Hydrologic Data Assimilation with the Ensemble Kalman Filter

ROLF H. REICHLE,* DENNIS B. MCLAUGHLIN, AND DARA ENTEKHABI

Ralph M. Parsons Laboratory, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Manuscript received 28 December 2000, in final form 22 March 2001)

ABSTRACT

Soil moisture controls the partitioning of moisture and energy fluxes at the land surface and is a key variable in weather and climate prediction. The performance of the ensemble Kalman filter (EnKF) for soil moisture estimation is assessed by assimilating L-band (1.4 GHz) microwave radiobrightness observations into a land surface model. An optimal smoother (a dynamic variational method) is used as a benchmark for evaluating the filter's performance. In a series of synthetic experiments the effect of ensemble size and non-Gaussian forecast errors on the estimation accuracy of the EnKF is investigated. With a state vector dimension of 4608 and a relatively small ensemble size of 30 (or 100; or 500), the actual errors in surface soil moisture at the final update time are reduced by 55% (or 70%; or 80%) from the value obtained without assimilation (as compared to 84% for the optimal smoother). For robust error variance estimates, an ensemble of at least 500 members is needed. The dynamic evolution of the estimation error variances is dominated by wetting and drying events with high variances during drydown and low variances when the soil is either very wet or very dry. Furthermore, the ensemble distribution of soil moisture is typically symmetric except under very dry or wet conditions when the effects of the nonlinearities in the model become significant. As a result, the actual errors are consistently larger than ensemble-derived forecast and analysis error variances. This suggests that the update is suboptimal. However, the degree of suboptimality is relatively small and results presented here indicate that the EnKF is a flexible and robust data assimilation option that gives satisfactory estimates even for moderate ensemble sizes.

1. Introduction

Near-surface soil moisture is a key variable in the atmospheric and hydrologic models that are used to predict weather and climate. Since soil moisture controls the partitioning of moisture and energy fluxes at the land surface, it has an important influence on the hydrologic cycle over timescales ranging from hourly to interannual. Land surface fluxes in turn affect the evolution of vertical buoyancy in the atmospheric column and also affect the baroclinicity that develops in the horizontal plane (Pan et al. 1995; Paegle et al. 1996). The formation and growth of clouds as well as the evolution of precipitating weather systems over land are affected by surface fluxes and surface soil moisture (Shaw et al. 1997). In fact, the timescale of soil moisture anomalies is at least on the order of several days, which is the forecast-lead horizon of operational weather forecasts.

At seasonal to interannual timescales, predictability of climatic variables such as precipitation is dependent on the land surface boundary conditions of the climate system. Koster et al. (2000) show that over the United States and other large continental regions soil moisture rivals sea surface temperature in explaining the variance in seasonal precipitation anomalies. Increasingly soil moisture and the memory associated with it are recognized to have important roles in the feedback mechanisms that intensify and prolong climate anomalies.

Despite the importance of soil moisture in weather and climate prediction there are currently no operational networks of in situ sensors that provide data suitable for these applications. Since such networks are logistically infeasible and prohibitively expensive, the focus has turned to remote sensing techniques that provide additional information about the land surface at large scales. In particular the L-band (1.4 GHz) microwave brightness temperature of the land surface is correlated with surface soil moisture because of the sharp contrast between the dielectric constants of water and soil minerals (Njoku and Entekhabi 1995).

Interpretation of remotely sensed passive and active microwave measurements is complicated by the effects of canopy microwave optical thickness, surface microroughness, and physical temperature. Remote sensing measurements are only one of many data sources that provide valuable information about soil moisture. Precipitation, soil texture, topography, land use, and a variety of meteorological variables influence the spatial

* Current affiliation: NASA GSFC, Greenbelt, Maryland.

Corresponding author address: Rolf H. Reichle, NASA GSFC, Bldg. 33, A-110, Code 974, Greenbelt, MD 20771.
E-mail: reichle@janus.gsfc.nasa.gov

© 2002 American Meteorological Society

distribution and temporal evolution of soil moisture. We can gain additional information from a coupled model of the soil–vegetation–atmosphere system that relates the measured variables to one another and to soil moisture. Yet uncertainties in the forcing data, the heterogeneity of the land surface at various scales, and the nonlinear nature of land–atmosphere interactions limit our ability to accurately model and predict the state of the land surface and the associated fluxes.

Modern data assimilation theory provides methods for optimally merging the information from uncertain remote sensing observations and uncertain land model predictions (Errico 1999; Errico et al. 2000). Among the prior work on large-scale soil moisture assimilation are studies by Bouttier et al. (1993) and Rhodin et al. (1999) who assimilate low-level air temperature and relative humidity to estimate soil moisture. This approach aims at improving numerical weather prediction and treats soil moisture as a tuning parameter. Houser et al. (1998) focus on the four-dimensional assimilation of in situ observations and soil moisture retrievals. The weak-constraint variational method of Reichle et al. (2001b) yields near-optimal estimates of the land surface states from direct assimilation of microwave observations. Reichle et al. (2001a) prove the concept of optimal downscaling for the case where soil moisture estimates are required at scales smaller than the scale of the microwave observations. They also show that soil moisture can be satisfactorily estimated even if quantitative precipitation estimates are not available.

In this paper we examine the feasibility of using the ensemble Kalman filter (EnKF) for soil moisture data assimilation. The EnKF is an attractive option for land surface applications because (i) its sequential structure is convenient for processing remotely sensed measurements in real time, (ii) it provides information on the accuracy of its estimates, (iii) it is relatively easy to implement even if the land surface model and measurement equations include thresholds and other nonlinearities, and (iv) it is able to account for a wide range of possible model errors. On the other hand, the EnKF relies on a number of assumptions and approximations that may compromise its performance in certain situations.

The EnKF and variants have been successfully applied to meteorological and oceanographic problems of moderate complexity in small- to medium-sized domains (Evensen and van Leeuwen 1996; Houtekamer and Mitchell 1998; Lermusiaux 1999; Madsen and Canizares 1999; Keppenne 2000). Hamill et al. (2000) provide an excellent discussion of the state of the art of ensemble forecasting and assimilation methods in the meteorological and oceanographic context. The models of geophysical flow used in most of these studies are chaotic in nature and typically have dominant modes that can grow rapidly within a certain subspace. Most such models also have an attractor and sample only a small subdomain of their phase space (Anderson and Anderson 1999). This greatly increases the potential to successfully apply ensemble filtering methods. By contrast, typical land surface models are dissipative in nature. Perturbations in the initial conditions tend to die out after a certain time rather than amplify. Consequently, the soil moisture ensemble filtering problem has certain distinctive aspects that merit closer investigation.

This paper evaluates the performance and computational burden of the EnKF for a synthetic experiment. As a benchmark for the EnKF we use a variational method that solves the optimal smoothing problem. We begin in section 2 with a brief review of the EnKF. The benchmark variational method is discussed in the cited references, including (Reichle et al. 2001b). In section 3 we briefly describe the land model and the setup of the synthetic experiments we use to investigate design issues. In section 4 we discuss the results of these experiments and compare the EnKF with the variational method. We conclude in section 5 with a summary of major findings.

2. The ensemble Kalman filter

The standard Kalman filter (KF) is the optimal sequential data assimilation method for linear dynamics and measurement processes with Gaussian error statistics (Gelb 1974). For nonlinear dynamics, the extended Kalman filter (EKF) can be used, although it is notoriously unstable if the nonlinearities are strong (Miller et al. 1994). Both the KF and the EKF explicitly propagate error information with a dynamic equation for the state error covariance matrix. However, the integration of this equation is not computationally feasible for large-scale environmental systems. To overcome these limitations, Evensen (1994) uses an ensemble of model trajectories from which the necessary error covariances are estimated at the time of an update. The technique has since become known as the ensemble Kalman filter. The method uses the nonlinear model to propagate the ensemble states. Some of the linearizations that make the EKF prone to failure are thereby avoided.

The nonlinear land surface model used for the assimilation can be expressed in a generic form if we assemble the spatially discretized state variables of interest (e.g., the soil moisture and soil temperature) at all computational nodes and at time t into the state vector $\mathbf{Y}(t)$ of dimension N_Y. The resulting model equation is

$$\frac{d\mathbf{Y}}{dt} = \mathcal{F}(\mathbf{Y}) + \mathbf{w}. \tag{1}$$

The nonlinear operator $\mathcal{F}(\cdot)$ includes all deterministic forcing data (e.g., observed rainfall). Uncertainties related to errors in the model formulation or the forcing data are summarized in the model error term $\mathbf{w}$.

The observations used for the assimilation are remotely sensed measurements of the microwave brightness temperature of the land surface. These observations are sparse in time and space and only indirectly related

to soil moisture and temperature. If we collect all observations taken at time t_k into the measurement vector $\mathbf{Z}_k$ of dimension N_{Zk}, we can express the measurement process as

$$\mathbf{Z}_k = \mathcal{M}_k\,[\mathbf{Y}(t_k)] + \mathbf{v}_k. \tag{2}$$

The nonlinear operator $\mathcal{M}_k(\cdot)$ relates the true state (e.g., soil moisture) to the measured variable (e.g., brightness temperature). The uncertainties of the measurement process are reflected in the measurement error $\mathbf{v}_k$.

We adopt a probabilistic interpretation of uncertainty and assume that $\mathbf{w}$ and $\mathbf{v}_k$ are zero mean random variables with covariances $\mathbf{C}_w$ and $\mathbf{C}_{vk}$, respectively. This provides a full statistical description if these random variables are normally distributed. To keep the notation simple, we assume that $\mathbf{w}$ and $\mathbf{v}_k$ are mutually uncorrelated and white (uncorrelated in time). In the appendix we show how temporally correlated model error can be accomodated with the technique of state augmentation. These various statistical assumptions collectively convey our prior knowledge about the measurement and model errors.

The EnKF moves sequentially from one measurement time to the next and divides naturally into two steps: a forecast step and an update (or analysis) step. We initialize the EnKF by generating an ensemble of initial condition fields $\mathbf{Y}^i(t = 0)$, $i = 1, \ldots, N_e$, around a mean $\overline{\mathbf{Y}(t = 0)}$ with covariance $\mathbf{C}_{Y0}$. This reflects our prior knowledge of the state at the initial time. In the forecast step, the ensemble is propagated forward in time with the nonlinear model Eq. (1) using a corresponding ensemble of N_e (synthetic) model error fields $\mathbf{w}^i$. To generate these spatially correlated random fields, we use a fast Fourier transform method that is very computationally efficient (Robin et al. 1993). The state estimate $\hat{\mathbf{Y}}(t)$ is computed as the mean of the ensemble states $\mathbf{Y}^i(t)$. If full dynamic consistency is a requirement, one could also define the estimate as the particular ensemble member that is closest to the mean in some sense.

At each measurement time t_k, $k = 1,2, \ldots$, when one or more observations become available we update each ensemble member (Evensen 1994):

$$\mathbf{Y}^i_+ = \mathbf{Y}^i_- + \mathbf{K}_k[\mathbf{Z}^i_k - \mathcal{M}_k(\mathbf{Y}^i_-)]. \tag{3}$$

Here, $\mathbf{Y}^i_-$ and $\mathbf{Y}^i_+$ are the state of the ith ensemble member (at time t_k) before and after the update, respectively. The state estimate just before and just after the update are also known as state forecast and analysis, respectively. The perturbed data $\mathbf{Z}^i_k = \mathbf{Z}_k + \mathbf{v}^i_k$ are particular to each ensemble member. They are obtained from the original data by adding a (synthetic) random realization $\mathbf{v}^i_k$ of the measurement error. This ensures that the spread of the updated ensemble is consistent with the true (posterior) estimation error covariance (Burgers et al. 1998). The matrix of weights (or Kalman gain) $\mathbf{K}_k$ for the optimal update is given by

$$\mathbf{K}_k = [\mathbf{C}_{YM}(\mathbf{C}_M + \mathbf{C}_v)^{-1}]_{t=t_k}, \tag{4}$$

and depends on the forecast error covariances, which are obtained directly from the ensemble prior to the update. In particular, $\mathbf{C}_M(t_k)$ is the (forecast) error covariance matrix of the measurement predictions $\mathcal{M}_k[\mathbf{Y}(t_k)]$, and $\mathbf{C}_{YM}(t_k)$ is the (forecast) cross covariance between the state $\mathbf{Y}(t_k)$ and the measurement predictions $\mathcal{M}_k[\mathbf{Y}(t_k)]$. These forecast error covariances vary in time because they depend on the dynamics and all the data included in the previous updates. In the EnKF the state error covariance $\mathbf{C}_Y$ is never explicitly needed, but parts or all of it can be computed at any time from the ensemble.

There are various slightly different approaches for propagating and updating the ensemble. Houtekamer and Mitchell (1998) suggest splitting the ensemble into two parts and updating each part with error covariances derived from the other part. This helps to prevent the collapse of the ensemble for small ensemble sizes when no model error is added. Lermusiaux and Robinson (1999) combine the ensemble approach with a dynamic rotation and compression of the state space. We follow the implementation of Keppenne (2000), who builds on Evensen (1994) and Burgers et al. (1998).

3. Land model and synthetic experiments

By its design the EnKF can easily be used with a variety of land surface models. In this paper, we use the land surface model of Reichle et al. (2001b). This choice allows us to use the variational method described by Reichle et al. (2001b) as a benchmark against which we can compare the EnKF.

To assess the merits of the EnKF we conduct a series of experiments with synthetically generated L-band microwave data. These tests are based on the 1997 Southern Great Plains (SGP97) Hydrology Experiment (Jackson et al. 1999) to ensure realistic conditions. Figure 1 shows the model domain, which covers an area of 80 km by 160 km with 16 by 32 pixels at 5-km resolution. The micrometeorological inputs to the model are interpolated from the Oklahoma Mesonet station data. The stations are shown in Fig. 1 together with the land cover and soil texture classes. The synthetic experiment covers a two-week period from 18 June 1997 (day of year 169) to 2 July 1997 (day of year 183). Time steps in the nonlinear model are adaptive and vary from a few seconds up to 30 min depending on soil conditions and forcing data. The tangent–linear and adjoint models of the benchmark variational method use a basic time step of 30 min.

Our model of coupled moisture and heat transport is a typical soil–vegetation–atmosphere transfer scheme. Vertical soil moisture and temperature dynamics are modeled with Richards' equation and the force–restore approximation, respectively, while the vegetation layer is treated with diagnostic variables, and fluxes through the canopy are described with a resistance network. The L-band brightness temperature is related to the land sur-

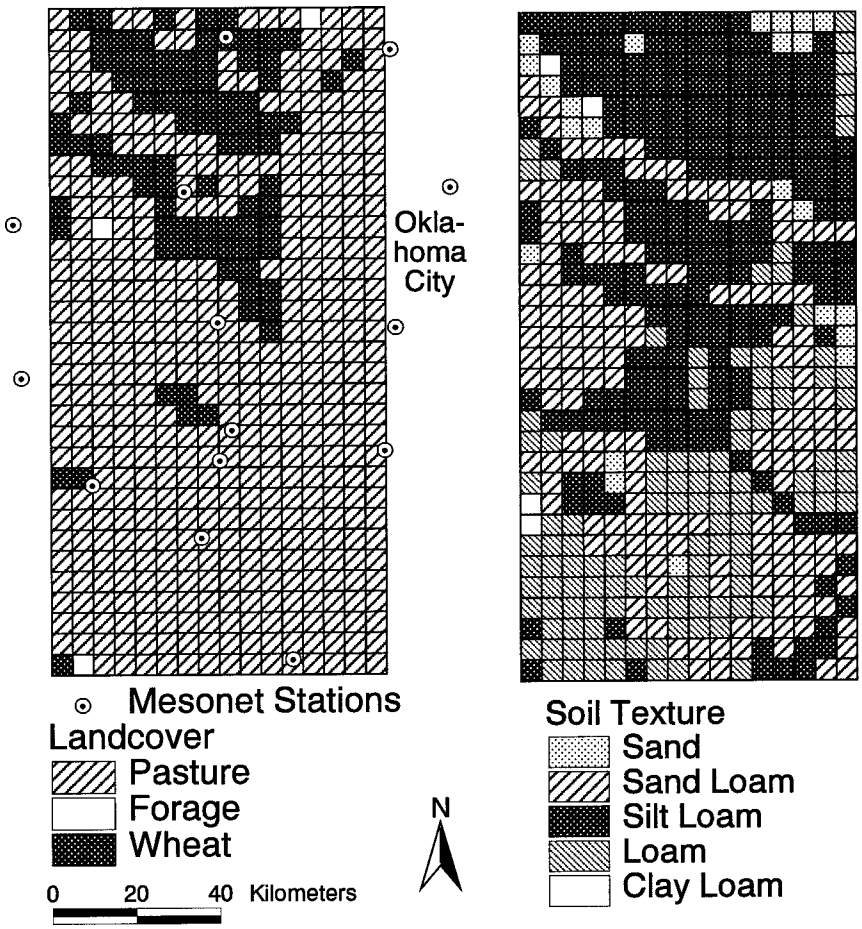

FIG. 1. Area for the synthetic experiment (from Reichle et al. 2001a).

face states with a graybody radiative transfer model. For details, see Reichle (2000) and Reichle et al. (2001a). Here, we use six vertical nodes at 0, −5, −15, −30, −55, and −90 cm for the soil moisture and a single layer of 5-cm thickness for the soil temperature. We assume that lateral moisture and heat fluxes in the unsaturated zone are negligible. As a result, horizontal structure in our soil moisture estimates reflects structure in micrometeorological inputs, land cover, soil texture, and spatial correlation of the respective errors.

We start the synthetic experiment by generating one set of ''true'' initial condition fields for soil moisture and temperature as well as time-dependent random model error fields. The corresponding model trajectory is defined to be the set of true system states. The ''open loop'' or ''prior'' state is the solution of Eq. (1) obtained when the initial conditions and model errors are set to their prior values $\overline{\mathbf{Y}(t = 0)}$ and $\overline{\mathbf{w}(t)} \equiv 0$, respectively. This open loop (or prior) solution can be viewed as a ''first guess'' of the true states available without the benefit of microwave measurements. Recall that the forcing data (including observed precipitation) are represented in the deterministic part of the model Eq. (1). The model errors account for uncertainties in the forcing data.

The spatial and temporal correlation functions of the uncertain inputs are unknown a priori and very difficult to characterize. Their determination for a given model and field setting constitutes a research project in its own right and is well beyond the scope of this paper (but see section 4c). Here, we only aim to prove the concept of soil moisture assimilation with the EnKF. This does not critically depend on the exact shapes and scales of the correlation functions, and we specify conditions that in our experience are appropriate for the experiment area and our model.

The initial condition and model errors generated in our synthetic experiments are normally distributed random fields with Gaussian and exponential correlation functions for space and time, respectively. While the total amount of water that is stored across the column at the initial time is uncertain, we prescribe the shape of the initial soil moisture profile. This makes the benchmark variational estimation more robust (Reichle et al. 2001b). The shape of the initial profile can be specified arbitrarily. For convenience, we choose a hydrostatic profile (no vertical moisture flow, pressure gradient balances gravity). In this experiment, the initial condition of the top node saturation has a sample standard devi-

ation of 0.084 and a horizontal correlation length of 10 km. The initial upper-layer soil temperature is set equal to the initial air temperature and is assumed to be known perfectly. The memory of the upper-layer soil temperature is only a few hours and its initial condition has little impact on longer-term estimates.

Model errors are represented as unknown fluxes in the near-surface soil moisture and soil energy balance equations. We assume that each of these errors is zero mean with a standard deviation of 50 W m^{-2}. We further assume that the model error has a Gaussian spatial covariance with a horizontal correlation length of 15 km and an exponential temporal covariance with a correlation time of 3 days. Note that we use the state augmentation technique described in the appendix to take these temporal correlations into account. In each pixel we have six soil moisture nodes, one soil temperature layer, and two temporally correlated model error components, resulting in an augmented state vector dimension of $9 \times 512 = 4608$.

The true brightness temperatures are obtained by running the true states through the radiative transfer model. The vectors $\mathbf{Z}_k$ of synthetic brightness temperature measurements are obtained by adding random measurement errors. A daily synthetic brightness temperature value is generated at every pixel in the model domain at 1000 local time for a total of 14 observation times (days 169.67, 170.67, . . . , 182.67). This yields 7168 scalar data points. The random measurement errors added to the brightness temperature values are spatially and temporally uncorrelated with a standard deviation of 5 K. This observation time and level of uncertainty are typical of the SGP97 field experiment. Note that measurement errors of satellite observations are likely to be spatially correlated. The absence of such spatial correlations in our synthetic experiment is not a constraint imposed by the algorithm but is a simplification adopted for convenience.

4. Results and discussion

The performance of the EnKF may be measured in a number of ways. One of the most straightforward is to compare the estimate to the ''true'' state, which is known in the synthetic experiment. Figure 2 shows the true (top row), open loop (second row), and estimated (third row) top node saturation across the domain just after selected updates. The estimates are derived with the EnKF using 500 ensemble members. Note that in our definition the soil saturation varies between zero and one. Volumetric soil moisture (volume of water per unit volume of soil) can be obtained from the saturation through multiplication by the porosity of the soil.

A comparison of the first and third rows in Fig. 2 shows that the EnKF is able to recover the true top node saturation from the observations and the prior information. The prior information includes the micrometeorological data and the correct statistics for the initial condition and the model errors. The corresponding prior fields (second row in Fig. 2) are poor estimates of the true conditions. This indicates that auxiliary data alone (soil texture, land cover, micrometeorology) are not sufficient for soil moisture estimation. It should also be noted that the excellent estimates obtained in our synthetic experiment reflect the fact that the model error statistics (mean values and covariances) supplied to the filter are identical to those used to generate the true errors. Since this is unlikely to occur in field applications the actual performance will probably not be as good as observed here (see also section 4c). The synthetic experiment serves primarily to establish a lower bound on the estimation error and to provide a controlled environment for testing the effect of ensemble size and distributional approximations.

The ensemble provides useful statistical information about the filter's internal assessment of the accuracy of its estimates. In particular, Fig. 2 shows the forecast (fourth row) and analysis (bottom row) estimation error standard deviations for the top node saturation as derived from the ensemble just before and just after the update, respectively. The estimation error standard deviations are typically on the order of a few percent saturation, or about 1%–2% in volumetric soil moisture. We discuss the relationship between the actual errors and the error standard deviations generated within the filter later in this section.

The forecast and analysis error standard deviations also demonstrate the value of dynamic covariance propagation. Figure 2 indicates that the forecast error standard deviation varies significantly across the domain and with time. It is also anticorrelated with the saturation. The smallest standard deviation occurs just after rainstorms and generally in wet areas. The standard deviation increases as the soil dries out. For extremely dry conditions, the standard deviation decreases again as for most ensemble members the soil moisture reaches the lower bound. Although the details depend on the actual model error statistics (which we prescribe) the general behavior is mostly governed by nonlinearities in the hydrologic model. For a given uncertainty in the rainfall, for instance, the uncertainty in surface soil moisture is greater for dry conditions, when a small amount of rainfall has a greater impact on surface soil moisture.

The results presented in Fig. 2 indicate that the forecast error standard deviation depends in a complex way on soil moisture, previous observations, and recent forcing. All of these vary over both time and space. The full forecast error covariance needed for the optimal update Eq. (4) is even more complex than the forecast error standard deviation. This suggests that it is unrealistic to expect that forecast error variances can be specified a priori (i.e., without dynamic propagation), as is required in statistical interpolation algorithms (Daley 1991).

a. Convergence with ensemble size

It is useful to consider how the EnKF estimates converge both to the true state and to the benchmark

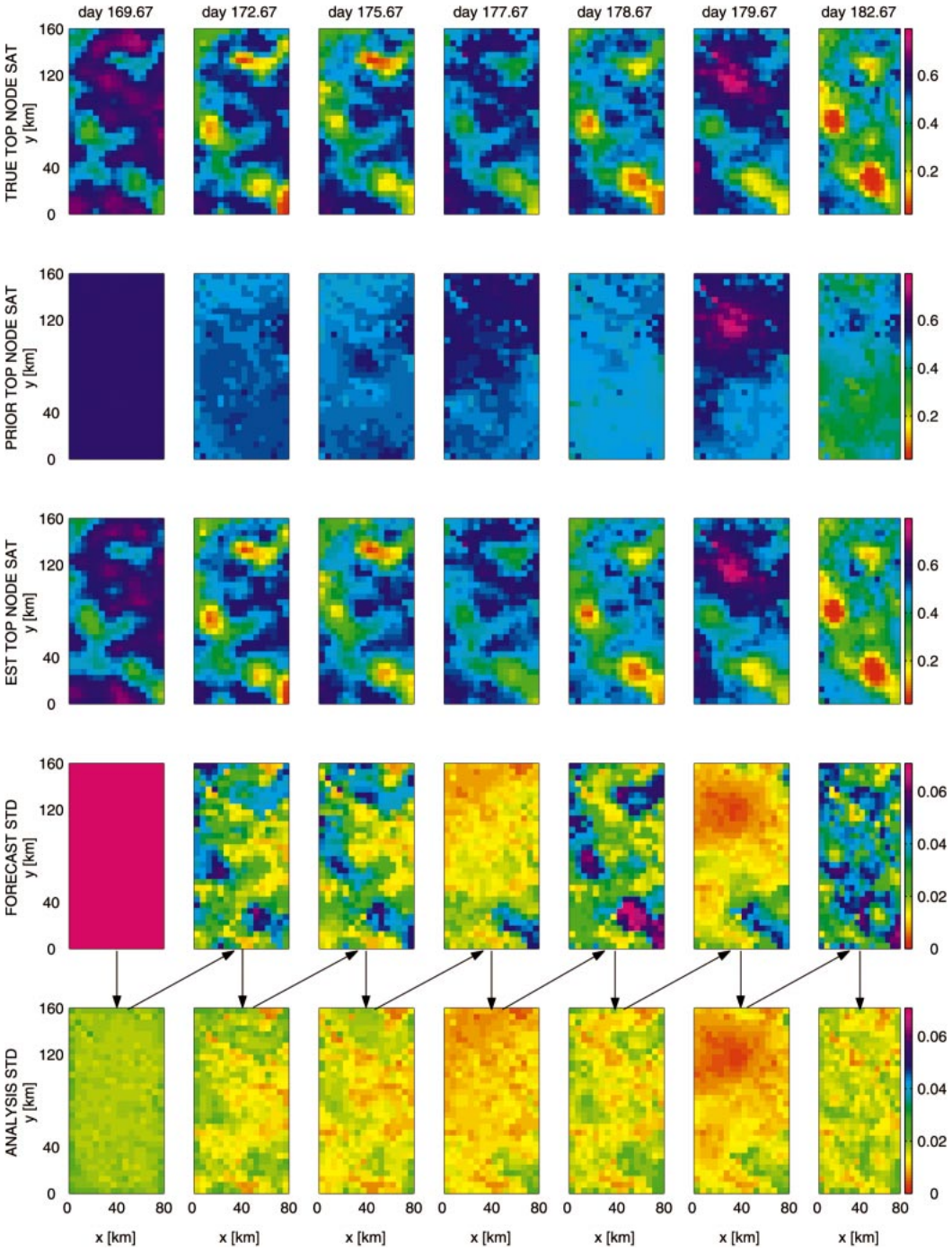

FIG. 2. Top node saturation. (first row) True, (second row) open loop, (third row) EnKF estimate with 500 ensemble members, and corresponding (fourth row) forecast and (bottom row) analysis error standard deviation (std dev). The arrows indicate the temporal order of the error std dev plots.

solution as the ensemble size increases. The benchmark variational approach (Reichle et al. 2001b) uses the iterated indirect representer method to solve the weak-constraint optimal smoothing problem over a specified time period. The state and measurement equations as well as the error statistics used in this optimal smoother are identical to those used in the EnKF experiments.

The conceptual differences between the variational optimal smoother and the EnKF are (i) the smoother

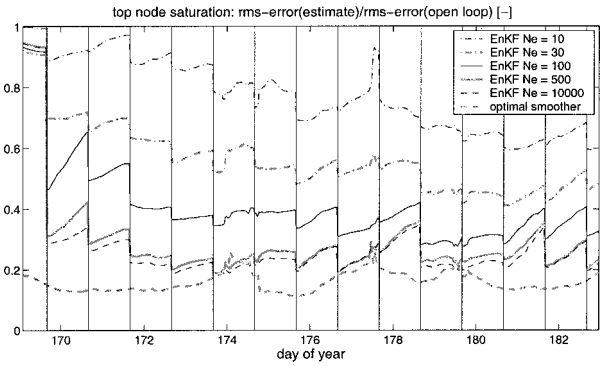

FIG. 3. Area average of the actual top node saturation rms error relative to the open loop error.

uses past and future measurements to estimate the state at any time t within the assimilation interval while the EnKF uses only past measurements obtained through time t, (ii) the EnKF estimate depends on the ensemble size and converges to an optimum only when the number of ensemble members is large, and (iii) in nonlinear problems the variational approach yields the conditional mode while the EnKF yields an approximation to the conditional mean (McLaughlin and Townley 1996). Since the optimal smoother relies on more information than the filter (except at the final time) and its performance is not limited by sampling considerations, we can effectively use it as a benchmark. Moreover, we can isolate the effects of nonlinearity and non-Gaussian forecast errors by comparing the EnKF and variational estimates at the final measurement time (when they both use all the data) and for a large ensemble size.

Our basic measure of estimation error is the difference between the true states and the estimate at any given time and location. A convenient spatially aggregated version of this measure is the root-mean-square (rms) error averaged over all pixels in the study area. The algorithm's estimation error should be compared to the open loop (or prior) error, which is the difference between the true and open loop (or prior) states. As can be expected from the estimation error standard deviation plotted in Fig. 2, the actual estimation errors depend strongly on the state of the system, with errors generally increasing as the soil dries out. This applies even in the open loop case when there is no assimilation. We can filter out the influence of precipitation transients on estimation error if we normalize by the open loop error. More precisely, Fig. 3 shows the rms error of the estimated top node saturation divided by the open loop rms error for 10, 30, 100, 500, and 10 000 ensemble members and for the optimal smoother.

Figure 3 indicates that the actual errors decrease and converge with increasing ensemble size. The area average error for the EnKF with only 30 ensemble members is less than half of the open loop error at the final time. An ensemble of 500 or more yields estimates that are very close to the optimal smoothing result. More quantitatively, for a relatively small ensemble size of 10 (or 30; or 100; or 500) the actual errors in the top node saturation at the final update time are reduced by 42% (or 55%; or 70%; or 80%) from the value obtained without assimilation (as compared to 84% for the optimal smoother).

The actual error of the smoothing estimate remains fairly constant throughout the assimilation interval, because the smoother processes all observations at once for all times. In contrast, the error of the EnKF (filtering) estimates shows an overall decrease with time as more and more observations are assimilated. While the smoothing error does not change predictably between update times, the error of the EnKF estimates typically increases after the update. In other words, the EnKF forecast deteriorates until a new observation becomes available. Since the smoothing estimate relies on both past and future observations it is better able to interpolate between measurement times. As expected, the smoother errors are nearly always smaller than the filter errors.

It is useful to examine the effect of the ensemble size on the spatial distribution of the rms estimation errors. This is done in Fig. 4, which shows contour plots of the actual errors for the top node soil saturation at the final update (day 182.67) for 10, 30, 100, and 500 ensemble members. The area average (rms) error is 0.087, 0.068, 0.046, and 0.030, respectively, compared to an open loop error of 0.15 and an optimal smoothing error of 0.024 (recall that the saturation varies between zero and one). For 100 ensemble members, the errors are confined to a few small areas that are evenly distributed across the domain and the EnKF estimates capture most of the large-scale features of the soil moisture field. The contour plots for more than 500 ensemble members (not shown) are almost identical to the case with 500 ensemble members.

For yet another diagnostic of the filter's performance we can look at the innovations sequence $\boldsymbol{\nu}_k \equiv \mathbf{Z}_k - \mathcal{M}_k[\hat{\mathbf{Y}}(t_k)]$ (actual minus predicted measurements). This sequence must be white (uncorrelated in time) if the filter operates in accordance with its underlying statistical assumptions. We have conducted a test for whiteness based on the autocorrelation function of $\boldsymbol{\nu}_k$ (Jenkins and Watts 1968). For small ensemble sizes ($N_e = 10$, 30), we must reject the null hypothesis that the $\boldsymbol{\nu}_k$ sequence is white. For larger ensembles ($N_e \geq 100$), there is no indication that the $\boldsymbol{\nu}_k$ are temporally correlated. Although we have a limited sequence with only 14 update times, we believe that the result is indicative of the near-optimal behavior of the EnKF for modest ensemble sizes.

b. Nonlinearities and deviations from Gaussian distributions

It is instructive to note that the optimal smoother generates slightly better estimates even at the final up-

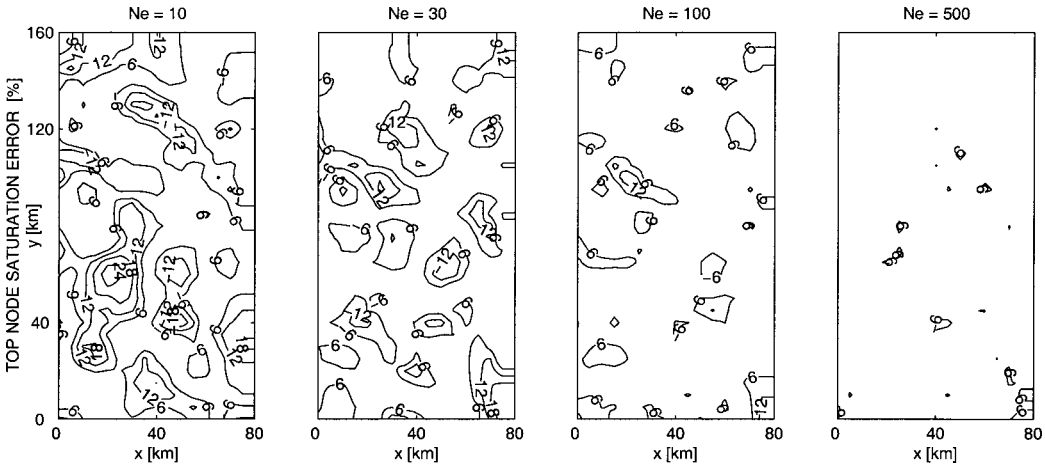

FIG. 4. Actual top node saturation error at the final update (day 182.67) for 10, 30, 100, and 500 ensemble members.

date, when the EnKF and the smoother have both used all observations. The question arises whether this difference is due to the effect of nonlinearities (difference iii in section 4a) or whether it is just statistical noise owing to the finite ensemble size (difference ii in section 4a). If for a moment we suppose that the smoothing estimates are ultimately optimal estimates, we can extrapolate the errors of the EnKF estimates to get a rough idea of the number of ensemble members that would be necessary to eliminate the statistical noise as the source of the discrepancy. Figure 5 plots the difference of the mean-square-errors (mse) of the EnKF and the optimal smoothing estimates versus ensemble size, where the mse is the area average square error of the top node saturation at the last update.

Figure 5 shows two regression lines to the error data. The first line applies for small and medium ensemble sizes (the first six data points) and has a slope of -0.81 ($R^2 = 0.99$). For larger ensemble sizes, however, the error difference does not decrease any further, suggesting that the EnKF estimates at the final time do not converge to the optimal smoothing solution when the ensemble size becomes very large. This is confirmed by the second regression line, which applies for large ensemble sizes and includes the last three data points. The slope of this line is 0.034 ($R^2 = 0.97$), which is not significantly different from zero.

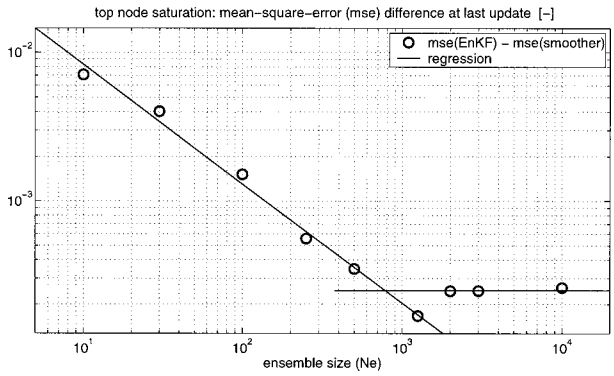

FIG. 5. Difference of the actual mse of the EnKF and the smoothing estimates at the final update vs ensemble size.

Since the effects of smoothing versus filtering and the limited ensemble size have been eliminated, we attribute the residual difference between the optimal smoothing and EnKF errors to the nonlinear nature of the model and measurement processes. The difference could reflect the combined influence of two factors. First, the variational method is designed to estimate the conditional mode while the EnKF is designed to estimate the conditional mean. These two distributional properties are generally different if the conditional probability density function (pdf) of the forecast is not symmetric. Second, the EnKF estimate may be an inaccurate approximation of the conditional mean because the update step Eq. (3) (via the weights $\mathbf{K}_k$) relies on only the first two moments of the conditional forecast pdf. This leads to estimates that are less than optimal (in the sense that they do not use all the information available in the ensemble).

Further insight on this issue can be gained if we examine the performance of the EnKF by comparing actual area average errors of the top node saturation with the error variances computed internally by the filter. For 500 ensemble members the filter yields forecast and analysis error variances that are consistently lower than the corresponding actual errors (Fig. 6). In other words, the filter is too optimistic about the quality of the estimates. While the actual area average errors decrease with increasing ensemble size, the forecast and analysis error variances (i.e., the ensemble spread) increase with increasing ensemble size. But the discrepancy between the actual and the expected errors does not disappear even when 10 000 ensemble members are used (not shown). The likely reason for this behavior is again the

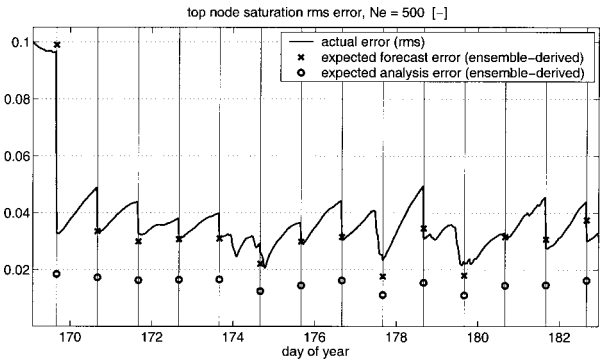

FIG. 6. Area average of the actual top node saturation error and corresponding forecast and analysis error standard deviations for 500 ensemble members.

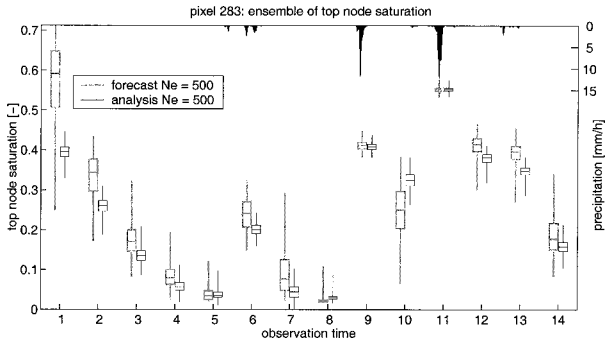

FIG. 7. Box plot of the ensemble distribution of the top node saturation for a representative pixel and 500 ensemble members. Each box shows the lower quartile, the median, and the upper quartile, while the whiskers indicate the minimum and the maximum. Precipitation is sketched from the top (right-hand scale).

effect of model nonlinearities, which produce non-Gaussian conditional pdf's. As mentioned above, the EnKF update is suboptimal in this case. Consequently, the actual errors are larger than expected.

We can further examine the significance of non-Gaussian behavior by looking at the distribution of the saturation across the ensemble at a particular location just before and just after each update. Figure 7 shows corresponding box plots for the top node saturation at a representative pixel. Each box shows the lower quartile, the median, and the upper quartile, while the whiskers show the minimum and the maximum. For reference, the time series of precipitation is sketched in the top of Fig. 7. In general, the boxes (and distributions) appear to be reasonably symmetric. For dry conditions, however, the distribution becomes skewed because of the lower bound of the saturation. For very dry conditions, the ensemble practically collapses. While the EnKF is able to fully propagate any asymmetry that might arise between update times, the update step ignores this asymmetry and relies on only the first two moments of the distribution. This eventually results in estimates that are less than optimal. To remedy the situation, a fully nonlinear filter would have to be used. One could, for instance, compute a Bayesian update of the conditional pdf (Anderson and Anderson 1999). Such an approach, however, is not computationally feasible for large-scale applications.

c. *Model error estimates and model error covariance*

Throughout this paper, we have used temporally correlated model error and modified the filter by augmenting the state in order to take the temporal correlations into account (appendix). This implies that we can also estimate the time series of the model error $\mathbf{w}$ from the ensemble. Figure 8 shows the model error estimates for the moisture flux upper boundary condition at a representative pixel. Although the optimal smoother clearly produces a better estimate of the model error, the EnKF model error estimates are quite reasonable. At each observation time, the EnKF model error estimate is updated, but after that the estimate necessarily tends toward zero (the mean model error) until the next observation becomes available.

Any data assimilation approach that provides for model error (including the two approaches considered in this paper) faces the serious challenge of determining the true model error covariance $\mathbf{C}_w$ in operational applications. The task of deriving an appropriate model error covariance is complicated by the scarcity of validation data. Since the model error partly represents uncertainties in the observed forcing data (such as precipitation), measurement and interpolation errors in these data can be incorporated into the model error covariance. But the key to this problem is more likely the innovations sequence $\boldsymbol{\nu}_k$ of actual minus predicted measurements, which can be computed in an operational setting. Through close examination of the innovations we may be able to derive and validate estimates of the model error covariance in a field setting (Dee 1995). Encouraging results on the sensitivity of the soil moisture estimates to the quality of the model error covariance can be found in Reichle, (2000) and Reichle et al. (2001a).

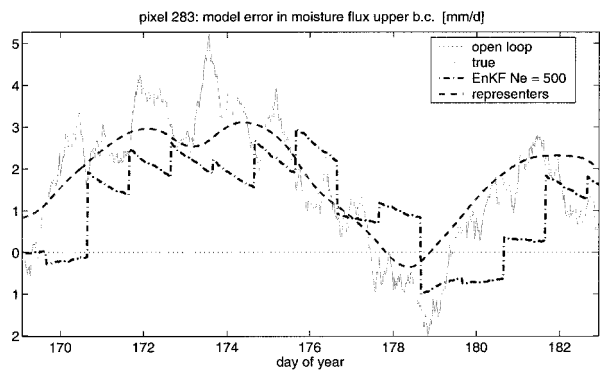

FIG. 8. Model error estimates for a representative pixel.

TABLE 1. Computational effort of the EnKF relative to the optimal smoother for our soil moisture example.

Ensemble size N_e	10	30	100	250	500	1250	2000	3000	10 000
Relative effort [−]	0.014	0.034	0.11	0.29	0.59	1.6	2.8	4.7	22

d. Operational prospects of the EnKF and the representer method

Since we have used a four-dimensional variational approach to benchmark the performance of the EnKF, it is reasonable to ask how the two methods are likely to compare in an operational setting. It is obviously difficult to generalize from a particular synthetic experiment. However, the EnKF and the variational (representer) approach each have distinctive features that can be expected to apply over a range of different problems. These are summarized in the following paragraphs.

The EnKF is an inherently sequential algorithm that is easy to use in real-time forecasting applications because measurements are processed as they become available. Reinitialization of the algorithm at measurement times is an inherent part of the EnKF and does not require any special treatment. In contrast, the variational approach simultaneously processes measurements taken at different times. In a long-term application the time period of interest must be partitioned into shorter smoothing intervals, which are processed as separate datasets. In order to obtain optimal estimates at the final time, each interval must be initialized with the correct initial condition error covariance, which is equal to the estimation error covariance at the final time of the preceding interval. Unfortunately, the computational effort required to obtain these covariances with the variational approach is overwhelming for problems of realistic size. Therefore, the reinitialization of the intervals must include approximations that can compromise the optimality of the variational estimates.

The analysis presented in section 4b suggests that the variational method may be better able to deal with nonlinearities in the state and measurement equations because it does not assume that the forecast conditional pdf is Gaussian. The EnKF is limited by its assumption that the first two moments adequately characterize the conditional forecast pdf. The practical consequences of this limitation can be expected to be application dependent. In our application, they probably have greatest impact when the soil is very dry and the forecast error distribution is highly skewed.

The EnKF offers great flexibility with respect to the form of model error that can be included. Such errors may be additive, multiplicative, or state dependent. Errors in model structure can even be accommodated if different ensemble members are generated from different models. The variational method can include model error but is less flexible because errors other than the additive errors generally complicate the formulation and considerably increase computational effort.

The EnKF has a number of implementation advantages that are worth mentioning. Since the EnKF does not rely on any linearizations, it does not require derivation of an adjoint equation or computation of model or measurement operator derivatives. This is in contrast to the extended Kalman filter and variational techniques, which rely on derivatives that need to be recomputed at each time step or iteration. The EnKF's freedom from an adjoint code is particularly important in soil moisture assimilation because there is no adjoint available to date for any of the commonly used land models. Deriving the adjoint of a land surface scheme is difficult because of the nonlinearity of the land processes and the many switches and discontinuities that are typically used (but not always well documented) in land models.

One of the attractive aspects of the EnKF is the computational advantage it may offer if the number of ensemble members can be kept sufficiently small. The computational load for the EnKF experiments is summarized in Table 1. The effort for each EnKF experiment is given relative to the effort for the optimal smoother, which reflects an equivalent of 1323 integrations of the tangent-linear or adjoint model (Reichle et al. 2001b). Table 1 shows that the EnKF with 100 ensemble members is more than nine times faster than the optimal smoother. With 500 ensemble members, the EnKF is still almost twice as fast as the smoother, while the EnKF with 3000 ensemble members takes about five times as long as the smoother. Approximately 10% of the computer time for the EnKF is taken up by the generation of model error replicates.

Computational requirements for the EnKF with moderate numbers of ensemble members are considerably less than for the variational smoother. However, it is still unclear how the ensemble size that is required for adequate estimates scales with the size of the problem. A similar comment applies to the representer method (Reichle et al. 2001b). The memory requirements for both methods are substantial. For the EnKF memory demand scales with $N_Y \times N_e$. The memory use of the representer approach is typically higher and scales with $N_Y \times N_t$, where N_t denotes the number of time steps within each smoothing interval.

Both methods are well suited for parallel implementation, especially in the soil moisture application where the land surface model is divided into many parallel one-dimensional columns. The advantages of parallel computation are likely to be even greater when updates are regionalized over space (Keppenne 2000; Reichle 2000). For the EnKF, there could be additional gains by developing very efficient methods for selecting (''breeding'') ensemble members. If the statistical information

needed to obtain accurate estimates can be captured by a small ensemble of informative members, the EnKF may be very attractive.

5. Summary and conclusions

In this paper, we discuss the application of the ensemble Kalman filter to hydrologic data assimilation and in particular to the estimation of soil moisture from L-band microwave brightness temperature observations. We also compare the performance of the EnKF to an optimal smoother (weak-constraint variational algorithm). Both methods are applied to the same problem and use identical state and measurement equations, error statistics, and synthetically generated measurements. We conclude that with relatively few ensemble members the EnKF yields reasonable soil moisture estimates. For a state vector dimension of 4608 and a relatively small ensemble size of 30 (or 100; or 500), the actual errors in surface soil moisture at the final update time decrease by 55% (or 70%; or 80%) from the value obtained without assimilation (as compared to 84% for the optimal smoother).

The EnKF significantly underestimates the forecast error variances for 100 ensemble members. However, the error variance estimates derived by the filter are reasonably good when the ensemble size is increased to 500 members. Our results indicate that the forecast error variances vary strongly with time and space. This implies that it is very important to account for dynamic error covariance propagation. Assimilation schemes that use static forecast error covariances (e.g., statistical interpolation) are unlikely to produce the desired near-optimal estimates.

More research is required to better understand the EnKF and its variants. In particular, better understanding is needed of the role of nonlinearities and related asymmetries in the conditional forecast probability density function. We have found that nonlinearities in the model and measurement processes contribute to differences in the filtering and smoothing estimates even at the final update. In our application the state (soil moisture) is bounded above and below and its distribution cannot always be well approximated by a Gaussian pdf. For very wet or dry conditions, in particular, the soil moisture pdf exhibits considerable skewness. It is likely that the variational smoother is superior to the EnKF when dealing with nonlinear and non-Gaussian effects. However, it is important to recognize that the variational approach is designed to estimate the mode of the conditional forecast density while the EnKF is designed to estimate the mean. So even if both approaches work as intended their estimates at the end of the smoothing interval can be expected to differ when the density is asymmetric.

In a practical application of the EnKF, it will probably be necessary to model the forecast covariance rather than to compute it in an exact sense. In the EnKF, covariance modeling could include smoothing out the ensemble-derived covariances before the update or applying the update to subregions of the computational domain. Ultimately, a hybrid filter that combines empirical forecast error covariances with dynamic error propagation via the EnKF (Hamill and Snyder 2000) may be the best approach.

A task that is closely related to the determination of the model error covariance is to find a good way to select or ''breed'' the members of the ensemble. In this paper, we have simply used a standard random field generator that produced synthetic model error fields based on our specified $\mathbf{C}_w$. In operational applications, one might also perturb key parameters of the model such as soil hydraulic parameters or even use a number of different models for different subsets of the ensemble. It may also be possible to reduce the ensemble size by using model compression or rank reduction techniques to generate ensemble members that effectively span an appropriate subspace of the state space.

There is no doubt that one of the most attractive features of the EnKF is its flexibility. It can be tested with many different state and measurement equations with no need to compute adjoint models or derivatives. It can handle a wide range of model errors. Users can readily trade off estimation accuracy and computational effort by simply adjusting the number of ensemble members. However, it is too early to say how the ensemble filtering approach will scale with problem size. It is also too early to make a definitive comparison between the ensemble and variational approaches. However, it is likely that the ''best'' approach to a given data assimilation problem will be application dependent and will combine aspects of ensemble and more traditional methods.

Acknowledgments. This research was sponsored by the NASA Land Surface Hydrology Program (NRA-98-OES-11). We would also like to thank the Earth System Science Center at The Pennsylvania State University and the Oklahoma Mesonet for their invaluable data.

APPENDIX

Model Error and State Augmentation

If the error $\mathbf{w}$ of the hydrologic model Eq. (1) is correlated in time, the EnKF described in section 2 must be modified with a technique known as state augmentation (Gelb 1974). Here, we assume that $\mathbf{w}$ is a first-order Markov process with covariance $\overline{\mathbf{w}(t+\tau)\mathbf{w}(t)} = \mathbf{C}_w \exp(-\rho|\tau|)$, where ρ^{-1} is the correlation time. This process can be represented with the differential equation $d\mathbf{w}/dt = -\rho\mathbf{w} + \mathbf{u}$, where $\mathbf{u}$ is white with appropriate spatial covariance. By augmenting Eq. (1) we get

$$\frac{d}{dt}\begin{bmatrix}\mathbf{Y}\\ \mathbf{w}\end{bmatrix} = \begin{bmatrix}\mathcal{F}(\mathbf{Y}) + \mathbf{w}\\ -\rho\mathbf{w}\end{bmatrix} + \begin{bmatrix}0\\ \mathbf{u}\end{bmatrix}, \tag{A1}$$

which serves as the new state equation with $[\mathbf{Y}\ \mathbf{w}]^{\mathrm{T}}$ as

the new state vector and $[0\ \mathbf{u}]^T$ as the new (white) model error. Obviously, the dimension of the state vector has increased. This requires additional computational effort when the update is computed. Note, however, that temporal correlations in the model error entail little additional effort for the propagation of the ensemble.

REFERENCES

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.,* **127,** 2741–2758.

Bouttier, F., J.-F. Mahfouf, and J. Noilhan, 1993: Sequential assimilation of soil moisture from atmospheric low-level parameters. Part II: Implementation in a mesoscale model. *J. Appl. Meteor.,* **32,** 1352–1364.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.,* **126,** 1719–1724.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.,* **123,** 1128–1145.

Errico, R. M., 1999: Workshop on assimilation of satellite data. *Bull. Amer. Meteor. Soc.,* **80,** 463–471.

——, G. Ohring, J. Derber, and J. Joiner, 2000: NOAA–NASA–DoD workshop on satellite data assimilation. *Bull. Amer. Meteor. Soc.,* **81,** 2457–2462.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99** (C5), 10 143–10 162.

——, and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.,* **124,** 85–96.

Gelb, A., Ed.,1974: *Applied Optimal Estimation.* The MIT Press, 374 pp.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter–3D variational analysis scheme. *Mon. Wea. Rev.,* **128,** 2905–2919.

——, S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81,** 2653–2664.

Houser, P. R., W. J. Shuttleworth, J. S. Famiglietti, H. V. Gupta, K. H. Syed, and D. C. Goodrich, 1998: Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resour. Res.,* **34,** 3405–3420.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

Jackson, T. J., D. M. Le Vine, A. Y. Hsu, A. Oldak, P. J. Starks, C. T. Swift, J. D. Isham, and M. Haken, 1999: Soil moisture mapping at regional scales using microwave radiometry: The Southern Great Plains Hydrology Experiment. *IEEE Trans. Geosci. Remote Sens.,* **37,** 2136–2151.

Jenkins, G. M., and D. G. Watts, 1968: *Spectral Analysis and Its Applications.* Holden-Day, 525 pp.

Keppenne, C. L., 2000: Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Wea. Rev.,* **128,** 1971–1981.

Koster, R. D., M. J. Suarez, and M. Heiser, 2000: Variance and predictability of precipitation at seasonal to interannual timescales. *J. Hydrometeor.,* **1,** 26–46.

Lermusiaux, P. F. J., 1999: Data assimilation via error subspace statistical estimation. Part II: Middle Atlantic Bight shelfbreak front simulations and ESSE validation. *Mon. Wea. Rev.,* **127,** 1408–1432.

——, and A. R. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.,* **127,** 1385–1407.

Madsen, H., and R. Canizares, 1999: Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling. *Int. J. Numer. Methods Fluids,* **31,** 961–981.

McLaughlin, D., and L. R. Townley, 1996: A reassessment of the groundwater inverse problem. *Water Resour. Res.,* **32,** 1131–1161.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.,* **51,** 1037–1056.

Njoku, E. G., and D. Entekhabi, 1995: Passive microwave remote sensing of soil moisture. *J. Hydrol.,* **184,** 101–130.

Paegle, J., K. C. Mo, and J. Nogués-Paegle, 1996: Dependence of simulated precipitation on surface evaporation during the 1993 United States summer floods. *Mon. Wea. Rev.,* **124,** 345–361.

Pan, Z., M. Segal, R. Turner, and E. Takle, 1995: Model simulation of impacts of transient surface wetness on summer rainfall in the United States Midwest during drought and flood years. *Mon. Wea. Rev.,* **123,** 1575–1581.

Reichle, R. H., 2000: Variational assimilation of remote sensing data for land surface hydrologic applications. Ph.D. dissertation, Massachusetts Institute of Technology, 192 pp.

——, D. Entekhabi, and D. B. McLaughlin, 2001a: Downscaling of radiobrightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. *Water Resour. Res.,* **37,** 2353–2364.

——, D. B. McLaughlin, and D. Entekhabi, 2001b: Variational data assimilation of microwave radiobrightness observations for land surface hydrologic applications. *IEEE Trans. Geosci. Remote Sens.,* **39,** 1708–1718.

Rhodin, A., F. Kucharski, U. Callies, D. P. Eppel, and W. Wergen, 1999: Variational analysis of effective soil moisture from screen-level atmospheric parameters: Application to a short-range weather forecast model. *Quart. J. Roy. Meteor. Soc.,* **125,** 2427–2448.

Robin, M. J. L., A. L. Gutjahr, E. A. Sudicky, and J. L. Wilson, 1993: Cross-correlated random field generation with the direct Fourier transform method. *Water Resour. Res.,* **29,** 2385–2397.

Shaw, B. L., R. A. Pielke, and C. L. Ziegler, 1997: A three-dimensional numerical simulation of a Great Plains dryline. *Mon. Wea. Rev.,* **125,** 1489–1506.