A Framework for Diagnosing Seasonal Prediction through Canonical Event Analysis

JOSHUA K. ROUNDY* AND XING YUAN

Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

JOHN SCHAAKE

Annapolis, Maryland

ERIC F. WOOD

Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

(Manuscript received 6 June 2014, in final form 2 January 2015)

ABSTRACT

Hydrologic extremes in the form of flood and drought have large impacts on society that can be reduced through preparations made possible by seasonal prediction. However, the skill of seasonal predictions from global climate models is uncertain, which severely limits their practical use. In the past, the skill assessment has been limited to a single temporal or spatial resolution for a short hindcast period, which is prone to sampling errors, and noise that leads to uncertainty. In this work a framework that uses ''canonical'' forecast events, or averages in space–time, to provide a more certain assessment of when and where models are skillful is developed. This framework is demonstrated by using NCEP's Climate Forecast System, version 2, hindcast dataset for precipitation and temperature over the contiguous United States (CONUS). As part of the canonical event analyses, the probabilistic predictability metric (PPM) is used to define spatial and seasonal variability of forecast skill and its attribution to El Niño–Southern Oscillation (ENSO) over the CONUS. The PPM indicates that there are clear seasonal and spatial patterns of model skill that provide a better understanding of when and where to have confidence in model predictions as compared to a skill metric based on a single temporal and spatial scale. Furthermore, the canonical event analysis also facilitates the attribution of spatiotemporal variations of precipitation predictive skill to the antecedent ENSO conditions. This work illustrates the importance of using canonical event analysis to diagnose seasonal predictions and discusses its extensions for model development.

1. Introduction

Extreme hydrologic events in the form of drought and flood have a large impact on society and can invoke large amounts of economical and social damage. In recent years there have been several of these events such as the drought of 2012, which greatly affected the primary corn and soybean regions of the United States (Karl et al. 2012), and a number of heavy rainfall events that caused record flooding throughout large segments of Iowa in 2008 (Smith et al. 2013). The ability to have a warning of such extremes would reduce the impact on society through preparation. One such preparation is the ability of water managers to make decisions to ensure sufficient water supply during droughts and sufficient reservoir storage for flood events. To achieve the intended benefit, these preparations require sufficient warning on a seasonal time scale. Because of the chaotic nature of the atmosphere, individual storms that cause hydrologic extremes usually cannot be predicted at seasonal time scales; however, some extremes are the result of a seasonal persistence into a wet or dry state. The 2008 flooding and the 2012 drought are examples of such extremes caused by a seasonal persistence (Smith et al. 2013; Karl et al. 2012) and not a single event. There have been several studies that have shown the potential of

* Current affiliation: Hydrological Services Branch, NASA GSFC, Greenbelt, Maryland.

Corresponding author address: Joshua K. Roundy, Goddard Space Flight Center, Building 33, Room G209, Greenbelt, MD 20771.
E-mail: joshua.roundy@nasa.gov

DOI: 10.1175/MWR-D-14-00190.1

© 2015 American Meteorological Society

seasonal predictability through a connection of large-scale climate patterns and land–atmosphere interactions (Trenberth and Guillemot 1996; Koster et al. 2010; Villarini et al. 2013; Goddard et al. 2001; Palmer and Anderson 1994).

The prospect of knowing the seasonal outlook of hydrologic drivers, such as precipitation and temperature, to aid in reservoir management has motivated water managers and forecasters to develop prediction models. Historically these models were statistically based and incorporated a large array of predictors based on observations (Goddard et al. 2001). As computational resources and our understanding of the climate have improved, the use of physically based general circulation models (GCMs) to make seasonal predictions has become available. Using the forecasts from GCMs is not always straightforward because of the coarse resolution and inherent model biases associated with their predictions. Even with these limitations, there have been many examples of using GCMs for seasonal hydrologic prediction (Yuan and Wood 2012; Luo et al. 2007; Wood et al. 2002) that have shown prediction skill for some areas and events (Luo and Wood 2007). There have been many studies that have compared the use of statistical methods against GCMs for hydrologic forecasts (Yuan et al. 2013), but the skill of the GCMs vary with space and time, and the skill of hydrologic forecast is greatly dependent on the precipitation forecasts. Furthermore, others argue that statistically based models are a better predictor when the GCM skill is low, but when the GCM precipitation forecasts is skillful there is clear advantage over statistical models (Mo et al. 2012). As a result the GCM forecasts are often underutilized, as it requires an understanding of when and where the models are skillful to achieve any benefit.

Determining when and where the GCM is skillful requires an assessment of the past performance of the model or its hindcasts. Because of data and computation limitations, current hindcasts from GCMs are usually less than 30 years in length. With such short sample sizes, the analysis of the skill can be heavily influenced by sampling error and the signal-to-noise ratio, which make the assessment of predictive skill uncertain (Kumar 2009). The uncertainty in predictive skill limits the practical use of GCMs, as users are unsure when and where to have confidence in GCM forecasts. One way of increasing the signal-to-noise ratio is to use temporal and spatial averaging, which can reduce the noise. This was shown by Luo and Wood (2006) who demonstrated that an anomaly is more predictable at larger temporal and spatial scales. Although the connection between forecast skill and spatiotemporal scale is well known, for simplicity, most evaluations of forecast skill only consider a single temporal and spatial scale and do not take into account the consistent skill of the forecasts in time and space. This remains largely an untapped resource for understanding model predictability. The aim of this study is to develop a framework that accounts for the consistency of the predictive skill across temporal and spatial averaging called ''canonical'' event analysis. As part of this analysis a simple metric that accounts for consistent forecast skill in time and space is developed to provide clarity as to when and where to have confidence in GCM forecasts. This framework is demonstrated by using the hindcasts from the NCEP Climate Forecast System, version 2 (CFSv2; Saha et al. 2014; Yuan et al. 2011), model over the contiguous United States (CONUS). Furthermore, this framework is used to attribute the influence of El Niño–Southern Oscillation (ENSO) on the predictive skill of CFSv2. The implications, limitations, and extensions of this work are also discussed.

2. Methodology

The fundamental hypothesis of this study is that seasons and locations where GCMs produce usable seasonal forecasts can be identified by consistent skill across temporal scales, spatial scales, and with lead times. This is done through temporal and spatial averaging of the forecasts to analyze the skill across multiple events (i.e., a single event is a specific time–space average). These events, which are simply time–space averages are called canonical events. The word canonical is based on the Merriam–Webster definition of ''canonical form'' meaning ''the simplest form of something'' (http://www.merriam-webster.com/dictionary/canonical) and refers to that fact these time–space averages represent the breakdown of the overall predictability into a series of simple events. To analyze the skill of the individual canonical events the Spearman rank correlation is used for its robustness in capturing nonlinear correlation structures regardless of biases. This correlation measure is calculated between the ensemble mean of the forecast and the validation dataset across all canonical events. Then, the correlation for individual canonical events is used to produce a probabilistic measure of predictability that captures the consistency of skill in time and space. The forecast model and the formation of its ensemble mean, the validation dataset, the temporal and spatial averaging that make up the canonical events, and the formation of probabilistic predictability metric are discussed below.

The forecast model used in this analysis is CFSv2, which has a 28-yr hindcast period from 1982 to 2009, called the Climate Forecast System Reanalysis and

TABLE 1. The number of ensembles and validation time for each forecast initialization in the CFSRR hindcast dataset.

Forecast initialization	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
No. of ensembles	28	20	23	24	28	24	24	23	24	21	18	18
Validation	Feb–Sep	Mar–Oct	Apr–Nov	May–Dec	Jun–Jan	Jul–Feb	Aug–Mar	Sep–Apr	Oct–May	Nov–Jun	Dec–Jul	Jan–Aug

Reforecast (CFSRR; Saha et al. 2014). During this reforecast period, 9-month seasonal hindcasts were made four times daily at 0000, 0600, 1200, and 1800 for every 5 days starting on 1 January of each year. CFSRR has a spatial resolution of T126 (~0.9375°), and includes the archival of certain variables at the 6-hourly temporal resolution. The forecast variables used in this study are those needed for hydrologic predictions, namely daily precipitation and daily maximum and minimum 2-m temperature. The daily values of these forcing variables are temporally upscaled from the 6-hourly data. The dataset that is used to validate the forecasts is from the North American Land Data Assimilation System, version 2 (NLDAS-2; Xia et al. 2012). The NLDAS-2 dataset incorporates gauge, radar, and reanalysis datasets to provide an observational based dataset over the CONUS that is routinely used for forcing land surface hydrologic models. The NLDAS-2 dataset has an hourly temporal resolution and a 0.125° spatial resolution. To be consistent with the forecasts from CFSv2, the NLDAS-2 dataset is temporally upscaled to a daily resolution and spatially upscaled to the forecast model resolution (~0.9375°) using bin averaging.

The Spearman rank correlation is used to assess the forecast skill of individual canonical events. Calculating the Spearman correlation between the observations (NLDAS-2) and the forecasts requires a single measure of prediction from the CFSRR ensembles given by the ensemble mean. The forecast ensemble that forms the ensemble mean is determined by the initialization time of the forecasts in order to preserve consistent forecast lengths. For example, all forecasts initialized in January run through the end of September. Likewise, all forecasts initialized in February run through October. Although there is a temporal overlap between the ensembles initialized in January and February, the combination of forecast ensembles by initialization month, with the validation beginning the following month, maintains ensemble consistency. However, because of the 5-day schedule of the hindcasts, the number of ensemble members initialized in each calendar month varies. In addition, as a result of missing data files, the number of ensemble members used in this analysis for each initialization month varies. The number of ensemble members used for each month along with its validation period is given in Table 1. According to the 5-day hindcast schedule, January and May have the largest number of ensemble members (28) and February has the fewest (20). Although November and December should each consist of 24 ensemble members, only 18 are used because of missing data files. Likewise, March, August, and October also have fewer ensemble members than the 5-day schedule dictates due to missing data files. The ensemble mean for each initialization is used to calculate the Spearman correlation across canonical events.

For each forecasts initialization, the canonical events consist of 52 different temporal averages that cover 9 different temporal scales. These temporal averages are shown in Fig. 1a, using the January initialization as an example. They range from 15 days to 8 months. The zero temporal scale is associated with submonthly temporal averages, namely the first 15 days and day 16 to the end of the months (shown in Fig. 1a by the number of days after the 15th). The following temporal scales are analogous to the number of months (i.e., the first temporal scale is for a temporal average of 1-month and the eighth temporal scale is for an 8-month temporal average). For each of the different 52 temporal averages there are 9 different spatial averages (see Fig. 1b) that also range from a spatial scale of 0 to 8. The spatial scale is associated with the radius of grid cells used in the spatial averaging of the variable. Therefore, the zero scale is at the native resolution of the CFSv2 and consists of no spatial averaging. The first spatial scale is based on a one-grid radius of averaging and includes the center grid cell along with the surrounding eight grid cells (3×3). This pattern continues up to the eighth spatial scale, which consists of the center grid, and the surrounding 288 grid scales (17×17). The extents of the nine spatial scales are represented in Fig. 1b for a grid in Texas (32.6°N, 96.7°W). Spatial scales that fall outside of the CONUS landmass (i.e., coastlines or borders) will not be used in the spatial averaging. As a result, cells close to the boundaries will incorporate fewer grids in the spatial averaging than fully land-locked grid cells.

For a single grid cell there is a total of 5616 canonical events at 9 different spatial averages and 624 temporal averages (52×12 initialization months). Although there

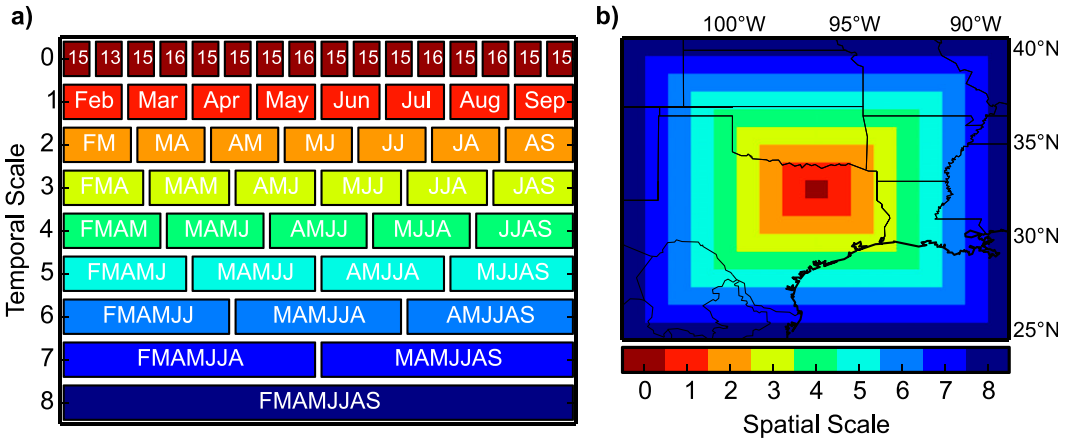

FIG. 1. Diagrams illustrating (a) the temporal scales (shown for forecasts initiated in January) and (b) the spatial scales (shown for a single grid in Texas, 32.6°N, 96.7°W) used in this study.

are 624 temporal averages throughout the calendar year, only 120 of them are unique temporal events (10 per month). For example, all time averages that start in February consists of the first 15 days of February, the last 13 days of February, the whole month of February, February through March, February through April, February through May, February through June, February through July, February through August, and February through September. Forecasts initialized from June to January predict a range of these 10 events that start in February, but vary in terms of the lead time of the forecast. For instance, predictions initialized in January are made at 0-month lead time and consist of all 10 events starting in February. In comparison forecasts made in June are done at a 7-month lead time and only consist of the first 15 days of February, the last 13 days of February and the whole month of February. Therefore, the redundant predictions of the 120 unique temporal events make up the third dimension of the canonical events, the forecast lead time. The validation of the forecasts is organized by the month the forecast begins to provide a consistent seasonal evaluation across all the canonical events.

The fundamental hypothesis of this framework is that usable forecast skill is consistent across canonical events. The correlation across canonical events for a single grid cell (same as the one depicted in Fig. 1b) is given in Fig. 2 by a matrix of correlations, where the x axis is the spatial scale and y axis is temporal scale (for the zero scale only the first 15 days are shown), for each of the eight lead times for precipitation forecasts that begin in April and October. The number of temporal events decreases as the lead time increases so that only the fine temporal scales are predicted at the longest lead times, as discussed above. For both months and lead times the skill is the greatest at the higher spatial and temporal scales and illustrates the utility of temporal and spatial averaging in order to reduce the forecast noise. This is similar to the result shown by Luo and Wood (2006). Although the skill is the highest at the larger temporal and spatial scales, there is a consistent decrease in skill as the lead time increases. In particular, the skill is low for all seasons at the longer (4–7 month) lead times. This is due to the separation between the initial conditions and target forecast (Chen et al. 2013). Furthermore, because of the fixed length of the forecasts, the temporal averaging is limited to smaller temporal scales as the lead time increases. Although the skill is low for the longer lead times, there is consistency in the skill across lead times for each of the months. For instance, forecasts that begin in April show low skill across most spatial and temporal scales and with lead time. In contrasts, forecasts that begin in October show consistently higher skill across all lead times, with the exception of the longest lead time. This demonstrates that there is consistent forecast skill across canonical events in October, while April shows little or inconsistent skill.

The results in Fig. 2 motivate the development of a measure of forecast skill that considers the consistency across canonical events. This can be simply summarized by using the conditional probability of significant correlation between the forecast and observations at a given confidence level across canonical events. This metric is defined as the probabilistic predictability metric (PPM) and is calculated by

$$\mathrm{PPM} = \frac{N_s}{N_t},$$

where N_s is the number of events with statistically significant correlation at the 95% confidence level and N_t is the total number of events. Therefore, the PPM is the

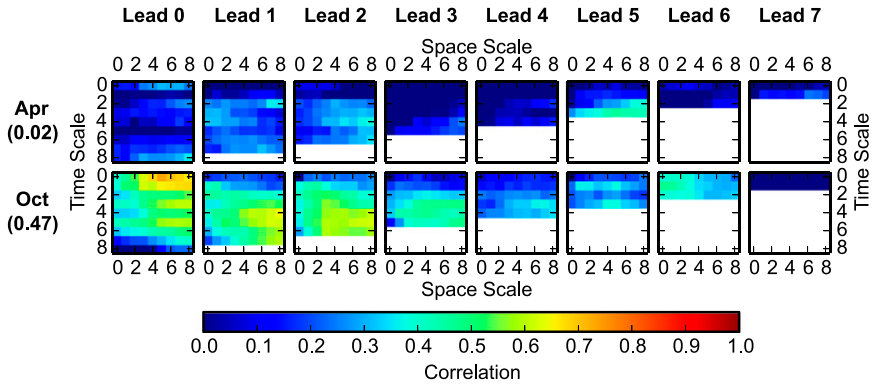

FIG. 2. Changes in the Spearman correlation across temporal and spatial scales at various lead times for the CFSv2 precipitation forecasts with validation beginning in April and October for a grid cell in east Texas (32.6°N, 96.7°W). The PPM (see text for details) is given in parentheses below each month.

probability of a forecast being statistically significant for a chosen range of canonical events, or simply stated, is the fraction of the number of forecasts for which the predictive skill is statistically significant. For example, the PPM for April and October can by calculated by considering all of the canonical events shown in Fig. 2. For April, of the 396 canonical events only 9 have statistically significant correlations at the 95% confidence level, which gives a PPM of 0.02. In contrast, in October there are 187 events that have statistically significant correlation, which gives a PPM of 0.47. The difference in the PPM of these two months captures the difference in the consistency of the skill across all canonical events and is, therefore, a more robust measure of forecast skill compared to a metric at a single space–time lead. Furthermore, since a 95% significance level is used to determine statistical significance, a PPM around 0.05 is within the margin of error for correctly defining statistical significance. This argues that the few statistically significant events in April are likely due to random variation and not due to the skill of the model. In contrast, the PPM in October is much higher than 0.05, so it is likely that the high skill in October is due to the predictive signal of the model. Although the PPM for forecasts that begin in October is much larger, it is still less than 0.5. From a probabilistic point of view, this suggests that it is more likely that a given forecast is not skillful. This is due to the long lead-time events, which have low correlations in October, but have the same weights as shorter lead times on the PPM. This illustrates that a PPM around 0.5 indicates a strong predictability for short lead-time events, but no predictability for long lead-time events. Including the longer lead-time events as part of the PPM provides a measure of the overall predictive skill across all lead times, but also creates ambiguity in the exact events that are contributing to the PPM. In this study, the focus is on the overall predictive skill of the model and will, therefore, use all spatial and temporal scales and lead times in calculating the PPM. Although this study focuses on the predictive skill across all canonical events, the PPM could be adapted to only consider a small range of temporal and spatial scales and lead time in order to provide a more precise metric.

The consistency in prediction captured by the PPM is due to the predictability of particular temporal events regardless of lead time. Reorganizing the correlations in Fig. 2 by forecast initialization and forecast period for selected temporal and spatial events demonstrates this. The forecasts of 3- and 4-month temporal average at ~1° (grid resolution) and ~5° spatial averaging are given in Fig. 3 along with the PPM. The correlation plots show a distinct diagonal structure of higher correlation and white areas (correlation <0). The diagonal structure indicates consistent predictability of events regardless of lead time. For instance, the 3-month temporal average indicates that the 1–3-month forecast initiated in September has a similar correlation as the 2–4-month forecast initialized in August and the 3–5-month forecast initialized in July, all of which are predicting October–December. For this location, it seems that the predictability is more dependent on the seasonality than the forecast leads. In addition, the four-month temporal average shows a similar diagonal correlation structure, although it is slightly diminished. Furthermore, both the 3- and 4-month forecasts show an enhancement of the correlation with a higher spatial averaging. In comparison, the PPM provides a summary of the consistent skill across the diagonal of these correlation plots for all nine spatial and temporal averages and clearly indicates that for the grid cell in east Texas the forecasts that start in the fall are most skillful, particularly for October. This

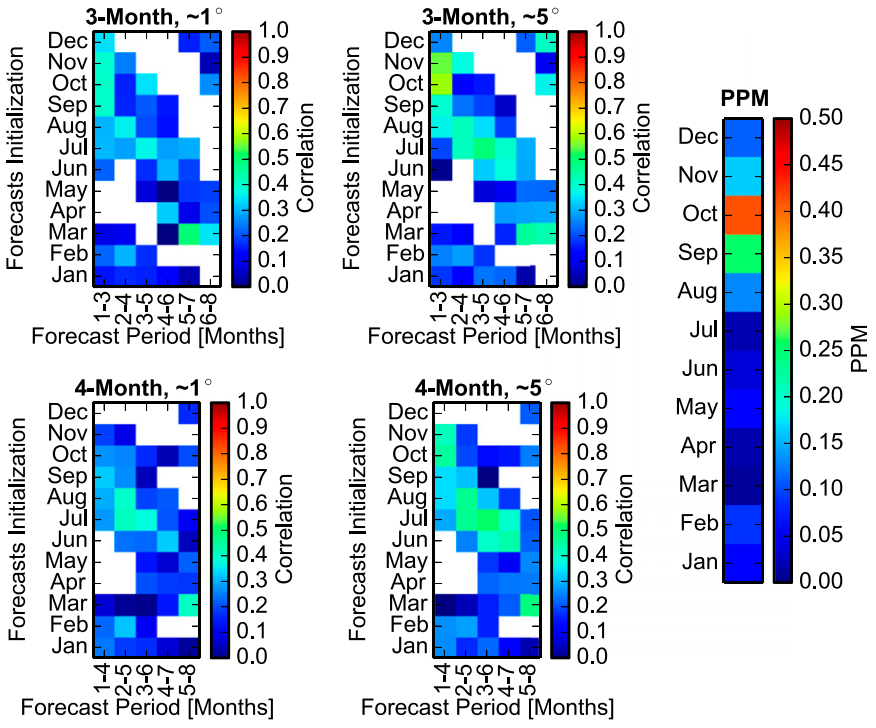

FIG. 3. The precipitation Spearman correlation with forecast initialization and forecast period for 3- and 4-month forecasts at the zeroth (~1°) and second (~5°) spatial resolution compared to the PPM for a grid cell in east Texas (32.6°N, 96.7°W). The white areas indicate correlations less than zero.

demonstrates the utility of the PPM for understanding when and where there is useable skill in the forecasts. As such, the remainder of this paper uses the PPM to understand the importance of temporal and spatial averaging and the seasonal and spatial variability of the forecast skill and its attribution.

3. Results: Seasonal and spatial variation of predictive skill

The PPM captures the consistent skill across canonical events, or in other words the consistent skill across temporal and spatial averaging. The temporal and spatial averaging can reduce the noise thus increasing the signal-to-noise ratio, but can also reduce the signal if the averaging in space and time extends past an isolated predictive event. If the rate of reducing the signal is greater than or equal to the noise reduction then there will be no increase in the predictive skill. Therefore, spatial and temporal averaging does not always result in an increase in skill. The PPM provides a means to understand the seasonal variability of temporal averaging, spatial averaging, and lead time on forecasts skill by individually isolating each component. The PPM conditioned on temporal averaging (time) gives an indication of the importance of temporal averaging and is calculated by considering all spatial scales and lead times for each temporal scale. Likewise, the importance of spatial averaging (space) is assessed over temporal scales and lead times at each spatial scale. Similarly the importance of lead time (lead) is analyzed over all spatial and temporal scales conditioned on lead time. In this manner the average PPM over the CONUS is shown in Fig. 4 for all seasons for precipitation, maximum daily temperature, and minimum daily temperature. For the precipitation (Fig. 4a) the PPM increases as the temporal scale increases. The largest increase occurs during the winter and spring months and the smallest increase during the August–October season. The PPM also increases with spatial scale for all months, with the highest increase during the summer months and a very small increase during the winter months. In terms of lead time, the zero lead time shows the highest PPM, but also the greatest seasonal variability. The seasonal variability of PPM decreases as the lead time increases, with very little difference in the PPM for the longer lead times. Relative to each other, time averaging of precipitation forecasts provides the greatest increase in skill over the CONUS across all seasons and is followed by lead time. The spatial averaging provides the lowest relative increase in

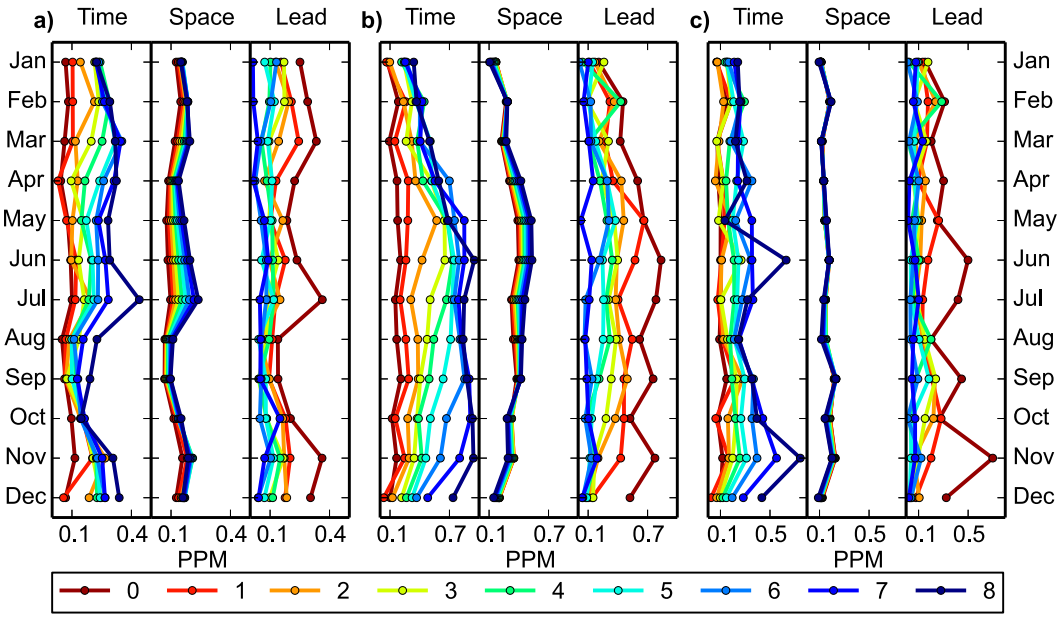

FIG. 4. The seasonal change in the average PPM over the CONUS across temporal scales (time), spatial scales (space), and lead time (lead) of the forecast for (a) precipitation, (b) daily maximum temperature, and (c) daily minimum temperature.

precipitation forecast skill, but also has the clearest seasonal cycle.

The daily maximum temperature (Fig. 4b) shows similar characteristics in that the PPM increases with temporal averaging for all months. However, the overall skill is much higher than precipitation, and the temporal averaging shows the lowest increase in skill in the winter and the highest in the summer. In terms of spatial averaging, the daily maximum temperature also shows an increase in skill during the spring and summer; however, spatial averaging during the winter months slightly reduces the skill. The lead time of the forecast also follows similar patterns to precipitation except that there is larger differentiation at longer lead times. In terms of the daily minimum temperature (Fig. 4c), temporal averaging also increases the skill for all seasons, but is much smaller in magnitude than daily maximum temperature. Furthermore, the spatial averaging of the daily minimum temperature shows no improvement in skill even during the summer months. This suggests that any spatial signal of predictability is much weaker for daily minimum temperature during the summer months as compared to the daily maximum temperature and precipitation. The daily minimum temperature has similar characteristics to that of precipitation and daily maximum temperature with respect to lead time, in that the skill and seasonal variability are the highest for the zero lead both of which decrease as lead time increases.

The CONUS average seasonal response to forecast skill using the PPM shows distinct seasonal characteristics, but it does not consider the spatial variability of the skill. The spatial patterns of the forecast skill using a single canonical event are very uncertain, however, using the PPM to quantify forecast skill can reduce the uncertainty. This is illustrated in Fig. 5, which shows seasonal spatial maps of the CONUS for precipitation in terms of the correlation of a 3-month temporal average at the grid scale and zero lead (Fig. 5a) along with the PPM (Fig. 5b). Although the 3-month temporal average at the grid resolution at zero lead has general consistency with the PPM, the skillful areas are somewhat erratic and there is a lot of noise (Fig. 5a). Accounting for the significance level of the correlation can reduce some of this noise, but it still differs from the PPM in the details. Furthermore, there is little interseasonal consistency and the skillful areas appear to jump from location to location throughout the year. In contrast the PPM shows strong spatial patterns of skill and seasonal evolution of that skill (Fig. 5b). Specifically, January and February have high skill in the U.S. Southwest and Southeast that dissipates and moves to the West Coast in March and becomes virtually nonexistent by April. Then in May, the skill in the northern intermountain area (Idaho, northern Utah, western Wyoming) begins to develop and strengths and spreads to the north-central Great Plains into June and July. The summertime skill then dissipates in August and September. The

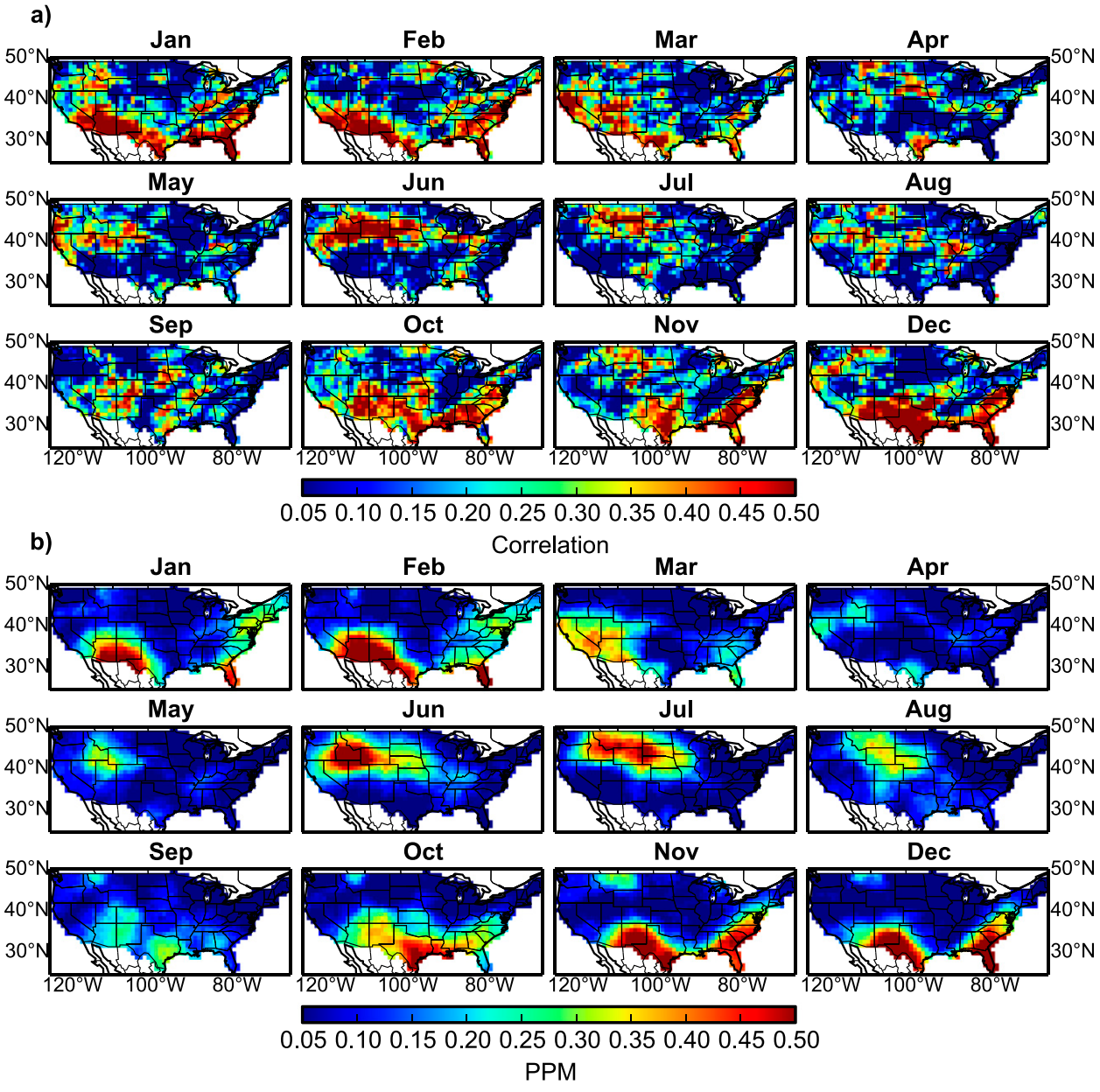

FIG. 5. The (a) Spearman correlation of the 3-month forecasts at the grid scale and 0-month lead time (i.e., Mar is the Mar–May forecast initialized in Feb) and (b) the PPM for forecast validation beginning at each month over the CONUS.

skill in the South reappears in October and moves to the Southwest and Southeast of the CONUS for November and December, similar to January. From this it can be seen that the precipitation forecasts in CFSv2 have two distinct seasons of skill: November–February in the Southwest and Southeast, and June–August in the north-central United States. March–May and August–October appear to be transitional months with less predictive skill. The difference in the skill between the correlation (Fig. 5a) and the PPM (Fig. 5b) is due to considering all the canonical events that brings out the spatial and seasonal patterns.

Clear spatial and seasonal patterns for daily maximum and minimum temperature are also found by using the PPM (Figs. 6a and 6b), although they are different from the patterns for precipitation. For the daily maximum temperature the skillful area in January is confined to the southernmost part of the CONUS, but expands to the northeast from February to May. The skill remains low for the Northwest. The June–August season shows

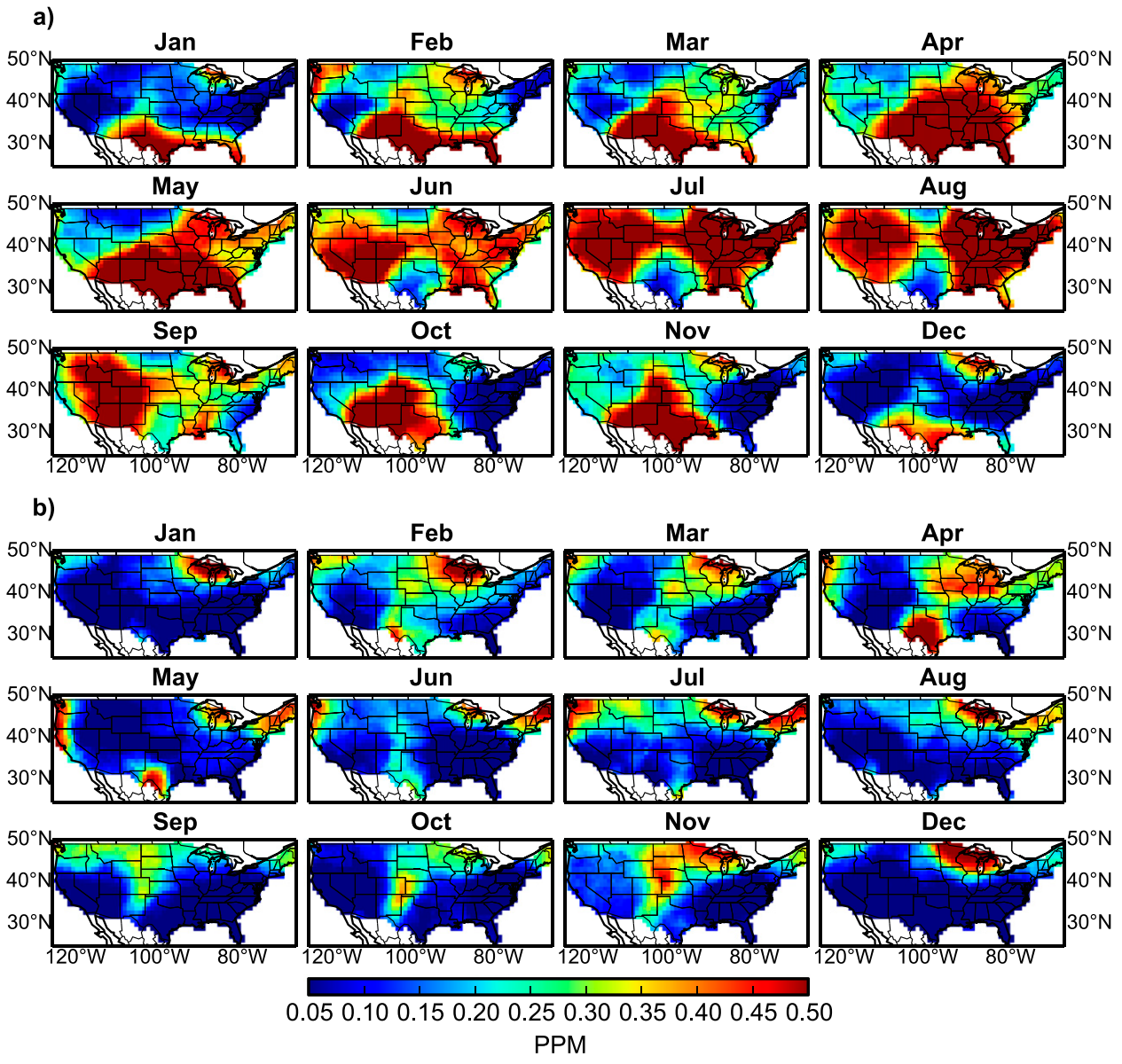

FIG. 6. The PPM for forecast validation beginning at each month over the CONUS for (a) daily maximum temperature and (b) daily minimum temperature.

an increase in skill in the Northwest, but a decline in the skill in the central United States to the point that by August the east and west are completely separated by an area of low predictability. The predictive skill persists for the west in September but decreases in the east, and continues to decrease in both areas in October and November to the south-central United States. By December the predictive skill is primarily confined again to the southern United States and to a small portion of the Great Lakes similar to January. Overall the daily maximum temperature shows a greater extent of predictive skill over the CONUS and different spatial and seasonal characteristics compared to precipitation. In comparison, the daily minimum temperature (Fig. 6b) shows the lowest skill among the variables. For daily minimum temperature, the predictability in January is confined to the Great Lakes area, but develops in Texas and the West Coast by April. The predictive skill then decreases from May to August with the northern West Coast and the Great Lakes area still showing some predictability, but by September–October the predictive skill is low for the entire CONUS. In November the skill increase again across the north-central and Great Lakes area, but is again confined to the Great Lakes by

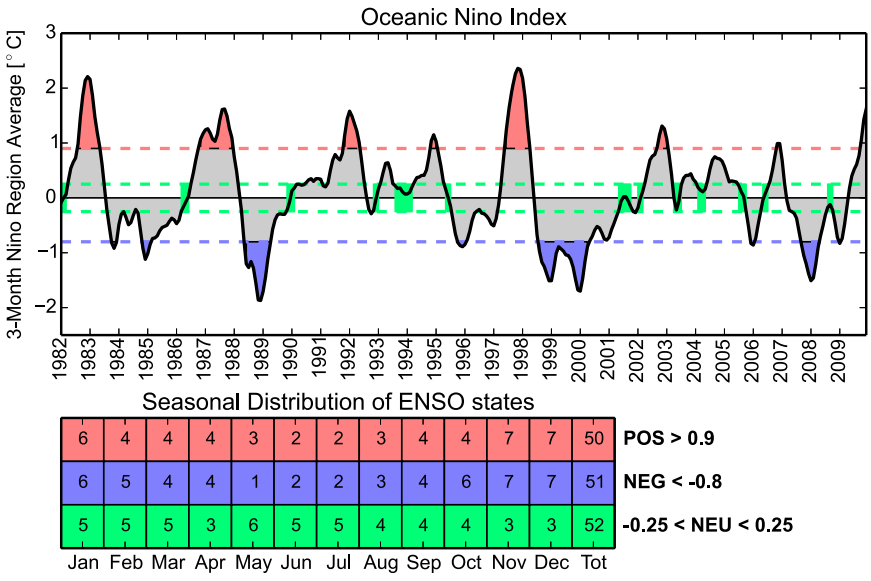

FIG. 7. The (top) oceanic Niño index (ONI) used to define the ENSO state into positive (POS), negative (NEG), and neutral (NEU) conditions and (bottom) the seasonal distribution of each category over the 1982–2009 hindcast period.

December. Although there is some consistency between the predictive skill of the daily maximum temperature and the daily minimum temperature (Great Lakes region), the spatial and seasonal patterns yet remain very different for these two temperature variables.

4. Results: Attribution of forecast skill

The above analysis shows that consistent forecast skill in CFSv2 varies with season and location. In particular, the precipitation forecasts showed two seasons of predictive skill: November–February and June–August. Attributing the skill in these seasons will provide additional information to decision-makers and enable further use of the forecasts. In this section the utility for using the PPM to provide insights to the attribution of the ENSO state at forecast initialization on the predictive skill is demonstrated. ENSO is used because it is considered as one of the strongest climate drivers at seasonal-to-interannual scales (Trenberth et al. 1998). The ENSO state is defined by using the oceanic Niño index (ONI) from the Climate Prediction Center (http://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt), and is used to define three states of ENSO: positive (POS), negative (NEG), and neutral (NEU). The typical threshold values used by the Climate Prediction Center for defining positive and negative ENSO is ±0.5. In this work, the classification is based on at least three consecutive months of ONI greater than 0.9, less than −0.8, and between −0.25 and 0.25 for positive, negative, and neutral, respectively. These adjusted threshold values isolate the strongest ENSO events and allow for a similar number of occurrences for each ENSO state. The ONI from 1982 to 2009 is shown in Fig. 7 and indicates periods of positive, negative and neutral ENSO. Furthermore, the number of months classified as positive, negative, and neutral ENSO is also given. Based on the defined thresholds, there are 50, 51, and 52 months classified into positive, negative, and neutral ENSO over the time period. The sample size of each ENSO state equates to roughly 4 of the 28 years in the hindcast and represents a level of sampling uncertainty that could be overcome with a larger hindcast dataset. These months are well distributed across all seasons for the neutral state, but there are fewer positive and negative ENSO states from May to August. For each ENSO state, all forecasts, which are initiated in a month classified by that state, are removed from the calculation of the Spearman correlations across all the canonical events used for calculating the PPM. This gives three different estimates of the PPM and can be compared to the PPM based on the full record in order to determine the impact of the ENSO state at initialization on the predictive skill. To demonstrate the utility of the PPM, the impact of the ENSO state at initialization is also attributed using the Spearman correlation at a single temporal and spatial average (3 month at the grid resolution). The attribution of the ENSO state on the predictive skill requires identifying statistically significant differences between the metrics based on the full

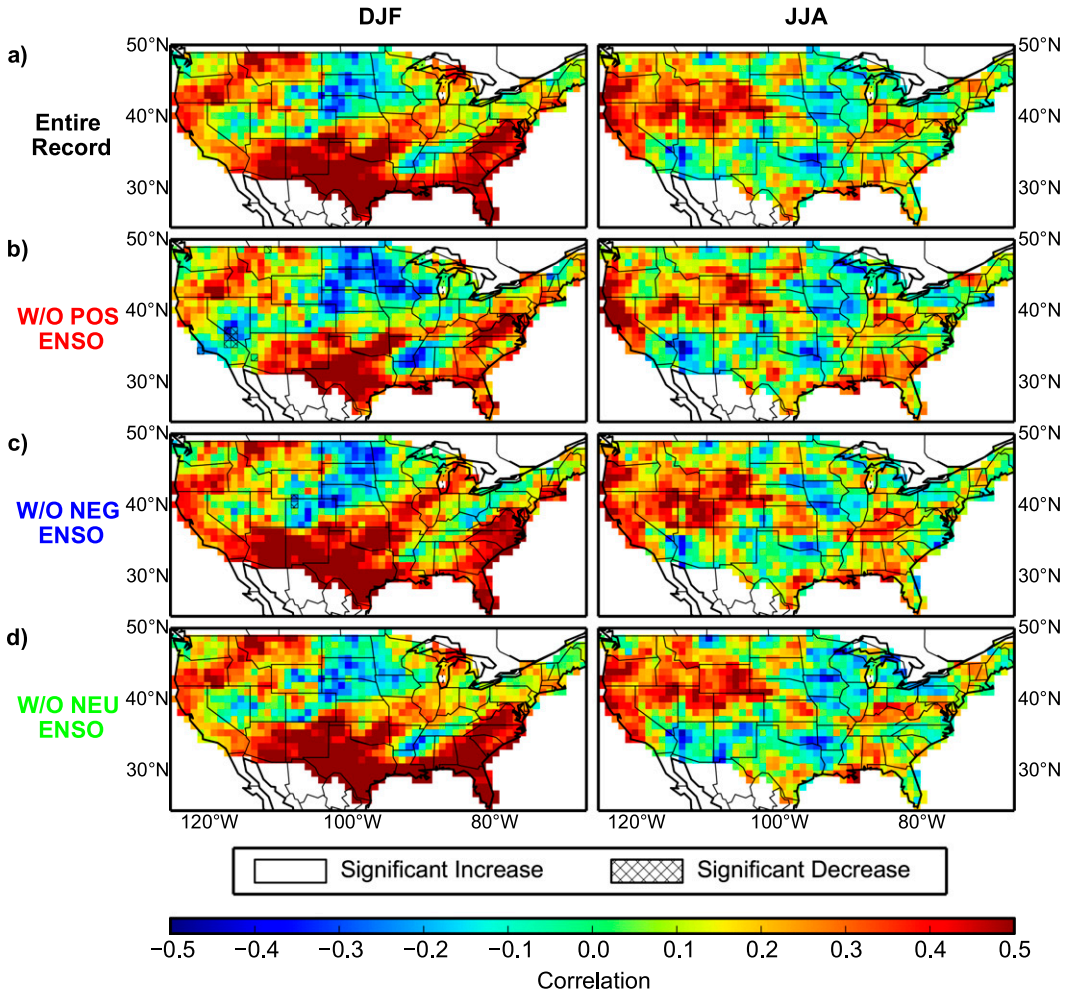

FIG. 8. The Spearman correlation for precipitation using (a) the entire record and without the forecasts initialized during (b) POS, (c) NEG, and (d) NEU ENSO conditions for DJF and JJA with hatching indicating statistically significant differences from the entire record.

record with those based on a subset. The statistical significance of the difference between the Spearman correlations is estimated by using the Fisher transformation (Fieller et al. 1957) to define 95% confidence intervals for the correlation based on the full record. The correlation based on the subset is considered to be significantly different if it lies outside of the 95% confidence intervals. The same approach is also used for the PPM, except that the confidence intervals are estimated from a Poisson distribution, since the PPM is simply a count of statistically significant Spearman correlations.

The 3-month average correlation and the PPM for precipitation with the whole record and removing forecasts initiated during a positive, negative, and neutral ENSO state are compared for the two predictive seasons shown in Fig. 5. Although the PPM indicates that the winter predictive season goes from November to February, only December–January–February (DJF) are used in order to produce a 3-month period similar to the summer season of June–July–August (JJA). The results for the 3-month temporal average Spearman correlation and the PPM are given in Figs. 8 and 9, respectively. In comparing the correlation in Fig. 8, it can be seen for this specific temporal and spatial average, there is little difference between the entire record and composite results based on the ENSO condition. Specifically there are only 11 grid cells that have statistically significant differences from the correlation given by the entire record. Nine of those grid cells show a statistically significant decrease in the correlation during DJF when removing the forecasts made from a positive ENSO state. The largest concentration of these grid cells (6) is located in southern Nevada. There are also grid cells in

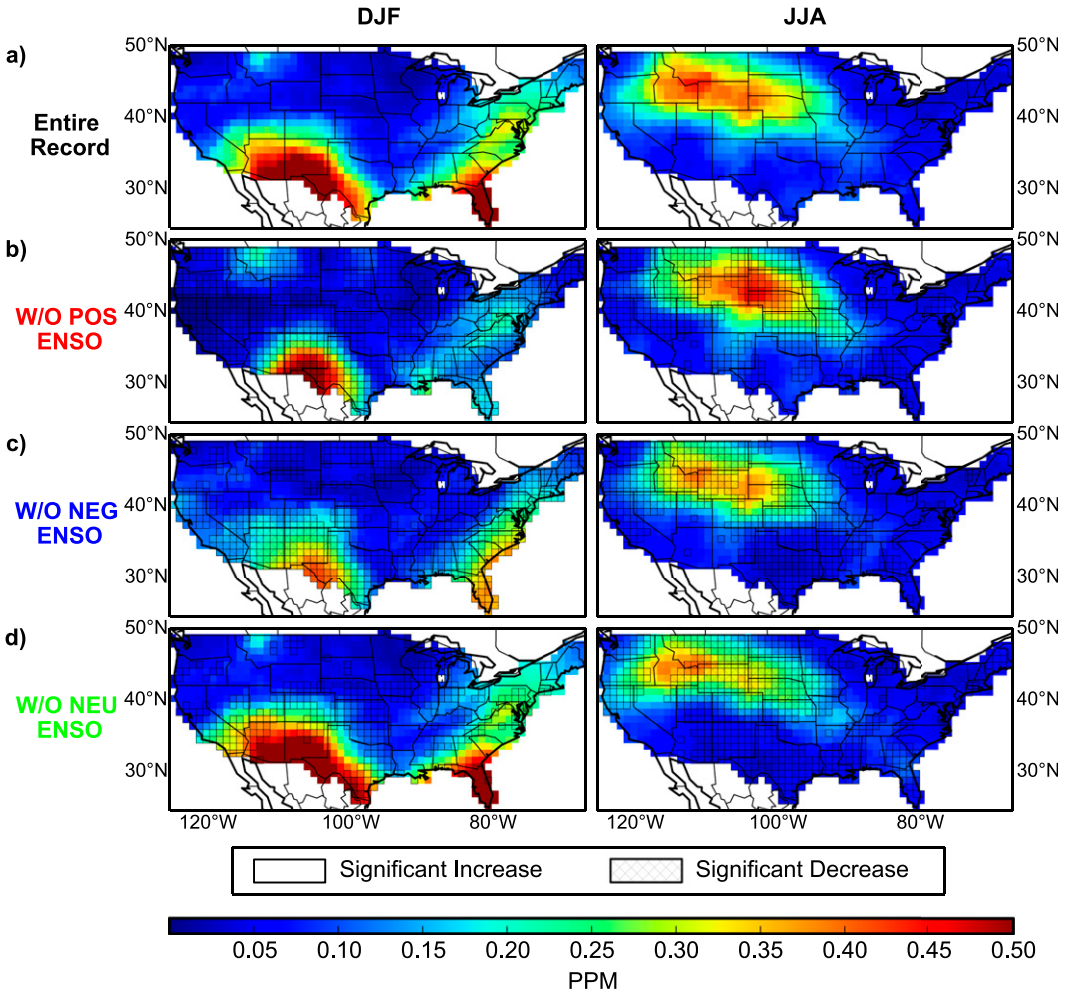

FIG. 9. As in Fig. 8, but for the PPM for precipitation.

northern Montana, western Arizona, and the on the southern coast of California that also show a significant decrease in the correlation of DJF from removing the forecasts initialized in a positive ENSO. There are also two grid cells in northern Colorado that show a significant decrease in the DJF correlation due to removing the forecasts initiated from a negative ENSO. In contrast, removing forecasts made in neutral ENSO resulted in no significant differences for DJF correlations. Furthermore, for JJA, none of the ENSO states resulted in a significant difference from the entire record. This shows that the impact of ENSO at forecast initialization is not picked up by examining the Spearman correlation for a single temporal and spatial average. This is likely due to the fact that there is much uncertainty in the correlation due to the small sample size.

In comparison, Fig. 9 gives the difference for the PPM and shows there are very large and spatially coherent areas that are significantly impacted by removing forecasts initiated from different ENSO states. Specifically for the DJF season, removing the forecasts initiated from a positive ENSO resulted in a significant decrease in the PPM over most of the United States especially for the southern portion where the PPM was the highest. The only exception is in the Northwest to north-central portion of the United States, which has some areas where there is an increase in the PPM. Removing the forecasts that were initialized during a negative ENSO state resulted in a similar pattern, with most of the United States showing a decrease in the PPM, with some areas having a small yet significant increase in the PPM for DJF. The two areas that show the highest PPM, the Southwest and the Southeast both show a significant decrease in the PPM for both positive and negative ENSO conditions. However, the Southeast is more affected by positive ENSO and the Southwest is more affected by negative ENSO conditions. In comparison, excluding neutral ENSO conditions result in

small yet significant increase in PPM for many parts of the United States. This indicates that the neutral ENSO years inhibit the predictive skill by adding noise that masks out the predictive signal due to ENSO. Overall these results indicate that ENSO plays a substantial role in the predictive skill in the CFSv2 model for the DJF season, specifically for the Southwest and Southeast United States.

Although there is a strong predictive signal in DJF from ENSO, JJA shows a mixed response to the ENSO state on the predictive skill. Removing the forecasts initialized from a positive ENSO state results in a decrease in the PPM for a large area of the Northwest and an increase in the PPM over the central Great Plains. This mixed response indicates that the positive ENSO state at initialization provides predictive skill for portions of the United States, but also inhibits the skill for others. The inhibition of predictive skill from positive ENSO conditions could be due to precipitation in the Great Plains being too strongly controlled by ENSO in the model and acts to mask out the signal from other important predictors like land–atmosphere feedbacks. In contrast, the negative ENSO conditions result in a decrease in the skill for both the Northwest and the central Great Plains. This indicates that the negative ENSO conditions provide a means for predictive skill for both of these areas. The neutral ENSO conditions also have a mixed response, in that it results in an increase in the PPM for the northwestern part of the country, but a significant decrease over a large portion of the central United States. This indicates that the predictive skill in the Northwest is greatly dependent on the ENSO state, while the central Great Plains shows a strong connection to neutral ENSO conditions, which suggests that there are drivers other than ENSO that contribute to the predictability for JJA.

5. Discussion and conclusions

This paper demonstrated the importance of considering consistent predictive skill metrics across multiple temporal and spatial scales and the clarity it brings to assessing the seasonal and spatial variation of seasonal forecast skill from a climate model, and its attribution. This was done through canonical event analysis, which considers the predictive skill for different temporal and spatial averages or canonical events. Using canonical event analysis provides a means to overcome the influence of sampling errors and enhances signal detection through temporal and spatial averaging. This provides a clearer assessment of consistent forecasts skill that is needed for practical use of seasonal forecast from GCMs. This was demonstrated through the probabilistic predictability metric (PPM), which is a frequency count of statistically significant Spearman correlations across temporal and spatial averages and showed clear spatial and temporal patterns of predictive skill from CFSv2 for precipitation, daily maximum, and minimum temperature. For precipitation, the PPM showed that the CFSv2 has two distinct seasons and areas of predictability, the cold season (November–February) for the southern United States and the warm season (June–August) for the north-central United States. The daily maximum temperature showed a higher predictability than precipitation overall; however, the predictability is weakest in the cold season (December and January) and the strongest in the summertime (June–August), which covered most of the United States except the central Great Plains. The daily minimum temperature showed the lowest predictability that peaked in April and November and roughly extended from the Great Lakes area down to west Texas. The difference between the PPM of the daily maximum and minimum temperature could be due to the long-term positive trend in the temperature and its variable impact on daily maximum and minimum temperature. A full treatment of the impact in predictive skill due to a long-term trend in precipitation and daily maximum and minimum temperature is outside the scope of this study and is left for future work.

One of the limitations of the PPM is that it is only a measure of statistically significant correlations and does not account for the magnitude of the correlation. For example, a forecast that is statistically significant with a correlation of 0.9 is more skillful than a statistically significant correlation of 0.5; however, this is not accounted for in the PPM. This needs to be kept in mind when comparing the PPM across space, seasons, and variables, as areas with the same PPM could provide very different prediction levels. This limitation could be overcome by calculating a companion metric to the PPM by averaging the correlations for all statistically significant events that would provide a measure of actual predictability. However, the average correlation over statistically significant events would not provide a sense of consistent skill across spatial and temporal scales and would merely be a companion statistic to the PPM to further distinguish between areas with a similar PPM.

The canonical event analysis and the PPM presented in this study provide clarity over metrics that rely on a single temporal and spatial scale, as was demonstrated in Fig. 5. The PPM provides a means to increase the practical use of the CFSv2 through instilling confidence to when and where there is predictive skill, which has practical application for hydrologic forecasting. In particular, downscaling precipitation and temperature

forecasts from GCMs for hydrologic prediction relies on a single or multiple spatiotemporal averages for most methods. If there is no predictive skill across canonical events, then the GCM provides little benefit for hydrologic forecasting. These results identify seasons and areas where there is the potential for skillful downscaled predictions. For example, a farmer in Idaho trying to predict the outlook of summer precipitation can have confidence that there is benefit from using the forecasts from CFSv2, while a farmer in the Northeast should know that there is little benefit for summertime seasonal predictions in his area. This is not as clearly evident when considering a single correlation in space and time (Fig. 5a). Although the PPM does not provide any benefit to the actual downscaling method, it provides the needed information about the temporal and spatial skill of the GCM to determine if downscaling provides any real benefit.

The practical use of seasonal forecasts can be further improved by understanding the attribution of the skill. The canonical event analysis and PPM demonstrated more sensitivity to the influences of the ENSO condition at forecast initialization on the predictive skill compared to a correlation based on a single temporal and spatial average (Figs. 8 and 9). Specifically, the PPM indicated that the predictive skill of precipitation in DJF is influenced by the ENSO condition at initialization, while the Southwest is more heavily influenced by negative ENSO conditions, which is consistent with the work of Hoerling et al. (2009). In contrast, the Southeast is more influenced by positive ENSO conditions. This attribution of the skill provides more utility and instills greater confidence in using seasonal predictions from the CFSv2. For instance, a farmer in Arizona should have more confidence in a DJF precipitation forecast initialized during a negative ENSO condition as compared to a neutral condition. This demonstrates only a few simple examples of how these results can be used to provide more confidence in seasonal predictions and provide a greater benefit for hydrologic forecasting. Furthermore, the canonical event analysis could be extended to additional climate indices and/or initial conditions and across multiple climate models to provide a better understanding of when and where the models provide useful forecasts. In addition, different climate models could demonstrate unique seasonal and spatial patterns of forecast skill that can be identified by the PPM. Having a better understanding when and where individual forecast models are skillful would be a great benefit for a multimodel framework such as the North American Multimodel Ensemble (Kirtman et al. 2014). The PPM analysis could provide information on redundant models and provide a means for generating model weights to optimally combine unique model forecasts specific to the climate state, spatial location, and season.

The PPM was also used to understand the effect of temporal averaging, spatial averaging, and lead time on the resulting skill of the forecast over the CONUS. The results showed that for precipitation, daily maximum and minimum temperature temporal averaging improved the skill across all seasons, although the seasonal increase was different across the variables (Fig. 4). This indicates that for all variables there is temporal noise in the predictions that can be reduced through temporal averaging, while maintaining the signal. In comparison, spatial averaging only increased forecast skill for precipitation and daily maximum temperature during the summer months. This suggests that only during the summertime and for precipitation and daily maximum temperature, does spatial averaging reduce the noise while maintaining the signal. The signal is most likely related to a physical process that occurs during the summer that only affects rainfall generation and the evolution of the daily maximum temperature. One such process could be land–atmosphere interactions, which affects both precipitation and daily maximum temperature through the latent and sensible heat fluxes at the surface. The spatial locations of the convective events driven by heat fluxes at the surface in the model are likely not exact, but through spatial averaging, the overall response is captured. This could also be why the spatial averaging does not improve forecast skill for daily minimum temperature. The daily minimum temperature usually occurs during the nighttime when land heat fluxes are the smallest and, therefore, is not directly controlled by a small-scale process. Furthermore, the absence of summertime predictability for precipitation (Fig. 5b) and daily maximum temperature (Fig. 6a) for large areas of the central Great Plains, which is considered as a ''hot spot'' for land–atmosphere interactions (Koster et al. 2004), is interesting. It could be that the biases demonstrated by Roundy et al. (2014) due to the tendency of the CFSv2 to persist in a wet coupling state could be inhibiting predictability. Another aspect that could be limiting the predictability over the central Great Plains is that the precipitation is overly controlled by the ENSO state at initialization in the model. This was suggested by the observation that the skill over the central Great Plains increased by removing all forecasts that were initiated from a positive ENSO condition (Fig. 9). Although fully understanding why the prediction is so low in this area is outside of the scoop of this study, this work demonstrates the utility of PPM to identify possible limitations of model physics and can provide vital information for modelers that can lead to model improvements and ultimately to better model predictions.

Although there are many different extensions to this work, the most important finding is that considering multiple spatial and temporal events when evaluating model skill in a canonical event analysis framework provides a clearer understanding of when and where the model is skillful. This understanding is vital for end users of seasonal forecasts as it provides them with a measure of confidence in the forecast. Even though the results showed that there are only limited areas and seasons where the CFSv2 forecasts can provide consistently skillful predictions, this skill should be utilized to provide the most benefit to society. It is also our hope that through continued improvement of climate models and research of forecast uncertainty, seasonal forecasts from climate models will be used as widely as weather forecasts to the ultimate benefit of society.

Acknowledgments. J. K. Roundy was supported through NASA Earth and Space Science Fellowship NNX08AU28H (Understanding Hydrologic Sensitivity and Land–Atmosphere Coupling through Space-Based Remote Sensing), and through support from the NOAA Climate Program Office (Grants NA10OAR4310246 and NA12OAR4310090). The support for this research is gratefully acknowledged. We would also like to thank the three anonymous reviewers for their helpful comments.

REFERENCES

Chen, M., W. Wang, and A. Kumar, 2013: Lagged ensembles, forecast configuration, and seasonal predictions. *Mon. Wea. Rev.,* **141,** 3477–3497, doi:10.1175/MWR-D-12-00184.1.

Fieller, E. C., H. O. Hartley, and E. S. Pearson, 1957: Tests for rank correlation coefficients. I. *Biometrika,* **44,** 470–481, doi:10.2307/2332878.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal to interannual climate predictions. *Int. J. Climatol.,* **21,** 1111–1152, doi:10.1002/joc.636.

Hoerling, M., X.-W. Quan, and J. Eischeid, 2009: Distinct causes for two principal U.S. droughts of the 20th century. *Geophys. Res. Lett.,* **36,** L19708, doi:10.1029/2009GL039860.

Karl, T. R., and Coauthors, 2012: U.S. temperature and drought: Recent anomalies and trends. *Eos, Trans. Amer. Geophys. Union,* **93,** 473–474, doi:10.1029/2012EO470001.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585–601, doi:10.1175/BAMS-D-12-00050.1.

Koster, R. D., and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science,* **305,** 1138–1140, doi:10.1126/science.1100217.

——, and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.,* **37,** L02402, doi:10.1029/2009GL041677.

Kumar, A., 2009: Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Mon. Wea. Rev.,* **137,** 2622–2631, doi:10.1175/2009MWR2814.1.

Luo, L. F., and E. F. Wood, 2006: Assessing the idealized predictability of precipitation and temperature in the NCEP Climate Forecast System. *Geophys. Res. Lett.,* **33,** L04708, doi:10.1029/2005GL025292.

——, and ——, 2007: Monitoring and predicting the 2007 U.S. drought. *Geophys. Res. Lett.,* **34,** L22702, doi:10.1029/2007GL031673.

——, ——, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.,* **112,** D10102, doi:10.1029/2006JD007655.

Mo, K. C., S. Shukla, D. P. Lettenmaier, and L.-C. Chen, 2012: Do Climate Forecast System (CFSv2) forecasts improve seasonal soil moisture prediction? *Geophys. Res. Lett.,* **39,** L23703, doi:10.1029/2012GL053598.

Palmer, T. N., and D. L. T. Anderson, 1994: The prospects for seasonal forecasting—A review paper. *Quart. J. Roy. Meteor. Soc.,* **120,** 755–793, doi:10.1002/qj.49712051802.

Roundy, J., C. Ferguson, and E. Wood, 2014: Impact of land-atmospheric coupling in CFSv2 on drought prediction. *Climate Dyn.,* **43,** 421–434, doi:10.1007/s00382-013-1982-7.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate,* **27,** 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Smith, J. A., M. L. Baeck, G. Villarini, D. B. Wright, and W. Krajewski, 2013: Extreme flood response: The June 2008 flooding in Iowa. *J. Hydrometeor.,* **14,** 1810–1825, doi:10.1175/JHM-D-12-0191.1.

Trenberth, K. E., and C. J. Guillemot, 1996: Physical processes involved in the 1988 drought and 1993 floods in North America. *J. Climate,* **9,** 1288–1298, doi:10.1175/1520-0442(1996)009<1288:PPIITD>2.0.CO;2.

——, G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.,* **103,** 14 291–14 324, doi:10.1029/97JC01444.

Villarini, G., J. A. Smith, R. Vitolo, and D. B. Stephenson, 2013: On the temporal clustering of US floods and its relationship to climate teleconnection patterns. *Int. J. Climatol.,* **33,** 629–640, doi:10.1002/joc.3458.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* **107,** 4429, doi:10.1029/2001JD000659.

Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.,* **117,** D03109, doi:10.1029/2011JD016048.

Yuan, X., and E. F. Wood, 2012: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resour. Res.,* **48,** W12519, doi:10.1029/2012WR012256.

——, ——, L. F. Luo, and M. Pan, 2011: A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophys. Res. Lett.,* **38,** L13402, doi:10.1029/2011GL047792.

——, ——, J. K. Roundy, and M. Pan, 2013: CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. *J. Climate,* **26,** 4828–4847, doi:10.1175/JCLI-D-12-00683.1.