A Multiscale Ensemble Filtering System for Hydrologic Data Assimilation. Part I: Implementation and Synthetic Experiment

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1175/2009JHM1088.1">http://dx.doi.org/10.1175/2009JHM1088.1</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>American Meteorological Society</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Jan 10 21:32:59 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/51791">http://hdl.handle.net/1721.1/51791</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
A Multiscale Ensemble Filtering System for Hydrologic Data Assimilation.  
Part I: Implementation and Synthetic Experiment

MING PAN AND ERIC F. WOOD
Princeton University, Princeton, New Jersey

DENNIS B. MCLAUGHLIN AND DARA ENTEKHABI
Massachusetts Institute of Technology, Cambridge, Massachusetts

LIFENG LUO
Princeton University, Princeton, New Jersey

(Manuscript received 2 September 2008, in final form 29 December 2008)

ABSTRACT
The multiscale autoregressive (MAR) framework was introduced in the last decade to process signals that exhibit multiscale features. It provides the method for identifying the multiscale structure in signals and a filtering procedure, and thus is an efficient way to solve the optimal estimation problem for many high-dimensional dynamic systems. Later, an ensemble version of this multiscale filtering procedure, the ensemble multiscale filter (EnMSF), was developed for estimation systems that rely on Monte Carlo samples, making this technique suitable for a range of applications in geosciences. Following the prototype study that introduced EnMSF, a strategy is devised here to implement the multiscale method in a hydrologic data assimilation system, which runs a land surface model. Assimilation experiments are carried out over the Arkansas–Red River basin, located in the central United States (−645 000 km²), using the Variable Infiltration Capacity (VIC) model with a computing grid of 1062 pixels. A synthetic data assimilation experiment is performed, driven by meteorological forcing fields downscaled from the ensemble forecasts made by the NOAA/National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS). The classic full-rank ensemble Kalman filter is used as the benchmark to evaluate the multiscale filter performance, and comparisons are also made with a horizontally uncoupled filter. It is demonstrated that the multiscale filter is able to closely approximate the full-rank solution with a low computational cost (~1/20 of the full-rank filter) in an experiment in which the top-layer soil moisture is assimilated, whereas the horizontally uncoupled filter fails to approximate the full-rank solution.

1. Introduction

Data assimilation is essentially to optimally merge information from different sources. As hydrology becomes increasingly data driven in the recent decades, data assimilation also emerges as a powerful tool for hydrologists to integrate a vast amount of information and make improved model forecasts and projections (McLaughlin 2002). Many factors contribute to this trend, including the fast-growing capacity of numerical models to simulate hydrologic processes (Mitchell et al. 2004), and the significant expansion of available remote sensing techniques and measurements, led by the National Aeronautics and Space Administration's (NASA) Earth Observing System (EOS; McCabe et al. 2008), which provide estimates of hydrologic variables at unprecedented large scales and fine resolutions. At the same time, challenges are presented by their use in hydrologic data assimilation, and they are mostly tied to the nature of the hydrologic dynamic systems, for example, nonlinear dynamics, non-Gaussian error distributions (Pan et al. 2008), and high dimensionality (Zhou et al. 2008).

Efforts have been made in many methodology studies, such as on the nonlinear dynamics with the introduction of the ensemble Kalman filter (EnKF; Evensen...
1994; Burgers et al. 1998) and on non-Gaussian errors with particle filters (Zhou et al. 2006). Our study focuses on the high-dimensionality issue. In hydrology, land surface models (LSMs) usually discretize the target area (e.g., a catchment, a river basin, or a continent) into many small pixels (computing grids) and solve for the moisture/energy states on each grid. Similarly, in remote sensing, data are collected pixel by pixel as the sensor antenna scans across the earth’s surface. The number of pixels multiplied by the number of states on each pixel, or the total number of discrete states on these pixels, determines the dimension of the data assimilation problem (denoted as \(d\) hereafter), which may reach \(10^3–10^6\) in various applications (Zhou et al. 2008).

Mathematically, data assimilation is usually posed in a probabilistic framework as an optimal estimation problem (e.g., maximum a posteriori probability), or in a variational framework as a cost function minimization problem. Either way, the solution procedure involves mathematical computations (inversion, square root operations, decomposition, etc.) of error covariance matrices with a dimension equal to the problem size \(d\), for example, in the EnKF algorithm (Evensen 1994). Such operations generally have the complexity \(O(d^3)\), which scales up rapidly as \(d\) increases, making large problems enormously difficult to solve. However, most \(d\)-dimensional state estimation problems from natural processes would not have a full-rank \(d\)-dimensional error covariance matrix. Instead, there is always some structure in these states, and such structure may be formulated in a low-dimensional framework. Researchers have been taking advantage of these special structures in the states to greatly simplify high-dimensional problems.

The multiscale structure of states is perhaps the most studied area because a wide range of natural processes show multiscale properties in space and/or time. The wavelets theory (Daubechies et al. 1992) is arguably the best example. Among many, the multiscale autoregressive (MAR) framework introduced in the 1990s (Willsky 2002; Frakt and Willsky 2001) is a very flexible multiscale model for describing states ranging from digital images to maps of hydrologic variables (Willsky 2002). What makes the MAR framework more powerful is the filtering procedure developed for multiscale state systems (Kumar 1999), which replaces a high-dimensional filtering problem with many low-dimensional localized filtering problems defined across scales and provides an extremely efficient solution to high-dimensional filtering problems. More recently, Zhou et al. (2008) adapted the multiscale filter such that it works in the ensemble form, that is, the ensemble multiscale filter (EnMSF), fusing the power/advantages of both the EnKF and multiscale methods.

The EnMSF is very suitable for hydrologic applications, because optimal state estimation problems in hydrology have nonlinear dynamics, which are usually solved using an ensemble (or Monte Carlo) algorithm (Crow and Wood 2003). The MAR algorithm is fully established in Zhou et al. (2008), including both the identification of the multiscale structure from input state ensembles using the method of maximum predictive efficiency (MPE; Frakt and Willsky 2001) and the two-sweep filtering procedure operating on state ensembles across scales. The formulation to handle measurements from coarser scales is also introduced, and a data assimilation experiment on a numerical model solving Navier–Stokes equations in a rectangular domain is performed in Zhou et al. (2008). Our study presented here aims at developing and implementing the EnMSF for hydrologic land surface model–driven applications and solving the challenges in applying EnMSF for such systems.

In the following sections, we will review the background on the ensemble-based multiscale methodology and then propose ways to solve issues related to implementing the EnMSF with hydrologic land surface models. A set of EnMSF data assimilation experiments follows.

2. Background on the multiscale model: MAR and EnMSF

We present a brief review on the multiscale methodology used in this study. For additional details not addressed here, refer to Zhou et al. (2008) and Willsky (2002).

a. MAR model and its identification using MPE

For a state variable \(x\) of dimension \(n\) that is distributed over \(n\) discrete locations (i.e., a computing grid), the MAR model summarizes it at different “scales.” That is, the original state variables are aggregated over windows of a certain size to form a “coarser” scale state signal, and this aggregation continues at coarser and coarser scales until the entire state vector is summarized as one “coarsest” state variable. This creates a “multiscale tree,” as shown in Fig. 1a. The conceptual depiction of a multiscale tree in Fig. 1a is usually drawn upside down: the root (coarsest-scale node) is on the top and the “leaves” (finest-scale nodes), which are mapped to the original state variable in space/time, are at the bottom.

We denote a node on the multiscale tree as \(s\), its parent as \(s_j\), and its \(q\) children as \(s_{q1}, s_{q2}, \ldots, s_{qq}\). The state at node \(s\) is denoted as \(\chi(s)\) (see Fig. 1a). The dimension of vector \(\chi(s)\) can be arbitrarily chosen, but we always use a small dimension for \(\chi(s)\) (which is much smaller than the total dimension of its children), because the purpose of this methodology is to break the large...
problem into many small problems. Every finescale node is mapped to one or more locations in the original computing grid domain. Suppose the multiscale tree has $M$ scales and we use $x$ for all the nodes at the finest scale, then $x/x^M$. Mapping examples can be found in Figs. 2 and 3. The dimension of $x$ is large, and so is the dimension of its error covariance $\text{cov}(x)$. Note that for simplicity, we use the notation $\text{cov}(x)$ for the "error covariance" of $x$, not for the covariance of $x$ itself. But both $x(s)$ and $\text{cov}[x(s)]$ are small.

A few more notations are as follows: the scale of node $s$ is $m(s)$, with scale 0 for the root and larger numbers for finer scales; the state at all children of node $s$, that is, $x(s_1), x(s_2), \ldots, x(s_q)$, is written collectively as $x^m_{m(s)}$; the state at the $i$th child of node $s$, that is, $x(s_i)$, is also written as $z_i(s)$; and the complement to $z_i(s)$, that is, the state at all nodes at scale $m(s) + 1$ except $z_i(s)$, is written collectively as $z_{ic}(s)$. All of them are illustrated in Fig. 1a.

Until now, how the coarse-scale states are "aggregated" has been unspecified. Suppose the aggregations are already done in some way and the tree is fully populated, and we have a tree-indexed process. The MAR model represents the state dynamics across scales on the tree using two state transition equations—the downward recursion:

$$x(s) = A(s)x(s) + w(s)$$

and the upward recursion:

$$x(s_\gamma) = F(s)x(s) + w'(s).$$

Here, $A(s)$ and $F(s)$ are downward and upward transition matrices, and $w(s)$ and $w'(s)$ are zero-mean random noises with covariances $Q(s)$ and $Q'(s)$. The MAR model requires a Markov property on the tree, which means that for any node $s$ above the finest node, the $q$ subtrees below $s$ and the rest of the tree (i.e., other than the subtree rooted at $s$) are conditionally uncorrelated given $x(s)$ (Frakt and Willsky 2001). In other words, node $s$ conditionally decorrelates the $q + 1$ sections into which it partitions the tree. This Markov property on the tree is a generalization of the Markov property for a regular state space model in time, and it allows us to filter information across scales just as a regular filter does in time.

Then how do we aggregate from finer to coarser scales such that the resulted tree-indexed process will satisfy the above Markov property? Frakt and Willsky (2001) found that the tree will satisfy the Markov property if it meets two conditions: (i) a "locally internal" condition and (ii) "scale-recursive Markov property." Locally

![Multiscale tree](adapted from Zhou et al. 2008)
internal means that the parent state $\chi(s)$ is a linear function of its children’s state $\chi^{m(s)+1}$:

$$\gamma(s) = V(s)\chi^{m(s)+1},$$  \hspace{1em} (3)$$

where $V(s)$ is a coefficient matrix. The problem can be further simplified by ignoring its off-diagonal blocks, so $V(s) = \text{diag}[V_1(s), V_2(s), \ldots, V_q(s)]$, where $V_i(s)$ is the matrix coefficient multiplied on $\chi(s \alpha_i)$.

Scale-recursive Markov property means that for any node $s$ above the finest, that is, the $q + 1$ states, its $q$ children states $\chi(s \alpha_1), \chi(s \alpha_2), \ldots, \chi(s \alpha_q)$ plus the state at its nonchildren combined state (see Fig. 1a) are conditionally uncorrelated given $\chi(s)$.

The above two conditions bring the global Markov requirement down to a local Markov requirement between a parent and its children, and they form the basis for the algorithm that identifies the multiscale tree from the finest-scale state vector. In other words, we create the parent by linearly combining its children and look for $V(s)$ such that the parent decorrelates each one of the children and the nonchildren. With the simplification

---

**FIG. 2.** Steps 1–3: Top-down (coarse to fine) creation of a 4-scale 4-adic tree topology for a 32 × 32 computing grid through recursive NG-driven dividing. (Bottom right) Multiscale tree creation when measurement pixels (black) are partially missing and 4 times coarser than that of the computing grid (gray).

**FIG. 3.** Multiscale tree topology (q = 4) for the Arkansas–Red River basin.
of $V(s)$ into small $V_f(s)$ blocks, the MPE method introduced in Frakt and Willsky (2001) estimates $V_f(s)$ one by one for each child. In MPE, $V_f(s)$ is the matrix coefficient that best decorrelates states at child $s_0$, and the rest of nodes at scale $m(s) + 1$, that is, to decorrelate $z(s)$ and $z_u(s)$ in Fig. 1a. The MPE method (Arun and Kung 1990; Rao 1964) does not perform a direct decorrelation through canonical correlations but instead (to reduce computational complexities) looks for $V_f(s)$, such that the conditional estimate of $z_u(s)$ given $V_f(s) z(s)$ has a minimum error, that is, a maximum predictive efficiency estimator for $z_u(s)$. Since the dimension of $x(s)$ is predetermined to be small, the $V_f(s)$ is constrained to provide only the best possible decorrelation with the possibility of an incomplete decorrelation. We can further reduce the computations by stripping down $z_u(s)$. For those nodes in $z_u(s)$ that are geographically far away from $z(s)$, most are naturally decorrelated with $z(s)$. So in MPE, we only need to include nodes in $z_u(s)$ within a reasonable “neighborhood” of that child, for example, within a certain range $r$. With $V_f(s)$ and the computed coarse-scale states, it is then easy to estimate the state recursion parameters $A(s)$, $F(s)$, $Q(s)$, and $Q'(s)$ (Zhou et al. 2008).

In effect, this conditional decorrelation decomposes the information in the child $s_0$ into two parts: the part that is unique to the child itself (individuality) and the part shared by all others at scale $m(s) + 1$ (commonality). As much as possible (if not all), commonality is passed onto the parent [constrained by the dimension of $x(s)$] and the individuality is kept with the child itself. In the end, the parent state $x(s)$ is a collection of information that is owned by its children and common to all at scale $m(s) + 1$. In other words, this fine-to-coarse tree identification recursively strips local information out of global information at different scales. Any two finest-scale nodes are no longer directly related on the tree as they would be in the large $\text{cov}(x)$; instead, they are connected through their common ancestor on the tree, which records the commonality shared by the two. The multiscale strategy approximates a full-rank $\text{cov}(x)$ by imposing a multiscale structure as “prior” knowledge on the signals. This may also help reduce sampling errors due to small sample size (Zhou 2006). For example, spurious teleconnection caused by insufficient sampling will be suppressed because two pixels far away from each other will be also far away on the tree.

### b. EnMSF—a two-sweep filtering procedure

Given a measurement at the finest scale, a two-sweep filtering procedure can be defined to propagate it across the tree (Willsky 2002). The two-sweep propagation is illustrated in Fig. 1b. This procedure is analogous to a Rauch–Tung–Striebel smoother (Anderson and Moore 1979) but performed on a tree. The upward sweep (filtering step) propagates measurement information from fine to coarse scales. At the end of the upward sweep, all the coarse nodes are updated with the finest pixels they cover. Then the downward sweep (smoothing step) spreads coarse-scale updates down to fine scales, and every finest pixel is updated with measurements at all pixels. EnMSF basically follows the classic tree filtering scheme summarized in Willsky (2002) and estimates covariances from node state ensembles and updates each ensemble member, as does an EnKF (Evensen 1994; Burgers et al. 1998).

Besides operating on ensembles, EnMSF also makes a significant change to the classic tree filtering procedure. In the upward sweep, the classic procedure produces an update of the parent node state given to every one of the $q$ children, and then it merges the $q$ updated estimates into one. EnMSF updates the parent using all its $q$ children at once, and it does this by augmenting the measurement vector such that the innovation at the parent includes all the filter increments (innovation times the Kalman gain) from all its children (Zhou et al. 2008). This new procedure is mathematically equivalent to the classic algorithm (Zhou 2006), but it facilitates the incorporation of coarse-scale measurements, which is a significant feature in hydrologic applications. A coarse-scale measurement can simply be the average of finest pixels it covers, or any linear function of the finest:

$$y(s) = h(s) x_M.$$  \(\text{(4)}\)

EnMSF incorporates the coarse measurement $y(s)$ by adding it to the augmented measurement vector at the first coarse-scale node that just covers all the finest pixels involved. That is to say, the measurement enters the tree at the finest possible scale to minimize the uncertainties incurred when we propagate the finest scale states to the coarse node in upward recursion.

### 3. Implementation of a multiscale hydrologic data assimilation system

The major issue in implementing EnMSF in hydrologic applications arises in creating the topology of the multiscale tree and mapping pixels from the computing grid onto nodes on the multiscale tree, which is a challenge because the computing grid for a land surface model may have an arbitrary shape and arbitrary number of pixels (e.g., a river basin or continent). Square-shaped computing grid with $4^k$ number of pixels is easiest for multiscale tree creation, and one just needs to evenly divide the pixels into $2 \times 2$ blocks recursively and map
them onto a 4-adic tree ($q = 4$ for all nodes). Even dividing is important for this multiscale method, because the children of one parent should share approximately the same area unless specific reasons exist to suggest one part of the area is more important than another. The tree topology represents the prior knowledge imposed on the data and specifying it properly is necessary. We developed an automated procedure for dividing a computing grid of irregular shapes to create a tree of “balanced” topology.

To achieve this, we use the neural gas (NG) algorithm (Martinetz et al. 1993), a simple but efficient way for clustering sample data points. The clustering problem tries to identify the locations in sample data space where the data points are concentrated, that is, “clusters,” and label each data point with the cluster it belongs to. The clustering technique is a useful tool for class identification plus data classification, data segmentation, dimensionality reductions, and so forth. The creation of the multiscale tree topology directly translates to a recursive clustering of pixels on the computing grid. For instance, the root node on the multiscale tree maps to the entire computing grid, thus mapping pixels to its $q$ children—$q$ children per child—which is exactly what NG does. Then this dividing process is carried out on the children and moves recursively down to the children’s children and down to the finest scale. Figure 2 gives an illustration of this recursive procedure to create a 4-scale 4-adic tree topology for a $32 \times 32$ computing grid. In the application presented in section 4, the multiscale tree topology is created for the Arkansas–Red River basin study area using this strategy, and the result is illustrated in Fig. 3. NG is an incremental learning procedure, that is, to adjust the cluster locations repeatedly until they fit the data well. This category of techniques is sometimes referred to as self-organizing maps (SOMs; Kangas et al. 1990). The NG algorithm follows the procedure given below to divide a parent’s pixels for its $q$ children:

1) Randomly sample $q$ pixels (i.e., their latitude–longitude) from the parent’s pixel set as the initial locations of $q$ children clusters;

2) Randomly sample one pixel from the parent’s pixel set, and adjust the location of all the $q$ clusters toward the random pixel according to their distance to it—that is, the nearest cluster will be moved toward the random pixel most and the farthest cluster will be moved toward the random pixel least;

3) Repeat step 2 for a large number of times, and the magnitude of the cluster location adjustment is gradually tuned down as the number of iterations increase; and

4) After cluster location adjustment stops, each pixel is labeled to its nearest cluster and assigned to the corresponding child.

Martinetz et al. (1993) provides a useful discussion on different formulations to adjust the cluster locations in step 2 and to tune down the adjustment in step 3. In this study, the adjustment is exponentially related to the distance rank (closest cluster ranks as 1 and farthest ranks as $q$) and exponentially tuned down with the number of iterations. As shown in Fig. 3, NG is able to achieve two goals: to divide evenly and to have the pixels of the same child stay clustered. Also, NG is a supervised algorithm. That is, the number of children (clusters) is prescribed, but this not an issue here because we predetermine the scaling factor $q$. For this tree topology creation problem, NG runs very fast and stays very robust, and the tree creation computational effort is negligible compared to the subsequent state filtering computations. NG clusters data in any dimension, thus this tree creation can be performed over 1D or 3D computing grids.

Another issue is to create the multiscale tree topology, such that coarser measurements can be conveniently assimilated and missing measurement pixels conveniently handled. The goal is to make sure the same coarse measurement pixel does not fall onto different finest-scale nodes on the multiscale tree, such that the coarse measurement can be included in the finest possible scale on the tree, as suggested in section 2. First of all, the measurement resolution should be an exact multiple of that of the computing grid, and the two grids should be perfectly aligned. If not, we preprocess (resample) the measurement data to satisfy the above. Then, as shown in the dashed box in the bottom right of Fig. 2, the fine pixels (the gray fine grid) on the computing grid are aggregated into groups (the gray thick grid), such that the pixel groups have the same resolution as the measurement and perfectly aligned with coarse measurement pixels (the black coarse grid). Finally, the multiscale tree topology is created from these groups instead of the finer pixels.

4. Synthetic data assimilation experiments and results

The study area chosen is the Arkansas–Red River basin in central United States, which covers about 645 000 km$^2$ (latitude–longitude range can be found in Fig. 3). The climate is relatively wet in the eastern part of the basin, and it gradually becomes drier toward the west. The vegetation covers also change, from forest in the east to grass/crop land in the west. The Variable
Infiltration Capacity (VIC) hydrologic model (Liang et al. 1994; Liang et al. 1996) is employed to simulate the land surface processes. The VIC model is driven by meteorological forcing inputs and solves for various land surface states/fluxes, including soil moisture and temperature at different vertical depths, surface and subsurface runoff, evaporation and transpiration, snow depth and its properties. It has been tested and validated over different land cover types and climate regimes, and implemented in a number of large-scale hydrologic and climate studies (Mitchell et al. 2004; Maurer et al. 2002). In all the assimilation experiments to follow, VIC is set up to run over a computing grid of 1062 pixels (see Fig. 3), with a pixel size of 0.25° × 0.25°.

The surface soil moisture assimilation experiments are driven by synthetic satellite measurements. To evaluate the relative performance (accuracy, speed, etc.) of the multiscale filter, two additional filter algorithms are implemented and applied alongside of EnMSF:

1) The classic EnKF (Evensen 1994; Burgers et al. 1998) with large state vector \( x \) (1062-dimensional) and large \( \text{cov}(x) \) (1062 × 1062). This is referred as the “full rank” EnKF, or simply EnKF. In the synthetic assimilation experiment, this provides a “benchmark” solution to the filtering problem, because the \( \text{cov}(x) \) is fully preserved, whereas the EnMSF simplifies and approximates \( \text{cov}(x) \) using the multiscale structure. The computational cost of the full-rank method is high. In fact, more efficient/accurate sampling/analysis strategies have been developed for EnKF, for example, the square-root analysis scheme (Anderson 2001; Whitaker and Hamill 2002), rank reduction (Pham 2001), and a very efficient square-root scheme that uses a low-rank representation of measurement error covariance (Evensen 2004). But in our experiments, the measurement error is uncorrelated in space, that is, no rank loss in sampling, so the classic EnKF should converge to the same solution as the above (though more slowly) and will suffice as the benchmark.

2) The same classic EnKF, but treat each element in \( x \) (i.e., the pixels on the computing grid) as completely uncoupled, so each pixel is filtered independently (1062 1D filters). This is equivalent to ignoring all the off-diagonal terms in \( \text{cov}(x) \), so any statistical error correlations across pixels are ignored. We refer to this experiment as the “uncoupled” filter. Many data assimilation studies in hydrology adopted this strategy (Crow et al. 2001; Reichle et al. 2008; Pan et al. 2008) when there is evidence suggesting the dynamics is horizontally uncoupled. This strategy is fast, but the effect of horizontal decoupling needs to be investigated by comparing its results to the benchmark full-rank EnKF and EnMSF.

The synthetic experiment intends primarily to verify the method and implementation strategy. To simplify the generation of the meteorological forcing ensembles and the measurements that will be assimilated, which can enormously complicate the interpretation of the results, we use ensemble predictions from a seasonal climate model. The seasonal forecast products come from the National Oceanic and Atmospheric Administration’s (NOAA) Climate Forecasting System (CFS) model (Saha et al. 2006), which was developed and is deployed at the National Centers for Environmental Predictions (NCEP). CFS forecasts have been used for the seasonal hydrologic forecasts by Luo and Wood (2008). We take the processed forcing ensembles from this seasonal hydrologic forecast study, reaggregate to 0.25°, and use them as ensemble inputs for the assimilation experiment. Note that not all the spatial correlation in the rainfall ensembles is from climate model physics and only the part at or above 1.875° is, because the downscaling procedure adds uncertainties at finer scales according to local rainfall climatology. Among the 20 downscaled ensemble realizations available, we pick at random one as the “synthetic truth” and add perturbations to create the “synthetic measurement.” This approach is based on the assumption that the ensemble fully represents all the uncertainties in the climate system (Koster et al. 2000). Thus, the ensemble size \( N = 20 - 1 = 19 \).

In the experiment, all the ensemble simulations are initialized with land surface states from using observations for 31 January 2008. Then the VIC hydrologic model is run at a daily time step with the ensemble inputs and without assimilation for two months, from 1 February to 31 March 2008. This two-month “free run” establishes a stable spread across ensemble members such that the effect of initial conditions can be minimized. Then the assimilation is performed daily from 1 April to 31 July in 2008. A zero-mean and spatially uncorrelated Gaussian noise with a standard deviation (STD) of 5% (volumetric, all soil moisture values will be volumetric herein) is added to the synthetic truth top-layer soil moisture to create the synthetic measurements. Some remotely sensed soil moisture products (passive microwave sensors) are reported to have
2%–3% root-mean-square-errors (RMSEs; Gao, et al. 2006; Bindlish et al. 2003); however, considerable difference is also reported between the dynamics of remotely sensed soil moisture and model simulations. Thus, a more conservative STD value is used.

Figure 4 plots the top-layer soil moisture in all prior ensemble members and the synthetic truth on the first day of assimilation (1 April 2008). The prior ensemble spread in Fig. 4 reflects the hydrological consequence of the uncertainties in the CFS ensemble forecast—that is, the uncertainties in precipitation forecast in the period before the first day of assimilation. Figure 5 shows the assimilation results by all the filters implemented on the first day of assimilation. Since no assimilation is performed prior to this date, the prior ensembles in all the filters are identical; the prior ensemble mean and standard deviation are shown in the first column in Fig. 5. The posterior ensemble means and standard deviations in the second, third, and fourth columns in Fig. 5 show the difference between filters. In general, EnMSF matches the benchmark EnKF very closely in terms of both the mean and standard deviation, while the uncoupled filter fails to do so. For example, the area in the southwest portion of the basin is wet in the measurement (second column in the bottom row of Fig. 5). This wet area is missed in the prior ensemble mean as a result of smaller precipitation in the ensemble members other than the truth (see the third and fourth columns in the bottom row of Fig. 5). This missing wet spot is largely recovered in the posterior ensemble means by both the benchmark EnKF and EnMSF (third and second columns in the first row of Fig. 5) but not by the uncoupled filter (fourth column). The second row of Fig. 5 shows how an individual ensemble member is updated by the different filters. The posterior ensemble spread (standard deviation) updates differently across the filter experiments (third row in Fig. 5): EnMSF approximates the benchmark EnKF closely, while the uncoupled filter significantly overestimates the spread. The basin averaged RMSE in top-layer soil moisture are also calculated with respect to the synthetic truth for the posterior ensemble means from the different filters: 2.25% in EnMSF, 1.97% in full-rank EnKF, and 5.41% in the uncoupled EnKF. The RMSE for the prior ensemble mean is 7.44%.

One important research goal is to understand the effect of horizontal coupling to the filters. The VIC is a distributed model (soil and vegetation parameters differ from pixel to pixel), and the horizontal coupling in soil moisture predictions is dominated by the coupling in the forcing fields. This is perhaps true for all large-scale
hydrologic models, because the land surface is driven mostly by boundary conditions (inputs) instead of internal dynamics. The strength of the horizontal coupling in the forcing data directly affects the comparison between the full-rank and uncoupled filters. We believe ensemble climate model forecasts offer a reasonable representation of horizontal coupling in the meteorological forcing fields, given the fact that all known physics about the climate system is built into the climate model. Figure 6a gives a grayscale illustration of the state error correlation matrix \( \text{cor}(x) \) [i.e., the covariance \( \text{cov}(x) \)] with entries \( \text{cov}(x_i, x_j) \) normalized by square root of \( \text{cov}(x_i) \text{cov}(x_j) \), computed from the prior ensemble on the first day of assimilation. The off-diagonal terms reflect the strength of this coupling and such coupling is fairly significant in Fig. 6a. This explains why the uncoupled filter fails to match the full-rank solution in Fig. 5, because it ignores these off-diagonal terms. Horizontal coupling allows the filter to update one pixel based on a measurement from other pixels. The uncoupled filter does not use this information. Consequently, the ensemble standard deviation in the uncoupled filter is larger compared to the benchmark EnKF. This also explains the larger ensemble spread by the uncoupled filter. Figure 6b plots the correlation between two pixels versus their actual distance on the computing grid and provides a measure of how the correlation decays with distance. Large spread exists in this correlation-versus-distance relationship, but there is a systematic decay from a correlation of 1 at distance 0 km down to 0 at distance of about 1000 km. This decay suggests that the correlation structure in \( \text{cov}(x) \) relates to spatial distance, and the multiscale...
structure is a reasonable approach for the soil moisture fields. Note that the amount of horizontal coupling reflected in Figs. 6a and 6b only represents the particular day and event in this experiment, and such coupling may extend out to larger or smaller distances for other days.

Setting a particular CFS ensemble member as the synthetic truth may pose bias in evaluating the filter performance. Therefore, the above assimilation experiment is repeated 20 times, with the synthetic truths set to different members (as in Fig. 4). Figure 7 plots the
basin averaged prior and posterior RMSE in top-layer soil moisture in these experiments. The RMSE comparisons are fairly consistent in all experiments; that is, EnMSF outperforms uncoupled EnKF and closely approximates the full-rank EnKF.

Figure 8 gives the time series of basin-averaged ensemble means and standard deviations for all the filters. The time series confirms that what happens on the first day (of assimilation) continues throughout the assimilation period. For example, the basin-averaged ensemble spread is consistently larger in the uncoupled filter and close to the benchmark filter spread in the EnMSF experiment. The basin-averaged mean soil moisture in the benchmark EnKF and EnMSF has a slightly sharper dynamics as a result of the stronger influence from the measurements, but the difference is not dramatic because the averaging procedure tends to mask some differences in spatial distribution of soil moisture, like that seen in Fig. 5.

It is also observed that EnMSF runs much faster than the benchmark EnKF. Under the specific settings in this study (state vector size, ensemble size, multiscale tree topology, and horizontal coupling strength), the EnMSF runs about 20–30 times faster, including the time spent on tree creation and identification. However, no comparisons are made to the more efficient low-rank EnKF methods, such as the one introduced in Evensen (2004). In terms of methodology, multiscale methods simplify the high-dimensional problems by imposing a multiscale structure made from low-dimensional nodes, while low-rank EnKF methods try to manipulate nontrivial eigen components in the covariance only. The highest possible rank of ensemble error covariance is $N^2$, and in many hydrologic applications $N \ll d$. Both approaches share the same goal of dimensionality reduction; one approach does it through the identification of the multiscale structure and the other approach finds the effective rank through singular value decomposition (SVD). Their methodological difference suggests that EnMSF
may be more resilient to large ensemble size $N$ (or $d$) because it is insensitive to it, and low-rank EnKF methods may be more efficient with smaller ensemble sizes because SVD produces a more concise low-dimensional model.

5. Conclusions and future work

Two major components, the VIC land surface model and the EnMSF, form the basis for the hydrologic data assimilation system in this study. Both the land surface model and the multiscale filter are developed for large-scale land surface problems and so is the assimilation system. With the developed methodology, we implemented and verified the VIC–EnMSF assimilation system by applying it to a synthetic assimilation experiment over the 645,000 km$^2$ Arkansas–Red River basin. Because EnMSF is based the MAR model, which relies on a reasonable multiscale tree topology, the automated creation of such a tree topology for an arbitrary land surface modeling grid was a major challenge. A NG-based tree creation procedure was designed that effectively created well-balanced multiscale tree topologies. Mechanisms to handle coarse and missing measurements are also programmed into the tree topology creation (bottom right, Fig. 2).

The multiscale assimilation system is designed to fully account for the horizontal coupling in surface hydrology, and the subsequent horizontal error correlations in measurements and state variables. To test whether this was achieved, three ensemble-based assimilation experiments are carried out where the precipitation ensembles that force the land surface model are based on downscaled seasonal climate model predictions. This should assure that the precipitation fields are spatially and temporally consistent with weather physics and should capture the horizontal coupling. This coupling turns out to be relatively strong in the seasonal climate model–driven hydrologic simulations, and the strength of coupling makes a large difference between a fully coupled EnKF and an uncoupled EnKF. EnMSF closely approximates the result of the fully coupled EnKF, which is a significant improvement over uncoupled methods, and at a much lower computational cost than the fully coupled filter. As for computational efficiency with other reduced-rank approaches, such as a square root filter, no strong conclusions can be made because comparisons have not been made with the more concise ensemble filtering methods. Nonetheless, the study presented here demonstrates that the VIC–EnMSF assimilation system presents an efficient and scalable solution to large-scale hydrologic estimation problems. This achieved the first major goal of the research, namely, to establish that a multiscale data assimilation approach can be effectively used for large-scale hydrologic applications related to the assimilation of soil moisture, say, from satellite retrievals. The assimilation system is now being implemented and tested with forcing precipitation fields and soil moisture retrievals from the Tropical Rainfall Measurement Mission (TRMM) and NASA’s Advanced Microwave Scanning Radiometer for EOS (AMSR-E).

Acknowledgments. This research is supported by the National Aeronautic and Space Administration’s (NASA) research projects Estimating Continental-Scale Water Balances through Modeling and Assimilation of EOS Terra and Aqua Data, and Assessment of the Predictive Skill of GPM-ERA Precipitation Estimates for Hydrologic Applications.

REFERENCES


