Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States

LIFENG LUO

Environmental Engineering and Water Resources, Department of Civil and Environmental Engineering, and Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey

ERIC F. WOOD

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

(Manuscript received 12 October 2007, in final form 18 March 2008)

ABSTRACT

Skillful seasonal hydrologic predictions are useful in managing water resources, preparing for droughts and their impacts, energy planning, and many other related sectors. In this study, a seasonal hydrologic ensemble prediction system is developed and evaluated over the eastern United States, with a focus on the Ohio River basin. The system uses a hydrologic model (i.e., the Variable Infiltration Capacity model) as the central element for producing ensemble predictions of soil moisture, snow, and streamflow with lead times up to six months. One unique feature of this system is in the method for generating ensemble atmospheric forcings for the forecast period. It merges seasonal climate forecasts from multiple climate models with observed climatology in a Bayesian framework, such that the uncertainties related to the atmospheric forcings can be better quantified while the signals from individual models are combined. Simultaneously, climate model forecasts are downscaled to an appropriate spatial scale for hydrologic predictions. When generating daily meteorological forcing, the system uses the rank structures of selected historical forcing records to ensure reasonable weather patterns in space and time.

Seasonal hydrologic predictions are made with this system, using seasonal climate forecast from the NCEP Climate Forecast System (CFS), and from a combination of the NCEP CFS and seven climate models in the European Union’s Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (CFS/H11001 DEMETER). Forecasts of these two types are made for the summer periods (May to October) of 1981–99 and are compared to forecasts produced with the traditional Ensemble Streamflow Prediction (ESP) approach used in operational seasonal streamflow predictions. The forecasts from this system for the summer of 1988 show very promising skill in precipitation, soil moisture, and streamflow over the Ohio River basin, especially the multimodel CFS/H11001 DEMETER forecast. The evaluation with all 19 summer forecasts shows that the multimodel CFS/H11001 DEMETER forecast is significantly better than the ESP forecast during the first two months of the forecasts. The advantage is marginal to moderate when only the CFS forecast is used. This study validates the approach of using seasonal climate predictions from dynamic climate models in hydrological predictions, and it also emphasizes the need for international collaborations to develop multimodel seasonal predictions.

1. Introduction

Seasonal predictions of precipitation, soil moisture, streamflow, and other hydrologic variables have great economic value in sectors such as water resource management, reservoir operation, hydropower plants, and agriculture. Reliable predictions of watershed inflow several months in advance, for instance, can help a hydropower plant better determine its potential seasonal hydropower generation and improve its seasonal scheduling, thereby increasing its operating profits. Hamlet et al. (2002) estimated that the annual increase in revenue is about 40 million U.S. dollars (USD) on average if an operational strategy based on reliable seasonal forecast is implemented for the hydropower plants in
the Columbia River basin. On the other hand, drought conditions, in the form of a long-term deficit of soil moisture, have a significant impact on the economy—especially crop production. The annual cost of U.S. droughts has been estimated to be in the range of 6–8 billion USD (Witt 1995). Seasonal prediction of soil moisture, especially droughts, can provide guidance on planning agricultural production and water usage so as to minimize drought losses. Therefore, high-quality seasonal hydrologic predictions are greatly needed.

Studies over the last two decades have demonstrated the feasibility of seasonal climate predictions with dynamical climate models. The skill of seasonal predictions is believed to come from the slowly varying components of the climate system, mainly tropical Pacific sea surface temperature, although more recently, surface soil moisture has also shown certain contributions over transition zones between dry and wet climatic regions (Koster et al. 2000, 2004). At present, seasonal climate predictions are made routinely at several weather and climate prediction centers and research institutes, including the European Centre for Medium-Range Weather Forecasting (ECMWF) and in the United States, the National Centers for Environmental Prediction (NCEP) and the International Research Institute for Climate Prediction (IRI) at Columbia University. The predictions have shown significant skill over the tropics, whereas their skill is improving in the midlatitudes, with some models showing skill comparable to the skill from statistical models (Saha et al. 2006). There is the expectation that these seasonal dynamical climate forecasts can contribute to the development of seasonal hydrologic prediction capabilities. However, significant challenges must be overcome in using seasonal climate forecasts from dynamical climate models in a seasonal hydrologic prediction system. The first challenge is to correct the biases in climate model predictions, especially those related to precipitation and temperature. The second challenge is to resolve the disparity in spatial scales between the ones resolved in climate models and the ones needed for hydrologic applications. For instance, the current operational NCEP global coupled ocean–atmosphere model, the Climate Forecast System (CFS; Saha et al. 2006), runs at T62L64 resolution (∼1.875° in longitude). The climate models in the European Union (EU) DEMETER project (Palmer et al. 2004) provide hindcasts at a resolution of 2.5° × 2.5°. However, the hydrologic predictions need atmospheric forcing at a much finer resolution. As an example, the North America Land Data Assimilation System (NLDAS; Mitchell et al. 2004), which provides real-time hydrologic simulations across the continental United States, has adopted a spatial scale of 1/8°. These disparities require a seasonal hydrologic forecast system to spatially downscale the information provided by the climate models to the finer hydrologic scale so that the information can be properly used. The third challenge is to create realistic daily atmospheric forcing for hydrologic modeling from the monthly information provided by the climate models. Climate model forecasts are generally only available as a monthly forecast, whereas the hydrologic models are run at daily or sub-daily time steps. To make skillful seasonal hydrologic predictions, a good strategy is needed to overcome these challenges.

Wood et al. (2002) presented a strategy for making ensemble seasonal hydrologic predictions using ensemble seasonal forecasts from the NCEP Global Spectral Model (the name of the older version of the current global forecast system). They used an equal-quantile transfer method to correct the biases in monthly precipitation and 2-m air temperature at the climate model scale. The correction factors for the monthly forecasts, at the climate model spatial scale, are interpolated spatially to a finer scale for the hydrologic predictions. The temporal downscaling from monthly to daily is achieved by sampling from high-resolution historical daily time series and adjusting them to match the bias-corrected monthly forecast. As one of the first groups to produce seasonal hydrologic predictions using seasonal climate model forecasts, Wood et al. demonstrated some skill in predicting soil moisture and streamflow over the eastern United States. However, their methodology has a few limitations. First, such an approach implicitly assumes that the climate models are skillful in predicting precipitation and temperature. If the climate models lack skill in precipitation and temperature, which is true for the seasonal forecasts over the midlatitudes, the bias-corrected precipitation and temperature forcing do not necessarily represent the uncertainties in the forecasts. In an extreme case when a climate model only gives a random forecast, the approach will not remove the randomness, resulting in a random hydrologic prediction with no skill. Second, because all ensemble members of climate model forecasts are treated individually and equally, one member in the climate model forecast is used to produce one ensemble member in the hydrologic prediction. The number of hydrologic ensembles is thus constrained. Third, this approach does not provide a flexible framework for combining forecasts from multiple climate models. The multimodel ensemble concept has been demonstrated to have the potential for producing more reliable and skillful forecasts with better estimates of the uncertainties, so we need to take advantage of this concept in seasonal hydrologic predictions.
This study presents a seasonal hydrologic ensemble prediction system that implements new approaches to overcome the aforementioned challenges and difficulties. In section 2, we present the methodology behind the prediction system, including a Bayesian merging procedure for combining climate forecasts from multiple climate models that also allows for the spatial and temporal downscaling of merged atmospheric forcings. Forecast evaluations for selected cases are presented in section 3, followed by a discussion on the limits and possible future improvements of the forecast system in section 4.

2. Methodology

Figure 1 illustrates the basic structure of our seasonal hydrologic ensemble prediction system that also fits into the framework proposed by the Hydrological Ensemble Prediction Experiment (HEPEX) project (Franz et al. 2005). As proposed by HEPEX, a seasonal hydrologic ensemble prediction system should consist of at least four elements: (i) a land data assimilator (or preforecast processor), (ii) an atmospheric ensemble preprocessor, (iii) a hydrologic ensemble processor, and (iv) a product generator, although actual implementations of this framework can vary drastically in complexity. In this section, our implementation of these elements and their functionality are described. Attention should be given to the atmospheric ensemble preprocessor, for it is the main focus of this study. Descriptions of datasets used in developing and running the system can be found in the appendix.

2a. Preforecast processor

The goal of having a preforecast processor is to create suitable hydrologic initial conditions of the land surface for the prediction. Suitable means that the hydrologic initial conditions need to be as accurate as possible and consistent with the hydrologic model used in the forecast system. One general strategy is to force the hydrologic model with (accurate) observed atmospheric forcings for a period prior to the forecast—usually referred to as an observation-based spinup. In our seasonal prediction system, we spin up the Variable Infiltration Capacity (VIC) land surface model using observed atmospheric forcing for at least two years prior to each forecast. For hindcasts before 2000, we use the atmospheric forcing dataset described by Mauzer et al. (2002). In our real-time forecasts, the real-time atmospheric forcing is from the NLDAS (Mitchell et al. 2004; Cosgrove et al. 2003; Luo et al. 2003). Evaluations have shown that the two datasets provide equivalent forcings. The generated initial conditions are never perfectly accurate, and their uncertainties need to be incorporated into the prediction system. For seasonal hydrologic predictions over snow-free regions, the uncertainties in the atmospheric forcing are likely to be much larger than the uncertainties associated with initial conditions. Therefore, in our eastern U.S. application presented here, we ignore the uncertainties associated with initial conditions and provide only one initial condition for each forecast, keeping in mind that uncertainties in the initial condition can become important and need to be addressed for forecasts covering snowmelt seasons.

2b. Atmospheric ensemble preprocessor

In seasonal hydrologic prediction, the uncertainties associated with the atmospheric forcings are likely to be the major contributor to the total forecast uncertainty. Thus, the accuracy of the atmospheric forcings is essential to the skill of the hydrologic predictions. Because of the nature of the forecast problem, atmospheric forcings cannot be accurately predicted; instead, their uncertainties need to be accurately quantified via an ensemble approach. The function of the atmosphere ensemble preprocessor is to create an ensemble of atmospheric forcings that reflects our best estimate of future conditions and their uncertainties. In our forecast system, this is achieved in the following two steps.
1) Bayesian merging of climate model forecast with the observed climatology

This step estimates the probability of precipitation and temperature at a monthly time scale at each 1/8° grid, using information from dynamic climate model forecasts. This is achieved within the Bayesian framework introduced by Luo et al. (2007). A brief description about this method is provided here.

The basic principle of using Bayes’ theorem, as expressed by Eq. (1), in seasonal hydrologic prediction is to update the probability of the forecast variable \( \theta \) based on the forecast of \( y \) from climate models.

\[
p(\theta | y) = \frac{p(\theta)p(y|\theta)}{p(y)} \quad (1)
\]

In Eq. (1), \( p(\theta) \) on the right-hand side is the prior distribution of \( \theta \) that reflects our prior belief about \( \theta \) and is usually based on observed climatology. The conditional distribution \( p(y|\theta) \) on the right-hand side of Eq. (1) is the likelihood function that expresses the probabilistic relationship between \( y \) and \( \theta \) based on “observations” that can consist of model outputs and in-situ data. The distribution \( p(\theta|y) \) on the left-hand side of Eq. (1) is the posterior distribution, which is the updated probability of \( \theta \) given the knowledge about \( y \). In applying Bayes’ theorem in seasonal hydrologic prediction, we estimate the probability distribution of monthly precipitation/temperature at the spatial scale for hydrologic modeling, given the forecasts of monthly precipitation/temperature from dynamic climate models that are at a much coarser resolution. As described in Luo et al. (2007), the prior distribution is estimated as the climatological distribution from historical observations. In this study, the dataset provided by Maurer et al. (2002) is used to estimate the climatological distribution of monthly precipitation/temperature over each 1/8° grid within the forecast domain.

Because Bayes’ theorem does not limit the spatial and temporal scales of \( \theta \) and \( y \), it is possible to allow \( \theta \) and \( y \) to be variables at two different scales, thus effectively downscaling information statistically. This solves the disparity in spatial scales. In the context of seasonal hydrologic prediction, if \( \theta \) is the monthly precipitation over a 1/8° grid, as in our forecast system, then \( y \) can be the monthly precipitation from a climate model forecast over a collocated climate model grid. The likelihood function is modeled as a weighted linear regression between the two from hindcasts. To make the linear regression stable, the length of the hindcasts should be maximized as much as possible, with 15–20 yr being an acceptable minimum length. To make the linear regression meaningful, the forecast quantity \( y \) from climate models has to be skillful. Luo and Wood (2006) discuss the idealized predictability of precipitation and temperature in the NCEP CFS, and one of the conclusions is that the model forecast skill decreases with lead time, even when predicting the model itself. Perhaps the only way to keep the same level of forecast skill is to increase the spatial and temporal scale of \( y \) as lead time increases. We have implicitly included this concept in our prediction system, by letting \( y \) be the spatially averaged precipitation/temperature forecast from the model, and the average is over a larger area when lead time increases.

After the prior distribution and the likelihood function have been estimated, the posterior distribution can be derived. In our Bayesian merging approach, both the prior and the likelihood function are assumed to follow normal distributions. Although this makes the mathematical derivation straightforward and produces a normal posterior distribution, the assumption may not be satisfied for seasonal prediction variables such as precipitation. Daily precipitation over a small region tends to not follow a normal distribution nor does monthly precipitation because the daily distribution tends to have a large mass at zero precipitation, and both daily and monthly distributions are positively skewed. Although the general Bayesian concept for merging information, as expressed by Eq. (1), still holds, the mathematical derivation of the posterior distribution becomes more difficult. In this case, when neither \( \theta \) nor \( y \) is normally distributed, an equal-quantile transformation scheme, as introduced by Luo et al. (2007), is used to convert the nonnormal distribution of precipitation to a standard normal distribution before the Bayesian merging. The same method was used by Wood et al. (2002) in their bias correction scheme. As shown by Luo et al. (2007), the posterior distribution is derived as a normal distribution in the transformed standard normal space. If climate forecasts from multiple climate models are to be used, this update can be done sequentially to form a multimodel posterior distribution with the assumption that these models provide independent pieces of information. For precipitation, the posterior distribution in the standard normal space is transferred back to the original space using the same equal-percentile transformation. The outputs from the Bayesian merging are probability density functions that express the distributions of monthly precipitation/temperature for each 1/8° grid and for each month during the six-month forecast period.

The Bayesian merging approach efficiently takes care of several problems when using climate model forecasts in a hydrologic prediction system. These include the following: (i) possible biases in climate model.
forecasts are removed during the merging, (ii) the disparity in spatial scales is resolved through the likelihood function that links the two scales, and (iii) the forecasts from multiple climate models can be combined in a consistent framework, and the weights of each model are dynamically assigned based on their hindcast performance. Hence, the strengths of all models are combined to form a posterior distribution that is superior to any raw model forecast (Luo et al. 2007). The forecast skills of climate models are explicitly considered when using their forecast information; therefore, at long lead times when climate models cannot make skillful predictions, the posterior distribution from the Bayesian merging will not be much different from the prior distribution, so the prediction automatically falls back to the observed climatology because the dynamical model forecasts are given near-zero weights in the merging.

2) Creating ensemble atmospheric forcing from historical data

The probability distributions obtained from the first step need to be sampled to produce individual members to form a hydrological ensemble forecast. In principle, an unlimited number of samples can be drawn from each distribution because the distributions are continuous. In practice, we only take 20 samples from each distribution for our hydrologic prediction, which we believe to be sufficient and practical. Because seasonal distribution for our hydrologic prediction, which we be-

annual number of samples can be drawn from each grid-based distribution then associated in the first step. Therefore, instead of randomly sampling from each grid-based distribution then associating the samples in space and time and across variables, the approach predefined the associated relationship using the observed patterns from a specific historical period over the forecast region. With this strategy, we only need to decide which historical period to select from the entire historical dataset, with the constraint that only the same months as in the forecast period are considered as candidates. This eliminates the influence of seasonality on the precipitation–temperature statistics. For example, if our forecast period is from May to October, then the candidates are observations over the forecast region from May to October of all available years.

The 20 historical forcing time series are selected in a hybrid fashion; that is, half of the ensemble members are selected using a historical-analog criterion based on precipitation and the other half are randomly selected. In the historical-analog criterion, we compare each historical pattern (both in space and time) with the predicted pattern and look at their similarities. If we denote \( P_{i,j} \) as the mean of the predicted precipitation at grid \( i (i = 1, N, \text{where } N \text{ is the total number of } 1/8^\circ \text{ grids in a forecast region}) \) and month \( j (j = 1, 6) \) and denote \( H_{i,j,k} \) as the observed precipitation from year \( k \) for the same grid \( i \) and same month \( j \), the similarity between each historical forcing time series and the predicted precipitation anomaly can be expressed as

\[
S_k = \sum_{j=1}^{6} w_j \sum_{i=1}^{N} (H_{i,j,k} - P_{i,j})^2, \tag{2}
\]

where \( w_j \) is the weight assigned to month \( j \) that decreases with lead time, so that patterns with shorter lead time receive higher weights in the selection. Then historical years are sorted according to their \( S_k \) values, and the forcings from the 10 historical years with the smallest \( S_k \) values are selected, with the smallest \( S_k \) being the first member. Although simple and empirical, this selection criterion considers the similarity in spatial and temporal patterns in monthly precipitation anomalies, so it can provide more suitable forcing under the same climatic conditions. However, because all members are similar to the predicted pattern, the ensemble members can cluster too closely potentially reducing the ensemble spread, which then leads to the underestimation of uncertainties. The random selection of the other 10 members acts to compensate for this and adds back uncertainties. The combination of these two selection methods will have a better handle on both the mean and uncertainties of the atmospheric forcing.
The precipitation and temperature fields in the selected 20 atmospheric forcing time series are then adjusted according to their respective climatological distribution and the forecast distribution. As illustrated in Fig. 2, the adjusted precipitation will have the same percentile value in the forecast distribution as the historical forcing time series in the climatological distribution at the monthly level. For example, if the monthly precipitation in the selected observational record at one grid is at the 90th percentile of its climatological distribution, then the actual monthly precipitation used for the forecast is the value that will be the 90th percentile in the forecast distribution. Applying the same principle on both precipitation and temperature for all grids every month, it is easy to see that the rank structure of historical observations is kept, not their actual values. If the forecast distribution is the same as the observed climatological distribution, this method will make no adjustment to the historical record. In this case, the forecast falls back to the traditional Ensemble Streamflow Prediction (ESP) method (Day 1985). In most cases, adjustments are needed, especially for the first few months of the forecast because the forecast distributions of precipitation and temperature are likely to be different from their climatological distributions. At each grid, when adjusting the historical record, precipitation is scaled up (down) and temperature is shifted up (down) to match the forecast monthly values. The same scaling factor (shifting factor) is applied throughout the month on each daily value to ensure consistency.

c. Hydrologic ensemble processor

The hydrologic ensemble processor transfers the ensemble meteorological information (precipitation and temperature in our system) to land surface hydrologic states such as soil moisture, snow extent and water equivalent, and streamflow. In our prediction system, this is achieved using the VIC land surface model (Liang et al. 1996; Cherkauer et al. 2003) and the simple linear routing model developed by Lohmann et al. (2004).

The VIC land surface model is a semidistributed grid-based hydrologic model with distinguishing hydrologic features such as the representation of subgrid variability in soil storage capacity as a spatial probability distribution, elevation bands to handle orographic precipitation and snow, and partial grid coverage by precipitation, among other features. It has been widely applied to large continental river basins and regions (e.g., see Nijssen et al. 1997; Maurer et al. 2002; Mitchell et al. 2004; Sheffield et al. 2004) and on global scales (Nijssen et al. 2001; Sheffield and Wood 2007). Its ability to simulate hydrologic processes has been well recognized. The VIC model can be run in full energy balance mode at subdaily time steps as well as in a simple water balance mode with minimum inputs of daily total pre-
cipitation, daily minimum and maximum temperature, and daily average wind speed. For the purpose of seasonal hydrologic prediction, running VIC in the water balance mode with daily forcing is sufficient and efficient.

The routing model used for this study is essentially the same as used in the NLDAS project (Lohmann et al. 2004). The distributed routing model (Lohmann et al. 1998) calculates the timing of the runoff reaching the outlet of a grid box as well as the transport of the water through the river network. The river flow direction velocity masks used in this study are the same as used in the NLDAS project. Readers are referred to Lohmann et al. (2004) for a detailed description.

Once streamflow is produced by the routing model, two necessary adjustments are made. The first adjustment is to scale the streamflow according to the ratio between the observed drainage area and the drainage area used in the routing model for the basin. The drainage area of a basin, calculated by the routing models, is often slightly different from the observed value, and the difference can be significant enough to cause bias in the modeled streamflow. The adjustment is simply to correct for this area bias. The other adjustment is to correct for bias as a result of the insufficient calibration of the VIC land surface model and the routing model. It is achieved again through the equal-quantile transformation method, so that the distribution of the long-term streamflow simulation at each basin outlet, or gauging station, matches the observed distribution.

d. Product generator

When the ensemble hydrologic predictions of soil moisture, snow properties, and streamflow are produced, it is up to the product generator to postprocess and interpret the forecast. Users will likely have different information needs from a forecast. For example, some might be interested in the total volume of river discharge from a basin over a season, whereas others might want to know the possible peak flow (or low flow) during the period and its probability. Based on users’ needs, different forecast products can be derived from the ensemble forecast. Both deterministic and probabilistic forecasts can be issued based on the ensemble forecast, but caution is needed when interpreting the results.

Currently, the following forecasts are issued in our forecast system: mean of the ensemble and the probability of each tercile (above normal, normal, and below normal) for monthly precipitation, soil moisture and streamflow, and the exceedance probability of monthly streamflow for selected thresholds.

3. Evaluation

To evaluate the performance of the seasonal hydrologic prediction system—in particular, to evaluate the improvement in forecast skill relative to the current forecast method at river forecast centers based on the traditional ESP approach (Day 1985)—we carried out hindcasts over selected basins in the eastern United States (Fig. 3). Our analysis focuses on the summer season, for example, forecasts were initialized on 1 May and spanned six months, from May to October. Hindcasts from the NCEP CFS and seven climate models in the European Union DEMETER project (Palmer et al. 2004) are available for the summer season for 1981-1999. This multimodel seasonal climate forecast dataset provides a unique opportunity for testing the Bayesian merging approach implemented in our hydrologic forecast system and for evaluating the hydrologic forecast skill with this new hydrologic forecast approach.

For each of these 19 forecast periods, two types of forecasts were produced with the Bayesian merging forecast approach: (i) only using seasonal forecast from
the NCEP CFS and (ii) using seasonal forecasts from the NCEP Climate Forecast System and seven DEMETER models (CFS+DEMETER). In addition, the ESP forecast is produced by simply using 20 randomly selected historical atmospheric forcing time series to drive the hydrologic model. Table 1 briefly lists the different characteristics among these three types of forecasts. The major differences are in the way they obtain the atmospheric forcing distribution and create daily atmospheric forcing from these distributions. The ESP forecast does not use any forecast information from the dynamical climate models, whereas the CFS and CFS+DEMETER forecasts follow the proposed forecast approach described earlier in this study. The ESP forecast serves as the baseline for the comparison because it is the forecast approach used operationally at river forecast centers. If the proposed forecast approach is superior to the ESP method, then we should see better forecast skill from CFS and/or CFS+DEMETER than from ESP. As we evaluate the new forecast approach in a hindcast setting, the target year along with two random years are removed from the training data for building the likelihood function. Although this reduces the sample size in the linear regression, it removes the impact of the target year hindcast on the regression, which is necessary to ensure the independence between the evaluation of the hindcast and the training data for producing the hindcast. This procedure is not necessary for real-time forecast.

The evaluation of seasonal hydrologic predictions focuses on precipitation, soil moisture, and streamflow. Precipitation forecasts are verified against the Maurer et al. (2002) in situ dataset, and streamflow forecasts are verified against observations reported by the U.S. Geological Survey (USGS). Although some long-term in situ soil moisture data are available for parts of the United States (Robock et al. 2000), they do not provide sufficient spatial and temporal coverage for forecast verification. Therefore, a retrospective offline simulation was performed with the VIC model, using atmospheric forcing from Maurer et al. (2002) to create the model simulated soil moisture dataset in a manner described in Sheffield et al. (2004). As shown in Fig. 4, the streamflow from this offline simulation agrees well with observations, which infers that the variation of the simulated soil moisture is reasonably represented. More importantly, in the context of seasonal hydrological prediction, it is the soil moisture anomalies that provide important forecast information rather than absolute model or observed values. Thus, we refer to the model-simulated soil moisture fields, based on observationally based forcings, as observations in the remaining portions of this paper.

With these ensemble predictions, the ensemble mean (or ensemble median) can be verified as a single-valued deterministic forecast, whereas the complete ensemble set can be verified probabilistically. Verification metrics can then be chosen accordingly, as shown in the following examples.

a. Forecasts for summer 1988

One of the goals for seasonal hydrologic predictions is to predict climate extremes such as drought with reasonable accuracy several months in advance. The successful prediction of the onset, development, and recovery of droughts is very valuable for drought preparation and mitigation. The summer of 1988 was very dry over the northern part of United States. The severe drought conditions lasted more than five months, causing significant damages to agriculture and the local economy (American Meteorological Society 1997). Here, we use the the summer of 1988 forecasts for over the Ohio River basin as an example to illustrate the skill of our prediction system.

Figure 5 shows the month 1 (May 1988) precipitation forecast (ensemble mean for each model as a single-valued deterministic prediction) from the NCEP CFS and seven DEMETER models. The original model
forecasts are shown at their original grid without any bias correction and spatial downscaling. Clearly, these climate models do not agree with each other on their predictions, and some of the predictions are significantly different from observations. For instance, the NCEP CFS predicts more precipitation over the southern portion of the basin, whereas two DEMETER models predict less precipitation across the region, with a dry spot over the southern part of the region. Figure 5 also shows the precipitation from the CFS/H11001 DEMETER forecast, in which precipitation forecasts from the NCEP CFS and seven DEMETER models pass through the atmospheric ensemble preproces-
sor as described in section 2, and therefore it has been bias corrected and downscaled. This precipitation field is much more similar to the observational field, especially in magnitude, suggesting a more accurate precipitation forcing for the hydrological modeling in the CFS+DEMETER forecast system.

Instead of comparisons with the original climate model forecasts, Fig. 6 presents the differences in precipitation forecasts from the three approaches, CFS, CFS+DEMETER, and ESP. During the first month of the forecast period (May 1988), all three forecasts predict the dry anomaly. However, the pattern and magnitudes of the anomaly are different. ESP is expected to
provide only climatology because it randomly selects historical forcings, but a weak anomaly can appear as a result of an insufficient number of ensemble members. Both CFS and CFS+DEMETER forecasts predict the dry anomaly, and the CFS+DEMETER forecast gives a much stronger signal. As lead time increases, the CFS+DEMETER forecast seems to be a little closer to the observation than the others, but the differences are getting smaller and all three forecasts are approaching the climatology. This suggests, and it supports other research (e.g., Barnston et al. 1994; Kumar and Hoerling 1995; Rowell 1998; Luo and Wood 2006), that seasonal climate models are not skillful in predicting precipitation at longer lead times. Because the weights given to these model forecasts are close to zero at the longer lead times, the posterior distribution from the Bayesian merging hardly differs from the prior distribution, that is, the climatological precipitation distribution. In this case, all three forecasts will virtually give the same answer.

With this predicted precipitation forcing, it is not surprising to see that CFS+DEMETER gives the best soil moisture (Fig. 7) and streamflow predictions (Fig. 8) among the three approaches in this case. The soil moisture initial condition has some impact on the soil moisture evolution and, consequently, streamflow predictions. At the beginning of May 1988, much of the region was already in drought, so all three approaches predicted dry anomaly in soil moisture, as indicated by the low percentile values of the ensemble mean. However, the development of the drought is predicted differently mainly as a result of the difference in precipitation forecasts. The drought as predicted by CFS+DEMETER persists much longer (to the end of the forecast period), whereas the drought almost recovers in June according to the ESP forecast. In Fig. 8, we compare the streamflow forecast distributions from the three approaches with observations and the offline simulations. The locations of these stream gauges are indicated in Fig. 3. As discussed earlier, the simulated streamflow agrees well with observations, but there are periods with significant differences that may arise from water management, especially reservoir operations. One way to avoid these complicating effects is to verify the seasonal hydrologic predictions against offline streamflow simulated with our best-known, in situ forcing data. An analysis of Fig. 8 leads to similar results, as does the precipitation and soil moisture comparisons but here we look at the whole ensemble instead of the ensemble mean. The box-and-whisker plots summarize the ensemble distribution of each forecast, and the ensemble spread tells how much confidence we should give to the mean forecast. Ideally, we would like a skillful forecast to have a narrow distribution and to have the observation fall near the middle of the forecast distribution. As shown in Fig. 8, CFS and CFS+DEMETER forecasts are shifted toward low flow conditions for most of the months at every gauge during summer 1998. In particular, the CFS+DEMETER forecast has much narrower distributions and the observations are within the distributions in most cases. The differences among these three forecast decrease with the increase in lead time, and in the end, they are all closely centered on the streamflow climatology.

b. Evaluation over the 19-yr hindcast period

The forecasts for summer 1988 show encouraging improvements in forecast skills from the ESP approach to the CFS forecast, then further to the CFS+DEMETER
Fig. 6. Comparison of precipitation forecasts (ensemble mean of monthly average precipitation expressed as an anomaly in mm day$^{-1}$) from (left column) observations and (right columns) three forecast approaches. The rows are forecasts at different lead times.
FIG. 7. Comparison of soil moisture forecasts (ensemble mean of monthly average precipitation expressed as the percentile value within the climatological distribution) from (right three columns) three forecast approaches and observations. (left column) The retrospective offline simulation of VIC, driven by the Maurer et al. (2002) dataset that serves as a proxy for observations.
approach. This section summarizes the hindcast skill, with selected metrics using all forecasts during the 19-yr period.

Figure 9 summarizes the comparison between predicted precipitation (ensemble mean) and observations for the first three months for all 19 forecast periods over every grid within the Ohio River basin. Overall, the CFS+DEMETER forecast shows skill in precipitation forecast for the first month, with the skill decreasing quickly in the second and subsequent months. Although the CFS forecast shows some skill in the 1988 forecast, its overall performance is probably only marginally better than ESP. Consequently, the soil moisture predictions agree well with observations during the first month because of the persistence of initial conditions, and they start to differ in the second month in the CFS and ESP forecasts (Fig. 10). The CFS+DEMETER forecast has its skills prolonged to the second month and probably some to the third month. The last three months of the forecast (August, September, and October) do not show significant skill and hence are not shown here. This comparison suggests that seasonal cli-
mate forecasts from some of the DEMETER models provide useful additional information to the information already provided by the NCEP CFS seasonal climate forecast; hence, the posterior distribution from the Bayesian merging procedure tends to be more skillful.

Streamflow forecasts are evaluated as probabilistic forecasts, with the ranked probability score (RPS; Epstein 1969; Murphy 1971). RPS is the sum of the squared difference between the cumulative forecast and observation vectors in multiple categories. A perfect probabilistic forecast with this measure is 0 and the worst score is $J/J$, where $J$ is the number of categories in consideration. For monthly streamflow at each gauge, we divide the historical distribution from the offline simulation into three categories, denoted as above normal, normal, and below normal, with each category having a probability of 1/3. The RPS is then calculated for each type of the three forecasts and should range between 0 and 2. Figure 11 compares the ranked probability scores averaged over 19 periods from all three forecast types for the first three months at selected USGS gauging stations. During the first month of all 19 forecasts, ESP has the largest average RPS value, indicating that it performs the worst in providing probabilistic forecast among the three, and CFS+DEMETER gives a much better forecast with a much smaller average RPS value. This is because the CFS+DEMETER forecast places the entire ensemble distribution more closely around the target observation, as seen in Fig. 8. The CFS forecast is generally better than ESP but not as skillful as CFS+DEMETER. As lead time increases, the difference in skill between CFS and ESP becomes less noticeable, but the CFS+DEMETER forecast still appears to be better than the other two. This is true with many gauging stations when lead time is longer than three months (not shown here).

4. Conclusions and discussion

We presented a newly developed seasonal hydrologic prediction system that is based on the VIC model and
seasonal forecasts from multiple dynamic seasonal climate models. The performance of the forecast system is evaluated against the traditional ESP forecast approach that is currently used by U.S. river forecast centers. The system uses the major development and success in land surface hydrological modeling from the North American Land Data Assimilation System (NLDAS; Mitchell et al. 2004) that provided real-time and retrospective forcing data and from advances in seasonal prediction from dynamic climate models.

The key challenge in seasonal hydrological prediction is the accurate quantification of the uncertainties associated with precipitation and other forcing variables at spatial and temporal scales relevant to catchment hydrology. We took on this challenge by implementing a Bayesian merging approach for the atmospheric ensemble preprocessor that provides the atmospheric forcings to the hydrologic model. The major advantage of the Bayesian merging approach, as illustrated in this study and by Luo et al. (2007), is that it combines information from multiple sources and merges them to form the best estimate of future conditions including the mean and its uncertainties. As climate models are developed more or less independently, their forecast skills vary geographically and by season. The Bayesian merging approach evaluates the skills of these models and combines their forecasts based on their hindcast performance, so that the strengths from each model are combined. Simultaneously, the information from the low-resolution seasonal climate model forecast is downscaled statistically during the merging to the smaller scale that is appropriate for hydrological predictions.

As shown in Luo et al. (2007) and in this study, the Bayesian merging of multiple climate model forecasts provides better predictions of future conditions in both the mean forecast and the associated distribution, which measures its uncertainty. In this study, forecasts from the NCEP CFS and seven DEMETER models are used for the hindcast over the Ohio River basin. The evaluation of the hindcast shows that better forecast skills in soil moisture and streamflow predictions are achieved through using seasonal forecasts from multiple climate models. This emphasizes the need for a multiple-models platform for real-time seasonal hydrological predictions. Therefore, it is important for the

![Fig. 10. Same as Fig. 9, but for total soil moisture (anomaly normalized by the climatological mean).](http://journals.ametsoc.org/jhm/article-pdf/9/5/866/4164564/2008jhm980_1.pdf)
The current study does not address the uncertainties associated with the hydrologic model or the initial hydrologic conditions (e.g., soil moisture and snow extent). For the latter variable, problems can arise over mountainous regions when snow starts to melt during the forecast period. The impact on hydrologic prediction skill as a result of uncertainties in the hydrologic model, model calibration, and initial hydrologic conditions will be addressed in a future paper.

Acknowledgments. This research was supported by NOAA’s Climate Prediction Program for the Americas hydrometeorological community to have seasonal predictions from multiple climate models available on a regular basis from operational centers. An additional avenue of research is the possibility of merging predictions from dynamic climate models with predictions from statistical models such as the canonical correlation analysis (CCA) and the optimal climate normal (OCN) method, which are used operationally at NOAA’s Climate Prediction Center for seasonal outlook. This will likely enable us to further improve the estimates of future conditions by pooling together more independent information.
of precipitation and satellite retrievals of the surface observations including rain gauges and radar observations. This hourly forcing is a combination of Eta Data Assimilation System and observations, which is different from the reanalysis dataset. The temperature and precipitation in the dataset are gridded product derived directly from station observations, which is different from the reanalysis that is normally used as observations in many studies. The dataset is aggregated at daily and monthly time scales. The monthly values are further aggregated to the climate model grid. In this study, this dataset serves two purposes: (i) the daily 1/8° observation is used as the base for spatial and temporal downscaling during the forecast period, and (ii) the land surface states from a VIC offline simulation forced by this daily forcing dataset serve as a proxy of observations because large-scale continuous observations of soil moisture do not exist.

b. NLDAS real-time forcing

The North American Land Data Assimilation System (NLDAS) is a multi-institutional project, with the goal of providing better land surface initial conditions to operational weather and climate forecasts. One important contribution of NLDAS to the community is the creation and archive of a good-quality near-real-time atmospheric forcing at 1/8° resolution for the conterminous United States (Mitchell et al. 2004; Cosgrove et al. 2003; Luo et al. 2003). This hourly forcing is a combination of Eta Data Assimilation System and observations including rain gauges and radar observations of precipitation and satellite retrievals of the surface shortwave radiation. In our seasonal hydrologic forecast system, the real-time NLDAS forcing is used to drive the VIC hydrologic model to produce the hydrologic initial conditions for near-real-time forecasts.

c. NCEP Climate Forecast System and its hindcast

The NCEP Climate Forecast System (CFS) is a fully coupled ocean–land–atmosphere dynamical seasonal prediction system, and it is the current operational system used in NCEP for seasonal prediction. It consists of an interactive ocean component, Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model, version 3 (MOM3; Pacanowski and Griffies 1998), and an atmospheric component that is the same as the one used for operational weather forecasting at NCEP, without additional tuning for climate applications except for having a coarser resolution (T62L64). The atmospheric and oceanic components are coupled and exchange daily averaged quantities, such as heat and momentum fluxes once a day, with no flux correction. Full interaction between these two components is confined to 66°S to 50°N, because of the difference in latitudinal domain.

In real time, the CFS is run twice a day from 0000 UTC and 1200 UTC initial conditions for the atmosphere and ocean of seven days ago (Saha et al. 2006). Atmospheric initial conditions are obtained from NCEP–Department of Energy Global Reanalysis 2 (Kanamitsu et al. 2002) and the oceanic initial conditions are obtained from the NCEP global ocean data assimilation system (GODAS). The forecast period covers the first partial month and nine full months into the future. With this setup, there are about 60 members available for each nine-month forecast period. Besides the daily forecasts, a set of fully coupled retrospective predictions spanning the period of 1981–2004, with 15 members per month out to nine months into the future, has been produced with the CFS (Saha et al. 2006). These retrospective forecasts are important for bias correction and calibration of the operational seasonal prediction. More details about CFS and its forecast products can be found in Saha et al. (2006) and on NCEP’s Web sites. In this study, the 24-yr hindcasts are used to build the correlation model in the Bayesian merging then individual hindcasts are used to evaluate the performance of the hydrologic prediction system.

d. European Union DEMETER seasonal forecast

DEMETER is the acronym for the EU-funded project Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction. The objective of the project is to develop a well-
validated European coupled multimodel ensemble forecast system for reliable seasonal-to-interannual prediction (Palmer et al. 2004). Seven atmospheric general circulation models (GCMs) from operational and research institutes in Europe participated in the project. These atmospheric GCMs were coupled with an ocean model to produce a multimodel seasonal forecast six months into the future with nine ensemble members from each model. A hindcast dataset was created with the DEMETER system and is available from the European Centre for Medium-Range Weather Forecasts (ECMWF). The hindcasts start from the first day of February, May, August, and November initial conditions. More information about DEMETER can be found in Palmer et al. (2004) and on the ECMWF Web site. In this study, the 19 hindcasts (1981–1999) are used to build the correlation model in the Bayesian merging. Then individual hindcasts are used to evaluate the performance of the hydrologic prediction system.

e. Streamflow data from the U.S. Geological Survey

Long-term observations of streamflow from selected gauges in the Ohio River basin are obtained from the U.S. Geological Survey. Similar data were used to calibrate hydrologic models including the VIC model over large basins in the United States (Maurer et al. 2002). Although only four gauges are shown in this study, data from more stations are used by the forecast system for the following purposes: 1) the long-term observations are used to construct the observed streamflow climatological distribution, which is further used to adjust (postprocessing) streamflow predictions from the routing model; and 2) the adjustment is based on the equal-quantile transformation method, which takes care of possible bias in the streamflow forecast when calibration is not sufficient. The streamflow observations are used to evaluate the streamflow forecast from the prediction system.

REFERENCES


—, R. Schnur, and D. P. Lettenmaier, 2001: Global retrospec-


