**Hydrology and Earth System Sciences**

# Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill

**S. Shukla[1], N. Voisin[2], and D. P. Lettenmaier[1]**

[1]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, 98195, USA
[2]Pacific Northwest National Laboratory, Richland, WA, 99352, USA

*Correspondence to:* D. P. Lettenmaier (dennisl@uw.edu)

**Abstract.** We investigated the contribution of medium range weather forecasts with lead times of up to 14 days to seasonal hydrologic prediction skill over the conterminous United States (CONUS). Three different Ensemble Streamflow Prediction (ESP) based experiments were performed for the period 1980–2003 using the Variable Infiltration Capacity (VIC) hydrology model to generate forecasts of monthly runoff and soil moisture (SM) at lead-1 (first month of the forecast period) to lead-3. The first experiment (ESP) used a resampling from the retrospective period 1980–2003 and represented full climatological uncertainty for the entire forecast period. In the second and third experiments, the first 14 days of each ESP ensemble member were replaced by either observations (perfect 14-day forecast) or by a deterministic 14-day weather forecast. We used Spearman rank correlations of forecasts and observations as the forecast skill score. We estimated the potential and actual improvement in baseline skill as the difference between the skill of experiments 2 and 3 relative to ESP, respectively. We found that useful runoff and SM forecast skill at lead-1 to -3 months can be obtained by exploiting medium range weather forecast skill in conjunction with the skill derived by the knowledge of initial hydrologic conditions. Potential improvement in baseline skill by using medium range weather forecasts for runoff [SM] forecasts generally varies from 0 to 0.8 [0 to 0.5] as measured by differences in correlations, with actual improvement generally from 0 to 0.8 of the potential improvement. With some exceptions, most of the improvement in runoff is for lead-1 forecasts, although some improvement in SM was achieved at lead-2.

## 1 Introduction

Droughts are among the most expensive natural disasters (Ross and Lott, 2003). Proactive risk-based approaches to drought management that include better monitoring, early warning and prediction are essential for mitigating drought losses (Schubert et al., 2007). Seasonal hydrologic and drought prediction systems, such as the NOAA Climate Prediction Center's (CPC) seasonal drought outlook, derive their skill from knowledge of initial hydrologic conditions (IHCs) and weather/climate information during the forecast period. The contribution of IHCs and climate forecast skill in seasonal hydrologic prediction varies seasonally, spatially and with lead time. Over the conterminous United States (CONUS), Shukla and Lettenmaier (2011) found that IHCs generally dominate at short leads (i.e. 1–2 months), while climate forecast skill dominates for longer leads; although, IHCs can account for a substantial part of the total hydrologic forecast skill under some conditions for leads of as long as 6 months.

Macro-scale land surface models (LSMs) provide a reasonably accurate estimate of IHCs at the time of forecast initialization for seasonal hydrologic prediction. For example, seasonal hydrologic/drought prediction systems, such as The National Centers for Environmental Prediction's (NCEP) drought monitor (http://www.emc.ncep.noaa.gov/mmb/nldas/forecast/TSM/prob/) and the University of Washington's Surface Water Monitor (http://www.hydro.washington.edu/forecast/monitor/outlook/index.shtml), use IHCs generated by LSMs. Within the multi-institutional North American Land Data Assimilation System project (Mitchell et al., 1999, 2004), a suite of large scale hydrologic

models have been developed and tested over the CONUS for their ability to simulate various hydrometeorological processes (Cosgrove et al., 2003; Luo et al., 2003; Pan et al., 2003; Schaake et al., 2004; Sheffield et al., 2003; Xia et al., 2011a, b)

Simultaneously, major strides have been made toward understanding the sources of predictability of seasonal precipitation and temperature in the US (Higgins et al., 2000), and improving climate forecasts (O'Lenic et al., 2008). Statistical and physical modeling approaches can exploit predictability in the climate system primarily via the thermal inertia present in sea surface temperatures (Barnston et al., 1999), especially during strong El Niño/La Niña-Southern Oscillation years. Otherwise, precipitation forecast skill beyond a month or so is quite limited (Quan et al., 2006; Wilks, 2000). Precipitation forecast skill is generally lower than the skill of forecasts for temperature or atmospheric circulation patterns for the same location and time (Barnston et al., 2010; Gong et al., 2003; Lavers et al., 2009; Wilks and Godfrey, 2002). Since precipitation is the major driver of drought conditions, seasonal drought prediction skill is severely limited by the lack of precipitation forecast skill under most conditions. The difficulty of forecasting rainfall, mainly during summer, has been a major stumbling block for the CPC's seasonal drought outlook as well (Hayes et al., 2005).

Various statistical and dynamical methods of seasonal hydrologic forecasting have been developed and are being used operationally (see e.g. Day, 1985; Wood et al., 2002; Luo et al., 2007; Wang et al., 2009; Wood and Lettenmaier, 2006). Most previous studies have found that due to limited seasonal climate forecast skill, seasonal hydrologic forecast skill comes in substantial part from IHCs (Wood et al., 2002, 2005; Lavers et al., 2009, Lettenmaier and Wood, 2009). One potential means for improving seasonal hydrologic prediction is to better exploit medium range weather forecasts (MRWFs) for the first 14 days of a seasonal forecast period. MRWFs have greatly improved in the last two decades as increased computer power and more integrated observation systems have allowed general circulation models to run at finer resolutions with improved initializations (Pappenberger et al., 2005). MRWFs have been coupled with LSMs to provide flood and streamflow forecasts for lead times of up to 2 weeks, using both deterministic and probabilistic approaches (Clark and Hay, 2004; Hou et al., 2009; Thielen et al., 2009; Voisin et al., 2011; Werner et al., 2005). Werner et al. (2005) found that incorporating 14-day precipitation and temperature forecasts from a MRWF model into the National Weather River Forecast System's traditional ESP forecast system generally improved the streamflow forecast skill for up to 18 days. Hou et al. (2009) evaluated the Global Ensemble Forecast System of NCEP coupled with the Noah LSM for its ability to provide useful streamflow forecast skill. They concluded that the coupled system had some positive streamflow forecast skill at lead times varying from 1–3

days for smaller basins and more than 7–10 days for large river basins.

The use of MRWFs has been mostly limited to up to two weeks in lead time, and their value in improving hydrologic prediction at seasonal scale is largely unexplored so far. By merging MRWFs ($\sim$ 14 day lead) with seasonal climate forecasts, seasonal hydrologic prediction skill could potentially be (i) improved at short lead times ($\sim$ 1–2 months) and (ii) extended in time beyond what is derived solely from the IHCs, particularly in those cases when climate forecasts at even short lead times have skill that is no better than climatological.

The goal of this study is to assess the contribution of MRWFs in seasonal hydrologic prediction. Specifically, we evaluate the potential of MRWFs to improve seasonal hydrologic forecast skill relative to that achievable by the Ensemble Streamflow Prediction (ESP) approach. ESP (Day, 1985; Franz et al., 2003) is a method that involves running an LSM up to the forecast initialization date using observed forcings, and then producing ensembles by resampling time sequences of forcings from years in the historic record. Hence, its skill is derived solely from knowledge of IHCs. We evaluate the additional forecast skill derivable from MRWFs in the context of hydrologic ensembles of monthly runoff and mean monthly soil moisture (SM) at leads from one to several months.

## 2 Approach

Three ESP-based experiments were conducted. The basic framework for each experiment was the same: IHCs were derived by running an LSM using observed meteorological forcings until the date of forecast initialization, i.e. on the first of each month in the 1980–2003 period. In forecast mode, the LSM was forced with 3-month long observed meteorological forcings resampled from the historical period (23 ensemble members from 1980–2003) using a leave-one-year-out approach and starting on the day of the forecast, i.e. on the first of each month. The experiments differed in the forcings for the first 14 days of the forecast periods as follows:

- The first experiment (hereafter referred to as ESP) used the conventional ESP framework (Fig. 1a; as in Wood and Lettenmaier 2006, 2008; Wood et al., 2002; Li et al., 2009; Shukla and Lettenmaier, 2011). It defines the baseline seasonal hydrologic prediction skill.

- In the second experiment (hereafter referred to as OBS_Merged_ESP), the first 14 days of each ESP ensemble member were replaced with observations (i.e. perfect MRWF). For example, as shown in Fig. 1b, the forcings used for days 1 to 14 were the observations during that period (deterministic perfect forecast), beyond which the forecast ensemble members were the same
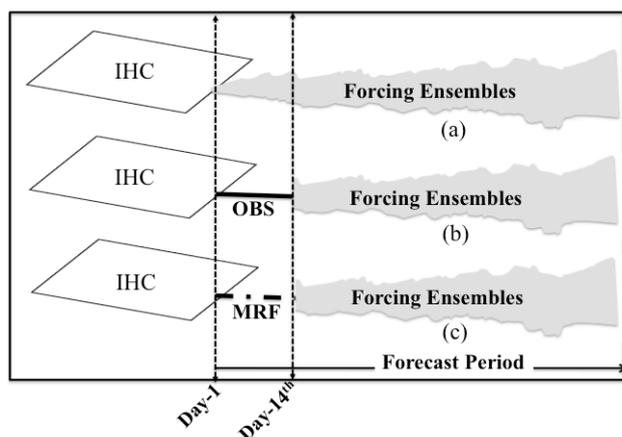
**Fig. 1.** Schematic showing the framework for **(a)** Experiment 1 (ESP) **(b)** Experiment 2 (OBS_Merged_ESP) and **(c)** Experiment 3 (MRF_Merged_ESP).

as in ESP. OBS_Merged_ESP defines the maximum improvement in seasonal hydrologic prediction skill that can be obtained if perfect knowledge of the LSM forcings could be extended to 14 days in the future.

– The third experiment (hereafter referred to as MRF_Merged_ESP) is similar to the second experiment, but observations for the first 14 days in each ensemble member were replaced with a deterministic MRWF (Fig. 1c). This experiment defines the actual improvement in seasonal hydrologic prediction skill that can be derived from use of realistic weather forecasts over those 14 days. The skill contributed by these forecasts may also be limited by the need to downscale the MRWF to the spatial resolution of the hydrologic model (one-half degree in the case of our experiments).

The skill of each experiment was estimated with respect to the "simulated observed" values (hereafter referred to as reference values) of runoff and SM, which were treated as surrogates for observations. The reference runoff and SM were obtained from a consistent long-term (1980–2003) simulation of the Variable Infiltration Capacity (VIC) LSM (Sect. 2.1.1) forced with observed gridded station data (see Sect. 2.1.2).

## 2.1 LSM and forcing data

### 2.1.1 The Variable Infiltration Capacity (VIC) model

The VIC macro-scale hydrology model (Liang et al., 1994; 1996; Cherkauer et al., 2003) was run at a daily time step and 1/2 degree latitude-longitude spatial resolution. The VIC model includes a parameterization for spatial variability of the infiltration capacity (and hence variability of runoff) and evaporation from different vegetation types, as well as bare soil evaporation. It provides for non-linear dependence of the

partitioning of precipitation into infiltration and direct runoff as determined by soil moisture in the upper layer and its spatial heterogeneity. The subsurface is partitioned into three layers. The first layer has a fixed depth of $\sim 10$ cm and responds quickly to changes in surface conditions and precipitation. Moisture transfers between the first and second, and second and third soil layers are governed by gravity drainage, with diffusion from the second to the upper layer allowed in unsaturated conditions. Base flow is a non-linear function of the moisture content of the third soil-layer (Liang et al., 1994; Todini, 1996). The model was run in water balance mode; which means that the surface temperature is assumed equal to the surface air temperature, and is not iterated for energy balance closure (this also implies zero ground heat flux). The VIC model represents the snowpack as a two-layer medium (a thin surface, and a thick deeper layer), and solves an energy and mass balance as part of its computation of pack ablation (Andreadis et al., 2009).

### 2.1.2 Retrospective simulation (Control Run)

Given the lack of observed spatially distributed runoff (a variable produced by the model, in contrast to streamflow, which is an observed variable) and the absence of spatially distributed soil moisture observations (note that satellite observations of soil moisture do exist, but they do not fully coincide with our period of analysis and furthermore are limited to the upper few cm of the soil column), we chose to use a historic reference VIC model simulation as the basis for evaluation of both runoff and soil moisture. Maurer et al. (2002) have shown that hydrologic variables and fluxes derived from the VIC model generally are in good agreement with available observations across the CONUS domain.

A consistent data set of runoff and mean monthly SM over the analysis period (1980–2003) to be used as the reference was generated by forcing the VIC model with observed gridded meteorological forcings over the analysis period. This simulation also included a lengthy model spinup to remove the effects of initial conditions. The model forcings (daily precipitation, and maximum ($T_{\max}$) and minimum ($T_{\min}$) temperature) were taken from Cooperative Observer Program stations, and gridded at 1/2 degree spatial resolution using methods outlined in Maurer et al. (2002). Additional model forcings (downward solar and longwave radiation, and humidity) were estimated from the daily air temperature and temperature range following methods outlined in Maurer et al. (2002). Surface wind was taken from the lowest level of the NCEP/NCAR reanalysis (Kalnay et al., 1996). The IHCs for each forecast initialization day used in the experiments were provided by this control run.

### 2.1.3 Weather forecasts

We used the 1979–2005 15-day 12-hourly 2.5-degree NCEP/Climate Diagnostics Center Medium Range Forecast
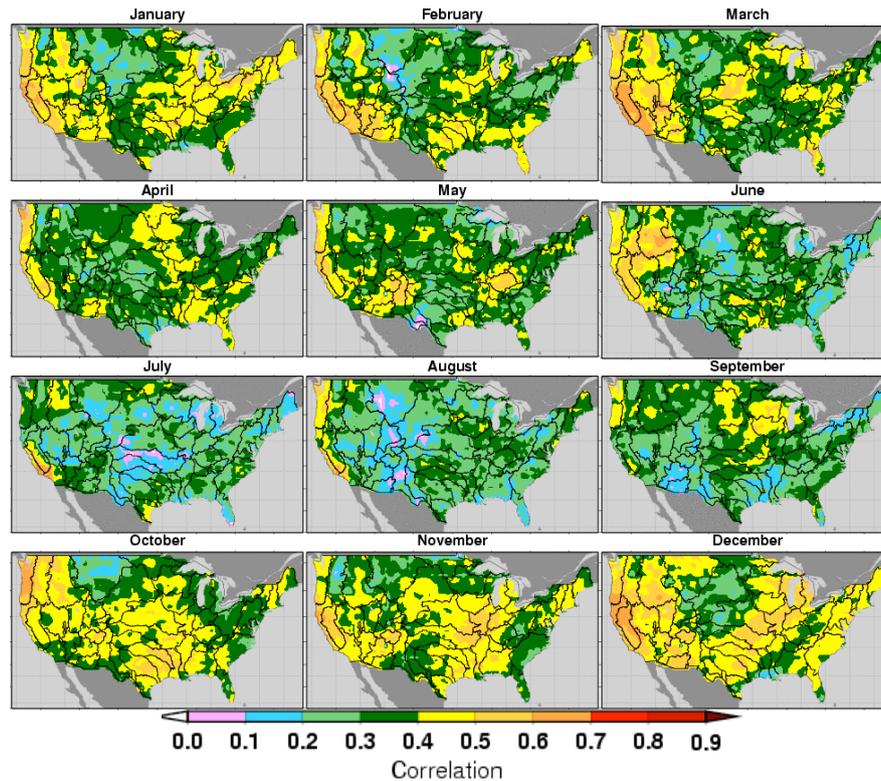
**Fig. 2.** Correlation between observed and forecasted (MRF) 14-day accumulated precipitation during each month.

(MRF) reforecast data set from Hamill et al. (2006). The Hamill et al. (2006) data set uses a fixed version (1998) of the NCEP global forecast model and hence should have nearly consistent (aside from some differences in the data that were available for assimilation) forecast skill over the period of analysis. The reforecasts were downscaled from their native resolution (2.5 degree) to the 0.5-degree scale of the hydrology model and bias corrected to be consistent with the meteorological forcings used in the LSM spinup and reference simulation. The downscaling was performed by first aggregating the 12-hourly ensemble mean forecasts to 14 days, then interpolating the ensemble averages using an inverse squared distance interpolation scheme (Shepard, 1984). Figures 2 and 3 show the Spearman rank correlation between the observed and downscaled forecasts (at 1/2 degree resolution) of 14-day accumulated precipitation and 14-day mean average daily temperature. In general, 14-day average daily temperature forecast skill is much higher than the 14-day accumulated precipitation skill for any given month, and the weather forecast skill (both precipitation and average temperature) is lowest in summer months (June, July and August). The precipitation forecast skill is highest over the Pacific coastal regions and parts of the eastern US during winter, a pattern that was observed by Clark and Hay (2004) and Hamill et al. (2006) as well (however, for daily precipitation totals for lead times of up to 5 days or so only).

The downscaled and accumulated 14-day weather forecasts were subsequently bias corrected by rescaling so that the long term 14-day accumulated mean precipitation and average $T_{max}$ and $T_{min}$ matched the corresponding values from the observed gridded forcings over the 1980–2003 period (Wood et al., 2002; Voisin et al., 2010). We used all years in the observed and forecast data sets to estimate the climatology, notwithstanding that the year of forecast was included in the probability distribution estimates (the small effect of reducing the number of observations used to estimate the probability distributions by one year is not likely to have much effect on the results). The 14-day spatially downscaled and bias corrected forecasts were then temporally disaggregated to daily forecasts using the observed sequencing of precipitation and temperature. More elaborate weather pre-processors could have been used (e.g. Schaake et al., 2007; Voisin et al., 2010; Wu et al., 2011); however, given our focus on seasonal hydrological forecasts, the daily sequencing of events is less important than the aggregate quantities.

We chose to merge the 14-day ensemble mean forecasts, rather than each ensemble member, into the post-day 14 ESP ensembles to avoid complications in merging two sets of ensembles (Clark et al., 2004). Furthermore, this limits the impact of the calibration and downscaling approach on the seasonal hydrologic ensemble forecast skill. Our approach is similar to that of Clark and Hay (2004) and Werner et
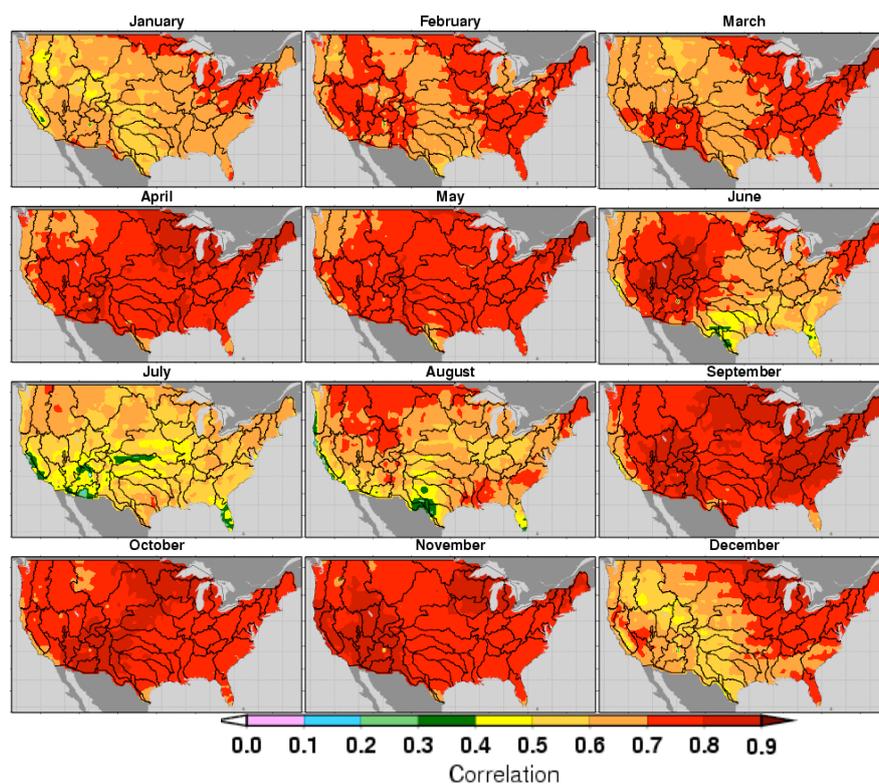
**Fig. 3.** Correlation between observed and forecasted (MRF) 14-day mean average-temperature during each month.

al. (2005); however, unlike in those studies we use the ensembles average, not the probabilistic MRF forecasts, to merge with ESP forecasts. In contrast to the merged forecasts, the ESP approach, which simply takes the first 14 days of the forecast period from random resampling of past observations, derives its skill entirely from the knowledge of the IHCs (whereas in MRF_Merged_ESP some skill comes from weather forecasts during the first 14 days).

The bias correction and statistical disaggregation approach in general reduces or eliminates biases, but does not preserve probabilistic information inherent in the ensemble forecasts (Voisin et al., 2010). Here, we evaluate the potential improvement in seasonal hydrologic prediction from merging MRF with ESP, assuming that the information in the MRF ensemble is not calibrated and only the ensemble mean forecast is useful for our application.

## 2.2 Forecast skill score

For simplicity, daily spatially distributed runoff and SM forecasts and reference values (obtained from the control run) were aggregated in time to monthly accumulations or averages, and to the spatial scale of 18 hydrologic sub-regions across the CONUS domain (Table 1). These sub-regions are the same as the sub-regions used in Shukla and Lettenmaier (2011) and were created by merging the 221 USGS

**Table 1.** List of USGS water resources regions.

| Region 01 | New England (NE) |
|---|---|
| Region 02 | Mid-Atlantic (MA) |
| Region 03 | South Atlantic-Gulf (SAG) |
| Region 04 | Great Lakes (GL) |
| Region 05 | Ohio (OH) |
| Region 06 | Tennessee (TN) |
| Region 07 | Upper Mississippi (UM) |
| Region 08 | Lower Mississippi (LM) |
| Region 09 | Souris-Red-Rainy (SRR) |
| Region 10 | Missouri (MO) |
| Region 11 | Arkansas-White-Red (AR) |
| Region 12 | Texas-Gulf (TX) |
| Region 13 | Rio Grande (RG) |
| Region 14 | Upper Colorado (UC) |
| Region 15 | Lower Colorado (LC) |
| Region 16 | Great Basin (GB) |
| Region 17 | Pacific Northwest (PNW) |
| Region 18 | California (CA) |

hydrologic sub-regions. Each of the sub-regions is named after the water resources region in which it is located (Table 1).

To evaluate the prediction skill of each experiment, we estimated Spearman rank correlation coefficients (Wilks, 2006) between the ensemble mean forecasts (over years)
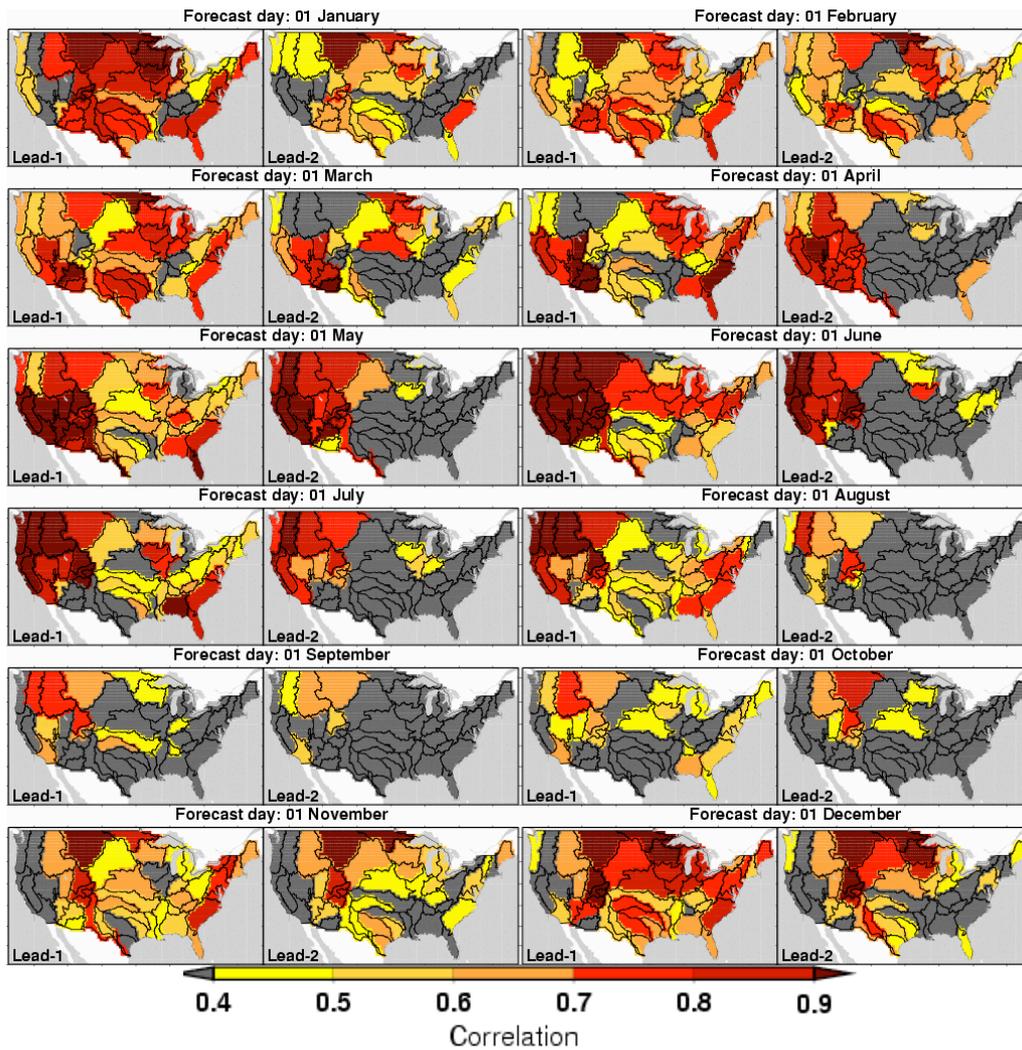
**Fig. 4.** Baseline skill (i.e. skill of ESP experiment) for runoff forecasts at leads 1–2 months (dark grey color shows the sub-regions where the baseline skill is not significant at the 95 % significance level).

and the reference simulations. The Spearman rank correlation is a measure of monotonic associations (both linear and non-linear) between forecasts and observations (Jolliffe and Stephenson, 2003), unlike the Pearson correlation that is a measure of linear associations only. Furthermore, the Spearman rank correlation is calculated using the ranks of the data so outliers in the sample (and zeros in particular for the variables of interest) do not impact the Spearman rank correlation coefficient (in contrast to the Pearson correlation coefficient, Wilks, 2006). The skill (rank correlation) of the ESP experiment is considered to be the "Baseline skill". We considered the difference between the skill of OBS_Merged_ESP and the ESP experiment as the potential improvement, and the difference between the skill of MRF_Merged_ESP and the ESP experiment as the actual improvement in baseline skill.

## 3   Results

We present the results for a forecast period of 2 months only (Figs. 4, 5, 6, 7, 8 and 9). Although in a few cases we observed improvements in seasonal hydrologic prediction skill due to use of MRWFs for three-month lead, generally the improvement in skill was limited to lead-1 and lead-2.

First, we show the baseline skill (skill of the ESP experiment). The sub-regions where the baseline skill is not significant at the 95 % significance level (i.e. given the degrees of freedom of the sample, the correlation value is lower than it would have to be to reject the hypothesis of the correlation being different from 0 at the significance level of 95 %) have been masked and are shown in dark grey (the critical value of the Spearman rank correlation was estimated using the table given in Zar (1972)). We then show the potential improvement in the baseline skill (the difference between the skill
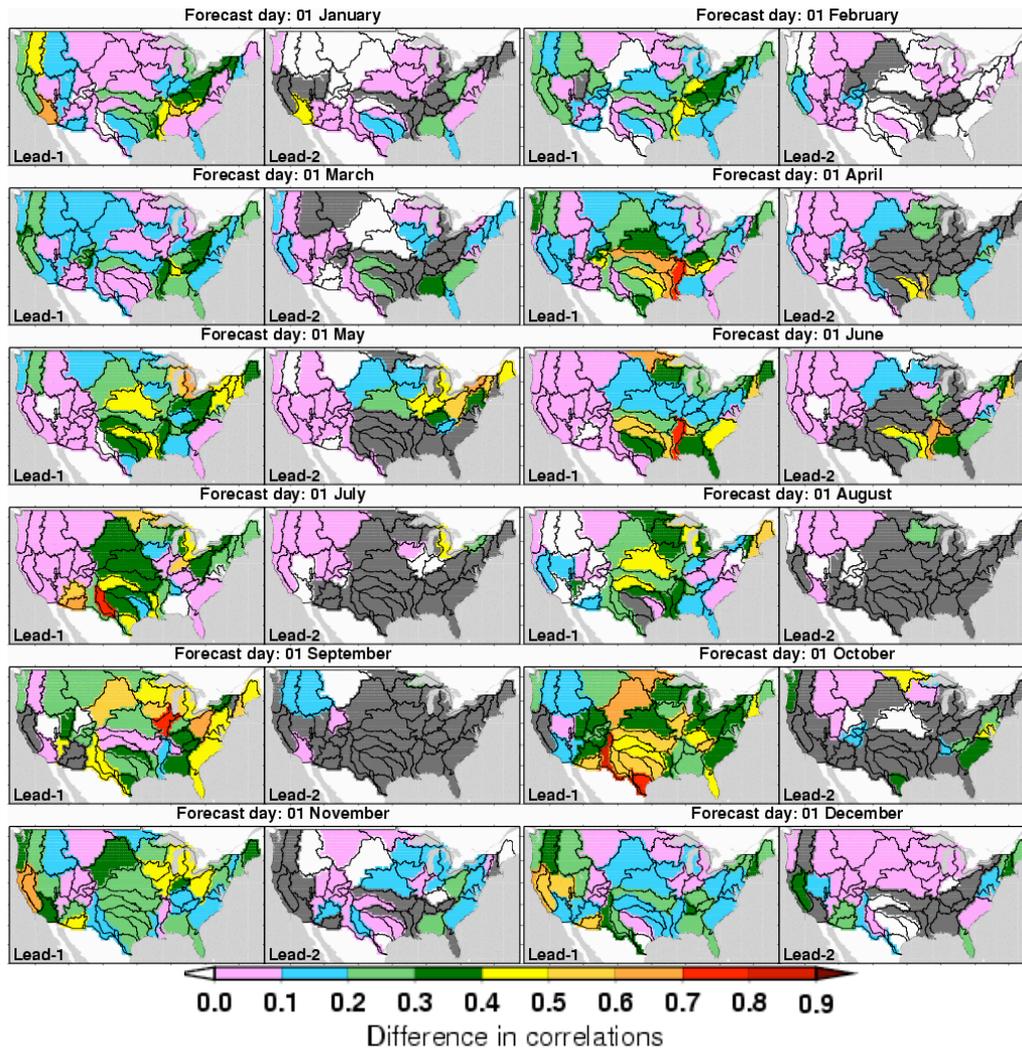
Fig. 5. Potential improvement in runoff forecast skills at leads 1–2 months (dark grey color shows the sub-regions where the skill of OBS_Merged_ESP is not significant at the 95 % significance level).

of OBS_Merged_ESP and ESP experiments). Again, the improvement is shown over those sub-regions where the skill of OBS_Merged_ESP is significant at the 95 % level. Finally, we show the ratio of the actual improvement in skill (difference between the skill of MRF_Merged_ESP and ESP experiment) and the potential improvement in skill, to highlight the level of the improvement in skill actually recovered by using realistic MRWFs. We show the actual improvement in skill over those sub-regions only where the potential improvement in skill is > 0.1, and the skill of OBS_Merged_ESP is significant at the 95 % level.

## 3.1 Monthly runoff forecasts

The correlations of ensemble mean monthly runoff forecasts from ESP initialized (baseline skill) on day 1 of each month with the reference runoff at leads 1 to 2 months are shown

in Fig. 4. In general, the baseline skill is highest at lead-1. Overall, across the CONUS, the baseline skill for runoff forecasts is highest during forecast periods starting in winter months (i.e. December, January, and February (DJF)) and lowest during forecast periods starting in fall months (mainly September and October). During forecast periods starting in spring (March, April, and May (MAM)) and early summer months (June and July), the western US stands out with relatively high runoff forecast skill up to lead-2 (and beyond, not shown here). This is mostly attributable to the effects of snow, which provides substantial IHC-related forecast skill for forecast periods starting in late winter to early summer.

Figure 5 shows the potential improvement in baseline skill of monthly runoff forecasts (i.e. the difference between the skill score of runoff forecasts from OBS_Merged_ESP and ESP). Not surprisingly, the greatest improvement in runoff
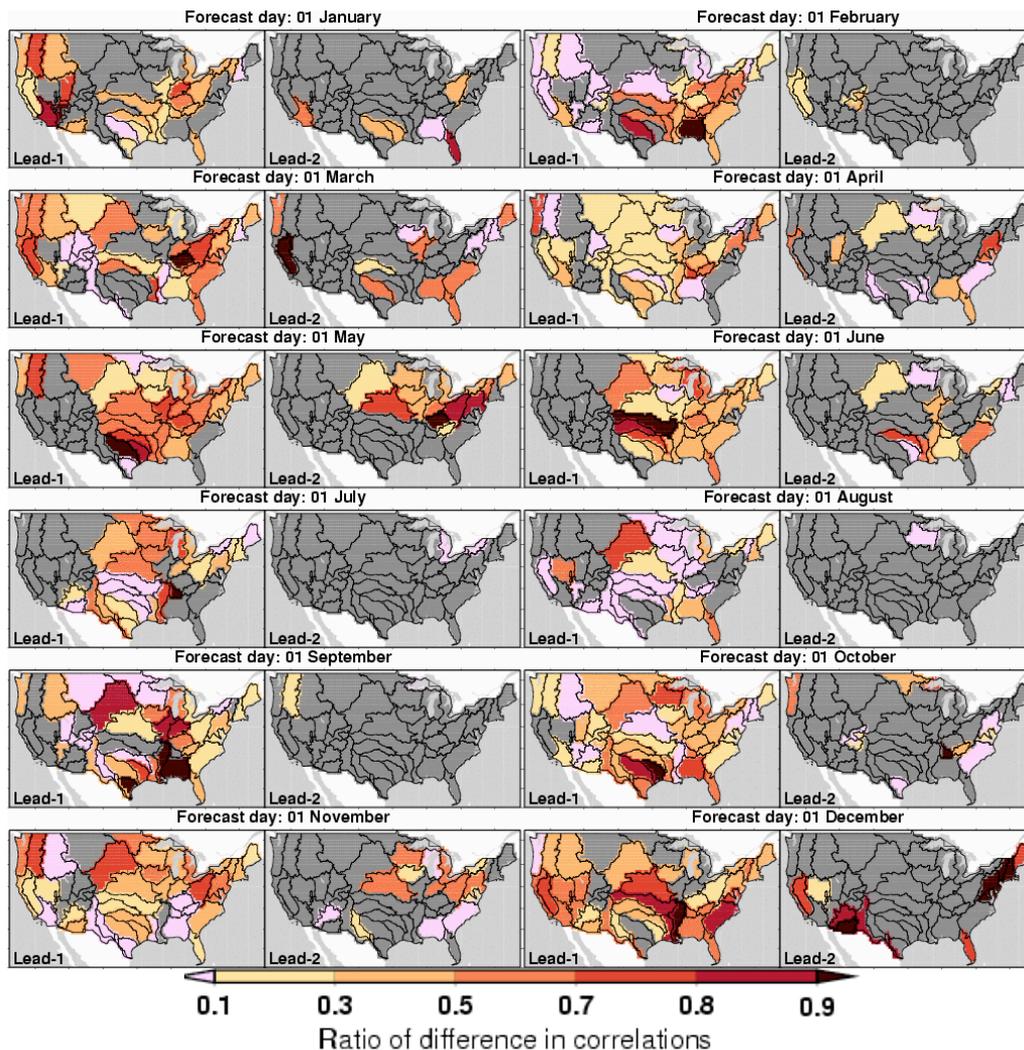
**Fig. 6.** The ratio of actual improvement and potential improvement in baseline runoff forecast skill at leads 1–2 months (dark grey color shows the sub-regions where either the potential improvement in skill is < 0.1 or the skill of OBS_Merged_ESP is not significant at the 95 % significance level).

forecast skill is at lead-1, and the effect decreases with lead time. The largest improvement in skill for any given sub-region at lead-1 is generally in those cases where the first month of the forecast period is climatologically wet. This is the case, for example, for sub-regions in the Great Plains (i.e. the area of generally low relief east of the Rocky Mountains and west of the Mississippi River), Midwest (most of the Mississippi Basin) and Lower Mississippi sub-regions for forecasts starting in April through October, and for the Pacific coastal sub-regions for forecast periods starting in November through the winter months (i.e. DJF). On the other hand, the improvement in skill at lead-1 is small for sub-regions for which the first month of the forecast period is climatologically dry or the initial moisture variability is much higher than the precipitation variability during the forecast period (small $\kappa$ values according to the convention of Ma-

hanama et al., 2011); such conditions lead to high baseline skill. This is the case for instance in the interior of the western US during spring and summer months.

In some cases, the improvements in skill due to use of perfect MRWFs persists into leads-2 and -3 (not shown). These cases likely correspond to better knowledge of IHCs at the end of the 14 days in the OBS_Merged_ESP experiment than in the ESP experiment.

The potential improvement in skill shown in Fig. 5 clearly is optimistic relative to what is achievable in practice, because weather forecast skill is imperfect even for the smallest (e.g. one day) leads, and declines thereafter throughout the 14-day MRWF period.

Figure 6 shows the ratio of actual improvement in skill (differences in correlations for runoff forecasts derived by MRF_Merged_ESP and ESP) to potential improvement in

skill (as discussed above), and indicates the improvement in runoff forecast skill that can be achieved realistically by using MRF medium range weather forecasts for the first 14 days of the forecast period. (It should be noted that these results may be slightly pessimistic as the MRF model has been retired, and MRWF skill for current generation weather forecast models may be slightly higher. However, the MRF reforecast data set is unique in providing a consistent set of reforecasts appropriate for the type of analysis we have performed; a newer version of this data set is planned but has not yet been released.) Two main factors control the actual improvement in runoff forecast skill: (i) the potential improvement in skill (as shown in Fig. 5, derived from the use of perfect MRWFs) and (ii) the forecast skill of the MRWFs themselves. In other words, the improvement in skill due to use of MRWFs will be highest when both the potential improvement in hydrologic forecast skill and the MRWF skill (primarily for precipitation) are high. Therefore, in Fig. 6 we show the actual improvement over those sub-regions only where the skill of OBS_Merged_ESP is significant at 95 % level and the potential improvement in baseline skill is greater than 0.1.

In general, Fig. 6 shows that the actual improvement in skill due to use of the MRF forecasts is highest for those sub-regions and times of the year where the first month is climatologically wet. Overall, the actual improvement in skill is extensive over the Great Plains, Midwest, Texas-Gulf and parts of the northern and southeastern US at lead-1 during the forecast periods starting in spring (mainly April and May), summer (mainly June and July) and fall (September, October, and November) months. Over the mountainous western sub-regions, the actual improvement in skill is highest during the forecast periods initialized on 1 November, December and January. Again, those are also the forecast periods when the baseline skill is low over those regions (Fig. 4), whereas during the forecast periods starting in spring and summer months (when the baseline skill is high) both the potential and actual improvement in skill is generally negligible (Figs. 5 and 6). The sub-regions shown in white during each forecast period show potential improvement but little or no actual improvement, likely due to limited MRF precipitation forecast skill.

## 3.2 Soil moisture (SM) forecasts

Figure 7 shows the baseline skill for SM forecasts for lead-1 and lead-2. In general, the baseline skill for SM is much higher than for runoff (Figs. 5 and 7). Shukla and Lettenmaier (2011) also showed that at lead-1 IHCs generally dominate SM forecast skill.

Similar to the case of runoff forecasts across the CONUS, baseline skill for SM is generally highest during forecast periods starting in the winter, with higher skill over the western as compared with the eastern US. The baseline skill at leads-2 (and -3, not shown here) is high over the interior of the

western US for forecast periods starting on day 1 of spring (MAM) and summer (June, July, and August (JJA)) months.

The potential improvement in the baseline skill of SM forecasts for each forecast period is shown in Fig. 8. Overall, the potential improvement in SM forecast skill at lead-1 is lower than the corresponding values for monthly runoff forecast skill (Figs. 5 and 8). This appears to be a result of the high baseline skill for SM at lead-1 (i.e. high contribution of IHCs in SM forecast skill), hence leaving less room for improvement than for the case of runoff (since the maximum correlation value or the value of skill is 1). As for runoff, the greatest potential improvement in skill is for sub-regions and forecast periods where the lead-1 month is climatologically wet. Improvements at lead-1 are mostly limited to the southwestern and eastern US (and Great Plains in a few cases), where the contribution of IHCs to SM forecast skill is lower than for the western US. Mainly in the forecast periods starting in April, May, June and fall months (September and October), relatively large potential improvements can be seen over those regions. The potential improvement in skill at lead-2, however, seems more extensive in the case of SM forecasts than runoff. There could be a few explanations for this pattern. First, more sub-regions show significant levels of OBS_Merged_ESP skill at lead-2 in the case of SM forecast skill than in that of runoff skill (therefore fewer regions are shown in dark grey at lead-2 in Fig. 8 than Fig. 5). Second, the baseline skill for SM forecasts (i.e. skill of ESP experiments) at lead-2 is smaller than at lead-1, leaving more room for improvement in skill. Finally, the improvement in SM forecast skill at lead-2 could be a result of persistence of the contribution of MRWF skill at lead-1. Once again, the potential improvement in SM forecast skill at lead-2 is generally prominent over the eastern half of the country.

The ratio of actual to potential improvement in SM forecast skill is shown in Fig. 9. The actual improvement in skill is shown only over the regions where potential improvement in SM forecast skill is greater than 0.1 and the skill of OBS_Merged_ESP is significant at the 95 % level. Since the baseline skill of ESP (and hence skill of OBS_Merged_ESP) is generally significant across the CONUS at lead-1, the sub-regions shown in grey in Fig. 9 are mostly those regions where the potential improvement in skill is lower than 0.1. Overall, for the most part the actual improvement in skill is limited to the sub-regions in the eastern half of the US, mostly during the forecast periods starting in April, May, June, September and October. Actual improvement in skill, however, can be seen over Pacific coastal regions at lead-1 for forecast periods starting in November and December. Again, following the pattern of potential improvement, actual improvement at lead-2 in SM forecast skill also seems more extensive than in runoff forecast skill. This could be due to the persistence of the contributions of MRWFs at lead-1.
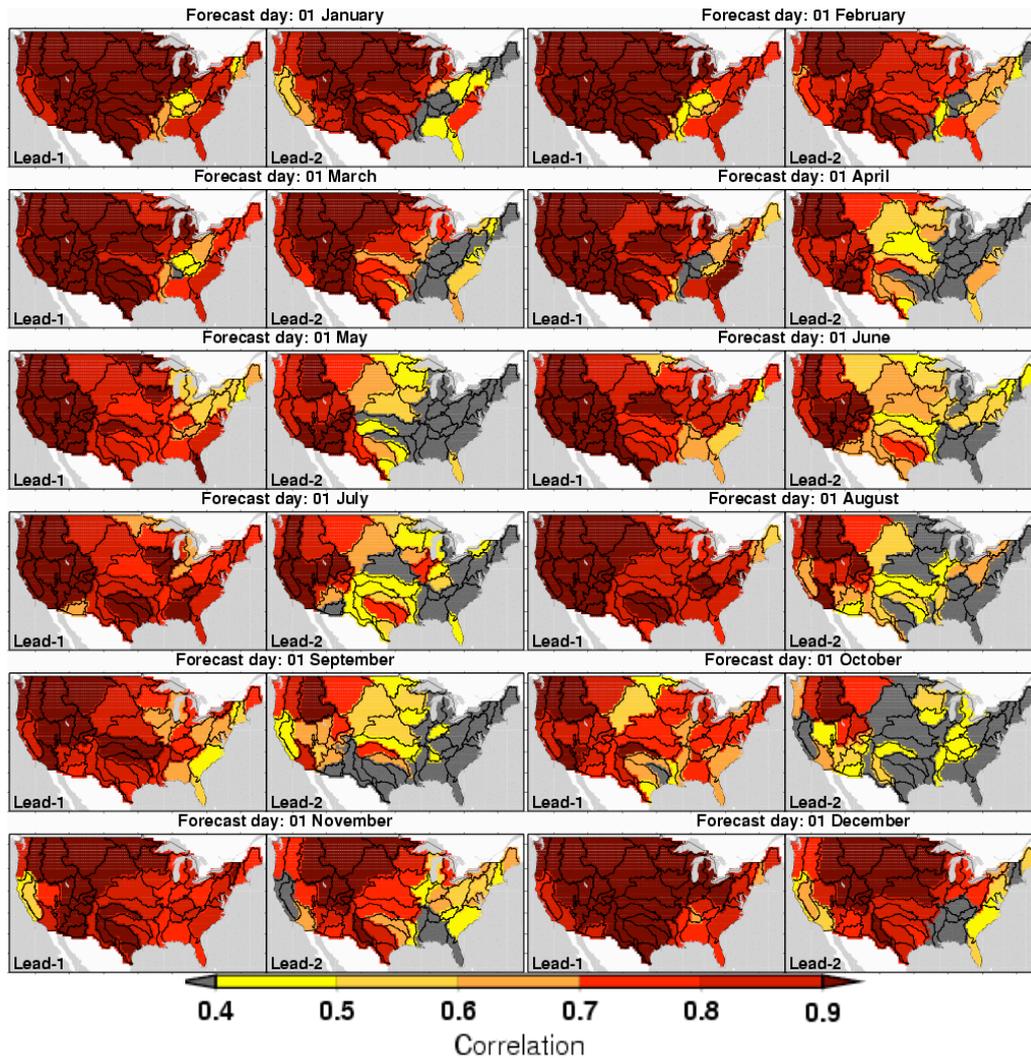
**Fig. 7.** Baseline skill (i.e. skill of ESP experiment) for SM forecasts at leads 1–2 months (dark grey color shows the sub-regions where the baseline skill is not significant at the 95 % significance level).

## 4   Discussion

Not surprisingly, the OBS_Merged_ESP experiment that used observed forcings over the first 14 days showed the greatest improvement in baseline skill, while the MRF_Merged_ESP experiment that used realistic weather forecasts (i.e. MRF forecasts) showed smaller or no improvement. However, further improvement in MRWF skill will presumably lead to improvement in seasonal hydrologic prediction skill in those sub-regions and forecast periods where the use of perfect medium range weather forecasts yields most improvement in seasonal hydrologic prediction skill. For example, during summer months (JJA), when the potential improvement for interior western US regions and much of the eastern US is greater than 0.2, the actual improvement is limited due to the limited MRWF skill (Figs. 2 and 3).

We used a simple bias correction and disaggregation approach in the MRF_Merged_ESP experiment (Sect. 2.1.3). Our focus was on the removal of bias in the 14-day accumulated forecast. Our analyses were performed at the monthly time scale for each grid cell (not routed), and as such the daily sequencing should not change the monthly results significantly.

## 5   Conclusions

Our analysis indicates the following:

1.  There is potential to improve monthly runoff and SM forecast skill beyond the IHC effect at lead-1 (and up to 3 months in a few cases) by exploiting MRWF skill. In general, the Great Plain regions, Midwest, parts of the southwestern US (sub-regions in Texas) and eastern
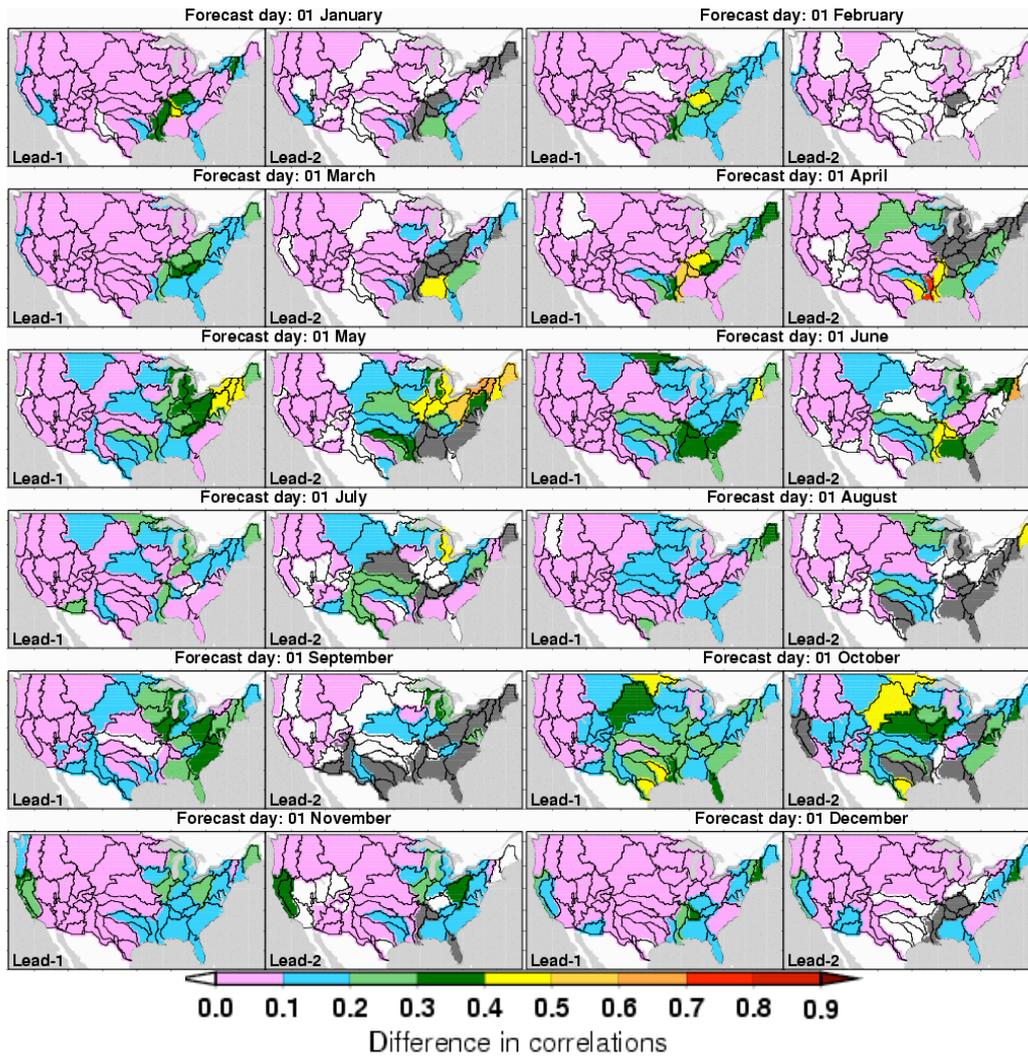
**Fig. 8.** Potential improvement in SM forecast skills at leads 1–2 months (dark grey color shows where the skill of OBS_Merged_ESP is not significant at the 95 % significance level).

US would benefit most during forecast periods starting in April through November. On the other hand, sub-regions in the mountainous western US would benefit most during forecast periods starting in November and the winter months (DJF).

2. The potential (and actual) improvement in runoff forecast skill, as contrasted with SM skill, is larger at lead-1, mostly due to high baseline skill for SM (i.e. stronger IHC effect in SM), whereas the improvement at lead-2 is more extensive for SM forecasts than for runoff.

3. Potential improvement in baseline skill for runoff forecasts generally varies from 0 to 0.8, whereas for SM it varies from 0 to 0.5. However, the space–time patterns of improvements are similar for runoff and SM.

4. The actual improvement in skill due to use of MRF forecasts is limited by modest forecast skill for precipitation. The ratio of actual skill to potential skill improvement generally varies from 0 to 0.8. Sub-regions in the Great Plains, Midwest, Texas, and northeastern and southeastern US could potentially benefit most from improvement in MRF skill during forecast periods starting in the summer months (JJA).

Our findings could have significant implications for the improvement of seasonal hydrologic predictions at short lead times (i.e. lead-1 to -3 months). Present protocols for generation of ensemble hydrologic forecasts from seasonal climate forecasts (e.g. Luo et al., 2007) make use of climate forecast ensembles that are generated through use of temporal offsets. The temporal offsets are mainly used to exploit predictability from different initial sea surface temperature
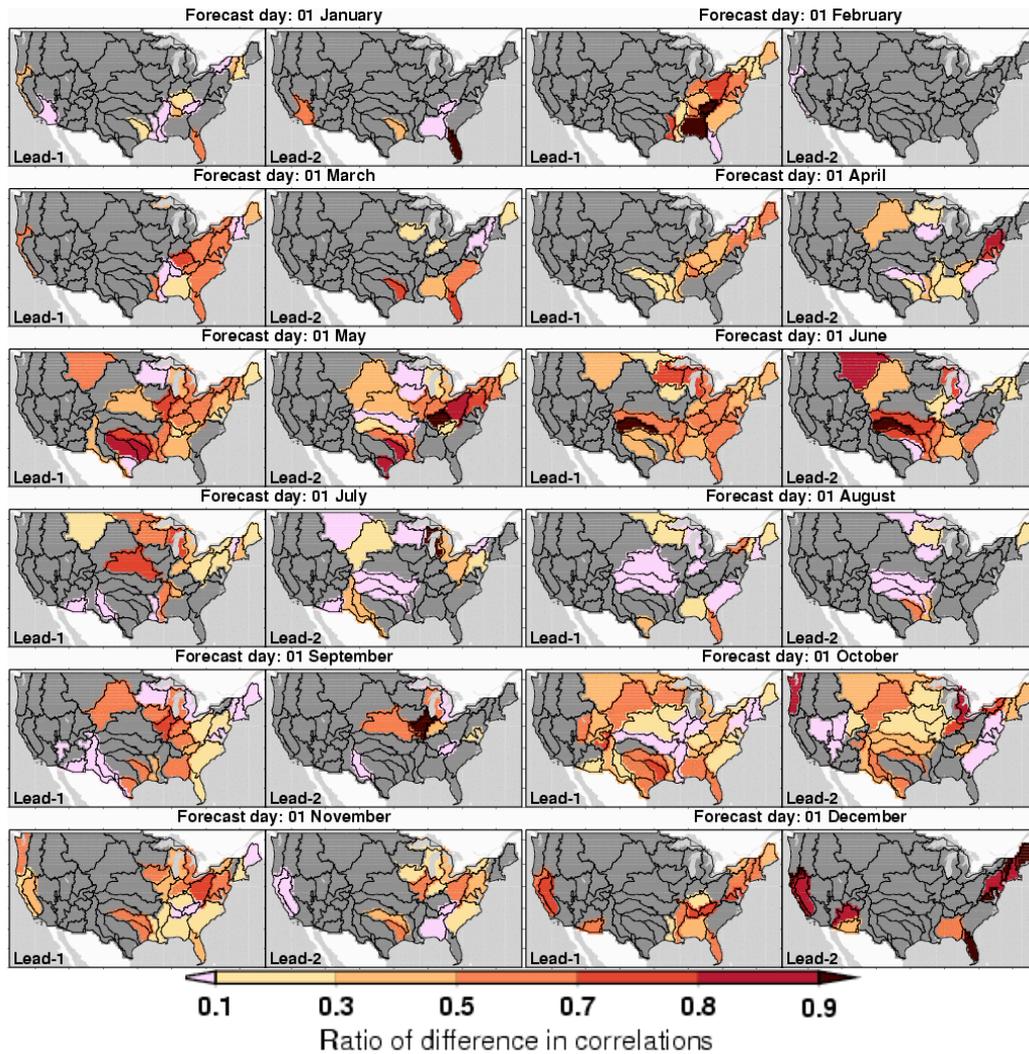
**Fig. 9.** The ratio of actual improvement and potential improvement in baseline SM forecast skill at leads 1–2 months (dark grey color shows the sub-regions where either the potential improvement in skill is < 0.1 or the skill of OBS_Merged_ESP is not significant at the 95 % significance level).

conditions. For example, real-time operational climate forecasts such as the International Research Institute's seasonal climate forecasts are generated using seven atmospheric global circulation models (forced by the predicted global tropical sea surface temperature). However, the forecast integration occurs 3–4 weeks in advance of the seasonal forecast period; hence, the models do not exploit the skill from the observed atmospheric initial conditions (as well as the land surface conditions) at the beginning of the forecast period (Barnston et al., 2010). Likewise, the Climate Forecast System's (Saha et al., 2006) real-time seasonal forecasts make use of initial conditions of the last 30 days. As a result, the effects of MRWFs at the beginning of the forecast period are not reflected in the seasonal climate forecasts. This could be resolved either by (a) use of shorter temporal offsets or (b)

merging deterministic weather forecasts for the first 14 days (or perhaps shorter, given that most forecast skill comes from the first 5 days or so) with seasonal climate model forecasts thereafter.

Finally, improvement in drought prediction skill at short lead times could potentially help with decisions that involve identification of regions with the potential for drought recovery. This often occurs over much shorter lead times than drought onset; hence, better use of weather forecasts could provide practical benefits in this arena as well.

# References

Andreadis, K. M., Storck, P., and Lettenmaier, D. P.: Modeling snow accumulation and ablation processes in forested environments, Water Resour. Res., 45, W05429, doi:10.1029/2008WR007042, 2009.

Barnston, A. G., Glantz, M. H., and He, Y.: Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–98 El Nino episode and the 1998 La Nina onset, B. Am. Meteorl. Soc., 80, 217–224, 1999.

Barnston, A. G., Li, S., Mason, S. J., DeWitt, D. G., Goddard, L., and Gong, X.: Verification of the first 11 years of IRI's seasonal climate forecasts, J. Appl. Meteorol. Clim., 49, 493–520, 2010.

Cherkauer, K. A., Bowling, L. C., and Lettenmaier, D. P.: Variable infiltration capacity cold land process model updates, Global Planet. Change, 38, 151–159, 2003.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake Shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, J. Hydrometeorol., 5, 243–262, 2004.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Sheffield, J., Duan, Q., Luo, L., Higgins, W. R., Pinker, R. T., and Tarpley, J. D.: Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS), J. Geophys. Res., 108, 8845, doi:10.1029/2002JD003316, 2003.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Res. Pl., 111, 157–170, 1985.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin, J. Hydrometeorol., 4, 1105–1118, 2003.

Gong, X., Barnston, A. G., and Ward, M. N.: The effect of spatial aggregation on the skill of seasonal precipitation forecasts, J. Clim., 16, 3059–3071, 2003.

Hamill, T. M., Whitaker, J. S., and Mullen, S. L.: Reforecasts: An important dataset for improving weather predictions, B. Am. Meteorl. Soc., 87, 33–46, 2006.

Hayes, M., Svoboda, M., Le Comte, D., Redmond, K., and Pasteris, P.: Drought monitoring: New tools for the 21st century, in: Drought and Water Crises: Science, edited by: Wihite, D. A., Technology, and Management Issues, 53–69, 2005.

Higgins, R. W., Leetmaa, A., Xue, Y., and Barnston, A.: Dominant factors influencing the seasonal predictability of US precipitation and surface air temperature, J. Clim., 13, 3994–4017, 2000.

Hou, D., Mitchell, K., Toth, Z., Lohmann, D., and Wei, H.: The Effect of Large-Scale Atmospheric Uncertainty on Streamflow Predictability, J. Hydrometeorol., 10, 717–733, 2009.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, 2003.

Kalnay, E. C., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, B. Am. Meteorol. Soc., 77, 437–472, 1996.

Lavers, D., Luo, L., and Wood, E. F.: A multiple model assessment of seasonal climate forecast skill for applications, Geophys. Res. Lett., 36, L23711, doi:10.1029/2009GL041365, 2009.

Lettenmaier, D. P. and Wood, E.: Water in a changing climate: Challenges for GEWEX, Gewex News, 19, 3–5, 2009.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, doi:10.1029/2008JD010969, 2009.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99, 14415–14428, doi:10.1029/94JD00483, 1994.

Liang, X., Wood, E. F., and Lettenmaier, D. P.: Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification, Global Planet. Change, 13, 195–206, 1996.

Luo, L., Robock, A., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Lohmann, D., Cosgrove, B., Wen, F., Sheffield, J., Duan, Q., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: Validation of the North American Land Data Assimilation System (NLDAS) retrospective forcing over the southern Great Plains, J. Geophys. Res., 108, 8843, doi:10.1029/2002JD003246, 2003.

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

Mahanama, S., Livneh, B., Koster, R., Lettenmaier, D., and Reichle, R.: Soil Moisture, Snow, and Seasonal Streamflow Forecasts in the United States, J. Hydrometeor, 13, 189–203, doi:10.1175/JHM-D-11-046.1, 2011.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States, J. Clim., 15, 3237–3251, 2002.

Mitchell, K., Houser, P., Wood, E., Schaake, J., Tarpley, D., Lettenmaier, D., Higgins, W., Marshall, C., Lohmann, D., Ek, M., Cosgrove, B., Entin, J., Duan, Q., Pinker, R., Robock, A., Habets, F., and Vinnikov, K.: GCIP land data assimilation system (LDAS) project now underway, Gewex News, 9, 3–6, 1999.

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Cosgrove, B. A., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., Lettenmaier, D. P., Marshall, C. H., Entin, J. K., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B. H., and Balley A. A.: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, J. Geophys. Res., 109, D07S90, doi:10.1029/2003JD003823, 2004.

O'Lenic, E. A., Unger, D. A., Halpert, M. S., and Pelman, K. S.: Developments in operational long-range climate prediction at CPC, Weather Forecast., 23, 496–515, 2008.

www.hydrol-earth-syst-sci.net/16/2825/2012/

Hydrol. Earth Syst. Sci., 16, 2825–2838, 2012

Pan, M., Sheffield, J., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., Robock, A., Lohmann, D., Cosgrove, B. A., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: Snow process modeling in the North American Land Data Assimilation System (NLDAS). Part II: Evaluation of model simulated snow water equivalent, J. Geophys. Res., 108, 8850, doi:10.1029/2003JD003994, 2003.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrol. Earth Syst. Sci., 9, 381–393, doi:10.5194/hess-9-381-2005, 2005.

Quan, X., Hoerling, M., Whitaker, J., Bates, G., and Xu, T.: Diagnosing Sources of U.S. Seasonal Forecast Skill, J. Clim., 19, 3279–3293, doi:10.1175/JCLI3789.1, 2006.

Ross, T. and Lott, N.: A climatology of 1980–2003 extreme weather and climate events, Tech. Rep. 300-1, National Climatic Data Center, 2003.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., Van den Dool, H., Pan, H. L., Moorthi, S., Behringer, D., Stokes, D., Pena, M.m Lord, S., White, G., Ebisuzaki, W., Peng, P., and Xie, P.: The NCEP climate forecast system, J. Clim., 19, 3483–3517, 2006.

Schaake, J. C., Duan, Q., Koren, V., Mitchell, K. E., Houser, P. R., Wood, E. F., Robock, A., Lettenmaier, D. P., Lohmann, D., Cosgrove, B., Sheffield, J., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS), J. Geophys. Res., 109, D01S90, doi:10.1029/2002JD003309, 2004.

Schubert, S., Koster, R., Hoerling, M., Seager, R., Lettenmaier, D., Kumar, A., and Gutzler, D.: Predicting drought on seasonal-to-decadal time scales, B. Am. Meteor. Soc., 88, 1625, doi:10.1175/BAMS-88-10-1625, 2007.

Sheffield, J., Pan, M., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., Robock, A., Lohmann, D., Cosgrove, B., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., and Ramsay, B. H.: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model-simulated snow cover extent, J. Geophys. Res., 108, 8849, doi:10.1029/2002JD003274, 2003.

Shepard, D. S.: Computer mapping: The SYMAP interpolation algorithm, Spatial Statistics and Models, 133–145, 1984.

Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529–3538, doi:10.5194/hess-15-3529-2011, 2011.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, doi:10.5194/hess-13-125-2009, 2009.

Todini, E.: The ARNO rainfall–runoff model, J. Hydrol., 175, 339–382, 1996.

Voisin, N., Schaake, J. C., and Lettenmaier, D. P.: Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts, Weather Forecast., 25, 1603–1627, 2010.

Voisin, N., Pappenberger, F., Lettenmaier, D. P., Buizza, R., and Schaake, J. C.: Application of a Medium-Range Global Hydrologic Probabilistic Forecast Scheme to the Ohio River Basin, Weather Forecast., 26, 425–446, 2011.

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resour. Res., 45, W05407, doi:10.1029/2008WR007355, 2009.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S.: Incorporating medium-range numerical weather model output into the ensemble streamflow prediction system of the National Weather Service, J. Hydrometeorol., 6, 101–114, 2005.

Wilks, D. S.: Diagnostic verification of the climate prediction center long-lead outlooks, 1995–98, J. Clim., 13, 2389–2403, 2000.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic press., New York, 54–55, 2006.

Wilks, D. S. and Godfrey, C. M.: Diagnostic verification of the IRI net assessment forecasts, 1997–2000, J. Clim., 15, 1369–1377, 2002.

Wood, A. W. and Lettenmaier, D. P.: A test bed for new seasonal hydrologic forecasting approaches in the western United States, B. Am. Meteorol. Soc., 87, 1699–1712, 2006.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, L14401, doi:10.1029/2008GL034648, 2008.

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

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, J. Geophys. Res., 110, D04105, doi:10.1029/2004JD004508, 2005.

Wu, L., Seo, D.-J., Demargne, J., Brown, J., Cong, S., and Schaake, J.: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast via meta-Gaussian distribution-based models, J. Hydrol., 399, 281–298, 2011.

Xia, Y., Mitchell, K. E., Ek, M. B., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C. J., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-Scale Water and Energy Flux Analysis and Validation for North-American Land Data Assimilation System Project Phase 2, Part 2: Validation of Model-Simulated Streamflow, J. Geophys. Res., 117, D03110, doi:10.1029/2011JD016051, 2011a.

Xia, Y., Mitchell, K. E., Ek, M. B., Sheffield, J., Cosgrove, B., Wood, E. F., Luo, L., Alonge, C. J., Wei, H., Meng, J., Livneh, B., Lettenmaier, D. P., Koren, V. I., Duan, Q., Mo, K. C., Fan, Y., and Mocko, D. M.: Continental-Scale Water and Energy Flux Analysis and Validation for the North-American Land Data Assimilation System Project Phase 2 (NLDAS-2), Part 1: Intercomparison and Application of Model Products, J. Geophys. Res., 117, D03109, doi:10.1029/2011JD016048, 2011b.

Zar, J. H.: Significance testing of the Spearman rank correlation coefficient, J. Am. Stat. Assoc., 67, 578–580, 1972.