A Multiscale Ensemble Filtering System for Hydrologic Data Assimilation. Part II: Application to Land Surface Modeling with Satellite Rainfall Forcing

MING PAN AND ERIC F. WOOD

Princeton University, Princeton, New Jersey

(Manuscript received 2 February 2009, in final form 12 May 2009)

ABSTRACT

Part I of this series of studies developed procedures to implement the multiscale filtering algorithm for land surface hydrology and performed assimilation experiments with rainfall ensembles from a climate model. However, a most important application of the multiscale technique is to assimilate satellite-based remote sensing observations into a land surface model—and this has not been realized. This paper focuses on enabling the multiscale assimilation system to use remotely sensed precipitation data. The major challenge is the generation of a rainfall ensemble given one satellite rainfall map. An acceptable rainfall ensemble must contain a proper multiscale spatial correlation structure, and each ensemble member presents a realistic rainfall process in both space and time. A pattern-based sampling approach is proposed, in which random samples are drawn from a historical rainfall database according to the pattern of the satellite rainfall and then a cumulative distribution function matching procedure is applied to ensure the proper statistics for the pixel-level rainfall intensity. The assimilation system is applied using Tropical Rainfall Measuring Mission real-time satellite rainfall over the Red–Arkansas River basin. Results show that the ensembles so generated satisfy the requirements for spatial correlation and realism and the multiscale assimilation works reasonably well. A number of limitations also exist in applying this generation method, mainly stemming from the high dimensionality of the problem and the lack of historical records.

1. Introduction

Addressing continental-to-global-scale hydrologic problems that are at the core of the World Climate Research Programme (WCRP) Global Energy and Water Cycle Experiment (GEWEX; Schaake 1994) and the National Aeronautics and Space Administration (NASA) Energy and Water Cycle Study (NEWS; Houser and Entin 2006) requires the use of remote sensing data (McCabe et al. 2008; Rodell et al. 2004). However, remote sensing observations are insufficient to understand the spatial and temporal mean and variability in the water cycle variables because of varying sensor spatial scales, different observing times for different variables, and remote sensing retrieval errors. This results in inconsistencies in the water and energy cycle budgets when estimated solely using remote sensing retrievals (McCabe et al. 2008; Pan et al. 2008). Although land surface models (LSMs) have consistency by construct, they require inputs that are often unavailable in many areas of the globe (e.g., Africa and the Arctic); therefore, there is a need to merge at continental scales remote sensing retrievals into land surface models.

There has been considerable work in assimilating remote sensing data into land surface models (e.g., Crow and Wood 2003; Reichle and Koster 2005; Margulis et al. 2006; Slater and Clark 2006). To date, the assimilation of actual remote sensing retrievals have been done grid by grid, ignoring the spatial structure in the retrieval errors. Pan et al. (2009) developed a multiscale data assimilation system to dynamically merge continental-scale remotely sensed retrieved variables into an LSM. The focus of Pan et al. (2009) is the implementation of the multiscale filtering algorithm (Zhou et al. 2008; Zhou 2006; Frakt and Willsky 2001; Willsky 2002) for applications in land surface hydrology by developing the necessary techniques to automatically build a proper multiscale tree that is 1) suitable for an arbitrary computing grid, 2) topologically balanced, and 3) structurally efficient for assimilating measurements at different resolutions. The goal of Pan et al. (2009) is primarily to

Corresponding author address: Ming Pan, Dept. of Civil and Environmental Engineering, Princeton University, EQuad, Olden St., Princeton, NJ 08544.
E-mail: mpan@princeton.edu

DOI: 10.1175/2009JHM1155.1

prove the concept of multiscale assimilation for land surface hydrology—that is, its feasibility, applicability, and strength. Pan et al. (2009) used precipitation ensembles from a climate model to force the land surface model, which worked well to test the multiscale ensemble assimilation; however,the feasibility of implementing such a system that relies on remotely sensed precipitation has not been investigated. Here in this paper, we use remotely sensed precipitation retrievals from the TRMM satellite in the multiscale assimilation system presented in Pan et al. (2009), resolving issues in generating the ensemble from the satellite data and evaluating the proposed assimilation system for regional-scale hydrologic studies.

The major challenge in constructing such an ensemble assimilation system is the generation of input forcing ensembles (e.g., for hydrology, the rainfall) given a single map of remotely sensed values. These input ensembles are fed into an LSM to produce the prior (forecast) state estimate and the information on state errors (magnitude and correlation structure). The quality of the input ensemble determines the quality of prior state ensembles—that is, its error correlation structure—and thus the quality of the assimilation. The multiscale approach works best with large-scale problems—that is, when the state vector has a very high dimension (equal to the number of pixels in the computing grid times the number of states at each pixel) that may reach 10^4–10^6. For ensemble data assimilation, the same high-dimensional spatial rainfall fields need to be generated as ensembles. There are two major requirements for the ensembles: 1) ensembles must contain a proper multiscale cross-member spatial correlation structure and 2) each ensemble member must present a realistic rainfall process in both space and time. The latter means that there should be a reasonable time evolution of storms. This property is important because if storms that exist in one time step are unrelated to the next then the LSM would recharge the soil at random places in different time steps and eventually the moisture would be smeared across the domain as the model integrates forward, making ensemble members indistinguishable from each other. In the sections to follow, the ensemble generation will be first discussed with a novel generation method proposed, and the ensembles so generated will be used to force the assimilation experiments. The multiscale spatial correlation structure in the rainfall ensemble will be tested in the experiments, as well as how this correlation structure is translated into the correlation structure of the soil moisture state errors by the LSM and how it affects the assimilation of soil moisture observations.

2. Rainfall ensemble generation

a. Background and previous approaches

Because of the nonstationary and intermittent nature of rainfall processes, ensemble generation of rainfall over large domains for assimilation purposes has been a difficult problem, and no standard or de facto approach exists. Usually, two classes of approaches are taken: generating rainfall fields using stochastic models or using dynamic models. The first approach assumes a probabilistic behavior of rainfall and uses Monte Carlo simulation to generate random realizations (e.g., Sivapalan and Wood 1987). The probabilistic models can be parametric or nonparametric. The parametric models are often complicated, with space–time correlations that capture the multiscale structure of synoptic scales/mesoscales with embedded convective cells that are often represented as occurring as a Poisson process in space (Sivapalan and Wood 1987). For example, storm dynamics can be modeled with an exponential decay of rainfall in time and Gaussian clustering of storms in space (Chatdarong 2006), the conditional rain rate probability can be based on a multiple-point process (Wójcik et al. 2008), and the spatial correlation of rain rates can be distance dependent (Villarini et al. 2009). Nonparametric rainfall models often resort to random sampling of historical fields and rescaling of these rain fields to adjust the distribution of rainfall totals. Examples of this approach are the Schaake shuffle (Clark et al. 2004), cumulative distribution function (CDF) matching, and similar procedures being used in downscaling research (Luo and Wood 2008).

The second approach for generating rainfall ensembles is to run multiple realizations of a dynamic atmospheric model—for example, a weather model—with perturbed initial conditions or forcing inputs to produce ensembles of rainfall forecasts. Such an approach provides rainfall ensembles that follow the physics of the underlying model and thus should be physically realistic both in the spatial distribution of rainfall depths and in the temporal evolution of the storms. This approach should provide very good spatial correlation among ensemble members because the ensemble spread and its intermember correlation come directly from the uncertainties that are built into the initial conditions, forcing fields, and model physics. However, it is computationally very expensive to generate ensembles using a regional atmospheric model. Ensemble simulations of atmospheric models are being carried out at all weather and seasonal climate centers, such as the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction. The spatial resolution of seasonal climate models like the

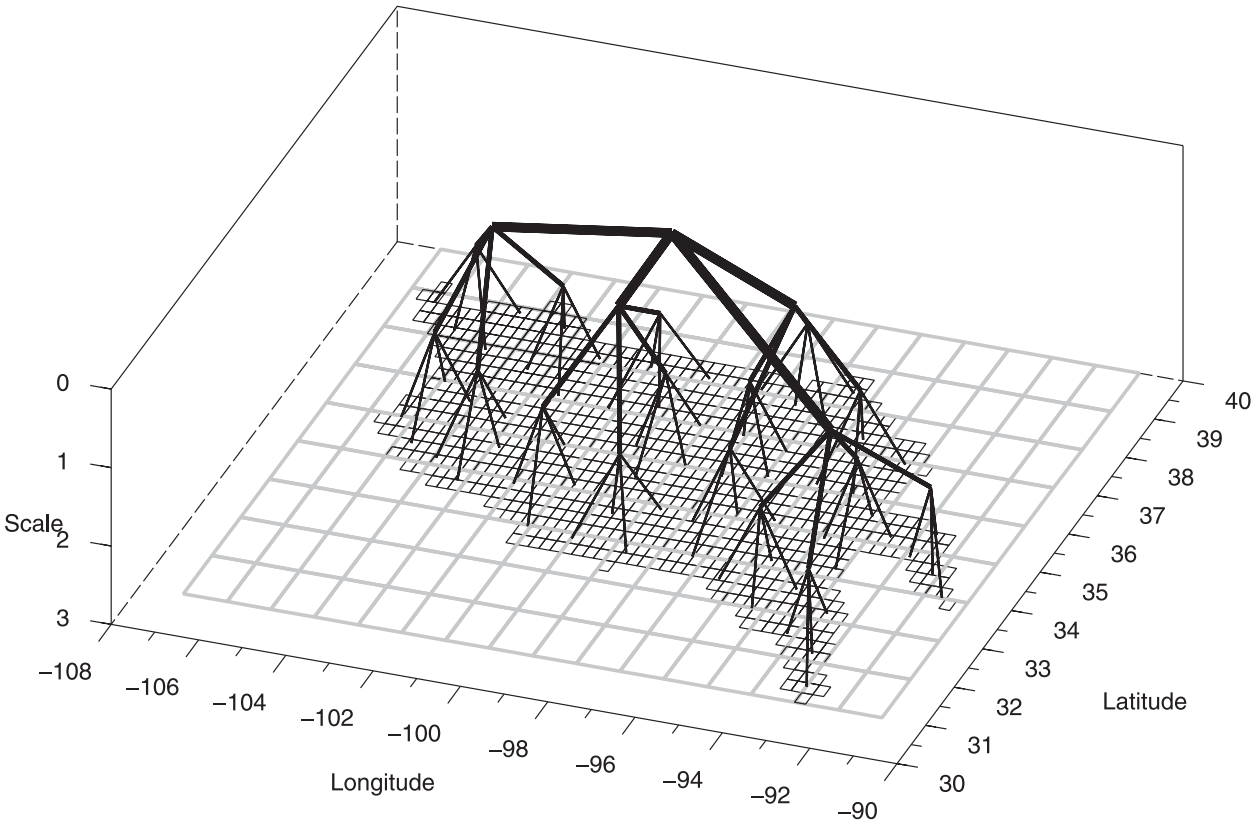

FIG. 1. The 1062-pixel computing grid at 0.25° resolution for the study area (fine pixels within the basin) and the "multiscale tree" defined for the grid. The 16 × 9 rectangular coarse grid at 1° resolution used in rainfall ensemble generation is delineated by the thick gray lines.

Climate Forecast System (Saha et al. 2006) or the NOAA global weather model (the Global Forecast System) are too coarse relative to the needs of LSM remote sensing applications, and their higher temporal fields are often not archived for application users. Regional weather models do provide a comparable resolution—for example, the Weather Research and Forecasting Model (Skamarock et al. 2008)—though the ensemble simulations are computationally expensive at large scales. Even when ensemble dynamic modeling is possible, it is so far impossible to condition the predictions upon the satellite observations—that is, to force the dynamic model to produce forecasts that are scattered around the observations.

After a review of these previous studies on ensemble generation, we determine that the approach of random sampling from historical rainfall fields offers for our current study the best physical consistency of rainfall storms in individual ensemble members. The main challenge with this approach is related to the high dimensionality of the rainfall field that must match the modeling grid, which makes the requirement on sample size very high and consequently the sampling process slow.

b. Rainfall ensemble generation with pattern-based sampling and CDF matching

In this study, the remote sensing assimilation experiment is carried out over the same domain as was used in Pan et al. (2009)—the Red–Arkansas River basin in the central United States. Figure 1 shows the latitude and longitude range of the basin. This is a relatively large basin that covers ~645 000 km^2—a size that is ideal for this study because it is large enough to observe multiscale phenomena in hydrology and to test the efficiency of the assimilation algorithm; however, it is not too large so that no more than one mesoscale precipitation system is expected to exist at any one time. Satellite-derived rainfall fields from the Tropical Rainfall Measuring Mission (TRMM) project are used—specifically the real-time TRMM product version 3B42RT. This product combines multiple satellite data sources (both microwave and infrared) in the estimation of the rainfall fields (Huffman et al. 2007). TRMM-3B42RT is available at a 0.25° spatial resolution and 3-h temporal resolution.

The LSM used in this study is the Variable Infiltration Capacity Model (VIC; Liang et al. 1994, 1996), and it is

configured to run at a 0.25° computing grid that results in 1062 pixels over the Red–Arkansas River basin study domain (see Fig. 1) The model time step is set to hourly because the model parameters available are calibrated at hourly step (Troy et al. 2008), with the assimilation of the 24-hourly soil moisture fields occurring at 1900 UTC, which matches the ascending overpasses of the Advanced Microwave Scanning Radiometer on board the Earth Observing System (AMSR-E) for soil moisture retrievals.

Given the above experiment setting, an ensemble of rainfall fields needs to be generated for every 24-h period for input into the LSM. A particular random sample (i.e. ensemble member) consists of 24-hourly fields covering a 16° × 9° rectangular area that encompasses the basin, because rainfall features are not limited by the basin boundary. This results in each rainfall field consisting of 64 × 36 = 2304 pixels (0.25°), with one 24-h random sample having the dimension 2304 × 24 = 55 296. Such a high-dimensional sample not only makes it difficult to compare among ensemble samples or to search for good samples but it requires a very large rainfall database from which to sample. However, rainfall events do not take arbitrary shapes, and the effective dimension is not that high, because most events follow specific patterns. Therefore, it is feasible to reduce the dimensionality of the sample by aggregation. Also, by categorizing rainfall events into different patterns, a long rainfall history can be divided into different groups according to their patterns, which can be used to limit the number of days from which samples are drawn.

1) SPATIAL/TEMPORAL AGGREGATION TO REDUCE DIMENSIONALITY

The purpose of aggregation is to reduce the dimensionality of data for the pattern classification and pattern matching later such that these operations can be performed faster and, at the same time, most of the rainfall pattern information is being used (only very fine details are ignored). The hourly rainfall database used in this study comes from the North America Land Data Assimilation System project phase 2 (NLDAS-2; Cosgrove 2007). NLDAS-2 hourly data are based on both gauge measurements and the North American Regional Reanalysis (Mesinger et al. 2006) and cover the contiguous United States region at 0.125° resolution from 1979 to near–real time. We denote an NLDAS-2 rainfall map for the study region as

$$\mathbf{r}^{(i)(j)}_{0.125^\circ,1\mathrm{hr}}$$

for the ith day and jth hour in the record. The complete NLDAS-2 dataset can be written as

$$\{\mathbf{r}^{(i)(j)}_{0.125^\circ,1\mathrm{hr}}\}_{\substack{i=1,\ldots,N_R\\ j=1,\ldots,24}},$$

where $N_R = 10\,591$ is the number of days in the record. We aggregate the rainfall up to 1° resolution and daily level:

$$\{\mathbf{r}^{(i)(j)}_{0.125^\circ,1\mathrm{hr}}\}_{\substack{i=1,\ldots,N_R\\ j=1,\ldots,24}} \xrightarrow{\text{aggregate}} \{\mathbf{r}^{(i)}_{1^\circ,24\mathrm{hr}}\}_{i=1,\ldots,N_R}, \quad (1)$$

resulting in the dimension of each sample day becoming $16 \times 9 \times 1 = 144$. At this resolution, most major features of the rainfall events are still captured. (Examples will be presented later.) Because the aggregation was to a daily total, variations in the hourly amounts are ignored. This is reasonable because at the daily scale the dynamics in soil moisture rely more on total daily rainfall than the hourly variations. For each day, the TRMM-3B42RT rainfall, noted as

$$\{\mathbf{m}^{(j)}_{0.25^\circ,3\mathrm{hr}}\}_{j=1,\ldots,8},$$

is also aggregated to the same spatial/temporal resolution:

$$\{\mathbf{m}^{(j)}_{0.25^\circ,3\mathrm{hr}}\}_{j=1,\ldots,8} \xrightarrow{\text{aggregate}} \mathbf{m}_{1^\circ,24\mathrm{hr}}. \quad (2)$$

2) PATTERN CLASSIFICATION TO LIMIT SAMPLE SIZE

The main purpose of the pattern classification is to predivide the entire sample space into several smaller areas, such that the random sampling can be performed over a smaller subset of the entire historical record. Such classification also helps to control the ensemble spread, which will be further discussed in section 2b(4). The classification is performed upon the reduced 1° daily samples using the neural gas (NG) clustering algorithm (Martinetz et al. 1993), which was previously applied by Pan et al. (2009) for multiscale classifications. The number of classification patterns N_P is predetermined. The NG algorithm finds N_P rainfall fields

$$\{\mathbf{p}^{(k)}_{1^\circ,24\mathrm{hr}}\}_{k=1,\ldots,N_P},$$

that is, patterns around which the rainfall events are clustered, and each day in the historical data record is assigned to a pattern. In NG classification, the distance between one rainfall field and another is measured by their Euclidean distance. For example, the distance between ith record and kth pattern is

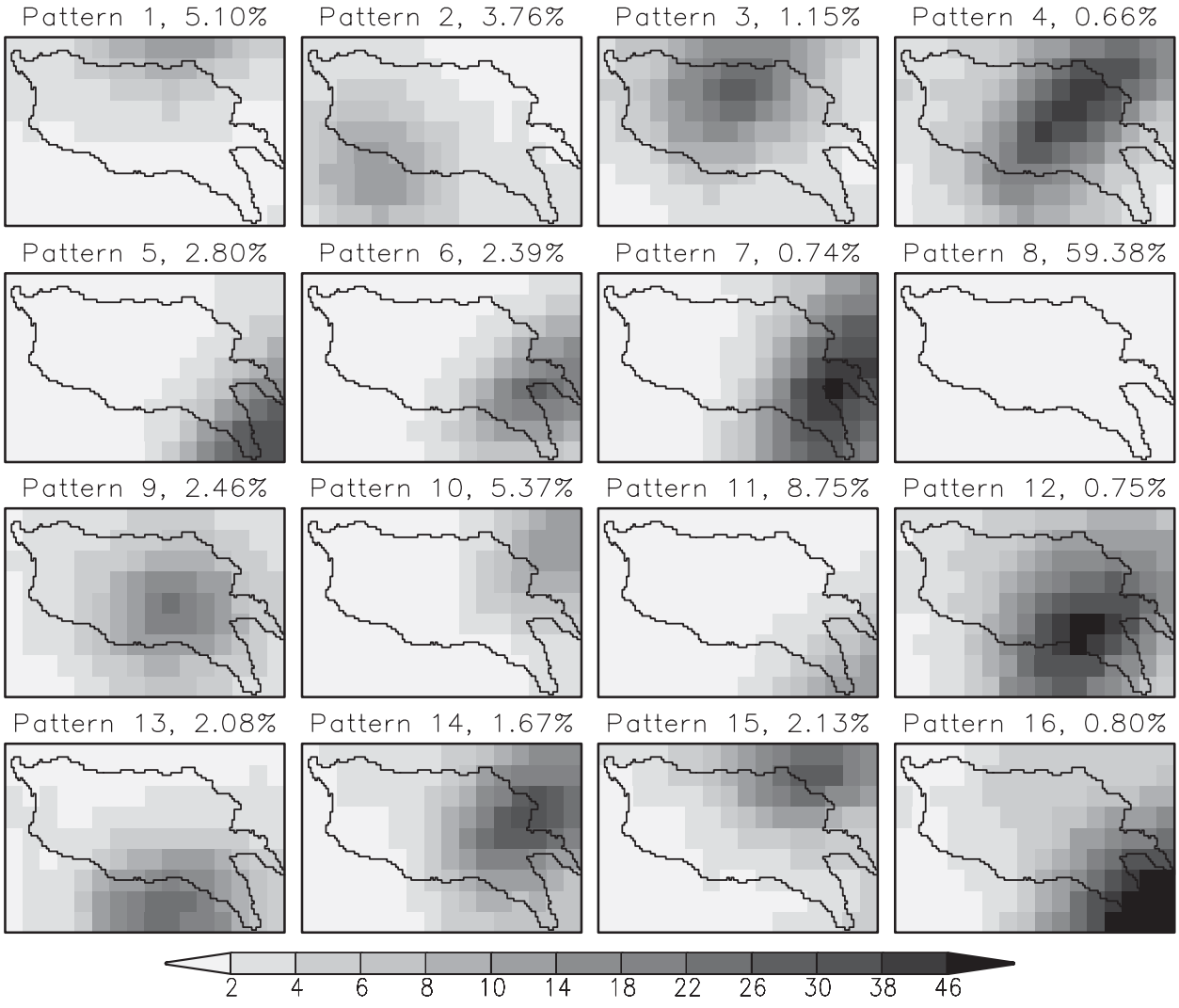

FIG. 2. Rainfall patterns identified in the experiment, with $N_P = 16$. The percentage value next to the pattern number in each panel is the fraction of samples that are classified as that pattern. Note that the eighth pattern, the no-rain pattern, contains the most samples (59.38%).

$$d^{(i)(k)} = \|\mathbf{r}^{(i)}_{1°,24hr} - \mathbf{p}^{(k)}_{1°,24hr}\| = \{[\mathbf{r}^{(i)}_{1°,24hr} - \mathbf{p}^{(k)}_{1°,24hr}]^{\mathrm{T}}[\mathbf{r}^{(i)}_{1°,24hr} - \mathbf{p}^{(k)}_{1°,24hr}]\}^{1/2}. \quad (3)$$

Note that rainfall fields

$$\mathbf{r}^{(i)}_{1°,24hr} \quad \text{and} \quad \mathbf{p}^{(k)}_{1°,24hr}$$

are treated as 144-dimensional vectors. Rainfall fields are classified to the closest pattern in terms of the Euclidean distance. For example, if the ith day is classified as the kth pattern, then

$$d^{(i)(k)} \le d^{(i)(l)} \quad \text{for any} \quad l \ne k. \quad (4)$$

Figure 2 shows an example of pattern classification for $N_P = 16$, together with the fraction of the total records that belongs to each of the patterns. These 16 patterns in Fig. 2 basically summarize the possible location/distribution of storms in this area. Note that the eighth pattern, the no-rain scenario, makes up ~60% percent of the days, which is reasonable because most days are nonrain days. Figure 3 shows the daily rainfall patterns for some typical days that are classified as the fourth pattern, labeled with dates.

Given one 1° daily TRMM-3B42RT rainfall field $\mathbf{m}_{1°,24hr}$, we first determine to which pattern it belongs. We then sample N_E days (N_E is the size of the ensemble) randomly from all the days in the historical records (i.e., NLDAS-2) that belong to the same pattern to form an ensemble of 1° daily rainfall

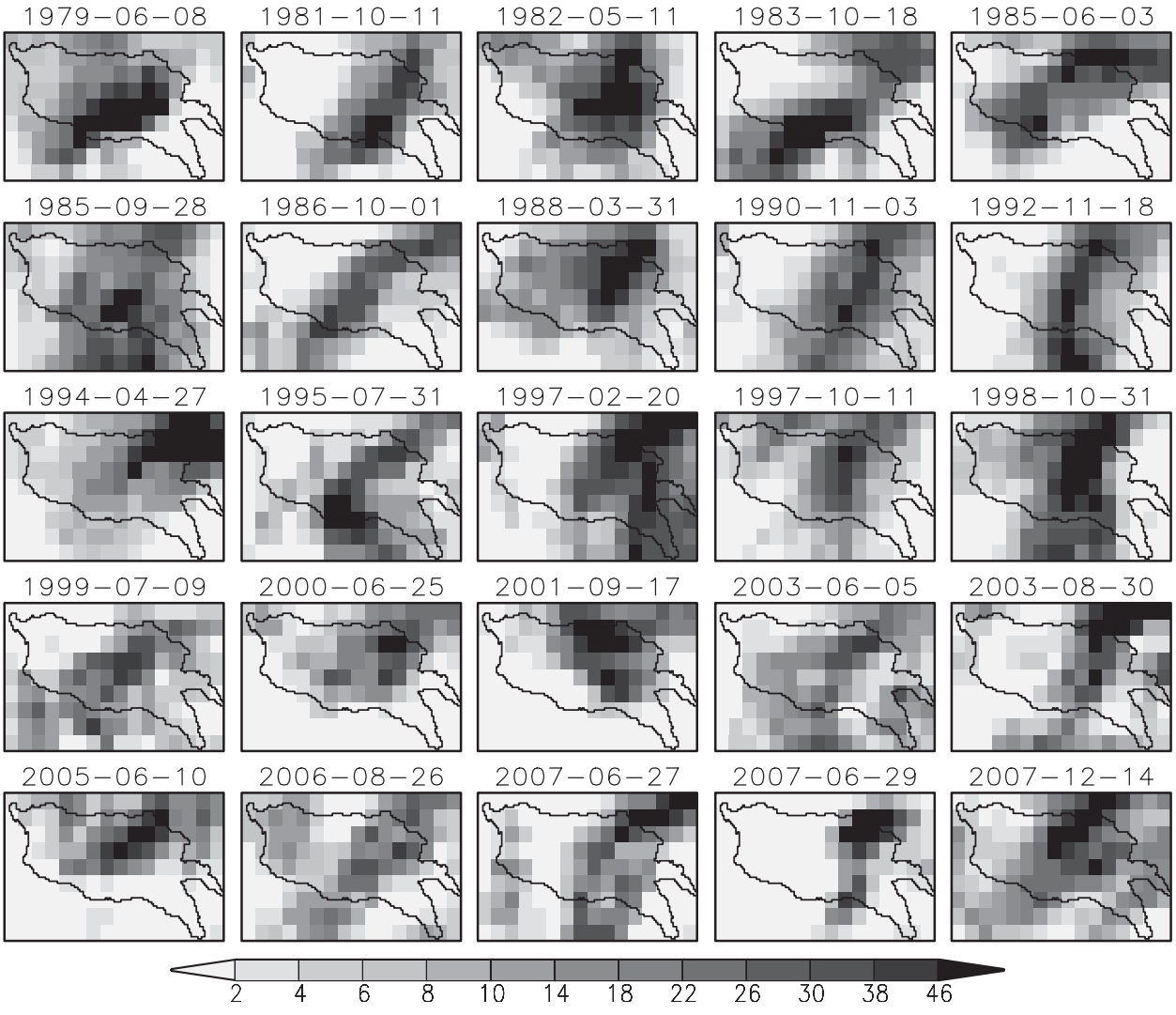

FIG. 3. Days in the historical record that are classified as the fourth pattern in the experiment, with $N_P = 16$ (only 25 of them are shown in the figure because of limited space). The shading scale is the same as in Fig. 2.

$$\{\mathbf{r}^{(l)}_{1^\circ,24\mathrm{hr}}\}_{l=1,\ldots,N_E}.$$

The 0.25° hourly data for the same days are retrieved from the NLDAS-2 data records to create an ensemble of hourly rainfall inputs

$$\{\mathbf{r}^{(l)(j)}_{0.25^\circ,1\mathrm{hr}}\}_{\substack{l=1,\ldots,N_E\\ j=1,\ldots,24}}$$

for VIC land surface model simulations.

3) CDF MATCHING TO ADJUST RAINFALL INTENSITY DISTRIBUTION

The 0.25° hourly ensemble

$$\{\mathbf{r}^{(l)(j)}_{0.25^\circ,1\mathrm{hr}}\}_{\substack{l=1,\ldots,N_E\\ j=1,\ldots,24}}$$

obtained in the last step may not have a distribution of rainfall intensities that best fits the TRMM-3B42RT rainfall on that specific day. This is resolved by matching the ensemble rainfall intensity distribution to the TRMM-3B42RT rainfall for that day using a CDF matching procedure (Reichle et al. 2008). That is to say, values in

$$\{\mathbf{r}^{(l)(j)}_{0.25^\circ,1\mathrm{hr}}\}_{\substack{l=1,\ldots,N_E\\ j=1,\ldots,24}}$$

will be replaced with values in

$$\{\mathbf{m}^{(j)}_{0.25^\circ,3\mathrm{hr}}\}_{j=1,\ldots,8}$$

of the same quantile in their respective distributions, with the result that the intensity distribution in

$$\{\mathbf{r}^{(l)(j)}_{0.25°,1\text{hr}}\}_{\substack{l=1,\ldots,N_E \\ j=1,\ldots,24}}$$

will be the same as in

$$\{\mathbf{m}^{(j)}_{0.25°,3\text{hr}}\}_{j=1,\ldots,8}.$$

Note that it is the intensity distribution of all ensemble members, in all 24 time steps, and collectively in all 2304 pixels that is matched to the TRMM-3B42RT measurements during a specific day. Individual members and hours in the ensemble may not follow the same distribution, but the ensemble as a whole does. This ensures that the ensemble, as a whole, does not present any bias or any other distributional differences against the TRMM-3B42RT input on any given day. Figure 4a shows an example of the final ensemble generated for the 24-h period from 1900 UTC 11 July 2004 to 1900 UTC 12 July 2004. These are used later in the assimilation experiment (see section 3) as maps of daily total rainfall ($N_P = 25$ and $N_E = 20$ are used throughout the assimilation experiment). Figure 4b gives the time series of basin total for the same ensemble set.

4) ENSEMBLE SPREAD CONTROL, CORRELATION STRUCTURE, AND OTHER ISSUES

A fundamental concern is what kind of uncertainty/error this pattern-based sampling method generates and how to control it. Some generation schemes (Wójcik et al. 2008; Villarini et al. 2009) try to parameterize and simulate the sensor errors, whereas others target model forecast uncertainties (Luo and Wood 2008). The sampling approach presented here makes no comparison between retrievals or forecasts to any ''truth'' reference, and therefore it is unable to simulate any type of errors. Instead, it focuses on the uncertainty of the rainfall process itself, with the long-term record providing the rainfall climatological values. If the sampling is performed over the entire record, then the ensemble spread will reflect the climatological uncertainty of the rainfall process over the study domain. When sampling is constrained within a specific pattern, the ensemble spread reflects the conditional uncertainty of the rainfall process given one event type. The number of patterns N_P determines the sampling spread across events from one classification. We can estimate the spread of one pattern (one rainfall event type) as $1/N_P$ times the climatological spread. Note that the spread differs from one type to another; for example, the no-rain type would have a much smaller spread than the spread for a heavy-storm type. Here the ensemble spread is measured relative to the climatological uncertainty only; therefore, if a certain amount of uncertainty is needed in the ensemble, then we need to translate it to the corresponding N_P or $1/N_P$ first. Once N_P is determined, then the ensemble spread is fixed no matter how large or small an ensemble will be drawn from it (i.e., it is not related to ensemble size N_E), because samples are always drawn from within the same pattern. The correlation structure should be fixed for a specific pattern. Different from some parametric methods, in which stationarity and isotropy are assumed for the errors, our sampling approach generates a correlation structure that is event-type specific and nonstationary. Figure 5 gives the correlation matrices—that is, normalized covariance $\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2}$—computed for each pattern (using all the samples in that pattern; $N_P = 16$). All of the matrices are 144×144 with the 144 pixels in the 2D field stretched along the side (i.e., at 1° resolution). Although it is difficult to match the matrices to what happens in the field, we can still see that high/low correlation happens at different places among different patterns, indicating the nonstationarity in time (event) and space.

The many advantages mentioned above rely on a long record length for a complete sampling of the rainfall climatological values. However, for such a high-dimensional problem, 10 591 days is still relatively short. Therefore, in practice, N_P cannot be too large; otherwise some low-probability event types get very few samples and the sampling becomes unstable. This poses a lower limit of the ensemble spread that can be achieved. To obtain tighter ensembles, one can set a maximum distance measure $d_{\max}$ such that all the days sampled are not farther than $d_{\max}$ away from the TRMM-3B42RT observations. The tightest ensemble so generated would be the N_E nearest neighbors in the historical record—the hard limit for pure sampling. For even tighter ensembles, additional manipulations must be performed—for example, linear or nonlinear rescaling to shrink the ensemble toward TRMM-3B42RT.

3. Data assimilation experiments and results

Many assimilation experiment settings have been described in section 2, and some of them are similar to those used in Pan et al. (2009). The same domain (Red–Arkansas River basin), computing grid (1062 pixels at 0.25°), and hydrologic model (VIC) are used. Major differences include using satellite-retrieved rainfall for the estimation experiment and that no ''synthetic'' truth is assumed. Instead, the LSM soil moisture fields, driven by ground-based rainfall, are used as the reference for a truth check. The ground-based rainfall comes from

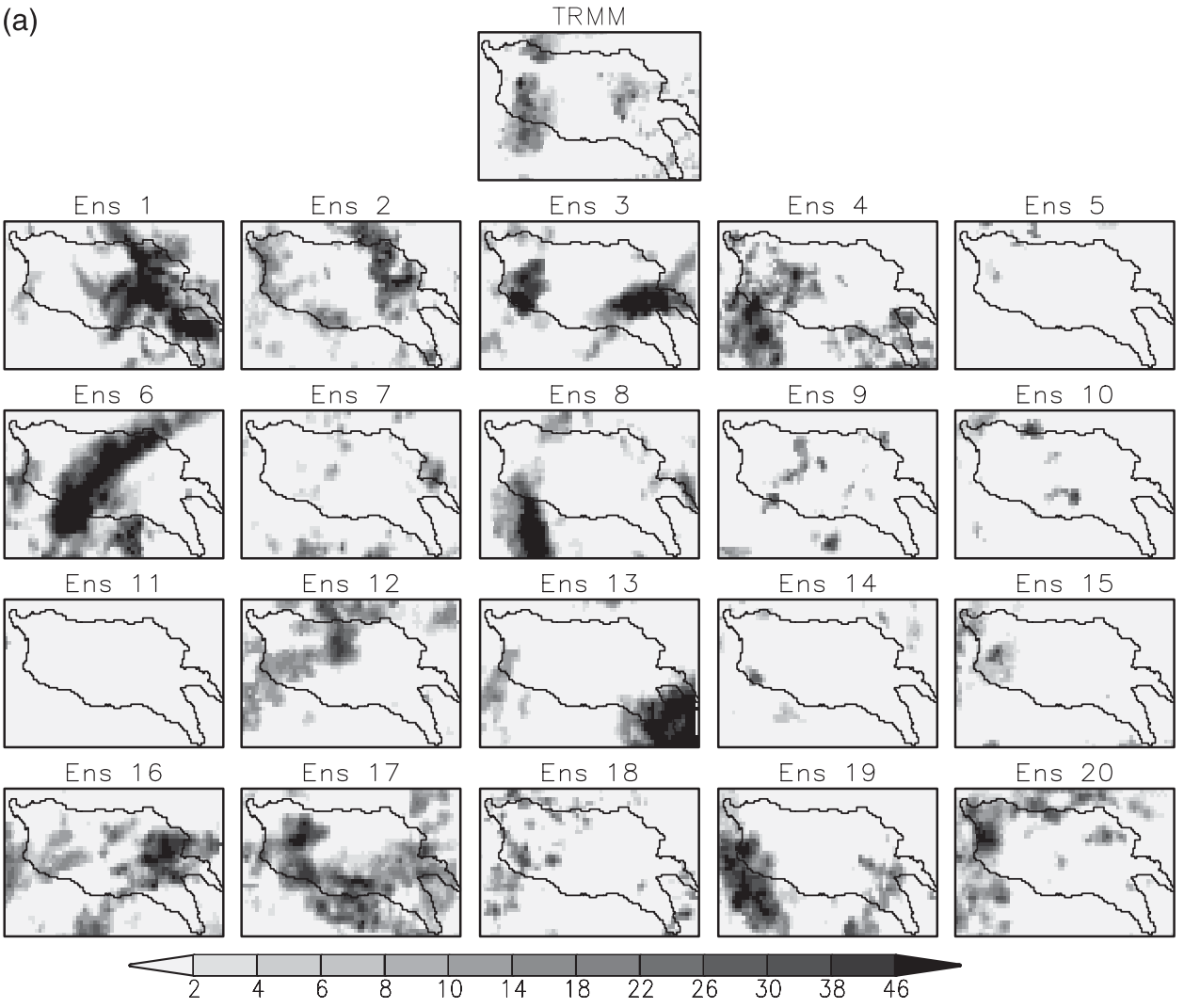

FIG. 4. (a) The daily total TRMM-3B42RT observations (center-top panel) and rainfall ensemble members (20 panels below) generated for the 24-h period from 1900 UTC 11 Jul to 1900 UTC 12 Jul 2004, using $N_P = 25$ and $N_E = 20$. (b) The time series of basin average rainfall corresponding to the same TRMM-3B42RT observation (center-top panel) and rainfall ensemble members shown in (a).

the retrospective forcing fields prepared for phase 1 of the NLDAS project (Mitchell et al. 2004), named "NLDAS" as opposed to "NLDAS-2", which is a combination of gauges and radar (Cosgrove et al. 2003). These rainfall data draw from more than 13 000 daily gauge reports and gauge-corrected hourly Weather Surveillance Radar-1988 Doppler (WSR-88D) radar estimates (for time disaggregation) over the contiguous United States. It is considered to be the best ground truth we can obtain in our study and should be sufficient to provide a truth check for this remote sensing assimilation experiment. VIC simulation forced with NLDAS-based rainfall is called a "reference" run here instead of a truth. Other significant differences in this experiment relative to Pan et al. (2009) are that the model predictions are hourly instead of daily to work with the remote sensing measurements and that the measurements are assimilated every 24 time steps (once per day) instead of every step as in the earlier study. The experiment period is the three summer months, 1 June–31 August 2004. Ensembles of 20 members ($N_E = 20$ and $N_P = 25$) are generated and used in the assimilation experiment. Two kinds of initializations are used prior to the assimilation: LSM spinup and an ensemble spinup. VIC is first run using a single observation-based forcing from 1 January to 31 May 2004 to minimize the effect of inaccurate initial conditions, and then it is run in ensemble form from 1 June to 30 June 2004 without any assimilation to achieve a stable ensemble spread. Measurements for the top-layer soil moisture (10 cm deep) are generated by

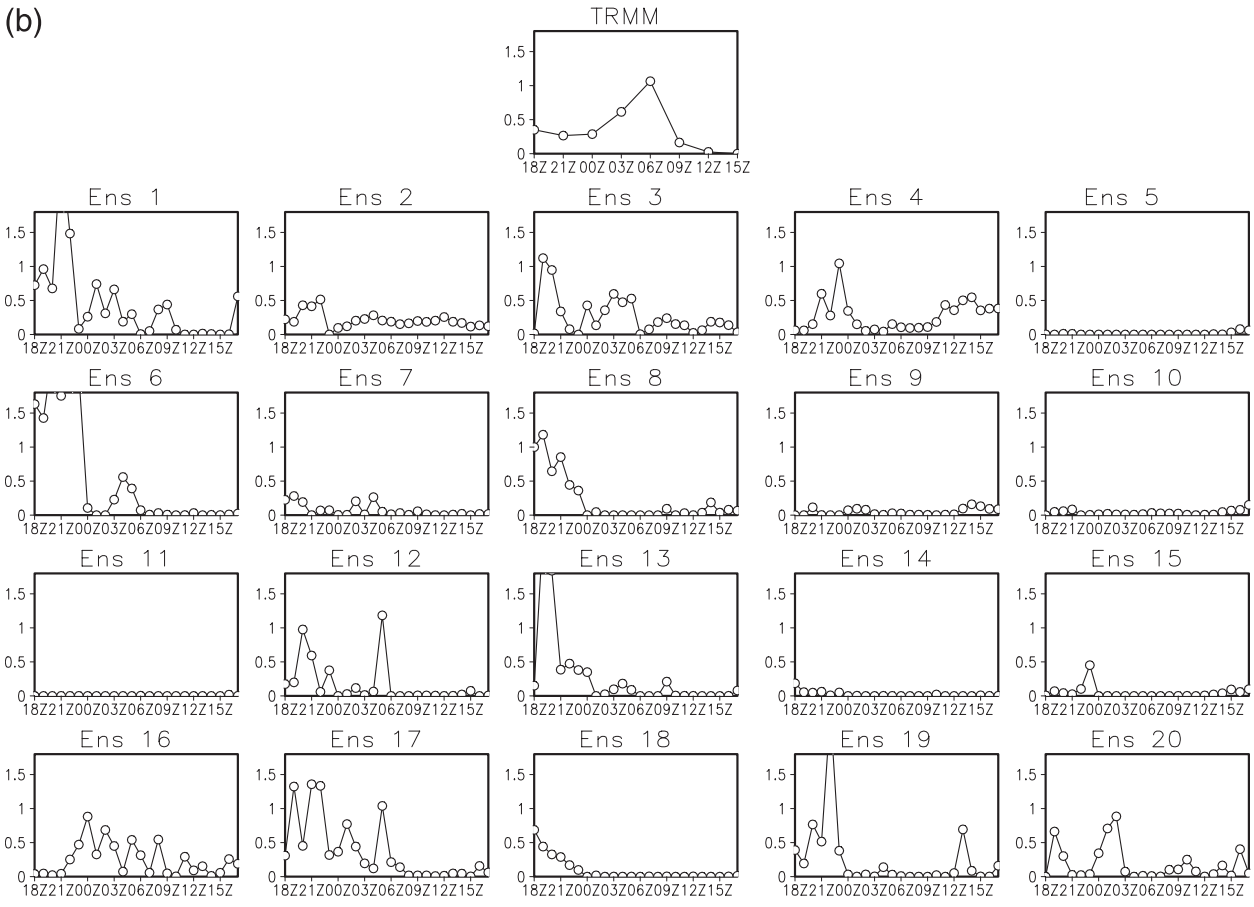

FIG. 4. (*Continued*)

adding Gaussian perturbations (zero mean and 3% standard deviation) to the reference simulation forced with NLDAS-based rainfall. These measurements are then assimilated into the model at 1900 UTC every day (at the AMSR-E ascending overpass) from 1 July to 31 August. Although no synthetic truth is assumed, the experiments are not purely ''real'' either, because no real AMSR-E soil moisture estimates are used.

Because one important goal of the study is to test the multiscale spatial correlation structure in the rainfall and consequently in soil moisture state errors, and how it affects the assimilation, we first look at the magnitude of the error correlations in the top-layer soil moisture ensemble. Figure 6 shows an example of the error correlation in top-layer soil moisture versus interpixel distances. Because the error correlation matrix is a large (1062 $\times$ 1062) symmetric matrix, only the statistics (median, 25th, and 75th percentiles) are computed from the lower triangle (diagonal included) of the correlation matrix using 200-km bins. The distance-lagged correlation structures before and after assimilation—that is, both the prior correlation (solid line/boxes) and posterior correlation (dashed line/boxes)—at 1900 UTC 30 July 2004 are computed. An immediate observation is that a significant amount of spatial correlation exists in the prior soil moisture errors. The median correlation at interpixel distance of ~600 km is about 0.2, large enough to have an effect on the assimilation. This relatively strong horizontal coupling in soil moisture errors suggests that a measurement taken as far as 600 km away may still carry some information about the pixel being updated. Thus, such horizontal coupling should be considered in soil moisture assimilation practice, which low-dimensional uncoupled filters may fail to do, as was also found in Pan et al. (2009). Second, the overall strength of the spatial correlation drops sharply after the update—the mean error correlation at all distances drops from 0.35 to 0.05. This suggests that the spatial correlation in soil moisture errors is more associated with the rainfall errors than with the intrinsic characteristics of the soil moisture field, and once the assimilation update reduces the ensemble spread around the measurement, the error correlation also diminishes.

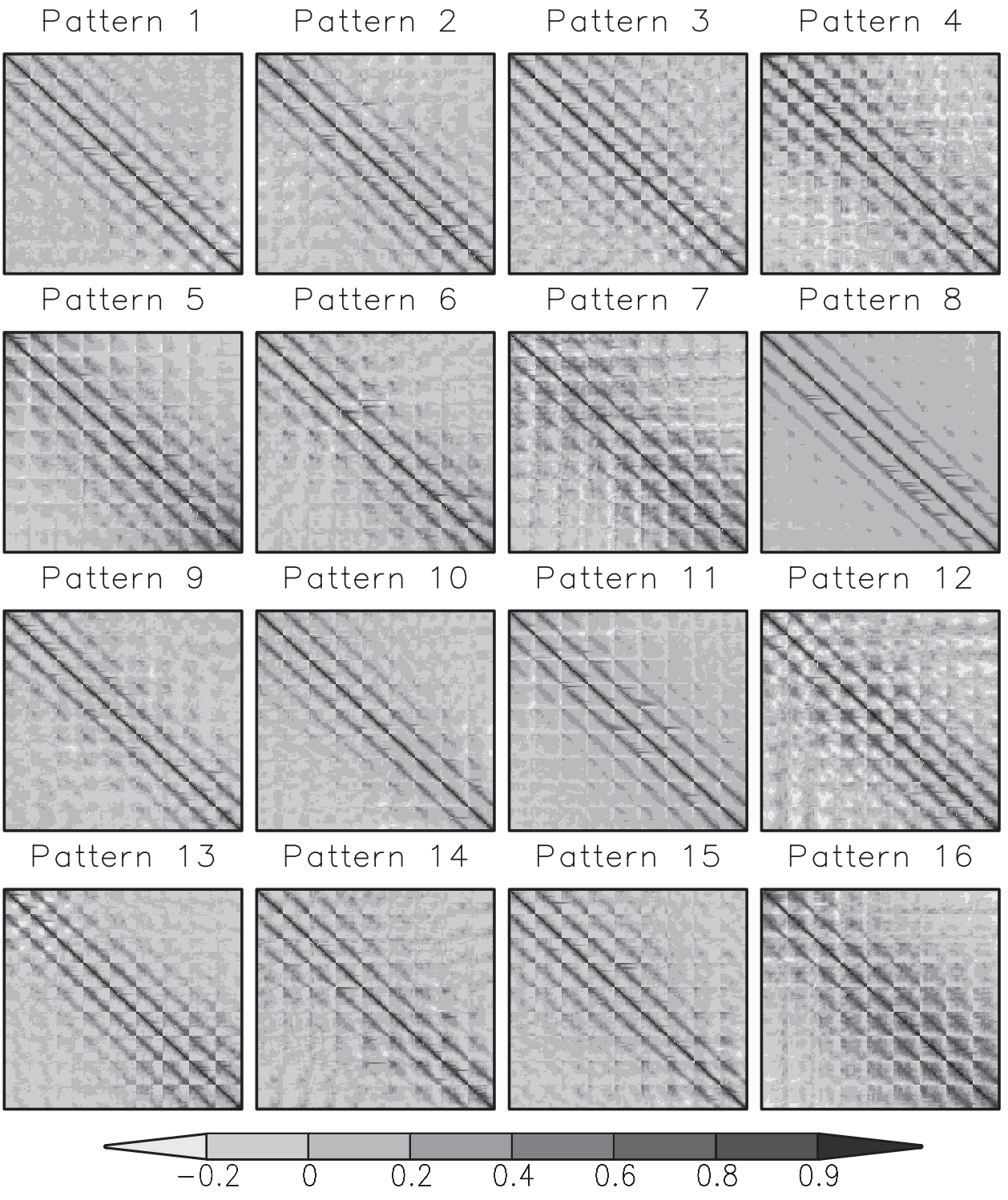

FIG. 5. Correlation (normalized covariance) matrices computed for each different rainfall pattern (using all of the samples in that pattern) in the $N_P = 16$ experiment. Each matrix is 144 × 144 with the 144 columns/rows corresponding to the 16 × 9 1° pixels.

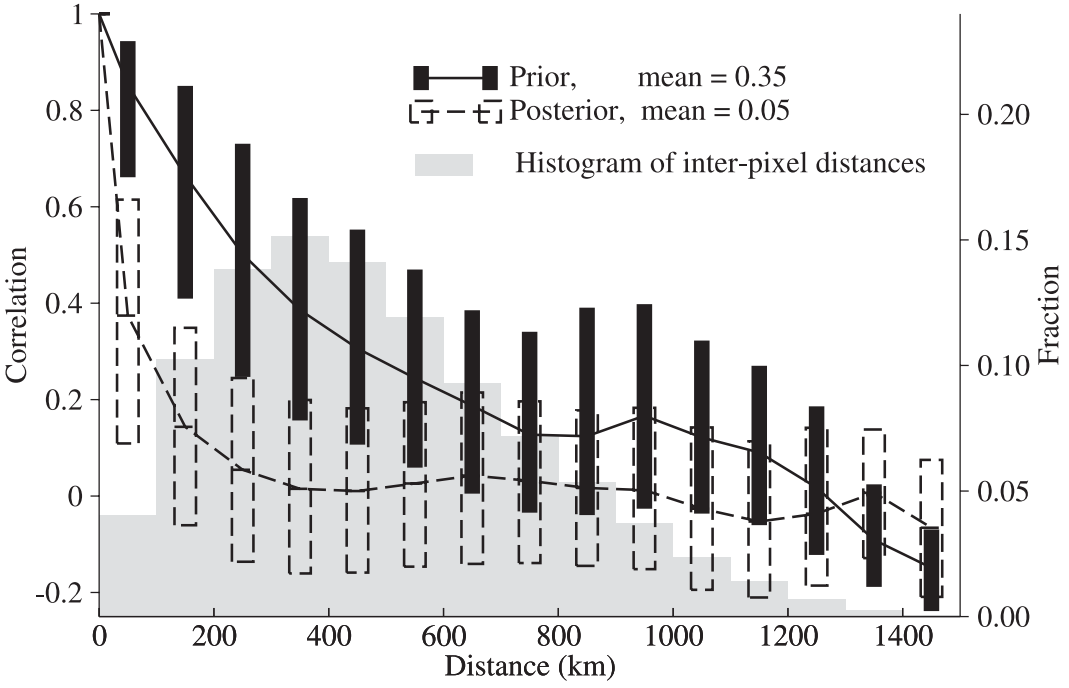

FIG. 6. Distribution of error correlation in top-layer soil moisture along different interpixel distances. The 1062 × 1062 error correlation matrix is computed for the model time step at 1900 UTC 30 Jul both before and after the update (prior/posterior), and the statistics in this figure are computed using the values in the lower triangle (with the diagonal included) of the 1062 × 1062 correlation matrix (symmetric) and 200-km bins. The lines and the upper/lower edges of the boxes indicate the median, 75th, and 25th percentiles of the distribution in that bin, respectively. Solid-line (dashed line) boxes are for the prior (posterior) error correlation. The mean error correlation at all distances is 0.35 and 0.05 for the prior and posterior, respectively. The histogram of interpixel distances is plotted in light gray in the background.

Figure 7 summarizes the assimilation in time series. Because of the relatively long length of the experiment (2208 time steps), only a period of approximately three weeks (26 July–13 August) is illustrated. The period covers a couple of significant rainfall events and a long dry-down. The top panel in Fig. 7 shows the basin-averaged ensemble-mean top-layer soil moisture and the basin-averaged ensemble standard deviation (STD). The growth of the ensemble spread during the 24-h periods between every two filter updates is very clearly expressed by the error bars (STD), as is the reduction of the soil moisture ensemble spread at every update time (1900 UTC). The ensemble spread grows between assimilation periods and during large rainfall events (e.g., 28–30 July and 10–11 August). The ensemble spread becomes much smaller during the dry-down period in the beginning of August. This behavior reflects the nonstationary nature of the rainfall ensemble generation, which determines the extent of spread based on event type and climatological values (i.e., a large spread for large storm events and small spread for dry days), and the response of the LSM to rainfall or to dry conditions.

The middle panel compares the open-loop simulation (no measurements assimilated; thick dashed line), the ensemble mean (thick solid line), and the NLDAS rainfall–driven reference simulation (thin solid line). The improvement in the soil moisture estimation made by the assimilation against the open-loop run can be seen in this panel. Such an improvement is significant, especially during the period from 25 July to 4 August, when the biased low soil moisture is well corrected through assimilation.

The bottom panel draws the time series of mean spatial correlation in soil moisture errors (over all interpixel distances). The correlation magnitude grows in the 24-h periods between updates and decreases significantly after assimilating new measurements—confirming the observation that the spatial correlation in soil moisture errors comes primarily from the rainfall-forcing ensemble and will diminish as the soil moisture ensemble spread decreases upon the filter update.

The basin-averaged root-mean-square errors (RMSE) in top-layer soil moisture computed against the NLDAS-driven reference run is 2.93% for the open-loop run and 2.29% for the assimilation run (22% error reduction).

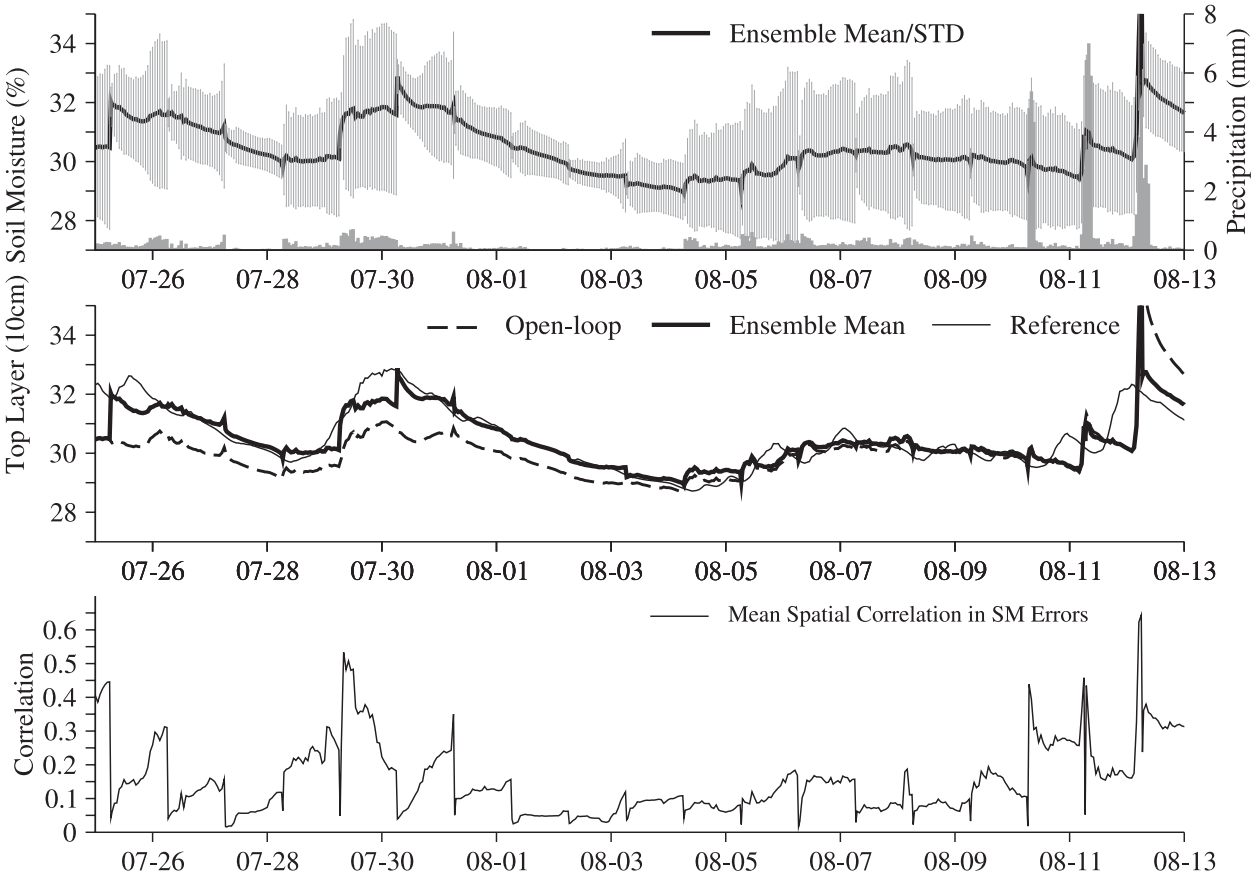

FIG. 7. (top) Basin-averaged ensemble-mean top-layer soil moisture (thick solid line), basin-averaged ensemble STD (error bars), and basin-averaged ensemble-mean precipitation (gray bars). (middle) Basin-averaged ensemble mean (thick solid line) vs the open-loop run (thick dashed line) and NLDAS rainfall–driven reference run (thin solid line). (bottom) Mean spatial correlation in soil moisture errors, computed from the same lower triangle of error correlation matrices as was used in Fig. 6.

The results are based on the one-overpass-per-day configuration, whereas AMSR-E can provide two overpasses per day. The assimilation experiment with two overpasses per day (0700 and 1900 UTC) brings the RMSE down to 2.09% (29% error reduction). All of the absolute RMSE values—2.93%, 2.29% and 2.09%—are small, as are the 3% error added to create the synthetic soil moisture measurements when compared with the error levels reported for soil moisture retrievals (Bindlish et al. 2003; Gao et al. 2006). Note that all these small numbers are not unreasonable because they are all relative to the soil moisture dynamics of VIC LSM simulations. The VIC model drains its 10-cm top layer only by gravity and evapotranspiration, and its soil moisture value is usually between 20% and saturation (~45%). The dynamic range of VIC surface moisture is considerably smaller than remotely sensed, X-band (10.67 GHz), thin (~1 cm) top-layer moisture, which sometimes reaches <5% (Gao et al. 2006). The RMSEs observed in this study should be rescaled by a factor of 1.5–2 to obtain the equivalent values in remote sensing terms.

4. Conclusions

We design and test a nonparametric rainfall ensemble generation scheme for a multiscale land surface hydrologic assimilation system such that remotely sensed rainfall data can be used. The random sampling/CDF matching approach proposed here provides a computationally efficient way for generating realistic rainfall ensembles. The rainfall uncertainties (i.e., errors) so generated are nonstationary in both time and space, and the uncertainty in the rainfall errors are no longer assumed to be normal or lognormal but are measured relative to the climatological uncertainties of the rainfall in the region. Assimilation experiments conducted show that this generation method works reasonably well with the remote sensing rainfall products being used (TRMM-3B42RT). In addition to using remotely sensed rainfall, all of the assimilation experiments are

configured to work with the soil moisture retrievals from an operational sensor like the AMSR-E onboard NASA's EOS *Aqua* satellite. Although no real AMSR-E retrievals are used in this paper, the experiments with the LSM-created measurements illustrate good potential of having a remote sensing–based multiscale assimilation system for improving soil moisture estimations.

The ensemble generation method proposed has its limitations. It relies on the existence of a long-term subdaily rainfall database, which poses a lower bound on the ensemble spread that can be achieved and also limits its application in some data-limited areas. The potential use of nonobservational rainfall databases—for example, the European Centre for Medium-Range Weather Forecasts ''ERA-Interim'' global reanalysis (Simmons et al. 2006) or the statistically interpolated 50-yr global surface meteorological records developed in Sheffield et al. (2006)—may help to alleviate this limitation. The length of the historical record also puts a limit on domain size (number of pixels), and our experiments are all at regional scale so far. Therefore, this method may only be applied to larger-scale (continental or global) applications on a region-by-region basis. The uncertainties built into the ensembles are relative to the regional rainfall climatological values and do not directly reflect the satellite sensor errors, which is sometimes expected in data assimilation. Overall, this study serves as a good proof of concept for the remote sensing–based multiscale assimilation system, even though additional issues—for example, the sensitivity of assimilation performance on retrieval accuracy, spatial availability, and revisit frequency—need to be resolved.

Acknowledgments. This research is supported by National Aeronautics and Space Administration (NASA) Grants NNG06GD79G (''Estimating Continental-Scale Water Balances through Modeling and Assimilation of EOS *Terra* and *Aqua* Data,'' in collaboration with the University of Washington) and NNX07AK41G (''Assessment of the Predictive Skill of GPM-ERA Precipitation Estimates for Hydrologic Applications'').

REFERENCES

Bindlish, R., T. J. Jackson, E. Wood, H. Gao, P. Starks, D. Bosch, and V. Lakshmi, 2003: Soil moisture estimates from TRMM Microwave Imager observations over the Southern United States. *Remote Sens. Environ.,* **85,** 507–515.

Chatdarong, V., 2006: Multi-sensor rainfall data assimilation using ensemble approaches. Ph.D. thesis, Massachusetts Institute of Technology, 203 pp.

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

Cosgrove, B., 2007: Forcing files for NLDAS Phase 2 (NLDAS-2), version 1.0. NASA Rep., 7 pp. [Available online at http://ldas.gsfc.nasa.gov/LDAS8th/LDASdocs/nldas2forcing.pdf.]

——, and Coauthors, 2003: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.,* **108,** 8842, doi:10.1029/2002JD003118.

Crow, W. T., and E. F. Wood, 2003: The assimilation of remotely sensed soil brightness temperature imagery into a land-surface model using ensemble Kalman filtering: A case study based on ESTAR measurements during SGP97. *Adv. Water Resour.,* **26,** 137–149.

Frakt, A. B., and A. S. Willsky, 2001: Computational efficient stochastic realization for internal autoregressive models. *Multidimens. Syst. Signal Process.,* **12,** 109–142.

Gao, H., E. F. Wood, T. J. Jackson, M. Drusch, and R. Bindlish, 2006: Using TRMM/TMI to retrieve surface soil moisture over the southern United States from 1998 to 2002. *J. Hydrometeor.,* **7,** 23–38.

Houser, P. R., and J. Entin, 2006: The NASA Energy- and Water-Cycle Study (NEWS). *Geophysical Research Abstracts,* Vol. 8, Abstract 08456. [Available online at http://www.cosis.net/abstracts/EGU06/08456/EGU06-J-08456.pdf.]

Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.,* **8,** 38–55.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.,* **99,** 14 415–14 428.

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

Luo, L., and E. F. Wood, 2008: Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeor.,* **9,** 866–884.

Margulis, S. A., E. F. Wood, and P. A. Troch, 2006: The terrestrial water cycle: Modeling and data assimilation across catchment scales. *J Hydrometeor.,* **7,** 309–311.

Martinetz, T. M., S. G. Berkovich, and K. J. Schulten, 1993: Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Trans. Neural Network,* **4,** 558–569.

McCabe, M. F., E. F. Wood, R. Wójcik, M. Pan, J. Sheffield, H. Gao, and H. Su, 2008: Hydrological consistency using multi-sensor remote sensing data for water and energy cycle studies. *Remote Sens. Environ.,* **112,** 430–444.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.,* **87,** 343–360.

Mitchell, K. E., and Coauthors, 2004: 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.

Pan, M., E. F. Wood, R. Wójcik, and M. F. McCabe, 2008: Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. *Remote Sens. Environ.,* **112,** 1282–1294.

——, ——, D. B. McLaughlin, D. Entekhabi, and L. Luo, 2009: A multiscale ensemble filtering system for hydrologic data assimilation. Part I: Implementation and synthetic experiment. *J. Hydrometeor.,* **10,** 794–806.

Reichle, R. H., and R. D. Koster, 2005: Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. *Geophys. Res. Lett.,* **32,** L02404, doi:10.1029/2004GL021700.

——, W. T. Crow, R. D. Koster, H. O. Sharif, and S. P. P. Mahanama, 2008: Contribution of soil moisture retrievals to land data assimilation products. *Geophys. Res. Lett.,* **35,** L01404, doi:10.1029/2007GL031986.

Rodell, M., and Coauthors, 2004: The Global Land Data Assimilation System. *Bull. Amer. Meteor. Soc.,* **85,** 381–394.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

Schaake, J. C., 1994: Science strategy of the GEWEX Continental-Scale International Project (GCIP). *Adv. Water Resour.,* **17,** 117–127.

Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling. *J. Climate,* **19,** 3088–3111.

Simmons, A. J., S. M. Uppala, D. Dee, and S. Kobayashi, 2006: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter,* No. 110, ECMWF, Reading, United Kingdom, 25–35.

Sivapalan, M., and E. F. Wood, 1987: A multidimensional model of nonstationary space–time rainfall at the catchment scale. *Water Resour. Res.,* **23,** 1289–1299.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 125 pp. [Available online at http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Slater, A. G., and M. P. Clark, 2006: Snow data assimilation via an ensemble Kalman filter. *J. Hydrometeor.,* **7,** 478–493.

Troy, T. J., E. F. Wood, and J. Sheffield, 2008: An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.,* **44,** W09411, doi:10.1029/2007WR006513.

Villarini, G., W. F. Krajewski, G. J. Ciach, and D. L. Zimmerman, 2009: Product-error-driven generator of probable rainfall conditioned on WSR-88D precipitation estimates. *Water Resour. Res.,* **45,** W01404, doi:10.1029/2008WR006946.

Willsky, A. S., 2002: Multiresolution Markov models for signal and image processing. *Proc. IEEE,* **90,** 1396–1458.

Wójcik, R., D. McLaughlin, A. G. Konings, and D. Entekhabi, 2008: Conditioning stochastic rainfall replicates on remote sensing data. *IEEE Trans. Geosci. Remote Sens.,* **47,** 2436–2449.

Zhou, Y., 2006: Multisensor large scale land surface data assimilation using ensemble approaches. Ph.D. thesis, Massachusetts Institute of Technology, 234 pp.

——, D. McLaughlin, D. Entekhabi, and G. C. Ng, 2008: An ensemble multiscale filter for large nonlinear data assimilation problems. *Mon. Wea. Rev.,* **136,** 678–698.