Spatial and Temporal Variation of Subseasonal-to-Seasonal (S2S) Precipitation Reforecast Skill Across CONUS

Jessica Rose Levey[1] and Arumugam Sankarasubramanian[1]

[1]North Carolina State University

September 13, 2023

Abstract

Precipitation forecasts, particularly at subseasonal-to-seasonal (S2S) time scale, are essential for informed and proactive water resources management. Although S2S precipitation forecasts have been evaluated, no systematic decomposition of the skill, Nash-Sutcliffe Efficiency (NSE) coefficient, has been analyzed towards understanding the forecast accuracy. We decompose the NSE of S2S precipitation forecast into its three components – correlation, conditional bias, and unconditional bias – by four seasons, three lead times (1–12-day, 1-22 day, and 1-32 day), and three models (ECMWF, CFS, NCEP) over the Conterminous United States (CONUS). Application of dry mask is critical as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. Further, a west-to-east gradient in S2S forecast skill exists and forecast skill was better during the winter months and for areas closer to the coast. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during fall and winter months. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in in-land areas. Post-processing using simple Model Output Statistics could reduce both unconditional and conditional bias to zero, thereby offering better skill for regimes with high correlation. Our decomposition results also show efforts should focus on improving model parametrization and initialization schemes for climate regimes with low correlation values.

Hosted file

`973472_0_art_file_11364153_s027fm.docx` available at https://authorea.com/users/663671/articles/665762-spatial-and-temporal-variation-of-subseasonal-to-seasonal-s2s-precipitation-reforecast-skill-across-conus

Hosted file

`973472_0_supp_11364154_s0d7fm.docx` available at https://authorea.com/users/663671/articles/665762-spatial-and-temporal-variation-of-subseasonal-to-seasonal-s2s-precipitation-reforecast-skill-across-conus

Spatial and Temporal Variation of Subseasonal-to-Seasonal (S2S) Precipitation Reforecast Skill Across CONUS

J. R. Levey[1], A. Sankarasubramanian[1]

[1]Department of Civil, Construction and Environmental Engineering, North Carolina State University.

Corresponding author: Jessica Levey (jrlevey@ncsu.edu)

Key Points:

- NSE decomposition of S2S reforecast skill shows the spatio-temporal variations in correlation, conditional and unconditional bias.
- Longitudinal gradient of forecast skill exists from the West (higher) to East (lower).
- Regression based model-output statistics provide correlation as the lower bound of NSE as the marginal and conditional bias reduces to zero.

Abstract

Precipitation forecasts, particularly at subseasonal-to-seasonal (S2S) time scale, are essential for informed and proactive water resources management. Although S2S precipitation forecasts have been evaluated, no systematic decomposition of the skill, Nash-Sutcliffe Efficiency (NSE) coefficient, has been analyzed towards understanding the forecast accuracy. We decompose the NSE of S2S precipitation forecast into its three components – correlation, conditional bias, and unconditional bias – by four seasons, three lead times (1–12-day, 1-22 day, and 1-32 day), and three models (ECMWF, CFS, NCEP) over the Conterminous United States (CONUS). Application of dry mask is critical as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. Further, a west-to-east gradient in S2S forecast skill exists and forecast skill was better during the winter months and for areas closer to the coast. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during fall and winter months. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in in-land areas. Post-processing using simple Model Output Statistics could reduce both unconditional and conditional bias to zero, thereby offering better skill for regimes with high correlation. Our decomposition results also show efforts should focus on improving model parametrization and initialization schemes for climate regimes with low correlation values.

1.0 Introduction

Global climate change and regional anthropogenic disturbances, including urbanization and deforestation, are driving shifts in the hydrologic cycle, and impacting water resources (Konapala et al, 2020; Milly et al., 2008). Consequently, extreme precipitation events, including prolonged droughts or flooding, are expected to be more frequent, further threatening water supply and variability (Milly et al., 2008). In conjunction with hydroclimatic changes, population changes also stress surface and groundwater resource withdrawals in many regions across the Conterminous US (CONUS) (Sankarasubramanian et al., 2017). Reservoir releases, during both floods and droughts, are modified for human needs, downstream ecological health, and for ensuring watershed resilience (Chalise et al., 2021). Mismanagement of water resources, both surface water and groundwater, may pose threats to agriculture, supply chains, human and environmental health, and regional economies. Hence, reliable and accurate subseasonal-to-seasonal (S2S) precipitation forecasts are essential in an age of a changing climate for improving water management strategies and preparing for near-future hydroclimatic extremes.

Compared to the skill of short-range weather forecasts (less than 15 days) and long-range seasonal forecasts, which are reasonably good, the skill of S2S forecasts, ranging between 15 to 60 days, is low and is often referred to as the 'predictability desert' (Vitart et al., 2012). Understanding the current S2S precipitation forecasts skill, as well as highlighting the potential avenues – initialization, parametrization, and post-processing schemes – for improvement are critical for accurate S2S precipitation forecasts for operational use (White et al., 2017). Known contributing factors that influence S2S model forecasting performance include the parametrization and initialization schemes, large-scale atmospheric circulation modes, and coupled models (Vitart et al., 2018). The model initialization scheme, including land surface and soil moisture representation, are also crucial for accurately representation of geophysical fluxes.

Climate oscillations, such as El Nino Southern Oscillations (ENSO) and Madden-Julian Oscillations (MJO) also influence seasonal forecast prediction skill (Zhang, 2013). ENSO's influence on United States' winter hydroclimatology is well-known, particularly over the Southeast and west coast, accounting for roughly a third of US winter forecasting skill (Quan et al., 2006).

Previous studies have attributed S2S skill between ENSO and MJO (Sun et al., 2022; Wang et al., 2019) and have compared S2S skill across models, lead times and seasons (Zhang et al 2021; de Andrade et al, 2019). However, these studies that examined S2S models' forecasting performance did not apply a threshold on dry mask prior to calculating the model's skill. Zhang et al (2021) have evaluated S2S forecast skill by filtering extreme precipitation events, but did not apply a dry mask threshold for evaluating the overall skill. Without a dry mask threshold, the S2S skill will be inflated, especially in regions with a pronounced dry season, as zero rainfall days is included in these skill calculations (Wilks, 2006). The ability to predict days without precipitation is important for drought prediction and planning, but the skill will be inflated for wetter and normal conditions; therefore, the dry mask application was used to filter out areas of inflated skill based on the climatological means. Several studies focused on extreme precipitation forecasts have applied percentile filters (Zhang et al., 2021), which reduces the sample size particularly while evaluating monthly/seasonal skill. Given the pronounced seasonality in precipitation over the CONUS (Petersen et al., 2012), we systematically evaluate the S2S forecasting skill across CONUS by applying a dry mask before considering the skill for each lead time, season and region. Evaluating the forecast skill after applying the dry mask could potentially affect the source of model skill, and the associated biases that could be obtained from decomposition.

S2S precipitation forecast skill has been compared considering both probabilistic and deterministic metrics to evaluate the forecast skill (Zhang et al., 2021; de Andrade et al., 2019). S2S models' skill have been evaluated using Mean Square Error (MSE), mean square skill score (MSSS), root mean square error (RMSE), anomaly correlation coefficient (ACC), Pearson's correlation coefficient, and ranked probability skill score (RPSS) (Zhang et al., 2021; de Andrade et al., 2019). de Andrade, et al., (2019) evaluated hindcast skill using linear correlation coefficient and analyzed the sources of bias and variability; however, this study was a large-scale global analysis of forecast skill and did not consider the seasonal skills and the associated errors. Decomposing the MSSS three components – correlation coefficient, condition bias and marginal bias – would provide information on the regions and seasons over which the selected models have the ability to capture the variability in observed precipitation but have significant biases in estimation. Further, the hindcast assessment of (de Andrade et al., 2019) was performed without the dry mask application, which may inflate forecast skill particularly for regions with pronounced dry season.

The Nash-Sutcliffe Efficiency (NSE), also known as the coefficient of determination, is a metric that measures the skill of hydrologic models (Nash & Sutcliffe, 1970). Li et al., (2022) used to evaluate S2S forecast skill performance based on Kling-Gupta Efficiency (KGE) metric, which provides a different decomposition of NSE, without applying the dry mask across the CONUS or considering seasonality. However, decomposing the Nash-Sutcliffe Efficiency (NSE) for precipitation hindcasts after applying the dry mask provides critical information without inflating the skill of the model. Furthermore, implementing new parametrizations and

initialization schemes could be costly and take additional time to develop reforecasts. One effective way to improve the forecasting skill is to consider post-processing schemes (Carter et al., 1998; Glahn et al., 2003). Further, post-processing could also be implemented over reforecasts from multiple models to develop multi-model ensembles which have been shown to improve the forecast skill compared to the best individual model (Weigel et al., 2008). Past work on statistical post-processing has considered both parametric and non-parametric approaches (Hamill et al., 1997; Schefzik et al., 2013; Scheuerer et al., 2015). Although many studies have used post-processing schemes on S2S precipitation forecasts, understanding the components of S2S forecast skill could provide additional insights on how post-processing schemes can be used and could also indicate potential regions where improvements in models will be needed to further improve the forecast skill.

Several S2S models that contribute multi-model ensembles have been run for reforecasts. Historically, some S2S multi-model datasets have only been running for a period of short time, limiting the ability to capture the interannual variability in precipitation. Other multi-model ensembles have primarily focused on generating monthly forecasts for seasonal prediction with infrequent model initialization. This study uses three individual models hindcasts from the World Weather Research Programme (WWRP) and World Climate Research Programme (WCRP) S2S prediction project (Vitart et al., 2012). The S2S project, originating in 2013, has a long record of forecasts and reforecasts that are initialized multiple times a week (Vitart et al., 2017). The longer range of data allows for larger sample sizes for robust estimation of NSE and decomposition metrics. Comparing model performance is important because forecast skill varies between S2S models as each model has different parameterization schemes, number of ensembles, and resolution. This study will consider decomposition of NSE of S2S reforecasts over the CONUS for three models – European Centre of Medium-Range Weather Forecast's (ECWMF) National Centre for Environmental Prediction Climate Forecast System (NCEP CFS) and Environment and Climate Change Canada (ECCC) – after applying the dry mask. Previous studies have shown ECMWF S2S hindcast models have outperformed both CFS and ECCC models on a global basis (de Andrade et al, 2019), but the performance of these three models have not been compared after the dry mask threshold has been applied. The North American Multi-Model Ensemble (NMME) forecasts have proved to perform better than individual models by pooling the ensemble members from several models (Krakauer, 2019). However, for this study, the NMME was not considered because the number of ensemble members varies between individual models, giving more weight to some models. Additionally, to improve multi-model performance, understanding individual models' type of errors and potential for correcting the biases before pooling the ensembles, which could further improve the multi-model forecast performance. Hence, this study will compare the decomposed NSE and associated errors of S2S precipitation forecasts of three individual models by season and lead time under three Koppen climate regimes across the CONUS.

The main intent of this study is to decompose the S2S forecasting skill as a function of lead time over the CONUS after applying the dry mask. To our knowledge, limited/no work has been performed on systematically decomposing the NSE over various seasons after applying the dry mask. In addition to applying the dry mask, evaluating model skill regionally is also critical as the precipitation has pronounced seasonality over the CONUS (Petersen et al., 2012).

Analyzing forecasting skills regionally can also provide insights on how land surface conditions, low-frequency oscillations, and regional hydroclimate influence the model performance.

The manuscript is organized into the following sections: S2S precipitation hindcast and observed databases from three different models are provided in the next section. Then, the dry mask threshold application procedure is presented along with the NSE decomposition. The following section provides the results from the full decomposition of ECMWF and the results from different regimes along with the skill comparison from three S2S reforecasts.

2.0 Data

This section provides the S2S hindcast database and observed data along with the details to calculate and decompose the NSE for S2S forecasts over various lead times and seasons.

Observed Precipitation

For calculating the S2S reforecasts skill, we used the CPC Global Unified Precipitation dataset provided by the NOAA Physical Science Laboratory (PSL), with a resolution of (0.5°x0.5°) (Chen, et al., 2008). Upon comparing the accuracy of various precipitation datasets, the CPC Unified dataset performed particularly well in areas that have dense areas of rain gauges (Beck et al., 2017). This study focused on the CONUS, which has a dense system of rain gauges, and has been used in other forecast verification studies (Becker et al, 2020).

S2S Hindcast Database

For S2S model skill evaluation, three hindcast models were assessed: 1.) European Centre of Medium-Range Weather Forecasts (ECMWF), 2.) National Center for Environmental Prediction's (NCEP) Climate Forecast System (CFS) model, and 3.) Environment and Climate Change Canada (ECCC). For full decomposition of ECMWF, the S2S hindcasts were evaluated for the full 20-year hindcast period (Table 1) and up to the longest available lead time of 45 days. The ensemble means were averaged over three different lead times: 1) 1-15 days, 2) 1-30 days, and 3) 1-45 days, and compared with the observed average precipitation corresponding to the three lead times. Additionally, the average forecasts and corresponding observed average daily precipitation values were pooled by the date of hindcast initialization into the following seasons: a) January, February, March (JFM), b) April, May, June (AMJ), c) July, August, September (JAS), d) October, November, December (OND). Thus, the evaluation for each season provides the skill of forecasts issued during the months within the considered four seasons.

For the model comparison section, the three models were assessed for lead times of 1-12 days, 1-22 days, and 1-32 days for four different seasons between January 1st 2000 and December 30th 2010, the longest available overlapping date ranges and lead times for all three models. Additionally, ECMWF and NCEP were compared for lead times of 1-42 days. The ECMWF hindcasts are initialized twice a week and range from 2000-2019, NCEP CFS hindcasts are initialized daily and are available from 1999-2010, and ECCC are initialized weekly, and

reforecasts range from 1995-2012 (Vitart et al., 2017). The S2S precipitation hindcast model's information and specification are shown in Table 1 (Vitart et al., 2017).

Model	LEAD TIME	RESOLUTION	HINDCAST PERIOD	HINDCAST ENSEMBLE SIZE	FORECAST ENSEMBLE SIZE	HINDCAST FREQUENCY	OCEAN COUPLING	SEA ICE COUPLING
ECMWF	0-46 Days	0.25°x0.25°, days 0-10, 0.5°x0.5°, after day 10 L91	Past 20 Years	11	51	Twice a Week	Yes	No
NCEP CFS	0-44 Days	~1°x1°, L64	1999-2010	4	16	Daily	Yes	Yes
ECCC	0-32 Days	0.45°x0.45°, L40	1995-2012	4	21	Weekly	Yes	No

Table 1. Subseasonal-to-Seasonal Hindcast Models and Forecast model information

2.1 Dry Mask application and Skill Assessment and Decomposition

a. Seasonality of Rainfall and Dry Mask Application

Prior to calculating the NSE for each hindcast-initialized season, a dry mask was applied based on the observed precipitation dataset to filter out the areas that receive small amounts of rainfall, which may result in an inflated forecast skill because the forecasted and observed rainfall have no rainfall. Antolilk (2000) and Charba et al., (2011) considered daily precipitation less than 0.01 inches as no event for evaluating the skill. Based on that work, the dry mask was set at a threshold value for each individual grid cell, if the observed daily precipitation over the 20 years is less than 0.15 inches, 0.30 inches and 0.45 inches for 15-day, 30-day and 45-day lead times from the time of issued forecast, respectively. The NSE and the three components were evaluated for all the three models for each lead time over the CONUS. We also evaluate the forecast skill – NSE and its components – based on the climate regime. For this purpose, we considered three main regimes – desert (regime B), temperate (regime C) and continental (regime D) – over the CONUS based on Koppen climate classification. A small area in southern Florida fell into the tropical (regime A) Koppen climate group; however, since this regime corresponds to only one grid cell from the hindcast model, we combined this tropical area with the temperate regime (Supplemental Information (SI) - Figure SI-1). Using the aggregated Koppen Climate Regime (Beck, et. al, 2017) into three climate regimes, a regional analysis was performed for each of the S2S hindcast models (Supplemental Information (SI) - Figure SI-1).

b. Skill Assessment Metrics

Skill assessment metrics measure the performance of the model's forecast ability compared to the observed variable. Frequently used performance metrics include anomaly correlation, NSE and Kling Gupta Efficiency (Clark et al., 2021). The NSE measures the magnitude of error variance from the model prediction compared to the observed variance in the data and has an upper bound of 1 but has a lower bound of -∞ and is used to determine the 'goodness-of-fit' of a

model. NSE is related to MSE but is normalized by the standard deviation of the observed precipitation or data values (Gupta et al., 2009).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\overline{o}_{it})^2} \quad (1)$$

Where o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. The mean observed precipitation is $\overline{o}_{it}$. For a given i, NSE will be decomposed into three parts (Murphy 1988; Weglarczyk 1998): A) Pearson's correlation coefficient (equation 3), B) conditional bias (equation 4), and C) unconditional bias (equation 5) (Gupta et al., 2009).

$$NSE = A - B - C$$

$$NSE = \rho_{xo}{}^2 - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad (2)$$

$$A=\rho_{xo}{}^2 \quad where \quad \rho_{xo} = \frac{cov(x,o)}{\sigma_x * \sigma_o} \quad (3)$$

$$B=[\rho_{xo} - \frac{\sigma_x}{\sigma_x}]^2 \quad (4)$$

$$C=[\frac{\overline{x}-\overline{o}}{\sigma_o}]^2 \quad (5)$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 1. The pearson correlation coefficient between x and o is ρ_{xo} (equation 3). The first component of the decomposition, Pearson's correlation coefficient, shows the linear association between the forecast and the observation. The conditional bias is the difference in the slope of the regression line fitted between forecast and observation with a slope of 1 that indicates a perfect forecast. The unconditional bias, indicating a systematic bias, denotes the ratio of difference between the mean of the observation and the mean of the forecast to the observed standard deviation.

3.0 Results

Full Decomposition of ECMWF

A full NSE decomposition was performed on the ECMWF S2S hindcast model because the ECMWF model has the longest available reforecast time range and has the largest number of ensemble members. Prior to decomposing NSE, a dry-mask threshold was applied based on the lead time for the climatological means of each grid cell, to mask out areas with low precipitation values to avoid inflated skill values. Both NSE and correlation are lower across all seasons after the dry mask threshold was applied. Figure 1a illustrates the difference in Normalized Nash-Sutcliffe Efficiency (NNSE) of 30-day ahead S2S precipitation forecast skill with and without the dry mask threshold). For instance, a forecast issued on March 30, 2000 with a lead time of 45 days corresponds to the skill of the forecast in predicting precipitation from March 30, 2000 to May 15, 2000. Thus, the skill of the forecast issued in JFM can cover the observed precipitation in April and May. To reiterate, all the figures with seasonal S2S performance metrics denote the skill summary of the forecast issued during that season as opposed to the ability to forecast the observed precipitation during that season.

To understand the importance of dry masking, we first show the 1-30 day ahead S2S precipitation forecast skill with and without dry mask (Figure 1) based on Normalized NSE (NNSE). Lower NNSE (equation 6) values, the inverse of NSE, indicate better predictive performance.

$$NNSE = \frac{1}{2 - NSE} \qquad (6)$$

For the forecast issued in the four seasons, the mean NNSE values are lower for the grid cells below the dry mask threshold than for the grid cells that exceeded the threshold (Figure 1). Even though including "no-precipitation event" is expected to inflate the skill, dry masking by filtering out regimes rather than simply removing values below a given threshold, allows us to maintain the same sample size across all grid cells, thereby changing the masked areas based on both forecast-initialized seasons (Figure 1) and lead time.

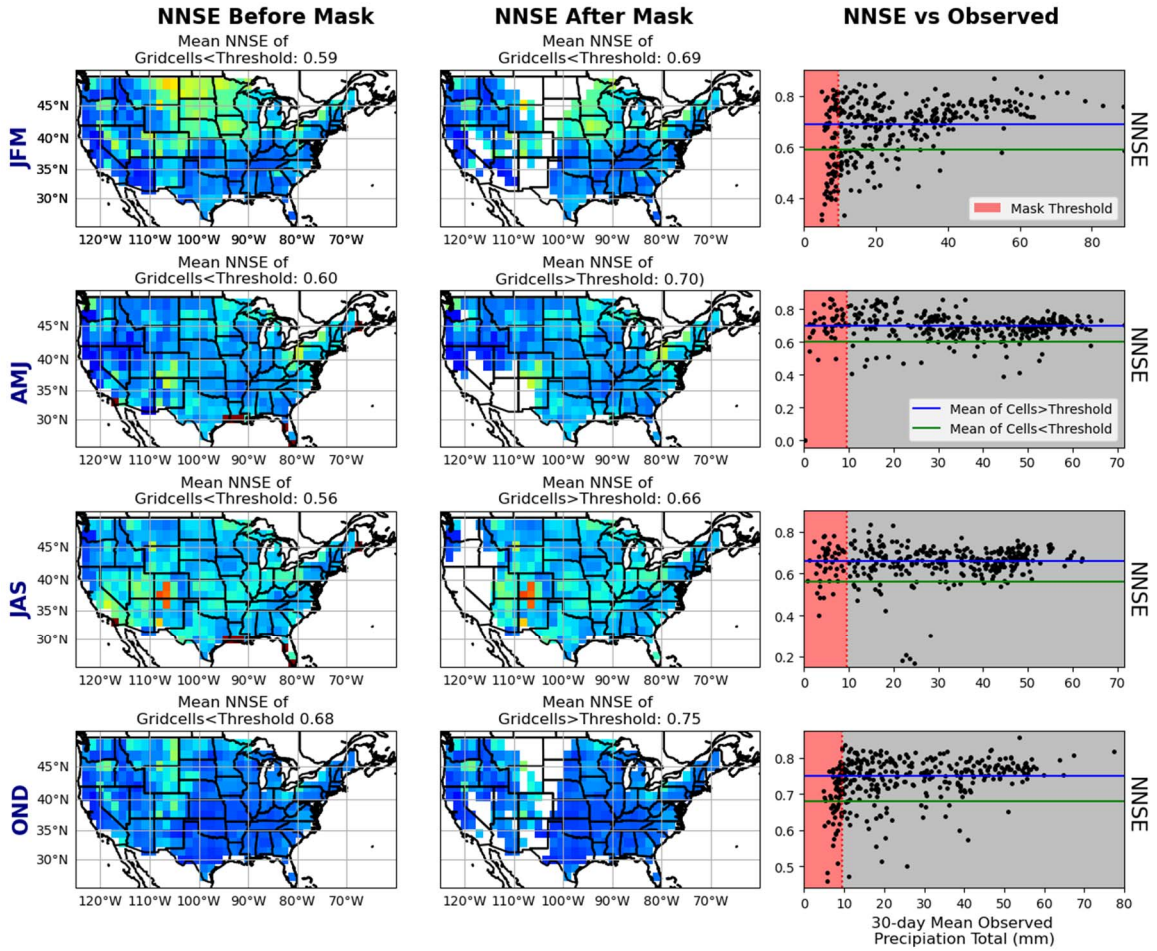


Figure 1. Normalized Nash Sutcliffe Efficiency (NNSE) of 1-30 days ahead ECMWF hindcast for the CONUS before dry mask is applied (left column) and after (middle column) dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS and OND for 1-30-day lead time. The scatter plot comparison of grid cell's 1-30-day climatological precipitation means and the corresponding Normalized NSE values (right column). The scatter plot shows the NNSE values that fall below the dry mask threshold (red region) and above (gray region). The average NNSE of the grid cells below the dry mask threshold (green) and above the dry mask threshold (blue). Since the NNSE is the inverse of the NSE, the lower NNSE values indicate better predictive performance.

The overestimation of S2S forecast skill occurs if no dry mask is applied, particularly for pronounced dry seasons (JFM and JAS). Studies that evaluated S2S precipitation forecasts skill did not consider dry mask application, which ignores the seasonality in precipitation, thereby indicating potential difference in forecast skill between regions (e.g., Li et al., 2022). However, after the dry mask application (Figure 1), we find that the skill was fairly similar between regimes. Thus, it is important to apply a dry mask which inherently considers the seasonality in precipitation for skill evaluation. Quantifying the forecast skill for critical events (e.g., peak rainfall seasons) is important particularly if the interest is to identify regions with limited skill.

a) NSE Spatial Patterns

We present results for the NSE and its decomposition (Figures 2-7) for the ECMWF model and then compare its performance with NCEP and ECCC later (Figures 8-10). Before assessing the components of the NSE, we first investigate the NSE over the CONUS, which shows the S2S forecasting skill of ECMWF for various lead times over the season (Figure 2). NSE is better in the winter and fall seasons (JFM and OND) in comparison to spring and summer seasons (AMJ and JAS) (Figure 2), which is partially due to El Nino Southern Oscillation (ENSO) being active during winter and fall months and ENSO dying or being at an incipient stage during AMJ and JAS (Ham et al., 2019). The NSE also tends to be better closer to the coasts indicating the local sea surface temperatures (SSTs) in influencing S2S forecasts. Additionally, the NSE shows a slight gradient from West Coast to East Coast (Figure 2). The NSE tends to be weaker around the Great Lakes. Further, the areas surrounding the dry mask regions tend to have a lower NSE.

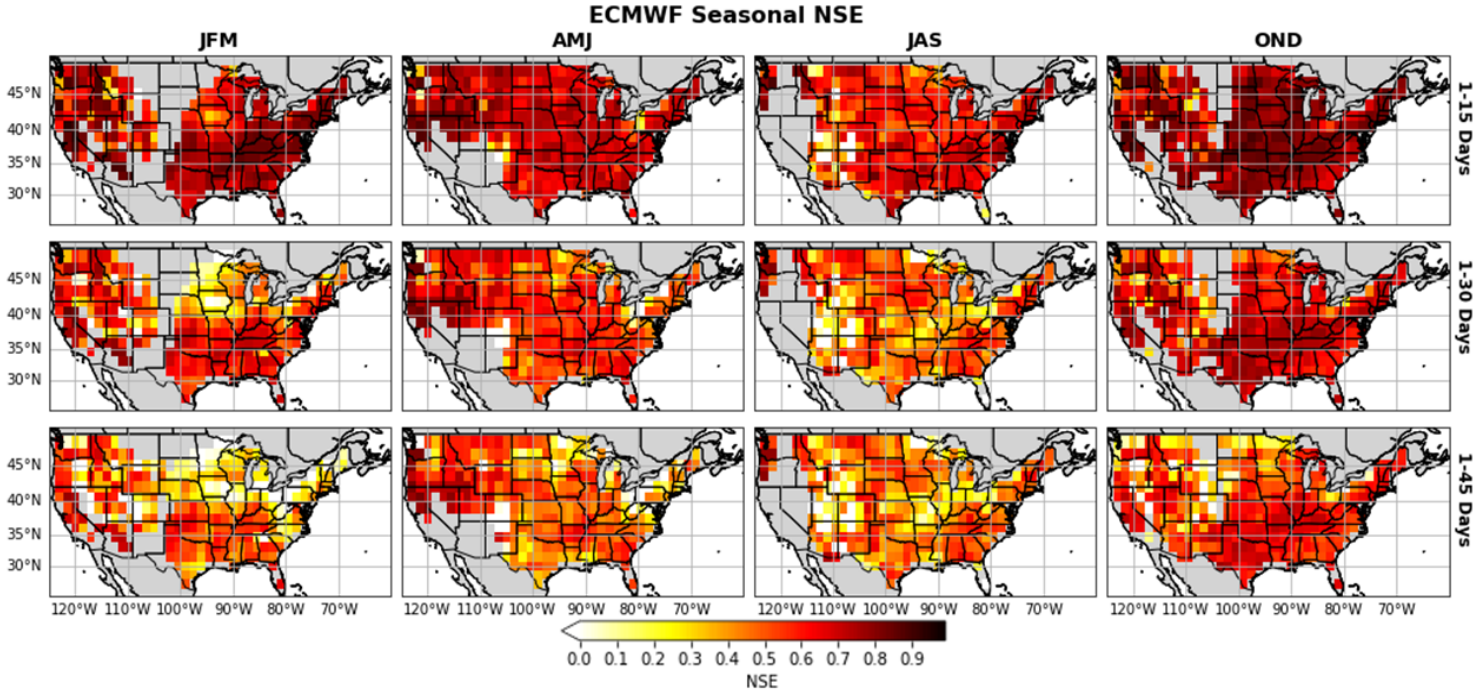


Figure 2. Nash Sutcliffe Efficiency (NSE) of ECMWF hindcast for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

b) Decomposition Plots

We decompose the NSE of ECMWF in Figure 2 into correlation (Figures 3), conditional bias (Figure 5) and unconditional bias (Figure 6) for each lead time for the four seasons.

i) Correlation and its longitudinal distribution

The first component of decomposition, Pearson's correlation coefficient, shows the innate model skill and the lower bound for explained variance in the model. The analysis of correlation shows that the skill decreases as lead time increases for all seasons (Figure 3.). Similar to the NSE, the correlation is also lower in the summer seasons and higher in the winter seasons. The correlation between S2S precipitation hindcasts and observed precipitation was averaged by longitude, for each season and lead time, after the dry mask threshold was applied. This longitudinal distribution more clearly illustrates the West to East coast gradient, where the correlation is higher in the West Coast and decreases towards the East Coast (Figure 3-4).

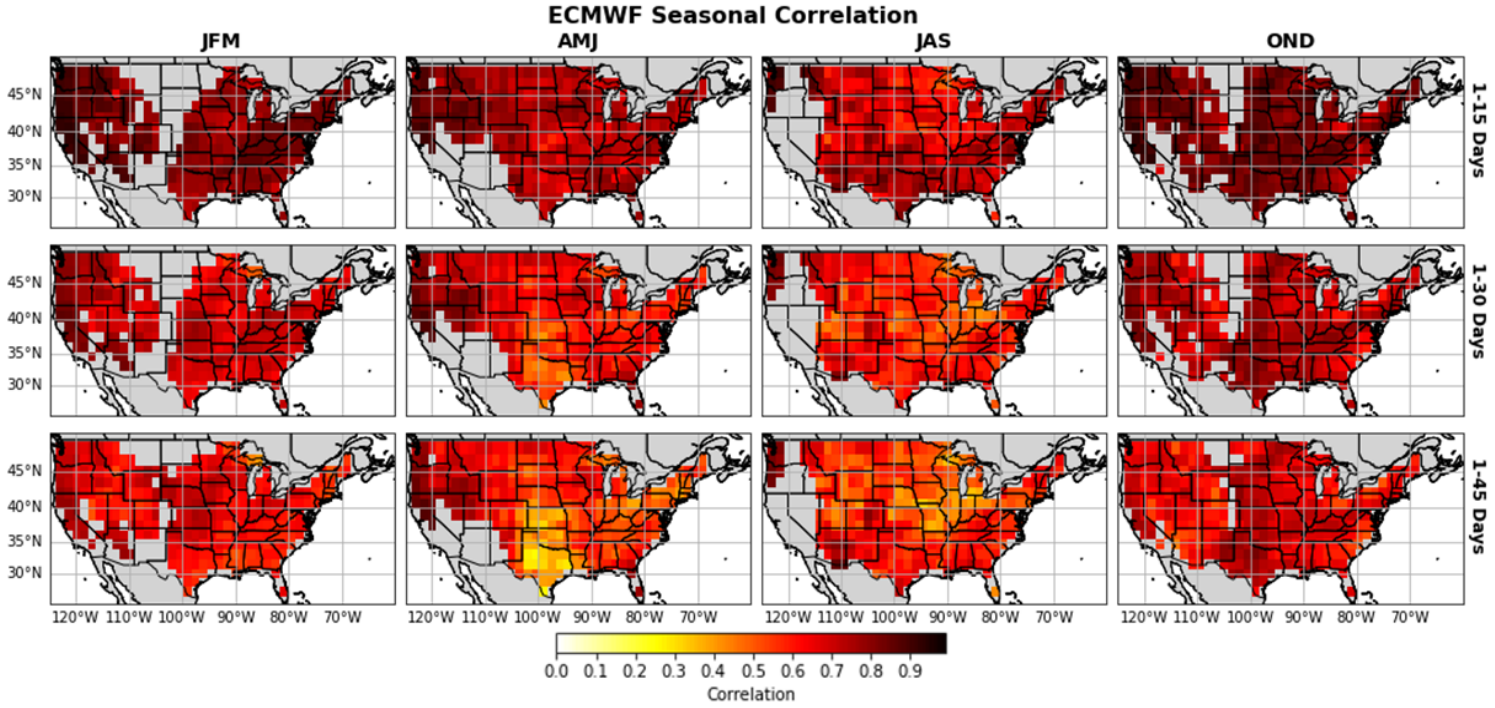


Figure 3. Correlation, the first component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS, and OND, and for three different lead times: 1-15 days, 1-30 days, and 1-45 days.

On the West Coast, correlation coefficients are higher than on the East Coast, which is partially due to the pronounced seasonality in precipitation over the West Coast that results in reduced number of grid cells being considered for evaluation after applying the dry mask. Additionally, correlation coefficients are higher towards the coasts and weaker further inland due to potential influence of local SSTs (Sankarasubramanian et al., 2017). Correlation coefficients are also lower towards the area surrounding the masked out regions.

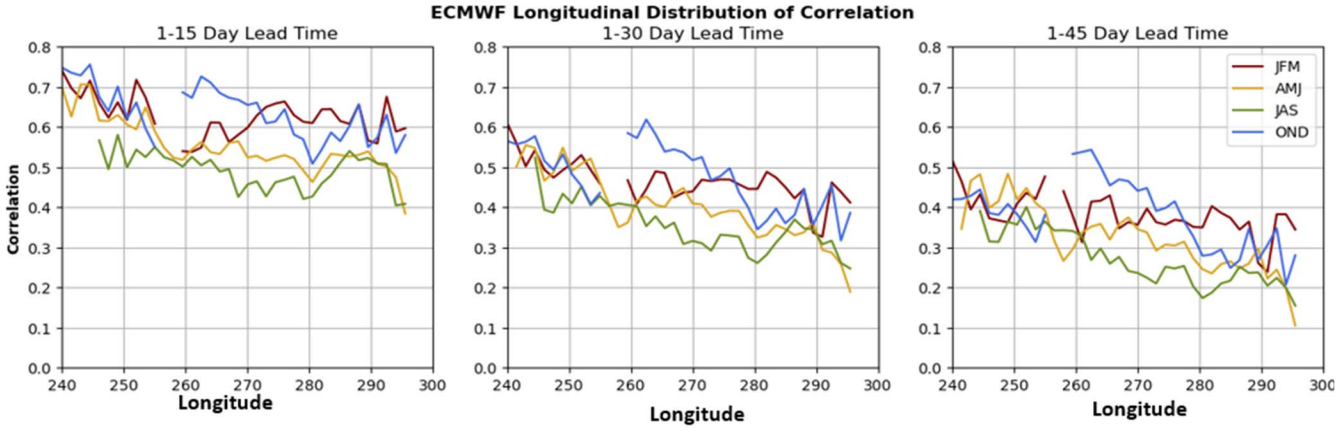


Figure 4. Longitudinal distribution of correlation by the average by latitude of the ECMWF hindcast data for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days

ii.) Conditional Bias

The second and third components, conditional bias, and unconditional bias, are expected to be zero for ideal forecasts. The conditional bias for the ECMWF decomposition increases as lead time increases and tends to be higher towards the coasts. Further, the conditional bias is higher during the summer season in comparison to the winter season (Figure 5). The Great Lakes Region and the central part of the US has a high conditional bias that increases with increasing lead times, whereas the Sunbelt has a low conditional bias during the winter and spring seasons. Conditional bias is also higher towards the areas that were masked out from the dry mask. Conditional bias is highest during JAS, specifically in the desert areas that were masked out during the other seasons and is lowest during OND.

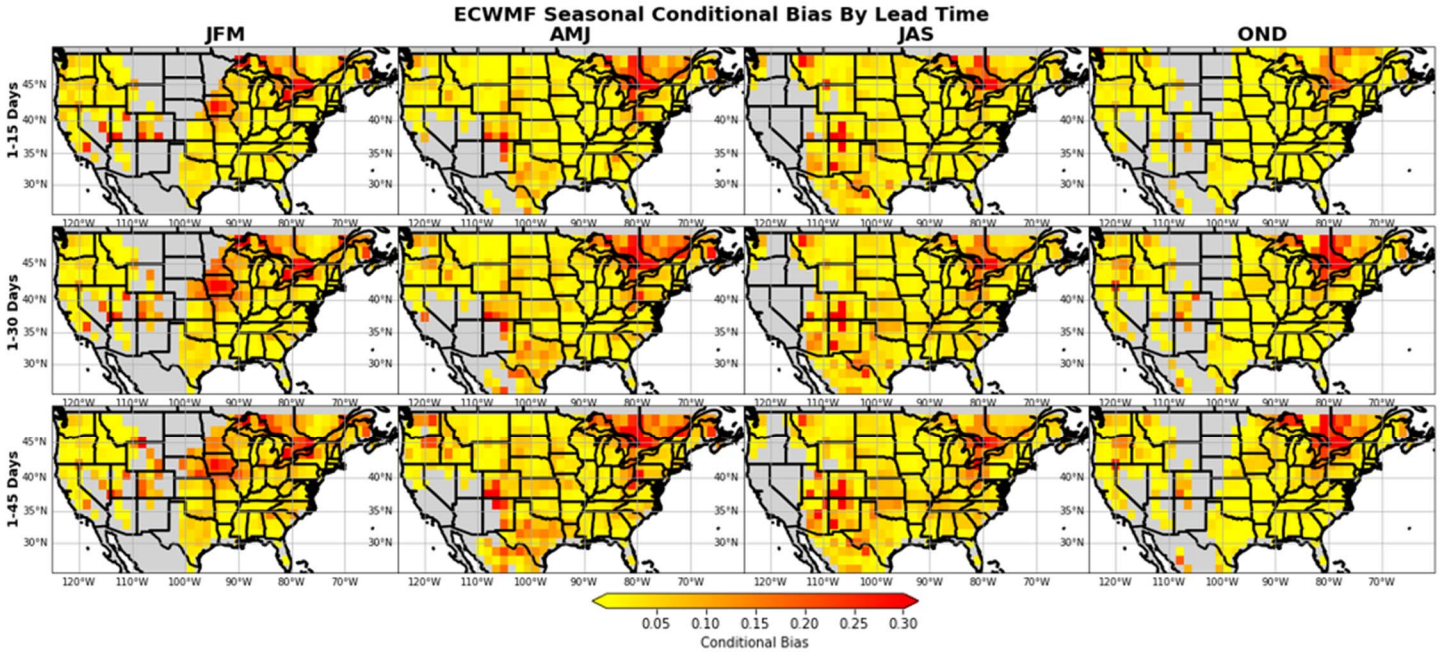


Figure 5. The second component, conditional bias, of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

iii.) Unconditional Bias

The third component, unconditional bias, represents the systematic bias in reproducing the long-term mean of the observed precipitation. Unconditional bias is high in the Great Lakes Region and in the central part of the US (Figure 6). Additionally, unconditional bias is high in the desert regions for JAS, which was masked during the other seasons, for JAS. Conditional bias and unconditional bias are generally correlated and have higher values in the same regions.

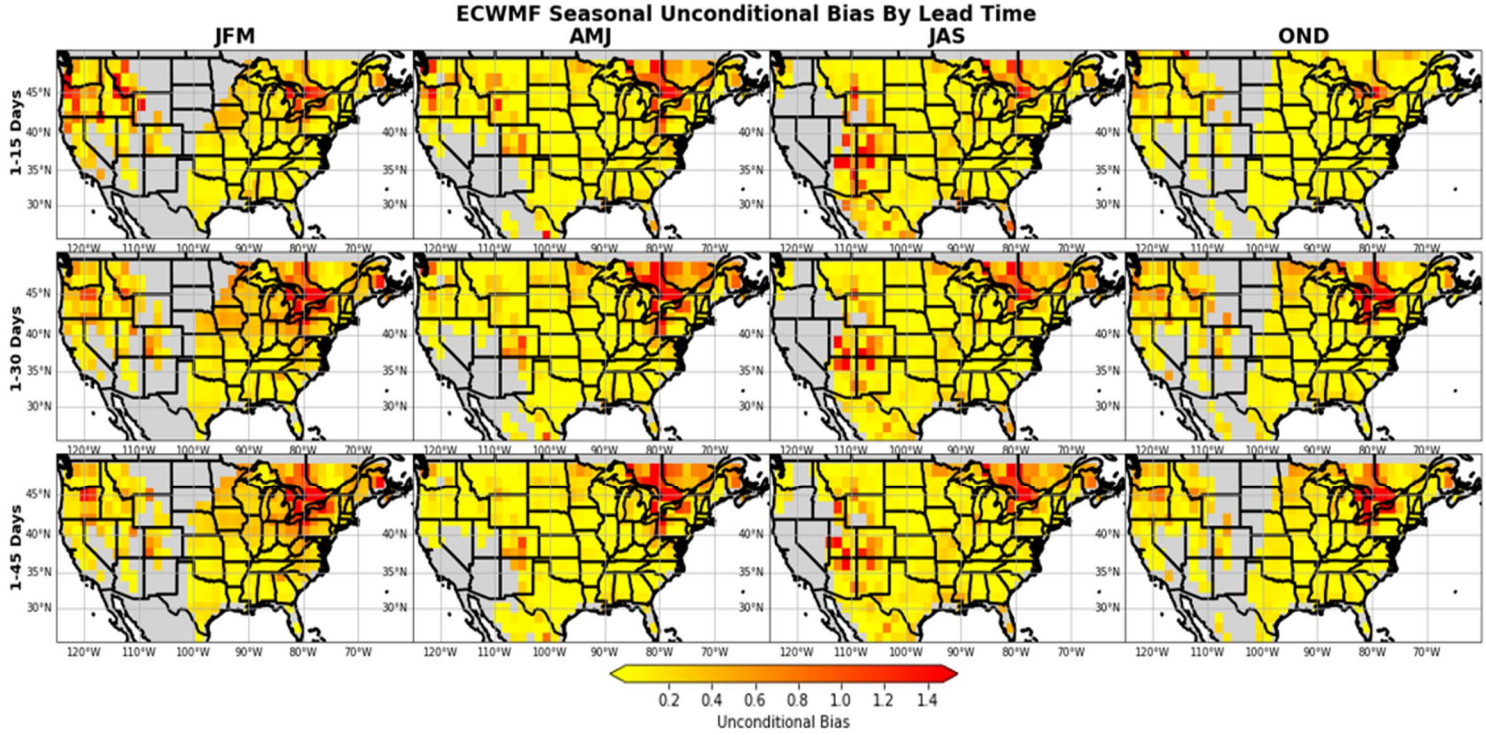


Figure 6. Unconditional bias, the third component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

c. Skill comparison across Koppen Climate Regimes

The skill of ECMWF S2S hindcast model was compared under three Koppen climate regimes: a) desert b.) temperate and c.) continental (Figure SI-1). For all lead times and climate regimes, the correlation varies by season and is lower in the summer months and is the highest in the winter months (Figure 7). Since the dry mask threshold was applied before the climate regime classification was considered, the correlation does not vary much between regimes within a given season. Conversely, if a dry mask had not been applied, the desert regimes may expect to have much better skill, because of inflated skill due to no-precipitation days.

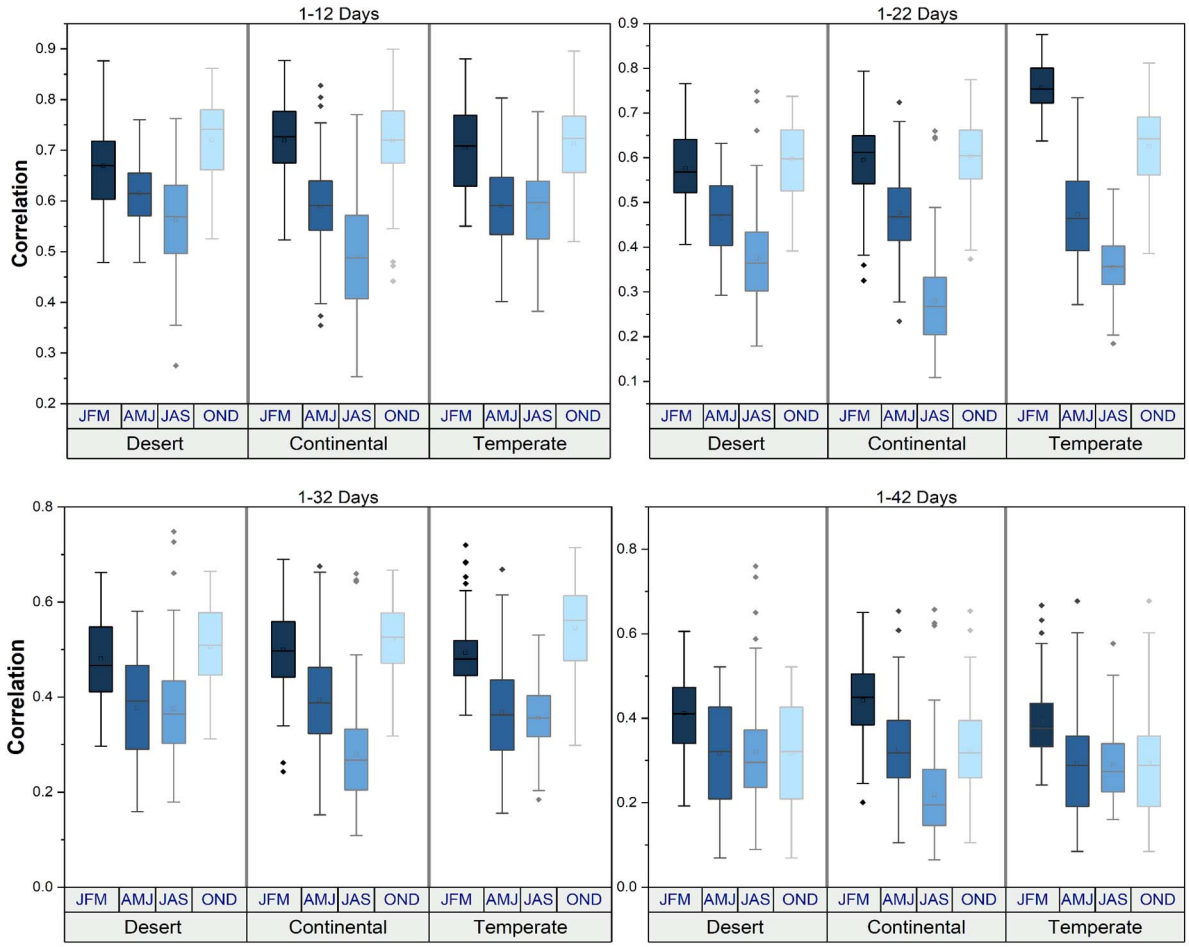


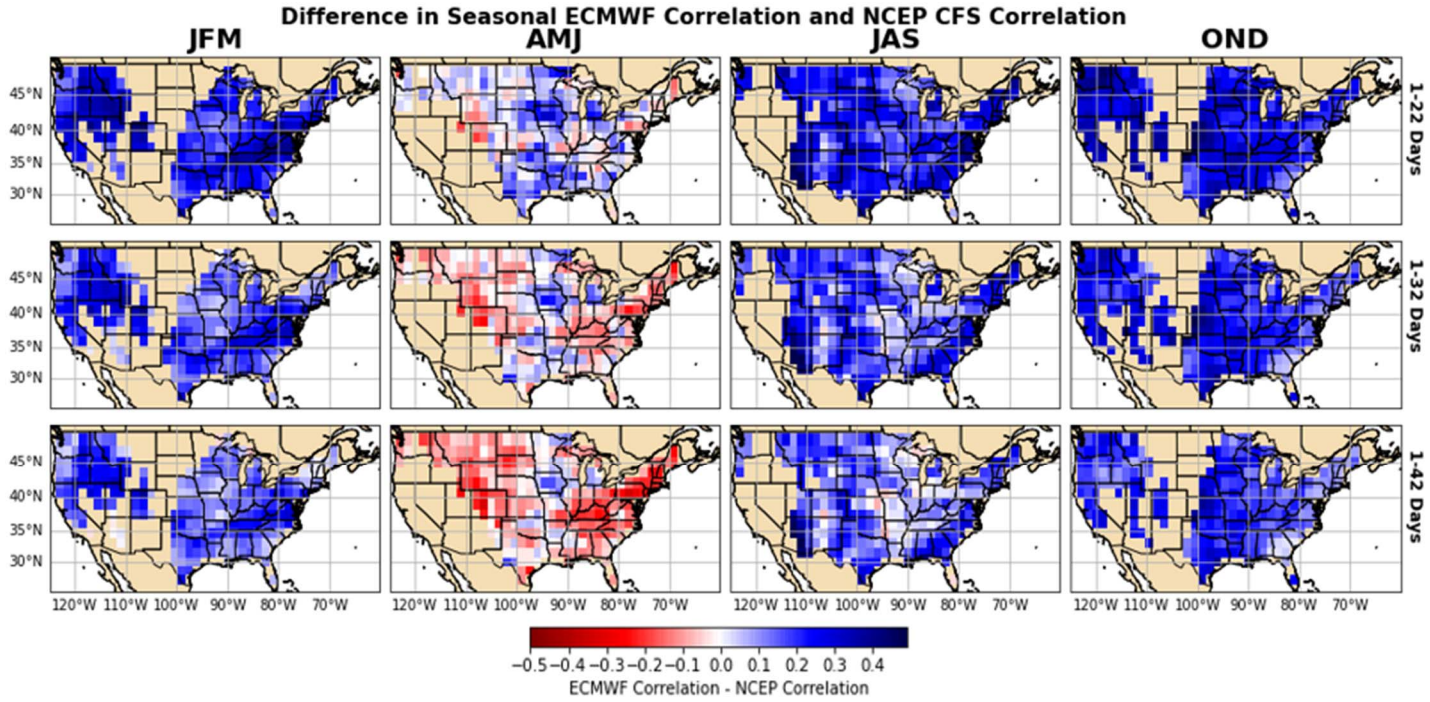
Figure 7. The box and whisker plot of correlation from the ECMWF hindcast model for three Koppen climate regimes: desert (red), temperate (blue) and continental (green) for lead times 1-12, 1-22, 1-32, and 1-42 days for all four seasons that the forecasts were initialized: JFM, AMJ, JAS, OND.

d. Model Comparison of NSE and Correlation

Comparing S2S hindcast models is important to understand the relative performance of the individual models. In this analysis, ECMWF's NSE was compared to NCEP CFS's NSE and next ECMWF's correlation was compared to all three models. The dry mask threshold may affect the model performance; therefore, forecast skill was not considered in areas where the historically observed precipitation did not exceed this threshold.

The blue regions in Figure SI-2 show where ECMWF's NSE outperforms the NSE of NCEP CFS for most lead times, regimes, and seasons, especially at shorter lead times, except for a few inland areas. Although ECMWF's NSE is higher than NCEP's in most regimes, seasons, and lead times, the ECMWF and NCEP CFS's correlation is closer in value (Figure 8). NCEP CFS has a higher NSE and correlation than ECMWF during AMJ. In comparison to ECMWF, NCEP's correlation improves with longer lead times during AMJ and is also higher in areas

365 further inland. Conversely, ECMWF has better performance around the coast (Figure 8) except
366 for OND, which may be due to the two different ocean models used in the initializations.



367 *Figure 8. Difference in Correlation values between ECMWF S2S hindcast and NCEP CFS for CONUS after dry*
368 *mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead*
369 *times: 1-12 days, 1-22 days, and 1-42 days.*
370
371 ECMWF and ECCC models' correlation differ by season but Figure 9 does not show a
372 clear inland-coastal differential in skill (Figure 9), which could be potentially due to ECMWF
373 and ECCC having the same ocean models. ECCC has a higher correlation than ECMWF during
374 the forecasts initiated in the summer months (JAS). However, since ECCC's lead time ranges
375 from 1-32 days, 1-42 day lead time between ECMWF and ECCC could not be compared.
376
377 Across seasons and lead times, NCEP CFS's correlation is higher than ECCC's
378 correlation for NCEP (Figure SI-3). NCEP CFS' model performance improves noticeably at
379 longer lead times and was not compared to 1-42 days lead time because of ECCC's shorter lead
380 time forecast availability. However, when comparing the first component, correlation, by
381 regime, season, and lead time, ECCC has higher correlation in AMJ, when compared to both
382 NCEP CFS as well as ECMWF. However, ECCC's performance tends to be worse in the
383 remaining three seasons.
384
385

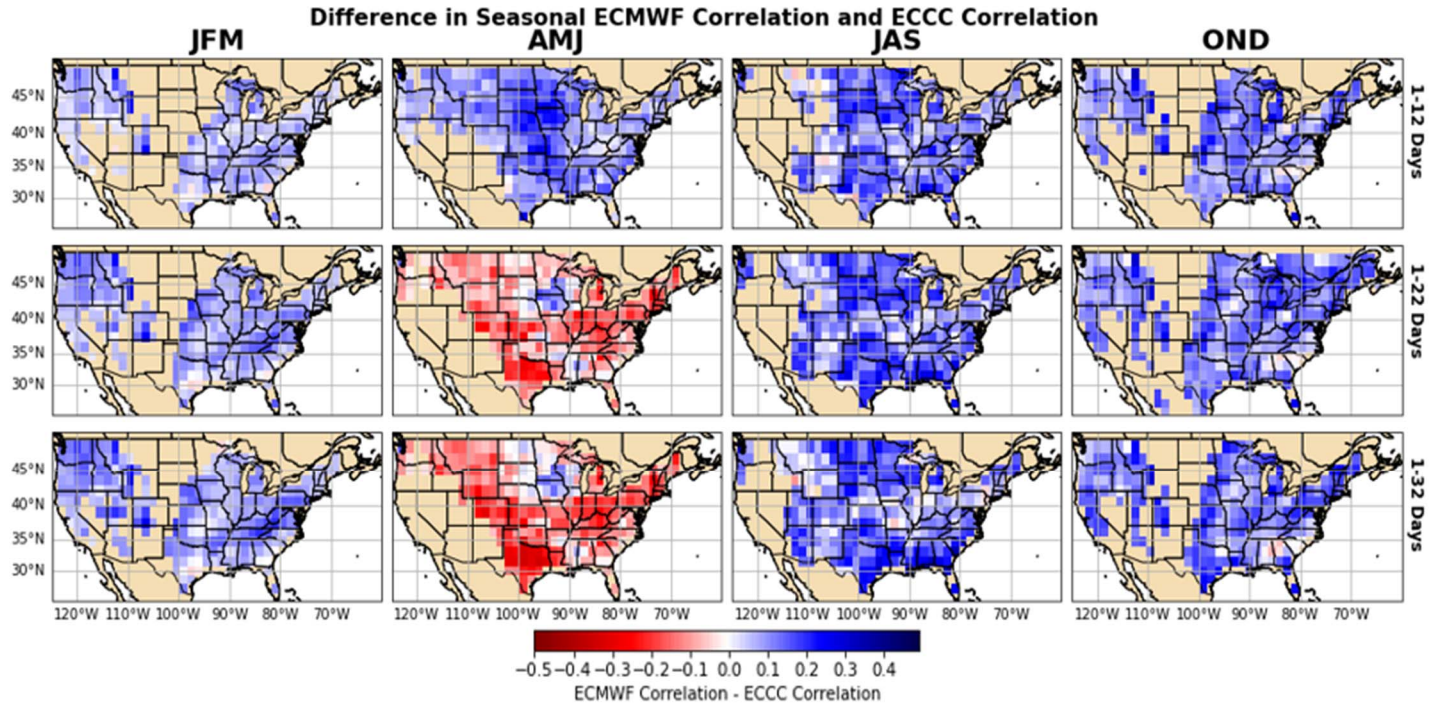


Figure 9. Difference in Correlation values between ECMWF S2S hindcast and ECCC for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-12 days, 1-22 days, and 1-32 days.

Overall, ECMWF's correlation for the forecast issued in seasons, JFM and OND, is higher than the other two models, but ECMWF's correlation is lower than the other models for the forecasts issued in AMJ (Figure 8-9). ECMWF has the highest NSE and correlation when solely considering the skill within the CONUS boundaries; however, NCEP CFS and ECCC hindcast models have much better forecast skill in the Great Lakes regime on and the Canadian regime just north of the Great Lakes, which although may not fall within the US boundaries, is still critical for the Midwest's water resources. ECMWF performs better towards the coasts and the skill may be higher in the winter seasons due to the areas that were masked out by the dry mask threshold. NCEP CFS and ECCC perform better in areas further inland, which is why the skill may be noticeably better in the spring and summer months (AMJ and JAS) where the inland regimes are not masked by the dry mask threshold since the regime receives higher precipitation during the summer. The differences in model skill could be due to the different ocean models and different initialization schemes, however this attribution has to be systematically analyzed further.

e. Model Skill comparison across Koppen Climate Regimes

The performance metrics for the three hindcast models were analyzed across the three Koppen climate regimes over the CONUS. Each model's NSE and the decomposed components were divided into climate regimes by season and lead times. At longer lead times, the differences in NSE reduces across seasons and climate regimes with NCEP CFS beginning to outperform

ECCC (Figure SI-4). ECMWF's NSE was higher than the NSE of ECCC and NCEP CFS across climate regime, season, and lead times (Figure SI-4), because NCEP and ECCC had high unconditional and conditional biases (Figure SI-4). Since these biases can be reduced to zero with simple post-processing techniques such as Model Output Statistics (Appendix A), we focused on comparing correlation (Figure 10).

The Pearson correlation coefficient is generally higher for ECMWF in comparison to ECCC and NCEP CFS models for all lead times, regimes, and seasons (Figure 10). There does not seem to be a consistent trend on how models perform for each climate regime across seasons and lead times even though both NCEP and ECCC perform better with forecasts issued in AMJ (Figure SI-4). For ECMWF and ECCC, the correlation is higher at shorter lead times, but NCEP's correlation remains relatively consistent across lead times (Figure 10). Across all models, lead times, and regimes the seasonal patterns illustrate that correlation is the highest during JFM and OND and lowest during AMJ and JAS.

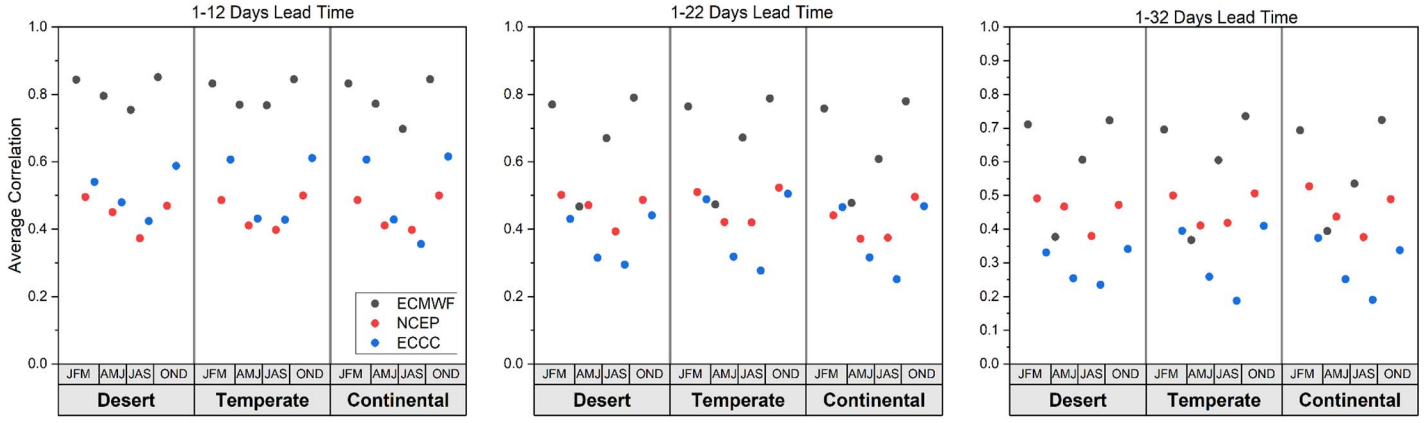


Figure 10. The average correlation for each regime: Regime B (desert), Regime C (temperate), and Regime D (continental) for each model: ECMWF (black), ECCC (blue), and NCEP CFS (red). The average correlation was calculated by lead time a) 1-12 days b.) 1-22 days and c.) 1-32 days for seasons JFM, AMJ, JAS, and OND.

The conditional bias is the lowest for ECMWF and highest for NCEP CFS particularly for AMJ and at shorter lead times (Figure SI-4). NCEP's median marginal bias was lower than ECMWF and ECCC, but one grid cell on the West Coast had a very high conditional bias causing the mean bias of all of the grid cells to be higher than the other two models.. ECCC has the highest conditional bias at the shorter lead times and ECMWF and NCEP CFS were comparable at 1-12 days for JFM, JAS, and OND. Conditional bias has the highest spread during spring months (AMJ). With longer lead times (e.g., 1-32 days), the unconditional bias across the selected models is similar, with ECCC being slightly higher than the other two models. No clear regional pattern of unconditional bias across all models and seasons was evident (Figure SI-4 g-i). The seasonality of unconditional bias seems to change based on lead times. We discuss in the next section how the conditional bias and unconditional bias could be potentially improved using post-processing techniques that focus on developing statistical relationships between model forecasts and the observed precipitation.

4.0 Discussion

Understanding the S2S precipitation forecasts skill across the CONUS over different seasons, as well as highlighting potential avenues for model improvement is critical for better forecast application. This study a) investigated and compared the spatial distribution of NSE for three S2S precipitation hindcast models across the CONUS, b) decomposed Nash-Sutcliffe Efficiency into correlation, conditional bias and unconditional bias based on the lead time and forecast issued in a season for each model and c) analyzed model skill across three (tropical, desert and temperate) Koppen Climate regimes. Our analysis shows that NSE of ECMWF was higher closer to the coast, most likely due to the influence of MJO and ENSO, and was also higher for the forecast issued during winter months and with shorter lead times. Decomposition of NSE shows that the first component, correlation, illustrates there is a gradient in skill from west coast (higher) to east coast (lower). Both the conditional and unconditional biases were also smaller during the winter months and in areas closer to the coast. The model comparison showed that ECMWF performs well in the winter seasons and towards the coasts, whereas NCEP CFS's performance is the best for forecasts issued during AMJ and in inland areas. The conditional and unconditional bias were high over the Midwest Great Lakes region. The conditional bias was higher for NCEP CFS, particularly for forecasts issued in AMJ and the unconditional bias was high for forecasts issued in JAS. ECCC's skill is high during AMJ and at short lead times, but decreases significantly with longer lead times. No clear trends were observed across the climate regimes across the three hindcast models' performances, but NSE and correlation was higher for the winter seasons than the summer seasons consistently for all the lead times, regimes and three models.

Potential for improving S2S forecasts

Even though our analysis, after application of dry mask, showed that conditional bias and unconditional bias are the primary reasons for low and negative NSE values for the S2S hindcasts, this could be overcome by selecting a proper post-processing scheme where the correlation is high across the CONUS. One of the commonly used post-processing scheme for correcting weather/climate forecasts is Model Output Statistics (MOS), which is a linear regression model that uses the forecast or a transformation of it (e.g., principal components) as a predictor and the observed precipitation as a predictand (Antolik et al., 2000; Sankarasubramanian et al., 2008). One advantage with a linear regression model is that it reduces the marginal bias to zero (Appendix A). Further, we also show analytically in Appendix A, a linear regression model reduces the conditional bias to zero which turns the NSE of the corrected forecasts from a MOS being equal to the square of the correlation coefficient (i.e., component A). Thus, a linear regression based MOS provides a lower bound on the NSE of the forecast to be decomposed component A, thereby providing a guidance on where post-processing schemes will be useful for a given location/regime. An example of where post-processing can be effective for correcting bias is NCEP CFS's 1-42 day forecasts. ECMWF did not have any grid cells where NSE was below zero, because the conditional and unconditional bias were low, so we

show NCEP, which has large sources of unconditional and conditional bias across all regimes, but relatively high correlation (SI-4).

Figure 11 shows locations where a) NCEP's NSE is less than zero and correlation is significant ($p<0.05$), b) NCEP's NSE is greater than zero and correlation is significant ($p<0.05$), and c) NCEP's NSE is less than zero, but correlation is not significant ($p>0.05$) for 1-42 day lead times. For the first case, where NSE is low and correlation is high, post-processing such as MOS can be effectively used to reduce conditional and unconditional biases to improve forecast skill, and a large portion of CONUS, mostly inland area and particularly for forecasts issued in seasons JFM and AMJ (Figure 11). For the second category, a large portion of the coastal region, particularly in forecast-initialized seasons AMJ and OND, have significant ($p<0.05$) correlation and high NSE, which means post-processing will not be effective as the model does not capture the observed variability. Similarly, post-processing will not be effective in areas with low NSE and correlation that is not significant ($p>0.05$), which includes a few grid points in AMJ and JAS (Figure 11). Even though linear-regression based MOS may not result in improved skill in areas where both NSE and correlation are low, other MOS post-processing schemes can be considered such as a semi-parametric model or machine learning models (Glahn et al., 1972; Taillardat et al., 2019), NSE of S2S forecasts could be potentially improved as such models are more flexible in reducing the mean square error in the forecast.

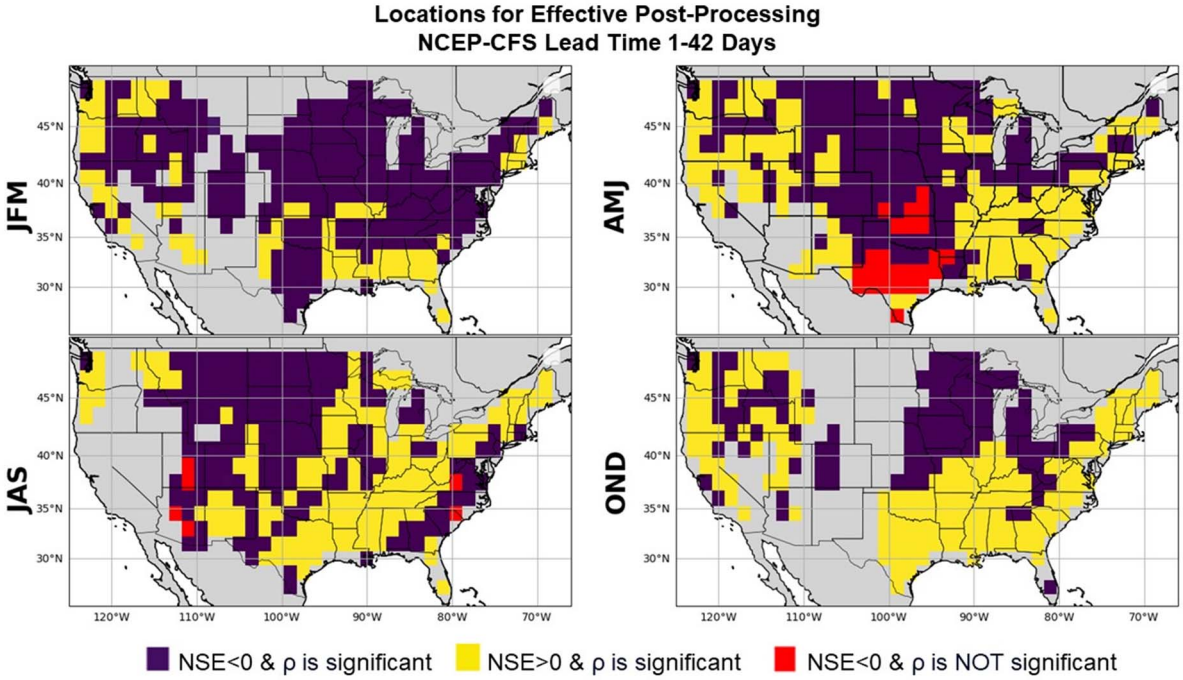


Figure 11. Post-processing will be effective in the locations where NSE<0 and correlation is significant (purple), but will not be necessary in places where NSE<0 but correlation is not significant (red) or in places where correlation is significant (yellow).

Even though the selected models had ensemble forecast, we considered only ensemble mean for forecast decomposition. We did not consider probabilistic forecasts such as Brier Skill score for skill evaluation and decomposition since the differences in ensemble members could

significantly affect the forecast evaluation. Similar decomposition on Brier score could reveal the forecast reliability and resolution of each model's performance in below-normal and above-normal conditions (Brier, 1950). Further, our analysis focused on decomposition without evaluating the model's performance during extreme conditions, which could be pursued further to understand the sources of bias. Our analysis also did not consider NMME because the number of ensemble members varies between models, giving more weight to some models. Additionally, the models within NMME have varying forecast issued frequencies, lead times, and issued dates. These varying model features within the multi-model need to be addressed before valid model comparisons can occur. Since the intent of this study was to show a systematic process of evaluating model skill and comparing across the models, we did not consider NMME for our study.

5.0 Conclusions

S2S precipitation forecasts are critical for operational and proactive water resource management and planning. Systematic S2S forecast skill assessment is essential for understanding existing model skill and how different errors contribute to it. Our evaluation of three S2S reforecasts – ECMWF, ECCC and NCEP – based on NSE decomposition primarily looked at the skill of forecasts issued during four seasons and under three different lead times. Our analysis shows the importance of applying dry mask as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. The full decomposition of ECMWF revealed a West to East coast longitudinal gradient in NSE and correlation. Decomposed components, conditional and unconditional bias, did not show any longitudinal trends. ECMWF's skill showed that seasonal trends in forecast skill occurred across all lead times and all seasons, but correlation did not differ by climate regimes.

The forecast skill and associated errors were also compared across models. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during the winter months, (JFM and OND). ECMWF had the highest NSE across the three climate regimes – temperate, desert and continental – considered. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in areas further away from the coast. Our decomposition efforts show S2S improvements in physical modeling efforts such as parameterization and initialization should be undertaken for ECMWF particularly for areas further from the coast, for forecasts issued in the spring months, AMJ, and for NCEP CFS and ECCC for the forecasts issued in the winter months over coastal areas.

Our analytical derivation on how MOS could help improve the forecast shows that a linear regression based MOS could ensure the NSE of the post-processed forecast to be component A, which is the square of the correlation coefficient between forecasts and the observation. This shows because simple linear regression based MOS can eliminate conditional

and marginal biases. This also provides information on regions (Figure 11, NSE <0 and ρ not significant) where S2S forecasting schemes can focus on improved model parameterizations and initializations including coupling with land surface models for improving the skill (Entekhabi et al., 1999).

Acknowledgments

The first author was supported by the National Science Foundation Fellowship (NSF) for the Graduate Research Fellowship Program (GRFP) support (award # DGE-2137100). Apart from that, this research was also supported by two NSF grants (award # CBET - 1805293 and IIE-2033607).

Open Research

The hindcast model data was accessed on the ECMWF S2S reforecast portal (https://apps.ecmwf.int/datasets/data/s2s/). The CPC Unified Gauged-Based observed precipitation dataset are available at https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html, and the Koppen climate classification data are available at www.gloh2o.org/koppen/.

Appendix A. Decomposition of NSE for Linear-Regression Based Model Output Statistics

For each grid cell, o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation value and y_{it} is the corrected precipitation value, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. Linear regression model 2 is used for the model to get the corrected precipitation value, which is the MOS estimate.

$$o_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \quad [1]$$

$$y_{it} = \beta_0 + \beta_1 x_{it} \quad [2]$$

For a given i, NSE is originally between o_{it} and x_{it} (equation 3), but a linear regression is used to estimate the corrected precipitation, y_{it}. For a given i , the NSE is calculated between o_{it} and y_{it} (equation 4) and then decomposed into parts A (equation 8-14), B (equation (15), and C (equation 16-17).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = \rho_{xo}^{\,2} - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad [3]$$

$$NSE_i(o_{it}, y_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-y_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = = \rho_{yo}^{\,2} - (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 - (\frac{\overline{y}-\overline{o}}{\sigma_o}))^2 \quad [4]$$

$$\beta_1 = \frac{cov(o-x)}{\sigma_x^{\,2}} \quad [5] \qquad \beta_0 = \overline{o} - \beta_1 * \overline{x} \quad [6] \qquad \beta_1 = \frac{\rho_{xo} * \sigma_x * \sigma_o}{\sigma_x^{\,2}} = \frac{\rho_{x\,o} * \sigma_o}{\sigma_x} \quad [7]$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 3. The pearson correlation coefficient between x and o is ρ_{xo}. For the corrected precipitation, y_{it}, the standard deviation and mean are σ_y and $\overline{y}$ respectively, when y_{it} is summed over time from 1 to n for lead time i in equation 4. The correlation coefficient between o and y is ρ_{yo} .

NSE of o_{it} and y_{it} is decomposed into the three corresponding parts a.) correlation, b.) conditional bias and c.) unconditional bias. It is important to note that correlation, Component A (o, y), will be the same as the Component A (o, x). Where

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [8]$$

$$\rho_{xo} = \frac{cov(x, o)}{\sigma_x * \sigma_o} \quad [9]$$

$$cov(y, o) = cov(\beta_0 + \beta_1 x, o) = \beta_1 cov(x, o) \quad [10]$$

$$var(y) = \beta_1^{\,2} * \sigma_o \quad [11] \qquad \sigma_y = \beta_1^{\,2} * \sigma_x \quad [12]$$

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [13]$$

$$\rho_{yo} = \frac{\beta_1 * cov(y, o)}{\beta_1 * \sigma_x * \sigma_o} = \frac{\beta_1 * cov(y, o)}{\sigma_y * \sigma_o} = \rho_{xo} \quad [14]$$

Conditional bias B (o, y) will be reduced to zero MOS estimates.

$$B(o, y) = (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 = (\rho_{xo} - \frac{\beta_1 * \sigma_x}{\sigma_o})^2 = (\rho_{xo} - \frac{\sigma_y}{\sigma_o})^2 \quad [15]$$

638 $$\rho_{xo} = (-(\frac{\rho_{xo} - \sigma_o}{x}) \cdot \frac{\sigma_x}{\sigma_o})^2 = 0$$

639 Unconditional bias C (o , y) will also reduce to zero for MOS estimates.

640 $$C(o, y) = (\frac{\overline{y} - \overline{o}}{\sigma_o}))^2 \qquad [16]$$

641 $$\overline{y} = \beta_0 + \beta_1 * \overline{x} = \overline{o} - \beta_1 * \overline{x} * + \beta_1 * \overline{x} \qquad [17]$$

642 $$C(o, y) \to 0$$

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

References

Antolik, M.S. (2000). An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *Journal of Hydrology*, 239, 306- 337.

Barbero, R., Fowler, H. J., Blenkinsop, S., Westra, S., Moron, V., Lewis, E., et al. (2019). A synthesis of hourly and daily precipitation extremes in different climatic regions. *Weather and Climate Extremes*, 26, 100219. https://doi.org/10.1016/j.wace.2019.100219.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, *Hydrol. Earth Syst. Sci*., 21, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017.

‘

Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. *Geophysical Research Letters,* 47, e2020GL087408. https://doi.org/10.1029/2020GL087408

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

Carter, G. M., Dallavalle, J.P., & Glahn, H.R. (1989). Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401– 412.

685 Chalise, D. R., Sankarasubramanian, A., Olden, J. D., & Ruhi, A. (2023). Spectral signatures of
686 flow regime alteration by dams across the United States. *Earth's Future,* 11, e2022EF003078.
687 https://doi.org/10.1029/2022EF003078
688
689 Charba, J. P., and F. G. Samplatsky, 2011: High-Resolution GFS-Based MOS Quantitative
690 Precipitation Forecasts on a 4-km Grid. *Mon. Wea. Rev.,* 139, 39–68,
691 https://doi.org/10.1175/2010MWR3224.1.
692
693 Chen, M., W. Shi, P. Xie, V. B. S. Silva, V E. Kousky, R. Wayne Higgins, & J. E. Janowiak
694 (2008), Assessing objective techniques for gauge-based analyses of global daily precipitation, *J.*
695 *Geophys. Res*., 113, D04110, doi:10.1029/2007JD009132.
696
697 Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al.
698 (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources*
699 *Research*, 57, https://doi.org/10.1029/2020WR029001.
700
701 de Andrade, F.M., Coelho, C.A.S. & Cavalcanti, I.F.A. (2019). Global precipitation hindcast
702 quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Clim Dyn,*
703 52, 5451–5475. https://doi.org/10.1007/s00382-018-4457-z.
704
705 Entekhabi, D., and Coauthors (1999). An agenda for land-surface hydrology research and a call
706 for the second International Hydrological Decade. *Bull. Amer. Meteor*. Soc., 80, 2043–2058,
707 https://doi.org/10.1175/1520-0477(1999)080<2043:AAFLSH>2.0.CO;2.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, 11, 1203–1211.

Glahn, H. R., & Ruth D. P. (2003). The New Digital Forecast Database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, 84 , 195–201.

Goddard, L., Kumar, A., Solomon, A. et al. (2013). A verification framework for interannual-to-decadal predictions experiments. Clim Dyn, 40, 245–272. https://doi.org/10.1007/s00382-012-1481-2.

Gupta, H. V., Kling, H., Yilmaz, K. K., & G. F. Martinez (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.,* 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

Hamill, T.M. & Colucci, S.J., (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6), pp.1312-1327.

Ham, YG., Kim, JH. & Luo, JJ. (2019). Deep learning for multi-year ENSO forecasts. *Nature* 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7

729 Konapala, G., Mishra, A.K., Wada, Y. et al. (2020). Climate change will affect global water
730 availability through compounding changes in seasonal precipitation and evaporation. *Nat*
731 *Commun*, 11, 3044. https://doi.org/10.1038/s41467-020-16757-w
732
733 Krakauer, N.Y. (2019). Temperature trends and prediction skill in NMME seasonal forecasts.
734 Clim Dyn, 53, 7201–7213. https://doi.org/10.1007/s00382-017-3657-2
735 Li, Y., Tian, D., & Medina, H. (2021). Multimodel Subseasonal Precipitation Forecasts over the
736 Contiguous United States: Skill Assessment and Statistical Postprocessing. J. Hydrometeor., 22,
737 2581–2600, https://doi.org/10.1175/JHM-D-21-0029.1.
738
739 Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier,
740 D.P., Stouffer, R.J. (2008). Stationarity is dead: Whither water management?. *Science,* 319
741 (5863): 573-574. https://doi.org/10.1126/science.1151915
742
743 Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the
744 correlation coefficient. *Monthly Weather Review*, 116, 2417– 2424.
745
746 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I—
747 A discussion of principles. *Journal of Hydrology,* 10(3), 282–290. https://doi.org/10.1016/0022-
748 1694(70)90255-6
749
750 Petersen, T., Devineni, N., & A. Sankarasubramanian, A. (2012). Seasonality of monthly runoff
751 over the continental United States: Causality and relations to mean annual and mean monthly

distributions of moisture and energy, *J. Hydrol*., 468–469, 139–150, 10.1016/j.jhydrol.2012.08.028.

Quan, X., Hoerling, M., Whitaker, J., Bates, G., & T. Xu, T. (2006). Diagnosing Sources of U.S. Seasonal Forecast Skill. J. Climate, 19, 3279–3293, https://doi.org/10.1175/JCLI3789.1.

Sankarasubramanian, A., Sabo, J.L., Larson, K.L., Seo, S.B., Sinha, T., Bhowmik, R., Vidal, A.R., Kunkel, K., Mahinthakumar, G., Berglund, E.Z. & Kominoski, J. (2017), Synthesis of public water supply use in the United States: Spatio-temporal patterns and socio-economic controls. *Earth's Future*, 5: 771-788. https://doi.org/10.1002/2016EF000511

Sankarasubramanian, A., Lall, U., & Espinueva, S. (2008). Role of Retrospective Forecasts of GCMs Forced with Persisted SST Anomalies in Operational Streamflow Forecasts Development. *J. Hydrometeor*., 9, 212–227, https://doi.org/10.1175/2007JHM842.1.

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T., (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, 28(4), pp.616-640.

Scheuerer, M., & Hamill, T.M., (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Rev*.w, 143(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

775 Sun, L., Hoerling, M.P., Richter, J.H., Hoell, A., Kumar, A. & Hurrell, J.W., (2022). Attribution
776 of North American Subseasonal Precipitation Prediction Skill. *Weather and Forecasting*, 37(11),
777 pp.2069-2085.
778
779 Taillardat, M., Fougères, A., Naveau, P., & Mestre, O. (2019). Forest-Based and Semiparametric
780 Methods for the Postprocessing of Rainfall Ensemble Forecasting. *Wea. Forecasting*, 34, 617–
781 634, https://doi.org/10.1175/WAF-D-18-0149.1.
782
783 Vitart, F., Robertson, A. & Anderson, D. (2012). Sub-seasonal to Seasonal Prediction Project:
784 bridging the gap between weather and climate. WMO Bull. 61, 23–28.
785
786 Vitart, F., Robertson, A.W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the
787 prediction of extreme events. *npj Clim Atmos Sci* 1, 3. https://doi.org/10.1038/s41612-018-0013-
788 0
789
790 Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti,
791 E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P.
792 Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami,
793 R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A.W. Robertson, P. Ruti, C. Sun, Y.
794 Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. Won, H. Xiao, R.
795 Zaripov, and L. Zhang, (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database.
796 gi. *Amer. Meteor. Soc*., 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.
797

798 Vitart, F., Robertsn, A. W., & S2S Steering Group (2015). Sub-seasonal to seasonal prediction:
799 Linking weather and climate. In: Seamless Prediction of the Earth System: From Minutes to
800 Months. (pp. 385– 401). WMO-No.1156 (Chapter 20). Retrieved from
801 http://library.wmo.int/pmb_ged/wmo_11
802
803 Wang, L., & Robertson, A.W. (2019). Week 3–4 predictability over the United States assessed
804 from two operational ensemble prediction systems. *Clim Dyn,* 52, 5861–5875.
805 https://doi.org/10.1007/s00382-018-4484-9.
806
807 Weglarczyk S (1998), The interdependence and applicability of some statistical quality measures
808 for hydrological models. *Journal of Hydrology,* 206: 98-103. https://doi.org/10.1016/S0022-
809 1694(98)00094-8.
810
811 Weigel, A.P., Liniger, M.A., & Appenzeller, C. (2008), Can multi-model combination really
812 enhance the prediction skill of probabilistic ensemble forecasts?. *Q.J.R. Meteorol. Soc.,* 134:
813 241-260. https://doi.org/10.1002/qj.210
814
815 White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F.,
816 Coughlan de Perez, E., Ray, A.J., Murray, V., & Bharwani, S., 2017. Potential applications of
817 subseasonal-to-seasonal (S2S) predictions. *Meteorological applications*, 24(3), pp.315-325.
818
819 Wilks, D. Statistical Methods in the Atmospheric Sciences (Academic, 2006).
820

821 Wood, A. W., Maurer, E.P., A. Kumar, & Lettenmaier, D.P. (2002). Long-range experimental
822 hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* 107, 4429, doi:
823 10.1029/2001JD000659.
824
825 Zhang, C. (2013). Madden–Julian Oscillation: Bridging weather and climate. *Bulletin of the*
826 *American Meteorological Society*, 94, 1849–1870. https://doi.org/10.1175/BAMS-D-12-00026.1
827
828 Zhang, L., Kim, T., Yang, T., Hong, Y. & Zhu, Q. (2021). Evaluation of Subseasonal-to-
829 Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II
830 (NMME-2) over the contiguous US. *Journal of Hydrology*, 603, p.127058.
831 https://doi.org/10.1016/j.jhydrol.2021.127058
832

Spatial and Temporal Variation of Subseasonal-to-Seasonal (S2S) Precipitation Reforecast Skill Across CONUS

J. R. Levey[1], A. Sankarasubramanian[1]

[1]Department of Civil, Construction and Environmental Engineering, North Carolina State University.

Corresponding author: Jessica Levey (jrlevey@ncsu.edu)

Key Points:

- NSE decomposition of S2S reforecast skill shows the spatio-temporal variations in correlation, conditional and unconditional bias.
- Longitudinal gradient of forecast skill exists from the West (higher) to East (lower).
- Regression based model-output statistics provide correlation as the lower bound of NSE as the marginal and conditional bias reduces to zero.

Abstract

Precipitation forecasts, particularly at subseasonal-to-seasonal (S2S) time scale, are essential for informed and proactive water resources management. Although S2S precipitation forecasts have been evaluated, no systematic decomposition of the skill, Nash-Sutcliffe Efficiency (NSE) coefficient, has been analyzed towards understanding the forecast accuracy. We decompose the NSE of S2S precipitation forecast into its three components – correlation, conditional bias, and unconditional bias – by four seasons, three lead times (1–12-day, 1-22 day, and 1-32 day), and three models (ECMWF, CFS, NCEP) over the Conterminous United States (CONUS). Application of dry mask is critical as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. Further, a west-to-east gradient in S2S forecast skill exists and forecast skill was better during the winter months and for areas closer to the coast. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during fall and winter months. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in in-land areas. Post-processing using simple Model Output Statistics could reduce both unconditional and conditional bias to zero, thereby offering better skill for regimes with high correlation. Our decomposition results also show efforts should focus on improving model parametrization and initialization schemes for climate regimes with low correlation values.

1.0 Introduction

Global climate change and regional anthropogenic disturbances, including urbanization and deforestation, are driving shifts in the hydrologic cycle, and impacting water resources (Konapala et al, 2020; Milly et al., 2008). Consequently, extreme precipitation events, including prolonged droughts or flooding, are expected to be more frequent, further threatening water supply and variability (Milly et al., 2008). In conjunction with hydroclimatic changes, population changes also stress surface and groundwater resource withdrawals in many regions across the Conterminous US (CONUS) (Sankarasubramanian et al., 2017). Reservoir releases, during both floods and droughts, are modified for human needs, downstream ecological health, and for ensuring watershed resilience (Chalise et al., 2021). Mismanagement of water resources, both surface water and groundwater, may pose threats to agriculture, supply chains, human and environmental health, and regional economies. Hence, reliable and accurate subseasonal-to-seasonal (S2S) precipitation forecasts are essential in an age of a changing climate for improving water management strategies and preparing for near-future hydroclimatic extremes.

Compared to the skill of short-range weather forecasts (less than 15 days) and long-range seasonal forecasts, which are reasonably good, the skill of S2S forecasts, ranging between 15 to 60 days, is low and is often referred to as the 'predictability desert' (Vitart et al., 2012). Understanding the current S2S precipitation forecasts skill, as well as highlighting the potential avenues – initialization, parametrization, and post-processing schemes – for improvement are critical for accurate S2S precipitation forecasts for operational use (White et al., 2017). Known contributing factors that influence S2S model forecasting performance include the parametrization and initialization schemes, large-scale atmospheric circulation modes, and coupled models (Vitart et al., 2018). The model initialization scheme, including land surface and soil moisture representation, are also crucial for accurately representation of geophysical fluxes.

Climate oscillations, such as El Nino Southern Oscillations (ENSO) and Madden-Julian Oscillations (MJO) also influence seasonal forecast prediction skill (Zhang, 2013). ENSO's influence on United States' winter hydroclimatology is well-known, particularly over the Southeast and west coast, accounting for roughly a third of US winter forecasting skill (Quan et al., 2006).

Previous studies have attributed S2S skill between ENSO and MJO (Sun et al., 2022; Wang et al., 2019) and have compared S2S skill across models, lead times and seasons (Zhang et al 2021; de Andrade et al, 2019). However, these studies that examined S2S models' forecasting performance did not apply a threshold on dry mask prior to calculating the model's skill. Zhang et al (2021) have evaluated S2S forecast skill by filtering extreme precipitation events, but did not apply a dry mask threshold for evaluating the overall skill. Without a dry mask threshold, the S2S skill will be inflated, especially in regions with a pronounced dry season, as zero rainfall days is included in these skill calculations (Wilks, 2006). The ability to predict days without precipitation is important for drought prediction and planning, but the skill will be inflated for wetter and normal conditions; therefore, the dry mask application was used to filter out areas of inflated skill based on the climatological means. Several studies focused on extreme precipitation forecasts have applied percentile filters (Zhang et al., 2021), which reduces the sample size particularly while evaluating monthly/seasonal skill. Given the pronounced seasonality in precipitation over the CONUS (Petersen et al., 2012), we systematically evaluate the S2S forecasting skill across CONUS by applying a dry mask before considering the skill for each lead time, season and region. Evaluating the forecast skill after applying the dry mask could potentially affect the source of model skill, and the associated biases that could be obtained from decomposition.

S2S precipitation forecast skill has been compared considering both probabilistic and deterministic metrics to evaluate the forecast skill (Zhang et al., 2021; de Andrade et al., 2019). S2S models' skill have been evaluated using Mean Square Error (MSE), mean square skill score (MSSS), root mean square error (RMSE), anomaly correlation coefficient (ACC), Pearson's correlation coefficient, and ranked probability skill score (RPSS) (Zhang et al., 2021; de Andrade et al., 2019). de Andrade, et al., (2019) evaluated hindcast skill using linear correlation coefficient and analyzed the sources of bias and variability; however, this study was a large-scale global analysis of forecast skill and did not consider the seasonal skills and the associated errors. Decomposing the MSSS three components – correlation coefficient, condition bias and marginal bias – would provide information on the regions and seasons over which the selected models have the ability to capture the variability in observed precipitation but have significant biases in estimation. Further, the hindcast assessment of (de Andrade et al., 2019) was performed without the dry mask application, which may inflate forecast skill particularly for regions with pronounced dry season.

The Nash-Sutcliffe Efficiency (NSE), also known as the coefficient of determination, is a metric that measures the skill of hydrologic models (Nash & Sutcliffe, 1970). Li et al., (2022) used to evaluate S2S forecast skill performance based on Kling-Gupta Efficiency (KGE) metric, which provides a different decomposition of NSE, without applying the dry mask across the CONUS or considering seasonality. However, decomposing the Nash-Sutcliffe Efficiency (NSE) for precipitation hindcasts after applying the dry mask provides critical information without inflating the skill of the model. Furthermore, implementing new parametrizations and

initialization schemes could be costly and take additional time to develop reforecasts. One effective way to improve the forecasting skill is to consider post-processing schemes (Carter et al., 1998; Glahn et al., 2003). Further, post-processing could also be implemented over reforecasts from multiple models to develop multi-model ensembles which have been shown to improve the forecast skill compared to the best individual model (Weigel et al., 2008). Past work on statistical post-processing has considered both parametric and non-parametric approaches (Hamill et al., 1997; Schefzik et al., 2013; Scheuerer et al., 2015). Although many studies have used post-processing schemes on S2S precipitation forecasts, understanding the components of S2S forecast skill could provide additional insights on how post-processing schemes can be used and could also indicate potential regions where improvements in models will be needed to further improve the forecast skill.

Several S2S models that contribute multi-model ensembles have been run for reforecasts. Historically, some S2S multi-model datasets have only been running for a period of short time, limiting the ability to capture the interannual variability in precipitation. Other multi-model ensembles have primarily focused on generating monthly forecasts for seasonal prediction with infrequent model initialization. This study uses three individual models hindcasts from the World Weather Research Programme (WWRP) and World Climate Research Programme (WCRP) S2S prediction project (Vitart et al., 2012). The S2S project, originating in 2013, has a long record of forecasts and reforecasts that are initialized multiple times a week (Vitart et al., 2017). The longer range of data allows for larger sample sizes for robust estimation of NSE and decomposition metrics. Comparing model performance is important because forecast skill varies between S2S models as each model has different parameterization schemes, number of ensembles, and resolution. This study will consider decomposition of NSE of S2S reforecasts over the CONUS for three models – European Centre of Medium-Range Weather Forecast's (ECWMF) National Centre for Environmental Prediction Climate Forecast System (NCEP CFS) and Environment and Climate Change Canada (ECCC) – after applying the dry mask. Previous studies have shown ECMWF S2S hindcast models have outperformed both CFS and ECCC models on a global basis (de Andrade et al, 2019), but the performance of these three models have not been compared after the dry mask threshold has been applied. The North American Multi-Model Ensemble (NMME) forecasts have proved to perform better than individual models by pooling the ensemble members from several models (Krakauer, 2019). However, for this study, the NMME was not considered because the number of ensemble members varies between individual models, giving more weight to some models. Additionally, to improve multi-model performance, understanding individual models' type of errors and potential for correcting the biases before pooling the ensembles, which could further improve the multi-model forecast performance. Hence, this study will compare the decomposed NSE and associated errors of S2S precipitation forecasts of three individual models by season and lead time under three Koppen climate regimes across the CONUS.

The main intent of this study is to decompose the S2S forecasting skill as a function of lead time over the CONUS after applying the dry mask. To our knowledge, limited/no work has been performed on systematically decomposing the NSE over various seasons after applying the dry mask. In addition to applying the dry mask, evaluating model skill regionally is also critical as the precipitation has pronounced seasonality over the CONUS (Petersen et al., 2012).

Analyzing forecasting skills regionally can also provide insights on how land surface conditions, low-frequency oscillations, and regional hydroclimate influence the model performance.

The manuscript is organized into the following sections: S2S precipitation hindcast and observed databases from three different models are provided in the next section. Then, the dry mask threshold application procedure is presented along with the NSE decomposition. The following section provides the results from the full decomposition of ECMWF and the results from different regimes along with the skill comparison from three S2S reforecasts.

2.0 Data

This section provides the S2S hindcast database and observed data along with the details to calculate and decompose the NSE for S2S forecasts over various lead times and seasons.

Observed Precipitation

For calculating the S2S reforecasts skill, we used the CPC Global Unified Precipitation dataset provided by the NOAA Physical Science Laboratory (PSL), with a resolution of (0.5°x0.5°) (Chen, et al., 2008). Upon comparing the accuracy of various precipitation datasets, the CPC Unified dataset performed particularly well in areas that have dense areas of rain gauges (Beck et al., 2017). This study focused on the CONUS, which has a dense system of rain gauges, and has been used in other forecast verification studies (Becker et al, 2020).

S2S Hindcast Database

For S2S model skill evaluation, three hindcast models were assessed: 1.) European Centre of Medium-Range Weather Forecasts (ECMWF), 2.) National Center for Environmental Prediction's (NCEP) Climate Forecast System (CFS) model, and 3.) Environment and Climate Change Canada (ECCC). For full decomposition of ECMWF, the S2S hindcasts were evaluated for the full 20-year hindcast period (Table 1) and up to the longest available lead time of 45 days. The ensemble means were averaged over three different lead times: 1) 1-15 days, 2) 1-30 days, and 3) 1-45 days, and compared with the observed average precipitation corresponding to the three lead times. Additionally, the average forecasts and corresponding observed average daily precipitation values were pooled by the date of hindcast initialization into the following seasons: a) January, February, March (JFM), b) April, May, June (AMJ), c) July, August, September (JAS), d) October, November, December (OND). Thus, the evaluation for each season provides the skill of forecasts issued during the months within the considered four seasons.

For the model comparison section, the three models were assessed for lead times of 1-12 days, 1-22 days, and 1-32 days for four different seasons between January 1st 2000 and December 30th 2010, the longest available overlapping date ranges and lead times for all three models. Additionally, ECMWF and NCEP were compared for lead times of 1-42 days. The ECMWF hindcasts are initialized twice a week and range from 2000-2019, NCEP CFS hindcasts are initialized daily and are available from 1999-2010, and ECCC are initialized weekly, and

reforecasts range from 1995-2012 (Vitart et al., 2017). The S2S precipitation hindcast model's information and specification are shown in Table 1 (Vitart et al., 2017).

Model	LEAD TIME	RESOLUTION	HINDCAST PERIOD	HINDCAST ENSEMBLE SIZE	FORECAST ENSEMBLE SIZE	HINDCAST FREQUENCY	OCEAN COUPLING	SEA ICE COUPLING
ECMWF	0-46 Days	0.25°x0.25°, days 0-10, 0.5°x0.5°, after day 10 L91	Past 20 Years	11	51	Twice a Week	Yes	No
NCEP CFS	0-44 Days	~1°x1°, L64	1999-2010	4	16	Daily	Yes	Yes
ECCC	0-32 Days	0.45°x0.45°, L40	1995-2012	4	21	Weekly	Yes	No

Table 1. Subseasonal-to-Seasonal Hindcast Models and Forecast model information

2.1 Dry Mask application and Skill Assessment and Decomposition

a. Seasonality of Rainfall and Dry Mask Application

Prior to calculating the NSE for each hindcast-initialized season, a dry mask was applied based on the observed precipitation dataset to filter out the areas that receive small amounts of rainfall, which may result in an inflated forecast skill because the forecasted and observed rainfall have no rainfall. Antolilk (2000) and Charba et al., (2011) considered daily precipitation less than 0.01 inches as no event for evaluating the skill. Based on that work, the dry mask was set at a threshold value for each individual grid cell, if the observed daily precipitation over the 20 years is less than 0.15 inches, 0.30 inches and 0.45 inches for 15-day, 30-day and 45-day lead times from the time of issued forecast, respectively. The NSE and the three components were evaluated for all the three models for each lead time over the CONUS. We also evaluate the forecast skill – NSE and its components – based on the climate regime. For this purpose, we considered three main regimes – desert (regime B), temperate (regime C) and continental (regime D) – over the CONUS based on Koppen climate classification. A small area in southern Florida fell into the tropical (regime A) Koppen climate group; however, since this regime corresponds to only one grid cell from the hindcast model, we combined this tropical area with the temperate regime (Supplemental Information (SI) - Figure SI-1). Using the aggregated Koppen Climate Regime (Beck, et. al, 2017) into three climate regimes, a regional analysis was performed for each of the S2S hindcast models (Supplemental Information (SI) - Figure SI-1).

b. Skill Assessment Metrics

Skill assessment metrics measure the performance of the model's forecast ability compared to the observed variable. Frequently used performance metrics include anomaly correlation, NSE and Kling Gupta Efficiency (Clark et al., 2021). The NSE measures the magnitude of error variance from the model prediction compared to the observed variance in the data and has an upper bound of 1 but has a lower bound of -∞ and is used to determine the 'goodness-of-fit' of a

model. NSE is related to MSE but is normalized by the standard deviation of the observed precipitation or data values (Gupta et al., 2009).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\overline{o}_{it})^2} \quad (1)$$

Where o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. The mean observed precipitation is $\overline{o}_{it}$. For a given i, NSE will be decomposed into three parts (Murphy 1988; Weglarczyk 1998): A) Pearson's correlation coefficient (equation 3), B) conditional bias (equation 4), and C) unconditional bias (equation 5) (Gupta et al., 2009).

$$NSE = A - B - C$$

$$NSE = \rho_{xo}{}^2 - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad (2)$$

$$A=\rho_{xo}{}^2 \quad where \quad \rho_{xo} = \frac{cov(x,o)}{\sigma_x * \sigma_o} \quad (3)$$

$$B=[\rho_{xo} - \frac{\sigma_x}{\sigma_x}]^2 \quad (4)$$

$$C=[\frac{\overline{x}-\overline{o}}{\sigma_o}]^2 \quad (5)$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 1. The pearson correlation coefficient between x and o is ρ_{xo} (equation 3). The first component of the decomposition, Pearson's correlation coefficient, shows the linear association between the forecast and the observation. The conditional bias is the difference in the slope of the regression line fitted between forecast and observation with a slope of 1 that indicates a perfect forecast. The unconditional bias, indicating a systematic bias, denotes the ratio of difference between the mean of the observation and the mean of the forecast to the observed standard deviation.

3.0 Results

Full Decomposition of ECMWF

A full NSE decomposition was performed on the ECMWF S2S hindcast model because the ECMWF model has the longest available reforecast time range and has the largest number of ensemble members. Prior to decomposing NSE, a dry-mask threshold was applied based on the lead time for the climatological means of each grid cell, to mask out areas with low precipitation values to avoid inflated skill values. Both NSE and correlation are lower across all seasons after the dry mask threshold was applied. Figure 1a illustrates the difference in Normalized Nash-Sutcliffe Efficiency (NNSE) of 30-day ahead S2S precipitation forecast skill with and without the dry mask threshold). For instance, a forecast issued on March 30, 2000 with a lead time of 45 days corresponds to the skill of the forecast in predicting precipitation from March 30, 2000 to May 15, 2000. Thus, the skill of the forecast issued in JFM can cover the observed precipitation in April and May. To reiterate, all the figures with seasonal S2S performance metrics denote the skill summary of the forecast issued during that season as opposed to the ability to forecast the observed precipitation during that season.

To understand the importance of dry masking, we first show the 1-30 day ahead S2S precipitation forecast skill with and without dry mask (Figure 1) based on Normalized NSE (NNSE). Lower NNSE (equation 6) values, the inverse of NSE, indicate better predictive performance.

$$NNSE = \frac{1}{2 - NSE} \qquad (6)$$

For the forecast issued in the four seasons, the mean NNSE values are lower for the grid cells below the dry mask threshold than for the grid cells that exceeded the threshold (Figure 1). Even though including "no-precipitation event" is expected to inflate the skill, dry masking by filtering out regimes rather than simply removing values below a given threshold, allows us to maintain the same sample size across all grid cells, thereby changing the masked areas based on both forecast-initialized seasons (Figure 1) and lead time.

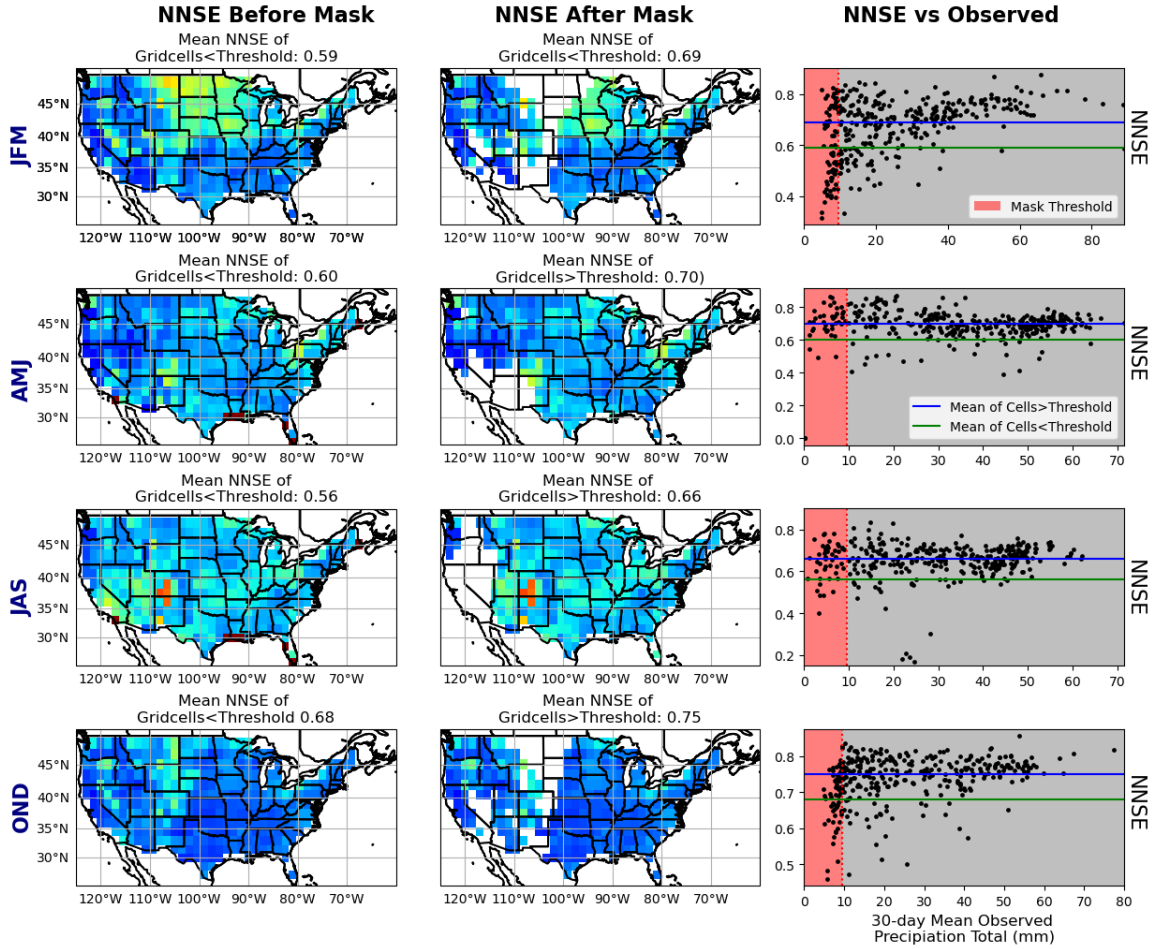


Figure 1. Normalized Nash Sutcliffe Efficiency (NNSE) of 1-30 days ahead ECMWF hindcast for the CONUS before dry mask is applied (left column) and after (middle column) dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS and OND for 1-30-day lead time. The scatter plot comparison of grid cell's 1-30-day climatological precipitation means and the corresponding Normalized NSE values (right column). The scatter plot shows the NNSE values that fall below the dry mask threshold (red region) and above (gray region). The average NNSE of the grid cells below the dry mask threshold (green) and above the dry mask threshold (blue). Since the NNSE is the inverse of the NSE, the lower NNSE values indicate better predictive performance.

The overestimation of S2S forecast skill occurs if no dry mask is applied, particularly for pronounced dry seasons (JFM and JAS). Studies that evaluated S2S precipitation forecasts skill did not consider dry mask application, which ignores the seasonality in precipitation, thereby indicating potential difference in forecast skill between regions (e.g., Li et al., 2022). However, after the dry mask application (Figure 1), we find that the skill was fairly similar between regimes. Thus, it is important to apply a dry mask which inherently considers the seasonality in precipitation for skill evaluation. Quantifying the forecast skill for critical events (e.g., peak rainfall seasons) is important particularly if the interest is to identify regions with limited skill.

a) NSE Spatial Patterns

We present results for the NSE and its decomposition (Figures 2-7) for the ECMWF model and then compare its performance with NCEP and ECCC later (Figures 8-10). Before assessing the components of the NSE, we first investigate the NSE over the CONUS, which shows the S2S forecasting skill of ECMWF for various lead times over the season (Figure 2). NSE is better in the winter and fall seasons (JFM and OND) in comparison to spring and summer seasons (AMJ and JAS) (Figure 2), which is partially due to El Nino Southern Oscillation (ENSO) being active during winter and fall months and ENSO dying or being at an incipient stage during AMJ and JAS (Ham et al., 2019). The NSE also tends to be better closer to the coasts indicating the local sea surface temperatures (SSTs) in influencing S2S forecasts. Additionally, the NSE shows a slight gradient from West Coast to East Coast (Figure 2). The NSE tends to be weaker around the Great Lakes. Further, the areas surrounding the dry mask regions tend to have a lower NSE.

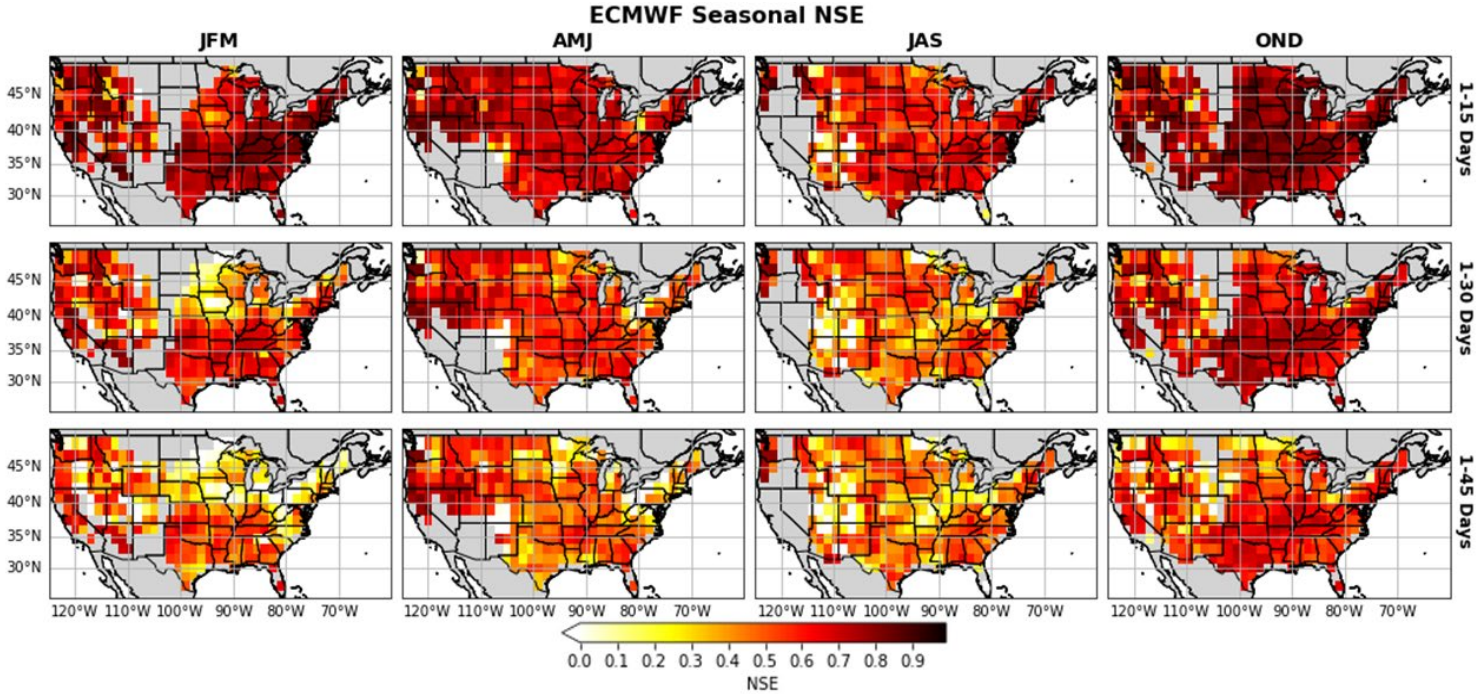


Figure 2. Nash Sutcliffe Efficiency (NSE) of ECMWF hindcast for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

b) Decomposition Plots

We decompose the NSE of ECMWF in Figure 2 into correlation (Figures 3), conditional bias (Figure 5) and unconditional bias (Figure 6) for each lead time for the four seasons.

i) Correlation and its longitudinal distribution

The first component of decomposition, Pearson's correlation coefficient, shows the innate model skill and the lower bound for explained variance in the model. The analysis of correlation shows that the skill decreases as lead time increases for all seasons (Figure 3.). Similar to the NSE, the correlation is also lower in the summer seasons and higher in the winter seasons. The correlation between S2S precipitation hindcasts and observed precipitation was averaged by longitude, for each season and lead time, after the dry mask threshold was applied. This longitudinal distribution more clearly illustrates the West to East coast gradient, where the correlation is higher in the West Coast and decreases towards the East Coast (Figure 3-4).

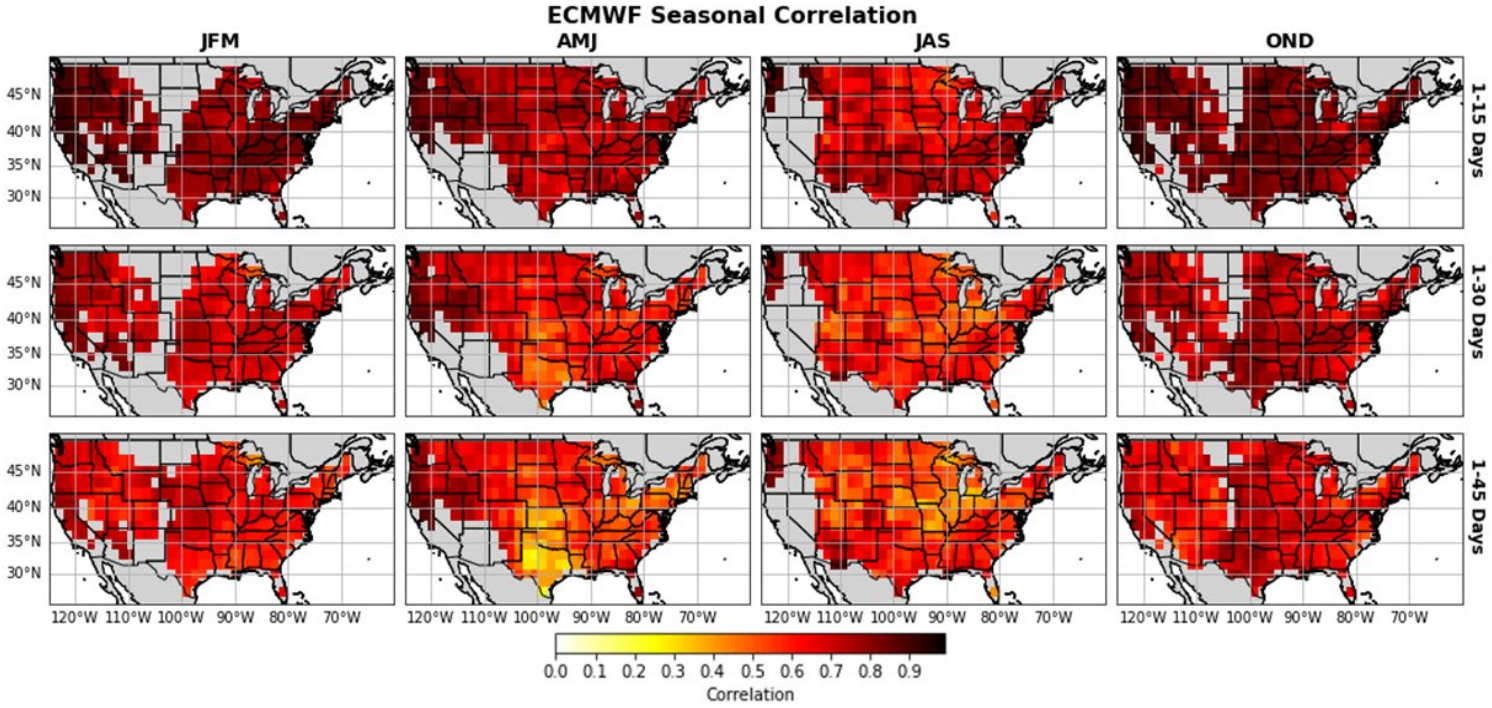
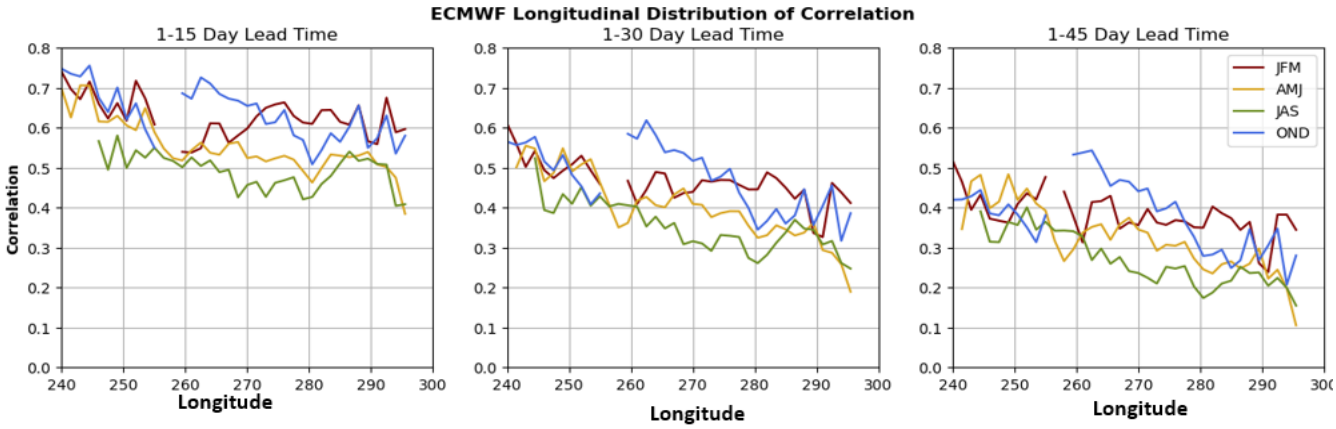


Figure 3. Correlation, the first component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS, and OND, and for three different lead times: 1-15 days, 1-30 days, and 1-45 days.

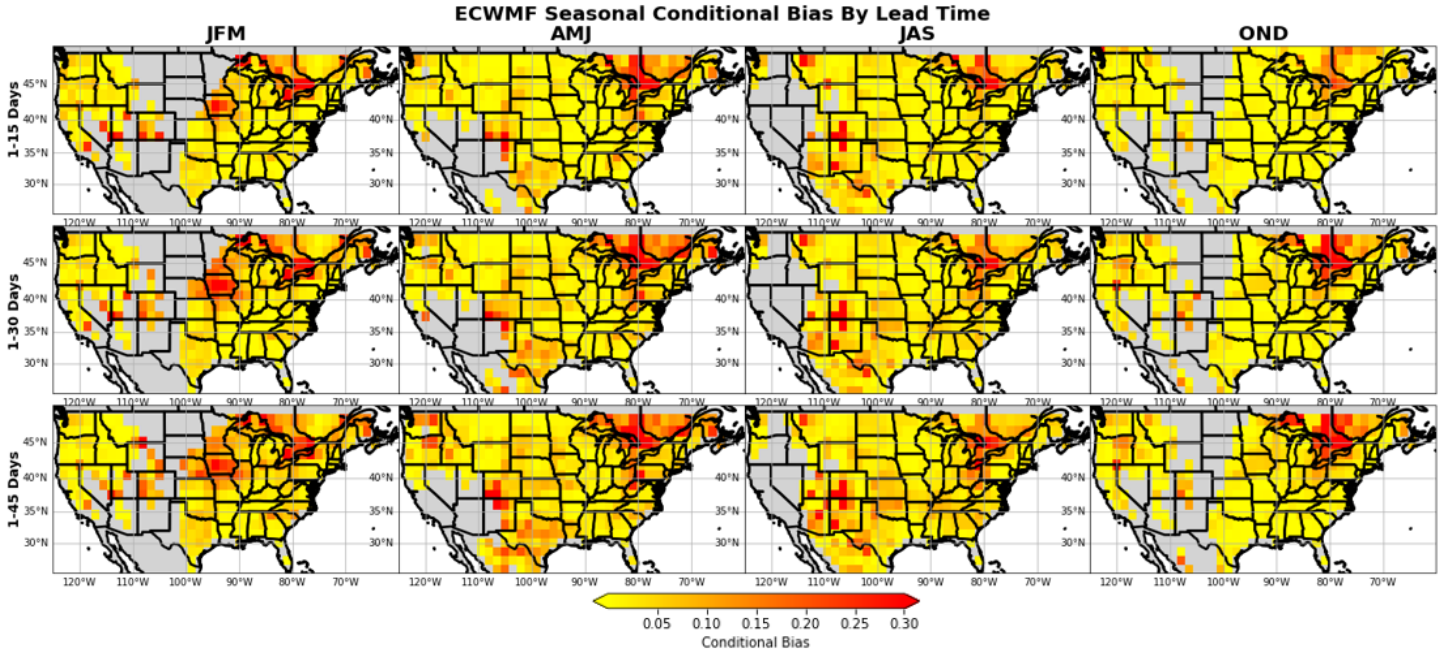
On the West Coast, correlation coefficients are higher than on the East Coast, which is partially due to the pronounced seasonality in precipitation over the West Coast that results in reduced number of grid cells being considered for evaluation after applying the dry mask. Additionally, correlation coefficients are higher towards the coasts and weaker further inland due to potential influence of local SSTs (Sankarasubramanian et al., 2017). Correlation coefficients are also lower towards the area surrounding the masked out regions.



313 *Figure 4. Longitudinal distribution of correlation by the average by latitude of the ECMWF hindcast data for*
314 *CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and*
315 *for three lead times: 1-15 days, 1-30 days, and 1-45 days*

316 *ii.) Conditional Bias*

317 The second and third components, conditional bias, and unconditional bias, are expected
318 to be zero for ideal forecasts. The conditional bias for the ECMWF decomposition increases as
319 lead time increases and tends to be higher towards the coasts. Further, the conditional bias is
320 higher during the summer season in comparison to the winter season (Figure 5). The Great
321 Lakes Region and the central part of the US has a high conditional bias that increases with
322 increasing lead times, whereas the Sunbelt has a low conditional bias during the winter and
323 spring seasons. Conditional bias is also higher towards the areas that were masked out from the
324 dry mask. Conditional bias is highest during JAS, specifically in the desert areas that were
325 masked out during the other seasons and is lowest during OND.



326 *Figure 5. The second component, conditional bias, of NSE decomposition, from the ECMWF hindcast data for*
327 *CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and*
328 *for three lead times: 1-15 days, 1-30 days, and 1-45 days.*

iii.) Unconditional Bias

The third component, unconditional bias, represents the systematic bias in reproducing the long-term mean of the observed precipitation. Unconditional bias is high in the Great Lakes Region and in the central part of the US (Figure 6). Additionally, unconditional bias is high in the desert regions for JAS, which was masked during the other seasons, for JAS. Conditional bias and unconditional bias are generally correlated and have higher values in the same regions.

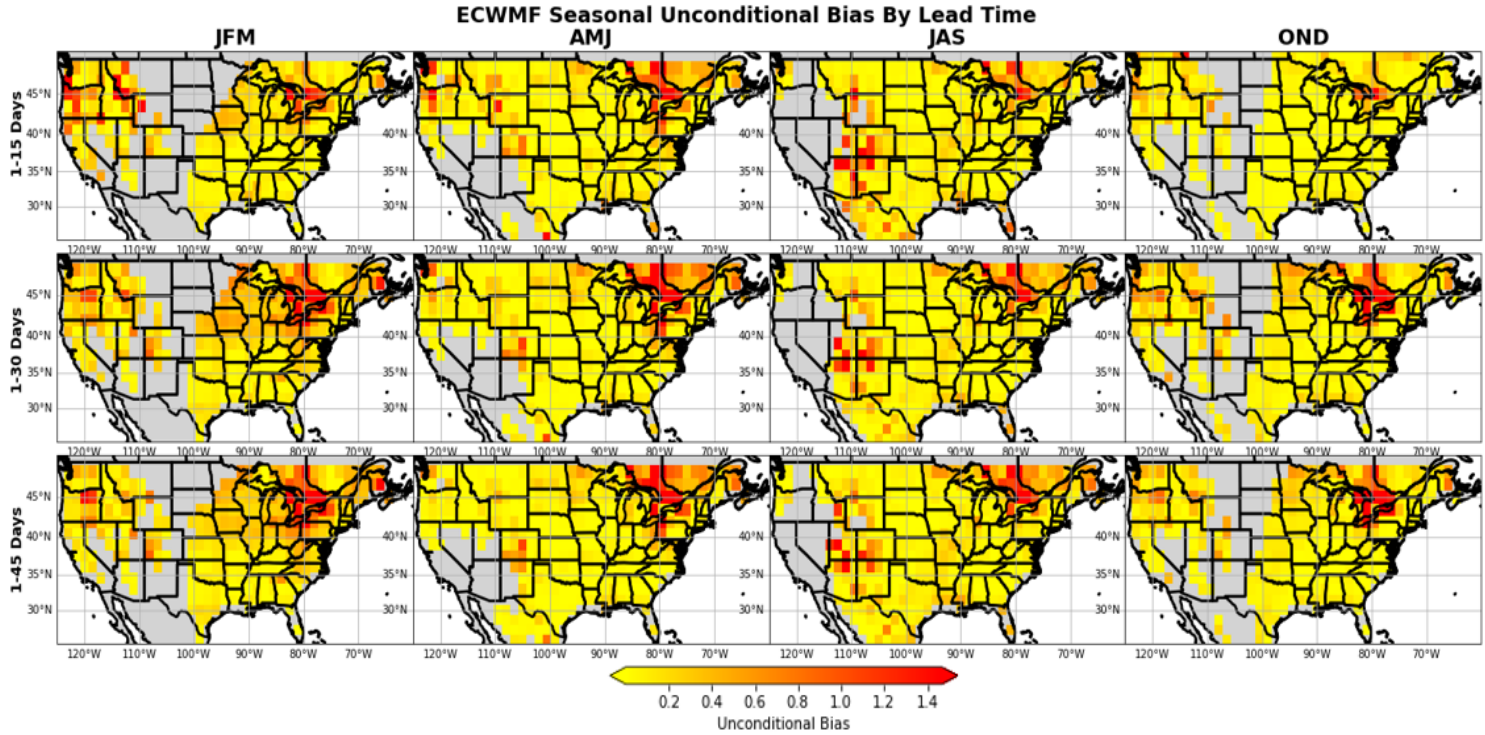


Figure 6. Unconditional bias, the third component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

c. Skill comparison across Koppen Climate Regimes

The skill of ECMWF S2S hindcast model was compared under three Koppen climate regimes: a) desert b.) temperate and c.) continental (Figure SI-1). For all lead times and climate regimes, the correlation varies by season and is lower in the summer months and is the highest in the winter months (Figure 7). Since the dry mask threshold was applied before the climate regime classification was considered, the correlation does not vary much between regimes within a given season. Conversely, if a dry mask had not been applied, the desert regimes may expect to have much better skill, because of inflated skill due to no-precipitation days.

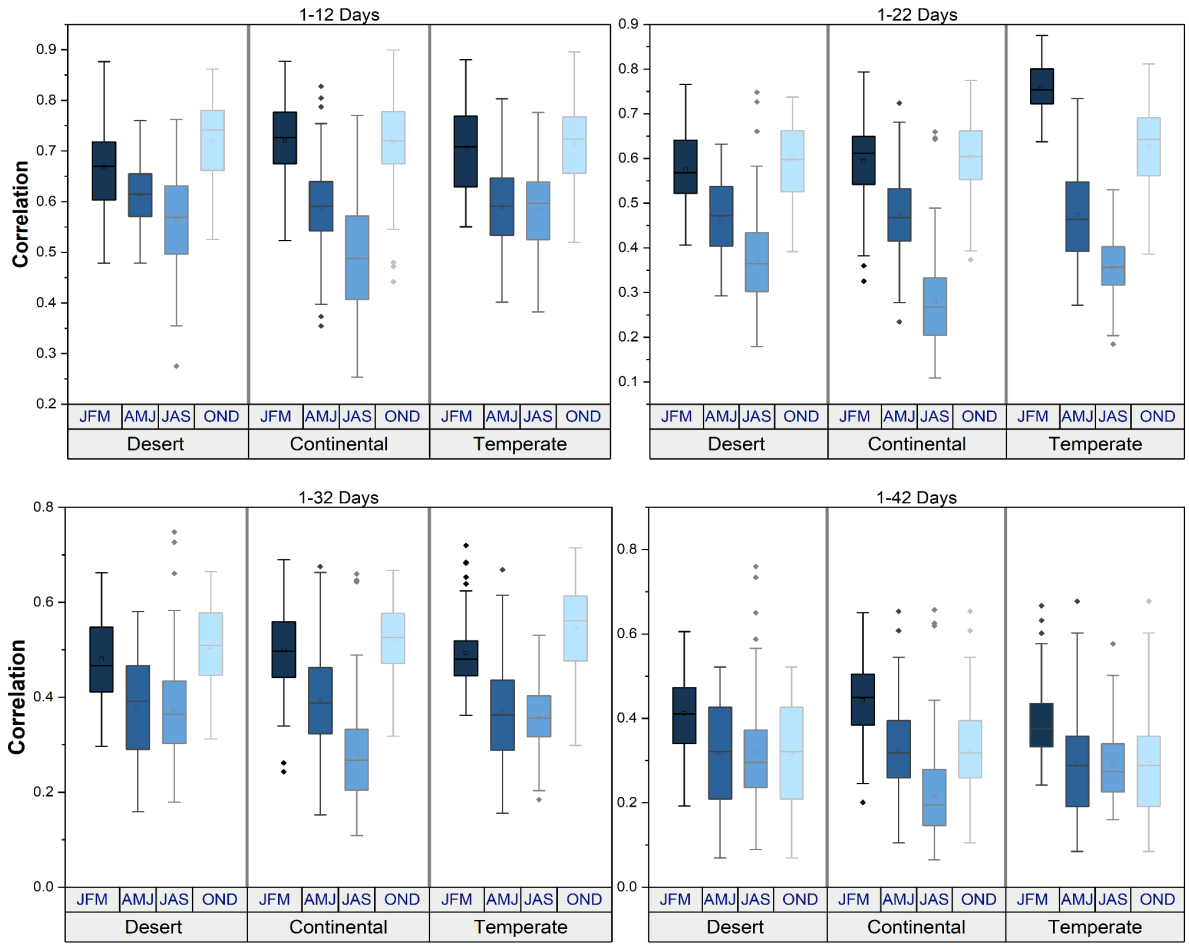


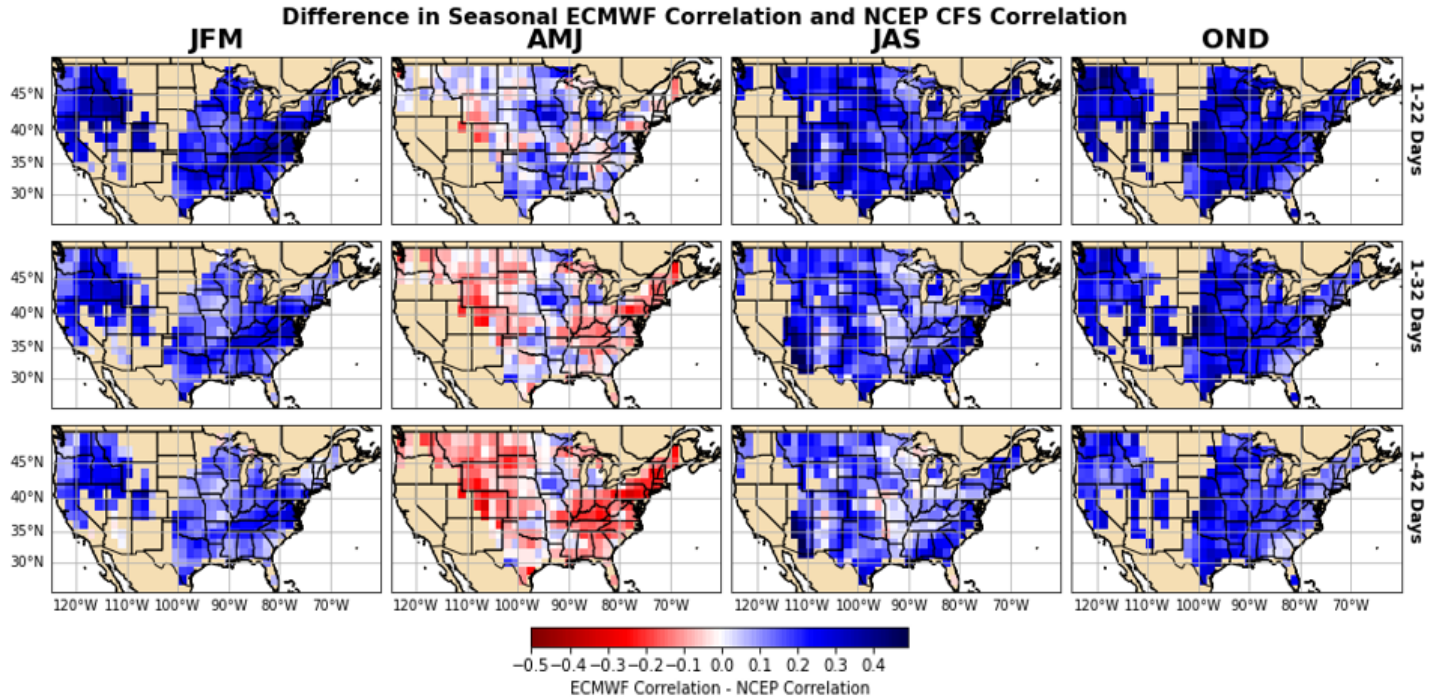
Figure 7. The box and whisker plot of correlation from the ECMWF hindcast model for three Koppen climate regimes: desert (red), temperate (blue) and continental (green) for lead times 1-12, 1-22, 1-32, and 1-42 days for all four seasons that the forecasts were initialized: JFM, AMJ, JAS, OND.

d. Model Comparison of NSE and Correlation

Comparing S2S hindcast models is important to understand the relative performance of the individual models. In this analysis, ECMWF's NSE was compared to NCEP CFS's NSE and next ECMWF's correlation was compared to all three models. The dry mask threshold may affect the model performance; therefore, forecast skill was not considered in areas where the historically observed precipitation did not exceed this threshold.

The blue regions in Figure SI-2 show where ECMWF's NSE outperforms the NSE of NCEP CFS for most lead times, regimes, and seasons, especially at shorter lead times, except for a few inland areas. Although ECMWF's NSE is higher than NCEP's in most regimes, seasons, and lead times, the ECMWF and NCEP CFS's correlation is closer in value (Figure 8). NCEP CFS has a higher NSE and correlation than ECMWF during AMJ. In comparison to ECMWF, NCEP's correlation improves with longer lead times during AMJ and is also higher in areas

365 further inland. Conversely, ECMWF has better performance around the coast (Figure 8) except
366 for OND, which may be due to the two different ocean models used in the initializations.



367 *Figure 8. Difference in Correlation values between ECMWF S2S hindcast and NCEP CFS for CONUS after dry*
368 *mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead*
369 *times: 1-12 days, 1-22 days, and 1-42 days.*
370
371 ECMWF and ECCC models' correlation differ by season but Figure 9 does not show a
372 clear inland-coastal differential in skill (Figure 9), which could be potentially due to ECMWF
373 and ECCC having the same ocean models. ECCC has a higher correlation than ECMWF during
374 the forecasts initiated in the summer months (JAS). However, since ECCC's lead time ranges
375 from 1-32 days, 1-42 day lead time between ECMWF and ECCC could not be compared.
376
377 Across seasons and lead times, NCEP CFS's correlation is higher than ECCC's
378 correlation for NCEP (Figure SI-3). NCEP CFS' model performance improves noticeably at
379 longer lead times and was not compared to 1-42 days lead time because of ECCC's shorter lead
380 time forecast availability. However, when comparing the first component, correlation, by
381 regime, season, and lead time, ECCC has higher correlation in AMJ, when compared to both
382 NCEP CFS as well as ECMWF. However, ECCC's performance tends to be worse in the
383 remaining three seasons.
384
385

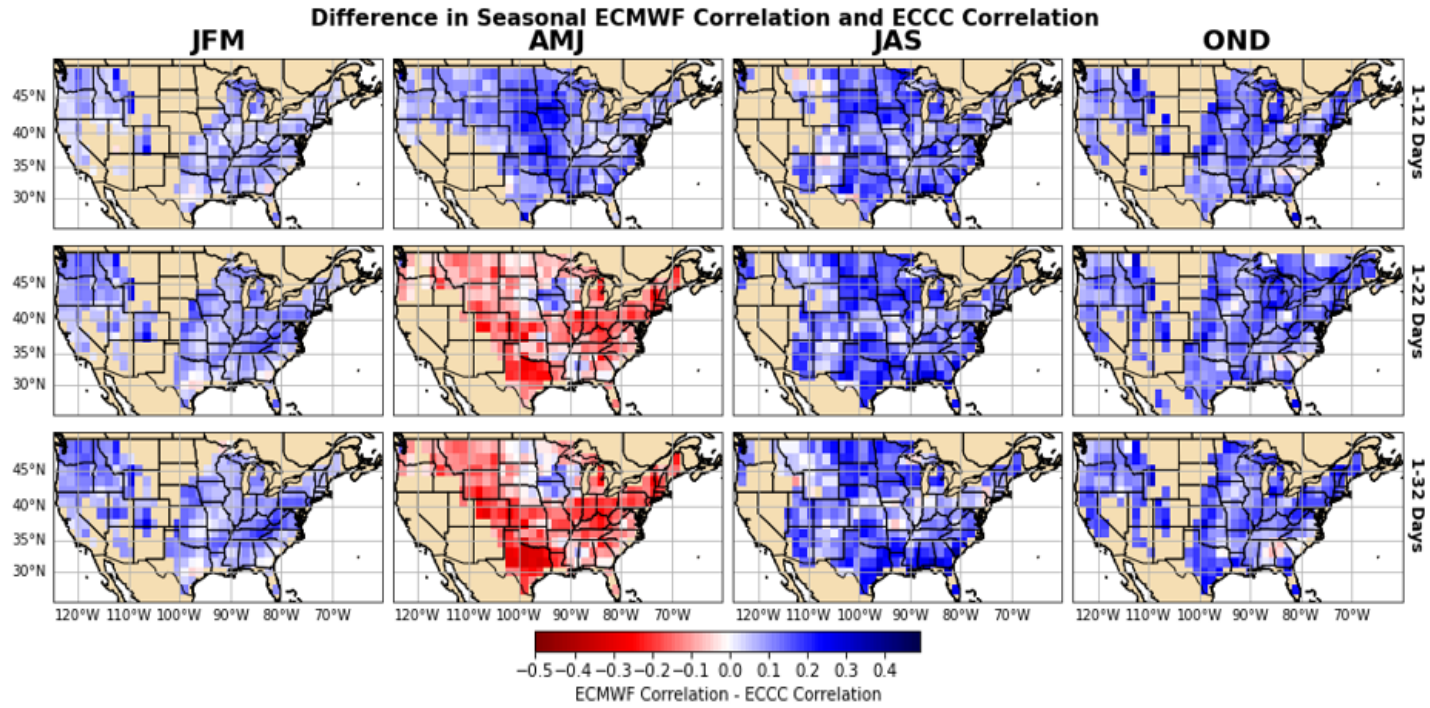


Figure 9. Difference in Correlation values between ECMWF S2S hindcast and ECCC for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-12 days, 1-22 days, and 1-32 days.

Overall, ECMWF's correlation for the forecast issued in seasons, JFM and OND, is higher than the other two models, but ECMWF's correlation is lower than the other models for the forecasts issued in AMJ (Figure 8-9). ECMWF has the highest NSE and correlation when solely considering the skill within the CONUS boundaries; however, NCEP CFS and ECCC hindcast models have much better forecast skill in the Great Lakes regime on and the Canadian regime just north of the Great Lakes, which although may not fall within the US boundaries, is still critical for the Midwest's water resources. ECMWF performs better towards the coasts and the skill may be higher in the winter seasons due to the areas that were masked out by the dry mask threshold. NCEP CFS and ECCC perform better in areas further inland, which is why the skill may be noticeably better in the spring and summer months (AMJ and JAS) where the inland regimes are not masked by the dry mask threshold since the regime receives higher precipitation during the summer. The differences in model skill could be due to the different ocean models and different initialization schemes, however this attribution has to be systematically analyzed further.

e. Model Skill comparison across Koppen Climate Regimes

The performance metrics for the three hindcast models were analyzed across the three Koppen climate regimes over the CONUS. Each model's NSE and the decomposed components were divided into climate regimes by season and lead times. At longer lead times, the differences in NSE reduces across seasons and climate regimes with NCEP CFS beginning to outperform

ECCC (Figure SI-4). ECMWF's NSE was higher than the NSE of ECCC and NCEP CFS across climate regime, season, and lead times (Figure SI-4), because NCEP and ECCC had high unconditional and conditional biases (Figure SI-4). Since these biases can be reduced to zero with simple post-processing techniques such as Model Output Statistics (Appendix A), we focused on comparing correlation (Figure 10).

The Pearson correlation coefficient is generally higher for ECMWF in comparison to ECCC and NCEP CFS models for all lead times, regimes, and seasons (Figure 10). There does not seem to be a consistent trend on how models perform for each climate regime across seasons and lead times even though both NCEP and ECCC perform better with forecasts issued in AMJ (Figure SI-4). For ECMWF and ECCC, the correlation is higher at shorter lead times, but NCEP's correlation remains relatively consistent across lead times (Figure 10). Across all models, lead times, and regimes the seasonal patterns illustrate that correlation is the highest during JFM and OND and lowest during AMJ and JAS.

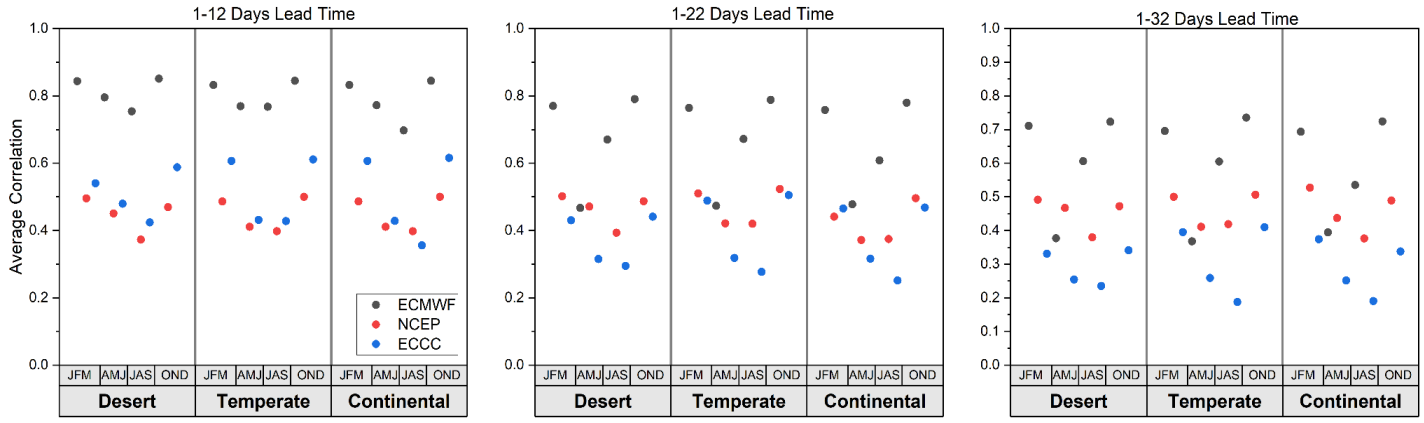


Figure 10. The average correlation for each regime: Regime B (desert), Regime C (temperate), and Regime D (continental) for each model: ECMWF (black), ECCC (blue), and NCEP CFS (red). The average correlation was calculated by lead time a) 1-12 days b.) 1-22 days and c.) 1-32 days for seasons JFM, AMJ, JAS, and OND.

The conditional bias is the lowest for ECMWF and highest for NCEP CFS particularly for AMJ and at shorter lead times (Figure SI-4). NCEP's median marginal bias was lower than ECMWF and ECCC, but one grid cell on the West Coast had a very high conditional bias causing the mean bias of all of the grid cells to be higher than the other two models.. ECCC has the highest conditional bias at the shorter lead times and ECMWF and NCEP CFS were comparable at 1-12 days for JFM, JAS, and OND. Conditional bias has the highest spread during spring months (AMJ). With longer lead times (e.g., 1-32 days), the unconditional bias across the selected models is similar, with ECCC being slightly higher than the other two models. No clear regional pattern of unconditional bias across all models and seasons was evident (Figure SI-4 g-i). The seasonality of unconditional bias seems to change based on lead times. We discuss in the next section how the conditional bias and unconditional bias could be potentially improved using post-processing techniques that focus on developing statistical relationships between model forecasts and the observed precipitation.

4.0 Discussion

Understanding the S2S precipitation forecasts skill across the CONUS over different seasons, as well as highlighting potential avenues for model improvement is critical for better forecast application. This study a) investigated and compared the spatial distribution of NSE for three S2S precipitation hindcast models across the CONUS, b) decomposed Nash-Sutcliffe Efficiency into correlation, conditional bias and unconditional bias based on the lead time and forecast issued in a season for each model and c) analyzed model skill across three (tropical, desert and temperate) Koppen Climate regimes. Our analysis shows that NSE of ECMWF was higher closer to the coast, most likely due to the influence of MJO and ENSO, and was also higher for the forecast issued during winter months and with shorter lead times. Decomposition of NSE shows that the first component, correlation, illustrates there is a gradient in skill from west coast (higher) to east coast (lower). Both the conditional and unconditional biases were also smaller during the winter months and in areas closer to the coast. The model comparison showed that ECMWF performs well in the winter seasons and towards the coasts, whereas NCEP CFS's performance is the best for forecasts issued during AMJ and in inland areas. The conditional and unconditional bias were high over the Midwest Great Lakes region. The conditional bias was higher for NCEP CFS, particularly for forecasts issued in AMJ and the unconditional bias was high for forecasts issued in JAS. ECCC's skill is high during AMJ and at short lead times, but decreases significantly with longer lead times. No clear trends were observed across the climate regimes across the three hindcast models' performances, but NSE and correlation was higher for the winter seasons than the summer seasons consistently for all the lead times, regimes and three models.

Potential for improving S2S forecasts

Even though our analysis, after application of dry mask, showed that conditional bias and unconditional bias are the primary reasons for low and negative NSE values for the S2S hindcasts, this could be overcome by selecting a proper post-processing scheme where the correlation is high across the CONUS. One of the commonly used post-processing scheme for correcting weather/climate forecasts is Model Output Statistics (MOS), which is a linear regression model that uses the forecast or a transformation of it (e.g., principal components) as a predictor and the observed precipitation as a predictand (Antolik et al., 2000; Sankarasubramanian et al., 2008). One advantage with a linear regression model is that it reduces the marginal bias to zero (Appendix A). Further, we also show analytically in Appendix A, a linear regression model reduces the conditional bias to zero which turns the NSE of the corrected forecasts from a MOS being equal to the square of the correlation coefficient (i.e., component A). Thus, a linear regression based MOS provides a lower bound on the NSE of the forecast to be decomposed component A, thereby providing a guidance on where post-processing schemes will be useful for a given location/regime. An example of where post-processing can be effective for correcting bias is NCEP CFS's 1-42 day forecasts. ECMWF did not have any grid cells where NSE was below zero, because the conditional and unconditional bias were low, so we

show NCEP, which has large sources of unconditional and conditional bias across all regimes, but relatively high correlation (SI-4).

Figure 11 shows locations where a) NCEP's NSE is less than zero and correlation is significant ($p<0.05$), b) NCEP's NSE is greater than zero and correlation is significant ($p<0.05$), and c) NCEP's NSE is less than zero, but correlation is not significant ($p>0.05$) for 1-42 day lead times. For the first case, where NSE is low and correlation is high, post-processing such as MOS can be effectively used to reduce conditional and unconditional biases to improve forecast skill, and a large portion of CONUS, mostly inland area and particularly for forecasts issued in seasons JFM and AMJ (Figure 11). For the second category, a large portion of the coastal region, particularly in forecast-initialized seasons AMJ and OND, have significant ($p<0.05$) correlation and high NSE, which means post-processing will not be effective as the model does not capture the observed variability. Similarly, post-processing will not be effective in areas with low NSE and correlation that is not significant ($p>0.05$), which includes a few grid points in AMJ and JAS (Figure 11). Even though linear-regression based MOS may not result in improved skill in areas where both NSE and correlation are low, other MOS post-processing schemes can be considered such as a semi-parametric model or machine learning models (Glahn et al., 1972; Taillardat et al., 2019), NSE of S2S forecasts could be potentially improved as such models are more flexible in reducing the mean square error in the forecast.

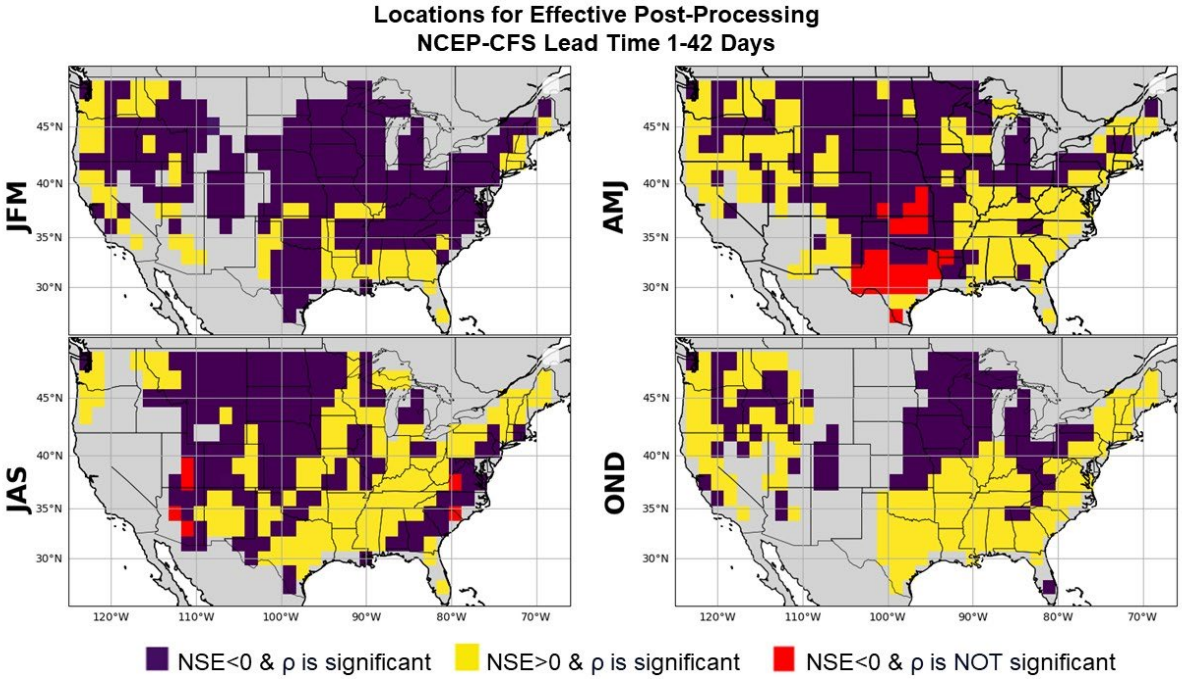


Figure 11. Post-processing will be effective in the locations where NSE<0 and correlation is significant (purple), but will not be necessary in places where NSE<0 but correlation is not significant (red) or in places where correlation is significant (yellow).

Even though the selected models had ensemble forecast, we considered only ensemble mean for forecast decomposition. We did not consider probabilistic forecasts such as Brier Skill score for skill evaluation and decomposition since the differences in ensemble members could

significantly affect the forecast evaluation. Similar decomposition on Brier score could reveal the forecast reliability and resolution of each model's performance in below-normal and above-normal conditions (Brier, 1950). Further, our analysis focused on decomposition without evaluating the model's performance during extreme conditions, which could be pursued further to understand the sources of bias. Our analysis also did not consider NMME because the number of ensemble members varies between models, giving more weight to some models. Additionally, the models within NMME have varying forecast issued frequencies, lead times, and issued dates. These varying model features within the multi-model need to be addressed before valid model comparisons can occur. Since the intent of this study was to show a systematic process of evaluating model skill and comparing across the models, we did not consider NMME for our study.

5.0 Conclusions

S2S precipitation forecasts are critical for operational and proactive water resource management and planning. Systematic S2S forecast skill assessment is essential for understanding existing model skill and how different errors contribute to it. Our evaluation of three S2S reforecasts – ECMWF, ECCC and NCEP – based on NSE decomposition primarily looked at the skill of forecasts issued during four seasons and under three different lead times. Our analysis shows the importance of applying dry mask as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. The full decomposition of ECMWF revealed a West to East coast longitudinal gradient in NSE and correlation. Decomposed components, conditional and unconditional bias, did not show any longitudinal trends. ECMWF's skill showed that seasonal trends in forecast skill occurred across all lead times and all seasons, but correlation did not differ by climate regimes.

The forecast skill and associated errors were also compared across models. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during the winter months, (JFM and OND). ECMWF had the highest NSE across the three climate regimes – temperate, desert and continental – considered. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in areas further away from the coast. Our decomposition efforts show S2S improvements in physical modeling efforts such as parameterization and initialization should be undertaken for ECMWF particularly for areas further from the coast, for forecasts issued in the spring months, AMJ, and for NCEP CFS and ECCC for the forecasts issued in the winter months over coastal areas.

Our analytical derivation on how MOS could help improve the forecast shows that a linear regression based MOS could ensure the NSE of the post-processed forecast to be component A, which is the square of the correlation coefficient between forecasts and the observation. This shows because simple linear regression based MOS can eliminate conditional

and marginal biases. This also provides information on regions (Figure 11, NSE <0 and ρ not significant) where S2S forecasting schemes can focus on improved model parameterizations and initializations including coupling with land surface models for improving the skill (Entekhabi et al., 1999).

Acknowledgments

The first author was supported by the National Science Foundation Fellowship (NSF) for the Graduate Research Fellowship Program (GRFP) support (award # DGE-2137100). Apart from that, this research was also supported by two NSF grants (award # CBET - 1805293 and IIE-2033607).

Open Research

The hindcast model data was accessed on the ECMWF S2S reforecast portal (https://apps.ecmwf.int/datasets/data/s2s/). The CPC Unified Gauged-Based observed precipitation dataset are available at https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html, and the Koppen climate classification data are available at www.gloh2o.org/koppen/.

Appendix A. Decomposition of NSE for Linear-Regression Based Model Output Statistics

For each grid cell, o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation value and y_{it} is the corrected precipitation value, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. Linear regression model 2 is used for the model to get the corrected precipitation value, which is the MOS estimate.

$$o_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \quad [1]$$

$$y_{it} = \beta_0 + \beta_1 x_{it} \quad [2]$$

For a given i, NSE is originally between o_{it} and x_{it} (equation 3), but a linear regression is used to estimate the corrected precipitation, y_{it}. For a given i , the NSE is calculated between o_{it} and y_{it} (equation 4) and then decomposed into parts A (equation 8-14), B (equation (15), and C (equation 16-17).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = \rho_{xo}^{\,2} - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad [3]$$

$$NSE_i(o_{it}, y_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-y_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = = \rho_{yo}^{\,2} - (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 - (\frac{\overline{y}-\overline{o}}{\sigma_o}))^2 \quad [4]$$

$$\beta_1 = \frac{cov(o-x)}{\sigma_x^{\,2}} \quad [5] \qquad \beta_0 = \overline{o} - \beta_1 * \overline{x} \quad [6] \qquad \beta_1 = \frac{\rho_{xo} * \sigma_x * \sigma_o}{\sigma_x^{\,2}} = \frac{\rho_{x\,o} * \sigma_o}{\sigma_x} \quad [7]$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 3. The pearson correlation coefficient between x and o is ρ_{xo}. For the corrected precipitation, y_{it}, the standard deviation and mean are σ_y and $\overline{y}$ respectively, when y_{it} is summed over time from 1 to n for lead time i in equation 4. The correlation coefficient between o and y is ρ_{yo} .

NSE of o_{it} and y_{it} is decomposed into the three corresponding parts a.) correlation, b.) conditional bias and c.) unconditional bias. It is important to note that correlation, Component A (o, y), will be the same as the Component A (o, x). Where

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [8]$$

$$\rho_{xo} = \frac{cov(x, o)}{\sigma_x * \sigma_o} \quad [9]$$

$$cov(y, o) = cov(\beta_0 + \beta_1 x, o) = \beta_1 \, cov(x, o) \quad [10]$$

$$var(y) = \beta_1^{\,2} * \sigma_o \quad [11] \qquad \sigma_y = \beta_1^{\,2} * \sigma_x \quad [12]$$

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [13]$$

$$\rho_{yo} = \frac{\beta_1 * cov(y, o)}{\beta_1 * \sigma_x * \sigma_o} = \frac{\beta_1 * cov(y, o)}{\sigma_y * \sigma_o} = \rho_{xo} \quad [14]$$

Conditional bias B (o, y) will be reduced to zero MOS estimates.

$$B(o, y) = (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 = (\rho_{xo} - \frac{\beta_1 * \sigma_x}{\sigma_o})^2 = (\rho_{xo} - \frac{\sigma_y}{\sigma_o})^2 \quad [15]$$

638 $$\rho_{xo} = (-(\frac{\rho_{xo} - \sigma_o}{x}) \cdot \frac{\sigma_x}{\sigma_o})^2 = 0$$

639 Unconditional bias C (o , y) will also reduce to zero for MOS estimates.

640 $$C(o, y) = (\frac{\overline{y} - \overline{o}}{\sigma_o}))^2 \qquad [16]$$

641 $$\overline{y} = \beta_0 + \beta_1 * \overline{x} = \overline{o} - \beta_1 * \overline{x} * + \beta_1 * \overline{x} \qquad [17]$$

642 $$C(o, y) \to 0$$

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

References

Antolik, M.S. (2000). An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *Journal of Hydrology*, 239, 306- 337.

Barbero, R., Fowler, H. J., Blenkinsop, S., Westra, S., Moron, V., Lewis, E., et al. (2019). A synthesis of hourly and daily precipitation extremes in different climatic regions. *Weather and Climate Extremes*, 26, 100219. https://doi.org/10.1016/j.wace.2019.100219.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, *Hydrol. Earth Syst. Sci*., 21, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017.

‘

Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. *Geophysical Research Letters,* 47, e2020GL087408. https://doi.org/10.1029/2020GL087408

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

Carter, G. M., Dallavalle, J.P., & Glahn, H.R. (1989). Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401– 412.

685 Chalise, D. R., Sankarasubramanian, A., Olden, J. D., & Ruhi, A. (2023). Spectral signatures of
686 flow regime alteration by dams across the United States. *Earth's Future,* 11, e2022EF003078.
687 https://doi.org/10.1029/2022EF003078
688
689 Charba, J. P., and F. G. Samplatsky, 2011: High-Resolution GFS-Based MOS Quantitative
690 Precipitation Forecasts on a 4-km Grid. *Mon. Wea. Rev.,* 139, 39–68,
691 https://doi.org/10.1175/2010MWR3224.1.
692
693 Chen, M., W. Shi, P. Xie, V. B. S. Silva, V E. Kousky, R. Wayne Higgins, & J. E. Janowiak
694 (2008), Assessing objective techniques for gauge-based analyses of global daily precipitation, *J.*
695 *Geophys. Res*., 113, D04110, doi:10.1029/2007JD009132.
696
697 Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al.
698 (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources*
699 *Research*, 57, https://doi.org/10.1029/2020WR029001.
700
701 de Andrade, F.M., Coelho, C.A.S. & Cavalcanti, I.F.A. (2019). Global precipitation hindcast
702 quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Clim Dyn,*
703 52, 5451–5475. https://doi.org/10.1007/s00382-018-4457-z.
704
705 Entekhabi, D., and Coauthors (1999). An agenda for land-surface hydrology research and a call
706 for the second International Hydrological Decade. *Bull. Amer. Meteor*. Soc., 80, 2043–2058,
707 https://doi.org/10.1175/1520-0477(1999)080<2043:AAFLSH>2.0.CO;2.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor*., 11, 1203–1211.

Glahn, H. R., & Ruth D. P. (2003). The New Digital Forecast Database of the National Weather Service. *Bull. Amer. Meteor. Soc*., 84 , 195–201.

Goddard, L., Kumar, A., Solomon, A. et al. (2013). A verification framework for interannual-to-decadal predictions experiments. Clim Dyn, 40, 245–272. https://doi.org/10.1007/s00382-012-1481-2.

Gupta, H. V., Kling, H., Yilmaz, K. K., & G. F. Martinez (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.,* 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

Hamill, T.M. & Colucci, S.J., (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6), pp.1312-1327.

Ham, YG., Kim, JH. & Luo, JJ. (2019). Deep learning for multi-year ENSO forecasts. *Nature* 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7

729 Konapala, G., Mishra, A.K., Wada, Y. et al. (2020). Climate change will affect global water
730 availability through compounding changes in seasonal precipitation and evaporation. *Nat*
731 *Commun*, 11, 3044. https://doi.org/10.1038/s41467-020-16757-w
732
733 Krakauer, N.Y. (2019). Temperature trends and prediction skill in NMME seasonal forecasts.
734 Clim Dyn, 53, 7201–7213. https://doi.org/10.1007/s00382-017-3657-2
735 Li, Y., Tian, D., & Medina, H. (2021). Multimodel Subseasonal Precipitation Forecasts over the
736 Contiguous United States: Skill Assessment and Statistical Postprocessing. J. Hydrometeor., 22,
737 2581–2600, https://doi.org/10.1175/JHM-D-21-0029.1.
738
739 Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier,
740 D.P., Stouffer, R.J. (2008). Stationarity is dead: Whither water management?. *Science,* 319
741 (5863): 573-574. https://doi.org/10.1126/science.1151915
742
743 Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the
744 correlation coefficient. *Monthly Weather Review*, 116, 2417– 2424.
745
746 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I—
747 A discussion of principles. *Journal of Hydrology,* 10(3), 282–290. https://doi.org/10.1016/0022-
748 1694(70)90255-6
749
750 Petersen, T., Devineni, N., & A. Sankarasubramanian, A. (2012). Seasonality of monthly runoff
751 over the continental United States: Causality and relations to mean annual and mean monthly

distributions of moisture and energy, *J. Hydrol*., 468–469, 139–150, 10.1016/j.jhydrol.2012.08.028.

Quan, X., Hoerling, M., Whitaker, J., Bates, G., & T. Xu, T. (2006). Diagnosing Sources of U.S. Seasonal Forecast Skill. J. Climate, 19, 3279–3293, https://doi.org/10.1175/JCLI3789.1.

Sankarasubramanian, A., Sabo, J.L., Larson, K.L., Seo, S.B., Sinha, T., Bhowmik, R., Vidal, A.R., Kunkel, K., Mahinthakumar, G., Berglund, E.Z. & Kominoski, J. (2017), Synthesis of public water supply use in the United States: Spatio-temporal patterns and socio-economic controls. *Earth's Future*, 5: 771-788. https://doi.org/10.1002/2016EF000511

Sankarasubramanian, A., Lall, U., & Espinueva, S. (2008). Role of Retrospective Forecasts of GCMs Forced with Persisted SST Anomalies in Operational Streamflow Forecasts Development. *J. Hydrometeor*., 9, 212–227, https://doi.org/10.1175/2007JHM842.1.

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T., (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, 28(4), pp.616-640.

Scheuerer, M., & Hamill, T.M., (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Rev*.w, 143(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

775 Sun, L., Hoerling, M.P., Richter, J.H., Hoell, A., Kumar, A. & Hurrell, J.W., (2022). Attribution
776 of North American Subseasonal Precipitation Prediction Skill. *Weather and Forecasting*, 37(11),
777 pp.2069-2085.
778
779 Taillardat, M., Fougères, A., Naveau, P., & Mestre, O. (2019). Forest-Based and Semiparametric
780 Methods for the Postprocessing of Rainfall Ensemble Forecasting. *Wea. Forecasting*, 34, 617–
781 634, https://doi.org/10.1175/WAF-D-18-0149.1.
782
783 Vitart, F., Robertson, A. & Anderson, D. (2012). Sub-seasonal to Seasonal Prediction Project:
784 bridging the gap between weather and climate. WMO Bull. 61, 23–28.
785
786 Vitart, F., Robertson, A.W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the
787 prediction of extreme events. *npj Clim Atmos Sci* 1, 3. https://doi.org/10.1038/s41612-018-0013-
788 0
789
790 Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti,
791 E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P.
792 Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami,
793 R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A.W. Robertson, P. Ruti, C. Sun, Y.
794 Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. Won, H. Xiao, R.
795 Zaripov, and L. Zhang, (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database.
796 gi. *Amer. Meteor. Soc*., 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.
797

798 Vitart, F., Robertsn, A. W., & S2S Steering Group (2015). Sub-seasonal to seasonal prediction:
799 Linking weather and climate. In: Seamless Prediction of the Earth System: From Minutes to
800 Months. (pp. 385– 401). WMO-No.1156 (Chapter 20). Retrieved from
801 http://library.wmo.int/pmb_ged/wmo_11
802
803 Wang, L., & Robertson, A.W. (2019). Week 3–4 predictability over the United States assessed
804 from two operational ensemble prediction systems. *Clim Dyn,* 52, 5861–5875.
805 https://doi.org/10.1007/s00382-018-4484-9.
806
807 Weglarczyk S (1998), The interdependence and applicability of some statistical quality measures
808 for hydrological models. *Journal of Hydrology,* 206: 98-103. https://doi.org/10.1016/S0022-
809 1694(98)00094-8.
810
811 Weigel, A.P., Liniger, M.A., & Appenzeller, C. (2008), Can multi-model combination really
812 enhance the prediction skill of probabilistic ensemble forecasts?. *Q.J.R. Meteorol. Soc.,* 134:
813 241-260. https://doi.org/10.1002/qj.210
814
815 White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F.,
816 Coughlan de Perez, E., Ray, A.J., Murray, V., & Bharwani, S., 2017. Potential applications of
817 subseasonal-to-seasonal (S2S) predictions. *Meteorological applications*, 24(3), pp.315-325.
818
819 Wilks, D. Statistical Methods in the Atmospheric Sciences (Academic, 2006).
820

821 Wood, A. W., Maurer, E.P., A. Kumar, & Lettenmaier, D.P. (2002). Long-range experimental
822 hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* 107, 4429, doi:
823 10.1029/2001JD000659.
824
825 Zhang, C. (2013). Madden–Julian Oscillation: Bridging weather and climate. *Bulletin of the*
826 *American Meteorological Society*, 94, 1849–1870. https://doi.org/10.1175/BAMS-D-12-00026.1
827
828 Zhang, L., Kim, T., Yang, T., Hong, Y. & Zhu, Q. (2021). Evaluation of Subseasonal-to-
829 Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II
830 (NMME-2) over the contiguous US. *Journal of Hydrology*, 603, p.127058.
831 https://doi.org/10.1016/j.jhydrol.2021.127058
832

Spatial and Temporal Variation of Subseasonal-to-Seasonal (S2S) Precipitation Reforecast Skill Across CONUS

J. R. Levey[1], A. Sankarasubramanian[1]

[1]Department of Civil, Construction and Environmental Engineering, North Carolina State University.

Corresponding author: Jessica Levey (jrlevey@ncsu.edu)

Key Points:

- NSE decomposition of S2S reforecast skill shows the spatio-temporal variations in correlation, conditional and unconditional bias.
- Longitudinal gradient of forecast skill exists from the West (higher) to East (lower).
- Regression based model-output statistics provide correlation as the lower bound of NSE as the marginal and conditional bias reduces to zero.

Abstract

Precipitation forecasts, particularly at subseasonal-to-seasonal (S2S) time scale, are essential for informed and proactive water resources management. Although S2S precipitation forecasts have been evaluated, no systematic decomposition of the skill, Nash-Sutcliffe Efficiency (NSE) coefficient, has been analyzed towards understanding the forecast accuracy. We decompose the NSE of S2S precipitation forecast into its three components – correlation, conditional bias, and unconditional bias – by four seasons, three lead times (1–12-day, 1-22 day, and 1-32 day), and three models (ECMWF, CFS, NCEP) over the Conterminous United States (CONUS). Application of dry mask is critical as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. Further, a west-to-east gradient in S2S forecast skill exists and forecast skill was better during the winter months and for areas closer to the coast. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during fall and winter months. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in in-land areas. Post-processing using simple Model Output Statistics could reduce both unconditional and conditional bias to zero, thereby offering better skill for regimes with high correlation. Our decomposition results also show efforts should focus on improving model parametrization and initialization schemes for climate regimes with low correlation values.

1.0 Introduction

Global climate change and regional anthropogenic disturbances, including urbanization and deforestation, are driving shifts in the hydrologic cycle, and impacting water resources (Konapala et al, 2020; Milly et al., 2008). Consequently, extreme precipitation events, including prolonged droughts or flooding, are expected to be more frequent, further threatening water supply and variability (Milly et al., 2008). In conjunction with hydroclimatic changes, population changes also stress surface and groundwater resource withdrawals in many regions across the Conterminous US (CONUS) (Sankarasubramanian et al., 2017). Reservoir releases, during both floods and droughts, are modified for human needs, downstream ecological health, and for ensuring watershed resilience (Chalise et al., 2021). Mismanagement of water resources, both surface water and groundwater, may pose threats to agriculture, supply chains, human and environmental health, and regional economies. Hence, reliable and accurate subseasonal-to-seasonal (S2S) precipitation forecasts are essential in an age of a changing climate for improving water management strategies and preparing for near-future hydroclimatic extremes.

Compared to the skill of short-range weather forecasts (less than 15 days) and long-range seasonal forecasts, which are reasonably good, the skill of S2S forecasts, ranging between 15 to 60 days, is low and is often referred to as the 'predictability desert' (Vitart et al., 2012). Understanding the current S2S precipitation forecasts skill, as well as highlighting the potential avenues – initialization, parametrization, and post-processing schemes – for improvement are critical for accurate S2S precipitation forecasts for operational use (White et al., 2017). Known contributing factors that influence S2S model forecasting performance include the parametrization and initialization schemes, large-scale atmospheric circulation modes, and coupled models (Vitart et al., 2018). The model initialization scheme, including land surface and soil moisture representation, are also crucial for accurately representation of geophysical fluxes.

Climate oscillations, such as El Nino Southern Oscillations (ENSO) and Madden-Julian Oscillations (MJO) also influence seasonal forecast prediction skill (Zhang, 2013). ENSO's influence on United States' winter hydroclimatology is well-known, particularly over the Southeast and west coast, accounting for roughly a third of US winter forecasting skill (Quan et al., 2006).

Previous studies have attributed S2S skill between ENSO and MJO (Sun et al., 2022; Wang et al., 2019) and have compared S2S skill across models, lead times and seasons (Zhang et al 2021; de Andrade et al, 2019). However, these studies that examined S2S models' forecasting performance did not apply a threshold on dry mask prior to calculating the model's skill. Zhang et al (2021) have evaluated S2S forecast skill by filtering extreme precipitation events, but did not apply a dry mask threshold for evaluating the overall skill. Without a dry mask threshold, the S2S skill will be inflated, especially in regions with a pronounced dry season, as zero rainfall days is included in these skill calculations (Wilks, 2006). The ability to predict days without precipitation is important for drought prediction and planning, but the skill will be inflated for wetter and normal conditions; therefore, the dry mask application was used to filter out areas of inflated skill based on the climatological means. Several studies focused on extreme precipitation forecasts have applied percentile filters (Zhang et al., 2021), which reduces the sample size particularly while evaluating monthly/seasonal skill. Given the pronounced seasonality in precipitation over the CONUS (Petersen et al., 2012), we systematically evaluate the S2S forecasting skill across CONUS by applying a dry mask before considering the skill for each lead time, season and region. Evaluating the forecast skill after applying the dry mask could potentially affect the source of model skill, and the associated biases that could be obtained from decomposition.

S2S precipitation forecast skill has been compared considering both probabilistic and deterministic metrics to evaluate the forecast skill (Zhang et al., 2021; de Andrade et al., 2019). S2S models' skill have been evaluated using Mean Square Error (MSE), mean square skill score (MSSS), root mean square error (RMSE), anomaly correlation coefficient (ACC), Pearson's correlation coefficient, and ranked probability skill score (RPSS) (Zhang et al., 2021; de Andrade et al., 2019). de Andrade, et al., (2019) evaluated hindcast skill using linear correlation coefficient and analyzed the sources of bias and variability; however, this study was a large-scale global analysis of forecast skill and did not consider the seasonal skills and the associated errors. Decomposing the MSSS three components – correlation coefficient, condition bias and marginal bias – would provide information on the regions and seasons over which the selected models have the ability to capture the variability in observed precipitation but have significant biases in estimation. Further, the hindcast assessment of (de Andrade et al., 2019) was performed without the dry mask application, which may inflate forecast skill particularly for regions with pronounced dry season.

The Nash-Sutcliffe Efficiency (NSE), also known as the coefficient of determination, is a metric that measures the skill of hydrologic models (Nash & Sutcliffe, 1970). Li et al., (2022) used to evaluate S2S forecast skill performance based on Kling-Gupta Efficiency (KGE) metric, which provides a different decomposition of NSE, without applying the dry mask across the CONUS or considering seasonality. However, decomposing the Nash-Sutcliffe Efficiency (NSE) for precipitation hindcasts after applying the dry mask provides critical information without inflating the skill of the model. Furthermore, implementing new parametrizations and

initialization schemes could be costly and take additional time to develop reforecasts. One effective way to improve the forecasting skill is to consider post-processing schemes (Carter et al., 1998; Glahn et al., 2003). Further, post-processing could also be implemented over reforecasts from multiple models to develop multi-model ensembles which have been shown to improve the forecast skill compared to the best individual model (Weigel et al., 2008). Past work on statistical post-processing has considered both parametric and non-parametric approaches (Hamill et al., 1997; Schefzik et al., 2013; Scheuerer et al., 2015). Although many studies have used post-processing schemes on S2S precipitation forecasts, understanding the components of S2S forecast skill could provide additional insights on how post-processing schemes can be used and could also indicate potential regions where improvements in models will be needed to further improve the forecast skill.

Several S2S models that contribute multi-model ensembles have been run for reforecasts. Historically, some S2S multi-model datasets have only been running for a period of short time, limiting the ability to capture the interannual variability in precipitation. Other multi-model ensembles have primarily focused on generating monthly forecasts for seasonal prediction with infrequent model initialization. This study uses three individual models hindcasts from the World Weather Research Programme (WWRP) and World Climate Research Programme (WCRP) S2S prediction project (Vitart et al., 2012). The S2S project, originating in 2013, has a long record of forecasts and reforecasts that are initialized multiple times a week (Vitart et al., 2017). The longer range of data allows for larger sample sizes for robust estimation of NSE and decomposition metrics. Comparing model performance is important because forecast skill varies between S2S models as each model has different parameterization schemes, number of ensembles, and resolution. This study will consider decomposition of NSE of S2S reforecasts over the CONUS for three models – European Centre of Medium-Range Weather Forecast's (ECWMF) National Centre for Environmental Prediction Climate Forecast System (NCEP CFS) and Environment and Climate Change Canada (ECCC) – after applying the dry mask. Previous studies have shown ECMWF S2S hindcast models have outperformed both CFS and ECCC models on a global basis (de Andrade et al, 2019), but the performance of these three models have not been compared after the dry mask threshold has been applied. The North American Multi-Model Ensemble (NMME) forecasts have proved to perform better than individual models by pooling the ensemble members from several models (Krakauer, 2019). However, for this study, the NMME was not considered because the number of ensemble members varies between individual models, giving more weight to some models. Additionally, to improve multi-model performance, understanding individual models' type of errors and potential for correcting the biases before pooling the ensembles, which could further improve the multi-model forecast performance. Hence, this study will compare the decomposed NSE and associated errors of S2S precipitation forecasts of three individual models by season and lead time under three Koppen climate regimes across the CONUS.

The main intent of this study is to decompose the S2S forecasting skill as a function of lead time over the CONUS after applying the dry mask. To our knowledge, limited/no work has been performed on systematically decomposing the NSE over various seasons after applying the dry mask. In addition to applying the dry mask, evaluating model skill regionally is also critical as the precipitation has pronounced seasonality over the CONUS (Petersen et al., 2012).

Analyzing forecasting skills regionally can also provide insights on how land surface conditions, low-frequency oscillations, and regional hydroclimate influence the model performance.

The manuscript is organized into the following sections: S2S precipitation hindcast and observed databases from three different models are provided in the next section. Then, the dry mask threshold application procedure is presented along with the NSE decomposition. The following section provides the results from the full decomposition of ECMWF and the results from different regimes along with the skill comparison from three S2S reforecasts.

2.0 Data

This section provides the S2S hindcast database and observed data along with the details to calculate and decompose the NSE for S2S forecasts over various lead times and seasons.

Observed Precipitation

For calculating the S2S reforecasts skill, we used the CPC Global Unified Precipitation dataset provided by the NOAA Physical Science Laboratory (PSL), with a resolution of (0.5°x0.5°) (Chen, et al., 2008). Upon comparing the accuracy of various precipitation datasets, the CPC Unified dataset performed particularly well in areas that have dense areas of rain gauges (Beck et al., 2017). This study focused on the CONUS, which has a dense system of rain gauges, and has been used in other forecast verification studies (Becker et al, 2020).

S2S Hindcast Database

For S2S model skill evaluation, three hindcast models were assessed: 1.) European Centre of Medium-Range Weather Forecasts (ECMWF), 2.) National Center for Environmental Prediction's (NCEP) Climate Forecast System (CFS) model, and 3.) Environment and Climate Change Canada (ECCC). For full decomposition of ECMWF, the S2S hindcasts were evaluated for the full 20-year hindcast period (Table 1) and up to the longest available lead time of 45 days. The ensemble means were averaged over three different lead times: 1) 1-15 days, 2) 1-30 days, and 3) 1-45 days, and compared with the observed average precipitation corresponding to the three lead times. Additionally, the average forecasts and corresponding observed average daily precipitation values were pooled by the date of hindcast initialization into the following seasons: a) January, February, March (JFM), b) April, May, June (AMJ), c) July, August, September (JAS), d) October, November, December (OND). Thus, the evaluation for each season provides the skill of forecasts issued during the months within the considered four seasons.

For the model comparison section, the three models were assessed for lead times of 1-12 days, 1-22 days, and 1-32 days for four different seasons between January 1st 2000 and December 30th 2010, the longest available overlapping date ranges and lead times for all three models. Additionally, ECMWF and NCEP were compared for lead times of 1-42 days. The ECMWF hindcasts are initialized twice a week and range from 2000-2019, NCEP CFS hindcasts are initialized daily and are available from 1999-2010, and ECCC are initialized weekly, and

reforecasts range from 1995-2012 (Vitart et al., 2017). The S2S precipitation hindcast model's information and specification are shown in Table 1 (Vitart et al., 2017).

Model	LEAD TIME	RESOLUTION	HINDCAST PERIOD	HINDCAST ENSEMBLE SIZE	FORECAST ENSEMBLE SIZE	HINDCAST FREQUENCY	OCEAN COUPLING	SEA ICE COUPLING
ECMWF	0-46 Days	0.25°x0.25°, days 0-10, 0.5°x0.5°, after day 10 L91	Past 20 Years	11	51	Twice a Week	Yes	No
NCEP CFS	0-44 Days	~1°x1°, L64	1999-2010	4	16	Daily	Yes	Yes
ECCC	0-32 Days	0.45°x0.45°, L40	1995-2012	4	21	Weekly	Yes	No

Table 1. Subseasonal-to-Seasonal Hindcast Models and Forecast model information

2.1 Dry Mask application and Skill Assessment and Decomposition

a. Seasonality of Rainfall and Dry Mask Application

Prior to calculating the NSE for each hindcast-initialized season, a dry mask was applied based on the observed precipitation dataset to filter out the areas that receive small amounts of rainfall, which may result in an inflated forecast skill because the forecasted and observed rainfall have no rainfall. Antolilk (2000) and Charba et al., (2011) considered daily precipitation less than 0.01 inches as no event for evaluating the skill. Based on that work, the dry mask was set at a threshold value for each individual grid cell, if the observed daily precipitation over the 20 years is less than 0.15 inches, 0.30 inches and 0.45 inches for 15-day, 30-day and 45-day lead times from the time of issued forecast, respectively. The NSE and the three components were evaluated for all the three models for each lead time over the CONUS. We also evaluate the forecast skill – NSE and its components – based on the climate regime. For this purpose, we considered three main regimes – desert (regime B), temperate (regime C) and continental (regime D) – over the CONUS based on Koppen climate classification. A small area in southern Florida fell into the tropical (regime A) Koppen climate group; however, since this regime corresponds to only one grid cell from the hindcast model, we combined this tropical area with the temperate regime (Supplemental Information (SI) - Figure SI-1). Using the aggregated Koppen Climate Regime (Beck, et. al, 2017) into three climate regimes, a regional analysis was performed for each of the S2S hindcast models (Supplemental Information (SI) - Figure SI-1).

b. Skill Assessment Metrics

Skill assessment metrics measure the performance of the model's forecast ability compared to the observed variable. Frequently used performance metrics include anomaly correlation, NSE and Kling Gupta Efficiency (Clark et al., 2021). The NSE measures the magnitude of error variance from the model prediction compared to the observed variance in the data and has an upper bound of 1 but has a lower bound of -∞ and is used to determine the 'goodness-of-fit' of a

model. NSE is related to MSE but is normalized by the standard deviation of the observed precipitation or data values (Gupta et al., 2009).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\overline{o}_{it})^2} \quad (1)$$

Where o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. The mean observed precipitation is $\overline{o}_{it}$. For a given i, NSE will be decomposed into three parts (Murphy 1988; Weglarczyk 1998): A) Pearson's correlation coefficient (equation 3), B) conditional bias (equation 4), and C) unconditional bias (equation 5) (Gupta et al., 2009).

$$NSE = A - B - C$$

$$NSE = \rho_{xo}{}^2 - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad (2)$$

$$A=\rho_{xo}{}^2 \quad where \quad \rho_{xo} = \frac{cov(x,o)}{\sigma_x * \sigma_o} \quad (3)$$

$$B=[\rho_{xo} - \frac{\sigma_x}{\sigma_x}]^2 \quad (4)$$

$$C=[\frac{\overline{x}-\overline{o}}{\sigma_o}]^2 \quad (5)$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 1. The pearson correlation coefficient between x and o is ρ_{xo} (equation 3). The first component of the decomposition, Pearson's correlation coefficient, shows the linear association between the forecast and the observation. The conditional bias is the difference in the slope of the regression line fitted between forecast and observation with a slope of 1 that indicates a perfect forecast. The unconditional bias, indicating a systematic bias, denotes the ratio of difference between the mean of the observation and the mean of the forecast to the observed standard deviation.

3.0 Results

Full Decomposition of ECMWF

A full NSE decomposition was performed on the ECMWF S2S hindcast model because the ECMWF model has the longest available reforecast time range and has the largest number of ensemble members. Prior to decomposing NSE, a dry-mask threshold was applied based on the lead time for the climatological means of each grid cell, to mask out areas with low precipitation values to avoid inflated skill values. Both NSE and correlation are lower across all seasons after the dry mask threshold was applied. Figure 1a illustrates the difference in Normalized Nash-Sutcliffe Efficiency (NNSE) of 30-day ahead S2S precipitation forecast skill with and without the dry mask threshold). For instance, a forecast issued on March 30, 2000 with a lead time of 45 days corresponds to the skill of the forecast in predicting precipitation from March 30, 2000 to May 15, 2000. Thus, the skill of the forecast issued in JFM can cover the observed precipitation in April and May. To reiterate, all the figures with seasonal S2S performance metrics denote the skill summary of the forecast issued during that season as opposed to the ability to forecast the observed precipitation during that season.

To understand the importance of dry masking, we first show the 1-30 day ahead S2S precipitation forecast skill with and without dry mask (Figure 1) based on Normalized NSE (NNSE). Lower NNSE (equation 6) values, the inverse of NSE, indicate better predictive performance.

$$NNSE = \frac{1}{2 - NSE} \qquad (6)$$

For the forecast issued in the four seasons, the mean NNSE values are lower for the grid cells below the dry mask threshold than for the grid cells that exceeded the threshold (Figure 1). Even though including "no-precipitation event" is expected to inflate the skill, dry masking by filtering out regimes rather than simply removing values below a given threshold, allows us to maintain the same sample size across all grid cells, thereby changing the masked areas based on both forecast-initialized seasons (Figure 1) and lead time.

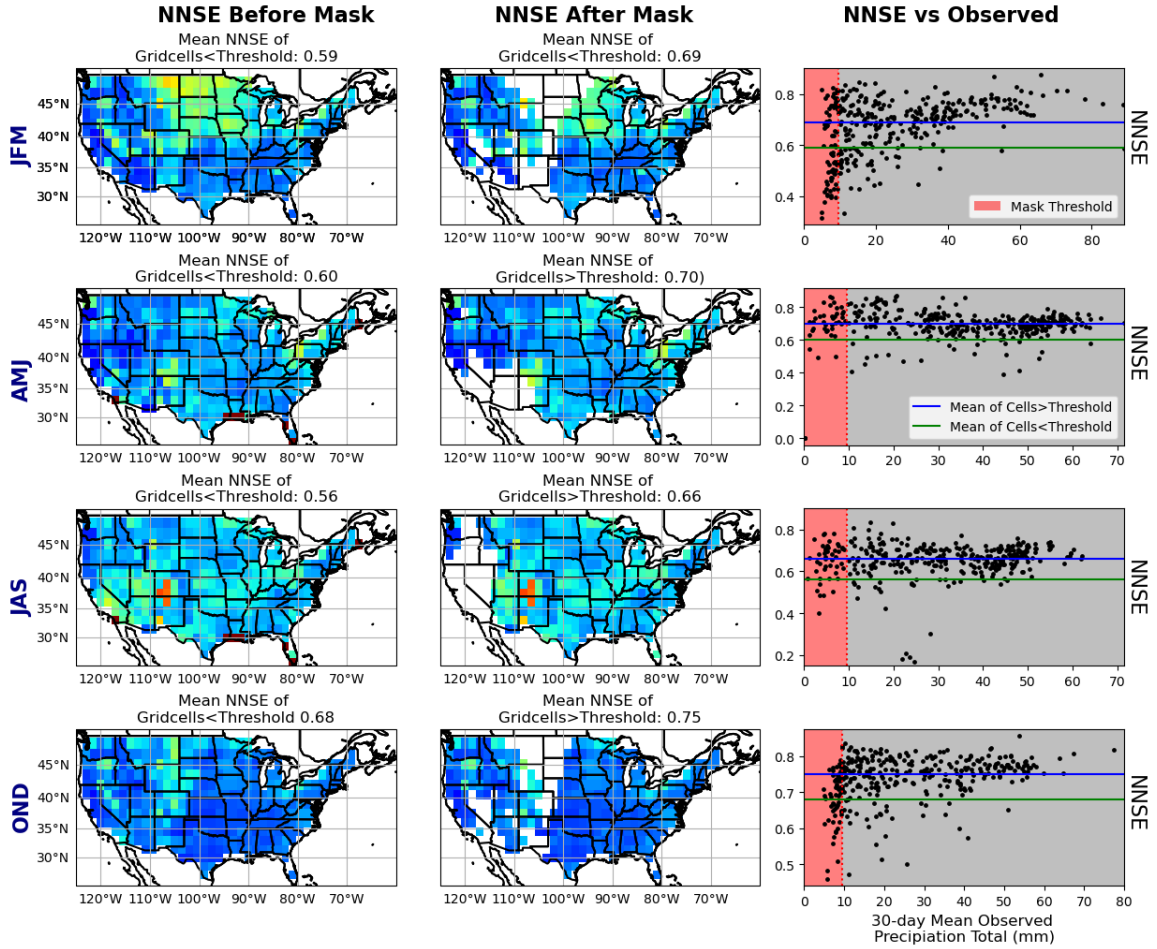


Figure 1. Normalized Nash Sutcliffe Efficiency (NNSE) of 1-30 days ahead ECMWF hindcast for the CONUS before dry mask is applied (left column) and after (middle column) dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS and OND for 1-30-day lead time. The scatter plot comparison of grid cell's 1-30-day climatological precipitation means and the corresponding Normalized NSE values (right column). The scatter plot shows the NNSE values that fall below the dry mask threshold (red region) and above (gray region). The average NNSE of the grid cells below the dry mask threshold (green) and above the dry mask threshold (blue). Since the NNSE is the inverse of the NSE, the lower NNSE values indicate better predictive performance.

The overestimation of S2S forecast skill occurs if no dry mask is applied, particularly for pronounced dry seasons (JFM and JAS). Studies that evaluated S2S precipitation forecasts skill did not consider dry mask application, which ignores the seasonality in precipitation, thereby indicating potential difference in forecast skill between regions (e.g., Li et al., 2022). However, after the dry mask application (Figure 1), we find that the skill was fairly similar between regimes. Thus, it is important to apply a dry mask which inherently considers the seasonality in precipitation for skill evaluation. Quantifying the forecast skill for critical events (e.g., peak rainfall seasons) is important particularly if the interest is to identify regions with limited skill.

a) NSE Spatial Patterns

We present results for the NSE and its decomposition (Figures 2-7) for the ECMWF model and then compare its performance with NCEP and ECCC later (Figures 8-10). Before assessing the components of the NSE, we first investigate the NSE over the CONUS, which shows the S2S forecasting skill of ECMWF for various lead times over the season (Figure 2). NSE is better in the winter and fall seasons (JFM and OND) in comparison to spring and summer seasons (AMJ and JAS) (Figure 2), which is partially due to El Nino Southern Oscillation (ENSO) being active during winter and fall months and ENSO dying or being at an incipient stage during AMJ and JAS (Ham et al., 2019). The NSE also tends to be better closer to the coasts indicating the local sea surface temperatures (SSTs) in influencing S2S forecasts. Additionally, the NSE shows a slight gradient from West Coast to East Coast (Figure 2). The NSE tends to be weaker around the Great Lakes. Further, the areas surrounding the dry mask regions tend to have a lower NSE.

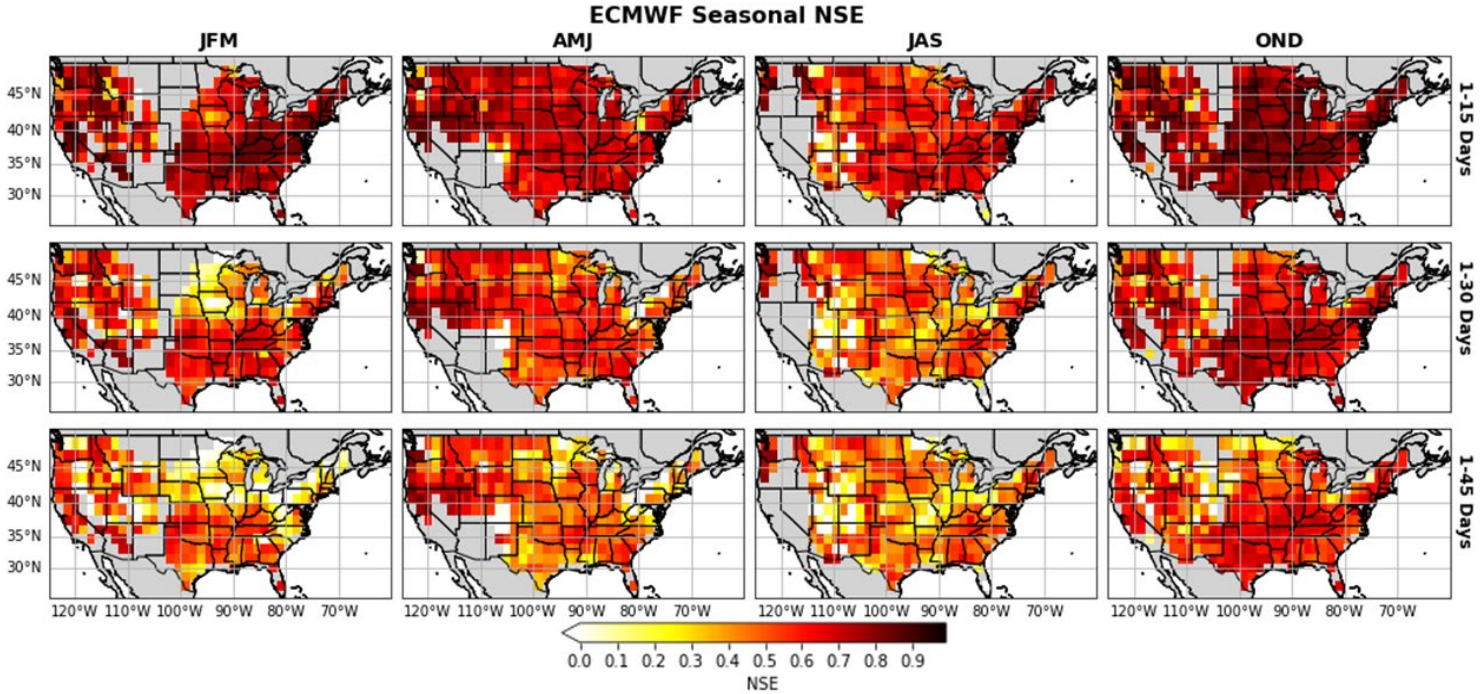


Figure 2. Nash Sutcliffe Efficiency (NSE) of ECMWF hindcast for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

b) Decomposition Plots

We decompose the NSE of ECMWF in Figure 2 into correlation (Figures 3), conditional bias (Figure 5) and unconditional bias (Figure 6) for each lead time for the four seasons.

i) Correlation and its longitudinal distribution

The first component of decomposition, Pearson's correlation coefficient, shows the innate model skill and the lower bound for explained variance in the model. The analysis of correlation shows that the skill decreases as lead time increases for all seasons (Figure 3.). Similar to the NSE, the correlation is also lower in the summer seasons and higher in the winter seasons. The correlation between S2S precipitation hindcasts and observed precipitation was averaged by longitude, for each season and lead time, after the dry mask threshold was applied. This longitudinal distribution more clearly illustrates the West to East coast gradient, where the correlation is higher in the West Coast and decreases towards the East Coast (Figure 3-4).

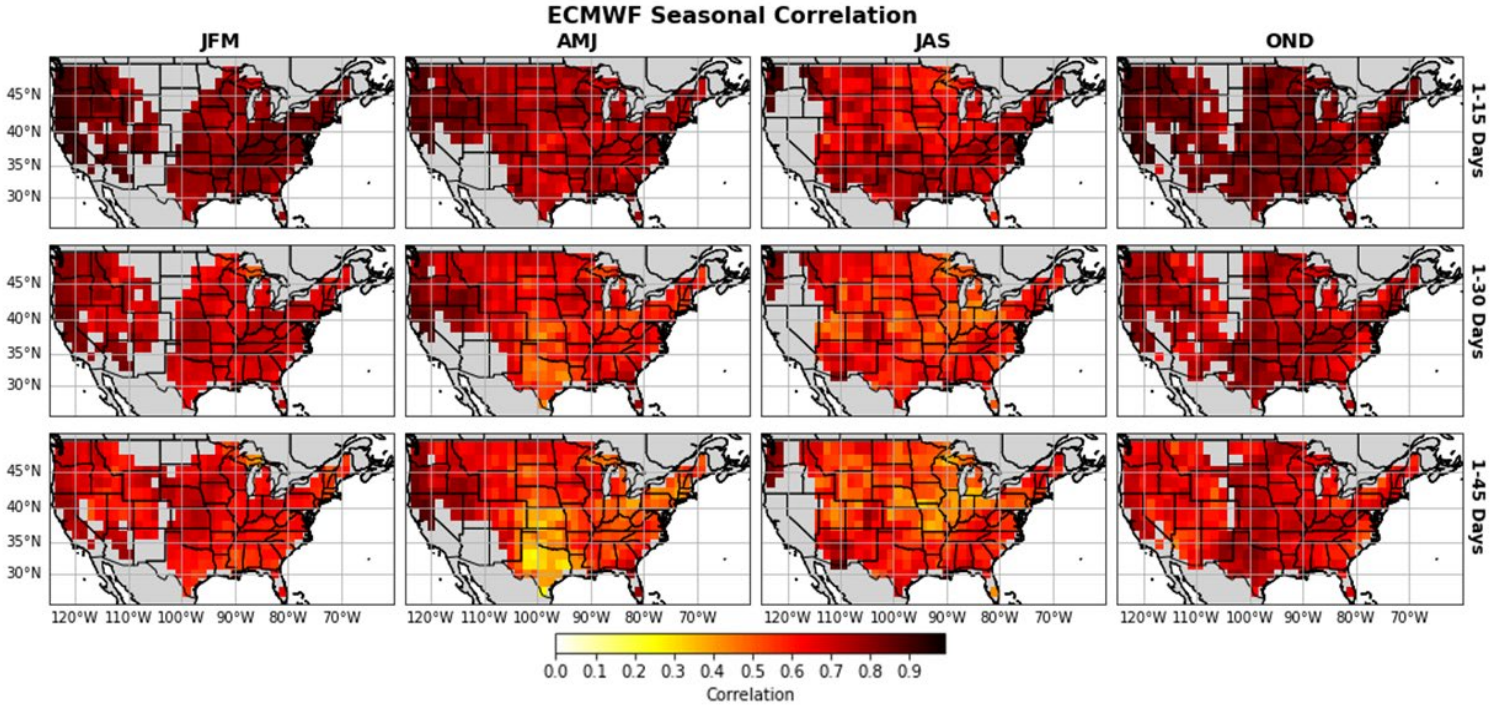
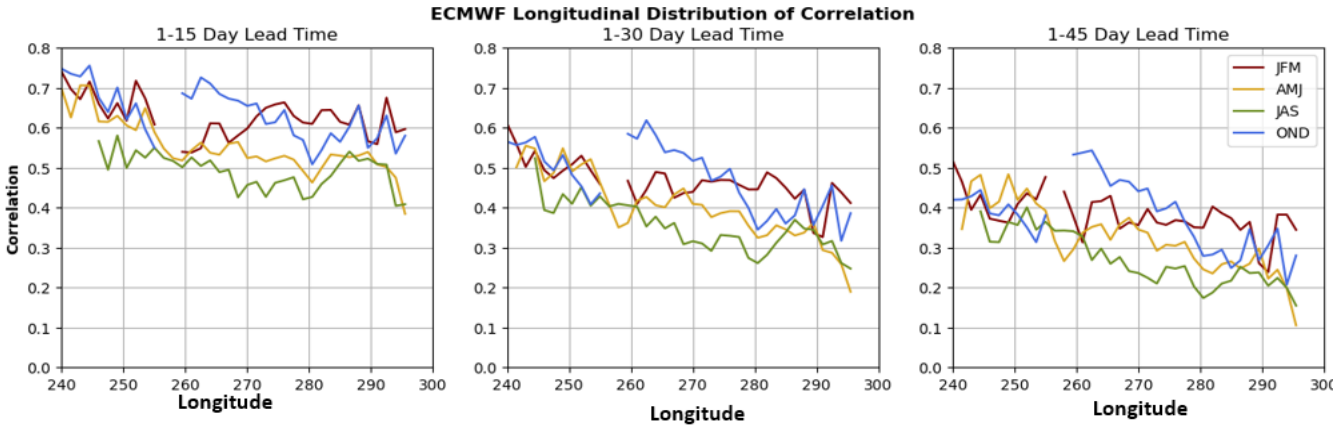


Figure 3. Correlation, the first component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four seasons of initialized forecasts: JFM, AMJ, JAS, and OND, and for three different lead times: 1-15 days, 1-30 days, and 1-45 days.

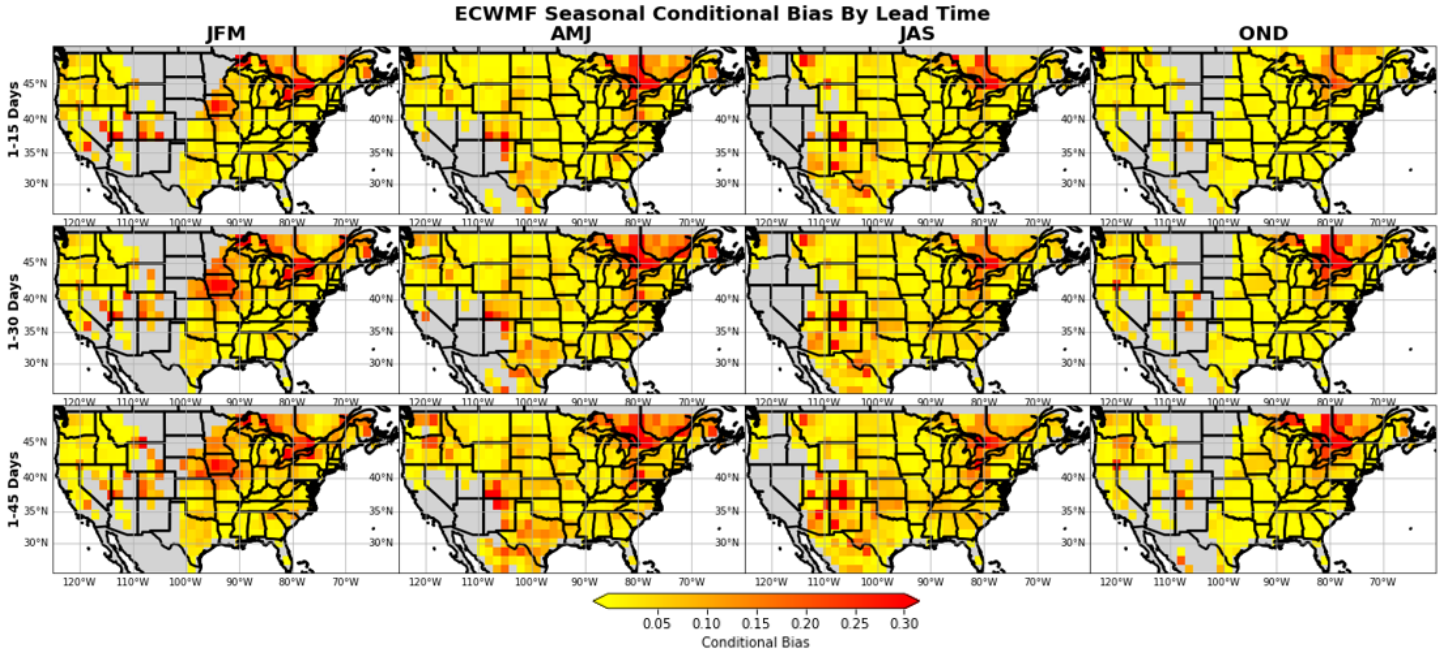
On the West Coast, correlation coefficients are higher than on the East Coast, which is partially due to the pronounced seasonality in precipitation over the West Coast that results in reduced number of grid cells being considered for evaluation after applying the dry mask. Additionally, correlation coefficients are higher towards the coasts and weaker further inland due to potential influence of local SSTs (Sankarasubramanian et al., 2017). Correlation coefficients are also lower towards the area surrounding the masked out regions.



313 *Figure 4. Longitudinal distribution of correlation by the average by latitude of the ECMWF hindcast data for*
314 *CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and*
315 *for three lead times: 1-15 days, 1-30 days, and 1-45 days*

316 *ii.) Conditional Bias*

317 The second and third components, conditional bias, and unconditional bias, are expected
318 to be zero for ideal forecasts. The conditional bias for the ECMWF decomposition increases as
319 lead time increases and tends to be higher towards the coasts. Further, the conditional bias is
320 higher during the summer season in comparison to the winter season (Figure 5). The Great
321 Lakes Region and the central part of the US has a high conditional bias that increases with
322 increasing lead times, whereas the Sunbelt has a low conditional bias during the winter and
323 spring seasons. Conditional bias is also higher towards the areas that were masked out from the
324 dry mask. Conditional bias is highest during JAS, specifically in the desert areas that were
325 masked out during the other seasons and is lowest during OND.



326 *Figure 5. The second component, conditional bias, of NSE decomposition, from the ECMWF hindcast data for*
327 *CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and*
328 *for three lead times: 1-15 days, 1-30 days, and 1-45 days.*

iii.) Unconditional Bias

The third component, unconditional bias, represents the systematic bias in reproducing the long-term mean of the observed precipitation. Unconditional bias is high in the Great Lakes Region and in the central part of the US (Figure 6). Additionally, unconditional bias is high in the desert regions for JAS, which was masked during the other seasons, for JAS. Conditional bias and unconditional bias are generally correlated and have higher values in the same regions.

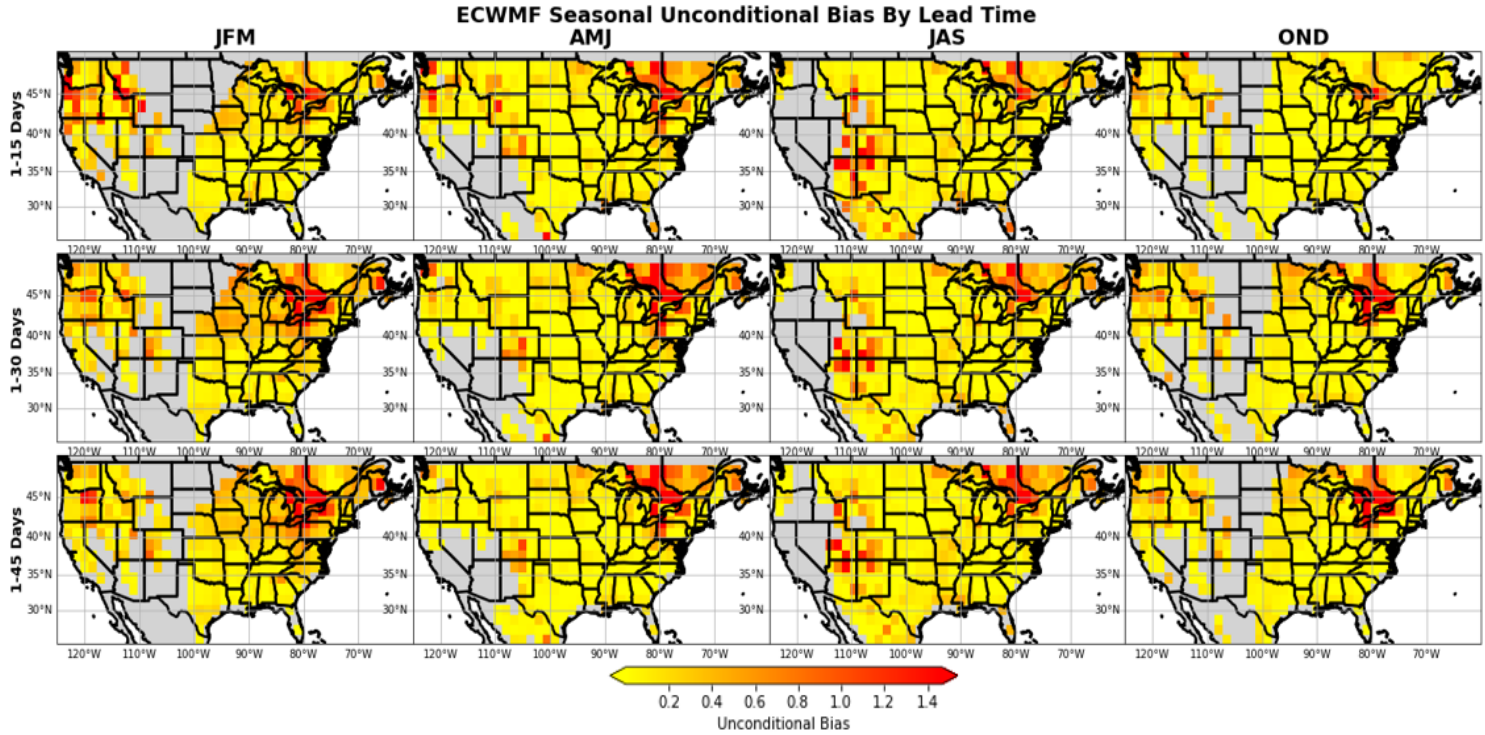


Figure 6. Unconditional bias, the third component of NSE decomposition, from the ECMWF hindcast data for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-15 days, 1-30 days, and 1-45 days.

c. Skill comparison across Koppen Climate Regimes

The skill of ECMWF S2S hindcast model was compared under three Koppen climate regimes: a) desert b.) temperate and c.) continental (Figure SI-1). For all lead times and climate regimes, the correlation varies by season and is lower in the summer months and is the highest in the winter months (Figure 7). Since the dry mask threshold was applied before the climate regime classification was considered, the correlation does not vary much between regimes within a given season. Conversely, if a dry mask had not been applied, the desert regimes may expect to have much better skill, because of inflated skill due to no-precipitation days.

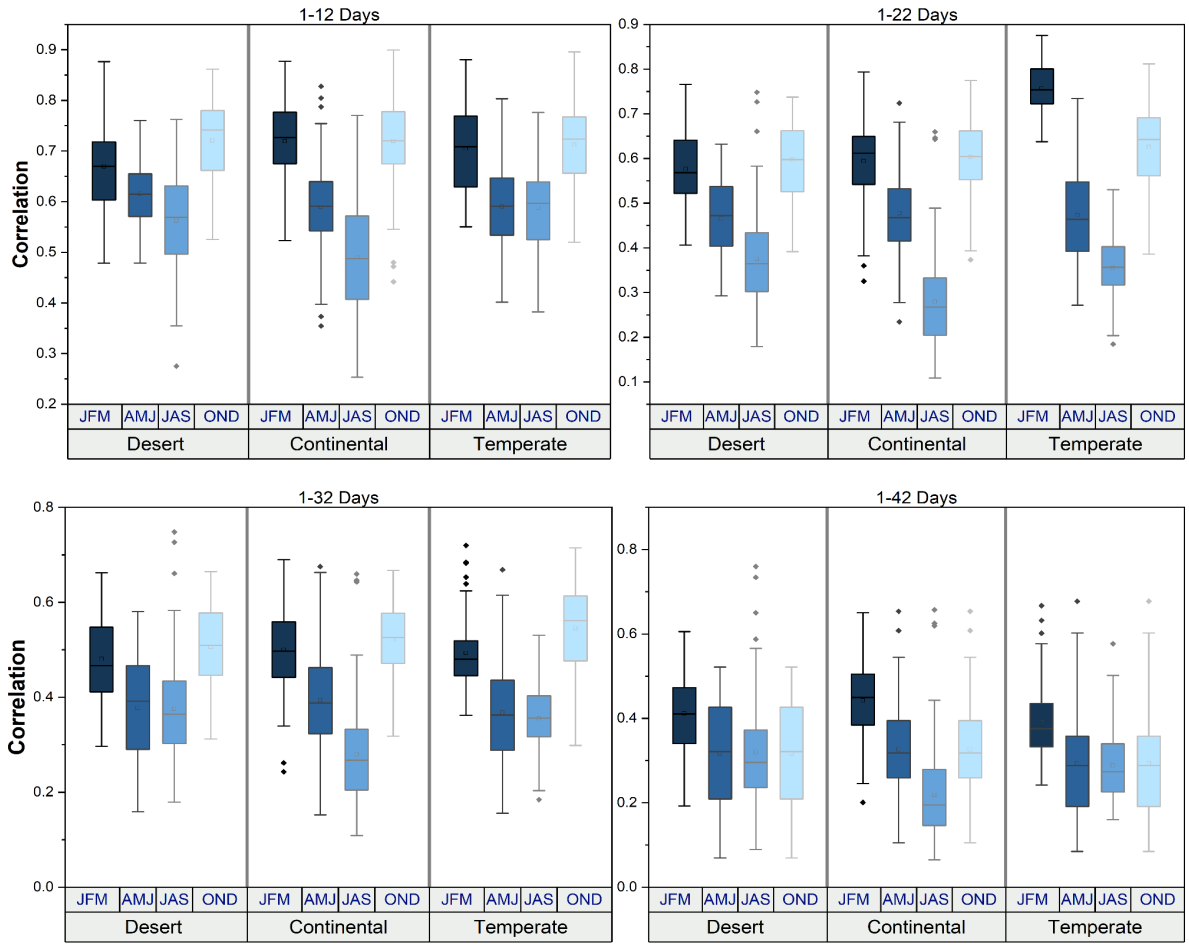


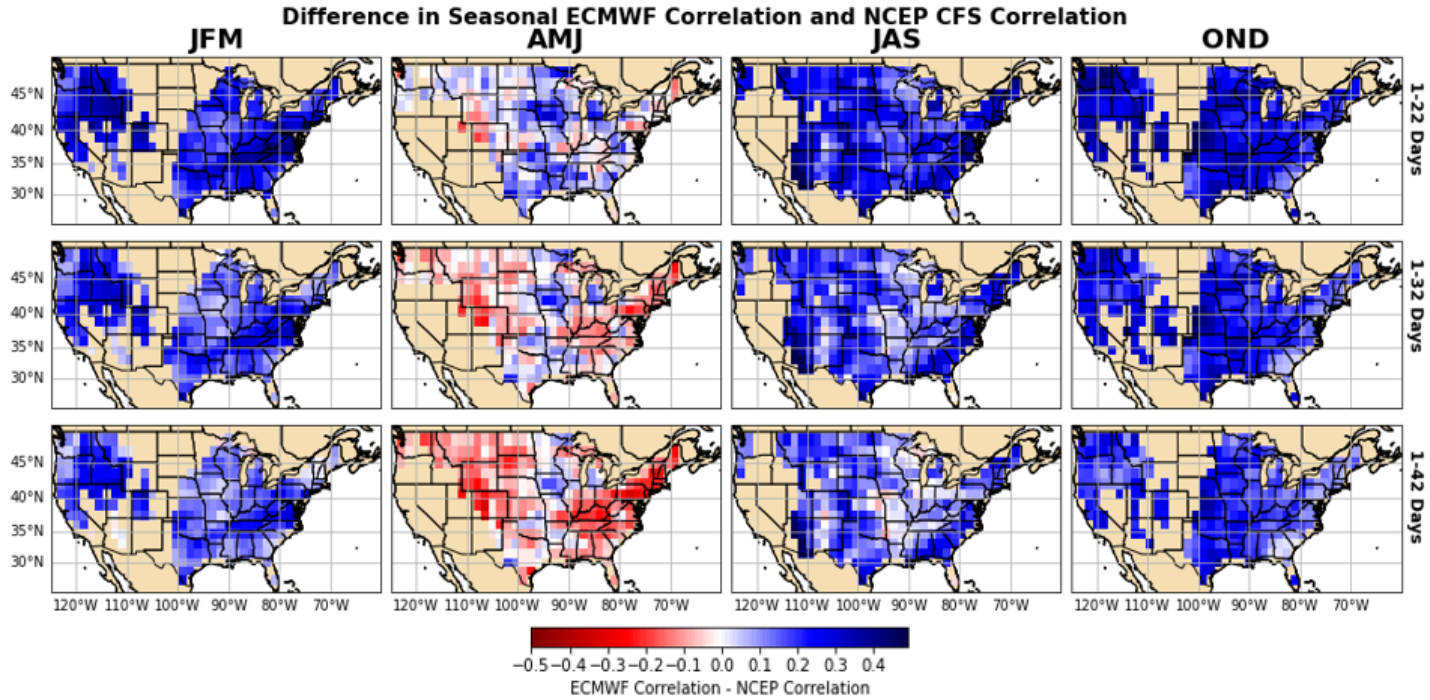
Figure 7. The box and whisker plot of correlation from the ECMWF hindcast model for three Koppen climate regimes: desert (red), temperate (blue) and continental (green) for lead times 1-12, 1-22, 1-32, and 1-42 days for all four seasons that the forecasts were initialized: JFM, AMJ, JAS, OND.

d. Model Comparison of NSE and Correlation

Comparing S2S hindcast models is important to understand the relative performance of the individual models. In this analysis, ECMWF's NSE was compared to NCEP CFS's NSE and next ECMWF's correlation was compared to all three models. The dry mask threshold may affect the model performance; therefore, forecast skill was not considered in areas where the historically observed precipitation did not exceed this threshold.

The blue regions in Figure SI-2 show where ECMWF's NSE outperforms the NSE of NCEP CFS for most lead times, regimes, and seasons, especially at shorter lead times, except for a few inland areas. Although ECMWF's NSE is higher than NCEP's in most regimes, seasons, and lead times, the ECMWF and NCEP CFS's correlation is closer in value (Figure 8). NCEP CFS has a higher NSE and correlation than ECMWF during AMJ. In comparison to ECMWF, NCEP's correlation improves with longer lead times during AMJ and is also higher in areas

365 further inland. Conversely, ECMWF has better performance around the coast (Figure 8) except
366 for OND, which may be due to the two different ocean models used in the initializations.



367 *Figure 8. Difference in Correlation values between ECMWF S2S hindcast and NCEP CFS for CONUS after dry*
368 *mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead*
369 *times: 1-12 days, 1-22 days, and 1-42 days.*
370
371 ECMWF and ECCC models' correlation differ by season but Figure 9 does not show a
372 clear inland-coastal differential in skill (Figure 9), which could be potentially due to ECMWF
373 and ECCC having the same ocean models. ECCC has a higher correlation than ECMWF during
374 the forecasts initiated in the summer months (JAS). However, since ECCC's lead time ranges
375 from 1-32 days, 1-42 day lead time between ECMWF and ECCC could not be compared.
376
377 Across seasons and lead times, NCEP CFS's correlation is higher than ECCC's
378 correlation for NCEP (Figure SI-3). NCEP CFS' model performance improves noticeably at
379 longer lead times and was not compared to 1-42 days lead time because of ECCC's shorter lead
380 time forecast availability. However, when comparing the first component, correlation, by
381 regime, season, and lead time, ECCC has higher correlation in AMJ, when compared to both
382 NCEP CFS as well as ECMWF. However, ECCC's performance tends to be worse in the
383 remaining three seasons.
384
385

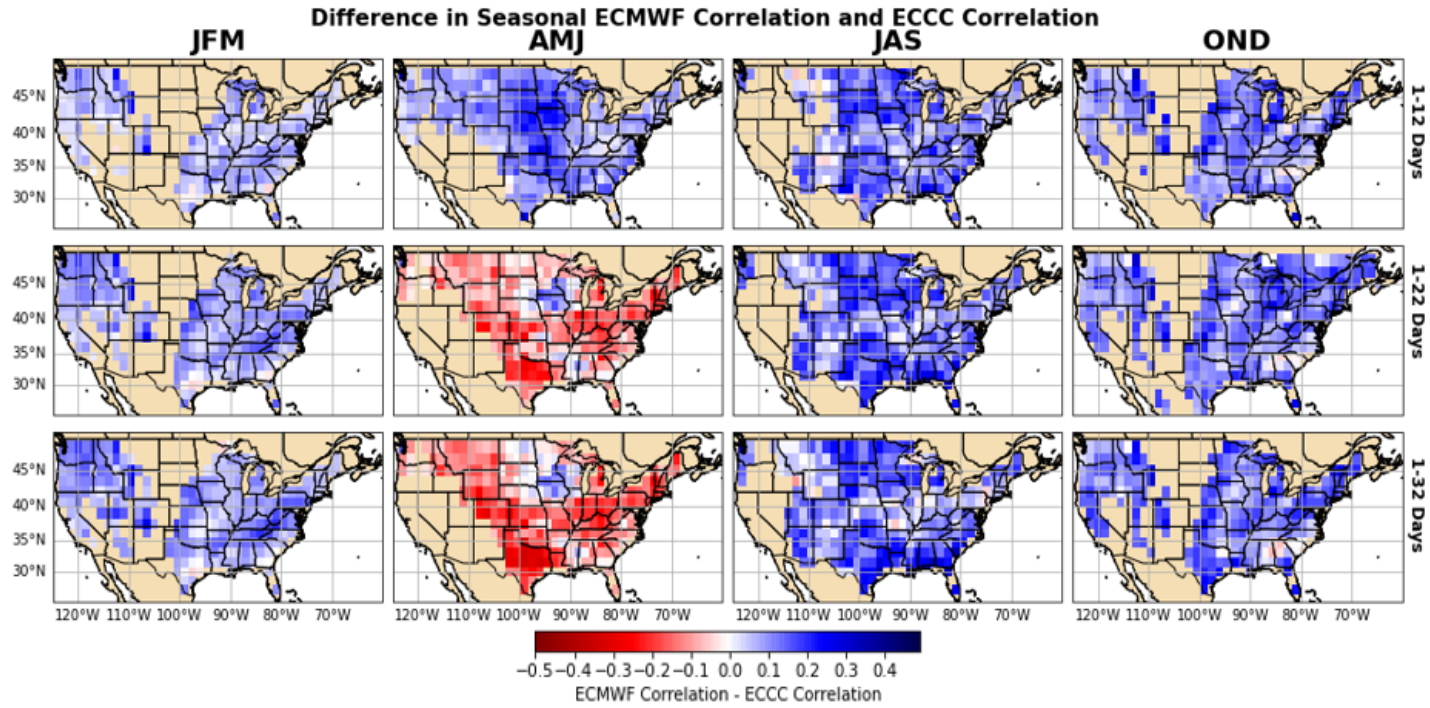


Figure 9. Difference in Correlation values between ECMWF S2S hindcast and ECCC for CONUS after dry mask threshold is applied for four season of initialized forecasts: JFM, AMJ, JAS, and OND, and for three lead times: 1-12 days, 1-22 days, and 1-32 days.

Overall, ECMWF's correlation for the forecast issued in seasons, JFM and OND, is higher than the other two models, but ECMWF's correlation is lower than the other models for the forecasts issued in AMJ (Figure 8-9). ECMWF has the highest NSE and correlation when solely considering the skill within the CONUS boundaries; however, NCEP CFS and ECCC hindcast models have much better forecast skill in the Great Lakes regime on and the Canadian regime just north of the Great Lakes, which although may not fall within the US boundaries, is still critical for the Midwest's water resources. ECMWF performs better towards the coasts and the skill may be higher in the winter seasons due to the areas that were masked out by the dry mask threshold. NCEP CFS and ECCC perform better in areas further inland, which is why the skill may be noticeably better in the spring and summer months (AMJ and JAS) where the inland regimes are not masked by the dry mask threshold since the regime receives higher precipitation during the summer. The differences in model skill could be due to the different ocean models and different initialization schemes, however this attribution has to be systematically analyzed further.

e. Model Skill comparison across Koppen Climate Regimes

The performance metrics for the three hindcast models were analyzed across the three Koppen climate regimes over the CONUS. Each model's NSE and the decomposed components were divided into climate regimes by season and lead times. At longer lead times, the differences in NSE reduces across seasons and climate regimes with NCEP CFS beginning to outperform

ECCC (Figure SI-4). ECMWF's NSE was higher than the NSE of ECCC and NCEP CFS across climate regime, season, and lead times (Figure SI-4), because NCEP and ECCC had high unconditional and conditional biases (Figure SI-4). Since these biases can be reduced to zero with simple post-processing techniques such as Model Output Statistics (Appendix A), we focused on comparing correlation (Figure 10).

The Pearson correlation coefficient is generally higher for ECMWF in comparison to ECCC and NCEP CFS models for all lead times, regimes, and seasons (Figure 10). There does not seem to be a consistent trend on how models perform for each climate regime across seasons and lead times even though both NCEP and ECCC perform better with forecasts issued in AMJ (Figure SI-4). For ECMWF and ECCC, the correlation is higher at shorter lead times, but NCEP's correlation remains relatively consistent across lead times (Figure 10). Across all models, lead times, and regimes the seasonal patterns illustrate that correlation is the highest during JFM and OND and lowest during AMJ and JAS.

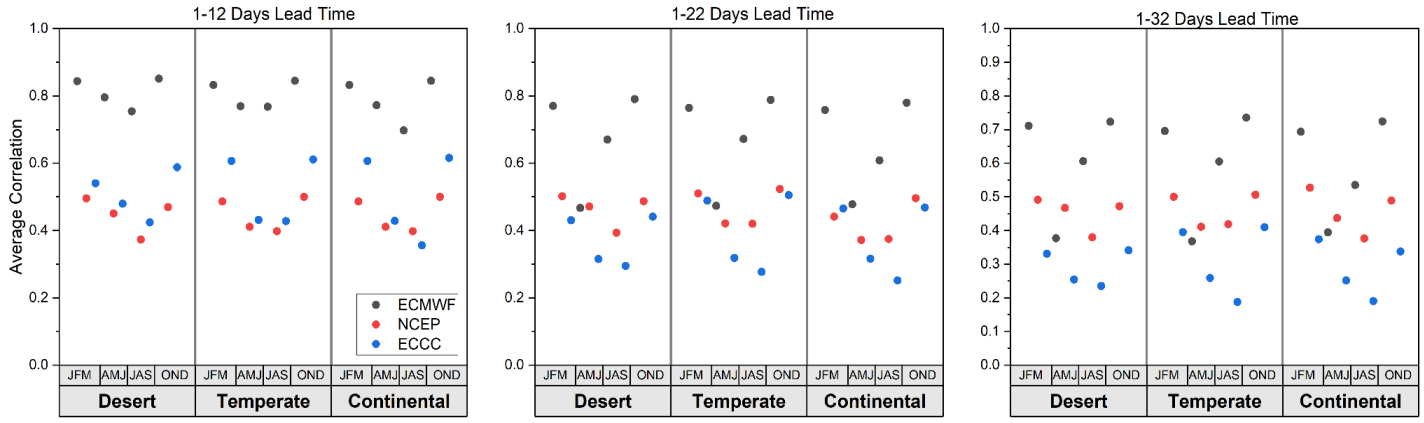


Figure 10. The average correlation for each regime: Regime B (desert), Regime C (temperate), and Regime D (continental) for each model: ECMWF (black), ECCC (blue), and NCEP CFS (red). The average correlation was calculated by lead time a) 1-12 days b.) 1-22 days and c.) 1-32 days for seasons JFM, AMJ, JAS, and OND.

The conditional bias is the lowest for ECMWF and highest for NCEP CFS particularly for AMJ and at shorter lead times (Figure SI-4). NCEP's median marginal bias was lower than ECMWF and ECCC, but one grid cell on the West Coast had a very high conditional bias causing the mean bias of all of the grid cells to be higher than the other two models.. ECCC has the highest conditional bias at the shorter lead times and ECMWF and NCEP CFS were comparable at 1-12 days for JFM, JAS, and OND. Conditional bias has the highest spread during spring months (AMJ). With longer lead times (e.g., 1-32 days), the unconditional bias across the selected models is similar, with ECCC being slightly higher than the other two models. No clear regional pattern of unconditional bias across all models and seasons was evident (Figure SI-4 g-i). The seasonality of unconditional bias seems to change based on lead times. We discuss in the next section how the conditional bias and unconditional bias could be potentially improved using post-processing techniques that focus on developing statistical relationships between model forecasts and the observed precipitation.

4.0 Discussion

Understanding the S2S precipitation forecasts skill across the CONUS over different seasons, as well as highlighting potential avenues for model improvement is critical for better forecast application. This study a) investigated and compared the spatial distribution of NSE for three S2S precipitation hindcast models across the CONUS, b) decomposed Nash-Sutcliffe Efficiency into correlation, conditional bias and unconditional bias based on the lead time and forecast issued in a season for each model and c) analyzed model skill across three (tropical, desert and temperate) Koppen Climate regimes. Our analysis shows that NSE of ECMWF was higher closer to the coast, most likely due to the influence of MJO and ENSO, and was also higher for the forecast issued during winter months and with shorter lead times. Decomposition of NSE shows that the first component, correlation, illustrates there is a gradient in skill from west coast (higher) to east coast (lower). Both the conditional and unconditional biases were also smaller during the winter months and in areas closer to the coast. The model comparison showed that ECMWF performs well in the winter seasons and towards the coasts, whereas NCEP CFS's performance is the best for forecasts issued during AMJ and in inland areas. The conditional and unconditional bias were high over the Midwest Great Lakes region. The conditional bias was higher for NCEP CFS, particularly for forecasts issued in AMJ and the unconditional bias was high for forecasts issued in JAS. ECCC's skill is high during AMJ and at short lead times, but decreases significantly with longer lead times. No clear trends were observed across the climate regimes across the three hindcast models' performances, but NSE and correlation was higher for the winter seasons than the summer seasons consistently for all the lead times, regimes and three models.

Potential for improving S2S forecasts

Even though our analysis, after application of dry mask, showed that conditional bias and unconditional bias are the primary reasons for low and negative NSE values for the S2S hindcasts, this could be overcome by selecting a proper post-processing scheme where the correlation is high across the CONUS. One of the commonly used post-processing scheme for correcting weather/climate forecasts is Model Output Statistics (MOS), which is a linear regression model that uses the forecast or a transformation of it (e.g., principal components) as a predictor and the observed precipitation as a predictand (Antolik et al., 2000; Sankarasubramanian et al., 2008). One advantage with a linear regression model is that it reduces the marginal bias to zero (Appendix A). Further, we also show analytically in Appendix A, a linear regression model reduces the conditional bias to zero which turns the NSE of the corrected forecasts from a MOS being equal to the square of the correlation coefficient (i.e., component A). Thus, a linear regression based MOS provides a lower bound on the NSE of the forecast to be decomposed component A, thereby providing a guidance on where post-processing schemes will be useful for a given location/regime. An example of where post-processing can be effective for correcting bias is NCEP CFS's 1-42 day forecasts. ECMWF did not have any grid cells where NSE was below zero, because the conditional and unconditional bias were low, so we

show NCEP, which has large sources of unconditional and conditional bias across all regimes, but relatively high correlation (SI-4).

Figure 11 shows locations where a) NCEP's NSE is less than zero and correlation is significant ($p<0.05$), b) NCEP's NSE is greater than zero and correlation is significant ($p<0.05$), and c) NCEP's NSE is less than zero, but correlation is not significant ($p>0.05$) for 1-42 day lead times. For the first case, where NSE is low and correlation is high, post-processing such as MOS can be effectively used to reduce conditional and unconditional biases to improve forecast skill, and a large portion of CONUS, mostly inland area and particularly for forecasts issued in seasons JFM and AMJ (Figure 11). For the second category, a large portion of the coastal region, particularly in forecast-initialized seasons AMJ and OND, have significant ($p<0.05$) correlation and high NSE, which means post-processing will not be effective as the model does not capture the observed variability. Similarly, post-processing will not be effective in areas with low NSE and correlation that is not significant ($p>0.05$), which includes a few grid points in AMJ and JAS (Figure 11). Even though linear-regression based MOS may not result in improved skill in areas where both NSE and correlation are low, other MOS post-processing schemes can be considered such as a semi-parametric model or machine learning models (Glahn et al., 1972; Taillardat et al., 2019), NSE of S2S forecasts could be potentially improved as such models are more flexible in reducing the mean square error in the forecast.

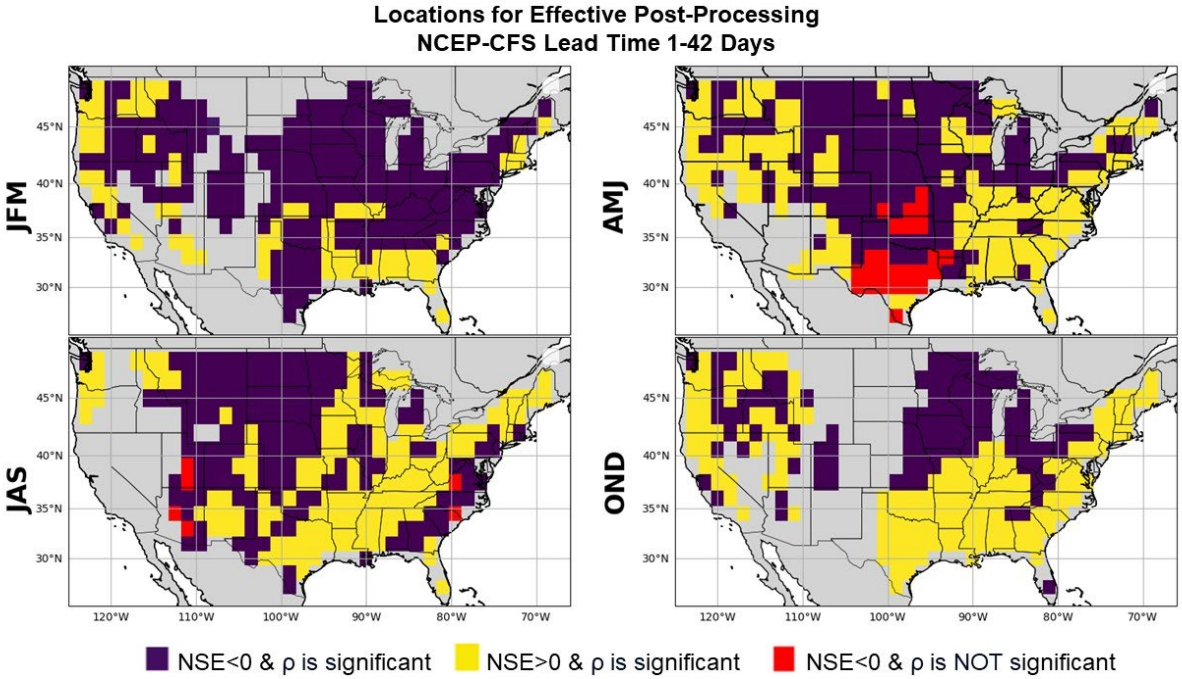


Figure 11. Post-processing will be effective in the locations where NSE<0 and correlation is significant (purple), but will not be necessary in places where NSE<0 but correlation is not significant (red) or in places where correlation is significant (yellow).

Even though the selected models had ensemble forecast, we considered only ensemble mean for forecast decomposition. We did not consider probabilistic forecasts such as Brier Skill score for skill evaluation and decomposition since the differences in ensemble members could

significantly affect the forecast evaluation. Similar decomposition on Brier score could reveal the forecast reliability and resolution of each model's performance in below-normal and above-normal conditions (Brier, 1950). Further, our analysis focused on decomposition without evaluating the model's performance during extreme conditions, which could be pursued further to understand the sources of bias. Our analysis also did not consider NMME because the number of ensemble members varies between models, giving more weight to some models. Additionally, the models within NMME have varying forecast issued frequencies, lead times, and issued dates. These varying model features within the multi-model need to be addressed before valid model comparisons can occur. Since the intent of this study was to show a systematic process of evaluating model skill and comparing across the models, we did not consider NMME for our study.

5.0 Conclusions

S2S precipitation forecasts are critical for operational and proactive water resource management and planning. Systematic S2S forecast skill assessment is essential for understanding existing model skill and how different errors contribute to it. Our evaluation of three S2S reforecasts – ECMWF, ECCC and NCEP – based on NSE decomposition primarily looked at the skill of forecasts issued during four seasons and under three different lead times. Our analysis shows the importance of applying dry mask as the NSE and correlation are lower across all seasons after masking areas with low precipitation values. The full decomposition of ECMWF revealed a West to East coast longitudinal gradient in NSE and correlation. Decomposed components, conditional and unconditional bias, did not show any longitudinal trends. ECMWF's skill showed that seasonal trends in forecast skill occurred across all lead times and all seasons, but correlation did not differ by climate regimes.

The forecast skill and associated errors were also compared across models. Overall, ECMWF's model performance was stronger than both ECCC and NCEP CFS's performance, mainly for the forecasts issued during the winter months, (JFM and OND). ECMWF had the highest NSE across the three climate regimes – temperate, desert and continental – considered. However, ECCC and NCEP CFS performed better for the forecast issued during the spring months, and also performed better in areas further away from the coast. Our decomposition efforts show S2S improvements in physical modeling efforts such as parameterization and initialization should be undertaken for ECMWF particularly for areas further from the coast, for forecasts issued in the spring months, AMJ, and for NCEP CFS and ECCC for the forecasts issued in the winter months over coastal areas.

Our analytical derivation on how MOS could help improve the forecast shows that a linear regression based MOS could ensure the NSE of the post-processed forecast to be component A, which is the square of the correlation coefficient between forecasts and the observation. This shows because simple linear regression based MOS can eliminate conditional

and marginal biases. This also provides information on regions (Figure 11, NSE <0 and ρ not significant) where S2S forecasting schemes can focus on improved model parameterizations and initializations including coupling with land surface models for improving the skill (Entekhabi et al., 1999).

Acknowledgments

The first author was supported by the National Science Foundation Fellowship (NSF) for the Graduate Research Fellowship Program (GRFP) support (award # DGE-2137100). Apart from that, this research was also supported by two NSF grants (award # CBET - 1805293 and IIE-2033607).

Open Research

The hindcast model data was accessed on the ECMWF S2S reforecast portal (https://apps.ecmwf.int/datasets/data/s2s/). The CPC Unified Gauged-Based observed precipitation dataset are available at https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html, and the Koppen climate classification data are available at www.gloh2o.org/koppen/.

Appendix A. Decomposition of NSE for Linear-Regression Based Model Output Statistics

For each grid cell, o_{it} is the observed precipitation value, x_{it} is the corresponding S2S precipitation value and y_{it} is the corrected precipitation value, where t = 1, 2…n is the time index with 'n' forecasts and i is the lead time of the forecast. Linear regression model 2 is used for the model to get the corrected precipitation value, which is the MOS estimate.

$$o_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \quad [1]$$

$$y_{it} = \beta_0 + \beta_1 x_{it} \quad [2]$$

For a given i, NSE is originally between o_{it} and x_{it} (equation 3), but a linear regression is used to estimate the corrected precipitation, y_{it}. For a given i , the NSE is calculated between o_{it} and y_{it} (equation 4) and then decomposed into parts A (equation 8-14), B (equation (15), and C (equation 16-17).

$$NSE_i(o_{it}, x_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-x_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = \rho_{xo}^{\,2} - (\rho_{xo} - (\frac{\sigma_x}{\sigma_o}))^2 - (\frac{\overline{x}-\overline{o}}{\sigma_o}))^2 \quad [3]$$

$$NSE_i(o_{it}, y_{it}) = 1 - \frac{\sum_{t=1}^{n}(o_{it}-y_{it})^2}{\sum_{t=1}^{n}(o_{it}-\bar{o}_{it})^2} = = \rho_{yo}^{\,2} - (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 - (\frac{\overline{y}-\overline{o}}{\sigma_o}))^2 \quad [4]$$

$$\beta_1 = \frac{cov(o-x)}{\sigma_x^{\,2}} \quad [5] \qquad \beta_0 = \overline{o} - \beta_1 * \overline{x} \quad [6] \qquad \beta_1 = \frac{\rho_{xo} * \sigma_x * \sigma_o}{\sigma_x^{\,2}} = \frac{\rho_{x\,o} * \sigma_o}{\sigma_x} \quad [7]$$

Where σ_x and σ_o represent the standard deviation of x and o, and $\overline{o}$ and $\overline{x}$ represent the mean of x and o once x_{it} and o_{it} were summed from 1 to n for lead time i in equation 3. The pearson correlation coefficient between x and o is ρ_{xo}. For the corrected precipitation, y_{it}, the standard deviation and mean are σ_y and $\overline{y}$ respectively, when y_{it} is summed over time from 1 to n for lead time i in equation 4. The correlation coefficient between o and y is ρ_{yo} .

NSE of o_{it} and y_{it} is decomposed into the three corresponding parts a.) correlation, b.) conditional bias and c.) unconditional bias. It is important to note that correlation, Component A (o, y), will be the same as the Component A (o, x). Where

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [8]$$

$$\rho_{xo} = \frac{cov(x, o)}{\sigma_x * \sigma_o} \quad [9]$$

$$cov(y, o) = cov(\beta_0 + \beta_1 x, o) = \beta_1 \, cov(x, o) \quad [10]$$

$$var(y) = \beta_1^{\,2} * \sigma_o \quad [11] \qquad \sigma_y = \beta_1^{\,2} * \sigma_x \quad [12]$$

$$\rho_{yo} = \frac{cov(y, o)}{\sigma_y * \sigma_o} \quad [13]$$

$$\rho_{yo} = \frac{\beta_1 * cov(y, o)}{\beta_1 * \sigma_x * \sigma_o} = \frac{\beta_1 * cov(y, o)}{\sigma_y * \sigma_o} = \rho_{xo} \quad [14]$$

Conditional bias B (o, y) will be reduced to zero MOS estimates.

$$B(o, y) = (\rho_{yo} - (\frac{\sigma_y}{\sigma_o}))^2 = (\rho_{xo} - \frac{\beta_1 * \sigma_x}{\sigma_o})^2 = (\rho_{xo} - \frac{\sigma_y}{\sigma_o})^2 \quad [15]$$

638 $$\rho_{xo} = (-(\frac{\rho_{xo} - \sigma_o}{x}) \cdot \frac{\sigma_x}{\sigma_o})^2 = 0$$

639 Unconditional bias C (o , y) will also reduce to zero for MOS estimates.

640 $$C(o, y) = (\frac{\overline{y} - \overline{o}}{\sigma_o}))^2 \qquad [16]$$

641 $$\overline{y} = \beta_0 + \beta_1 * \overline{x} = \overline{o} - \beta_1 * \overline{x} * + \beta_1 * \overline{x} \qquad [17]$$

642 $$C(o, y) \to 0$$

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

References

Antolik, M.S. (2000). An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *Journal of Hydrology*, 239, 306- 337.

Barbero, R., Fowler, H. J., Blenkinsop, S., Westra, S., Moron, V., Lewis, E., et al. (2019). A synthesis of hourly and daily precipitation extremes in different climatic regions. *Weather and Climate Extremes*, 26, 100219. https://doi.org/10.1016/j.wace.2019.100219.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, *Hydrol. Earth Syst. Sci*., 21, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017.

‘

Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. *Geophysical Research Letters,* 47, e2020GL087408. https://doi.org/10.1029/2020GL087408

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

Carter, G. M., Dallavalle, J.P., & Glahn, H.R. (1989). Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401– 412.

685 Chalise, D. R., Sankarasubramanian, A., Olden, J. D., & Ruhi, A. (2023). Spectral signatures of
686 flow regime alteration by dams across the United States. *Earth's Future,* 11, e2022EF003078.
687 https://doi.org/10.1029/2022EF003078
688
689 Charba, J. P., and F. G. Samplatsky, 2011: High-Resolution GFS-Based MOS Quantitative
690 Precipitation Forecasts on a 4-km Grid. *Mon. Wea. Rev.,* 139, 39–68,
691 https://doi.org/10.1175/2010MWR3224.1.
692
693 Chen, M., W. Shi, P. Xie, V. B. S. Silva, V E. Kousky, R. Wayne Higgins, & J. E. Janowiak
694 (2008), Assessing objective techniques for gauge-based analyses of global daily precipitation, *J.*
695 *Geophys. Res*., 113, D04110, doi:10.1029/2007JD009132.
696
697 Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al.
698 (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources*
699 *Research*, 57, https://doi.org/10.1029/2020WR029001.
700
701 de Andrade, F.M., Coelho, C.A.S. & Cavalcanti, I.F.A. (2019). Global precipitation hindcast
702 quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Clim Dyn,*
703 52, 5451–5475. https://doi.org/10.1007/s00382-018-4457-z.
704
705 Entekhabi, D., and Coauthors (1999). An agenda for land-surface hydrology research and a call
706 for the second International Hydrological Decade. *Bull. Amer. Meteor*. Soc., 80, 2043–2058,
707 https://doi.org/10.1175/1520-0477(1999)080<2043:AAFLSH>2.0.CO;2.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor*., 11, 1203–1211.

Glahn, H. R., & Ruth D. P. (2003). The New Digital Forecast Database of the National Weather Service. *Bull. Amer. Meteor. Soc*., 84 , 195–201.

Goddard, L., Kumar, A., Solomon, A. et al. (2013). A verification framework for interannual-to-decadal predictions experiments. Clim Dyn, 40, 245–272. https://doi.org/10.1007/s00382-012-1481-2.

Gupta, H. V., Kling, H., Yilmaz, K. K., & G. F. Martinez (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.,* 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

Hamill, T.M. & Colucci, S.J., (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6), pp.1312-1327.

Ham, YG., Kim, JH. & Luo, JJ. (2019). Deep learning for multi-year ENSO forecasts. *Nature* 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7

729 Konapala, G., Mishra, A.K., Wada, Y. et al. (2020). Climate change will affect global water
730 availability through compounding changes in seasonal precipitation and evaporation. *Nat*
731 *Commun*, 11, 3044. https://doi.org/10.1038/s41467-020-16757-w
732
733 Krakauer, N.Y. (2019). Temperature trends and prediction skill in NMME seasonal forecasts.
734 Clim Dyn, 53, 7201–7213. https://doi.org/10.1007/s00382-017-3657-2
735 Li, Y., Tian, D., & Medina, H. (2021). Multimodel Subseasonal Precipitation Forecasts over the
736 Contiguous United States: Skill Assessment and Statistical Postprocessing. J. Hydrometeor., 22,
737 2581–2600, https://doi.org/10.1175/JHM-D-21-0029.1.
738
739 Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier,
740 D.P., Stouffer, R.J. (2008). Stationarity is dead: Whither water management?. *Science,* 319
741 (5863): 573-574. https://doi.org/10.1126/science.1151915
742
743 Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the
744 correlation coefficient. *Monthly Weather Review*, 116, 2417– 2424.
745
746 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I—
747 A discussion of principles. *Journal of Hydrology,* 10(3), 282–290. https://doi.org/10.1016/0022-
748 1694(70)90255-6
749
750 Petersen, T., Devineni, N., & A. Sankarasubramanian, A. (2012). Seasonality of monthly runoff
751 over the continental United States: Causality and relations to mean annual and mean monthly

distributions of moisture and energy, *J. Hydrol*., 468–469, 139–150, 10.1016/j.jhydrol.2012.08.028.

Quan, X., Hoerling, M., Whitaker, J., Bates, G., & T. Xu, T. (2006). Diagnosing Sources of U.S. Seasonal Forecast Skill. J. Climate, 19, 3279–3293, https://doi.org/10.1175/JCLI3789.1.

Sankarasubramanian, A., Sabo, J.L., Larson, K.L., Seo, S.B., Sinha, T., Bhowmik, R., Vidal, A.R., Kunkel, K., Mahinthakumar, G., Berglund, E.Z. & Kominoski, J. (2017), Synthesis of public water supply use in the United States: Spatio-temporal patterns and socio-economic controls. *Earth's Future*, 5: 771-788. https://doi.org/10.1002/2016EF000511

Sankarasubramanian, A., Lall, U., & Espinueva, S. (2008). Role of Retrospective Forecasts of GCMs Forced with Persisted SST Anomalies in Operational Streamflow Forecasts Development. *J. Hydrometeor*., 9, 212–227, https://doi.org/10.1175/2007JHM842.1.

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T., (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, 28(4), pp.616-640.

Scheuerer, M., & Hamill, T.M., (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Rev*.w, 143(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

775 Sun, L., Hoerling, M.P., Richter, J.H., Hoell, A., Kumar, A. & Hurrell, J.W., (2022). Attribution
776 of North American Subseasonal Precipitation Prediction Skill. *Weather and Forecasting*, 37(11),
777 pp.2069-2085.
778
779 Taillardat, M., Fougères, A., Naveau, P., & Mestre, O. (2019). Forest-Based and Semiparametric
780 Methods for the Postprocessing of Rainfall Ensemble Forecasting. *Wea. Forecasting*, 34, 617–
781 634, https://doi.org/10.1175/WAF-D-18-0149.1.
782
783 Vitart, F., Robertson, A. & Anderson, D. (2012). Sub-seasonal to Seasonal Prediction Project:
784 bridging the gap between weather and climate. WMO Bull. 61, 23–28.
785
786 Vitart, F., Robertson, A.W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the
787 prediction of extreme events. *npj Clim Atmos Sci* 1, 3. https://doi.org/10.1038/s41612-018-0013-
788 0
789
790 Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti,
791 E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P.
792 Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami,
793 R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A.W. Robertson, P. Ruti, C. Sun, Y.
794 Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. Won, H. Xiao, R.
795 Zaripov, and L. Zhang, (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database.
796 gi. *Amer. Meteor. Soc*., 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.
797

798 Vitart, F., Robertsn, A. W., & S2S Steering Group (2015). Sub-seasonal to seasonal prediction:
799 Linking weather and climate. In: Seamless Prediction of the Earth System: From Minutes to
800 Months. (pp. 385– 401). WMO-No.1156 (Chapter 20). Retrieved from
801 http://library.wmo.int/pmb_ged/wmo_11
802
803 Wang, L., & Robertson, A.W. (2019). Week 3–4 predictability over the United States assessed
804 from two operational ensemble prediction systems. *Clim Dyn,* 52, 5861–5875.
805 https://doi.org/10.1007/s00382-018-4484-9.
806
807 Weglarczyk S (1998), The interdependence and applicability of some statistical quality measures
808 for hydrological models. *Journal of Hydrology,* 206: 98-103. https://doi.org/10.1016/S0022-
809 1694(98)00094-8.
810
811 Weigel, A.P., Liniger, M.A., & Appenzeller, C. (2008), Can multi-model combination really
812 enhance the prediction skill of probabilistic ensemble forecasts?. *Q.J.R. Meteorol. Soc.,* 134:
813 241-260. https://doi.org/10.1002/qj.210
814
815 White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F.,
816 Coughlan de Perez, E., Ray, A.J., Murray, V., & Bharwani, S., 2017. Potential applications of
817 subseasonal-to-seasonal (S2S) predictions. *Meteorological applications*, 24(3), pp.315-325.
818
819 Wilks, D. Statistical Methods in the Atmospheric Sciences (Academic, 2006).
820

821 Wood, A. W., Maurer, E.P., A. Kumar, & Lettenmaier, D.P. (2002). Long-range experimental
822 hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* 107, 4429, doi:
823 10.1029/2001JD000659.
824
825 Zhang, C. (2013). Madden–Julian Oscillation: Bridging weather and climate. *Bulletin of the*
826 *American Meteorological Society*, 94, 1849–1870. https://doi.org/10.1175/BAMS-D-12-00026.1
827
828 Zhang, L., Kim, T., Yang, T., Hong, Y. & Zhu, Q. (2021). Evaluation of Subseasonal-to-
829 Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II
830 (NMME-2) over the contiguous US. *Journal of Hydrology*, 603, p.127058.
831 https://doi.org/10.1016/j.jhydrol.2021.127058
832