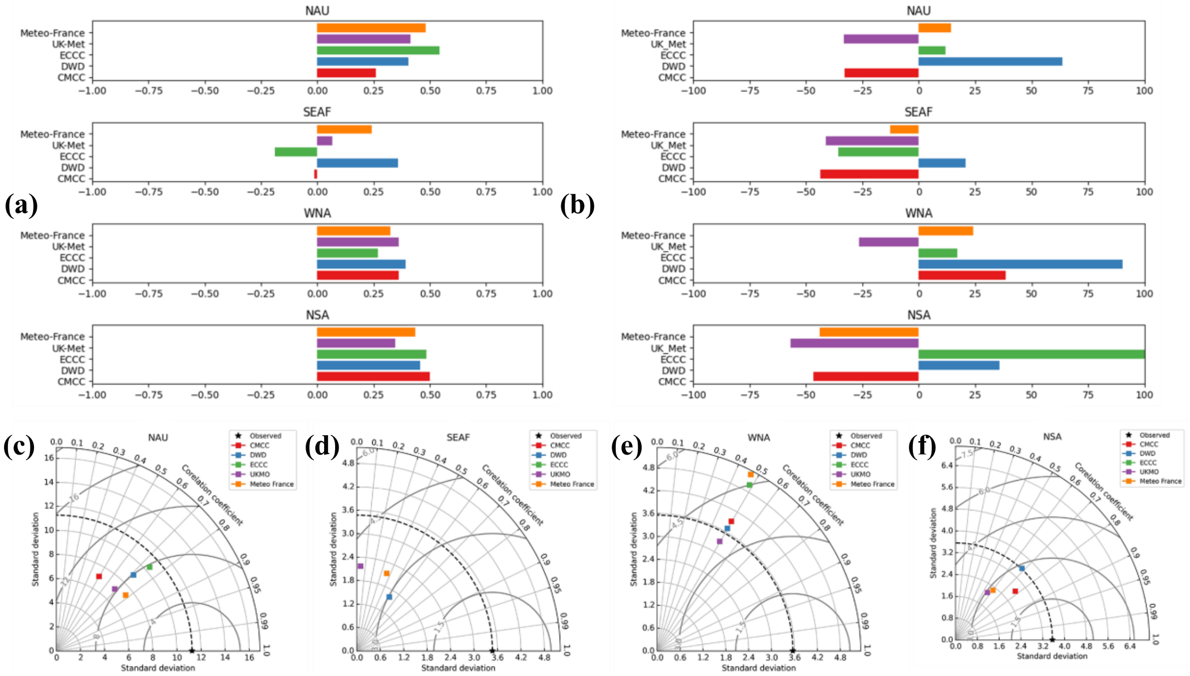

Figure 3. Performance metrics for the maximum consecutive dry days index with 1mm precipitation threshold (cdd1): a) Kendall's Tau coefficient, b) Percent Bias, and Taylor diagram for c) NAU, d) SEAF, e) WNA, and f) NSA regions respectively.

We extended our model selection framework to other three regions and the main findings reveal distinct model performance variations in different regions. The ECCC model is particularly strong in forecasting consecutive dry days in the NAU region and closely tracks reference data. In contrast, the DWD model emerges as the top performer in the SEAF and NSA regions, exhibiting the highest correlation, lower bias, and lower root mean square error. The UK-Met model excels in the WNA region, demonstrating a close match with the reference data's standard deviation. These variations in model performance are attributed to their abilities in simulating significant large-scale climate variabilities such as ENSO, IOD, and north Australian SSTs. Refer to section S3 for further details regarding the analysis of proportion of days with precipitation acceding 1mm threshold (propd1) index in wildfire prone regions.

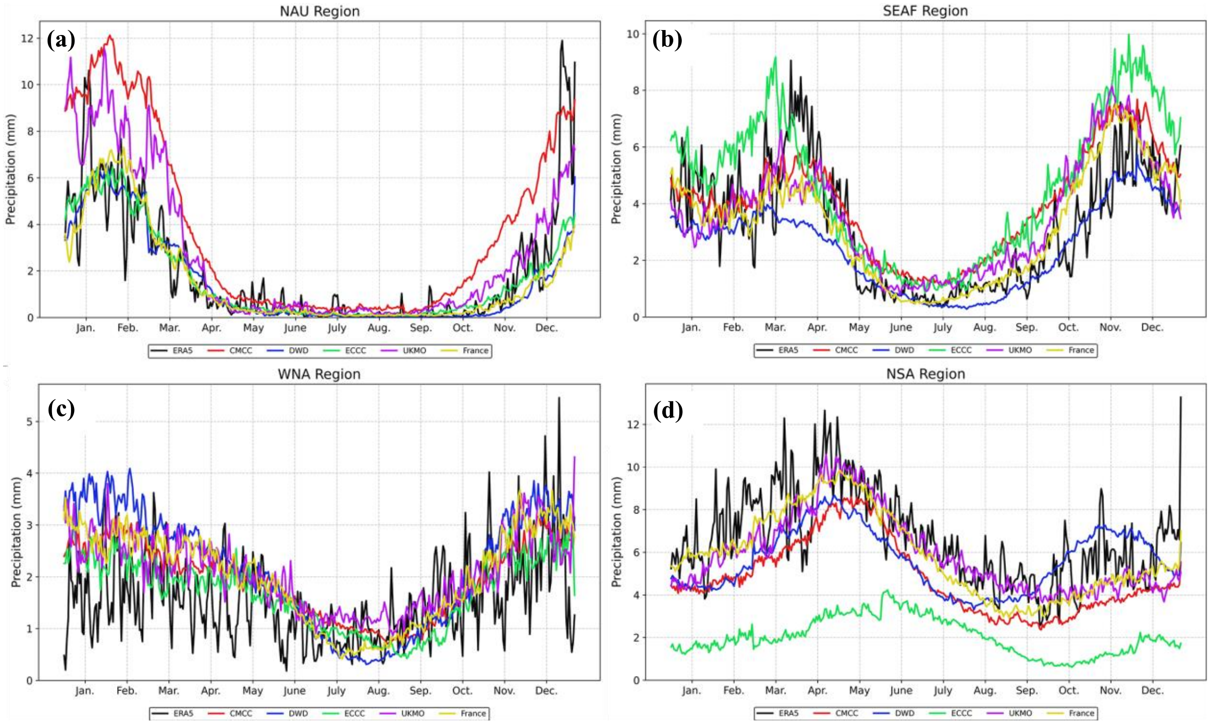

Figure 4. Annual climatology time series of the precipitation for five C3S models and the ERA5 reference data over NAU, SEAF, WNA, and NSA regions.

3.4 Flood-prone Regions: Targeted Forecast Performance Analysis

Consecutive occurrences of extreme precipitation over successive days can significantly elevate the probability of widespread flooding, and since extreme events like heavy precipitation days are becoming more frequent, reliable forecast models are pivotal. Many investigations have documented instances of substantial flooding due to consecutive multi-day extreme precipitation (Ávila et al., 2016; Du et al., 2022; Rivoire et al., 2023). In 2021, extreme precipitation events across central Europe caused severe flooding in many regions, resulting in more than 200 fatalities and significant damage to infrastructure (Tradowsky et al., 2023). To assess the capabilities of the C3S models across IPCC regions, where flooding is a predominant natural hazard, we employed the heavy precipitation days index (hpd) and number of wet days with 10mm precipitation threshold index, also exceeding the 75th percentiles of the reference dataset (nwd10q75) for further analysis. Refer to section S4 for detailed analysis regarding the analysis of flood-related indices on a global scale, while in the following we focus on specific regions with flood as a prominent natural hazard.

In the South-East Asia (SEA) monsoon region during JJA (flood season), the UK-Met demonstrates a superior performance compared to other models, exhibiting notably high correlation and lower bias values (Figure 5a and 5b). Although, the Taylor diagram indicates that UK-Met exhibits a larger standard deviation value compared to other models, the markedly lower bias values make this model the optimal choice here (Figure 5c). This is also evident in the Figure 6a where Météo-France shows overestimation of precipitation, ECCC shows underestimation, while UK-Met, DWD, and CMCC follow the reference precipitation very closely. Overall, in this region the prediction skill is mostly higher in the pre-monsoon (April–May) and post-monsoon (October–November) seasons, while during the monsoon season (JJA) forecast skill is lower because of the monsoon influences on precipitation predictability (Wanthanaporn et al., 2023).

In the Western and Central Europe (WCE) region during the DJF flooding season, the ECCC model exhibits higher significant correlation values with reference data compared to other models (Figure 5). However, all models struggle to adequately capture reference data variations, as indicated by high RMSE values and low correlation coefficients. In the South Asia (SAS) region during the JJA season, the UK-Met and CMCC models demonstrate higher correlation values. The UK-Met model outperforms others by exhibiting smaller bias. Therefore, the UK-Met model is favored for the SAS region. In the Central North America (CNA) region during the JJA season, both the UK-Met and Météo-France models exhibit significant correlation coefficients, while all models display large bias values. Once again, the UK-Met model stands out due to its lower bias compared to the other models.

Upon eliminating the constraint associated with the 75th percentile of the reference data in predicting heavy precipitation days, there is an observable reduction in bias values across SEA and SAS regions for the hpd index (Figure S9a and S9b). The correlation values are, however, very similar to those of nwd10q75 index. with removal of the 75^{th} percentile threshold, the standard deviation values also become more aligned with the reference data (Figure S9c and S9f). Despite these changes, the order of outperforming models remains consistent.

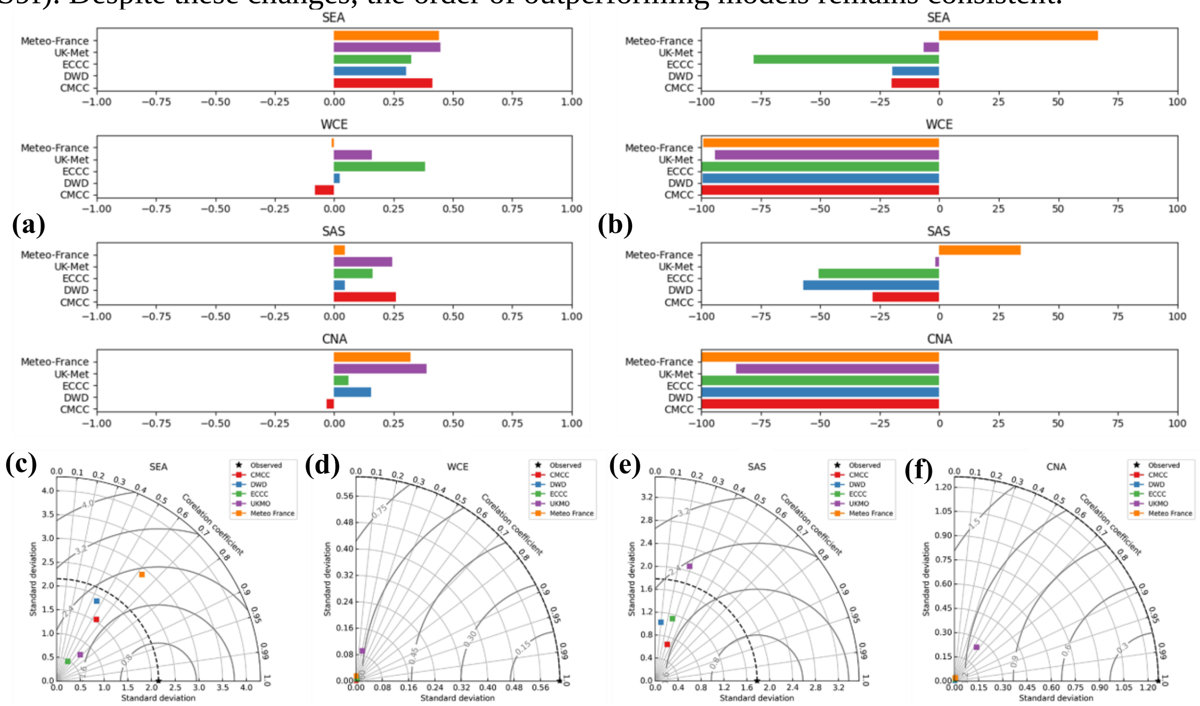

Figure 5. Model performance with respect to the number of heavy precipitation days exceeding 10 mm and the 75th percentile of the reference data (nwd10q75): a) Kendall's Tau coefficient, b) Percent Bias, and Taylor diagram for c) SEA, d) WCE, e) SAS, and f) CNA regions, respectively.

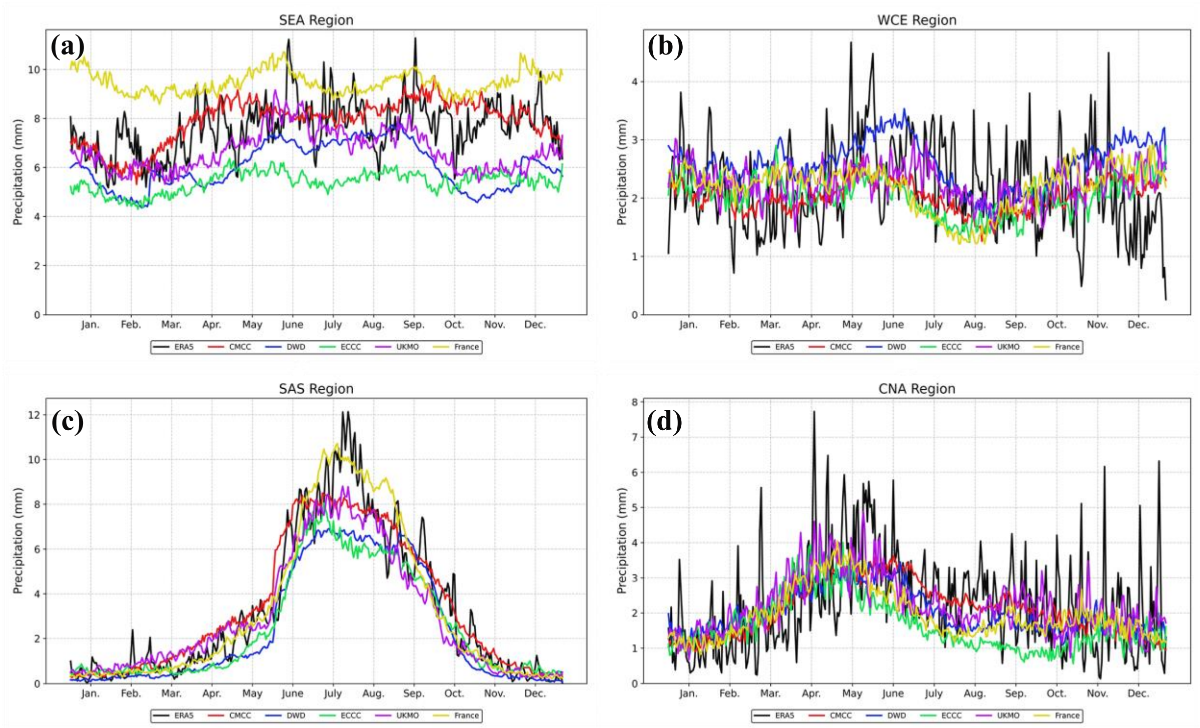

Figure 6. Annual climatology time series of the precipitation for five C3S models and the ERA5 dataset over a) SEA, b) WCE, c) SAS, and d) CNA region.

3.5 Populated and Densely Built-up Regions: Targeted Forecast Performance Analysis

We now focus on two regions where the population and built-up density is very high: East Asia (EAS), and South Asia (SAS) regions. Within each region, a particular season characterized by an elevated likelihood of consecutive days with rainfall occurrence (which has a significant impact on urban areas) was selected for further analysis. Moderate rainfall events may not be inherently destructive, but consecutive days of precipitation, even with low amounts, can have devastating impacts on the environment and urban areas. In EAS region with high population density, consecutive days of moderate-intensity rainfall serve as significant triggers for geological hazards such as landslides and mudslides. These events can lead to immense damages to lives and property (Zheng et al., 2020).

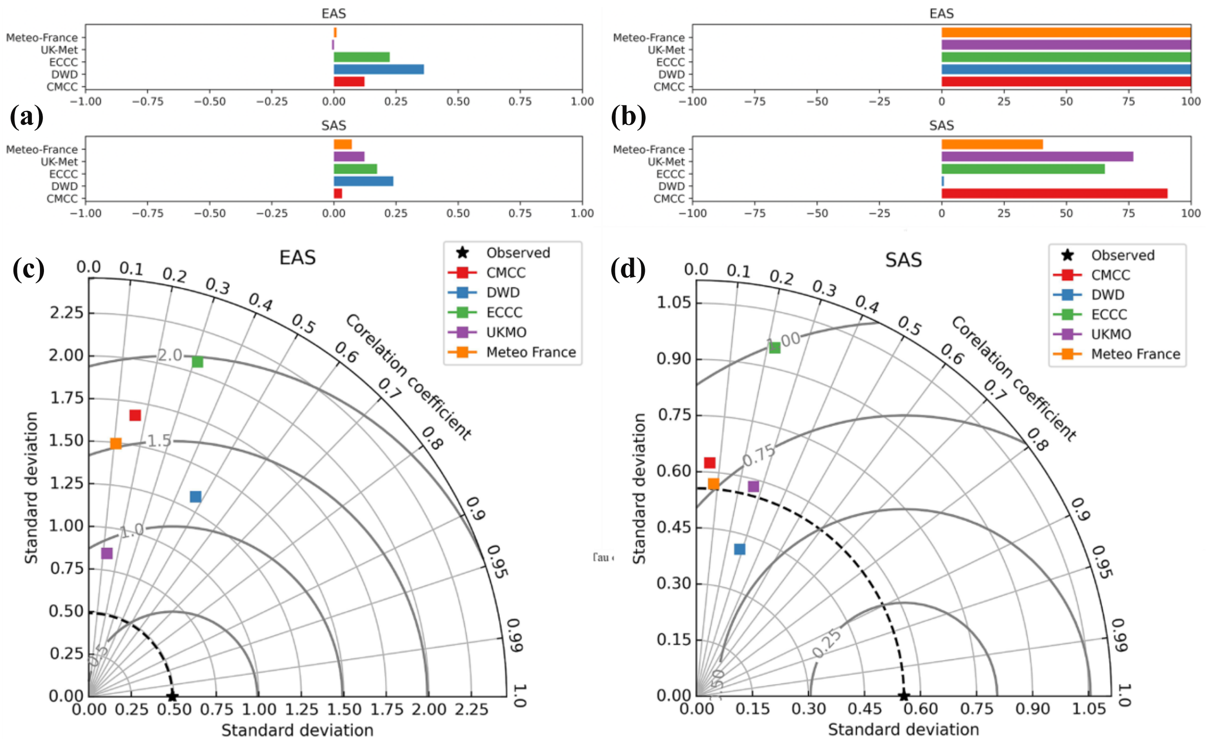

Figure 7 Model performance with respect to the consecutive wet days with 1mm precipitation threshold (cwd1): a) Kendall's Tau coefficient, b) Percent Bias, and Taylor diagram for c) EAS, d) SAS regions, respectively.

In East Asia (EAS) and during the DJF season, the DWD model demonstrates a superior performance compared to other models, exhibiting notably high correlation and lower bias values (Figure 7a and 7b). Although, the Taylor diagram indicates that the DWD model exhibits a larger standard deviation compared to UK-Met in the EAS region, still the markedly higher correlation values make this model the optimal choice here (Figure 7c). In the SAS region, the standard deviation of the DWD model is not aligned well with that of the reference data but considering its high correlation and lower bias values compared to the rest of models, DWD is the best model in representing underlaying processes of climate patterns in this region (Figure 7d).

3.6 Model Effectiveness in Process Representation Across Seasons and Regions

Over all five models, our findings reveal that with increase in the precipitation threshold the model's bias increases, suggesting a lack of skill in modelling severe precipitation events. Correlation scores are lower in extratropical regions as compared to the tropical regions, likely due to the inherent unpredictability of extratropical atmospheric variability and model limitations in replicating land surface processes and tropical-extratropical interactions, including the Pacific-South American (PSA) pattern and the Pacific-North American (PNA) pattern, both of which can be influenced by ENSO and the MJO (De Andrade et al., 2019). This is depicted in Figure 8, where extratropical regions struggled to accurately represent the underlying processes of climate and weather patterns (i.e., exhibiting statistically significant correlations while maintaining lower bias values) for most indices. The figure also highlights the superior performance of UK-Met and Météo-France in representing these processes across all four seasons. Refer to section S5 for further details regarding the Model effectiveness of models in process representation.

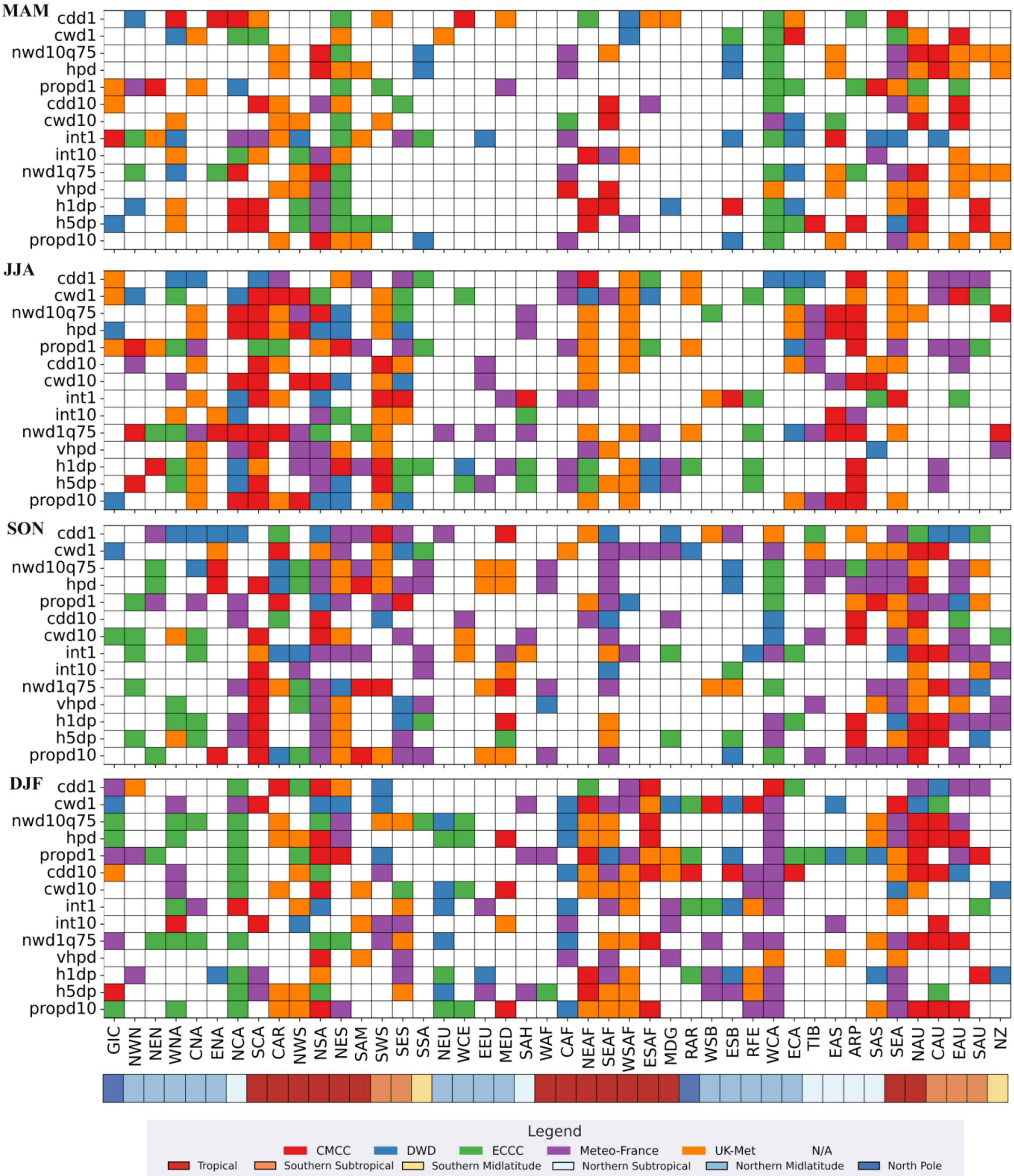

Figure 8. Models that best represent the underlying processes based on statistically significant Kendall's Tau at the 0.05 level and Percent Bias over IPCC regions across 14 climate extreme indices for a) MAM, b) JJA, c) SON, and d) DJF seasons. IPCC regions are color-coded based on their location within respect to latitudinal zones.

4. Summary and Discussion

This study's primary objective is to assess the performance of five C3S seasonal forecast models in predicting extreme precipitation events spanning the period from 1993 to 2016. To achieve this, the study assesses 14 extreme precipitation indices defined by the Expert Team on Climate Change Detection and Indices group. These indices are established based on specific precipitation thresholds of 1mm, and 10mm, as well as locally-specific 75th percentile of reference data. The ERA5 reanalysis precipitation dataset is used as reference.

Our goal is to identify the most reliable models for targeted precipitation risk assessments. We introduce an impact-based forecast model assessment framework designed to evaluate the models' effectiveness in predicting extreme weather events that have the potential to instigate hazardous conditions such as floods and wildfires. We employ performance metrics, including Percent Bias and the Kendall Tau Rank Correlation Score, to gauge the models' skill in representing the underlying processes that govern climate and weather pattern and generate extreme weather events. Furthermore, we evaluate the discrimination capacity of models in discerning extreme events from non-events.

While post-processing techniques play a crucial role in enhancing the reliability of seasonal forecast models, our study primarily aims to identify models that best represent underlying processes in seasonal climate and weather patterns. By evaluating raw forecasts, we aim to uncover models' inherent skills and deficiencies, free from the influence of statistical post-processing. In other words, we use raw model forecasts for our assessment purposes, and avoid statistical postprocessing that convolutes process-based forecasts with statistical correction. This choice is specifically useful in the face of extremes repeatedly breaking records in a changing climate, given statistical methods' limitations in reproducing out-of-sample data. This approach enables us to detect key differences between forecast models and assess their abilities in capturing extreme events.

One key finding is the consistent bias toward underestimation of most of the wet extreme climate indices for all models. Nevertheless, the UK-Met and Météo-France models are found to outperform others, which is consistent with the literature (De Andrade et al., 2019; McAdam et al., 2022). Despite the prevalent bias, statistically significant correlation are found in tropical and subtropical regions, indicating that the models can reasonably capture the variability of events even if they underperform in terms of the magnitude of the extremes (F. Vitart et al., 2017).

We assess model skill in forecasting extreme events in regions susceptible to cascading natural hazards like wildfires and floods. To evaluate performance in areas prone to wildfires, we employed indices such as maximum consecutive dry days with precipitation below 1mm (cdd1), and proportion of days with precipitation at or above 1mm (propd1). For flood-prone zones, we used number of wet days with precipitation above10mm and the 75th percentile of the reference data (nwd10q75), as well as heavy precipitation days (days with precipitation above 10 mm, hpd) as primary indices. Also, for areas with high population density and elevated built-up density, consecutive wet days with precipitation above 1mm (cwd1) threshold is used for evaluations.

In the context of wildfire risk analysis, notable differences in predictive capacities are observed, with specific models showcasing prowess in different regions and for different extreme precipitation indices. In the Northern Australia region, Météo-France and ECCC models display robust performance in predicting consecutive dry days. In the Southern Africa region, the DWD model emerged as a frontrunner for predicting extreme precipitation events. The UK-Met model shows promising results for Western North America. Lastly, the DWD model shows good performance for the North South America region.

For flood-prone regions, the UK-Met model demonstrates superior predictive capabilities in the South-East Asia during the monsoon season (JJA). In Western and Central Europe during the flood season (DJF), the ECCC model excels with notable correlation and comparable bias, despite challenges in capturing reference data variations. In South Asia and during the JJA season, the UK-Met and CMCC models excel, with the UK-Met showing favourable correlation and lower bias.

Forecast models showed superior results for tropical regions than others globally. Lower correlation scores in extratropical regions can be attributed to the inherent unpredictability of extratropical variability and the errors stemming from model deficiencies in representing teleconnections (De Andrade et al., 2019). Our analysis of extreme precipitation indices across multiple models reveals that higher precipitation thresholds used for model evaluation correspond to increased bias, indicating a lack of skill in modelling severe precipitation events. Finally, our results emphasized the superiority of UK-Met and Météo-France models throughout all four seasons. The ECCC and CMCC models demonstrate effectiveness, following UK-Met and Météo-France, across specific indices and regions. Model fusion emerges as a successful approach for predicting extreme events across different seasons. These findings highlight the effectiveness of the impact-based framework in thoroughly assessing forecast skills in representing climate and weather patterns across various models and seasons.

Open Research

The data used in this study were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) Copernicus Climate Change Service, specifically from the ERA5 reanalysis dataset and C3S seasonal forecasts. These datasets are publicly available through the Copernicus Climate Data Store (CDS) at https://cds.climate.copernicus.eu under an Open Data Commons Attribution 4.0 International (ODC-BY 4.0) license. To access the data, users can register for a free account on the Copernicus Climate Data Store platform and follow the provided guidelines for data retrieval. The specific seasonal model version numbers used in this study are detailed in the supplementary material of the paper.

References

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, *113*(42), 11770–11775. https://doi.org/10.1073/pnas.1607171113

AghaKouchak, A., Huning, L. S., Chiang, F., Sadegh, M., Vahedifard, F., Mazdiyasni, O., et al. (2018). How do natural hazards cascade to cause disasters? *Nature*, *561*(7724), 458–460. https://doi.org/10.1038/d41586-018-06783-6

AghaKouchak, A., Huning, L. S., Sadegh, M., Qin, Y., Markonis, Y., Vahedifard, F., et al. (2023). Toward impact-based monitoring of drought and its cascading hazards.

Nature Reviews Earth & Environment, *4*(8), 582–595. https://doi.org/10.1038/s43017-023-00457-2

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, *17*(3), 1161–1175. https://doi.org/10.5194/hess-17-1161-2013

Alizadeh, M. R., Abatzoglou, J. T., Luce, C. H., Adamowski, J. F., Farid, A., & Sadegh, M. (2021). Warming enabled upslope advance in western US forest fires. *Proceedings of the National Academy of Sciences*, *118*(22), e2009717118. https://doi.org/10.1073/pnas.2009717118

Alizadeh, M. R., Abatzoglou, J. T., Adamowski, J., Modaresi Rad, A., AghaKouchak, A., Pausata, F. S. R., & Sadegh, M. (2023). Elevation-dependent intensification of fire danger in the western United States. *Nature Communications*, *14*(1), 1773. https://doi.org/10.1038/s41467-023-37311-4

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., & Pineda, L. (2020). Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. *Hydrology and Earth System Sciences*, *24*(2), 535–559. https://doi.org/10.5194/hess-24-535-2020

Assessment of Intraseasonal to Interannual Climate Prediction and Predictability. (2010) (p. 12878). Washington, D.C.: National Academies Press. https://doi.org/10.17226/12878

Ávila, A., Justino, F., Wilson, A., Bromwich, D., & Amorim, M. (2016). Recent precipitation trends, flash floods and landslides in southern Brazil. *Environmental Research Letters*, *11*(11), 114029. https://doi.org/10.1088/1748-9326/11/11/114029

Baldwin, M. P., Stephenson, D. B., Thompson, D. W. J., Dunkerton, T. J., Charlton, A. J., & O'Neill, A. (2003). Stratospheric Memory and Skill of Extended-Range Weather Forecasts. *Science*, *301*(5633), 636–640. https://doi.org/10.1126/science.1087143

Becker, E., Den Dool, H. V., & Zhang, Q. (2014). Predictability and Forecast Skill in NMME. *Journal of Climate*, *27*(15), 5891–5906. https://doi.org/10.1175/JCLI-D-13-00597.1

Chervenkov, H., & Slavov, K. (2019). STARDEX and ETCCDI Climate Indices Based on E-OBS and CARPATCLIM: Part Two: ClimData in Use. In G. Nikolov, N. Kolkovska, & K. Georgiev (Eds.), *Numerical Methods and Applications* (Vol. 11189, pp. 368–374). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-10692-8_41

Chervenkov, H., Slavov, K., & Ivanov, V. (2019). STARDEX and ETCCDI Climate Indices Based on E-OBS and CARPATCLIM: Part One: General Description. In G. Nikolov, N. Kolkovska, & K. Georgiev (Eds.), *Numerical Methods and Applications* (Vol. 11189, pp. 360–367). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-10692-8_40

Chuvieco, E., Pettinari, M. L., Lizundia-Loiola, J., Storm, T., & Padilla Parellada, M. (2018). ESA Fire Climate Change Initiative (Fire_cci): MODIS Fire_cci Burned Area Pixel product, version 5.1 (Version 3.1) [Application/xml]. Centre for Environmental Data Analysis (CEDA). https://doi.org/10.5285/58F00D8814064B79A0C49662AD3AF537

Committee on Developing a U.S. Research Agenda to Advance Subseasonal to Seasonal Forecasting, Board on Atmospheric Sciences and Climate, Ocean Studies Board, Division on Earth and Life Studies, & National Academies of Sciences, Engineering, and Medicine. (2016). *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts* (p. 21873). Washington, D.C.: National Academies Press. https://doi.org/10.17226/21873

De Andrade, F. M., Coelho, C. A. S., & Cavalcanti, I. F. A. (2019). Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dynamics*, *52*(9–10), 5451–5475. https://doi.org/10.1007/s00382-018-4457-z

Du, H., Donat, M. G., Zong, S., Alexander, L. V., Manzanas, R., Kruger, A., et al. (2022). Extreme Precipitation on Consecutive Days Occurs More Often in a Warming Climate. *Bulletin of the American Meteorological Society*, *103*(4), E1130–E1145. https://doi.org/10.1175/BAMS-D-21-0140.1

Dunn, R. J. H., Donat, M. G., & Alexander, L. V. (2022). Comparing extremes indices in recent observational and reanalysis products. *Frontiers in Climate*, *4*, 989505. https://doi.org/10.3389/fclim.2022.989505

Friedl, M., & Sulla-Menashe, D. (2022). MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD12Q1.061

Gebrechorkos, S. H., Pan, M., Beck, H. E., & Sheffield, J. (2022). Performance of State-of-the-Art C3S European Seasonal Climate Forecast Models for Mean and Extreme Precipitation Over Africa. *Water Resources Research*, *58*(3). https://doi.org/10.1029/2021WR031480

Giuntoli, I., Fabiano, F., & Corti, S. (2022). Seasonal predictability of Mediterranean weather regimes in the Copernicus C3S systems. *Climate Dynamics*, *58*(7–8), 2131–2147. https://doi.org/10.1007/s00382-021-05681-4

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., et al. (2020). Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment*, *236*, 111510. https://doi.org/10.1016/j.rse.2019.111510

Guimarães, B. D. S., Coelho, C. A. D. S., Woolnough, S. J., Kubota, P. Y., Bastarz, C. F., Figueroa, S. N., et al. (2021). An inter-comparison performance assessment of a Brazilian global sub-seasonal prediction model against four sub-seasonal to seasonal

(S2S) prediction project models. *Climate Dynamics*, *56*(7–8), 2359–2375. https://doi.org/10.1007/s00382-020-05589-5

Hao, Z., Singh, V. P., & Xia, Y. (2018). Seasonal Drought Prediction: Advances, Challenges, and Future Prospects. *Reviews of Geophysics*, *56*(1), 108–141. https://doi.org/10.1002/2016RG000549

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., et al. (2020). An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*, *12*(4), 2959–2970. https://doi.org/10.5194/essd-12-2959-2020

Jie, W., Vitart, F., Wu, T., & Liu, X. (2017). Simulations of the Asian summer monsoon in the sub-seasonal to seasonal prediction project (S2S) database: Simulations of Asian Summer Monsoon in the S2S Database. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2282–2295. https://doi.org/10.1002/qj.3085

Khorshidi, M. S., Dennison, P. E., Nikoo, M. R., AghaKouchak, A., Luce, C. H., & Sadegh, M. (2020). Increasing concurrence of wildfire drivers tripled megafire critical danger days in Southern California between1982 and 2018. *Environmental Research Letters*, *15*(10), 104002. https://doi.org/10.1088/1748-9326/abae9e

King, A. D., Hudson, D., Lim, E., Marshall, A. G., Hendon, H. H., Lane, T. P., & Alves, O. (2020). Sub-seasonal to seasonal prediction of rainfall extremes in Australia. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2228–2249. https://doi.org/10.1002/qj.3789

Kumar, A., & Zhu, J. (2018). Spatial Variability in Seasonal Prediction Skill of SSTs: Inherent Predictability or Forecast Errors? *Journal of Climate*, *31*(2), 613–621. https://doi.org/10.1175/JCLI-D-17-0279.1

Lau, W. K.-M., & Waliser, D. E. (2012). *Intraseasonal Variability in the Atmosphere-Ocean Climate System*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13914-7

Lizundia-Loiola, J., Otón, G., Ramo, R., & Chuvieco, E. (2020). A spatio-temporal active-fire clustering approach for global burned area mapping at 250 m from MODIS data. *Remote Sensing of Environment*, *236*, 111493. https://doi.org/10.1016/j.rse.2019.111493

Mallakpour, I., Sadeghi, M., Mosaffa, H., Akbari Asanjan, A., Sadegh, M., Nguyen, P., et al. (2022). Discrepancies in changes in precipitation characteristics over the contiguous United States based on six daily gridded precipitation datasets. *Weather and Climate Extremes*, *36*, 100433. https://doi.org/10.1016/j.wace.2022.100433

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

McAdam, R., Masina, S., Balmaseda, M., Gualdi, S., Senan, R., & Mayer, M. (2022). Seasonal forecast skill of upper-ocean heat content in coupled high-resolution systems. *Climate Dynamics*, *58*(11–12), 3335–3350. https://doi.org/10.1007/s00382-021-06101-3

Modaresi Rad, A., Kreitler, J., Abatzoglou, J. T., Fallon, K., Roche, K. R., & Sadegh, M. (2022). Anthropogenic stressors compound climate impacts on inland lake dynamics: The case of Hamun Lakes. *Science of The Total Environment*, *829*, 154419. https://doi.org/10.1016/j.scitotenv.2022.154419

Moron, V., & Robertson, A. W. (2020). Tropical rainfall subseasonal-to-seasonal predictability types. *Npj Climate and Atmospheric Science*, *3*(1), 4. https://doi.org/10.1038/s41612-020-0107-3

Nobakht, M., Saghafian, B., & Aminyavari, S. (2021). Skill Assessment of Copernicus Climate Change Service Seasonal Ensemble Precipitation Forecasts over Iran. *Advances in Atmospheric Sciences*, *38*(3), 504–521. https://doi.org/10.1007/s00376-020-0025-7

Rivoire, P., Martius, O., Naveau, P., & Tuel, A. (2023). Assessment of subseasonal-to-seasonal (S2S) ensemble extreme precipitation forecast skill over Europe. *Natural Hazards and Earth System Sciences*, *23*(8), 2857–2871. https://doi.org/10.5194/nhess-23-2857-2023

Roy, T., He, X., Lin, P., Beck, H. E., Castro, C., & Wood, E. F. (2020). Global Evaluation of Seasonal Precipitation and Temperature Forecasts from NMME. *Journal of Hydrometeorology*, *21*(11), 2473–2486. https://doi.org/10.1175/JHM-D-19-0095.1

Rudisill, W. J., Flores, A. N., Marshall, H. P., Siirila-Woodburn, E., Feldman, D. R., Rhoades, A. M., et al. (2024). Cold-Season Precipitation Sensitivity to Microphysical Parameterizations: Hydrologic Evaluations Leveraging Snow Lidar Datasets. *Journal of Hydrometeorology*, *25*(1), 143–160. https://doi.org/10.1175/JHM-D-22-0217.1

Sadegh, M., Moftakhari, H., Gupta, H. V., Ragno, E., Mazdiyasni, O., Sanders, B., et al. (2018). Multihazard Scenarios for Analysis of Compound Extreme Events. *Geophysical Research Letters*, *45*(11), 5470–5480. https://doi.org/10.1029/2018GL077317

Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., et al. (2019). Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe. *Bulletin of the American Meteorological Society*, *100*(12), 2451–2472. https://doi.org/10.1175/BAMS-D-17-0274.1

Schiavina, M., Freire, S., & MacManus, K. (2023). GHS-POP R2023A - GHS population grid multitemporal (1975-2030) [Data set]. European Commission, Joint Research Centre (JRC). https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE

Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Sharma, A. R., Jain, P., Abatzoglou, J. T., & Flannigan, M. (2022). Persistent Positive Anomalies in Geopotential Heights Promote Wildfires in Western North America. *Journal of Climate*, *35*(19), 6469–6486. https://doi.org/10.1175/JCLI-D-21-0926.1

Shukla, J., Marx, L., Paolino, D., Straus, D., Anderson, J., Ploshay, J., et al. (2000). Dynamical Seasonal Prediction. *Bulletin of the American Meteorological Society*, *81*(11), 2593–2606. https://doi.org/10.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2

Spies, R. R., Franz, K. J., Hogue, T. S., & Bowman, A. L. (2015). Distributed Hydrologic Modeling Using Satellite-Derived Potential Evapotranspiration. *Journal of Hydrometeorology*, *16*(1), 129–146. https://doi.org/10.1175/JHM-D-14-0047.1

Tapiador, F. J., Roca, R., Del Genio, A., Dewitte, B., Petersen, W., & Zhang, F. (2019). Is Precipitation a Good Metric for Model Performance? *Bulletin of the American Meteorological Society*, *100*(2), 223–233. https://doi.org/10.1175/BAMS-D-17-0218.1

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., et al. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, *596*(7870), 80–86. https://doi.org/10.1038/s41586-021-03695-w

Thielen, J., Bartholmes, J., Ramos, M.-H., & De Roo, A. (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, *13*(2), 125–140. https://doi.org/10.5194/hess-13-125-2009

Tradowsky, J. S., Philip, S. Y., Kreienkamp, F., Kew, S. F., Lorenz, P., Arrighi, J., et al. (2023). Attribution of the heavy rainfall events leading to severe flooding in Western

Europe during July 2021. *Climatic Change*, *176*(7), 90. https://doi.org/10.1007/s10584-023-03502-7

Turco, M., Abatzoglou, J. T., Herrera, S., Zhuang, Y., Jerez, S., Lucas, D. D., et al. (2023). Anthropogenic climate change impacts exacerbate summer forest fires in California. *Proceedings of the National Academy of Sciences*, *120*(25), e2213815120. https://doi.org/10.1073/pnas.2213815120

Villarini, G., Vecchi, G. A., Knutson, T. R., Zhao, M., & Smith, J. A. (2011). North Atlantic Tropical Storm Frequency Response to Anthropogenic Forcing: Projections and Sources of Uncertainty. *Journal of Climate*, *24*(13), 3224–3238. https://doi.org/10.1175/2011JCLI3853.1

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, Frédéric, & Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Npj Climate and Atmospheric Science*, *1*(1), 3. https://doi.org/10.1038/s41612-018-0013-0

Wanders, N., & Wood, E. F. (2016). Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environmental Research Letters*, *11*(9), 094007. https://doi.org/10.1088/1748-9326/11/9/094007

Wanthanaporn, U., Supit, I., Van Hove, B., & Hutjes, R. W. A. (2023). *Analysis of seasonal climate and streamflow forecasts performance for Mainland Southeast Asia* (preprint). Water Resources Management/Modelling approaches. https://doi.org/10.5194/hess-2023-56

Xu, L., Chen, N., Chen, Z., Zhang, C., & Yu, H. (2021). Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. *Earth-Science Reviews*, *222*, 103828. https://doi.org/10.1016/j.earscirev.2021.103828

Xue, P., Malanotte-Rizzoli, P., Wei, J., & Eltahir, E. A. B. (2020). Coupled Ocean-Atmosphere Modeling Over the Maritime Continent: A Review. *Journal of Geophysical Research: Oceans*, *125*(6). https://doi.org/10.1029/2019JC014978

Zhang, Y., Wallace, J. M., & Battisti, D. S. (1997). ENSO-like Interdecadal Variability: 1900–93. *Journal of Climate*, *10*(5), 1004–1020. https://doi.org/10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2

Zheng, Y., Li, S., & Ullah, K. (2020). Increased Occurrence and Intensity of Consecutive Rainfall Events in the China's Three Gorges Reservoir Area Under Global Warming. *Earth and Space Science*, *7*(8), e2020EA001188. https://doi.org/10.1029/2020EA001188

References from the Supporting Information

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., et al. (2019). GEOS-5 seasonal forecast system. *Climate Dynamics*, *53*(12), 7335–7361. https://doi.org/10.1007/s00382-017-3835-2

Burke, M., Childs, M. L., De La Cuesta, B., Qiu, M., Li, J., Gould, C. F., et al. (2023). The contribution of wildfire to PM2.5 trends in the USA. *Nature*, *622*(7984), 761–766. https://doi.org/10.1038/s41586-023-06522-6

García-Franco, J. L., Lee, C.-Y., Camargo, S. J., Tippett, M. K., Kim, D., Molod, A., & Lim, Y.-K. (2023). Climatology of tropical cyclone precipitation in the S2S models. *Weather and Forecasting*. https://doi.org/10.1175/WAF-D-23-0029.1

Zhou, Y., Zaitchik, B. F., Kumar, S. V., Arsenault, K. R., Matin, M. A., Qamer, F. M., et al. (2021). Developing a hydrological monitoring and sub-seasonal to seasonal forecasting system for South and Southeast Asian river basins. *Hydrology and Earth System Sciences*, *25*(1), 41–61. https://doi.org/10.5194/hess-25-41-2021