Multi-sensor large scale land surface data assimilation using ensemble approaches

by

Yuhua Zhou

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006

Author ....
Department of Civil and Environmental Engineering
February 1, 2006

Certified by ....
Dennis McLaughlin
Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by ....
Dara Entekhabi
Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by ....
Andrew J. Whittle
Chairman, Department Committee for Graduate Students

Multi-sensor large scale land surface data assimilation using ensemble approaches

by

Yuhua Zhou

Submitted to the Department of Civil and Environmental Engineering
on February 1, 2006, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Field of Hydrology

Abstract

One of the ensemble Kalman filter's (EnKF) attractive features in land surface applications is its ability to provide distributional information. The EnKF relies on normality approximations that improve its efficiency but can also compromise the accuracy of its distributional estimates. The effects of these approximations are evaluated by comparing the conditional marginal distributions and moments estimated by the EnKF to those obtained from an SIR particle filter, which gives exact solutions for large ensemble sizes. The results show that overall the EnKF appears to provide a good approximation for nonlinear, non-normal land surface problems.

A difficulty in land data assimilation problems results from the high dimensionality of states created by spatial discretization over large computational grids. The high dimensionality can be reduced by exploiting the fact that soil moisture field may have significant spatial correlation structure especially after extensive rainfall while it may have local structure determined by soil and vegetation variability after prolonged drydown. This is confirmed by SVD of the replicate matrix produced in an ensemble forecasting experiment. Local EnKF's are suitable for problems during dry periods but give less accurate results after rainfall. The most promising option is to develop a generalized method that reflects structural changes in the ensemble.

A highly efficient ensemble multiscale filter (EnMSF) is then proposed to solve large scale nonlinear estimation problems with arbitrary uncertainties. At each prediction step realizations of the state variables are propagated. At update times, joint Gaussian distribution of states and measurements are assumed and the Predictive Efficiency method is used to identify a multiscale tree to approximate statistics of the propagated ensemble. Then a two-sweep update is performed to estimate the state variables using all the data. By controlling the tree parameters, the EnMSF can reduce sampling error while keep long range correlation in the ensemble. Applications of the EnMSF to Navier-Stokes equation and a nonlinear diffusion problem are demonstrated. Finally, the EnMSF is successfully applied to soil moisture and surface fluxes estimation over the Great Plains using synthetic multiresolution L-band passive and active microwave soil moisture measurements following HYDROS specifications.

Thesis Supervisor: Dennis McLaughlin
Title: Professor of Civil and Environmental Engineering

Thesis Supervisor: Dara Entekhabi
Title: Professor of Civil and Environmental Engineering

Acknowledgments

I would like to express my gratitude to all the people who made it possible for me to finish this thesis work. First of all I'm deeply indebted to my advisors Professor Dennis McLaughlin and Professor Dara Entekhabi. Their guidance lead me to the fascinating field of data assimilation fulfiled with interesting opportunities. They helped me transform from a master student without any idea about data assimilation to a PhD obsessed with estimation theory. They aslo give me a magic wand of research methods and problem solving ability, which would be invaluable for me forever. I'm also personally obliged to Dennis and Dara for their warm-hearted help with my family. I can't imagine how I could have survived the first two years without their help.

I would like to take the opportunity to thank all the help and stimulating support from research group members. In particular, I would like to thank Virat Chatdarong for giving me the great rainfall replicates that are essential to demonstrate the Great Plains example and his tolerance with my constantly bugging for replicates. In addition, I want to thank him since his introduction of multiscale tree approaches inspired me about EnMSF. I want to thank Susan Dunne for providing me lots of help with NCAR land surface model and the passive and active remote sensing models. I'm also gratureful to Sai Ravela for helping me with the ITR cluster and tutoring me Gerris solver and Navier-Stokes equations. I would also thank Gene-Hua Crystal Ng for helping me with the Richard's model, and Sara Friedman for the fluid mechanics discussion. I would also extend thanks to David Flagg for his fantastic job of preparing all the raw rainfall data sources. Without his help probably I would delay my thesis work a long time. Some other people I would like to thank are Guangda Li, who tutored me about some basics fluid mechanics which helps me with the Navier Stokes example; Steven Margulis, who gave me a jump start on implementing EnKF; Rolf Reichle, who is always there when I need any help.

Finally, I would give special thank to my wife Yeren Sun, whose love, support, and encouragement enabled me to accomplish this work.

Contents

List of Figures

List of Tables

Chapter 1

Background and Introduction

Over the past decades, tremendous effort has been focused on studying the land surface system at different temporal and spatial scales. The understanding about the land system has been incorporated into many land surface models (LSM), which usually include water, energy, and carbon balances. Theoretically, LSM's are able to provide continuous description of the physical processes in time and space. Since these land surface models make many assumptions about real physical processes, they are always imperfect and subject to various errors.

On the other hand, with the advancement of observation ability, especially the remote sensing technologies, more and more land surface datasets are becoming available. Remotely sensed precipitation, radiation, soil moisture, vegetation, and ground water, etc. may provide valuable information to better understand the land surface system. Some of the measurements are directly used as input to LSM. Unlike the land surface models, measurements are often discontinuous in time and space. For example, satellite tracks are discontinuous in space and time but with large footprint coverage; ground station measurements are continuous in time but discontinuous in space. Similar to LSM's, the measurements are often subject to errors which result from observation instrument precision inadequacy and retrieval model assumptions.

Taking all the errors into account, the land surface models and observations can be described in an uncertain, or probabilistic manner, which offers great convenience to better estimate or predict the variables of interest. Data assimilation provides an ideal framework to consistently and optimally/suboptimally meld all the available continuous or discontinuous information from the dynamic model prediction and observation data to describe the reality.

1.1 Brief Introduction

Generally speaking, data assimilation techniques are designed to characterize the uncertain state of an environmental system, using all relevant measurements. Data assimilation has a long history in meteorology for generating initial conditions for numerical weather prediction. Currently, operational numerical weather prediction centers all produce initial conditions through a statistical combination of observations and short-range forecasts, i.e. data assimilation. Oceanographers use data assimilation to merge large volumes of data, such as TOPEX/POSEIDON altimeter data.

The data assimilation techniques in meteorology and oceanography has evolved from successive correction method, nudging, to the 4-dimensional variational methods, and ensemble methods. There are several excellent reviews in the literature about data assimilation. Daley [24] gives a comprehensive description of methods for atmospheric data analysis and assimilation. Ghil and Malanotte-Rizzoli [44] gives a rigorous discussion of present data assimilation methods with special emphasis on sequential methods. Talagrand [93] reviews current methods of data assimilation.

Data assimilation in hydrology is still young compared to the the other two. Despite the the short history of data assimilation in hydrology, it has found a wide range of applications in estimation of soil moisture, rainfall, groundwater, energy fluxes, runoff, snow and ice. McLaughlin [71] gives an excellent review of hydrologic data assimilation. Some of the earlier work focusing on 1-dimensional estimation using synthetic data includes Entekhabi [31], Callies [11], Castelli [14], Houser [55]. Some lately work includes Dunne [29], Reichle [82], Crow [22], Rodell [86], Mitchell [74], Parada [79], Caparrini [13], Kumar [64], Boulet [8], Li [65], Boni [6].

The techniques used in hydrology community are much the same as those used in meteorology and oceanography. In much of the work mentioned above, variational methods and variants of Kalman filters are the two types of commonly used assimilation methods. Reichle [83] solves a soil moisture estimation problem using 4-dimensional variational methods for a land surface model with additive errors. To account for nonadditive uncertainties commonly existent in the land surface system, ensemble methods are beginning to gain popularity. Margulis [69] uses the ensemble Kalman filter to assimilate airborne L-band microwave observations and ground-based measurements of micrometeorological variables, soil texture, and vegetation type into NOAH (NCEP, Oregon State University, Air Force, Hydrologic Research Lab) land surface model. Assimilation is done in a pixel by pixel fashion and no horizontal correlation is considered. Dunne [29] presents an adaptive hybrid filter/smoother in which brightness temperature is used to break the study interval into a series of dry-down events. The smoother is used on dry-down events, and the filter is used when precipitation is evident between estimation times.

1.2 Land Surface System

Before reviewing any detailed data assimilation techniques, it is useful to look at a simple land surface system. For any land surface data assimilation application, the system of interest typically involves water, energy, and carbon balance and fluxes at different spatial and temporal scales. Existing micro or macro scale land surface models characterize these processes using some complex nonlinear dynamic equations.

Usually the equations in these models are composed of three major parts: the soil water and heat dynamics, the vegetation processes, and the atmospheric boundary layer dynamics. These dynamics are highly coupled, which often entails iterative methods in solving these equations. The unsaturated zone soil water and heat transport equations are the key in the coupled system, since it controls the incoming water and energy partition. The coupled soil water and heat dynamics are described by nonlinear advection diffusion equation. For the soil moisture part, the models are in the variant form of the simple 1-D Richards equation

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[D(\theta)\frac{\partial \theta}{\partial z} + K(\theta)\right] - S(\theta) \tag{1.1}$$

where z is the vertical axis, θ is volumetric soil moisture, $D(\theta)$ is diffusivity, $K(\theta)$ is hydraulic conductivity, $S(\theta)$ is sink term representing vegetation root uptake. The simplest heat equation can be described by the following diffusion equation:

$$C(\theta)\frac{\partial T}{\partial t} = \frac{\partial}{\partial z}\left[\lambda(\theta)\frac{\partial T}{\partial z}\right] \tag{1.2}$$

where $C(\theta)$ is the soil volume heat capacity, $\lambda(\theta)$ is soil heat conductivity. These two equations are mainly coupled through the pathway of evapotranspiration, which exerts the top boundary conditions of both equations. Since evapotranspiration is also controlled by roughness of vegetation and canopy characteristics, all the water, energy, and carbon fluxes are then coupled.

The variants of these 1-D water and heat equation are widely used in land surface modelling. One distinctive feature of the land surface system is that the lateral interactions of water or heat is usually not considered. The same 1-D equations above are used for different computational pixels but with different set of parameters and inputs determined by land surface characteristics and meteorological forcing which have lateral connections between pixels.

These equations are simply a set of nonlinear diffusion equations which forms a nonlinear dissipative system. The downward water and energy diffusion processes depends on the gradients and land surface parameters. The inputs to the system are exerted at the top boundary, where rainfall controls soil moisture and short/long wave radiation controls soil temperature. After each rainfall events, the rainfall passing through vegetation infiltrates into ground and the amount exceeding soil infiltration capacity becomes runoff. The bottom boundary of the soil water equation is controlled by the position of the saturated water level. At large scales, the variables in these equations such as soil moisture below surface, groundwater, evapotranspiration are hard to observe directly. However, they can be inferred from the relationships between the variables in the equations and other physical quantities.

Since these models are only approximations to the true physical processes, they are subject to all kinds of errors or uncertainties. In a typical land surface model, these uncertainties may come from model parameters such as soil properties, vegetation properties, or from model forcing input such as rainfall, and solar radiation, etc. For the land surface processes taking place in a 4D space, accurate quantification of these errors is not a trivial task, and error assumptions usually have to be made.

The estimate of the variables of interest as accurate as possible, or inference of the unobservable variables or model parameters of the land surface system, can be solved with estimation theory.

1.3 Estimation Theory

The data assimilation problem is essentially an estimation or inference problem in statistics. Based on different ways of using the available information, data assimilation can be categorized into filtering, smoothing, and prediction three problems. When the time at which an estimate is desired coincides with the last measurement point, the problem is referred to as filtering; when the time of interest falls within the span of available measurement data, the problem is termed smoothing; and when the time of interest occurs after the last available measurement, the problem is called prediction. McLaughlin [72] gives a complete set of these demonstration problems, where interpolation is illustrated with an example based on multiscale estimation of rainfall during the TOAGA-COARE field experiment; smoothing is illustrated with a variational soil moisture estimation algorithm applied to the SGP97 field experiment and filtering is illustrated with an ensemble Kalman filter, also applied to the SGP97 (Southern Great Plains 1997) experiment. In this thesis, filtering problem will be the main focus.

1.3.1 Bayesian Least Squares Update

The simplest estimation problem can be described in this way: given an indirect measurement vector y of a Gaussian random vector x with mean m_x and covariance matrix P_x, we want the estimate of x as close as possible to the truth. Suppose the measurement uncertainty v in y is a Gaussian noise with zero mean and covariance matrix R; y and x are related through

$$y = Hx + v \tag{1.3}$$

which is called the measurement equation, H is the measurement operator. Now we want $\hat{x}$, the best or optimal estimate of x, in the sense that $tr(E[(\hat{x} - x)^2])$, the trace of expected mean squared error matrix, is minimized for this $\hat{x}$. Here E represents the expectation with respect to the conditional distribution $p(x|y)$. The approach of using the expected mean squared error as the performance gauge is intuitive since it describes on average how far away the estimate is from the truth. Then the minimization problem becomes

$$\hat{x} = arg\ min\ E[(\hat{x} - x)^2] = arg\ min \int (\hat{x} - x)^2 p(x|y)dx \tag{1.4}$$

It can be derived [60] that the solution to $\hat{x}$ is given by

$$\hat{x} = m_{x|y} \tag{1.5}$$

where $m_{x|y}$ is the conditional mean of x given by the posterior distribution $p(x|y)$. According to Bayes' theorem

$$p(x|y) = cp(x)p(y|x) \tag{1.6}$$

where c is to ensure $p(x|y)$ integrates to one, $p(x)$ is the prior distribution, $p(y|x)$ is the likelihood function completely determined by the probability distribution of v and the value of x. It indicates the posterior distribution is only a function of the prior distribution and likelihood function. This estimate $\hat{x}$ is called the Bayesian Least Squares (BLS) estimate. It has the minimum mean squared error and is unbiased.

If a linear estimate of x from y satisfying the above minimum mean squared error criteria is desired, we write it in the following form

$$\hat{x} = By + C \tag{1.7}$$

where B and C are a constant matrix and vector respectively. We can prove the best estimate is given by

$$\hat{x} = m_x + K(y - Hm_x) \tag{1.8}$$

where

$$K = (P_x H^T)(HP_x H^T + R)^{-1} \tag{1.9}$$

The K matrix is termed as the gain matrix. It weights the prior mean of x and the new information in measurement y, i.e., $y - Hm_x$. This estimate is called the Linear Least Squares (LLS) estimate. It is unbiased and has the minimum error covariance among all the linear estimates.

For this Gaussian x and v case, the conditional mean of x is exactly the same as the LLS estimate. Also, the error covariance matrix $P_e = E((\hat{x} - x - E(\hat{x} - x))^2)$ for LLS estimate is the same as the posterior covariance matrix of $\hat{x}$ from the BLS estimate.

The BLS and LLS estimates described in this section form the foundations of the most data assimilation problems. In particular, they consider all the measurements and variables of interest at the same time. The optimal interpolation or Kriging approach used in data assimilation is simply LLS estimate by assuming the variable of interest is normally distributed. However, in a data assimilation problem it is usually impossible to process all the variables as a batch simply because the gigantic size of a spatial or temporal problem would prohibit such a direct calculation of the estimate. To make the computation feasible, strategies to deal with temporal or spatial problems often include sequential processing of the data y by exploiting the internal structure of the x. A universal framework for dealing such a problem is the graphical model approach [61].

1.3.2 Sequential Estimation Theory for a Dynamic System

As mentioned in the last section, the problem size of a data assimilation problem might be huge if collecting all the variables together. It is impossible to handle the big problem directly. Usually the spatial, temporal, or scale Markovian assumption of a process would dramatically simplify the problem. This section would address the state space approach of estimation, including introduction to sorts of Kalman filters.

In filtering problems, it is convenient to describe the discrete time state and measurement equations in the following state space form:

$$x_t = f(x_{t-1}, u_t) \tag{1.10}$$

$$y_t = h(x_t) + v_t \tag{1.11}$$

where x_t is a function of space or time or scale, called the system state vector, with an uncertain initial condition x_0, u_t is a vector of uncertain model inputs (not necessarily additive), y_t is the measurement vector, and v_t is a vector of additive random measurement errors. In a land surface problem x_t could be a vector of soil moisture values in different pixels and layers, u_t a vector of precipitation rates, and y_t a vector of microwave radiometer measurements indirectly related to soil moisture. The uncertain variables x_0 , u_t, and v_t are assumed to have known prior probability distributions and the measurement error vectors at different times are assumed to be independent. The functions $f(\cdot)$ and $h(\cdot)$ represent discretized models of the system dynamics and measurement process.

Using the state pace models to describe the system has the advantage of simplifying the complexity of the problem using Markovian assumption which assumes the current state vector summarizes all the information at previous times/locations/scales and the data before the current time/location/scale can be discarded after being used.

Also, the state space model is consistent with the dynamic system discretization used for solving partial differential system equations. In terms of describing the uncertainties, the state space model makes it easier to reformulate the full probability distribution of the state vectors with a set of conditional probability distributions.

For a temporal case, the filtering problem is to characterize the current state x_t from $y_{1:t}$, the set of all measurements obtained at discrete times in the interval [0, t]. The ideal probabilistic characterization is the conditional probability density $p(x_t|y_{1:t})$, the conditional probability density function (pdf) of variable x at time t, given all the measurements y between time interval [0 t]. Estimating $x_{t'}$ from $y_{1:t}$, where $t' < t$, is the smoothing problem. The ideal probabilistic characterization for this case is the probabilistic distribution of the full trajectory of the state variables conditioned on all the measurements. This is analogous to the optimal interpolation problem as if the time index is removed.

Another point worth making is that the multivariate density of the whole state vector is difficult to compute or interpret for large land surface problems, we typically focus on particular properties of $p(x_t|y_{1:t})$, such as its moments and univariate marginal densities of $p(x_t|y_{1:t})$.

1.4 The Kalman-Bucy Filter and the Extended Kalman Filter

In general, estimation problems can be formulated in a Bayesian framework. The process of estimating $p(x_t|y_{1:t})$ can be simplified to be described in the state space framework as in (1.10) and (1.11). However, the closed form solution for $p(x_t|y_{1:t})$ is nearly impossible to obtain except for some special situations such as for linear Gaussian system and measurement model, which will be discussed in this section.

For a linear system with Gaussian process noise and measurement noise, the state space model is

$$x_t = F_t x_{t-1} + u_t \tag{1.12}$$

$$y_t = H_t x_t + v_t \tag{1.13}$$

where x_t is the system state vector with an uncertain initial condition x_0 , u_t is a vector of additive uncertain process noise, y_t is the measurement vector, and v_t is a vector of additive Gaussian random measurement errors. The uncertain variables x_0 , u_t, and v_t are assumed to be independent at different times, i.e., they are white noise. F_t is the linear system operator and H_t represents the linear measurement operator. For this system, state variables and measurements are jointly Gaussian, which makes the first two orders of moment sufficient to describe the uncertainties.

To obtain the estimates $\hat{x}_{t|t}$ at different times for this linear Gaussian system, the traditional optimal Kalman filter sequentially update the mean of the state variables based on the covariance of the estimation errors. The uncertainty in error itself is propagated using Riccati equation describing the propagation of the error $\hat{x} - x$, the difference between the estimate and true value. The updating equations can be written as:

$$\hat{x}_{t|t} = x_{t|t-1} + K_t(y_t - F_t x_{t|t-1}) \tag{1.14}$$

where the gain matrix K_t is given by

$$K_t = P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1} \tag{1.15}$$

which is determined by the error covariance matrix $P_{t|t-1}$ and the covariance matrix R_t of the measurement at time t. The error covariance matrix $P_{t|t-1}$ of $\hat{x}_{t|t-1}$ is given by

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^T + Q_t \tag{1.16}$$

and

$$P_{t|t} = (I - K_t H_t) P_{t|t-1} \tag{1.17}$$

In equation (1.14), the last term $y_t - F_t x_{t|t-1}$ is called innovation, which represents the new information given by the current measurement y_t. The innovation time series is a white noise process, which indicates the previous measurements can be discarded after being used and thus allows the sequential update form in (1.14).

If $H_t P_{t|t-1} H_t^T + R_t$ in (1.15) is rank deficient, pseudoinversion can be used instead

$$K_t = P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^+ \tag{1.18}$$

This is equivalent to first projecting the observations onto the dominant eigenspace of $H_t P_{t|t-1} H_t^T + R_t$, denoted as L; and then use the projected innovation $L(y_t - F_t x_{t-1|t-1})$ for the estimation.

For a quasi-linear system, the state space equation is nonlinear and is written as

$$x_t = F_t(x_{t-1}) + u_t \tag{1.19}$$

$$y_t = H_t x_t + v_t \tag{1.20}$$

In this case, although the mean of the estimate can still be propagated using the nonlinear model

$$x_{t|t-1} = F_t(x_{t-1|t-1}) \tag{1.21}$$

the error covariance matrix must be propagated by the linearized system equation with linear operator $F_t^{'}$, called the tangent linear, using

$$P_{t|t-1} = F_t^{'} P_{t-1|t-1} F_t^{'T} + Q_t \tag{1.22}$$

All the other equations are the same as in the Kalman-Bucy filter described above.

This filter is usually called the Extended Kalman Filter.

It's worth noting that the covariance matrix of the error $\hat{x}_t|y_{1:t} - x_t$ is the same as the covariance of $\hat{x}_t|y_{1:t}$, i.e. the posterior covariance of x_t. For linear Gaussian system, Bayesian least squares estimate and linear least squares estimate are identical.

The extended Kalman filter has been widely applied in meteorology, oceanography, hydrology. A big issue for these applications is the computational demand caused by the high dimensionality of these problems. Firstly the extended Kalman filter requires the propagation of the error covariance matrix at every time step or few steps if necessary, the size of which might be on the order of $10^6 \times 10^6$. The memory required to store this matrix would become a serious problem. Secondly the inversion of the $HPH^T + R$ matrix for gain calculation is also a daunting task. Another big issue is the nonlinearity in these models. For highly nonlinear models, the linearization in the extended Kalman filter would make the error covariance calculation numerically unstable which can cause the filter divergence. The other issue is the error assumptions in the process noise u_t. In hydrological applications, the additive Gaussian assumption is insufficient to describe the reality.

To deal with the high dimensionality issues, many suboptimal Kalman filters have been proposed. One type of such filters is the reduced rank Kalman filters [95, 20, 30, 99]. These methods efficiently represent the error covariance and its propagator by truncating their singular vectors and eigenvectors expansions to a few leading terms, which then results in a low-dimensional system easier to solve. Some other reduced rank filters use coarse-grid approximation [38], EOFs (empirical orthogonal functions) [12], or wavelets [17, 94, 3] of the error covariance matrix. All of these methods can only deal with uncertain initial condition problems or systems with additive errors. Another approximation type is the parameterized flow-dependent error covariance [85] method that assumes the background error correlation has essentially the same shape as the background field. This ad hoc approach has the potential to estimate anisotropic correlations. The other type is the traditional Monte Carlo or ensemble Kalman Filter (EnKF) methods [33, 56] These traditional EnkF's avoid the expense of directly propagating error covariances by estimating them from an ensemble of forecast, thus it doesn't rely on the tangent linear model to evolve error statistics. The EnkF is a promising research direction especially for the nonlinear filtering problems which will be discussed in detail in the following section.

1.5 Ensemble Nonlinear Filtering

For linear or quasi linear systems with a limited number of state variables and additive system error, the Kalman-Bucy or the Extended Kalman filters can be easily applied. However, estimation problems of a nonlinear high dimensional dynamic system with nonnormal nonadditive uncertainties is challenging, which is very common in the land surface system. Traditional Kalman filters can't be directly used for these cases.

During the last decade nonlinear filtering theory has undergone a rapid development partly due to the dramatically reduced computational cost. Many popular nonlinear filtering algorithms such as particle fillers [2, 45], unscented Kalman filter [62], and ensemble Kalman filters [33, 10] require no linearization and additive Gaussian uncertainties assumption and have been proved promising. The basic idea behind these ensemble filters is to use Monte Carlo approach to approximate the propagation of $p(x_t|y_{1:t-1})$, the posterior probability density function (pdf) of the state variables and then use Bayes' theorem or an update equation to ingest y_t, the new information from observations into the propagated ensemble to form $p(x_t|y_{1:t})$. Ensemble-based data assimilation methods are becoming popular in many of the earth sciences, largely because they are easy to use, flexible, and make relatively few restrictive assumptions (see the review by [34]). They also have the advantage of providing distributional information about uncertain variables, including approximate marginal distributions, quantiles, and higher-order moments. This information is particularly useful in land surface applications, where variables such as soil moisture can be highly skewed towards the wet or dry ends and can even be bi-modal, depending on the time and space scale considered [88]. In such cases, means and covariances alone may not adequately characterize variability.

From the Bayesian estimation point of view, ensemble methods are able to provide a practical alternative to exact Bayesian solutions because they rely on discrete approximations of the densities $p(x_t|y_{1:t-1})$ and $p(x_t|y_{1:t})$. The approximations can be expressed as:

$$p(x_t|y_{1:t-1}) \approx \sum_{i=1}^{N} w^i_{t|t-1}\delta(x_t - x^i_{t|t-1}) \tag{1.23}$$

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w^i_{t|t}\delta(x_t - x^i_{t|t}) \tag{1.24}$$

These approximations replace each continuous density by a sum of N Dirac delta densities located at the randomly generated state vectors, or replicates, $x^i_{t|t-1}$or $x^i_{t|t}$, for $i = 1,\ldots,N$. The Dirac delta terms (and the corresponding replicates) for each approximation are assigned discrete probabilities (or weights) $w^i_{t|t-1}$or $w^i_{t|t}$, respectively. If the weights in each expansion sum to unity, the integrals of (1.23) and (1.24) yield stepwise approximations of the continuous cumulative distribution functions for $p(x_t|$ $y_{1:t-1})$ and $p(x_t|y_{1:t})$, respectively. The random replicates and corresponding weights can be generated in a variety of ways, e.g. as in the SIR particle filter or ensemble Kalman filters.

In filtering applications it is useful to distinguish two sequential estimation operations: 1) propagation of the state from one measurement time to the next (forecasting) and 2) updating of the propagated state with the new measurement (analysis). If the complete density $p(x_t|\ y_{1:t})$ is desired forecasting is carried out by deriving $p(x_t|\ y_{1:t-1})$ from $p(x_{t-1}|y_{1:t-1})$ (e.g using the Fokker-Planck equation) and analysis is carried out by deriving $p(x_t|\ y_{1:t})$ from $p(x_t|y_{1:t-1})$ (e.g using Bayes theorem) [60]. The required calculations are generally feasible only for very small problems.

The nonlinear ensemble filters share the same forecasting step. To examine the mechanics of this step, suppose that replicate i at t - 1 has the value $x^i_{t-1|t-1}$with weight $w^i_{t-1|t-1}$. The nonlinear state equation (1.10) can be used to compute the value of this replicate at time t from the value at t-1, giving:

$$x^i_{t|t-1} = f(x^i_{t-1|t-1}, u^i_t) \tag{1.25}$$

Note that this operation requires generation of a random input replicate u_t^i, which is a random sample drawn from the specified prior input probability density $p(u_t)$. Assuming

$$w_{t|t-1}^i = 1/N \tag{1.26}$$

then equations (1.25) yields the following approximation for the forecast probability density:

$$p(x_t|y_{1:t-1}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(x_t - x_{t|t-1}^i) = \frac{1}{N}\sum_{i=1}^{N}\delta[x_t - f(x_{t-1|t-1}^i, u_t^i)] \tag{1.27}$$

Note that the assumption of the even weight can be released. This forecasting step is just a Monte Carlo-based procedure for deriving $p(x_t|y_{1:t-1})$ from $p(u_t)$ and $p(x_t| y_{1:t-1})$.

The ensemble nonlinear filters differ in the the analysis step. Among all these filters, the ensemble Kalman filters are especially attractive for land surface problems. Since it will be used in all the following chapters it will be discussed in the following section and the SIR particle filter will be discussed in Chapter 2 in detail. In chapter 4 a new ensemble multiscale filter designed for large scale problems will be proposed.

1.5.1 The Ensemble Kalman Filter (EnKF)

The ensemble Kalman filter uses the Dirac expansions of (1.23) and (1.24) to approximate the conditional probability densities of x_t and it adopts the approximation of (1.25) and (1.26) during the forecasting step. However, the ensemble Kalman filter makes more assumptions at the analysis step. The Kalman filter analysis step can be derived from various perspectives. Here we take a Bayesian or distribution-oriented perspective because we are interested in the filter's ability to estimate properties of the conditional density $p(x_t|y_{1:t})$.

It is generally very difficult to derive an exact closed form expression for $p(x_t|y_{1:t})$ from Bayes theorem, especially for problems with nonlinear dynamics and measurement operators. However, it is possible to obtain an exact solution when the forecast states and measurements are jointly normal. This typically occurs only when the state and measurement equations are linear and all sources of uncertainty are normally distributed. In this special case the analysis density given by (2.1) is normal and completely defined by the following mean and covariance, which are the update expressions of the classical Kalman filter. In practice, adopting a joint normality assumption is equivalent to assuming that the forecast and measurement densities are adequately characterized by their means and covariances (i.e. higher-order moments are ignored in the analysis step). It is possible to use the Kalman update expressions even when the joint normality assumption does not apply. In this case the conditional statistics produced by the Kalman filter may not match the true values but they may be close enough to be useful.

In an ensemble version of the Kalman filter we need to generate an ensemble of analysis replicates at t, for propagation from t to $t+1$. The sample mean and covariance of this ensemble should converge to the mean and covariance of (1.14) and (1.17) in the limit as the number of replicates approaches infinity. There are many ways to generate analysis replicates that satisfy this requirement. In nonlinear applications it is best to use an ensemble generation method that preserves at least some of the non-normal characteristics of the forecast ensemble when normality assumptions do not apply.

One way to accomplish this is to generate an analysis ensemble directly from the forecast ensemble, using the following algorithm [34, 33]:

$$x^i_{t|t} = x^i_{t|t-1} + K_{s,t}[y_t + v^i_t - h(x^i_{t|t-1})] \tag{1.28}$$

$$w^i_{t|t} = w^i_{t|t-1} = \frac{1}{N} \tag{1.29}$$

Where $v^i_t$is a sample drawn from the measurement error probability density $p(v_t)$ and $K_{s,t}$ is a sample estimate of the Kalman gain K_t :

$$K_{s,t} = X_{t|t-1}Y^T_{t|t-1}[Y_{t|t-1}Y^T_{t|t-1} + C_{vv,t}]^{-1} \tag{1.30}$$

The columns of the sample matrices$X_{t|t-1}$ and$Y_{t|t-1}$ are constructed from the mean-removed replicates of $x_{t|t-1}$, $h(x_{t|t-1})$, and v_t.

The ensemble Kalman filter algorithm of (1.28) through (1.30) produces analysis replicates that converge to the exact Bayesian solution for normal states and measurements. When there are deviations from normality the filter is suboptimal but the replicates are able to inherit non-normal properties from the forecast.

1.5.2 Two variants of the EnKF's

1. The Square Root EnKF

There are a number of other versions of the ensemble Kalman filter that use different approaches for generating non-normal ensembles that conform to (12) and (13) [96]. For EnSRF, also called the deterministic EnKF, the update equation for mean follows

$$x^i_{t|t} = x^i_{t|t-1} + K_t(y_t - h(x^i_{t|t-1})) \tag{1.31}$$

where there is no artificially added measurement noise v^i_t, which differs from (1.28). To obtain the correct error covariance as ensemble size approaches infinity, Tippett (2003) presented a square root framework for updating the ensemble spread:

$$Z(t|t) = Z(t|t-1)(I - \beta V(t|t-1)V(t|t-1)^T) \tag{1.32}$$

where Z represents the mean removed ensemble, $V(t|t-1) = h(Z(t|t-1))$; $\beta = [D + (R_t D)^{1/2}]^{-1}$, and $D = V(t|t-1)^T V(t|t-1) + R_t$. So (1.32) is essentially a linear transform of$Z(t|t-1)$.

There haven not been any documented hydrological research using the square root filters. However, the basic concepts here are similar to [33] classical ensemble Kalman filter. We will focus on the classical version of the filter described above.

2. Hybrid Kalman filter

All the ensemble methods suffer from the sampling error issues for large problems simply because the affordable ensemble size is small. To overcome the sampling error issues, the hybrid Kalman filter [52] adopts the control variate idea which is a traditional variance reduction technique in Monte Carlo methods. At time t, it uses L^*, the square root form of error covariance matrix $P(t|t-1)$ from the extended Kalman filter, as control variate and then incorporates replicates E from the EnKF into the null space of L^* to form

$$E^{\perp} = (L^*(L^{*T}L^*)L^{*T})E \tag{1.33}$$

where the term $(L^*(L^{*T}L^*)L^{*T})$ is simply the complementary space of L^*. Using the constructed error space L^* and $E^{\perp}$, the error covariance matrix $P^*(t|t-1)$ is computed from

$$P^*(t|t-1) = \left[L^* \frac{E^{\perp}}{\sqrt{n-1}}\right] \left[L^* \frac{E^{\perp}}{\sqrt{n-1}}\right]^T \tag{1.34}$$

which is used for the gain calculation. All the operations involving $P^*(t|t-1)$ in the filter are in square root form so full matrix of $P^*(t|t-1)$ is never used. The drawback of this hybrid Kalman filter is similar to those of the extended Kalman filter: it requires additive model noise and linearization of the system model when calculating L^*.

1.6 Motivations and Thesis Organization

Even though the EnKF's have been studied in land surface problems, many of issues are still outstanding, such as the the nonlinearity and nonnormality effects of the land surface model on the performance of the EnKF is still unknown; the high dimensionality of the land surface problems needs more efficient data assimilation techniques; for accuracy reasons the sampling errors due to limited ensemble size needs to be eliminated as possible; utility of the efficient data assimilation techniques for land surface applications involving large scale correlation should be examined. These are the major motivations for this thesis that will be addressed in this section in more details.

1.6.1 Nonlinearity and Nonnormality Effects

Ensemble approaches can provide high order moments of the states through the uncertain state propagation using nonlinear dynamic model. However, the widely studied ensemble Kalman filters assumes normality in the ensemble and only use first two order moments to calculate the weights between prior information and observation. For land surface problems, the interactions between the model parameters and state variables make the system highly nonlinear. The resulting uncertain state variables might be nonnormal. Another contributing factor to the nonormality is the nonnormal inputs such as rainfall, the intermittency of which could create multimodal distributions. Skewed or bimodal ensemble of soil moisture would occur after a period of dry-down or rainfall event, which is largely due to the threshold constraint of soil moisture. For these cases, high order prior information is ignored when updating. How the ignored high order information would affect the performance of the update and the following error propagation is still unknown. Fortunately the particle filters for nonlinear filtering can be used as the benchmark to assess the extent to which nonlinearity and nonnormality affect the performance of ensemble Kalman filters. The second chapter would use the SIR (Sequential Importance Resampling) particle filter as the tool to address this issue.

1.6.2 High Dimensional State Estimation

The computational requirements of environmental data assimilation problems are a direct result of the wide range of time and space scales that need to be accommodated. For example, the number of states to be estimated in a three-dimensional dynamic land surface assimilation problem tends to be proportional to square of the ratio of the largest scale of interest (e.g. a regional weather system) to the smallest scale of interest (e.g. the scale of terrain or soil variations that affect local evapotranspiration and infiltration). This number can easily exceed 10^6, a value that is sufficiently large to render many popular estimation techniques impractical. Similar arguments can be made for oceanographic and meteorological data assimilation applications.

The high dimensionality of the state and measurement vectors does not necessarily reflect what might be called the intrinsic size of a land surface assimilation problem. This is because the states and measurements might be highly correlated or linearly dependent. For example, soil moisture tends to be highly correlated between upper soil layers after a prolonged wet period. There can also be significant horizontal correlation. In the limit when all surface layer pixels are saturated and soil properties are uniform, the surface soil moisture in an extended region can be described with a single number, the porosity. On the other extreme, after long drydown periods horizontal correlation tends to decrease, reflecting the effect of small-scale soil variability. In this case, the intrinsic problem size is high. But it may be possible to divide this large problem into many independent small problems, each associated with a single isolated soil column (at least until the next large rain event).

If we could account for such behavior in a systematic way we should be able to greatly reduce the computational effort required for large problems. The difficulty is that spatial correlation in a land surface problem changes over time, sometimes rapidly. Highly efficient estimation algorithms must be able to adapt to such changes, continually adjusting the intrinsic problem size. Development of such algorithms requires a careful look at the space-time structure of the system states. The third chapter would examine the evolution of the soil moisture spatial structure using the spectrum analysis of the forecasted ensemble from the Community Land Surface model. The implications of the dynamically changing spectrum for data assimilation would also be discussed.

The high dimensionality problem stated above is closely related to the sampling error in any data assimilation approach based on Monte Carlo methods. Ensemble approaches inevitably have error resulting from random sampling with the limited ensemble size. As in any Monte Carlo approaches, the sampling error depends on both the ensemble size and the intrinsic dimension of the sate vectors. For example, the highly correlated state variables require smaller ensemble size than the highly uncorrelated ensemble size to achieve the same level of sampling error. However, the sampling errors in the ensemble doesn't have to enter the filter. The trick is in how to exploit the characteristics of the ensemble to wisely use the information in the ensemble and eliminate the noise. An efficient ensemble filter should require less replicates while keeping the sampling error low, thus decreasing the spurious spatial correlation due to limited ensemble size or increasing the effective spatial correlation length. After having longer effective spatial correlation length, the influence radius for a particular pixel would be increased. Hence more information from the neighboring measurements can be utilized if available, which would ultimately increase the filter performance. The sampling error issue would also be discussed in this thesis.

1.6.3 Need for a Nonlinear Filter for High Dimensional Estimation

For a nonlinear non-Gaussian system with nonadditive errors, traditional estimation methods such as the extended Kalman filter can't work well due to its requirements for linearization of the system and additive error assumption. During the last decade nonlinear filtering theory has undergone a rapid development partly due to the reduced computational cost. Many popular nonlinear filtering algorithms such as particle fillers [2, 45], unscented Kalman filter [62], and ensemble Kalman filters [33, 10] require no linearization and additive Gaussian uncertainties assumption and have been proved successful. The basic idea behind these ensemble filters is to use Monte Carlo approach to approximate the propagation of probability density function (pdf) of the state variables and use Bayes' theorem or an update equation to ingest information from observations into the propagated ensemble. In the particle filters, the update is based on sequential importance sampling of the posterior pdf conditional on measurements. While in ensemble Kalman filters, the update equation is similar to the traditional Kalman filter using the Kalman gain and innovation from observation. The updated ensemble is then propagated forward using the nonlinear dynamics. However, the high dimensionality in both state and measurement as is common in land surface problems makes it very hard to implement the particle filters. It is impossible even to store the error covariance matrix in the traditional covariance based methods. To deal with this problem, some approaches have been developed.

Multiresolution method [101] provides a very powerful framework for efficient large scale estimation. It describes a covariance matrix using multiresolution autoregressive model, which leads to an efficient way to deal with high dimensionality encountered in a big problem. In general, a scale-recursive model can be identified using methods like canonical correlation [57] or predictive efficiency method [37]. For a general dynamical system the covariance structure may vary in time. For example, a land surface system has time varying spatial correlation across many scales in the states due to moving rainfall input and land surface system dynamics [73]. These identification methods are able to identify an approximate scale recursive model from any given second order statistics, which is especially ideal for a system with varying covariance structure.

Originally, the multiresolution method was developed for static estimation. In [53] it is extended to solve linear diffusion problems, where a tree model is used to approximate the error covariance matrix. After each update the tree model as an approximation of the error covariance matrix is then propagated forward. This is similar to the concept of dynamic bayesian network. However, the approach in [53] is not applicable for a general dynamic system with nonadditive uncertainties.

Another related model for large scale problems is the Bayesian hierarchical models used in the study of environmental processes [4]. These models usually assume tractable conditional distributions for the states or parameters. The posterior density can be iteratively obtained using Markov Chain Monte Carlo methods. But the uncertainties of parameters and inputs of a system model might be rather complex and can't be simply expressed in a simple form such as Gaussian additive noise. This method is not good for sequential estimation either.

The Ensemble Kalman filter (EnKF) as the other attempt to deal with problems with high dimensionality rely on a reduced rank approximation to the full error covariance matrix, thus helping with the computational feasibility [34]. But calculating the gain matrix in EnKF is still difficult for a problem with lots of measurements even using the reduced rank approximation. Another weakness in EnKF is that only a limited number of replicates can be generated for large problems so the resulting sampling errors in covariance matrix might be significant.

Building on the power of ensemble approach in describing the nonlinear high dimensional dynamics, and the strength of multiresolution methods in solving large scale problems, an ensemble multiscale filter (EnMSF) is then proposed in the fourth chapter to exploit the advantages in both methods. It is an efficient filter suitable for large scale nonlinear applications and can reduce the sampling error at the same time.

1.6.4 Large Scale Evapotranspiration (ET) Estimation

Up to now, there have not been many documented studies on how the data assimilation methodologies can help understand large scale land surface processes. One reason is that the big problem size and complicated uncertainties of the nonlinear land surface system hindered the research in this direction. Data assimilation would provide a unique tool to estimate ET which is hard to observe directly. The existence of high correlation between ET and soil moisture under certain conditions allows to estimate soil moisture first and then use a land surface model to project ET from these soil moisture estimate. The fifth chapter will attempt to use the proposed ensemble multiscale filter to assimilate synthetic multi-sensor surface data to obtain better soil moisture estimate and then the large scale evapotranspiration estimate.

1.7 Outline of Original Contributions

The major contributions of this thesis are closely tied to the thesis motivations. They are outlined as follows:

- For the first time, the effects of EnKF's normality approximations at update step on the distributions in land surface data assimilation problems are evaluated. This is done by comparing the conditional marginal distributions and moments estimated by the ensemble Kalman filter to those obtained from an SIR particle filter, which gives exact solutions for large ensemble sizes.

Comparisons for two land surface examples indicate that the ensemble Kalman filter is generally able to reproduce non-normal soil moisture behavior, including the skewness that occurs when the soil is either very wet or very dry. Its conditional mean estimates are very close to those generated by the SIR filter. Its higher-order conditional moments are somewhat less accurate than the means. Overall, the ensemble Kalman filter appears to provide a good approximation for nonlinear, non-normal land surface problems, despite its dependence on normality assumptions.

- Spectrum analysis of the soil moisture replicate matrices (covariance square root) reveals dynamic soil moisture pattern dependent on rainfall dynamics and soil properties. The spectrum implications for efficient data assimilation are also investigated for the first time.

  Singular value decompositions of the replicate matrices produced in an ensemble forecasting simulation experiment confirms the existence of dynamic spatial structure of surface soil moisture, which relies on the precedent cumulative rainfall and soil property structures. The singular value spectrum drops off quickly after rainfall events, when a few leading modes dominate the spatial structure of soil moisture. The spectrum is much flatter after a prolonged drydown period, when spatial structure is less significant. Different spatial structure of soil moisture indicates different update scheme should be used. Controlled experiments show that local ensemble Kalman filters are suitable for such problems during dry periods but give less accurate results after rainfall.

- A new efficient ensemble multiscale filter (EnMSF) is proposed to approximately solve general large scale nonlinear estimation problem with any dynamic system and arbitrary uncertainties. A new non-ensemble upward sweep update replacing the Willsky's version [101] is developed. For a linear gaussian system, the EnMSF is proved optimal in minimum mean squared error sense, giving conditional distributions.

At each prediction step realizations of the state variables are propagated using a fine resolution nonlinear dynamic model with any appropriate input or parameter uncertainties as in EnKF. At each update time, joint Gaussian distribution of states and measurements are assumed and the EnMSF uses predictive efficiency method to identify a multiscale tree to approximate the full covariance matrix given by the propagated ensemble. Then the upward and downward two-sweep updates are performed on the identified tree to estimate the state variables using all the available (multiresolution) data. By controlling the tree model parameters, the EnMSF can reduce sampling error while keep long range correlation in the prior ensemble. The EnMSF is highly computationally efficient and especially appealing to large scale estimation problems as compared to the covariance based Kalman filters.

- Soil moisture and evapotranspiration are estimated over the Great Plains at 5km resolution for June 2004, using synthetic 40km L-band passive and 5km active microwave soil moisture measurements following HYDROS specifications. This is the first high resolution soil moisture and evapotranspiration estimation (a high dimensional estimation problem) using multi-sensor data without local spatial correlation assumption.

  Soil moisture estimation results show that using both active and passive measurement is better than using either of them and using any measurement is better than the unconditional estimate. The results also prove that the soil moisture estimation improved by microwave measurements helps with the evapotranspiration and root zone soil moisture estimation.

Chapter 2

Assessing the Performance of the Ensemble Kalman Filter for Land Surface Data Assimilation

In sequential filtering the distributional properties of an uncertain state x_t , given a set of measurements $y_{1:t}$ taken through time t, are conveyed by the conditional probability density $p(x_t|y_{0:t})$. The random replicates generated by ensemble methods may be used to compute finite sample approximations to this density and its moments. When new measurements become available some version of Bayes theorem is typically used to update the replicates (and the corresponding distributional approximations). The accuracy of this update depends on the assumptions made when applying Bayes theorem as well as the number of replicates. The ensemble Kalman filter is particularly efficient because it relies on normality assumptions that greatly simplify the update process. But this simplification can also limit the filter's ability to provide accurate distributional information. Here we evaluate the accuracy of the ensemble Kalman filter by comparing its distributional estimates to those of a less efficient ensemble method that relies on an exact Bayesian update, i.e., the SIR particle filter. This is done for two examples that provide useful insight about the ensemble Kalman filter's performance in land surface applications.

2.1 The SIR Particle Filter

Particle filters are a class of sequential Bayesian ensemble algorithms that can be derived from a discrete version of Bayes theorem. Arulampalam et al. (2002) [2] provide a useful tutorial that shows how several different particle filtering algorithms may be developed from the perspective of Sequential Importance Sampling (SIS). We use the Sequential Importance Resampling (SIR) filter here because it is easy to implement and converges to the exact Bayesian Solution as the number of replicates approaches infinity. It is also well-suited for the land surface application, where uncertain time-dependent inputs are generally more important than initial condition errors. In other applications, other types of particle filters may give better performance for a given number of particles.

2.1.1 A Simple Interpretation

The SIR algorithm adopts the approximation of (1.25) and (1.26) during the forecasting step of filtering. The analysis step is based on the following form of Bayes theorem:

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} = cp(y_t|x_t)p(x_t|y_{1:t-1}) \tag{2.1}$$

where c is a normalizing constant that insures that $p(x_t|\ y_{1:t})$ integrates to one. If we substitute the Dirac expansions for $p(x_t|y_{1:t-1})$ and $p(x_t|y_{1:t})$ into (2.1) we can relate the analysis density replicate values and weights of the unknown analysis density (left-hand side) to those of the known forecast density (right-hand side). In the SIR filter the analysis replicate values are initially kept the same as the forecast values and only the analysis weights are changed. This gives:

$$x^i_{t|t} = x^i_{t|t-1} \tag{2.2}$$

$$w^i_{t|t} = cp(y_t|x^i_{t|t})w^i_{t|t-1} = cp(y_t|x^i_{t|t-1})w^i_{t|t-1} = \frac{c}{N}p(y_t|x^i_{t|t-1}) \tag{2.3}$$

where$p(y_t|x^i_{t|t-1})$is the likelihood function for the propagated replicate$x^i_{t|t-1}$. The likelihood function can be readily computed if the measurement error is additive (as assumed here) since:

$$p(y_t|x^i_{t|t-1}) = p_{yt}(y_t|x^i_{t|t-1}) = p_{vt}[y_t - h(x^i_{t|t-1})] \tag{2.4}$$

where p_{vt} is the known (e.g. normal) probability density of the measurement error v_t. The likelihood function can be viewed as a measure of the "closeness" of the replicate $x^i_{t|t-1}$ to the measurement y_t.

We could substitute (2.2) and (2.3) directly into (1.24) to obtain an approximation of the analysis probability density but the result may be unsatisfactory unless the number of replicates is very large. This is because (2.3) gives replicates "closer" to the measurements much more weight than those that are "further away". This can result in the "collapse" of the ensemble to a very small number of replicates with high weights, giving a very coarse discrete representation of the analysis probability density. In order to prevent this, the SIR filter resamples the ensemble with replacement N times. The probability that replicate i is selected on sample k is equal to its weight:

$$p(\text{replicate } i \text{ selected on sample } k) = w^i_{t|t} \tag{2.5}$$

By construction, this resampling operation creates a new analysis ensemble of N equally likely replicates with the following values$x^k_{t|t}$ and weights$w^k_{t|t}$ (for $k = 1,\dots,N$):

$x^k_{t|t}$ =replicate value selected on sample k

$$w^k_{t|t} = \frac{1}{N} \tag{2.6}$$

The new analysis ensemble is a subset of the old analysis ensemble. Old replicates with high weight are more likely to be repeated in the new ensemble and old replicates with low weight are more likely to be omitted. Once the resampling operation is completed 2.5) and (2.6) can be substituted into (1.24) to give:

$$p(x_t|y_{1:t}) \approx \frac{1}{N}\sum_{k=1}^{N}\delta(x_t - x_{t|t}^k) \tag{2.7}$$

The new equally weighted resampled replicates can then be propagated from t to $t + 1$, following the procedure given in (1.25) and (1.26) (with t replaced by $t + 1$ and k by i). Although many of the resampled analysis replicates at t have the same value, these values diverge in the subsequent propagation to $t + 1$ because of the influence of the random input noise u_{t+1}. This keeps the ensemble from collapsing and is why the SIR approach works best for problems with random inputs.

The SIR filter's ensemble statistics (marginal densities, moments, etc.) can be shown to converge to their exact counterparts as the number of replicates approaches infinity. The version of the SIR filter described here assumes that the measurement errors are additive and independent over time but does not restrict the form of the probability densities for x_t, u_t, or v_t. or the form of the functions $f(\cdot)$ and $h(\cdot)$. The primary disadvantage of the SIR filter is the large number of replicates required to accurately represent the multivariate conditional probability densities of x_t. When the number of measurements exceeds a few hundred the SIR filter is not practical for land surface problems. However, it provides a very useful performance benchmark for small problems since it yields optimal conditional densities (as well as conditional means and other moments) if the ensemble is sufficiently large.

2.1.2 Sequential Importance Resampling

More generally, the particle filter can be derived using the idea of sequential importance sampling and resampling, which allows more flexible and efficient sampling approaches. The simple interpretation derived in the last section is only a special case of the sequential importance sampling and resampling approach. To have a broader view of the particle filter, the complete derivation from sequential importance sampling is given here.

1) Importance sampling

Suppose we want samples x^i from $p(x)$that is difficult to sample, where i is index of sample:

$$p(x) \approx \sum_{i=1}^{N} w^i \delta(x - x^i), \tag{2.8}$$

in which w^i is the associated weight for x^i and $\sum_i w^i = 1$, δrepresents delta function. If $p(x) \propto \pi(x)$, where $\pi(x)$is easy to evaluate, and we have another $q(x)$(proposal density) that is easy to sample, we can sample x^i from $q(x)$ first and calculate the weight by $w^i \propto \frac{\pi(x^i)}{q(x^i)}$. This is the traditional importance sampling for variance reduction (Hammersley, 1964 [49]). For filtering problems, we want samples $x^i_{0:t}$ from $p(x_{0:t}|y_{1:t})$, where the subscript is time index, x is the state vector and y is measurement vector:

$$p(x_{0:t}|y_{1:t}) \approx \sum_{i=1}^{N} w^i_t \delta(x_{0:t} - x^i_{0:t}) \tag{2.9}$$

Use the importance sampling method, one can obtain:

$$w^i_t \propto \frac{\pi(x^i_{0:t}|y_{1:t})}{q(x^i_{0:t}|y_{1:t})} \tag{2.10}$$

where $\sum_i w^i_t = 1$, $x^i_{0:t}$is from proposal density function $q(x^i_{0:t}|y_{1:t})$.

2) Sequential importance sampling

Usually for a dynamic system, the state at previous time step includes a lot of information that can be exploited to give recursive estimation of the current state. That way replicates before the previous steps can be discarded to save memory. In order to sequentially update $x_{0:t}$ using $y_{1:t}$, write $p(x_{0:t}|y_{1:t})$ as

$$p(x_{0:t}|y_{1:t}) = \frac{p(y_t|x_{0:t}, y_{1:t-1})p(x_{0:t}|y_{1:t-1})}{p(y_t|y_{1:t-1})} \tag{2.11}$$

Since the second item in the numerator $p(x_{0:t}|y_{1:t-1}) = p(x_t|x_{0:t-1}, y_{1:t-1})p(x_{0:t-1}|y_{1:t-1})$, plug it into (2.11) we have

$$p(x_{0:t}|y_{1:t}) = \frac{p(y_t|x_{0:t}, y_{1:t-1})p(x_t|x_{0:t-1}, y_{1:t-1})p(x_{0:t-1}|y_{1:t-1})}{p(y_t|y_{1:t-1})}. \tag{2.12}$$

Assume $p(y_t|x_{0:t}, y_{1:t-1}) = p(y_t|x_t)$, $p(x_t|x_{0:t-1}, y_{1:t-1}) = p(x_t|x_{t-1})$, which means the current measurement is only dependent on the current state and the process is Markovian. Rewrite (2.12) as

$$\begin{aligned} p(x_{0:t}|y_{1:t}) &= \frac{p(y_t|x_t)p(x_t|x_{t-1})p(x_{0:t-1}|y_{1:t-1})}{p(y_t|y_{1:t-1})} && (2.13) \\ &\propto p(y_t|x_t)p(x_t|x_{t-1})p(x_{0:t-1}|y_{1:t-1}) && (2.14) \end{aligned}$$

Here we obtain $\pi(x) = p(y_t|x_t)p(x_t|x_{t-1})p(x_{0:t-1}|y_{1:t-1})$ as in importance sampling. Select a proposal density $q(x_{0:t}|y_{1:t})$, and suppose

$$q(x_{0:t}|y_{1:t}) = q(x_t|x_{0:t-1}, y_{1:t})q(x_{0:t-1}|y_{1:t-1}) \qquad (2.15)$$

If we have samples $x^i_{0:t}$ from $q(x_{0:t}|y_{1:t})$, then

$$\begin{aligned} w^i_t &\propto \frac{p(y_t|x^i_t)p(x^i_t|x^i_{t-1})p(x^i_{0:t-1}|y_{1:t-1})}{q(x^i_t|x^i_{0:t-1}, y_{1:t})q(x^i_{0:t-1}|y_{1:t-1})} && (2.16) \\ &= \frac{p(y_t|x^i_t)p(x^i_t|x^i_{t-1})}{q(x^i_t|x^i_{t-1}, y_t)} w^i_{t-1} && (2.17) \end{aligned}$$

In the sequential importance sampling and resampling (SIR) filter, we choose $q(x^i_t|x^i_{t-1}, y_t) = p(x^i_t|x^i_{t-1})$, i.e., the proposal density function is just the prior density and $x^i_{0:t}$ is the prior replicates. Hence,

$$w^i_t \propto p(y_t|x^i_t)w^i_{t-1} \qquad (2.18)$$

If x^i_{t-1} are evenly weighted, then $w^i_{t-1} = 1/N$. (2.18) reduces to

$$w^i_t \propto p(y_t|x^i_t) \qquad (2.19)$$

In this simple equation (2.19), the final weight is proportional to the likelihood function so it is easy to implement. To propagate forward the new information from observation y_t, it is necessary to eliminate replicates with low weights and only retain those with high weights. Otherwise the computation efforts would be wasted on those with low probability. Resampling step is the same as discussed in the last section. After resampling, each replicate has the same weight $1/N$. It's worth noting that the proposal distribution doesn't have to be the prior distribution, which allows for more efficient filter if using some problem dependent proposal distribution.

The procedure of implementing particle filter is depicted in Figure 2-1, where red ovals represent updated ensemble and blue ovals propagated ensemble. In this diagram, the initial condition is perturbed and propagated forward. When observation is available, the weights are calculated from the prescribed likelihood function. Then the resampling step is done to prevent the population degeneracy. In Figure 2-1 only two different replicates are retained and duplicated after resampling since compared to other two replicates they are closer to the observation which is indicated by a magenta cross.

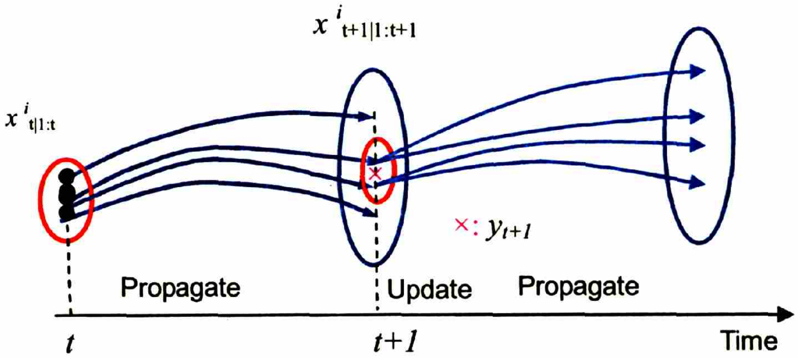


Figure 2-1: Schematic diagram of implementation procedures for particle filter. Red ovals represent updated ensemble and blue ovals propagated ensemble. Lines are trajectories of replicate. Magenta cross is observation at t+1.

As is shown, the particle filter doesn't assume any particular probability density function of $x_{0:t}$. As long as the likelihood function $p(y_t|x_t)$ is known, the posterior probability $p(x_{0:t}|y_{1:t})$ can be derived from (2.19). For a very large number of replicates, the true posterior PDF can be obtained without loss of high order prior information. Yet a caveat of the particle filter is that the resampling step might incur sample impoverishment, resulting in many identical states in the population. Sample impoverishment can lead to slow convergence of the particle filter. For the application here, it's not an issue so it will not be discussed.

It should also be pointed out that since particle filter approximates PDF by sampling approach it can't avoid the curse of dimensionality. For high dimensional problems, to represent a high dimensional PDF the required ensemble is prohibitively expensive if no special measure is taken such as Markov chain Monte Carlo. Hence only a small size problem is considered in this paper to make the assessment feasible. Another important fact about particle filter is that it only works well with system with process noise, because it's necessary to have a wide enough ensemble to encompass the measurement. Otherwise, it's possible that all of the replicates are not close enough to the measurement, which would cause many updated replicates sharing the same trajectory during the next propagation period. One remedy is to use other distributions, such as truncated Cauchy as in [98]. Also, the magnitude of observation error affects the convergence speed because for small error the number of good replicates would be less and more replicates would be discarded when resampling.

2.2 A Simple Nonlinear Land Surface Data Assimilation Example

Soil moisture is one of the key states controlling the partitioning of water and energy fluxes at the land-atmosphere boundary. It is likely to be skewed to the wet end (after precipitation) or the dry end (after a prolonged drydown period). Here we use a simple scalar example motivated by soil moisture behavior to illustrate the two nonlinear filters described in the previous section. Suppose a scalar soil moisture value x at a particular measurement time has the following forecast probability density:

$$p(x) = 27.7\exp(-\frac{x}{0.1}) \qquad 0.1 \leq x \leq 0.5 \tag{2.20}$$

This truncated exponential density is shown in Figure 2-2 a. The associated mean x is 0.19 and the variance is $C_{xx} = (0.08)^2$. We suppose that a single measurement $y = x + w$ is taken, where w is a zero mean normally distributed additive error independent of x with standard deviation $C_{vv} = (0.05)^2$.

For this problem the analysis probability density $p(x|y\ )$may be derived in closed

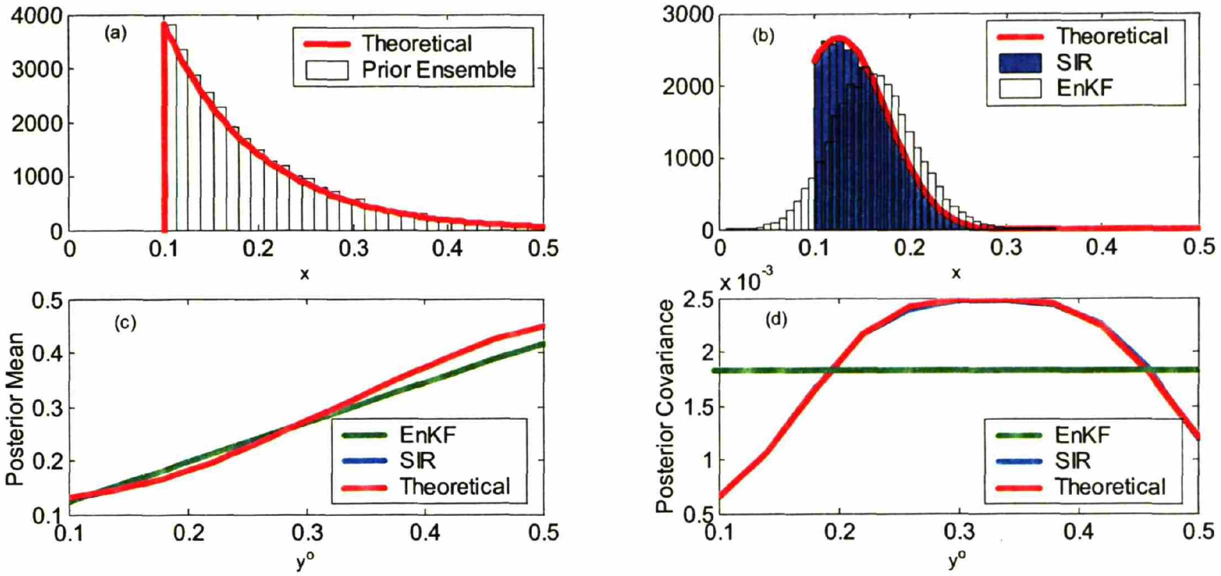


Figure 2-2: Estimates of scalar soil moisture state statistics for a skewed prior probability density where y^o is the actual observation. (a) Prior pdf: $p(x)$=$27.2exp(-x/0.1)$,0.1 $\leq x \leq$ 0.5; (b) Posterior pdf for y^o = 0.15, R= 0.05, estimated with SIR and Ensemble Kalman filters. Also plotted is the theoretical Bayesian solution; (c) Posterior mean vs. y^o ; (d) Posterior covariance vs. y^o.

form from Bayes theorem:

$$p(x|y) = cp(y|x)p(x) \tag{2.21}$$

This exact analysis density is plotted in Figure 2-2b for a measurement value y_0= 0.15, together with the results obtained from an SIR filter and an ensemble Kalman filter, each using 30,000 replicates (this large sample size essentially eliminates sampling error problems). The SIR filter closely approximates the skewed exact analysis density. The ensemble Kalman filter analysis density is much more normal in shape, reflecting the influence of the normally distributed measurement perturbations v^iadded in the update step. As the measurement error becomes larger the Kalman gain eventually becomes very small, the forecast replicates dominate, and the Kalman analysis density becomes more skewed.

The exact, SIR filter and ensemble Kalman filter analysis means are plotted vs. the measurement value y_0 in Figure 2-2c. The Kalman filter analysis mean deviates only slightly from the exact and SIR filter means. Figure 2-2d shows that the analysis standard deviations for the two filters behave quite differently. The SIR filter standard deviation depends on the measurement value while the ensemble Kalman filter standard deviation does not. So the Kalman filter underestimates uncertainty for mid-range measurements and overestimates uncertainty for low or high measurement values.

Although this scalar example is very simple it suggests that differences between the SIR and ensemble Kalman filters for land surface problems may be more apparent in the higher-order moments than in the analysis means. We investigate this hypothesis further in the next section.

2.3 Formulation of an Observing System Simulation Experiment (OSSE)

In this section we describe a land surface simulation experiment that enables us to compare the performance of the suboptimal ensemble Kalman filter to an optimal SIR filter for a realistic land surface application. The problem is to characterize soil moisture and evapotranspiration from remotely sensed passive microwave (radiometer) measurements. Land surface dynamics are described by the Community Land Model (CLM, version 2.0) (Bonan, 1996, 2002 [5]). Radiometer measurements are described by a nonlinear radiative transfer model (Njoku et al., 2002 [77]). Input uncertainties and measurement errors are described by statistical models that are intended to realistically represent natural variability. These models determine how the replicates of the ensemble filters are generated.

2.3.1 The Land Surface and Radiative Transfer Models

The CLM is a nonlinear spatially distributed model that describes energy, momentum, water, and CO2 exchange between the land and the atmosphere. Dynamic inputs to the model include precipitation, wind speed, air temperature, pressure, humidity, and solar radiation. Time-invariant inputs include soil and vegetation classifications. The model is discretized into square pixels that are each divided into several soil layers. Moisture and heat can only move vertically within individual pixels. Further details are discussed in Bonan (1996, 2002). Although moisture does not flow between pixels the states in different pixels are correlated by virtue of their dependence on spatially correlated inputs such as precipitation and vegetation.

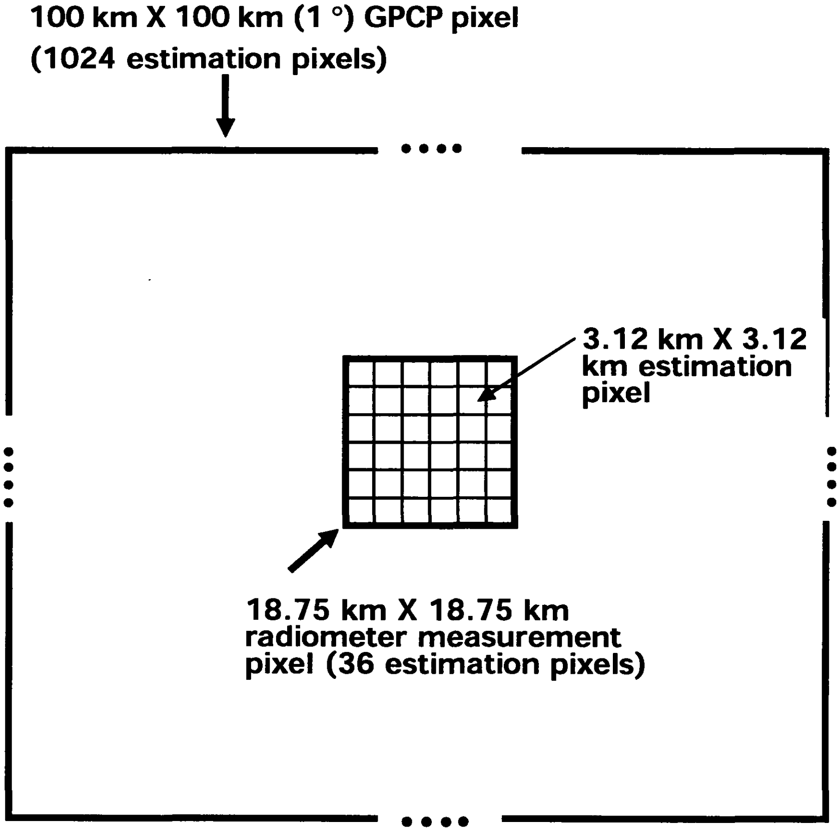

100 km X 100 km (1 °) GPCP pixel
(1024 estimation pixels)
3.12 km X 3.12 km estimation pixel
18.75 km X 18.75 km radiometer measurement pixel (36 estimation pixels)

Figure 2-3: Multiple scales used in the land surface OSSE. Precipitation data are available in a single 100 km by 100 km (GPCP) pixel, synthetic radiometer measurements are generated in a single 18.75 km. by 18.75 km pixel, and estimates are computed in 36 pixels, each 3.12 km by 3.12 km., nested inside the radiometer pixel.

The study region for our computational experiment reflects conditions at the Southern Great Plains (SGP97) site in eastern Oklahoma. This 18.75 km by 18.75 km region is shown in Figure 2-3. It is discretized over a 6 by 6 grid of (approximately) 3.12 km by 3.12 km estimation pixels with 8 soil layers in each pixel. The study region is small enough to be feasible for a SIR filter assimilation while large enough to reveal the impacts of horizontal correlation. The land use is assumed to be cropland with a loam soil and the soil layers have thicknesses (from top to bottom) of 2, 3, 5, 8, 12, 20, 57, 88 cm respectively. The CLM model states include soil moisture and soil temperature in the center of each soil layer as well as surface soil temperature, canopy temperature, and canopy intercepted water, for a total of 684 states in our 36 pixel grid. The CLM derives evapotranspiration from these states. The study period corresponds to a 28 day field campaign conducted from June 19, 1997at 0 hrs. UTC through July 16, 1997 at 15 hrs. UTC (Margulis, 2002, [70]). Input data are generated and the CLM is run for a 1 hr. time step.

Synthetic radiobrightness measurements can be related to soil moisture through soil reflectivity, as described by the Fresnel equation. For our experiment this process is described by the following expression for brightness temperature (Njoku et al., 2002 [77]):

$$T_b = T_s(1 - r_H)\exp(-\tau) + T_c(1 - \omega)[1 - \exp(-\tau)][1 + r_H \exp(-\tau)] \qquad (2.22)$$

where T_s and T_c are surface and canopy temperature (in ° K) and r_H is the horizontal polarization soil reflectivity. For L band (1.4 GHz) microwave, the vegetation can be considered predominantly absorbing with a small single scattering albedoω, and the vegetation opacity along the slant path is given by (Jackson and Schmugge, 1991 [59])

$$\tau = bw/\cos\theta \qquad (2.23)$$

where$w$is vegetation water content (Kg/m^2); b is vegetation-specific parameter; and θ is the incidence angle. The vegetation water content is derived from Normalized Difference Vegetation Index (NDVI) data (Jackson et al., 1999). Rough surface reflectivity is derived from the procedure described by (Choudhury et al., 1979 [19])

$$r_H' = r_H \exp(-h\cos^2\theta) \qquad (2.24)$$

where r_H is the smooth surface reflectivity and h is a vegetation-specific parameter. In our experiment, w, h, and b have the values 0.3kg/m^2, 0.1, and 0.04 respectively. The view angle θ is set to zero and the scattering albedo is 0.03.

2.3.2 Uncertain Model Inputs and Measurement Errors

The primary sources of uncertainty in land surface applications are time-invariant soil properties, time-dependent meteorological inputs, including precipitation, and initial conditions. In the ensemble approach random replicates for each of the uncertain inputs are provided to the CLM, which generates random replicates of the land surface states. Corresponding radiobrightness values at the estimation pixel scale are generated by the radiative transfer model of (2.22). The time-dependent random inputs can cause the ensemble to spread during the propagation step while assimilation of radiobrightness measurements can cause the ensemble to narrow at the analysis step. These effects are moderated by the physics of the problem, which constrains the states to lie in limited ranges (e.g. the volumetric soil moisture must lie between 0.0 and the porosity, which is less than 1.0).

The uncertain inputs are generated by transforming nominal input values to obtain sets of physically realistic replicates. This is done is various ways, depending on the variable. Table 2.1 lists the uncertain inputs and measurement errors considered in our simulation experiment. Note that different methods are used to introduce randomness for different inputs. The soil, vegetation, and precipitation inputs deserve some elaboration.

The nominal soil is assumed to be loam throughout the study region. Loam corresponds to a certain section of the classical silt-sand-clay soil triangle. The soil properties associated with different replicates and different pixels are obtained by selecting random points in the loam section and then reading off the corresponding silt, sand, and clay fractions, which are used by CLM. Different independent random samples are taken in different pixels and soil layers so soil property fluctuations are not correlated over space.

The vegetation type is assumed to be cropland through the study region. The

Variable	**Specified Nominal Value**	**Uncertainties in Replicates**
Soil fractions (sand-silt-clay)	Loam over entire study region	Uniformly distributed points in loam section of soil triangle
Vegetation	Cropland with: LAI=1.6, SAI=0.4 (June) LAI=1.3, SAI=0.8 (July)	Spatially uncorrelated multiplicative uniform noise $U[0.85, 1.15]$ for LAI.
Humidity, solar radiation, wind speed	Oklahoma Mesonet time series at El Reno, assumed to apply over entire study region.	Spatially and temporally uncorrelated multiplicative uniform noise: Relative humidity: $U[0.9, 1.1]$ Solar radiation: $U[0.9, 1.1]$ Wind speed: $U[0.7, 1.3]$
Air temperature	Oklahoma Mesonet time series at El Reno, assumed to apply over entire study region.	Spatially and temporally uncorrelated additive uniform noise $U[-4\ ^\circ\mathrm{K}, +4\ ^\circ\mathrm{K}]$
Precipitation	GPCP 1° daily data for SGP97 region	Nominal GPCP values downscaled in time from daily to hourly values with random pulses model. Temporally downscaled replicates downscaled in space from 100 km to 3.1 kilometer pixels with multiplicative cascade model
Initial soil moisture (at -10 days)	Specified soil moisture profiles	Spatially uncorrelated additive Gaussian noise $N(0.0, 0.3)$.
Initial soil temperature (at -10 days)	Specified temperature profiles	Spatially uncorrelated additive Gaussian noise $N(0.0, 4\ ^\circ\mathrm{K})$.
Radiobrightness measurement	Simulated "true" value at 18.3 km by 18.3 km scale	Temporally uncorrelated additive Gaussian noise $N(0.0, 3\ ^\circ\mathrm{K})$.

Table 2.1: Summary of uncertain inputs and measurement errors for the land surface simulation experiment

CLM characterizes land use types in terms of the leaf area index (LAI) and the stem area index (SAI). It uses these indices to compute various model vegetation parameters that control net radiation, energy partitioning, and intercepted water capacity. In our experiment the nominal LAI for cropland is 1.6 for June and 1.3 for July. The nominal SAI is 0.4 for June and 0.8 for July. Individual LAI replicates are generated by multiplying the nominal value by a uniformly distributed random variable in the range [0.85, 1.15]. SAI is treated as a deterministic input.

Precipitation displays significant correlation in time and space and has a patchy pattern that cannot be reproduced with simple multiplicative or additive perturbations to spatially uniform nominal values. A more realistic option is to generate small-scale replicates by downscaling (or disaggregating) larger-scale measurements of real precipitation over both time and space (Margulis, 2004 [68]). Downscaling relies on statistical models of small-scale variability.

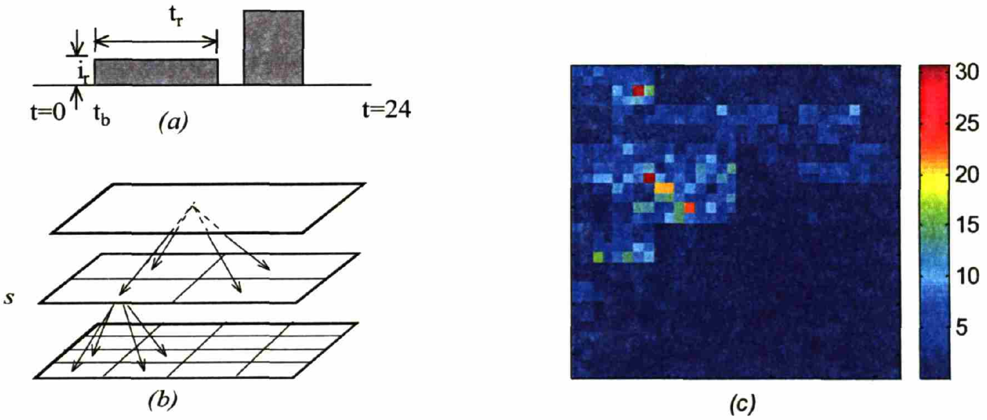


Figure 2-4: Spatial and temporal rainfall disaggregation model. (a) Rectangular pulses model (RPM) for temporal disaggregation of daily rainfall, (b) multiplicative cascade model for rainfall spatial disaggregation, and (c) one realization from the random cascade model over a 32x32 grid

Our simulation experiment uses nominal precipitation data from the Global Precipitation Climatology Project (GPCP). These daily data are available in the SGP97 region at a spatial resolution of 1 degree (100 km by 100 km). The GPCP time series for SGP97 during the 28 day time period of interest in our experiment is shown in

Figure 2-4d. We need to downscale this GPCP data from daily to hourly values in time and from 100 km by 100 km to 3.1 km by 3.1 km values in space, as indicated in Figure 2-3.

Our temporal downscaling procedure is based on a probabilistic rectangular pulses model (RPM) (Margulis and Entekhabi, 2001 [67]) that is constrained to reproduce observed daily totals (see Figure 2-4a). The RPM treats rainfall events as random rectangular pulses with an exponentially distributed constant intensity $i_{r,}$ and duration t_r and a uniformly distributed arrival time between 0 and 24 - t_r. Different RPM replicates have different hourly rainfall values. A given replicate may have no rainfall in any particular hour but its 24 hourly values must add up to the observed GPCP daily value. In our experiment the RPM mean intensity is 3.2 mm/hr. for June and 2.3 mm/hr. for July and the mean time between storms is 5.0 hrs. for June and 8.0 hrs. for July. These were estimated from climatological data (Hawk and Eagleson, 1992 [51]).

The temporal downscaling procedure provides 1 hr. precipitation replicates at the 100 km by 100 km GPCP measurement scale. These coarse resolution replicates can be downscaled to the 3.1 km by 3.1 km estimation pixel scale if we suppose that rainfall follows a multiplicative cascade model that relates intensities at different scales (Gupta and Waymire, 1993 [48]; Gorenburg et al., 2001 [46]). This model can be portrayed as a 6 level tree composed of groups of pixels (nodes) covering regions of different areas (see Figure 2-4b). The top (root) node defines the coarsest scale (one GPCP pixel) while the bottom nodes define the finest scale (the 1024 estimation pixels contained in the GPCP pixel). The rainfall value at a given node is obtained by multiplying the value at the next coarsest node (the parent) by a random lognormally distributed coefficient $W(s)$:

$$W(s) = \exp[w(s) - \sigma_w^2(s)/2] \tag{2.25}$$

where$w(s) \sim N[0, \sigma_w^2(s)]$, $\sigma_w(s) = 2^{-0.3(s-1)}$, and the scale index s increases from 1 at the root node of the cascade to 6 at the finest scale. A typical realization from this random multiplicative cascade is shown in Figure 2-4c. The cascade model generates rainfall that has a patchy pattern and is correlated over space.

In our simulation experiment the cascade model generates spatially downscaled rainfall on a 32x32 grid with a finest scale resolution of about 3.1 km, enforcing the same spatial pattern for each replicate in each hour of a given rainy day but allowing the rainfall intensity to change every hour. Rainfall intensities at the finest scale are normalized for each replicate to insure that the total rainfall at this scale is equal the total rainfall at the GPCP scale. A 6x6 portion of this grid provides the 3.1 km by 3.1 km rainfall data needed by the CLM model.

The CLM model is started at -10 days with random initial conditions generated by perturbing uniform soil moisture and temperature profiles. Each replicate is run forward with the model for 10 days to t = 0 to allow moisture in individual pixels to redistribute in accordance with local soil properties. The resulting soil moisture and temperature replicates initialize the SIR and Kalman filter ensemble simulations.

2.4 Simulation Experiment Specifications

For our synthetic experiment "truth" is defined by the state from a single CLM run obtained for a particular set of soil, vegetation, meteorological, and initial condition replicates, as described above. The CLM states and associated soil properties for this "truth" replicate are then used in (2.22) to generate a synthetic brightness temperature measurement at 15 UTC every day during the 28 day simulation period. This measurement is defined at a coarser scale than the model states, reflecting the lower resolution of anticipated satellite microwave radiometer measurements. In particular, we assume that the microwave measurement covers an 18.3 km by 18.3 km area (6 by 6 pixels) and is the average of the 36 pixel-scale brightness values computed from (2.22). At each measurement time a zero-mean normally distributed random perturbation is added to the averaged brightness temperature to account for the effect of measurement noise. This set of noisy measurements is provided to the two ensemble filters.

2.5 Results of the Simulation Experiment

It is useful to start our comparison of the SIR and ensemble Kalman filters by examining surface (top layer) soil moisture, surface soil temperature, and evapotranspiration replicates produced by the SIR filter at a typical pixel (pixel 9). These time series are shown in Figure 2-5, together with the applicable 1 degree daily GCPC precipitation record for the study period. The red line is the "true" replicate, the thick blue line is the mean of the SIR filter ensemble, and the thin cyan lines are the individual SIR filter replicates.

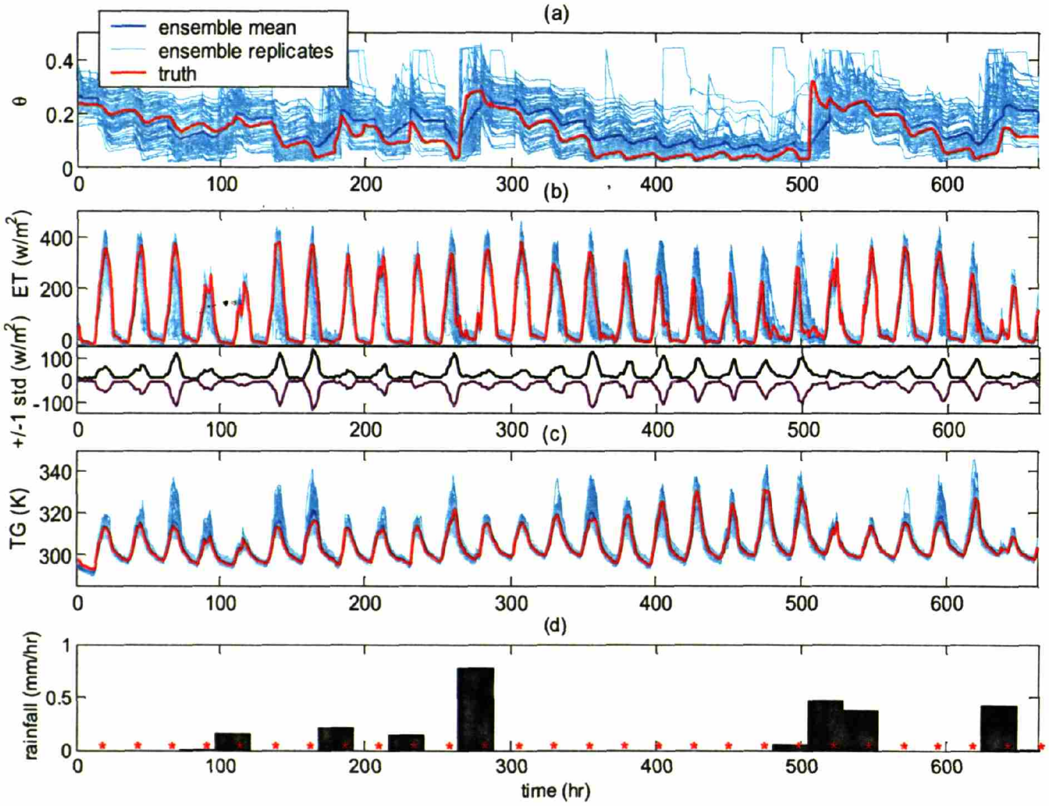


Figure 2-5: Ensembles from SIR solution at pixel 9 and the associated GPCP rainfall data: (a) ensemble of the first layer soil moisture θ; (b) ensemble of evapotranspiration and plus and minus one standard deviation of the ensemble; (c) ensemble of surface temperature, and (d) GPCP 1DD daily rainfall data time series. The asterisks on time axis of (d) represent the measurement times.

Before t = 100 uncertainty in soil moisture, indicated by the ensemble spread, is mainly due to uncertainties in initial conditions, soil properties, and LAI. After rainfall events occur, the uncertainties in surface layer soil moisture primarily reflect uncertainties in precipitation. Note that the ensemble spread is narrower during the dry periods, when the absence of rainfall makes it easy to infer that soil moisture values are low, even without the added information provided by radiobrightness measurements. By contrast, the ground temperature ensemble spread is wider during dry periods and narrower during wet periods. The spread of the evapotranspiration ensemble depends strongly on time of day, peaking just after noon. This is more apparent in the evapotranspiration ensemble standard deviation plot included just below the replicate plot. Replicates from the ensemble Kalman filter have a very similar structure.

It is important that our comparison of the SIR and ensemble Kalman filters is based on enough replicates to insure that sampling error is not a significant factor. Figure 2.5 shows the spatial root mean squared error (RMSE) for the top layer soil moisture, computed over all analysis times, where error is defined as the difference between the analysis ensemble mean and the "true" value. Also plotted are error bars that show plus or minus one standard deviation of the RMSE, computed over q filter runs started with different filter ensemble random seeds. The truth and measurements are kept the same for these runs. Clearly, the SIR filter needs more replicates to converge, although it eventually gives nearly the same RMSE as the Kalman filter. This is not surprising, considering that the converged SIR filter needs to resolve higher-order distributional properties that are ignored by the Kalman filter.

Figure 2-7 shows marginal forecast (left) and analysis (right) probability densities for pixel 9 surface soil moisture for some typical analysis times. Open loop (unconditional) densities are also shown for comparison. At the first analysis time, just prior to the first measurement ($t = 15$ hrs.), both filters and the open loop share the same forecast density with a skewness of 0.2 and kurtosis of 3.5 (since there have not yet been any measurements). The difference in the SIR and Kalman filter analysis densities at this time is minimal. The beneficial effect of the measurement is best revealed by a comparison of the open loop and analysis densities.

The densities plotted at times $t = 231$ and 279 show conditions during two rainy

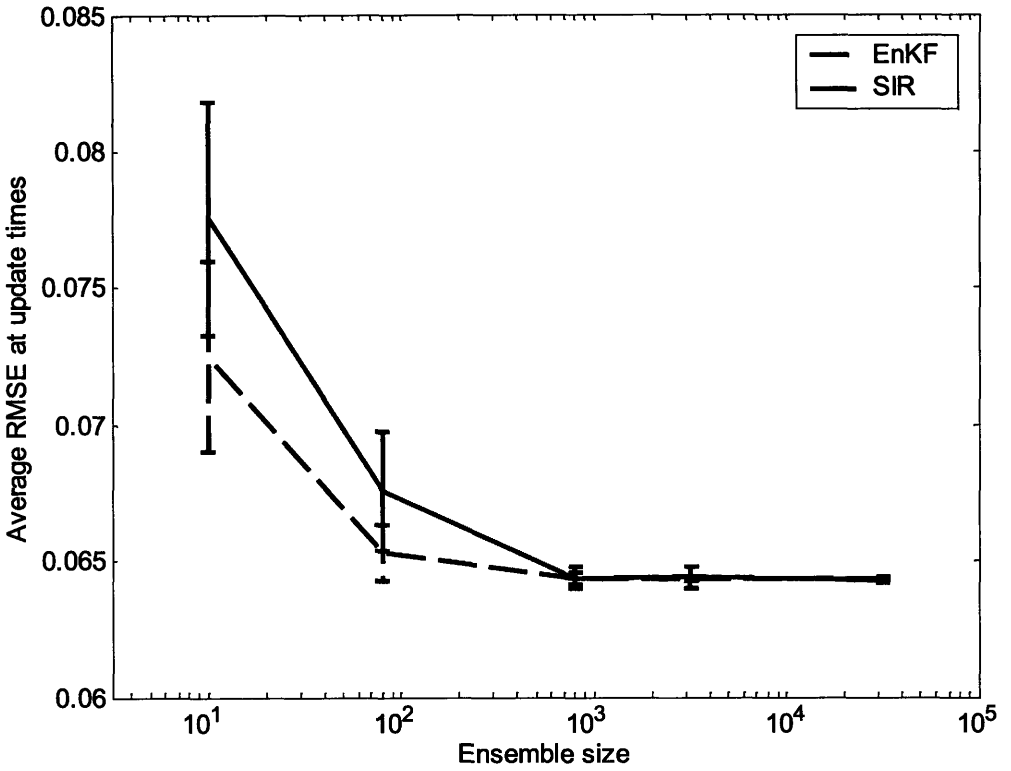


Figure 2-6: Averaged spatial RMSE of surface layer soil moisture at measurement times vs. replicate numbers. Error bars show plus or minus one standard deviation of the RMSE, computed over q filter runs started with different filter ensemble random seeds. For ensemble size n= 10, 80, 800, 3200, 32000, the run times q= 40, 10, 8, 6, 4 respectively.

periods. The skewness to the left in both of the forecast densities reflects the effects of the preceding drydown. The measurements at both times move the density noticeably toward the wet end producing significant differences between the open loop and filtered densities. Here again, differences between the SIR and Kalman filter analysis densities are minor. Also, at $t = 351$, after a drydown period all of the forecast densities reveal bimodal behavior. This bimodality is likely due to the properties of the multiplicative cascade rainfall model, which tends to produce replicates with wet or dry patches. The analysis densities for the SIR and ensemble filters at $t = 351$ are noticeably different.

After a long period of drydown, at $t = 471$, the forecast and analysis densities

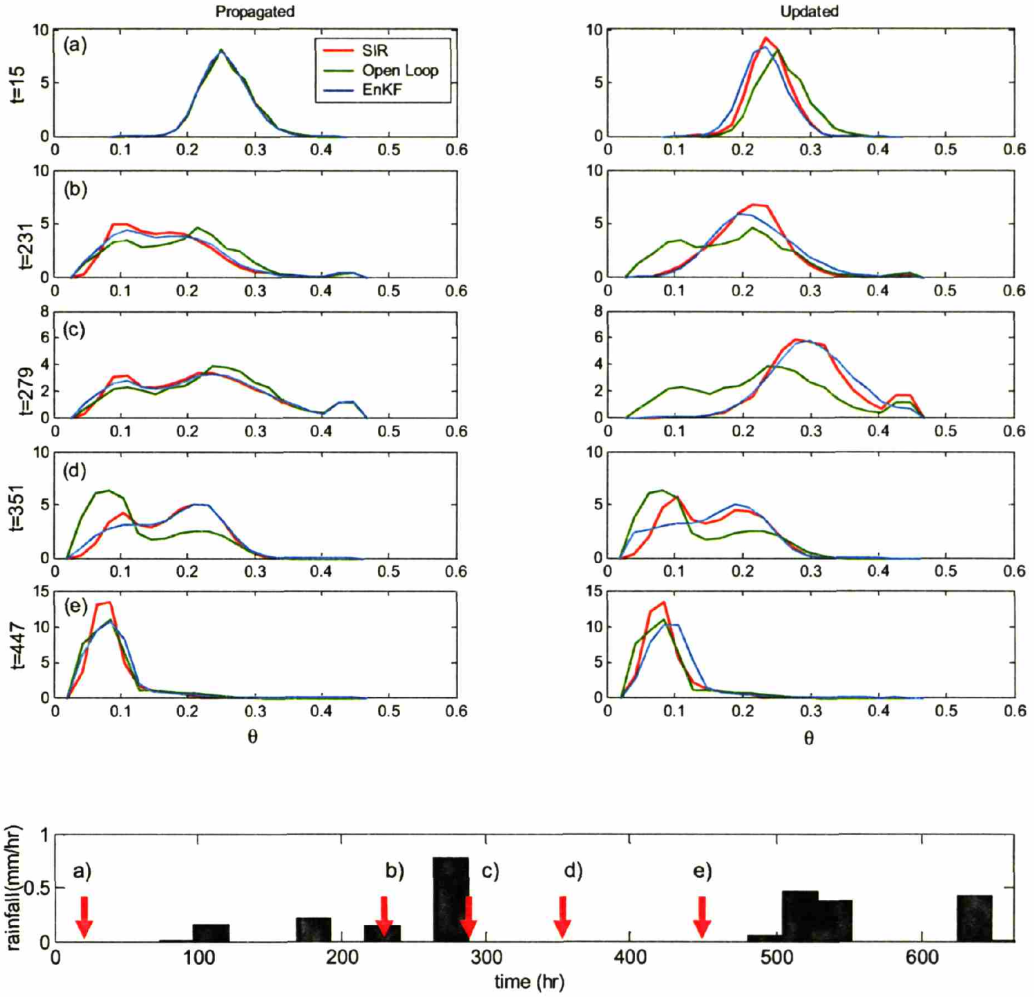


Figure 2-7: Marginal forecast (left) and analysis (right) probability densities for pixel 9 surface soil moisture for some typical analysis times. Open loop (unconditional) densities are also shown for comparison. Bottom panel shows daily rainfall series.

are all skewed to the dry end and the radiobrightness measurement does not provide much additional information about the surface soil moisture.

The marginal densities shown in Figure 2-7 illustrate the advantages of taking a distributional perspective in data assimilation. Ensemble means and even means plus variances do not always tell the whole story. Physical conditions such as prolonged wetting or drying can lead to skewed densities where the means are much different than the most probable values (modes). Although SIR filters provide accurate information on marginal distributions they are not practical for large problems. Fortunately, the ensemble Kalman filter seems able to convey much of this distributional information, despite its simplifying normality assumptions. This is a direct result of the ensemble Kalman filter's ability to update each replicate rather than just the ensemble mean. Individual replicate updating is able to preserve some skewness and multimodality, even when the analysis step is suboptimal.

In order to assess global performance, rather than performance at a single pixel, we examine in Figure 2-8 the time series of the differences between the ensemble mean and the "true" replicates for surface soil moisture, evapotranspiration, and ground temperature, all averaged over the entire domain. The errors are shown for the SIR filter, ensemble Kalman filter and open loop estimates. The abrupt change in soil moisture error due to assimilation of brightness temperature can be observed at analysis times for both filters but the impact of measurements is less clear for the ground temperature. This reflects the fact that brightness temperature is more sensitive to soil moisture than to ground temperature.

Although evapotranspiration is a diagnostic variable rather than an updated state we can see that the SIR and ensemble Kalman estimates of evapotranspiration benefit from radiobrightness measurements. Both of these generally have lower errors than the open loop estimate. A closer look at the plots suggest that the study period can be roughly divided into two stages, before and after at t=250. During the first stage, soil moisture is relatively high, so the evapotranspiration is controlled by available energy rather than soil moisture. Hence, the open loop estimates of evapotranspiration and ground temperature are nearly as good as the filter estimates. During the second stage, there is a long drydown period, evapotranspiration is moisture limited, and open loop errors are larger than filter errors. Assimilation of brightness temperature is clearly more beneficial during this stage.

Surface brightness temperatures can be used to estimate subsurface soil moisture profiles that are difficult to observe at large scales. Figure 2-9 shows the ensemble mean of the integrated soil water depth above 50cm deep over the entire domain. The integrated soil water depth could be viewed as a rough measure of the water available to a plant with a root depth of 50 cm. Here again, the ensemble Kalman filter gives results that are nearly as good as the SIR filter.

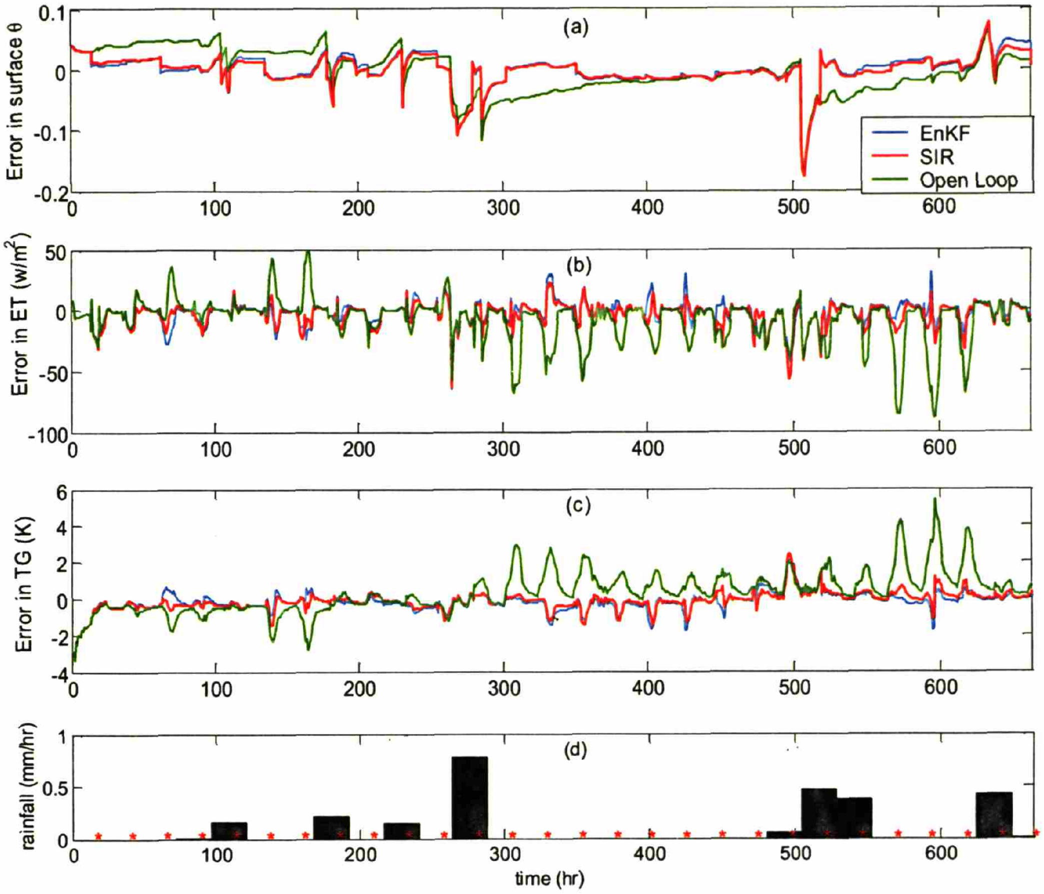


Figure 2-8: Time series of the differences between the ensemble mean and the "true" replicates for (a) surface soil moisture, (b) evapotranspiration, (c) ground temperature, all averaged over the entire domain, and (d) GPCP rainfall time series.

Figure 2-10 provides some indication of the distributional differences between the two filters by comparing time series of the higher order moments (standard deviation, skewness, and kurtosis) of the surface soil moisture for pixel 9. Differences between the higher-order moments produced by the two filters are greater than differences between the means. Both filters are able to capture the significant reduction in variance and increase in skewness experienced during the extended drydown period after $t = 300$. Heavy rainfall events seem to reduce differences between the two filters. It should be noted that the random pulse and multiplicative cascade models tend to generate very non-normal surface soil moisture density functions, as shown by the skewness and kurtosis in panel (b) and (c). The ensemble Kalman filter captures much of this non-normal behavior, at least for our application.

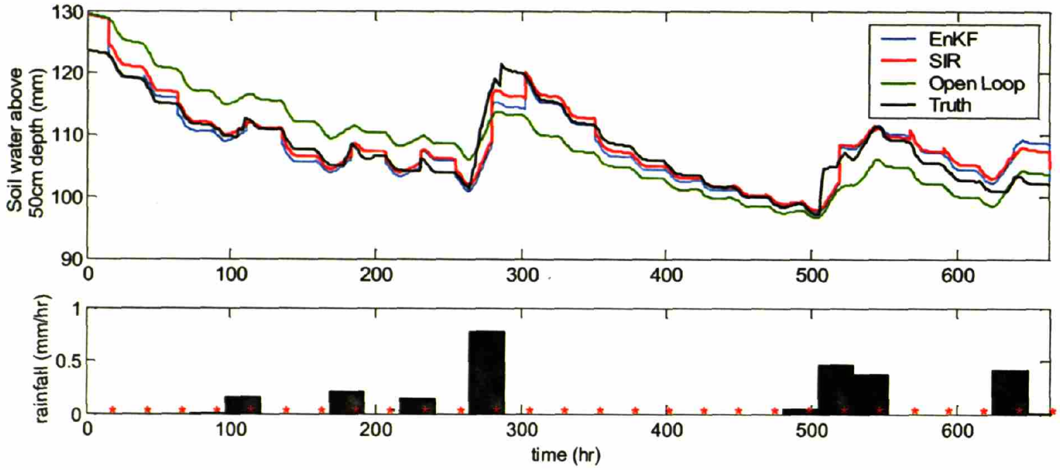


Figure 2-9: The ensemble mean of the integrated soil water depth above 50cm deep over the entire domain

The results of our land surface data assimilation experiment are summarized in Table 2.2, which lists root mean squared error values obtained from SIR filter, ensemble Kalman filter, and open loop means for the four variables of most interest. It is obvious that the SIR and ensemble Kalman filter errors are comparable in all cases. Taken together, our results strongly support the use of the Kalman approximation in land surface applications of ensemble data assimilation.

	SIR	**EnKF**	**Open loop**
Top layer soil moisture	0.026	0.027	0.036
Evapotranspiration (w/m2)	10.3	10.1	19.5
Ground temperature (K)	0.5	0.5	1.1
Water depth above 50cm (mm)	2.2	2.4	4.3

Table 2.2: Root mean squared error (over time) of spatially averaged top layer soil moisture, evapotranspiration, ground temperature, and water depth above 50cm with respect to the true

2.6 Discussion and Conclusions

This paper considers the performance of the ensemble Kalman filter in a particular context: land surface data assimilation. Land surface problems have several distinctive characteristics. In particular, the state equation is nonlinear and dissipative and

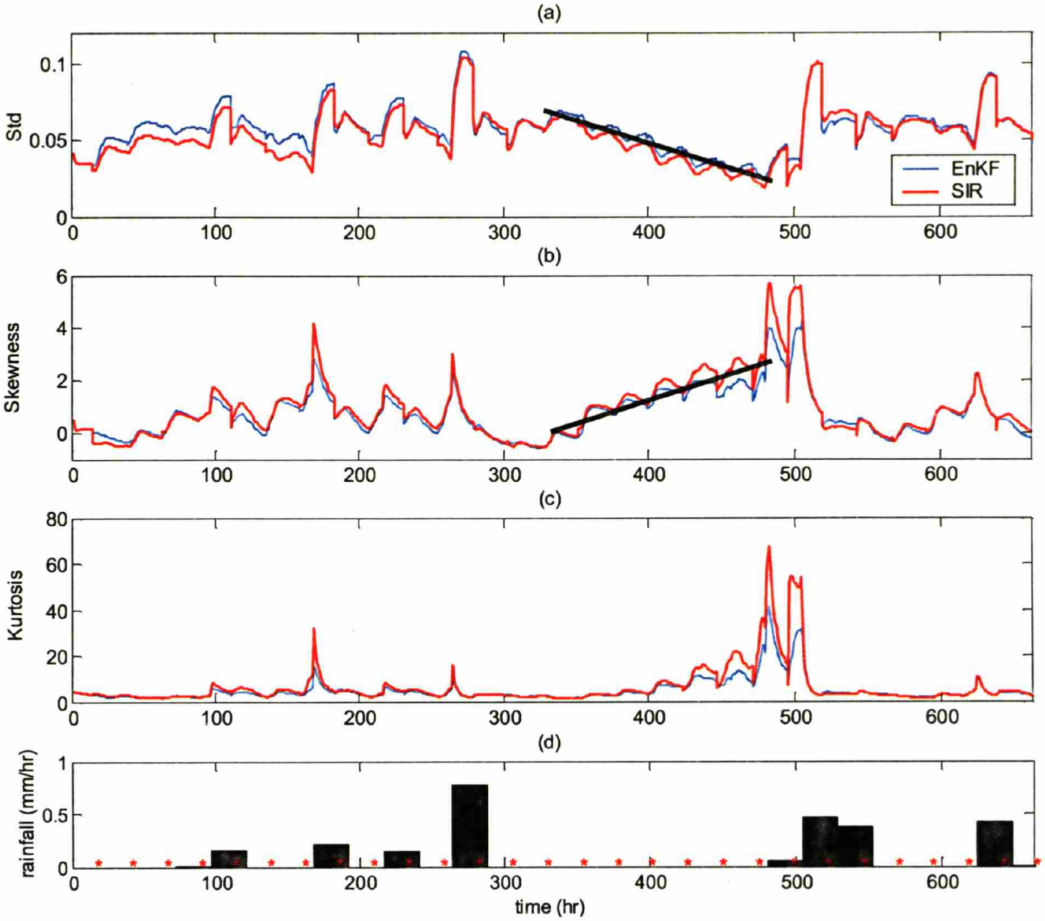


Figure 2-10: Time series of (a) standard deviation, (b) skewness, (c) kurtosis of surface layer soil moisture at pixel 9, and (d) GPCP rainfall time series. The thick straight line in (a) and (b) is to show the trend of standard deviation and skewness during drydown period.

the states are confined to relatively small ranges, with probability distributions that change over time and are often non-normal. Precipitation inputs are intermittent and highly variable over space and time and other inputs, such as soil properties, are uncertain and difficult to observe over large regions. The measurement equation is also nonlinear. Ensemble Kalman filters have been applied to land surface data assimilation with reasonable success, despite their dependence on assumptions that may not apply. Our objective here has been to better understand the reasons for this success and to obtain a more complete picture of the strengths and weaknesses of the ensemble Kalman filtering approach.

The simple example described in Section 3 shows that the SIR filter's conditional moment and marginal density estimates are very close to their exact counterparts if the replicate size is large enough. The ensemble Kalman filter's conditional mean estimate is also quite close to the exact value but its marginal density and variance are noticeably different. This example suggests that a converged SIR filter provides a good basis for evaluating the ensemble Kalman filter when an exact solution is not available.

Sections 4 and 5 describe a more realistic land surface estimation example that relies on state and measurement models used in operational settings. In this Observing System Simulation Experiment we generate hypothetical true states and measurements so that filter estimation errors can be evaluated exactly. The example problem is kept small so that the SIR filter is computationally feasible. The number of replicates needed for this filter to converge becomes very large when the state and measurement dimensions increase much beyond the values used in our example. This is why the ensemble Kalman filter, which converges for much smaller ensemble sizes, is preferable to the SIR filter in practical applications.

The results of our land surface example reveal that the ensemble Kalman filter performs nearly as well as the SIR filter for most conditions simulated. The surface soil moisture forecast densities obtained from the Kalman filter can be quite skewed and even multi-modal and are generally similar to those obtained from the SIR filter. The univariate densities of Figure 2-7 make it clear that the normality assumptions that must be met in order for the ensemble Kalman filter to yield optimal point estimates do not prevent it from generating non-normal ensembles. This is further emphasized in Figure 2-10, which shows that the skewness and kurtosis of the ensemble Kalman filter soil moisture ensembles can differ significantly from those associated with normally distributed variables,

The ensemble Kalman filter is especially good at reproducing the correct soil moisture conditional mean. This appears to be a consistent result at all times and pixels in our experiment and it is observed both at the surface (Figure 2-8) and integrated over the soil column (Figure 2-9). Similar performance is obtained for evapotranspiration, which benefits most from radiobrightness measurements when it is limited by soil moisture.

It is worth noting that the structure and timing of precipitation exert a dominant influence on the land surface system. This influence tends to reduce differences between alternative data assimilation algorithms that make similar assumptions about rainfall. The RPM and multiplicative cascade disaggregation models used here tend to create very non-normal soil moisture during rainy periods. In these periods soil moisture is skewed to the high end. As the surface moisture decreases through infiltration and evaporation, the skewness and kurtosis tend to decrease, making the ensemble filter's normality assumptions more appropriate. However, the skewness and kurtosis tend to increase again when soil dries and soil moisture is limited at the low end.

Soil properties also have a strong influence on the behavior of the land surface system and the performance of alternative filters. Open loop (unconditional) predictions of soil moisture are usually better for rapidly infiltrating sandy soils than for less permeable loam or clay soils. Also, soil moisture updates have less impact on evapotranspiration for sandy soils. In such situations differences between optimal and suboptimal filtering algorithms are less likely to be dramatic.

Even taking these distinctive problem features into account, our overall conclusion is that the ensemble Kalman filter provides surprisingly good performance in the land surface application. This applies both to the filter's ability to characterize non-normal distributional properties and its ability to provide accurate conditional means. We believe these results support previous studies that indicate the ensemble Kalman filter is a good estimation option for land surface applications. It would be useful to see the results of computational experiments similar to ours in other application areas. Such experiments could provide better understanding of when and why the ensemble Kalman filter can deal with nonlinearities and non-normal uncertainties.

Chapter 3

Spectrum Diagnostics of Land Surface System and its Implication for Data Assimilation

In practice, data assimilation generally requires processing of large data sets with computationally demanding models. The computational requirements of environmental data assimilation problems are a direct result of the wide range of time and space scales that need to be accommodated. The high dimensionality of the state and measurement vectors does not necessarily reflect what might be called the intrinsic size (i.e. number of independent variables) of a land surface assimilation problem. This is because the states and measurements might be highly correlated or linearly dependent. For example, soil moisture tends to be highly correlated both vertically and horizontally after a prolonged wet period. In this case the number of independent variables needed to properly characterize spatial variability over a specified grid is smaller than the number of grid cells. On the other extreme, after a long drydown period horizontal correlation tends to decrease, reflecting the effect of small-scale soil variability. Then the number of independent variables is large but the resulting data assimilation problem can be divided into many smaller independent problems, each associated with a single isolated soil column. In either case appropriate reformulation of the problem can yield a significant reduction in computational effort.

The difficulty is that spatial correlation in a land surface problem changes over time, sometimes rapidly. Highly efficient estimation algorithms must be able to adapt to such changes, continually adjusting between the two extremes mentioned above. Traditional model reduction algorithms are not able to do this, especially when the system is nonlinear (Crommelin and Majda, 2004 [21]). Development of more flexible and adaptive algorithms will require a careful look at the space-time structure of the system states.

The space-time structure of soil moisture and related land surface states depends strongly on meteorological forcing variables, particularly precipitation. Precipitation is highly uncertain and intermittent, with a complex space-time correlation structure that varies over a wide range of scales. Significant rainfall events tend to reinitialize the near surface soil system, diminishing the influence of initial conditions. This behavior is not compatible with the simplified additive white noise input models used in traditional estimation methods. In addition, land surface states such as soil moisture and temperature are confined to relatively narrow ranges of values, but variations within these ranges can have a significant impact on land-atmosphere fluxes. Connections between these states and external meteorological inputs are complex and nonlinear. The most obvious example is the two-way relationship between soil moisture and evapotranspiration.

Another distinctive aspect of land surface data assimilation relates to the nature of the measurements used to characterize system states. Traditional direct measurements of land surface states such as soil moisture and of inputs such as precipitation are expensive and sparsely distributed over time and space. Data sources such as airborne and satellite-based remote sensing have more extensive coverage but much coarser resolution. Remote sensing measurements are generally only indirectly related to the states of interest and can be much less reliable than their in situ counterparts. In either case, measurement errors depend on the measurement and estimation scales as well as the intrinsic accuracy of the sensor.

In order to properly characterize the space-time structure of land surface variables we need realistic descriptions of input and measurement uncertainty. This is one reason why there has been much interest recently in ensemble forecasting and data assimilation techniques for land surface applications (Reichle et al. 2002 [84], Margulis et al., 2002 [70],;Crow and Wood, 2003 [23]). Generally speaking, ensemble methods are able to accommodate a much richer set of uncertainty models than more traditional estimation alternatives. However, these methods become more difficult to use as dimensionality of the state vector increases since it becomes increasingly difficult to properly describe the distributional properties of the state with a moderate number of replicates (Silverman, 1986 [91]). This issue is especially problematic when the state is not normally distributed, as is often the case in nonlinear land surface problems.

This chapter uses ensemble-based simulation experiments to gain insight about connections between the space-time structure of soil moisture and computational efficiency. In the next section we provide a probabilistic problem formulation that sets the stage for our analysis. Then we describe the particular models used in the simulation experiments. We examine ensemble forecasts from these models for a particular land surface problem, focusing on the connection between spatial patterns in rainfall, soil properties, and soil moisture. This is followed by an assessment of two land surface data assimilation options, one that considers large-scale spatial correlation but is computationally demanding and another that only considers local correlation but is more computationally efficient. We conclude with a discussion of the need for a more general estimation approach that is able to handle a range of conditions without sacrificing either efficiency or accuracy.

3.1 Ensemble Analysis of Land Surface Estimation Problems

It is helpful to begin with a general problem description, using approximate models of the land surface system, associated sensors, and uncertain inputs. We assume that the land surface model is discretized over both space and time. The spatial discretization uses a computational grid of pixels composed of a number of distinct soil layers. The time discretization uses equal steps over a specified time interval [0, T]. The states of land surface model typically include the bulk soil moisture and average temperature in each layer of each pixel as well as various canopy states. External inputs include precipitation, solar radiation, air temperature, and other meteorological variables. Generally, speaking we have access to input measurements (e.g station data at scattered locations) that give a partial picture of the spatial and temporal variability of the uncertain inputs. We may also have output measurements (e.g microwave radiometer observations) that are related, usually indirectly, to the system states.

Suppose that the land surface states at time t are assembled in the n-dimensional state vector x_t and inputs are assembled in the vector u_t. Diagnostic variables such as evapotranspiration are assumed to be functions of x_t and u_t. Also, suppose that we have a composite m-dimensional vector of output measurements $y_{1:t}$, available at specified times over [0, T] and related to the states through models of the relevant sensors. We assume that the models can be expressed as follows:

$$x_{t+1} = f(x_t, u_t) \tag{3.1}$$

$$y_t = h(x_t) + v_t \tag{3.2}$$

Note that the measurement model of (3.2) includes an additive error v_t(the additivity assumption is made for simplicity and can be relaxed). We assume that the nonlinear functions $f(\cdot)$ and $g(\cdot)$ are known but the inputs u_t are uncertain. In practice, uncertainties in the functions themselves are accounted for (in an approximate way) by inflating input and measurement error uncertainties.

Given the importance of uncertainty it is reasonable to adopt a probabilistic characterization of the land surface system. In particular, the unconditional probability density $p(x_t)$ characterizes the state based only on prior statistical information and input measurements, without the benefit of output measurements. The conditional density $p(x_t|y_{1:t})$ characterizes the state given all measurements through T. Since these multivariate densities are cumbersome to work with we usually restrict our attention to marginal (univariate) densities and lower-order moments. Here we focus on the filtering problem, where $t = T$(i.e. the estimation time is at the end of the measurement interval). However, most of the points made also apply to smoothing problems, where $t < T$ (i.e. the estimation time is inside the measurement interval).

The unconditional density $p(x_t)$ can be derived from the state equation and specified input density and initial condition densities $p(u_1, ..., u_t)$ and $p(x_0)$. Note that there is no requirement here that the inputs at different times (say u_{t-1} and u_t) are independent. Although this independence assumption is frequently made in classical filtering methods it is not required for the ensemble filtering approaches described here. Calculation of $p(x_t)$ from $p(u_1, ..., u_t)$ and $p(x_0)$ is a classical derived distribution problem that can only be solved exactly in certain simple cases. When exact solutions are not feasible we can adopt an ensemble approach and randomly sample these densities to create an ensemble of equally likely initial condition and input replicatesx_0^i and u_t^i, for $i = 1,\ldots,N$. Starting with $t = 1$, we then use the nonlinear state equation (3.1) to compute replicates sequentially, at each time t from the value at the previous time t-1:

$$x_t^i = f(x_{t-1}^i, u_t^i) \quad ; \quad x_{t-1}^i \sim p(x_{t-1}) \tag{3.3}$$

This set of unconditional replicates is often called the "ensemble forecast" at t. Once the replicates are generated the unconditional density may be approximated by a sum of Dirac delta functions evaluated at the replicate values:

$$p(x_t) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(x_t - x_t^i) \tag{3.4}$$

The integral of (3.4) yields a stepwise approximation of the continuous cumulative distribution function x_t. The mean and other moments of $p(x_t)$ can be estimated directly from the ensemble of x_t^i.replicates. In particular, the sample unconditional mean and covariance are given by:

$$\bar{x}_t = \frac{1}{N} X_t 1_N \quad ; \quad Cov[x_t] = \frac{1}{N-1} \tilde{X}_t^T \tilde{X}_t \tag{3.5}$$

where$X_t$and $\tilde{X}_t$ are n by N dimensional matrices whose columns are the original and mean-removed unconditional replicates of x_t , respectively, and 1_N is a column vector of N ones.

Derivation of conditional densities such as $p(x_t|y_{1:t})$ is somewhat more complicated. The extra effort is usually worthwhile, however, since the measurements incorporated into a conditional density enables it to provide a more accurate characterization of the state. In filtering applications we construct $p(x_t|y_{1:t})$ recursively, in a series of alternating propagate (forecast) and update (analysis) steps. That is, the density $p(x_{t-1}|y_{1:t-1})$ of the state at t-1 conditioned on measurements through t-1 is propagated forward to measurement time t, giving $p(x_t|y_{1:t-1})$. This propagated density is updated with the new measurement at$t$ to give $p(x_t|y_{1:t})$ and the process is repeated.

The propagate step is similar to the unconditional derived distribution problem described above, except that $p(x_{t-1}|y_{1:t-1})$ replaces $p(x_{t-1})$. This problem can also be solved with an ensemble approach, using the same basic expressions as the unconditional ensemble forecast but with slightly different notation:

$$x^i_{t|t-1} = f(x^i_{t-1|t-1}, u^i_t) \quad ; \quad x^i_{t-1|t-1} \sim p(x_{t-1}|y_{1:t-1}) \tag{3.6}$$

$$p(x_t|y_{1:t-1} :) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(x_t - x^i_{t|t-1}) \tag{3.7}$$

The mean and other moments of $p(x_{t-1}|y_{1:t-1})$ can be estimated directly from the ensemble of $x^i_{t|t-1}$replicates

The update step of the conditional density derivation is based on the following version of Bayes theorem, which relates $p(x_t|\ y_{1:t})$ to $p(x_t|y_{1:t-1})$:

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} = cp(y_t|x_t)p(x_t|y_{1:t-1}) \tag{3.8}$$

where c is a normalization constant selected to insure that $p(x_t|\, y_{1:t})$ integrates to one and the likelihood function $p(y_t|\, x_t)$ is derived from the measurement equations and the measurement error probability densities. A Bayesian update that satisfies (3.8) exactly is difficult to implement for large problems, except for some important special cases. In particular, if the states and measurement errors are jointly normal and the measurement operator is linear $p(x_t|y_{1:t})$ is normal and completely characterized by its mean and covariance. In this case, (3.8) can be carried out by updating only the mean and covariance of the state. In an ensemble approach we update probability densities or moments by updating replicates and then computing sample statistics. That is, we begin by deriving $x^i_{t|t}$ from $x^i_{t|t-1}$. In the special jointly normal case mentioned above, we can obtain asymptotically exact conditional density and moments by updating each replicate with the ensemble Kalman filter described in section 1.5.1.

So we see that ensemble methods may be used to approximate the unconditional and the conditional statistics of uncertain states. In principle this solves the ensemble forecasting and data assimilation problems. But the computational requirements of ensemble methods are considerable and the accuracy of the approximations needed to obtain practical solutions is difficult to determine. It is convenient to investigate these with a controlled simulation experiment.

3.2 Models Used for the Land Surface Ensemble Experiments

In this section we describe the models used in our ensemble forecasting and data assimilation experiments. Our objective is to characterize soil moisture and evapotranspiration on hourly time scales over a region of approximately 10,000 km^2. Land surface dynamics are described by the Community Land Model (CLM, version 2.0) (Bonan, 1996, 2002 [5]). Radiometer measurements are described by a nonlinear radiative transfer model (Njoku et al., 2002 [77]). These provide indirect information on near surface soil moisture. Input uncertainties and measurement errors are described by statistical models that are intended to provide realistic representations of natural variability. These models determine how the replicates of the simulation experiment are generated. The input statistics are inferred from available meteorological and soil measurements.

3.2.1 The Land Surface and Measurement Models

The CLM is a nonlinear spatially distributed model that describes energy, momentum, water, and CO2 exchange between the land and the atmosphere. Dynamic inputs include precipitation, wind speed, air temperature, pressure, humidity, and solar radiation. Time-invariant inputs include soil and vegetation classifications. The model is discretized into square pixels that are each divided into several soil layers. Moisture and heat can only move vertically within individual pixels. Further details are discussed in Bonan (1996, 2002 [5]). Although moisture does not flow between pixels the states in different pixels are correlated by virtue of their dependence on spatially correlated inputs such as precipitation and vegetation.

The study region for our computational experiment reflects conditions at the Southern Great Plains (SGP97) site in eastern Oklahoma. This 10,000 km^2 (100×100km) region is defined by the corners (36 ° N, 99 ° W) and (37 ° N, 98 ° W), as shown in Figure 3-1a. It is discretized over a 64 by 64 grid of 1.56 km by 1.56 km estimation pixels with 8 soil layers in each pixel. The regions associated with each land use and soil type are indicated in Figure 3-1a. The soil layers have thicknesses (from top to bottom) of 2, 3, 5, 8, 12, 20, 57, 88 cm respectively. The CLM model states considered in our ensemble analysis are the soil moisture values at the centers of the top three layers, giving a total of 12,288 states in our 4096 pixel grid. The CLM derives evapotranspiration from these states. The study period corresponds to a 23 day field campaign conducted from June 16, 1997at 0 hrs. UTC through July 8, 1997 at 15 hrs. UTC (Margulis, 2002 [70]). Input data are generated and the CLM is run for a 1 hr. time step. Meteorological measurements are available at El Reno, in the SGP97 study area.

Synthetic radiobrightness measurements are calculated in the same way as that

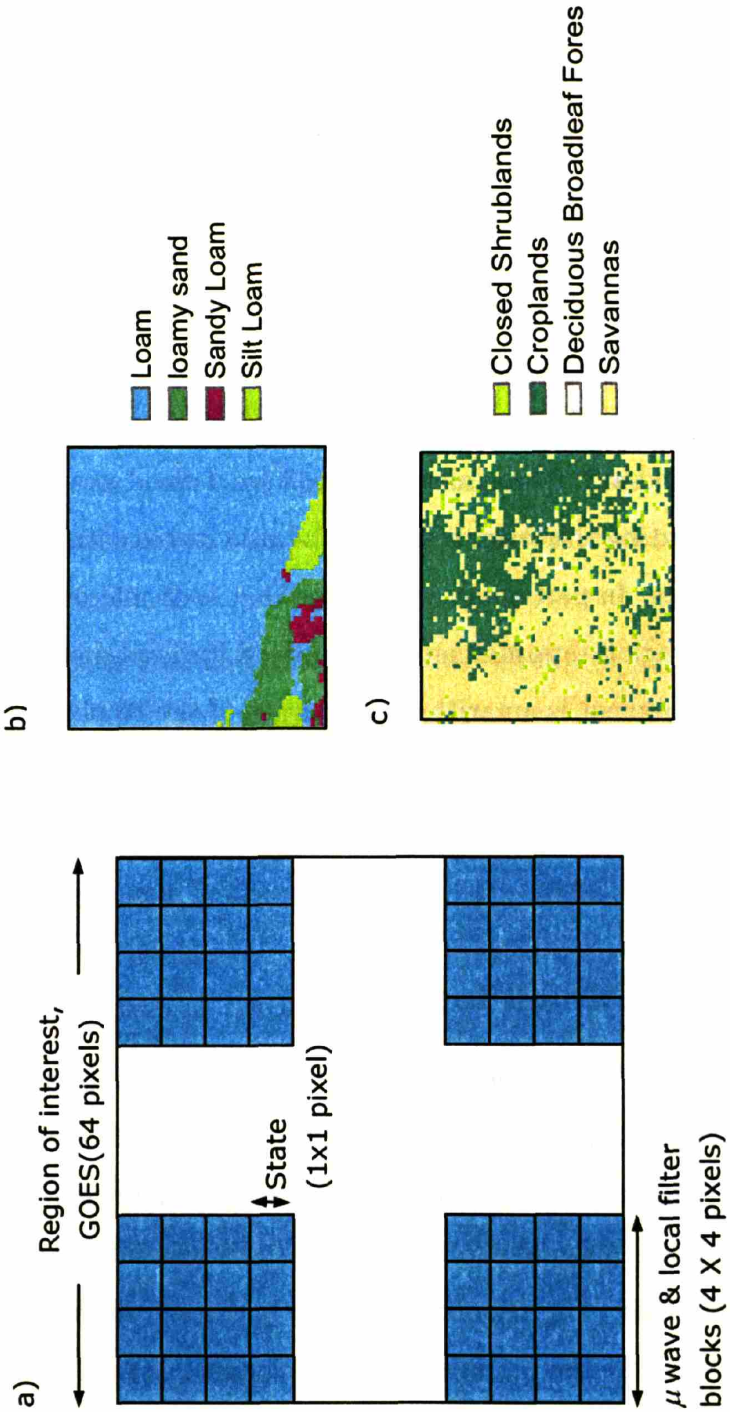


Figure 3-1: Inputs for the example simulation problem: a) 64x64 study domain, showing estimation and measurement (wave) pixels, b) soil type, c) vegetation classification.

described in section 2.31. In this experiment, w, h, and b have the values 0.3kg/m^2, 0.1, and 0.04 respectively. The view angle θ is set to 20 ° and the scattering albedo is 0.03.

In our ensemble forecasting experiment "truth" is defined by the state from a single CLM run obtained for a particular set of soil, vegetation, meteorological, and initial condition replicates, as described above. In this case, it's equivalent to assuming the error statistics of the system is perfectly known. The CLM states and associated soil properties for this "truth" replicate are then used in (2.22) to generate synthetic L band brightness temperature measurements at 15 hrs UTC at specified days during the 23 day simulation period. These measurements are defined at a coarser scale than the model states, reflecting the lower resolution of anticipated satellite microwave radiometer measurements. In particular, we assume that each microwave measurement covers a 4 by 4 pixel region (approximately 6 km by 6 km), as shown in Figure 3-1a. The microwave measurement is an arithmetic average of the 16 pixel-scale brightness values in this region. At each measurement time a zero-mean normally distributed random perturbation is added to the averaged brightness temperature to account for the effect of measurement noise.

3.2.2 Uncertain Model Inputs

The primary sources of input uncertainty in land surface applications are time-invariant soil properties, time-dependent meteorological inputs, including precipitation, and initial conditions. In the ensemble approach random replicates for each of the uncertain inputs are provided to the CLM, which generates random replicates of the land surface states. Corresponding radiobrightness values at the estimation pixel scale are generated by the radiative transfer model of (24). The time-dependent random inputs can cause the ensemble to spread during the propagation step while assimilation of radiobrightness measurements can cause the ensemble to narrow at the analysis step. These effects are moderated by the physics of the problem, which constrains the states to lie in limited ranges (e.g. the volumetric soil moisture must lie between 0.0 and the porosity, which is less than 1.0).

The uncertain inputs are generated by transforming nominal input values to obtain sets of physically realistic replicates. This is done is various ways, depending on the variable. Table 3.1 lists the uncertain inputs and measurement errors considered in our simulation experiment. Note that different methods are used to introduce randomness for different inputs. The soil, vegetation, and precipitation inputs deserve some elaboration.

In the CLM sand and clay fractions are used to calculate hydraulic parameters for Richard's equation, which controls the vertical movement of moisture. Replicates of these fractions are generated from the soil map shown in Figure 3-1b. Each soil type is assigned nominal sand and clay fractions based on the soil triangle. Deviations from these nominal values are normally distributed and uncorrelated. Details are described in Table 3.1.

The vegetation types used in our simulation are shown in Figure 3-1c. CLM uses these associates each type with particular values of the leaf area index (LAI) and the stem area index (SAI). It uses these indices to compute various model vegetation parameters that control net radiation, energy partitioning, and intercepted water capacity. In our experiment LAI is assumed to be a spatially uncorrelated random variable and SAI is deterministic. Nominal values for LAI and SAI and distributional properties for LAI are given in Table 3.1.

There are many options for generating rainfall replicates for land surface applications. These range from simple statistical models to complex primitive equation atmospheric models. Some of the basic requirements of a realistic rainfall model include the ability to reproduce intermittency (extensive periods or regions with no rainfall) and the ability to reproduce observed low-order statistics (such as means and covariances over time and space). The rainfall model used here is based on a Poisson cluster concept. It assumes that rainfall can be viewed as a superposition of circular storm cells generated at random times and locations (Rodriguez-Iturbe et al., 1987 [87]).

Chatdarong et al (unpublished manuscript) put the basic Poisson model of Rodriguez et al. [87] in a recursive form that is convenient for continuous time ensemble

Variable	Specified Nominal Value	Uncertainties in Replicates
Soil fractions (sand-silt-clay)	Soil type obtained from Figure 2. Fractions: Loam: 40/50/10 Loamy sand: 80/15/5 Sandy loam: 60/30/10 Silty loam: 22/61/17	Spatially uncorrelated additive zero-mean normal deviations from nominal. Standard deviations: Clay: 10% Sand: 20% 3 fractions constrained to sum to 1.0

Vegetation	Cropland with:
	LAI=1.6, SAI=0.4 (June)
	LAI=1.3, SAI=0.8 (July)
		Spatially uncorrelated multiplicative uniform noise U[0.85, 1.15] for LAI.

Humidity, solar radiation, wind speed & direction	Nominals from Oklahoma Mesonet time series at El Reno, assumed to apply over entire study region. Nominal wind direction constant at 15 ° from east.	Spatially and temporally uncorrelated multiplicative uniform noise: Relative humidity: U[0.9, 1.1] Solar radiation: U[0.9, 1.1] Wind speed: U[0.7, 1.3] Wind direction: U[0.7, 1.3]
Air temperature	Nominals from Oklahoma Mesonet time series at El Reno, assumed to apply over entire study region.	Spatially and temporally uncorrelated additive uniform noise U[-4 °K, +4 °K]
Precipitation	Poisson rainfall model parameters: β = 0.0024 cell km^{-2}hr^{-1}, D= 6.25 km, α = 0.2 hr^{-1}. Wind direction fixed, wind speed (m sec^{-1}), E[i_0] (mm hr^{-1}psgnd cloud cover obtained from meteorological time series at El Reno..	
Initial volumetric water content (at -23 days), by layer	[0.35 0.35 0.35 0.35 0.30 0.30 0.30 0.30]	Spatially uncorrelated additive Gaussian noise N(0.0, 0.06).
Initial soil temperature °K (at -23 days), by layer	[304 302 293 290 290 288 288 288]	Spatially uncorrelated additive Gaussian noise N(0.0, 4 °K).
Radiobrightness measurement	Simulated "true" value at 6 km by 6 km scale	Temporally uncorrelated additive Gaussian noise N(0.0, 4 °K).

Table 3.1: Summary of uncertain inputs and measurement errors for the land surface simulation experiment

analyses: This model can be written as:

$$r(t+1) = F[r(t)] + w(t) \tag{3.9}$$

where $r(t)$ is a vector of pixel rainfall intensities at time t; $w(t)$ is a vector of forcing terms that account for the birth of new raincells, and $F(\cdot)$ describes the effects of rain cell advection and temporal decay. Equation (3.9) can be viewed as a state equation for the rainfall intensity $r(t)$ but we treat this variable as an input and do not update it in the data assimilation process.

The rain cell birth process is best understood by first considering a Lagrangian (cell-based) description, as illustrated in Figure 3-2a. Cells are only allowed to be born in regions with cloud cover. Global cloud cover can be determined from Geostationary Operational Environmental Satellite (GOES) infrared sensors. For our experiment we inferred cloud cover from the rainfall time series at El Reno, as shown in Figure 3-2b. The center of rain cell k follows a two-dimensional spatial Poisson distribution with spatial density parameter β. The birth time of each cell is uniformly distributed over the time step $[t, t+1]$. The cell rainfall intensity $i_k(d_k,\ t)$ at a distance d_k from the cell center at time t follows a Gaussian distribution in space and decays exponentially in time at a rate α:

$$i_k(d_k, t) = i_{k0} \exp[-\alpha(t - t_{bk})] \exp[-d_k^2/2\pi D^2] \tag{3.10}$$

where t_{bk} is the cell birth time. The cell center intensity i_{k0} is exponentially distributed with mean $E[i_{k0}]$ and D is a specified constant Values for α, β, and D are given in Table 3.1. These are representative of the SGP97 site. The generated rain cells are rigidly transported at a specified wind velocity $(v_x,\ v_y)$. In our simulation experiment the wind direction is held fixed to emphasize the effect of raincell advection on the spatial correlation of soil moisture. The nominal wind speed is derived from meteorological data at El Reno, with multiplicative perturbations drawn from a uniform distribution, as indicated in Table 3.1.

The Eulerian rainfall intensity field r_i at pixel i is the superposition of the La-

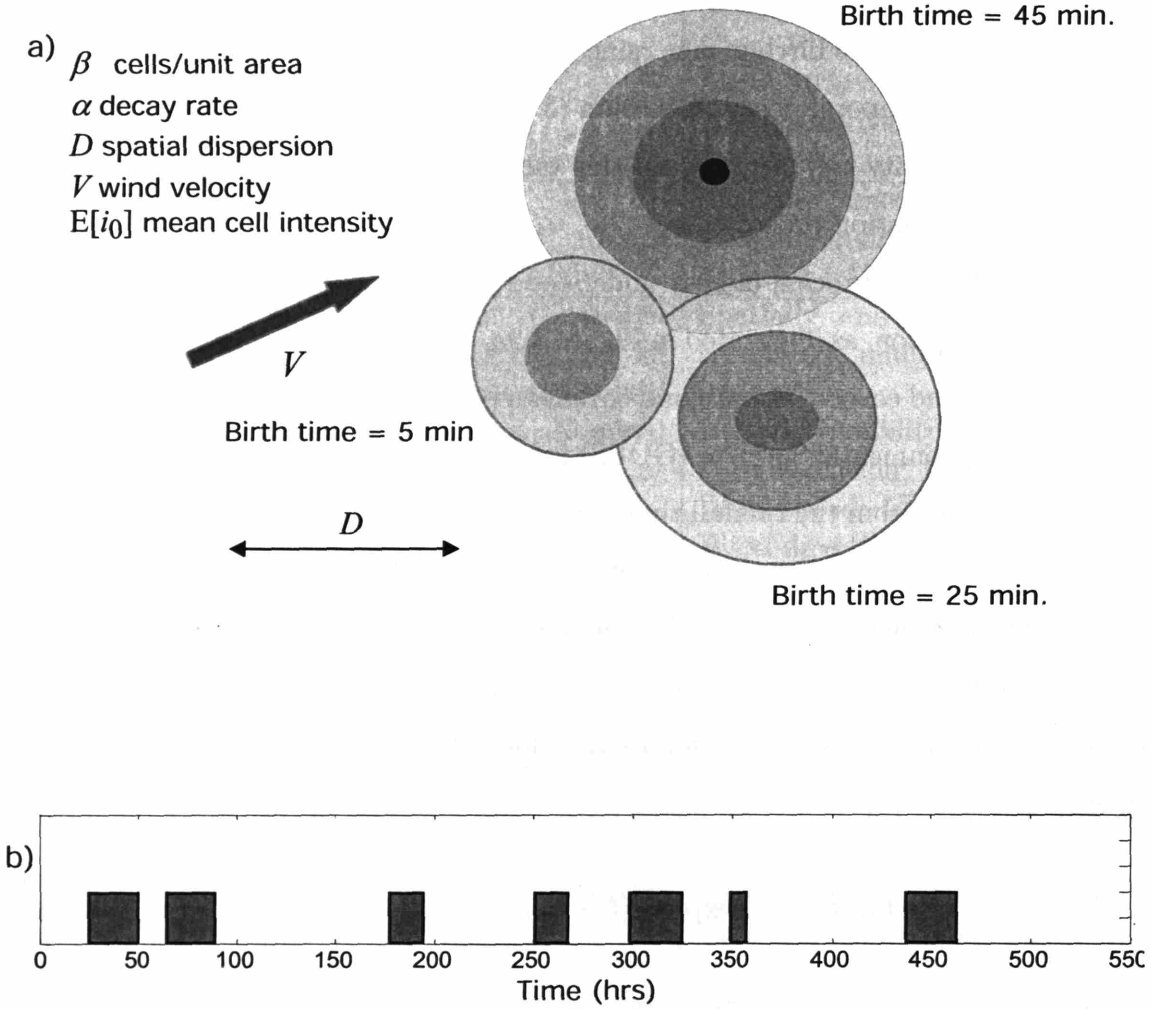


Figure 3-2: a) Typical rain cells generated by the Poisson rainfall model during a one hour period (each cell delineated by a 1 mm hr-1 outer contour). Rainfall intensity at any location is the sum of intensities contributed by the rain cells. Cell centers and birth times are random. Cells have a fixed characteristic distance and decay in intensity over time. Note that older cells are smaller and less intense. b) Cloudy periods during the simulation experiment.

grangian intensities produced by all cells in the domain, with the d_k values set equal to the distances between the center of pixel i and the center of cell k. The most recently born cells closest to the pixel have the greatest effect. A typical replicate generated from this Poisson rainfall model in a 64x64 cloudy domain is shown in the intensity contour plot of Figure 3-3a. The rainfall time series at a typical pixel is shown in Figure 3-3b. Note the spatial correlation revealed in the contour plot and the intermittency that occurs in both space and time. Although the Poisson model is simplified and cannot be expected to realistically represent all types of rainfall it does give a reasonable statistical description of convective storms such as those commonly encountered in Oklahoma during the summer (Rodriguez-Iturbe et al., 1987 [87]). It also illustrates some of the flexibility that is possible when generating replicates for ensemble land surface analyses.

The CLM is started at -23 days with random initial conditions generated by perturbing the soil moisture and temperature profiles listed in Table 3.1. Each replicate is run forward with the model for 23 days to t = 0 to allow moisture and temperature in individual pixels to redistribute in accordance with local soil properties. The resulting soil moisture and temperature replicates initialize the ensemble simulations.

3.3 The Ensemble Forecasting Experiment

Our ensemble forecasting experiment is intended to provide better understanding about the space-time structure of soil moisture under certain conditions. In particular, we consider the response of a region similar to SGP97 during a summer period when several storms move in a fixed direction across the study domain. This simulation enables us to investigate soil moisture at a level of detail and in a more controlled way than would ever be possible in a field experiment. Of course, our conclusions are limited by our models, which are approximate. But the ensemble analysis presented here is intended primarily to reveal qualitative effects that should be valid if the models provide at least a rough picture of reality.

One of the easiest ways to visualize land surface uncertainty is to examine a

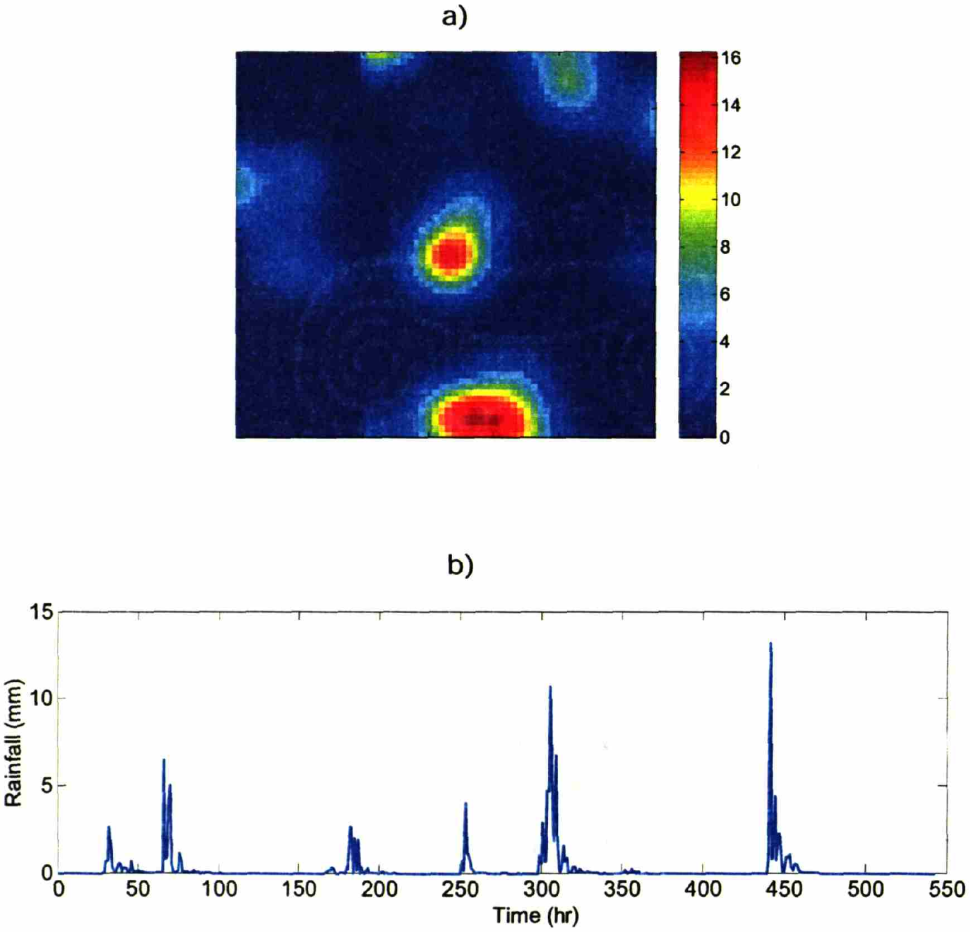


Figure 3-3: Typical rainfall replicate a) Rainfall intensity plot at a given time b) Rainfall time series at a typical pixel

time series of typical soil moisture replicates. Figure 3-4a shows the soil moisture replicate we have designated as truth (black soild line), the replicates of the ensemble (gray lines), and the ensemble mean (black dashed line) of the top layer volumetric soil moisture for a typical pixel. Figure 3-4b shows the corresponding ensemble of rainfall inputs (since these are clustered in cloudy periods it is difficult to distinguish individual replicates). The spread in the ensemble reflects uncertainties from soil properties, vegetation, rainfall, and other atmospheric forcing. During wet periods after rainfall, the soil moisture ensemble is wider, reflecting the strong impact of rainfall uncertainties in both timing and intensity. As the soil dries the replicates become almost parallel. This suggests that the variability in soil moisture at the beginning of the dry period is the dominant influence on ensemble spread during drydown.

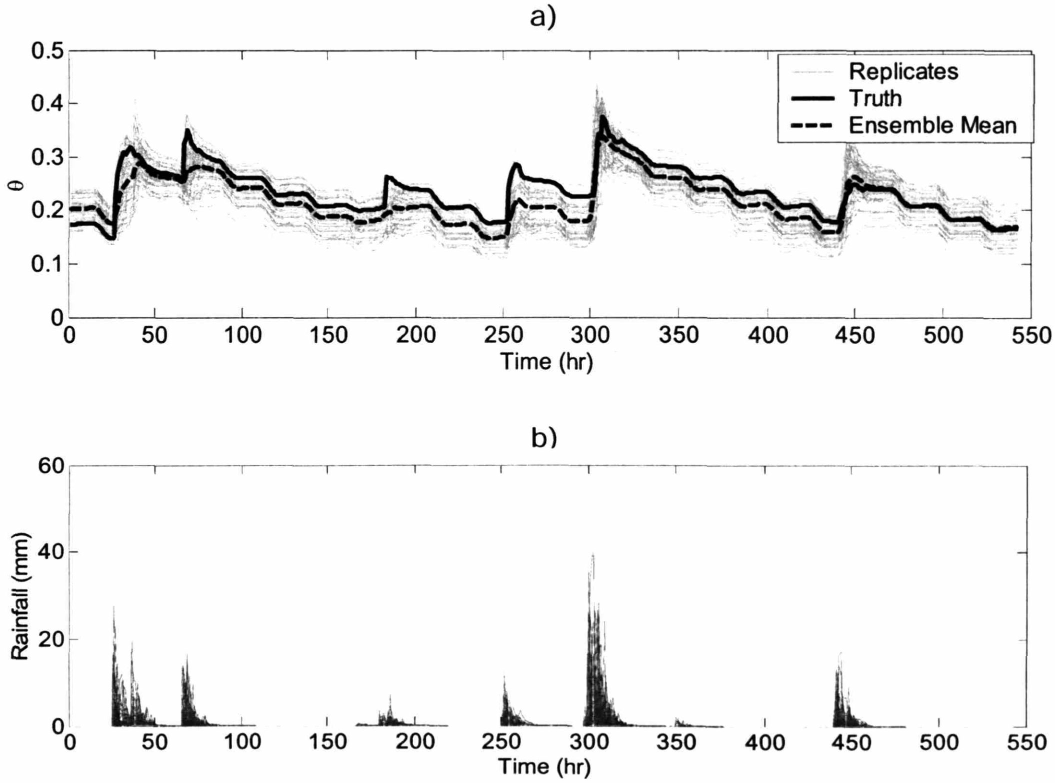


Figure 3-4: Ensemble time series at a typical pixel (a) The gray lines are individual replicates of top layer soil moisture from the ensemble forecast. The thick black line is the corresponding true top layer soil moisture. The dashed line is the ensemble mean (b) Plot of the rainfall ensemble at the same pixel.

Figure 3-5 provides useful insight on the connection between near surface soil moisture and rainfall. Here we compare the cumulative rainfall through time t = 90 to the top layer soil moisture at that time. The diagonal patterns in both plots clearly reveal the effect of advecting rainfall, which creates bands of wetter soil parallel to the fixed wind direction. However, there is also significant smaller-scale variability due to heterogeneity in soil properties (for example, areas with sandier soil are drier than those with finer soils). Soil moisture provides a cumulative record of recent rainfall, to varying degrees depending on soil type.

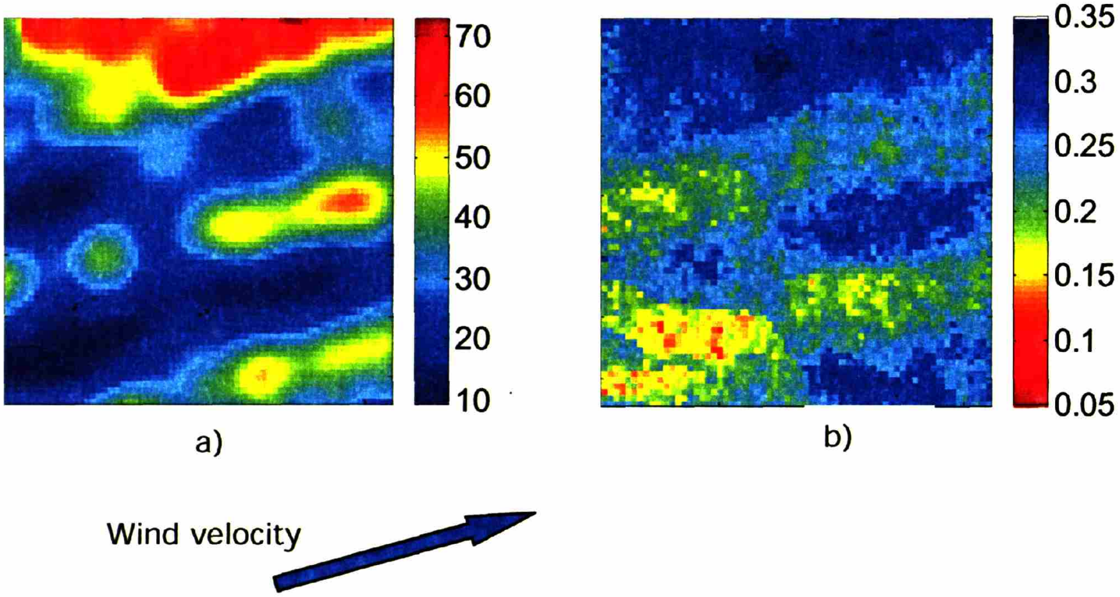


Figure 3-5: (a) True cumulative rainfall (mm) over the domain at t = 90. (b) Corresponding soil moisture.

The horizontal spatial structure of soil moisture changes significantly over time, as revealed by the sample spatial correlation plots shown in Figure 3-6. These plots display contours of the correlation between the top layer soil moisture in the center of the domain with all other soil moisture values. Correlation values are derived from the unconditional ensemble covariance matrix of (3.5). The correlation in all of the plots falls off from 1.0 in the center to near zero at the edges. Figure 3-6a, which applies at during the rainy period at $t = 320$, confirms the anisotropic diagonal pattern of soil moisture observed in Figure 3-5. By contrast, Figure 3-6b indicates that the soil moisture correlation is more localized and isotropic at $t = 430$, after a prolonged drydown. Figure 3-6c confirms that the rainfall correlation is isotropic. The correlation scale of rainfall is roughly 20km. The anisotropy in the wet soil moisture correlation plot reflects the soil's tendency to retain moisture as precipitation moves across the domain. As drydown takes over this memory effect disappears.

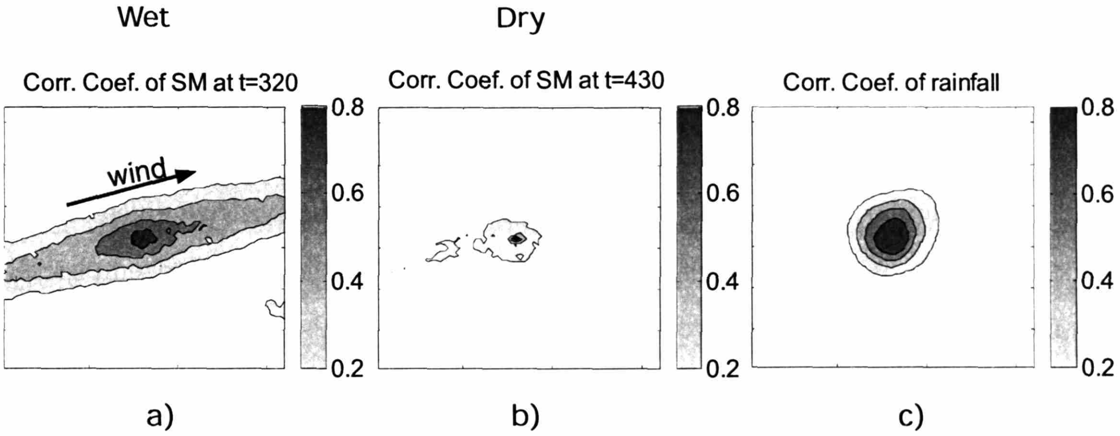


Figure 3-6: The spatial correlation coefficient between the pixel (33, 33) at the center of the domain and all other pixels. a) Top layer soil moisture at a wet time t=320; b) top layer soil moisture at a dry time t=430; c) rainfall correlation at a typical time.

The spatial structure evident in Figures 3-5 and 3-6 raises the question of whether we actually need 12,288 pixel-based moisture values to properly describe the soil moisture field. It is possible to examine this question further if we consider a singular value decomposition of the mean-removed ensemble matrix$\tilde{X}_t$. This decomposition provides a systematic way to develop a set of progressively accurate descriptions of an uncertain state vector. In particular, the state vector at time t may be expanded as follows:

$$x_t = \bar{x}_t + \sum_{j=1}^{n} \gamma_j u_{tj} \tag{3.11}$$

Where $\bar{x}_t$is the unconditional ensemble mean state vector, u_{tj} is the jth n-dimensional left singular vector of $\tilde{X}_t$and γ_{tj} is a random scalar. The u_{tj}'s form an orthonormal basis for the state space and the γ_{tj} are uncorrelated zero-mean random variables. The magnitude of the j^{th}singular value is the standard deviation ofgλ_{tj} . If the singular values are placed in order of decreasing magnitude $[\lambda_{1t}, \lambda_{2t}, ..., \lambda_{nt}]$ then the first term of the basis expansion has the largest variance and the remaining terms have progressively smaller variances. So, if the leading singular value is much larger than the others (3.11) may be approximated by:

$$x_t \approx \bar{x}_t + \gamma_{t1}\, u_{t1} \tag{3.12}$$

In this case the ensemble mean accounts for the overall trend and the singular vector term accounts for deviations around this mean Note that the left singular vectors of $\tilde{X}_t$ are equal to the eigen vectors of the sample covariance matrix C_{xx} defined in (3.5) and the singular values of $\tilde{X}_t$ are the square roots of the positive eigenvalues of this covariance. Also, the eigen/singular vectors are frequently called empirical orthogonal functions or principal components, depending on the field.

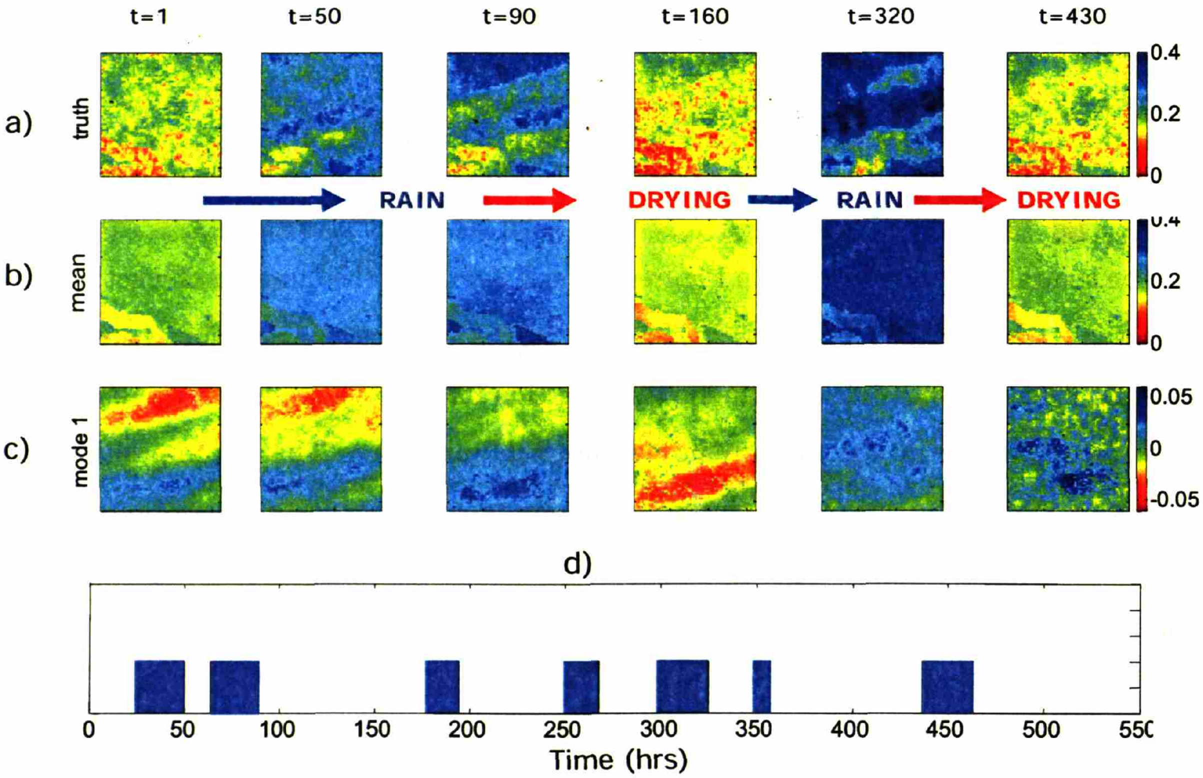


Figure 3-7: Spatial structure of unconditional top layer soil moisture field at 6 typical times. a): True soil moisture, b) Unconditional mean of soil moisture ; c) First singular vector of the unconditional ensemble, d) Cloudy periods shown for reference purposes. Truth is approximated by mean plus a random scaling of first singular vector

Figure 3-7 shows the true, ensemble mean, and first ensemble singular vectors of the unconditional top layer soil moisture, plotted for several times by pixel for over the domain of our simulation experiment. The ensemble mean reveals whether the soil as a whole is wet or dry but does not capture smaller scale variations, including the dominant diagonal structures mentioned earlier. At the first four times the first singular vector adds to the mean in the top half of the region and subtracts from it in the bottom half. This captures some of the anisotropic diagonal character of the soil moisture field. The last time is after a long drydown period when the effects of earlier rainstorms have dissipated. Consequently, the first singular vector does not convey much spatial structure.

The relative importance of the different singular vectors can be determined if the ordered singular values are all divided by λ_{t1} and plotted vs j. Figure 3-8 indicates that this singular value spectrum drops off more quickly at wet times, reflecting the fast that soil moisture structures can be better characterized by a few singular modes shortly after rainfall, when there are larger structures. After a long drydown at t=430 variations in soil moisture are smaller scale and the spectrum is flatter. Figure 3-8 indicates that about 50% of the total variability in soil moisture (as measured by the area under the spectrum) can be explained by the first 15 singular modes during wet periods. More modes are needed to achieve this explanatory level during dry periods. A special case here is at t=160, even though the true soil moisture field is at the same dry level as at t=430, the first singular mode is quite different. The reason is that for t=160, the soil moisture is not dry enough to eliminate the rainfall structure. While at t=430, there is a heavy rainfall at t=320, which is pretty uniform in space and thus can removes any streak rainfall patterns at t=430, leaving soil properties as the dominant factor. It's expected that after a long enough drydown period, all the rainfall pattern would disappear and only soil pattern dominates.

Although the CLM only allows the soil water to moves vertically within each pixel significantly horizontal correlation exists for some time after a rainfall event, especially when the storm moves, as in our simulation experiment. This correlation is not due to lateral subsurface water movement but is, rather, the consequence of

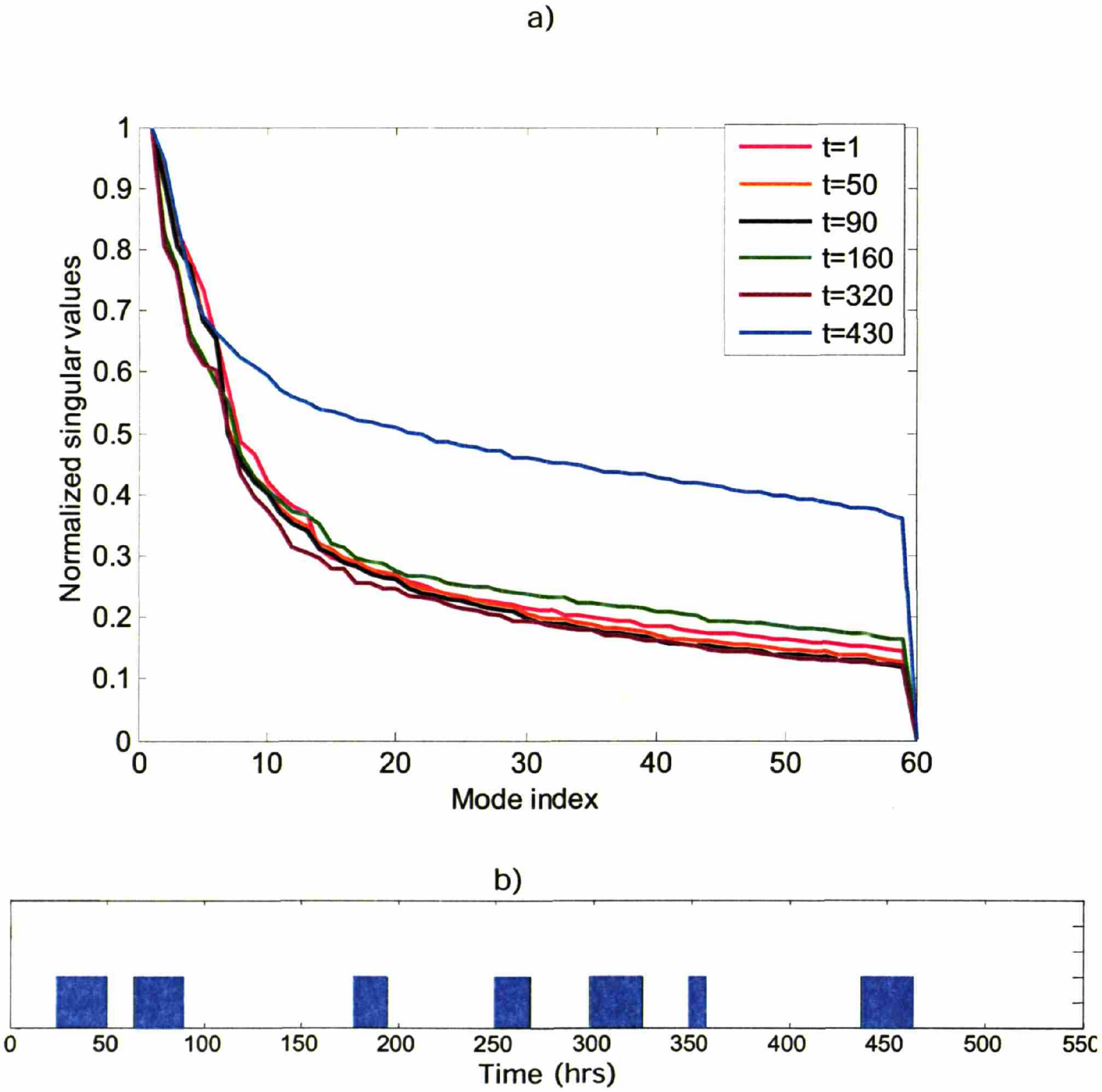


Figure 3-8: a) Normalized singular values of the mean-removed forecast ensemble (normalized by. the largest non-normalized singular value). b) Cloudy periods shown for reference purposes.

an extensive moving storm system that wets some areas and leaves others relatively dry. In answer to the question posed earlier, it appears that the essential structure of the soil moisture field during wet periods can be captured by the unconditional mean and a relatively small number of random singular vector coefficients. During dry periods the potential for problem size reduction is less. It remains to be seen how these features can be exploited in a land surface data assimilation system. We consider this issue further in the next section.

3.4 Ensemble Data Assimilation Experiments

In data assimilation applications we are frequently forced to make compromises between accuracy and computational efficiency. Here we examine this tradeoff in a simulated data assimilation experiment that is an extension of the ensemble forecasting experiment discussed above. In the forecasting experiment we considered the unconditional soil moisture ensemble produced by a particular set of uncertain inputs. Now we consider conditional soil moisture estimates derived from the synthetic L-band radiometer measurements described at the end of Section 3a. Our data assimilation algorithm uses the ensemble Kalman filtering approach described in Section 2.

The ensemble Kalman filter's use of a finite population of replicates introduces two sources of error: 1) sampling error and 2) rank deficiency. Sampling error enters through the mean-removed ensemble matrices $\tilde{X}_{t|t-1}$ and $\tilde{Y}_{t|t}$, which can be viewed as noisy sample estimates of the square roots of the conditional covariances $Cov[x_{t|t-1}]$ and $Cov[h(x_{t|t-1})]]$. Errors in the ensemble-based square root estimates influence the Kalman gain and state updates through (1.28) and (1.30). As the number of replicates increases the sampling error decreases and the filter's estimates converge to the values that would be obtained with exact covariances. In land surface applications ensemble sizes of hundreds appear to be satisfactory, especially if long range correlations are small (refs ....).

Rank deficiency is a source of error only when the number of replicates is less than the state dimension, $N < n$, which is usually the case in large problems. When $N < n$the matrix $\tilde{X}_{t|t-1}$ has rank N and only its first N singular values are non-zero. This has the effect of projecting the n dimensional state space onto the smaller N dimensional subspace spanned by the replicates in$\tilde{X}_{t|t-1}$. This subspace, which is random and constantly changing, is the same as the subspace spanned by the first N singular vectors. When $N < n$ the ensemble Kalman filter is a type of reduced rank filter. Its' estimates are based on a low dimensional approximation and will generally be less accurate than those of a full rank filter.

The analysis of the previous section suggests that the effects of rank deficiency in land surface problems depend on the state of the soil. When the soil is wet and the singular value spectrum falls off sharply a relatively small number of singular modes (or replicates) can provide an adequate description of spatial variability, so long as sampling errors are acceptable. On the other hand, when the soil is dry and the spectrum is flatter more singular modes (or replicates) are needed.

This suggests two different ways to apply ensemble Kalman filtering to land surface data assimilation (Reichle and Koster, 2004 [82]). In the first of these horizontal correlation is accounted for and the entire n dimensional state vector is estimated but the number of replicates is kept much smaller than n. This global filter can be expected to work best after rain, when the surface soil moisture is correlated across many pixels and the singular value spectrum falls off rapidly. It replaces the state covariance matrix with an approximation that is reduced rank but not sparse.

An alternative (Margulis et al., 2002 [70]) is to divide the region of interest into small blocks of pixels that are treated independently (see Figure 3-1a). That is, correlations within each block are considered but those between blocks are ignored (in the limit each block could consist of a single pixel). The data assimilation problem is then partitioned into n/n_l independent sub-problems, where $n_l$is the number of states associated with each block. Each of the small sub-problems is solved with a separate n_l dimensional filter. These independent filters will each be full rank if $N \quad \geq \quad n_l$, which will generally hold if the block are reasonably small. Assuming this is the case, the localized approach can be expected to work best during dry periods, when the singular value spectrum is relatively flat but there is minimal correlation between soil moisture in different blocks. The local filter replaces the complete state covariance matrix with a sparse blocked approximation.

A variant on local filtering is the Schur product approach (Gaspari and Cohn, 1999 [40]; Houtemaker and Mitchel, 1998 [56]), which attenuates correlations between distant points in accordance with a specified weighting function. This mitigates the effects of spurious long distance correlations caused by sampling error. If the weighting function goes to zero after some finite distance, the result is similar to the clustering technique, since the covariance matrix is again replaced by a sparse blocked approximation.

The computational demands of the global and local filters are:

Global filter: $O\left[n^3\right]$ Local filter:$O\left[\frac{n}{n_l}(n_l)^3\right] = O(nn_l^2)$

When n_l is much smaller than n the local filter is much more efficient. Also, this filter scales linearly with the size of the domain (since n_l is a constant). The cubic scaling of the global filter makes it impractical for very large problems.

We can investigate the performance of the global and local filters by carrying out two simulation experiments, with updates performed at wet and dry times, respectively. This design tends to emphasize the difference between the two estimation alternatives. The wet update times are at t = 50, 90, 320, and 460 while the dry update times are at t = 165, 240, 430, and 515 (all times are indicated in Figures 3-10 and 3-11 with asterisks). The local filter uses blocks of 4 by 4 pixels, the same size as the regions covered by the radiometer measurement, giving $n_l = 48$ states per block (See Figure 3-1a). In order to minimize the effects of sampling error we use an ensemble size of $N = 6000$, which is smaller than n but large enough to enable the global filter to properly account for horizontal correlation at all scales.

The top half of Figure 3-9 compares the spatial distribution of the true replicate with the conditional means from the local and global filters, before and after updates. The plots on the left half show results at t = 90 for the wet update strategy while those on the right show results at t = 430 for the dry update strategy. The bottom half of the figure shows similar plots for the conditional standard deviations.

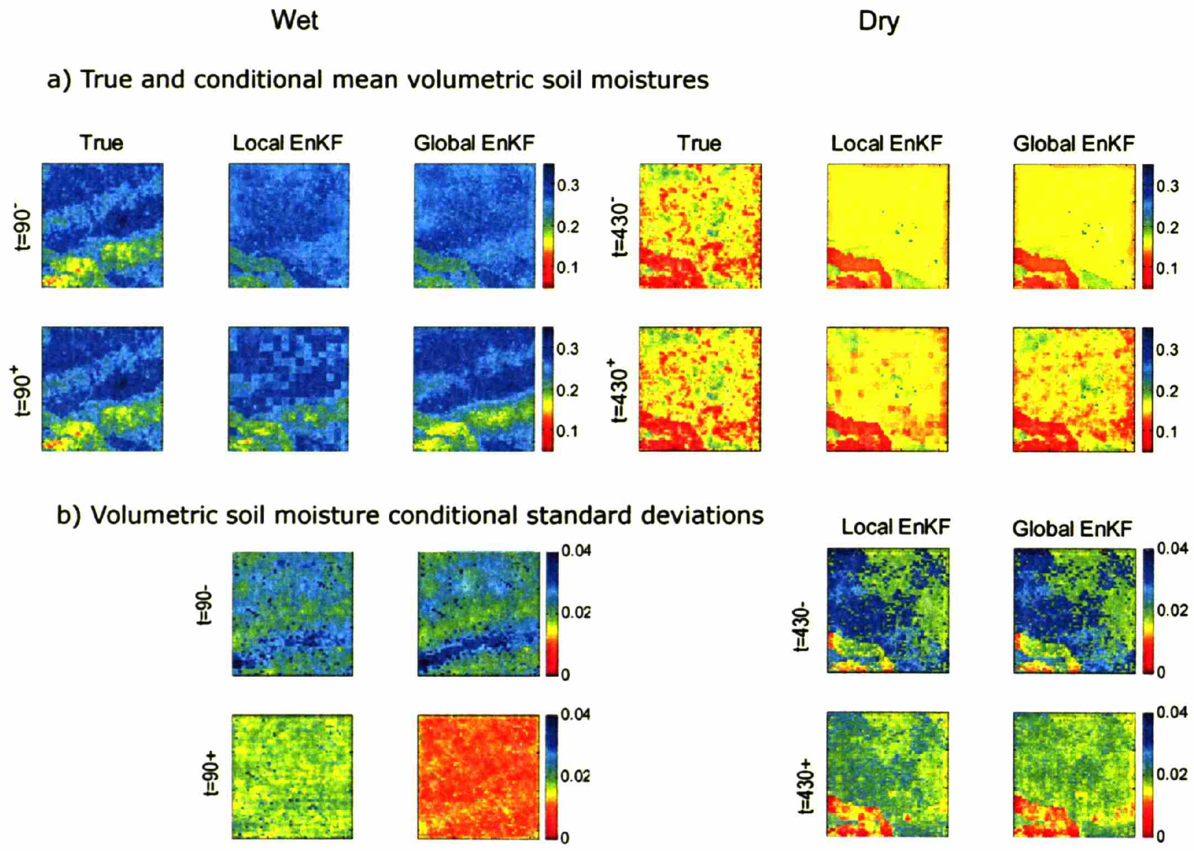


Figure 3-9: Comparison of global and local ensemble Kalman filters a) True soil moisture and conditional ensemble means before (superscript -) and after (superscript +) measurement updates at t = 90 (wet) and t = 430 (dry) b) Corresponding conditional ensemble standard deviations.

It is apparent that the difference between the two filters is greater at $t = 90$ (wet) than at t = 430 (dry). At the wet time the local filter updated mean does not match the true field as well as the global filter. It is also significantly more blocky, reflecting the presence of discontinuities between the independent block estimates. The local filter also gives a significantly higher updated conditional standard deviation than the global filter. This is because it cannot benefit from potentially useful correlations between states and measurements in different blocks. Such correlations are greater during wet periods.

By contrast, the local and global filters behave nearly the same at the dry time. After the long drydown the diagonal features induced by rainfall disappear and most of the variability that remains is related to soil and vegetation heterogeneity. The global filter's ability to account for horizontal correlation across longer distances is no longer of much benefit.

The time series of the spatial root mean squared error for the ensemble prediction (unconditional mean) and the filter estimates (updated conditional means) are shown in Figure 3-10. The top plot corresponds to the wet update strategy while the bottom panel corresponds to the dry update. As expected, the difference between the two filters is greater for the wet update strategy, when the global filter's ability to accommodate horizontal correlation is more advantageous. The two filters are barely distinguishable when all updates are at dry times. Figure 3-11 shows the increase in root mean squared evapotranspiration error of the local filter estimate and ensemble prediction over the error obtained from the global filter (the error increase plot removes the confounding effect of diurnal fluctuations in evapotranspiration, which are typically larger than the estimation errors). Here again, the difference between the two filters is much more significant for the wet update strategy.

Overall, the local filter does quite well in these simulation experiments, especially considering the substantial computational benefits it offers. It appears that most of the information to be gained from microwave measurements is local, in the region covered by the radiometer. Although some improvement can be achieved by accounting for longer range correlations at wet period updates, it is questionable whether

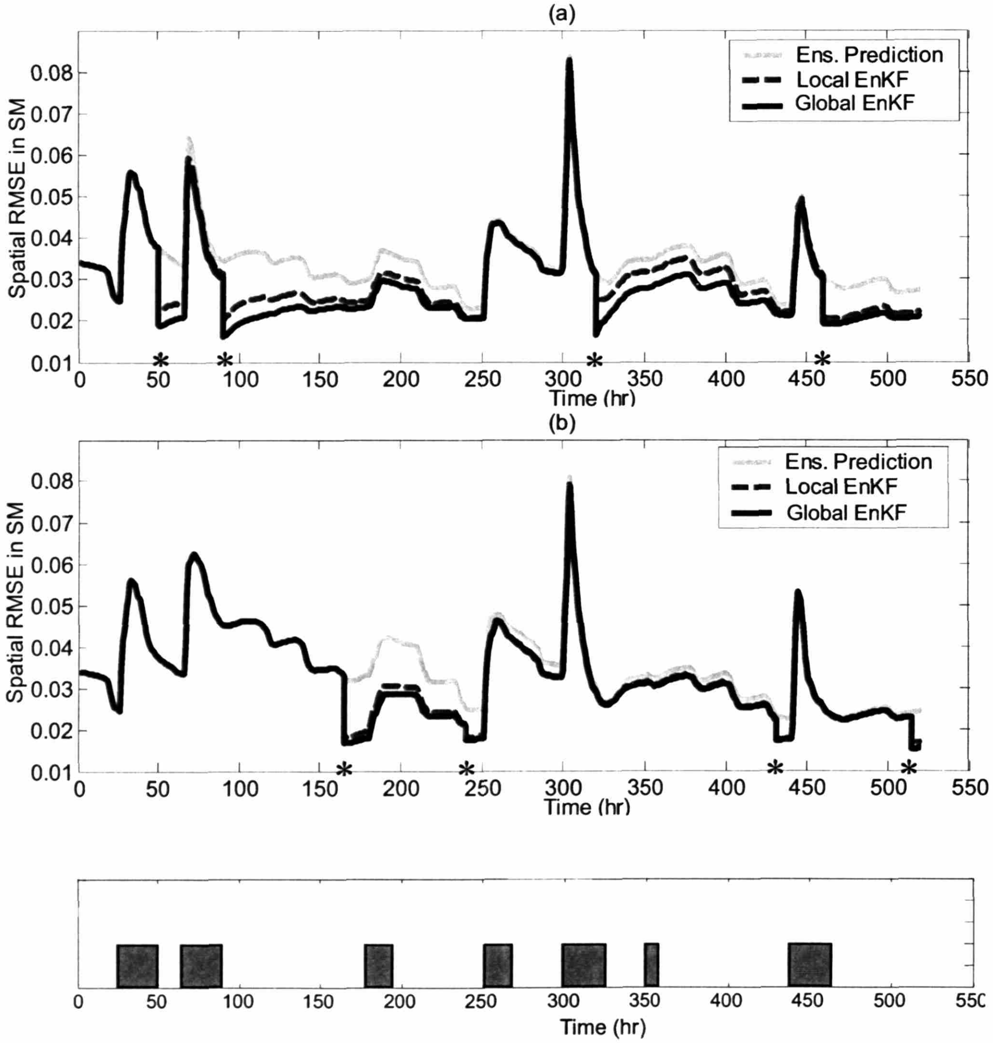


Figure 3-10: The spatial root mean squared soil moisture errors for unconditional ensemble predictions and for local and global ensemble Kalman filter estimates a) Updates at wet times b) Updates at dry times. Each of the update times is marked by a gray asterisk on the time axes. Error is defined as the difference between the respective ensemble mean and the true top layer soil moisture.

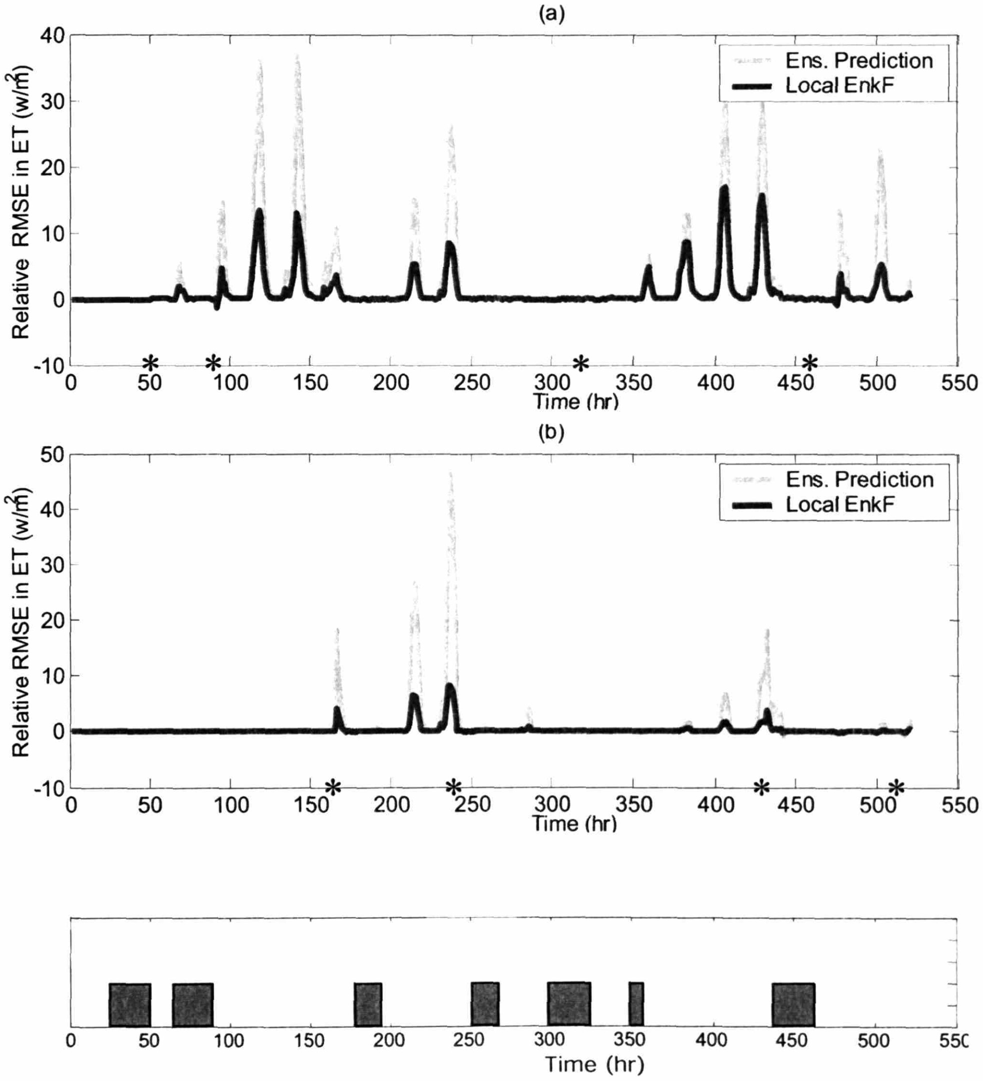


Figure 3-11: Increase in root mean squared evapotranspiration error over value obtained from global filter. Black line corresponds to local filter and gray line to unconditional ensemble prediction. a) Updates at wet times b) Updates at dry times. Each of the update times is marked by a gray asterisk on the time axes.

this justifies the substantial increase in computational effort required by the global filter.

3.5 Discussion and Conclusions

The simulation experiments presented here suggest that it should be possible to obtain accurate estimates of soil moisture and evapotranspiration with efficient data assimilation algorithms that adapt to changing land surface conditions. The alternative ensemble filters considered in Section 5 represent two extremes, one which is efficient but sometimes oversimplified and one which is more nearly optimal but too expensive to be practical for large problems. The relatively good overall performance of the local ensemble Kalman filter suggests that this option is a good place to start if we want to achieve performance improvements while preserving computational efficiency.

One possibility is to increase the size of the local blocks in the local filter during wet periods, following some empirical adjustment rule. While this could probably provide some performance improvement, it would also increase computational effort (since the time required for each local filtering computation increases with the square of the local state dimension n_l). Moreover, the correlations that need to be accommodated are typically anisotropic and not amenable to a simple blocking scheme. It is likely that some overall improvement could be obtained by using a climatological correlation function to derive a fixed (and possibly anisotropic) block geometry (Reichle and Koster, 2004 [82]).

It also worthwhile to mention that the results from this chapter are dependent on how the uncertainties in the system are assumed. Here soil properties with no local correlation is used while rainfall creates soil moisture error correlation scale spanning all over the domain. The rainfall storm size is large compared to the domain size (1/5 of the domain size). It's reasonable to argue that for a large scale problem, say, at scale of tens of thousands kilometers, this soil moisture error correlation structure across domain might rarely occur, in which case local correlation assumption might work well.

An ideal option would be an update algorithm that is able to automatically adapt to the correlation structure implicit in the propagated ensemble. Such an algorithm should be able to gradually make the transition from a reduced rank filter that only requires estimation of a few leading singular vector coefficients (when there is substantial horizontal correlation) to a local filter with a small block size (when there is very little horizontal correlation). It should also be able to attenuate the effects of spurious long distance correlations due to small ensemble sizes. A filter with these capabilities would be both accurate and efficient. It is possible that a wavelet-based or multiscale approach could provide the flexibility needed for a truly adaptive land surface estimation algorithm. The multiscale approach will be discussed in the next chapter.

Chapter 4

An Ensemble Multiscale Filter for Large Scale Nonlinear Estimation

4.1 Background and Motivation

Recent developments in remote sensing and numerical modeling are rapidly improving our understanding of the earth and our ability to predict the environmental consequences of human activities. By using models to combine many different sources of data we can obtain an internally consistent picture of global processes that extends well beyond the capabilities of any single instrument. Data integration carried out at global scales can be viewed as a large estimation problem, where the goal is to characterize a very high dimensional spatially distributed system state. This problem is challenging for a number of reasons. First, its sheer size places a high premium on efficiency. Second, process and instrument nonlinearities are common. Third, model and measurement uncertainties are significant but difficult to characterize. Together, these problem features make conventional estimation methods infeasible for global applications.

The estimation procedure described in this chapter deals with some of the issues posed by large environmental estimation problems. It relies on an ensemble (or Monte Carlo) approach that is sufficiently flexible to accommodate a wide range of nonlinearities and uncertainties. Since ensemble methods tend to be computationally demanding it may seem that they are not appropriate for global environmental applications. This is certainly true if methods developed for small problems are applied without modification to larger problems. But new variants of the ensemble approach may make this option computationally feasible. The significant benefits of ensemble methods, particularly for nonlinear systems, provide ample motivation for further work on this topic.

The computational demands of distributed estimation problems arise both from the cost of solving the model equations and from the cost of updating model predictions with measurements. Both increase rapidly with problem size. Distributed environmental models tend to be expensive because they need to capture a wide range of time and space scales. The desire to resolve small scale variability, which can have large scale effects in nonlinear problems, creates continual pressure for higher resolution and larger grids in disciplines such as meteorology and oceanography. The computational demands of estimation algorithms also depend on the amount of data to be processed. In the earth sciences there are always new sources of data to consider, acquired at higher rates and higher resolution. So the trend is definitely towards larger problems and more demanding computation.

Multiscale methods provide an attractive way to increase the efficiency of distributed modeling and estimation algorithms. Most of these methods have the ability to discriminate between different scales of variability, applying only the level of resolution required at each scale. There has been a considerable interest in multiscale methods for solving deterministic partial differential equations [9] and linear estimation problems [101]. However, there has been relatively little research on multiscale ensemble estimation, especially for nonlinear problems. This is the topic addressed in our paper.

The ensemble estimation method described here is a filtering algorithm that divides naturally into forecast and update steps. The forecast step proceeds in much the same way as in other ensemble filters, propagating individual random replicates forward in time with a nonlinear state equation discretized at a fixed fine scale. The update step is carried out on a multiscale tree that is identified from the forecast ensemble. This approach combines the flexibility of traditional ensemble methods, which can readily handle nonlinearities and complex error sources, and the efficiency of multiscale methods, which can greatly reduce computational effort. In order to make the multiscale update feasible for large problems certain approximations are required. The impacts of these approximations, which can be positive as well as negative, tend to be application-dependent.

Our discussion begins with a formulation of a general dynamic estimation problem and a brief review of the ensemble approach. This is followed by a survey of some relevant multiscale estimation concepts. Next we show how ensemble concepts and multiscale estimation can be combined and we illustrate the performance of the resulting ensemble multiscale filter with two examples. We conclude with a review of the advantages and limitations of the multiscale ensemble approach.

4.2 Ensemble Filtering

The inherent variability in natural processes as well as uncertainty in both models and measurements provide good reason to consider a probabilistic approach to environmental estimation problems. Bayesian estimation theory provides a convenient framework for such an approach. Suppose that we characterize the system of interest by a large number of spatially and temporally discretized states collected in the vector x_t, where the subscript t indicates time. For present purposes we suppose that each element of this vector corresponds to a distributed variable (e.g. pressure, temperature, velocity, etc.) evaluated at a particular cell on a fixed computational grid. The discrete times indexed by t need not be equally spaced.

The Bayesian approach treats the state x_t at any given time as a random vector which is fully characterized by a joint probability density. In the absence of measurements the appropriate choice is the unconditional density $p[x_t]$. When measurements of the states or related variables are also available we can more precisely characterize the uncertain state with conditional densities. Suppose the measurements available at a given time τ are collected in the vector y_τ. Then we write the conditional density of x_t, given all measurements with $\tau \in [0, T]$, as $p[x_t|y_{0:T}]$. This density characterizes everything we know about x_t, given $y_{0:T}$.

It is generally neither feasible nor desirable to derive the entire multivariate density of x_t for large problems. In practice, we focus on certain properties of this density, usually the mean, the mode, the variances and marginal univariate densities of particular elements of x_t, and the covariances between two different elements of x_t. Nevertheless, it is useful to consider $p[x_t|y_{0:T}]$ during problem formulation.

Bayesian estimation reduces to the need to derive $p[x_t|y_{0:T}]$ (or some of its properties) from specified probabilistic information about the state and the measurements. There are a number of ways to do this, depending on the problem at hand. Here we focus on filtering problems where the end of the measurement interval is the current time (so $T = t$). To simplify notation we seek estimates only at measurement times. In addition, we consider problems where the temporal evolution of the state and the measurement process are described as follows:

$$x_t = f_t(x_{t-1}, u_t) \tag{4.1}$$

$$y_t = h_t(x_t) + e_t \tag{4.2}$$

where u_t is a vector of random model inputs, which are not necessarily white or additive; e_t is a measurement noise vector, which we assume to be zero-mean and white in time; and x_t has a random initial condition x_0. The functions $f_t(\cdot)$ and $h_t(\cdot)$ represent time-dependent models of the system dynamics and measurement process. The random initial state, input vector, and measurement noise are all characterized by probability densities, which we assume to be given. We also assume that these random vectors are independent of one another.

The sequential structure of the state equation in (4.1) enables us to solve the Bayesian estimation problem recursively. In this case, the process of deriving the probability densities of x_t at t divides into two steps, a forecast from time $t-1$ to time t, and an update at time t. At any given time $t > 0$ the forecast step derives the forecast density $p[x_t|y_{0:t-1}]$ from the previously updated density $p[x_{t-1}|y_{0:t-1}]$. This can be done by solving the Fokker-Planck, or forward Kolmogorov, equation [60]. Exact solutions are possible only in certain special cases.

The update step at t derives the new updated density $p[x_t|y_{0:t}]$ from the forecast density $p[x_t|y_{0:t-1}]$ and the likelihood $p[y_t|x_t]$. This can be done by applying a sequential version of Bayes' theorem [60]. Here again, exact closed form results can only be obtained for special cases. In particular, exact results are obtainable when the specified densities for x_0, u_t, and e_t are Gaussian and the functions $f_t(\cdot)$ and $g_t(\cdot)$ are linear in all their arguments. In this case, x_t and y_t are jointly Gaussian with densities that are completely defined by their means and covariances. The associated sequential filtering algorithm is the well-known Kalman filter [41].

When Gaussian assumptions do not apply ensemble methods are the only solution options that are both general and computationally feasible. Examples suitable for filtering problems include particle filters [2, 45] and ensemble Kalman filters [33, 10]. Ensemble filtering techniques approximate $p(x_t|y_{0:t-1})$ and $p(x_t|y_{0:t})$ by weighted sums of Dirac delta densities located at the randomly generated replicate values $x^j_{t|t-1}$ or $x^j_{t|t}$.

$$\begin{aligned} p(x_t|y_{0:t-1}) &= \sum_{j=1}^{N} w^j_{t|t-1}\delta(x_{t|t-1} - x^j_{t|t-1}) && (4.3) \\ p(x_t|y_{0:t}) &= \sum_{j=1}^{N} w^j_{t|t}\delta(x_{t|t} - x^j_{t|t}) && (4.4) \end{aligned}$$

Here and in the following discussion $j = 1, \ldots, N$, where N is the number of replicates in the ensemble. The weights $w^j_{t|t-1}$ and $w^j_{t|t}$ given to the N replicates must sum to 1.0.

In the forecast step the replicate values rather than the probability densities are propagated with the state equation as follows:

$$x^j_{t|t-1} = f(x^j_{t-1|t-1}, u^j_t) \tag{4.5}$$

where x^j_0 and u^j_t are synthetically generated replicates drawn from the initial state and input probability densities. During the forecast step the weights generally remain unchanged. The forecast replicates obtained from (4.5) can be used to approximate $p(x_t|y_{0:t-1})$ and its properties.

The update step is more problematic since it is very difficult to generate an ensemble of replicates that adequately approximates the unknown density $p(x_t|y_{0:t})$. One attractive alternative is the ensemble Kalman filter. This filter uses uniform weights of $w^j_{t|t-1} = w^j_{t|t} = 1/N$ throughout both the forecast and update steps and adjusts only the replicates. To simplify the discussion we make the optional but convenient assumption that the measurement function is linear so that:

$$y_t = H_t x_t + e_t \tag{4.6}$$

The ensemble Kalman filter generates updated replicates from the following linear transformation of the forecast replicates:

$$x^j_{t|t} = x^j_{t|t-1} + K'_t[y_t + e^j_t - \hat{y}^j_{t|t-1}] \tag{4.7}$$

where $\hat{y}^j_{t|t-1}$ is a measurement prediction replicate defined by:

$$\hat{y}^j_{t|t-1} = H_t x^j_{t|t-1} \tag{4.8}$$

and K'_t is the Kalman gain defined by:

$$K'_t = \widehat{Cov}[x_t, H_t x_t|y_{0:t-1}]\left[\widehat{Cov}[H_t x_t|y_{0:t-1}] + r_t\right]^{-1} \tag{4.9}$$

The vector e^j_t is a synthetically generated zero-mean random measurement perturbation, drawn from the specified density of the measurement error e_t. The matrix r_t is the covariance of e_t.

Here and in the subsequent discussion $Cov[u, v|w] = E\{[u-E(u|w)][v-E(v|w)]^T|w\}$ indicates the covariance of the two random arguments u and v, conditioned on w and $Cov[u|w] = Cov[u, u|w]$. $\widehat{Cov}[u, v|w]$ is a sample estimate of $Cov[u, v|w]$ computed from N replicates of u and v, for a given w. This sample covariance can be written as a matrix product of the following form:

$$\widehat{Cov}[u, v|w] = \frac{1}{N-1}\tilde{U}\tilde{V}^T \tag{4.10}$$

where $\tilde{U}$ is a matrix with column j the mean-removed replicate $\tilde{u}^j$ and $\tilde{V}$ is a matrix with column j the mean-removed replicate $\tilde{v}^j$, both computed for a particular value of w.

In most ensemble estimation problems N is less than the dimensions of u and v and the sample covariance is rank deficient. However, the matrix inverted in (4.9) is full rank if r_t is full rank. Also, note that the Kalman gain in the ensemble version of the Kalman filter is a weak nonlinear function of past measurements since it depends, through the sample covariances, on replicates derived from these measurements.

The linear update of (4.7) yields an ensemble that converges to the exact $p(x_t|y_{0:t})$ if the state and measurement at t have a Gaussian joint conditional density $p(x_t, y_t|y_{0:t-1})$. If the state and measurement vectors are not jointly Gaussian an ensemble obtained from (4.7) will not converge to the exact conditional density. However, the mean of the updated ensemble converges to the minimum variance linear estimate and the covariance of the updated ensemble converges to the associated estimation error covariance, even for non-Gaussian problems.

For nonlinear problems the ensemble Kalman filter is a compromise that makes no assumptions during the forecast step but makes implicit linear Gaussian assumptions during the update step (note that the measurements and states are jointly Gaussian only if the state is Gaussian and the measurement function is linear). Despite this compromise, experience shows that the ensemble Kalman filter provides an acceptable approximation in many applications. This issue is discussed in more detail in citations provided in Evenson [34] as well as in [102] and in Chapter 2 of this thesis. There is, of course, no guarantee that the assumptions made in the ensemble Kalman filter will always be acceptable and caution must be used when applying this technique to highly nonlinear problems.

The ensemble Kalman filter has some features which can complicate its application to large problems. First, it requires computation and manipulation of sample covariance matrices which can be very large when the state and/or measurement vectors are large. This problem can be mitigated somewhat by using numerical implementations that are more efficient than the traditional approach given in (4.7) through (4.10). However, these refinements still have the disadvantage of working simultaneously with all the elements of a very large array of state replicates.

A second problem has to do with sampling error resulting from small ensemble sizes. In large problems the number of replicates that can be feasibly accommodated is small while the number of replicates needed to obtain accurate sample estimates of the forecast covariances is large. So sampling errors are often a serious problem, especially when estimating small but physically meaningful covariances [56]. One popular solution to the sampling problem is to impose a spatial filter that essentially ignores sample correlations over distances beyond some threshold. This so-called localization approach presumes that long distance correlations are small and that high magnitude sample correlations at such distances are spurious. Unfortunately, this is not always the case. In some problems, long-distance correlations arise as a natural result of physical processes. This is illustrated, for example, in [73] and in Chapter 3 of this thesis, where moving rain storms can induce long distance correlations in soil moisture.

The ensemble multiscale filter described here provides an alternative approach that deals with both the computational and sampling error issues of traditional ensemble Kalman filtering algorithms. The multiscale filter adopts the same forecast step as the ensemble Kalman filter but uses an update based on a multiscale tree model. The tree model describes global relationships between the elements of x_t in terms of local relationships between nodes on the tree. This leads to a multiscale ensemble update that is much more efficient than the classical covariance-based ensemble Kalman update. In the multiscale approach a different tree model is identified from the forecast ensemble at every update time, to provide for the time-dependent nature of the problem. The identification step introduces approximations that have the effect of filtering errors in sample correlations. Since this filtering procedure is accomplished with a truncation in scale rather than distance it is able to preserve physically legitimate long-range correlations.

The next section describes how tree models may be used to efficiently describe ensemble statistics for very large spatially distributed problems. This is followed by a detailed discussion of the ensemble multiscale filter.

4.3 Multiscale Models

4.3.1 Multiscale Descriptions of Random Fields

A tree model consists of a set of related nodes that may be visualized as shown in Figure 4-1. Groups of nodes are organized into scales distinguished as separate rows in the tree diagram. The scale with the most nodes (at the bottom)is called the finest scale while the scale with only one node (the root node at the top) is the coarsest scale. Each node s is associated with a relatively small nodal state vector $\chi(s)$ of dimension $n(s)$. The definition of each nodal state vector is a design decision that is part of the modeling process.

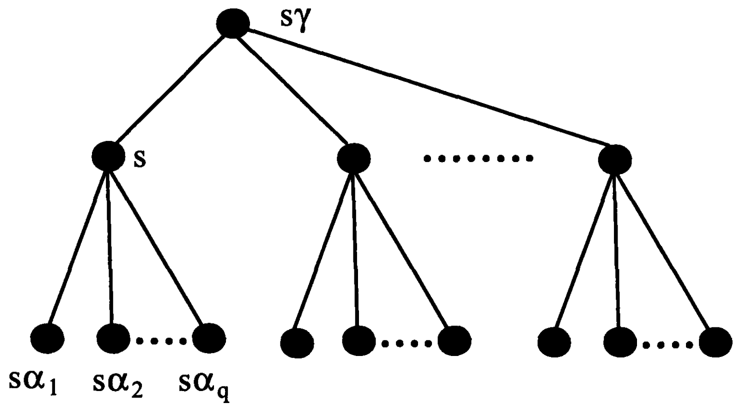


Figure 4-1: A multiscale tree with scaling factor q

In the applications considered here the nodal state vectors are first defined at the finest scale and then recursively identified for each of the coarser scale. Since the tree is designed to provide a compact representation of the ensemble propagated on the computational grid it is convenient to associate each of the finest scale nodes with a particular block of grid cells. The grid block is intentionally kept small (typically 4 to 32 cells) to limit the size of the nodal state vector. The nodal state vector $\chi(s)$ at finest scale node s, the $n(s)$ elements of the mean-removed global state vector associated with block s, are the elements of the global state vector x_t that correspond to grid block s.

In a tree model the finest scale nodes are not related directly to one another, as they are in a typical computational grid. Instead, they are related indirectly through their relationships with common nodes located higher up the tree. Each internal tree node is related to a parent and to several children. The parent-child relationships are indicated graphically by lines on the tree diagram. In order for the tree to serve as a model of the forecast ensemble the local parent-child relationships need to properly represent global correlations between the finest scale states. Methods for deriving these relationships are discussed below.

Unless indicated otherwise we suppose that every node (except the finest scale nodes) has q children and we represent the children of s by $s\alpha_1, s\alpha_2, \ldots, s\alpha_q$. Also, every node (except the root node) has a single parent $s\gamma$. The index $m(s)$ indicates the scale of node s (i.e. the row on the tree diagram containing s). This index increases from 0 at the top of the tree (coarsest scale) to M at the bottom of the tree (finest scale). The fluid mechanics example discussed in Section 5 has two scalar states (the two components of velocity) per computational grid cell in a grid of or $524,288$ (1024 by 512) cells giving a total of $n(x) = 1,048,576$ global states. The corresponding tree has 8 scales. The coarsest scale has two children while all others have 4 children, giving a total of $2 \times (4^7) = 32768$ finest scale nodes. Each finest scale node has $d_M = 32$ local states associated with a small block of 16 grid cells. So the tree also has a total of $n_M = 32 \times (32768)$ states at the finest scale. It is apparent that a tree with a relatively small number of scales and relatively small local state vectors can accommodate large problems with large global state vectors.

A certain class of tree models called multiresolution autoregressive (MAR) models can be used to derive particularly efficient scale-recursive estimation algorithms. The downward recursion used to describe an MAR model is:

$$\chi(s) \quad = \quad A(s)\chi(s\gamma) + w(s) \tag{4.11}$$

where $\chi(s\gamma)$ is the state at the parent $s\gamma$ of s, $A(s)$ is a downward transition matrix and $w(s)$ is a zero-mean random scale perturbation with covariance $Q(s)$. The random root node state $\chi(0)$ that initializes the recursion is zero mean with covariance $Cov[\chi(0)]$. Together these assumptions imply that all the states on the tree are zero mean. The $w(s)$ values at different nodes are uncorrelated with one another and with $\chi(0)$. Note that the scale of the parent node $s\gamma$ is $m(s\gamma) = m(s) - 1$.

An equivalent upward MAR model recursion is (4.11):

$$\chi(s\gamma) \quad = \quad F(s)\chi(s) + w'(s) \tag{4.12}$$

where $F(s)$ is an upward transition matrix and $w'(s)$ is a zero-mean random scale perturbation with covariance $Q'(s)$. This recursion is initialized with the random finest scale state χ_M, which is zero mean with covariance $Cov[\chi_M]$. If the model is multiresolution autoregressive the $w'(s)$ at different nodes must be uncorrelated with one another and with χ_M. The transition matrices $A(s)$ and $F(s)$ and the covariances $Q(s)$ and $Q'(S)$ are related to one another and to the specified finest scale prior covariance $Cov[x_M]$, which can be viewed as a terminal condition for the downward recursion and an initial condition for the upward recursion.

A tree that satisfies the scale recursions given in (4.11) and (4.12) has a number of properties convenient for estimation. In particular, when these recursions apply it is possible to carry out an optimal linear measurement update recursively, moving up and then down the tree from scale to scale. At each step of the update algorithm, it is only necessary to consider relationships between a particular node and its parent or children. The associated computations deal only with local state and measurement vectors, which are much smaller than their global counterparts. This makes the multiscale update algorithm very efficient.

4.3.2 Internal Models and the Scale-recursive Markov Property

The process of identifying efficient tree models is greatly facilitated if we constrain the model to have certain internal properties. To define these properties suppose that $\chi(s)$ is the state vector at node s at scale $m(s)$ and $\chi_{m(s)+1}(s)$ is the vector of all states at the children of s, which are all at scale $m(s)+1$. The tree is said to be a locally internal model if $\chi(s)$ is a linear combination of the states at its children, for all nodes on the tree. This requirement can be expressed concisely as follows [37]:

$$\chi(s) \;=\; V(s)\chi_{m(s)+1} = V(s)\begin{bmatrix} \chi(s\alpha_1) \\ \vdots \\ \chi(s\alpha_q) \end{bmatrix} \tag{4.13}$$

where $V(s)$ is an $n(s)$ by $n_{m(s)+1}$ dimensional local internal matrix associated with node s and $n_{m(s)+1}$ is the sum of the dimensions of the state vectors $\chi(s\alpha_i)$ for $i = 1,\ldots,q$. The set of $V(s)$ matrices defines, through (4.13), all the coarser scale states on the tree. Specification of the $V(s)$ matrices is therefore equivalent to specification of the scale transition and covariance matrices defined in (4.11) and (4.12).

The recursion of (4.13) may also be expressed as follows:

$$\chi(s) \;=\; W(s)\chi_M(s) \tag{4.14}$$

where $\chi_M(s)$ is the vector of states at all finest scale nodes (ie. at scale $m(s) = M$) descended from s and $W(s)$ is an $n(s)$ by $q^{[M-m(s)]}d_M$ dimensional finest scale internal matrix. This matrix can be derived from the local internal matrices for s and its descendants, using a recursion described in [101]. Note that (4.14) implies that all states on the tree are zero mean since the elements of the finest scale nodal state vector χ_M are constructed from mean-removed elements of the global state vector.

We impose the internality condition of (4.13) and describe tree structure in terms of the $V(s)$ matrices in order to facilitate the process of identifying the recursions of (4.11) and (4.12). The statistical assumptions that accompany (4.11) and (4.12) are equivalent to a multiscale extension of the well-known Markov property of time series analysis. The multiscale Markov property relies on the fact that any given node s at scale $m(s)$ partitions the nodes at the next finer scale $m(s)+1$ into $q+1$ sets. The first q sets consist of the q children of s. The final set consists of the complementary group of nodes that are not children of s. The multiscale Markov property holds for locally internal models if and only if the vector of all states in any one of these $q+1$ sets is conditionally uncorrelated with the vector of all the states in the remaining q sets, given $\chi(s)$ [37]. If the tree obtained from a particular set of $V(s)$ matrices meets this requirement it is possible to identify scale transition equations having the form given in (4.11) and (4.12). This leads us to consider practical methods for identifying the $V(s)$ matrices.

4.3.3 Identification of Multiscale Tree Models

The objective of tree identification is to obtain a locally internal model (i.e. a set of $V(s)$ matrices) that satisfies the decorrelation requirements of the multiscale Markov property. In order to obtain perfect decorrelation we often need to use high-dimensional coarser scale nodal state vectors. This defeats the purpose of the tree, which is to provide a concise and efficient alternative to traditional estimation methods. For practical applications we need to constrain state dimensionality at each coarser scale node or, equivalently, we need to limit the number of rows in the corresponding $V(s)$ matrix. Then the identification problem at node s reduces to a search for the $V(s)$ that minimizes the conditional covariance subject to the constraint $n(s) \leq d(s)$, where $d(s)$ is the maximum number of states allowed at s. This requires a measure of conditional covariance that can be readily minimized.

The identification problem is easier to solve if the set of $V(s)$ candidates is limited to block diagonal matrices having the following form:

$$V(s) = diag[V_1(s), ..., V_q(s)] \tag{4.15}$$

The submatrix $V_i(s)$ corresponds to child $s\alpha_i$ and has dimension $d_i(s)$ by $n(s\alpha_i)$. The block diagonal structure of $V(s)$ implies that each row of $\chi(s)$ is a linear combination of states at a particular child of s. It also enables us to divide the $V(s)$ identification problem into q smaller problems that each focus on the covariance between a particular pair of state vectors. Frakt [37] discusses some of the attractive properties of block diagonal $V(s)$ matrices.

Focusing for the moment on the i^{th} child of s, consider the conditional covariance, given $\chi(s)$, between the vector $z_i(s) = \chi(s\alpha_i)$ composed of the states at child $s\alpha_i$ and the complementary vector $z_{ic}(s)$ composed of the states at all other nodes at the same scale.:

$$\begin{aligned}
&Cov[z_i(s), z_{ic}(s)|\chi(s)] = \\
&\quad E\left[[z_i - E[z_i|\chi(s)]]\,[z_{ic} - E[z_{ic}|\chi(s)]]^T\right] = \\
&\quad E\left[[\chi(s\alpha_i) - E[\chi(s\alpha_i)|\chi(s)]]\,[z_{ic} - E[z_{ic}|\chi(s)]]^T\right]
\end{aligned} \tag{4.16}$$

This is one of q conditional covariances addressed in the scale-recursive Markov property.

Our objective is to select a $V_i(s)$ and the associated state $\chi(s) = V_i(s)z_i = V_i(s)\chi(s\alpha_i)$ that gives the smallest possible conditional covariance. Rather than minimize the conditional covariance directly the predictive efficiency method minimizes a more convenient surrogate performance measure. This measure is related to the mean-squared error of the following estimate of $z_{ic}(s)$, derived from a linear function of $z_i(s)$.

$$\hat{z}_{ic}(s) = E[z_{ic}(s)|\chi(s)] = E[z_{ic}(s)|V_i(s)z_i(s)] \tag{4.17}$$

When $V_i(s)$ is an identity with the same dimension as $z_i(s)$ the mean-squared error is minimized and the conditional covariance is zero (this is easily checked by substituting $\chi(s) = z_i(s)$ into (4.16)). For a general $V(s)$ the mean-squared error will be larger than the minimum value obtained when $V_i(s)$ is an identity and the conditional covariance will not be zero. In this case, the departure from optimality can be measured by the following relative mean-squared error expression [37]:

$$\begin{aligned}\bar{\epsilon}[z_{ic}(s)|\chi(s)] = \text{trace}\left\{Cov[z_{ic}, z_i]Cov^{-1}[z_i]Cov[z_i, z_{ic}]\right\} - \\ \text{trace}\left\{Cov[z_{ic}, z_i]V_i^T\left[V_iCov[z_i]V_i^T\right]^{-1}V_iCov[z_i, z_{ic}]\right\}\end{aligned} \tag{4.18}$$

In the predictive efficiency approach the best choice of $V_i(s)$ is taken to be the one that minimizes $\bar{\epsilon}[z_{ic}(s)|\chi(s)]$ subject to the constraint $n(s) \leq d_i(s)$. Frakt [37] shows that this $V_i(s)$ is given by the first $d_i(s)$ rows of the following matrix $V_i'(s)$:

$$V_i'(s) \quad = \quad U_i^T(s)Cov[z_i(s)]^{-1/2} \tag{4.19}$$

The columns of the matrix $U_i(s)$ are the eigenvectors of the following $n(s\alpha_i)$ dimensional square matrix:

$$Cov^{-1/2}[z_i(s)]Cov[z_i(s), z_{ic}(s)]Cov^T[z_i(s), z_{ic}(s)]Cov^{-T/2}[z_i(s)] \tag{4.20}$$

These eigenvectors are assumed to be arranged according to the magnitudes of the corresponding eigenvalues, from largest to smallest.

The predictive efficiency method outlined above provides a local internal matrix $V_i(s)$ for each child of s. These q $V_i(s)$ matrices can be assembled to form $V(s)$, as specified in (4.15). The total number of rows in the resulting $V(s)$ may exceed the total number of rows $d(s)$ originally specified for $V(s)$ in the nodal state vector size constraint. Following Frakt [37], we deal with this issue by retaining only the $d(s)$ rows in $V(s)$ that correspond to the $d(s)$ largest predictive efficiency eigenvalues. Note that this will generally result in some of the reduced $V_i(s)$ matrices having more rows than others. That is, some children will contribute more elements to $\chi(s)$ than others.

Once an appropriate set of $V(s)$ matrices has been identified it is possible to derive the scale transition and covariance matrices that appear in (4.11) and (4.12). These quantities can be written as functions of prior nodal state covariances, as follows:

[37]:

$$
\begin{aligned}
A(s) &= Cov[\chi(s), \chi(s\gamma)]Cov^{-1}[\chi(s\gamma)] && (4.21)\\
F(s) &= Cov[\chi(s\gamma)]A(s)^T Cov^{-1}[\chi(s)] && (4.22)\\
Q(s) &= Cov[\chi(s)] - A(s)Cov[\chi(s), \chi(s\gamma)]^T && (4.23)\\
Q'(s) &= Cov[\chi(s\gamma)] - F(s)A(s)Cov[\chi(s\gamma)] && (4.24)
\end{aligned}
$$

The covariances that appear on the right sides of these expressions can be derived recursively from the $V(s)$ matrices and the specified finest scale covariance [37]. This is particularly convenient in an ensemble implementation where we rely on sample estimates of the covariances. Further details are provided in the next section.

After getting the A matrices, the covariance between states at any two nodes s and l can be readily determined from the related A matrices and the covariance at their closest common ancestor $s\Lambda l$ according to

$$
\begin{aligned}
&Cov(s,l) = \Phi(s, s \wedge l)Cov(s \wedge l)\Phi^T(s, s \wedge l)\\
&\text{where } \Phi(s,\xi) = \begin{cases} I \text{ if } \xi = s \\ A(s)\Phi(s\gamma, \xi) \text{ where } \xi \text{ is an ancestor of } s \end{cases} && (4.25)
\end{aligned}
$$

The predictive efficiency identification process moves up the tree, starting at the next to finest scale ($m(s) = M - 1$) and ending at the root node ($m(s) = 0$). The $V(s)$ computations at scale $m(s)$ node rely on the covariances $Cov[z_i(s)]$ and $Cov[z_i(s), z_{ic}(s)]$ between states at the next lower scale $m(s) + 1$, as indicated in (4.20). When the size of the vector $z_{ic}(s)$ is large (as it typically is in practical applications) derivation of the cross-covariance $Cov[z_i(s), z_{ic}(s)]$ can be computationally demanding. The effort required can be reduced dramatically if the cross-covariance computation only considers correlations between states at nearby nodes. This is feasible when the tree model is internal and the finest scale nodes are associated with groups of nearby pixels, as is assumed in our application. Then the nodes at higher scales correspond to spatial regions of increasingly larger size, with the region of each node containing the regions associated with its children. This makes it possible to define a neighborhood $\mathcal{N}_h(s)$ of h nodes around any given node s at scale $m(s)$. Nodes associated with $z_{ic}(s)$ are included in the calculation of $Cov[z_i(s), z_{ic}(s)]$ only if they lie in $\mathcal{N}_h(s\alpha_i)$. Frakt [37] shows that the complexity of the tree identification algorithm is $\mathcal{O}[n(\chi)^2]$ if the entire cross-covariance is computed. This complexity reduces to $\mathcal{O}[n(\chi)]$ if the neighborhood restriction is enforced, with the same h value used at all scales. The result is a substantial savings in computational effort for large problems.

Note that the use of a neighborhood screen for deriving $Cov[z_i(s), z_{ic}(s)]$ and $V(s)$ is not the same as the spatial localization approach proposed in [20]. The neighborhood screening approximation suggested here preserves the ability to represent large-scale correlations since the tree still relates distant nodes through their common ancestors. The state dimension truncation and neighborhood screening operations used in the internal matrix calculations have the combined effect of filtering spurious fluctuations such as sampling errors while retaining dominant long distance correlations. The examples discussed at the end of this chapter illustrate this filtering action, which is particularly useful in ensemble applications. Of course, if the state truncation and neighborhood screening are too severe important information will be lost and there will be a decline in performance. So some judgement is involved in selecting the appropriate level of state truncation (d) and the neighborhood size (h) for a given application.

4.4 A Multiscale Ensemble Update Procedure

The update step of the multiscale ensemble filter carries out the classical ensemble Kalman filter measurement update described in (4.7) through (4.9) on a tree model identified from the forecast ensemble. The multiscale measurement update at time t consists of two basic tasks, each divided into a few subtasks:

1. Identify and initialize the tree model

    (a) Construct a prior finest scale mean-removed tree ensemble χ_M^j from the global forecast ensemble $x^j_{t|t-1}$.

(b) Use the predictive efficiency approach and prior finest scale replicates to identify the $V(s)$ matrices, nodal prior replicates $\chi^j(s)$, and scale transition and covariance matrices for all nodes on the tree.

2. Update the prior nodal states with measurements

   (a) Assign each measurement in the vector y_t to a tree node.

   (b) Carry out an upward sweep to obtain a set of updated replicates $\chi^j(s|s) = \chi^j(s|y(\mathcal{V}_s))$ where $y(\mathcal{V}_s)$ is the set of measurements at all nodes $\mathcal{V}_s$ that are at s or its descendants.

   (c) Carry out a downward sweep to obtain a set of smoothed replicates $\chi^j(s|S) = \chi^j(s|y(\mathcal{V}))$ where $y(\mathcal{V})$ is the set of measurements at all nodes $\mathcal{V}$ on the tree.

   (d) Construct the global updated ensemble $x^j_{t|t}$ from the smoothed finest scale tree ensemble $\chi^j(s|S)$, for $m(s) = M$, obtained at the end of the downward sweep.

In order to maintain consistency with published discussions of static multiscale estimation finest and coarser scale nodal replicates derived directly from the forecast ensemble are called "prior replicates" and their sample statistics are called "prior statistics". The above tasks are discussed in more detail in the following two sections.

4.4.1 Model Identification and Initialization

The ensemble version of the multiscale tree identification procedure is an upward recursion based on the predictive efficiency approach described in Section 3.3, with all the exact covariances replaced by sample estimates derived from replicates. The process is initialized by assigning elements of each forecast replicate to the finest scale nodes of the tree, using the grid blocking technique discussed earlier. The resulting nodal replicates are represented by $\chi^j(s)$ and the vector of all finest scale states is represented by χ^j_M.

The internal submatrix $V_i'(s)$ for each child $s\alpha_i$ of s is:

$$V_i'(s) \;=\; \hat{U}_i^T(s)\widehat{Cov}[z_i(s)]^{-1/2} \tag{4.26}$$

where the columns of the matrix $\hat{U}_i(s)$ are the eigenvectors of the following matrix:

$$\widehat{Cov}^{-1/2}[z_i(s)]\widehat{Cov}[z_i(s), z_{ic}(s)]\widehat{Cov}^{T}[z_i(s), z_{ic}(s)]\widehat{Cov}^{-T/2}[z_i(s)] \tag{4.27}$$

All of the replicates needed to derive the sample covariances appearing in (4.27) are available from the previous identification iteration at scale $m(s)+1$, since the components of the vectors z_i and z_{ic} depend only on states at scale $m(s)+1$.

The submatrices obtained for all q chidden of s are assembled in a larger internal matrix $V'(s)$ as follows:

$$V'(s) = diag[V_1'(s), ..., V_q'(s)] \tag{4.28}$$

The final internal matrix $V(s)$ contains only the $d(s)$ rows in $V'(s)$ that correspond to the $d(s)$ largest predictive efficiency eigenvalues, as discussed in Section 3.3. This insures that the size of the nodal state vector at s will not exceed the specified dimensionality limit of $d(s)$, while giving the best possible conditional decorrelation.

The identification calculations at node s are completed with the evaluation of the state replicates at this scale, using the ensemble version of (4.13):

$$\chi^j(s) \;=\; V(s)\chi^j_{m(s)+1} \tag{4.29}$$

When the $V(s)$ and $\chi^j(s)$ at all nodes at scale $m(s)$ have been evaluated the identification recursion moves to the next higher scale $m(s)-1$, moving up the tree. As the recursion proceeds coarser scale prior replicates obtained from (4.26) are used to derive sample estimates of the covariances appearing on the right sides of (4.21) through (4.24). These equations are then used to obtain approximate values for $A(s)$, $F(s)$, $Q(s)$ and $Q'(s)$.

4.4.2 Measurement Update

At the beginning of the update step at t we are given the prior finest scale ensemble χ^j_M, constructed from the global forecast ensemble $x^j_{t|t-1}$. This prior ensemble approximates the forecast density $p(x_t|y_{0:t-1})$ conditioned on all measurements taken through $t-1$. The goal of the update step is to produce a new ensemble $x^j_{t|t}$ that approximates the density $p(x_t|y_{0:t})$ conditioned on all measurements taken through t. It does this by updating each of the finest scale replicates to reflect the information contained in the new measurement y_t obtained at t. In order to use a tree-based update measurements contained in the vector y_t must be associated with particular nodes on the tree. There is generally a straightforward procedure, as discussed below.

Willsky [101] and a number of references he cites describe in detail a static multiscale estimation algorithm that derives the Gaussian conditional mean and covariance (or the minimum variance linear estimate and estimation error covariance) for states and measurements distributed on a multiscale tree. Here we present an adaptation of this algorithm suitable for ensemble applications. In such applications the primary focus is on the replicates of the ensemble, rather than the moments, which may always be estimated from the ensemble.

The ensemble multiscale update at t consists of two sweeps that are analogous to the forward and backward sweeps of the Rauch-Tung-Striebel algorithm used to compute smoothed time series estimates [41]. The first sweep is a recursion that moves upwards through the tree, starting at the finest scale nodes and ending at the root node, while the second sweep is a recursion that moves downward, from the root node back to the finest scale. Further details are provided in the following sections.

Upward Sweep

The ensemble Kalman filter, in either its traditional or multiscale form, does not require that the temporal dynamics described by the state equation of (4.1) be linear. However, its update relies on linear theory that is easier to apply if we assume that the measurements used for updating are linear functions of the global state vector x_t.

Consequently, we adopt the linear measurement function already introduced in (4.6):

$$y_t = H_t x_t + e_t \tag{4.30}$$

where H_t is a global measurement matrix and e_t is a zero mean random measurement noise vector with a specified covariance r_t. In the subsequent discussion we drop time subscripts unless required for clarification, since all update computations are carried out at time t.

In order to use a multiscale framework we need to locate the global measurements in the vector y_t on the tree constructed in the identification step. If the measurements used for updating have spatial supports that are coarser than the finest scale of the tree it is convenient to locate these measurements at coarser scale nodes. In particular, if a particular measurement depends on states at nodes that are descended from node s it may be expressed as:

$$y(s) = h(s)\chi_M(s) + e(s) \tag{4.31}$$

Here $y(s)$ is a subset of the global measurement vector y, located for convenience at node s, and $\chi_M(s)$ is the vector of finest scale states descended from s. The local measurement error vector $e(s)$ has a specified covariance $r(s)$.

The upward sweep of the ensemble multiscale update algorithm derives at each node s a set of updated replicates $\chi^j(s|s)$ that depend on measurements located at s and its descendants. This upward sweep is formulated here as a recursion that moves progressively from the finest to coarsest scales. Appendix A demonstrates that the sample mean and covariance of the replicates produced on the upward ensemble converge to the corresponding exact Gaussian conditional mean and covariance.

The update for the replicate at node s is proportional to the difference between an augmented perturbed measurement vector $Y^j(s)$ and an augmented measurement prediction vector $\hat{Y}^j(s)$. At the finest scale these are defined as in the traditional ensemble Kalman filter [34].

$$Y^j(s) = y(s) + e^j(s) \qquad ; m(s) = M \tag{4.32}$$

$$\hat{Y}^j(s) = h(s)\chi_M^j(s) \qquad ; m(s) = M \tag{4.33}$$

The zero mean random measurement perturbation $e^j(s)$ has the same covariance $r(s)$ as $e(s)$ in (4.31) and is included to insure that the update algorithm yields the correct conditional covariance.

At scales $m(s) < M$ above the finest scale the perturbed and predicted measurement vectors are constructed as follows:

$$Y^j(s) = \begin{bmatrix} K'(s\alpha_1)Y^j(s\alpha_1) \\ \vdots \\ K'(s\alpha_q)Y^j(s\alpha_q) \\ y(s) + e^j(s) \end{bmatrix} \qquad ; m(s) < M \tag{4.34}$$

$$\hat{Y}^j(s) = \begin{bmatrix} K'(s\alpha_1)\hat{Y}^j(s\alpha_1) \\ \vdots \\ K'(s\alpha_q)\hat{Y}^j(s\alpha_q) \\ h(s)\chi_M^j(s) \end{bmatrix} \qquad ; m(s) < M \tag{4.35}$$

where the $K'(s\alpha_i)$ are Kalman gain matrices defined below. Note that the perturbed measurement and measurement prediction vectors at s are augmented with linear functions of the corresponding vectors at the children of s. The augmented vectors convey, in an aggregate way, all the information from measurements at scales below $m(s)$. Also, note that $Y^j(s\alpha_i)$ and $K'(s\alpha_i)$ are available from previous iteration of the upward sweep.

The Kalman gains appearing in (4.34) and (4.35) are computed with the following recursion:

$$R(s) = r(s) \quad ; m(s) = M \tag{4.36}$$

$$K'(s) = \widehat{Cov}[\chi(s), \hat{Y}(s)] \left[\widehat{Cov}[\hat{Y}(s)] + R(s)\right]^{-1} ; m(s) = M \tag{4.37}$$

$$\begin{aligned} R(s) = diag[&K'(s\alpha_1)R(s\alpha_1)K'^T(s\alpha_1), \ldots, \\ &K'(s\alpha_q)R(s\alpha_q)K'^T(s\alpha_q), r(s)] \qquad ; m(s) < M \end{aligned} \tag{4.38}$$

$$K'(s) = \widehat{Cov}[\chi(s), \hat{Y}(s)] \left[\widehat{Cov}[\hat{Y}(s)] + R(s)\right]^{-1} ; m(s) < M \tag{4.39}$$

Here $diag[\cdot]$ represents a square matrix with $q+1$ by $q+1$ square blocks. Diagonal blocks $i = 1, \ldots, q$ have dimension $n(s\alpha_i)$ and diagonal block $q+1$ has dimension $n[Y^j(s)]$. All off-diagonal blocks are zero. The significance of the matrix $R(s)$ is discussed in Appendix A. This matrix depends only weakly on the replicate values (through its dependence on the sample Kalman gains and sample covariance matrices) and will generally be full rank. In this case the matrices to be inverted in (4.37) and (4.39) are non-singular, even for small N.

The prior replicates $\chi^j(s)$, Kalman gain $K'(s)$, perturbed measurement vector $Y^j(s)$, and measurement prediction vector $\hat{Y}^j(s)$ are combined to give the state update at node s:

$$\chi^j(s|s) = \chi^j(s) + K'(s)\left[Y^j(s) - \hat{Y}^j(s)\right] \quad ; m(s) \leq M \tag{4.40}$$

The upward sweep algorithm given here does not make explicit use of the upward multiscale transition equation but it does reply on the $V(s)$ matrices, which convey equivalent information.

Downward Sweep

The downward sweep of the ensemble multiscale update algorithm derives at each node s a set of smoothed replicates $\chi^j(s|S)$ that depend on all measurements on the tree. The downward sweep is formulated here as a recursion that moves progressively from the coarsest to finest scales. Appendix B demonstrates that the sample mean and covariance of the replicates produced on the downward sweep converge to the corresponding exact Gaussian conditional mean and covariance. The ensemble downward update is a direct extension of the moment-oriented downward update described in Willsky [101]:

At the end of upward sweep, the updated root node replicates $\chi^j(0|0) = \chi^j(0|S)$ incorporate all measurements on the tree and so already constitute a smoothed ensemble. At any node s below the root smoothed replicates $\chi^j(s|S)$ are obtained by adjusting the corresponding updated replicates $\chi^j(s|s)$ from the upward sweep. This requires computation of a set of projected replicates $\chi^j(s\gamma|s)$ at $s\gamma$ that characterize the state at the parent of s, given measurements at s and its descendants:

$$\chi^j(s\gamma|s) \quad = \quad F(s)\chi^j(s|s) + w'^j(s) \qquad ; m(s) > 0 \tag{4.41}$$

where the random perturbation is added to insure that the sample statistics are consistent with the scale transition equation of (4.12). Note that a different set of projected replicates is obtained at at $s\gamma$ from each of the q children of $s\gamma$.

The updated replicates at s, the projected replicates at $s\gamma$, and $F(s)$ are used to compute a smoothing gain $J'(s)$ as follows:

$$J'(s) \quad = \quad \widehat{Cov}[\chi(s)|s]F^T(s)\widehat{Cov}^{-1}[\chi(s\gamma)|s] \tag{4.42}$$

This gain is then used to derive the smoothed replicates at s:

$$\chi^j(s|S) \quad = \quad \chi^j(s|s) \qquad ; s = m(s) = 0 \tag{4.43}$$

$$\chi^j(s|S) \;=\; \chi^j(s|s) + J'(s)[\chi^j(s\gamma|S) - \chi^j(s\gamma|s)] \;;\; m(s) > 0 \qquad (4.44)$$

The adjustment at s to the replicate obtained from the upward sweep is proportional to the difference between the projected replicate $\chi^j(s\gamma|s)$ and the smoothed replicate $\chi^j(s\gamma|S)$ at $s\gamma$. This difference reflects the new information available from measurements at nodes that are not descendants of s. Note that $\chi^j(s\gamma|S)$ is available from the previous iteration of the downward sweep.

Computational Complexity

The complexity of the multiscale update depends on both tree identification (which must generally be performed at every update time) and the measurement update. The cost of model identification can be decomposed into two parts: 1) computation of the local covariances needed for the singular value decomposition and 2) the singular value decomposition proper.

In order to obtain some order of magnitude estimates of computational complexity suppose that the state dimension at every node is d, every state is measured, each parent has q children, and there are $M+1$ scales. At tree model identification step, the complexity of the local covariance calculation is $\mathcal{O}(qh(d^2N))$ for each parent node. The complexity for the singular value decomposition is $\mathcal{O}(qd^3)$ for each parent node. So the total complexity for identifying the entire tree is $\mathcal{O}(q^M(hd^2N + d^3))$. At the update step, the complexity for the upward sweep, which includes the Kalman gain calculation, is $\mathcal{O}(q^M(d^2N + d^3))$, while the complexity for the downward sweep is $\mathcal{O}(q^M(d^2N + d^3))$. Therefore, the overall complexity of the multiscale update scheme is $\mathcal{O}(q^M(hd^2N + d^3))$. The memory requirement for the update scheme is as much as $2q^M$ times of the storage requirement for one finest scale node.

For comparison, the traditional ensemble Kalman filter has complexity of $\mathcal{O}(q^{3M}d^3)$. This is much more expensive than the multiscale update, whose total complexity is on the order of $\mathcal{O}(q^{M+2}d^3)$ if we make the reasonable assumptions that $N = d$ and $h = q^2$.

4.5 Application of the EnMSF to Navier-Stokes Equations

4.5.1 State Equation: Navier-Stokes Equations

To demonstrate the performance of the ensemble multiscale filter, Navier-Stokes equation for a 2-D incompressible viscous fluid with constant density is used. For such a fluid the evolution of velocity field $[u\ v]^T$, pressure field p defined at location (x, y) and time t on some domain Ω follow

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + \eta(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) \tag{4.45}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \eta(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}) \tag{4.46}$$

These equations are highly nonlinear, so it's hard to be used as the state equation for the extended Kalman filter. However, the ensemble multiscale filter is capable of dealing with such high nonlinearity since it doesn't require linearization of the state equation. Also, the boundary layers and vorticity generation near solid boundaries governed by Navier-Stokes equations can give features at a wide range of spatial scales, which offers an advantage for using multiscale filters. The equations can be solved by Gerris [80], which is an Open Source Free Software library for the solution of the partial differential equations describing fluid flow. Gerris uses adaptive mesh refinement method to discretize the spatial domain, adapted to the the spatial scale and temporal evolution of flow structure. Computational resources are only focused on those regions with small scale features rather than those region with large smooth features. It combines a quad/octree discretization, a projection method and a multilevel Poisson solver. Advection terms are discretized using the robust second-order upwind scheme and complex solid boundaries are treated through a Cartesian volume-of-fluid approach.

4.5.2 Experiment Description

In this section, the first experiment is to demonstrate the tree model's ability to approximate covariance matrix and remove sampling error. The second experiment is to show that the multiscale filter's application to a high dimensional problem.

In the following discussion, the Naiver-Stokes equation is dimensionless. A rectangular domain enclosed by four corners located at (-0.5,-0.5),(-0.5,0.5),(1.5,0.5),(1.5,-0.5) is chosen for study. There is a solid rectangular barrier centered at (-0.22,0) with width 0.14 and length 0.23. The left boundary of the domain is a type of Dirichlet boundary defined as

$$u(0,y,t) = \begin{cases} 5 \text{ if } y \in [c(t)-0.12, c(t)+0.12] \\ 0 \text{ otherwise} \end{cases} \tag{4.47}$$

$$v(0,y,t) = 0 \tag{4.48}$$

where $c(t) = 0.9c(t-1) + w(t)$, $w(t)$ is white noise with zero mean and standard deviation 0.04. The right boundary is Neumann boundary with zero u and v gradient. The top and bottom boundaries are solid walls with slip conditions, i.e. $\frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} = 0$. Initial condition is zero u and v everywhere. The viscosity parameter η arises from numerical viscosity implicit in the solver.

4.5.3 Tree Representation of Sampling Covariance

Since the performance of the EnMSF depends on the realized tree representation of the covariance matrix, it is important to check the identified tree first. In order to compare the tree representation of a sampling covariance matrix against the true covariance matrix, the domain is discretized into a coarse 64 × 128 grid to allow for large ensemble calculation. At each time step u, v at the 64 × 128 grid points are derived from those over the irregular grid generated by the adaptive mesh refinement method in Gerris.

As stated in the previous sections, statistics approximation of the predicted ensemble using a tree model needs the following major parameters: the size of state vector at each finest scale node d_M, scaling factor q, neighbor screening size h, and state dimension d at coarse scales. The quality of the identified tree model and estimation largely depend on these parameters. As [37] pointed out, the state size at the finest scale should be large enough to reduce the difference between the realized covariance matrix and the original covariance matrix. In this experiment, state dimension of each node at the finest scale is set to 32, which corresponds to a 4x4 block. The used scaling factor $q = 4$ for all the non-root nodes, q=2 for the root node. The neighbor screening size h is set to 4 neighboring nodes. At different scales, the actual spatial length corresponding to the neighbor size is different.

For the 64 × 128 domain, using the tree model representation with parameters specified above, the total scale number 6. On the tree, the domain size at the finest scale becomes 16 × 32. The state dimension d of a node at coarse scales is 11. For comparison, a 6240 replicate ensemble is created to calculate the "true" correlation between pixel A at (8,8) and all the pixels at $t = 0.84$ (pixel (1,1) is the upper left corner of the domain). Then an ensemble with 52 members is run to provide an ensemble snapshot also at $t = 0.84$, which is used for the multiscale tree identification. On the tree, the correlation between this pixel and all the pixels in the domain is calculated from the realized tree using equation (4.25). The same procedure is repeated for pixel B at (88,56) at $t = 0.84$. The results are plotted in Figure 4-2 and Figure 4-3.

For pixel A, since it's close to the jet input disturbance at the left boundary, the spatial patterns of u and v tend to be smaller than those for pixels far away from the left boundary, such as pixel B. Thus, the spatial correlation length in u and v is relatively smaller than that for pixel B as shown in Figure 4-2 and Figure 4-3. In both figures, comparing to the true correlation as shown in the third rows, there exist some high frequency noise with relatively large magnitude in the first rows. These noise are sampling error due to inadequate ensemble size. However, in the second rows the tree representation of the ensemble reduces these spurious high frequency noise and keep large scale correlation in the prior ensemble, which is especially obvious for the v component. For u component, the tree model eliminates some sampling error but also introduces a little large scale structure error. This is consistent with the tree model identification procedure since only high frequency components are thrown away which helps reduce small scale error. Also, the realized correlations from the tree model bears error due to the selected tree structure, as is evident in the tree realized correlation of u as in Figure 4-3. However, the general spatial correlation pattern still follows the truth. The cross correlation between u and v at the same two pixels are shown in Figure 4-4 and 4-5. These cross correlation figures have the similar feature as discussed for the correlation figures. In Figure 4-4 the tree model can only pick up big features in the sampling covariance, which is different than the true one especially around point A and B.

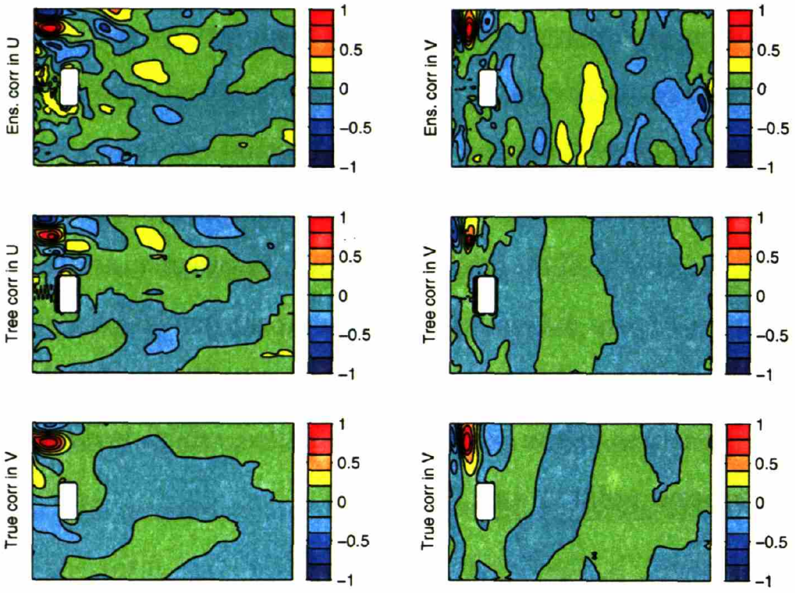


Figure 4-2: u and v correlation coefficients ρ between point (8,8) and all the pixels over a 64×128 domain at $t = 0.84$. First row: sampling ρ from an ensemble with 52 replicates; Second row: the realized ρ by the tree model using the same ensemble as in the first row; Third row: true ρ from an ensemble with 6240 replicates

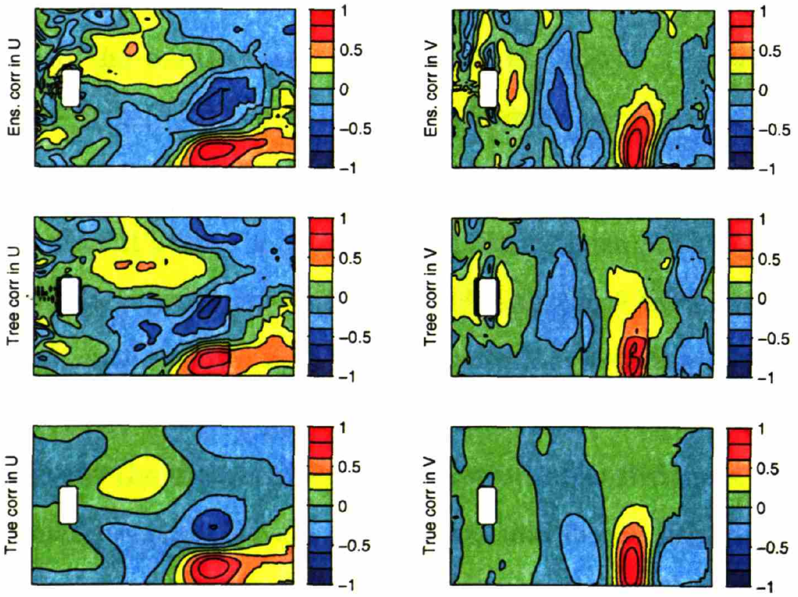


Figure 4-3: u and v correlation coefficients between point (88,56) and all the pixels. Configurations are the same as in Figure 4-2

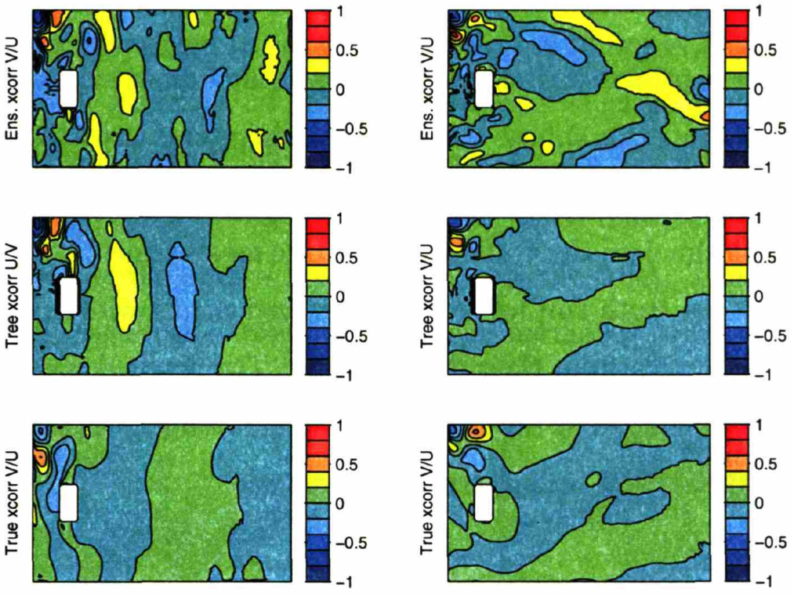


Figure 4-4: u and v cross correlation coefficients ρ_x between point (80,56) and all the pixels over a 64×128 domain at $t = 0.84$. First row: sampling ρ_x from an ensemble with 52 replicates; Second row: the realized ρ_x by the tree model using the same ensemble as in the first row; Third row: true ρ from an ensemble with 6240 replicates

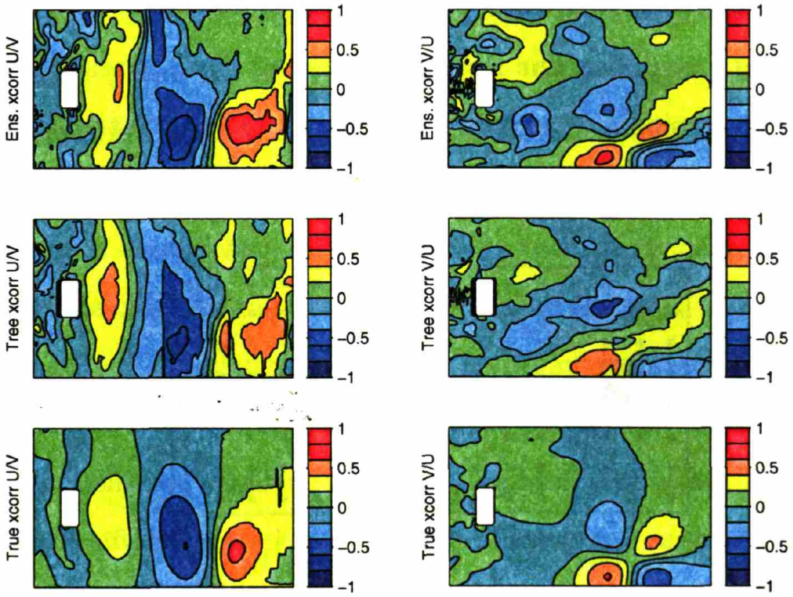


Figure 4-5: u and v cross correlation coefficients ρ_x between point (88,56) and all the pixels. Configurations are the same as in Figure 4-4

4.5.4 High Dimensional Estimation

The multiscale filter is primarily designed for efficient high dimensional nonlinear estimation. To demonstrate this ability, the domain is discretized into a 512 × 1024 grid, with 1048576 state variables. For experimental convenience, controlled experiment approach is taken to simulate the truth and synthetic measurements.

First conducted is a truth run using the discretized model with one replicate of boundary conditions drawn from their distributions. The resulted u, v are regarded as the truth, which is observed at the specified scattered locations as shown in Figure 4-6 every time interval of 0.21, using measurement model

$$y(t) = u(t) + e(t) \tag{4.49}$$

where $e(t)$ is a vector with uncorrelated elements, each of which follows Gaussian distribution with zero mean and standard deviation 1.5. Then the simulated measurements are assimilated with the multiscale filter. In this experiment, d_M, the state dimension of each node at the finest scale is set to 32, which corresponds to a 4×4 block with 2 states in each pixel. The scaling factor $q = 4$ for all the non-root nodes, $q = 2$ for the root node. The neighbor screening size h is set to 4 neighboring nodes. The state dimension d of a coarse scale node is 16. The estimation results are compared against the simulated truth to assess the filter's performance.

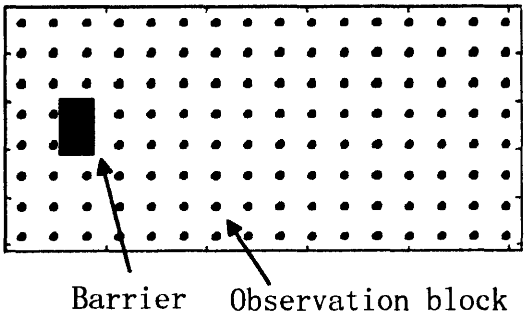


Figure 4-6: Experiment Domain and Observation Locations

For pixel-wise check, only the ensemble of u time series for pixel (276,472) and

(440, 296) are plotted in Figure 4-7. The ensemble mean and truth are also plotted. Since the first pixel is within the observation block as shown in Figure 4-6, measurement would provide sufficient information to adjust the ensemble towards the truth at each measurement time. While the second pixel is not measured, the update only comes through correlation between this pixel and those directly observed pixels. At $t = 0.63$ and $t = 0.84$ the update is even away from the truth. This might be due to the residual sampling error or tree model error or normality assumption in the prior.

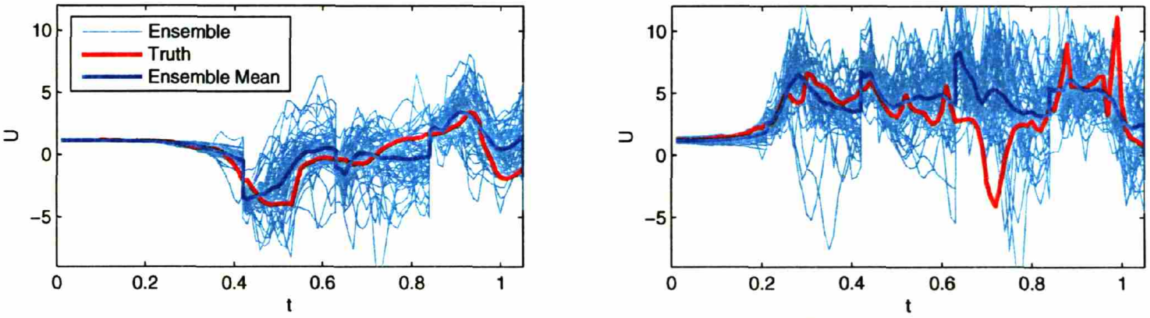


Figure 4-7: Filtered replicates, ensemble mean, and truth u time series at pixel (276,472) (left) and (440, 296) (right)

At every measurement time, the spatial distribution of mean and standard deviation of u and v are plotted in Figure 4-8 and 4-9. Before each update, the ensemble mean and standard deviation reflects all the information in the model prediction and all the previous measurements. After each update, the latest information from the new measurement is assimilated. We can see the estimated mean of u and v after each update become closer to the truth and the uncertainties in the estimate as indicated by the standard deviation plots in the 4th and 5yh column of these two figures also become smaller. It is also noticeable that for the directly measured blocks, the uncertainty reductions are smaller than those unobserved pixels. The blockiness in the ensemble mean after update is due to the tree structure error in representing the covariance. It can be solved using overlapping tree as discussed in [58]. While the blockiness in the standard deviation after update is due to the scattered measurement pattern. To look at the errors in the ensemble mean closer, the time series of spatial RMSE is calculated between the ensemble mean and truth for both u and v, which is plotted in Figure 4-10. The merit of updating u using data is obvious in the left panel. But for v, the unobserved state variable, although the spatial patterns shown in Figure 4-9 become closer to the truth after update, for the first three updates the benefits from data are not as obvious if using RMSE as a metric. Some investigation is needed in this aspect. Methods like field alignment might be helpful.

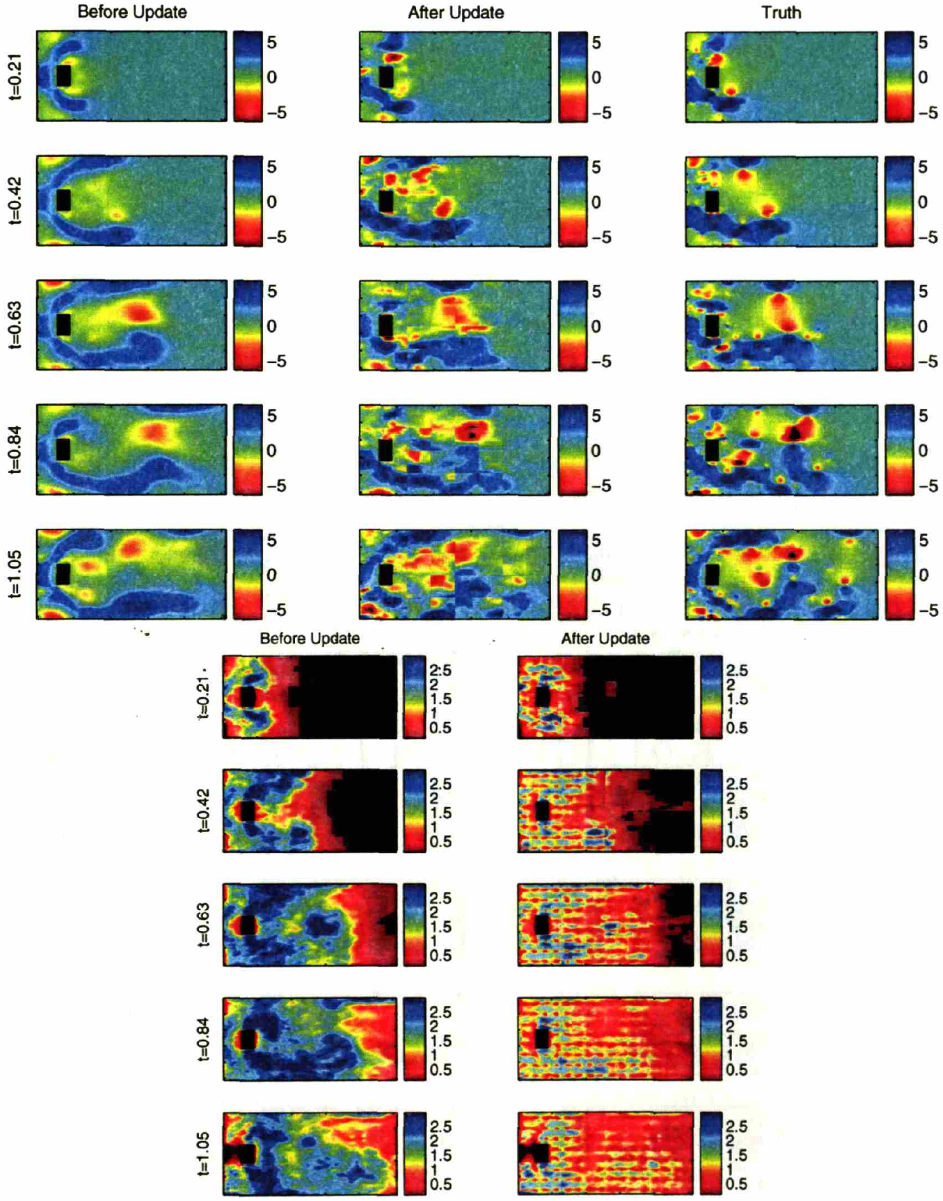


Figure 4-8: Ensemble mean of u before (1st column) and after (2nd column) update at measurement times, and the corresponding truth (3rd column); Also plotted are the ensemble standard deviation of u before (4th column) and after (5th column) update

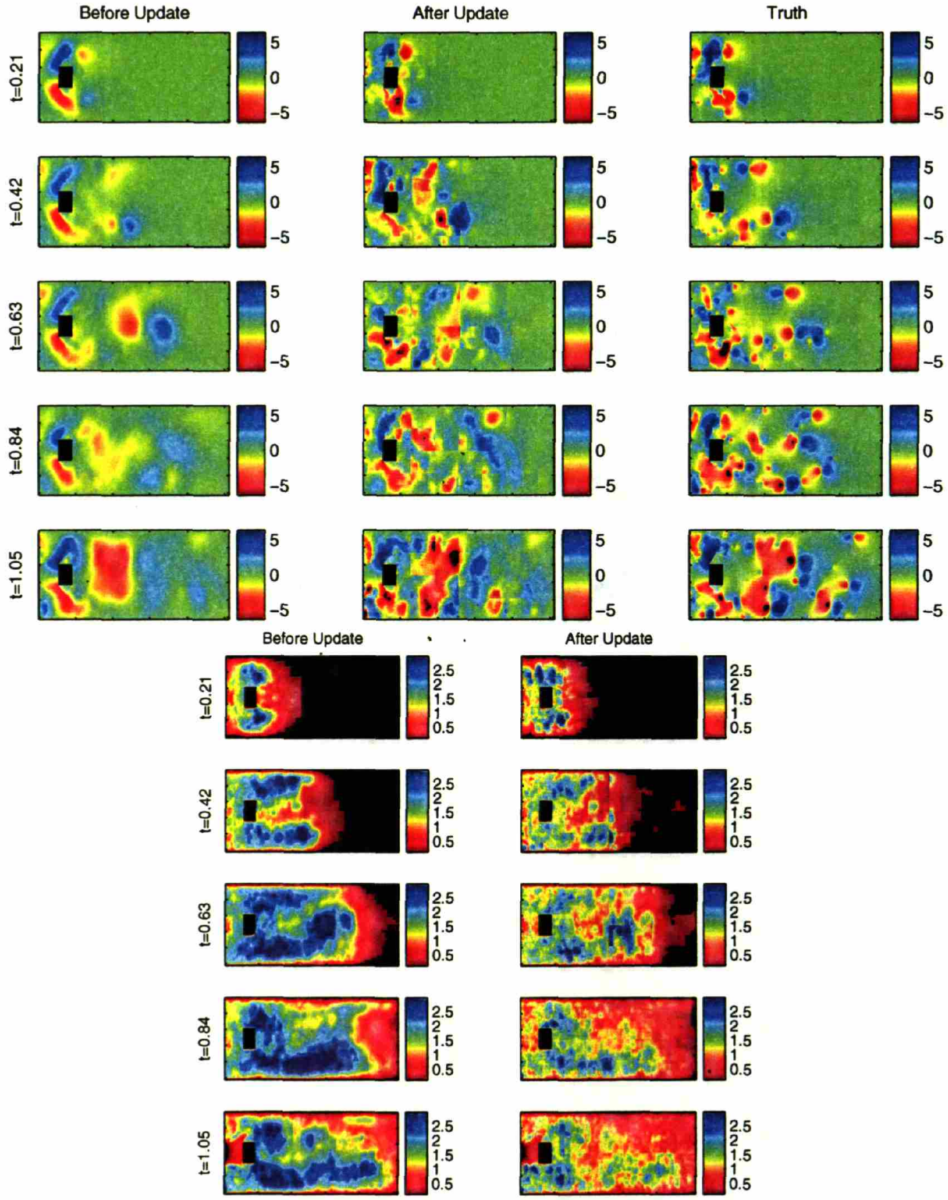


Figure 4-9: Ensemble mean of v before (1st column) and after (2nd column) update at measurement times, and the corresponding truth (3rd column); Also plotted are the ensemble standard deviation of v before (4th column) and after (5th column) update; Please note that v is unobserved and got updated through its correlation with u

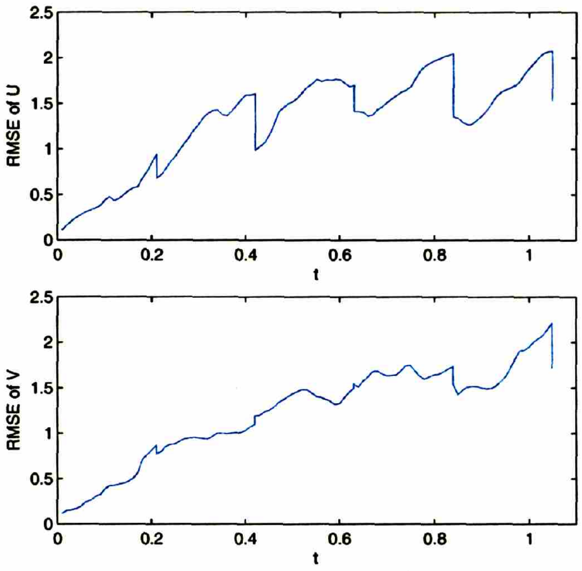


Figure 4-10: RMSE of the ensemble mean of u and v with respect to the truth over the entire domain; Please note that v is unobserved and got updated through its correlation with u

4.6 Application of the EnMSF to a Nonlinear Diffusion System

4.6.1 Nonlinear Diffusion Model with Random Boundary Conditions

The other dynamic system used to show the performance of EnMSF is a 2-D nonlinear diffusion equation of $\theta(x, y, t)$:

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial x}D(\theta)\frac{\partial\theta}{\partial x} + \frac{\partial}{\partial y}D(\theta)\frac{\partial\theta}{\partial y} + \frac{\partial K(\theta)}{\partial y} + w(t) \tag{4.50}$$

in which y is positive upward, $w(t)$ is Wiener process, and

$$K(\theta) = 2.44(\frac{\theta}{\theta_s})^{7.52} \tag{4.51}$$

$$D(\theta) = 0.0814\frac{\theta_s^{2.26}}{\theta^{3.26}}K(\theta) \tag{4.52}$$

θ_s is a static parameter and uniformly distributed between [0.422 0.443] and $0 \leq \theta \leq \theta_s$. The initial condition of $\theta(x, y, 0)$ follows Gaussian distribution $N(0.3, 0.02)$. The boundary condition is random and specified as for $i = 1, 2$ and $j = 1, 2, 3$,

$$\frac{\partial\theta(x,0,t)}{\partial y} = \begin{cases} P_i, & \text{if } x \in [x_{i,j}\ x_{i,j} + \frac{L}{4}] \text{ and } t \in [t_{i,j}\ t_{i,j} + d_{i,j}]; \\ 0, & \text{Otherwise.} \end{cases} \tag{4.53}$$

$$\frac{\partial\theta(x,L,t)}{\partial y} = \frac{\partial\theta(0,y,t)}{\partial x} = \frac{\partial\theta(L,y,t)}{\partial x} = 0 \tag{4.54}$$

where L is the width of the domain, P_i is uniformly distributed between [1.5 1.8], $x_{i,j}$ is uniformly distributed between $[0\ \frac{3}{4}L]$, $t_{i,j}$ is uniformly distributed between $[(j-1)T\ (j-1)T+DT]$, and $d_{i,j}$ is uniformly distributed between $[S_1\ S_2]$. T, DT, S_1, S_2 are specified parameters. The reason to use such boundary conditions is to create random fields with dynamically changing covariance structure. The non-Gaussianity of θ results from the nonlinear effects of parameters $K(\theta)$ and $D(\theta)$ in equation (4.50).

Two sets of the synthetic experiments similar to that in the last section are conducted. For both cases, diffusion equation (4.50) is explicitly discretized. The time step is set to 0.0021 and spatial step is set to 0.1 in both x and y directions. The first experiment has a 32×32 domain with $L = 3.2$ and the state dimension of 1024 which allows enough replicates to compare the EnMSF solution against the EnKF solution without worrying about the sampling errors, while the second one has a 512×512 domain with $L = 51.2$ and the state dimension of 262144 which is to show EnMSF can handle large scale problems.

For both cases, first conducted is a truth run using the discretized model with one replicate of θ_s, initial, and boundary conditions drawn from their distributions. The resulted $\theta(x, y, t)$ is regarded as the truth, which is observed at all the pixels every T steps using measurement model

$$y(t) = \theta(t) + e(t) \tag{4.55}$$

where $e(t)$ is white Gaussian noise vector with each element following $N(0, 0.05)$.

For the 32×32 domain, parameters in (4.53) are set as $T = 0.21, DT = 0.042, S_1 = 4, S_2 = 5$. For the 512×512 domain, $T = 1.26, DT = 0.42, S_1 = 40, S_2 = 50$. And the dimension of $y(t)$ is 1024 for the 32×32 domain and 262144 for the 512×512 domain at each measurement time.

4.6.2 Tree Model Performance

In the following two experiments, state dimension of each node at scale M is set to 16, which corresponds to a 4×4 grid at scale M. The scaling factor $q = 4$, which means the tree is a quadtree; The neighbor screening size h is 4 neighboring nodes.

For the 32×32 domain, using the tree model representation with parameters specified above, the total scale number is 4. The domain size at scale M is 8×8 with finest scale state dimension of 16. The state dimension d at coarse scales is set to 7. To examine how good the tree model is at representing the prior ensemble, the realized correlation coefficient ρ between point (17,17) and all the other points at two typical times are plotted in figure (4-11). Column (a) shows the sampling ρ with an ensemble of 40 replicates. Column (b) shows the true ρ with an ensemble size of 10000. Column (c) shows the realized ρ using a tree model for the same ensemble used in column (a). At $t = 0.42$, the spectrum is pretty sharp and long range correlation exists due to the apparent θ front. At $t = 1.26$, the spectrum of the covariance is flatter than at time $t = 0.42$, and variance is more evenly distributed among the modes. At both times, column (a) is much noisier than column (b). The high frequency noise are all due to sampling error. However, the tree representation of the ensemble obviously reduces the high frequency noise and the enhanced spatial correlation structure becomes closer to the true correlation structure as in column (b).

For a tree representation of the covariance matrix, the state dimension at the coarse scales determines how much variance is retained to decorrelate children. The percentage of the variance retained in a parent with a fixed state dimension at coarse scale may explain how the dynamics is changing and provide guidelines for controlling the tree structure. Also, by changing the state dimension, thus the percentage of the

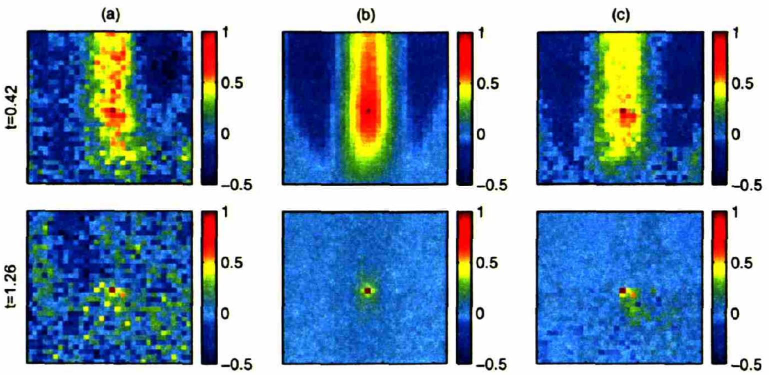


Figure 4-11: (a) Sampling correlation coefficient ρ between point (17,17) and the entire domain, from an ensemble with 40 replicates , (b) true ρ from an ensemble with 10000 replicates, and (c) the realized ρ by the tree model for a 32×32 domain at $t = 0.42$ and $t = 1.26$

retained variance, one can control the retained magnitude of the sampling error. Define the fraction of the retained variance in one parent as

$$\alpha = \frac{\sum_{l=1}^{q} \sum_{m=1}^{d(l)} \lambda_{l,m}}{\sum_{l=1}^{q} \sum_{m=1}^{n(l)} \lambda_{l,m}} \tag{4.56}$$

where $\lambda_{l,m}$ is the m^{th} diagonal elements of the eigenvalue matrix of (4.27) for the l^{th} child of this parent node. If $\alpha = 1$ then the parent node can completely decorrelate its children. If all the parents has $\alpha = 1$, then the tree model can exactly reproduce the covariance matrix. Since $rd(l) \leq n(l)$, α is usually less than one. For an ensemble with infinite replicates, we want α to be as close to 1 as possible for each parent node. But for an ensemble with limited replicates, α closer to 1 is not a good choice. In this case, an lower α can truncate the modes with small variance, which are usually the modes reflecting high variability in space.

To illustrate the points, two groups of lists showing α for each node at scales $s = 2, 1, 0$ for ensembles used in column (a) of Figure 4-11 are given below. The left list is for $t = 0.42$ and the right for $t = 1.26$:

$$\left[\underbrace{\begin{matrix}0.80 & 0.76 & 0.74 & 0.76\\ 0.75 & 0.72 & 0.72 & 0.72\\ 0.64 & 0.66 & 0.65 & 0.60\\ 0.48 & 0.49 & 0.47 & 0.48\end{matrix}}_{s=2}\ \underbrace{\begin{matrix}0.80 & 0.80\\ 0.77 & 0.76\end{matrix}}_{s=1}\ \underbrace{0.82}_{s=0}\right]$$

$$\left[\underbrace{\begin{matrix}0.32 & 0.25 & 0.23 & 0.29\\ 0.28 & 0.20 & 0.21 & 0.24\\ 0.23 & 0.20 & 0.19 & 0.22\\ 0.25 & 0.23 & 0.22 & 0.21\end{matrix}}_{s=2}\ \underbrace{\begin{matrix}0.50 & 0.53\\ 0.48 & 0.55\end{matrix}}_{s=1}\ \underbrace{0.65}_{s=0}\right]$$

At $t = 0.42$, at scale $s = 2$, the nodes corresponding to the top two rows in the left list keep more variance than those at the bottom two rows. The reason is that for the top two rows there exists more spatial correlation so the the spectrum of the covariance matrix is sharper thus the high modes thrown away don't contain much variance. While for the bottom two rows which correspond to the low half part of the domain, they are still under the influence of white noise. So the spectrum is flatter, and the high modes thrown away contain much variance. For the fixed state dimension of 7, α is then different depending on the dynamics. By checking the fraction we have a sense of how fast the front is moving. At scale $s = 1$ and $s = 0$ the retained variance is about 80%, which is good for the tree model to be able to describe long range correlation as is shown in the first row of Figure 4-11. While at $t = 1.26$, the retained variance is only about 20~30% at scale $s = 2$ as shown in the right list. This is because very little fine scale correlation exists. α at scale $s = 1$ and $s = 0$ are also low compared to those at $t = 0.42$, which means the large scale correlation is also very little. In both cases, if the state dimension is increased, then the retained variance would explain more sampling error which are revealed as the noise in Figure 4-11.

For the 32×32 domain, the tree model approximate the predicted ensemble very well and reduce the sampling error at the same time. For the much larger 512×512

domain, the tree performance is also examined. Using q=4 and the state dimension of 16 at the finest scale, the tree has the total scale number of 8. The domain size at the finest scale on the tree is 128 × 128. The coarse scale state dimension d is set to 10. From an ensemble with 80 replicates at $t = 1.26$, the sampling correlation coefficient ρ between pixel (128, 320) and all the other pixels is plotted in column (a) of Figure 4-12. The realized ρ is plotted in column (b). It is obvious that even though the used state dimension at the finest scale is 16, it is still able to capture both short and long range correlation. The realized ρ is smoother than the original sampling correlation coefficient, and most of the the sampling error located below the θ front is removed.

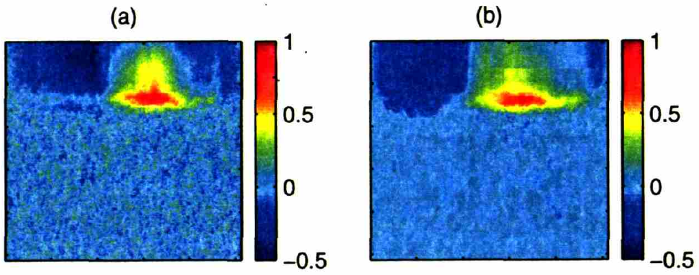


Figure 4-12: (a) Sampling correlation coefficient ρ between pixel (128,320) and the entire domain, from an ensemble with 80 replicates at $t = 1.26$, and (b) the realized ρ by the tree model for a 512 × 512 domain at $t = 1.26$

4.6.3 Data Assimilation Experiments

To ensure the EnMSF update is correct, EnMSF and EnKF are both applied to the 32x32 domain. An ensemble of 10000 replicates are used in this experiment, for which the sampling error can be ignored. All the tree model parameters are the same as those specified in the preceding section. Every 100 steps, the simulated observation is assimilated with both filters. The ensemble mean of the updated θ from EnMSF and EnKF at $t = 0.42$ and $t = 1.26$ compared to the truth are shown in Figure 4-13. The two filters almost give the same estimate at both times even though the spatial correlation for $t = 1.26$ is lesser than at $t = 0.42$. The resulting RMSE time series of the ensemble mean with respect to the truth are plotted in Figure 4-14. The black

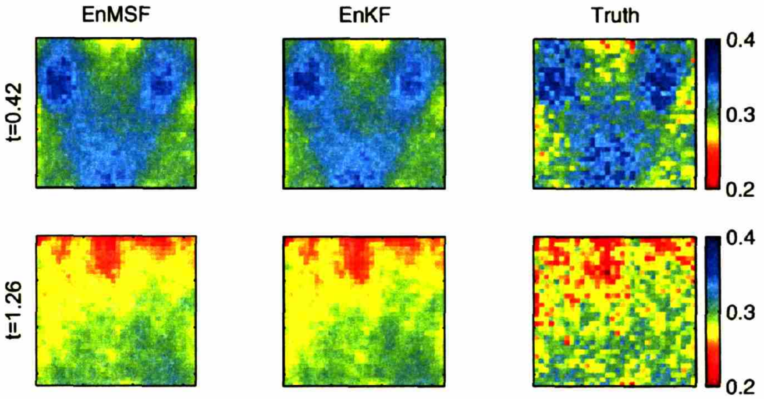


Figure 4-13: Ensemble mean of the updated θ from EnMSF and EnKF compared to the truth at $t = 0.42$ and 1.26

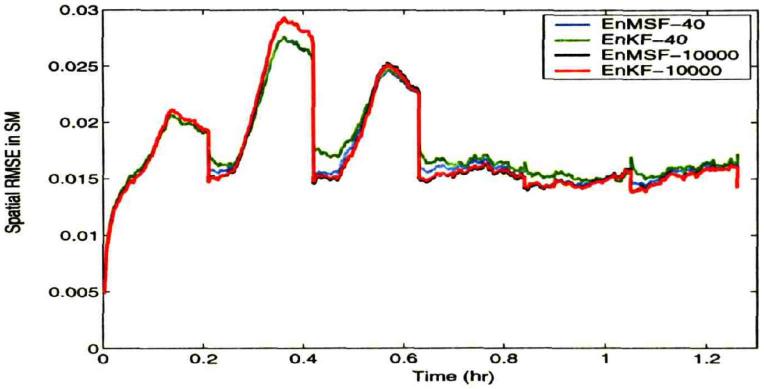


Figure 4-14: RMSE time series of the ensemble mean of the updated θ with respect to the truth for EnMSF and EnKF, 40 or 10000 replicates are used

line (EnMSF) and red line (EnKF) are almost indistinguishable from each other.

Since the tree representation of the state ensemble can capture the correlation structure and remove part of the sampling error for small ensemble size as shown in the preceding section, it is expected to perform well for small ensemble size. An ensemble of 40 replicates are then used to test the EnMSF performance for the 32×32 domain. The used truth and measurements are exactly the same as the one used for the large ensemble. The resulting RMSE time series of the ensemble mean with respect to the truth is plotted in Figure 4-14. After each update time, the estimated θ from EnMSF always performs better than that given by EnKF. At $t = 0.84, 1.05$, and 1.26, the RMSE from EnKF estimate even increases because of the sampling error. While for EnMSF estimate, the RMSE is always closer to the black line which has the least RMSE one can get using EnKF or EnMSF.

4.7 Discussion and Conclusions

A new ensemble multiscale filter is proposed to solve large scale nonlinear estimation problem. The general prediction and update two stages of EnMSF are similar to EnKF. Some replicates of the state variables are propagated using a fine resolution dynamic model to provide prior information. Any nonlinear dynamic model and parameter uncertainties model can be used at this step to generate a reasonable ensemble. All the statistical information required for prediction and updating are embedded in the generated ensemble, but only second order moments are used in the filter. When observation is available, multiscale update is used to assimilate all the available (multiresolution) data. At update times, first it uses Predictive Efficiency method to identify a multiscale tree to approximate the full covariance matrix given by the propagated ensemble. The specified tree model parameters d_M, h, d, q control the degree of approximation and sampling error. Then upward and downward sweep update is performed on the identified tree using the available data. Both model identification and two sweep update are based on the ensemble statistics. The estimation on the tree ensemble is helped with the scale recursive measurement prediction and the tree model internality. The performance from this new filter mainly depend on quality of the identified tree.

The application of the EnMSF to Navier-Stokes and nonlinear diffusion equations show that the EnMSF can eliminate some high frequency sampling error. The reason is that high frequency noise are ignored at the tree identification step. Also, local approximation of a covariance matrix in the tree identification step usually has full rank at any node on the tree, thus reducing the sampling error. A crucial advantage of tree model is that the local operations on the full covariance matrix is still able to capture the large scale correlation feature. So the applicability of the EnMSF to a dynamic system should be widely broad considering the ability of tree model to represent any covariance even if using limited state dimension at coarse scale nodes.

Based on the tree model ability to represent the full covariance matrix and reduce sampling error, the multiscale update is especially appealing to large scale filtering problem since it has the complexity of $O(q^M(hd^2n + d^3))$. The computational efficiency mainly depends on the specified state dimension d and ensemble size. Since the tree identification is essentially local, in practice it is highly parallelizable to implement. In fact, the state dimension d at coarse scales does not have to be specified. It can be adaptively changed based on a specified value of percentage of the retained variance at each parent node. This way it is possible to use low dimensional coarse resolution states to decorrelate the children nodes. Whether the tree structure parameters should be specified or dynamically changed based on a fixed retained variance ratio should be application dependent.

Chapter 5

Application of the EnMSF to U.S. Great Plains Evapotranspiration Estimation

Evapotranspiration (ET) is a key factor that links water and energy transfer between land surface and atmosphere. ET provides the boundary condition for water fluxes from earth to atmosphere, so it is of crucial importance to weather forecast. ET is controlled by available energy, soïl water supply, turbulent transport condition, and vegetation characteristics. The energy control on ET is usually parameterized in the form of potential ET in most of the land surface models. Actual ET is a fraction of the potential ET due to soil moisture stress. Soil and vegetation controls are usually dependent on land cover characteristics such as the density of plant coverage, root distribution, roughness,etc. Many antecedent studies have given accurate ET estimation using surface meteorological and radiosonde observations. Most of these studies usually focus on scattered locations. As for the large scale ET, it has been difficult to develop models suitable for its prediction. This is partly because of the lack of direct measurement of large scale ET processes.

Currently, the often used two methods for calculating the regional evapotranspiration are the Priestley and Taylor method [81] and the complementary relationship of Bouchet [7], further developed by Morton [75]. Both methods are based on heuristic arguments, but they have often been applied successfully [63]. Additionally, using satellite remote sensing data is another very attractive approach as it enables a widespread area coverage and a higher repetition rate. However, direct measuring of ET through satellite remote sensing is hardly possible as the phase change of water molecules produces neither emission nor absorption of an electromagnetic signal. So large scale ET has to be derived by some indirect means. A possible approach is to use a physical model that links ET with other physical proxy quantities that are more easily to obtain using satellite. By updating the related quantities, ET can be driven close to the truth.

As an important component of the hydrological cycle, soil moisture can be such a candidate proxy quantity. It determines evapotranspiration, infiltration, and percolation. Changes in soil moisture storage are caused by precipitation, evapotranspiration, lateral flow or vertical drainage. Complicating factors influencing soil moisture include vertical and horizontal changes in soil composition, variations in drainage and slope of the terrain, spatial and temporal changes in vegetation, and intermittent precipitation. Soil moisture varies both temporally and spatially in response to many processes acting over a variety of scales. It has a memory effect of the forcing states, especially rainfall. In other words, soil moisture at a given location and at any given time reflects all the historical hydrological processes that have occurred. This soil moisture memory may have profound implications for long-term weather prediction through land-atmosphere feedback [66]. Also for data assimilation problems, the effects of the soil moisture memory on generating spatial error correlation have also been demonstrated in the last chapter. By exploiting the memory effects, it's expected that soil moisture observation would provide valuable information of the antecedent precipitation so as to enhance the ET prediction.

To test the utility of data assimilation techniques for understand the large scale hydrological processes, this chapter will attempt to use EnMSF to assimilate multi-sensor soil moisture measurement to obtain better soil moisture estimate and then use CLM to predict the evapotranspiration.

5.1 Experiment Description

5.1.1 Computational Domain

The spatial domain of interest is the U.S. Great Plains region, located between $(25.86^oN, 114.07^oW)$ and $(49.01^oN, 90.12^oW)$, delineated by the USGS hydrological unit boundaries as shown in Figure 5-1.The time domain of interest is chosen between June 1 GMT 0:00 and June 31 GMT 23:00 in 2004. The large scale spatial span allows examination of the scales effects of rainfall on soil moisture and evapotranspiration. The time window is mainly chosen to avoid snow and ice.

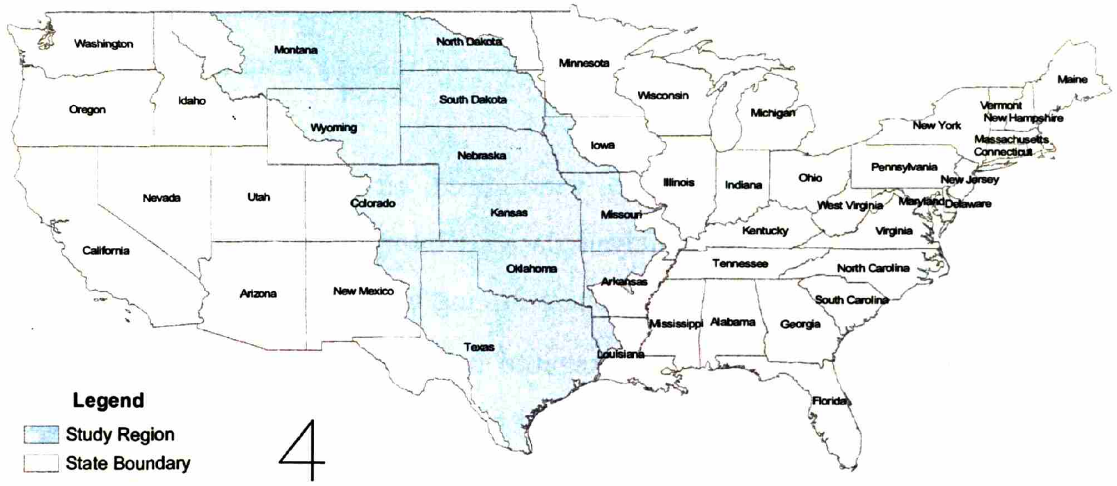


Figure 5-1: The Great Plains experiment region

5.1.2 Atmospheric Forcing and Land Surface Data

As a dominant forcing for land surface system, rainfall determines the top boundary of soil moisture transport. Temporally, the highs and lows of soil moisture reflects the antecedent rainfall history. Spatially, after rainfall events the spatial pattern of soil moisture may imply the cumulative rainfall spatial distribution. Due to rainfall intermittency, the soil moisture probability distribution would also bear some nonnormal features. Since data assimilation methods are built on quantification of various uncertainties, the rainfall error is critical in determining the data assimilation methods suitable for soil moisture estimation. As shown in the last chapter, the horizontal error correlation in the top layer soil moisture reveals dominant rainfall patters during wet periods. In chapter 2, a primitive rainfall model and wind velocity field demonstrated the rainfall error model influence on soil moisture spatial error correlation. The rain storm size, intensity, and moving direction are all recorded in soil and create anisotropic spatial error correlations. Realistic rainfall replicates generation is not a trivial task since it involves high dimensional measurement conditioning and strong non-Gaussianity. Chatdarong et al [15] developed a rainfall replicate generation system, conditioned on cloud measurements, ground station rain gauge measurements, remote sensing rainfall measurements from SSM/I, TRMM and AMSU, and a rainfall advection model. The replicates used in this thesis are directly from his work for the Great Plains region.

Other forcing such as solar radiation, wind speed, air temperature, and air humidity are from NCEP 6 hourly reanalysis data [1] interpolated to 0.5 degree grid. For simplicity there is no error added to these forcing variables.

The required land cover data are aggregated to the desired resolution (2km for truth generation/5km for data assimilation test) from the 1-km resolution global land cover characteristics database prepared by the U.S. Geological Survey, the University of Nebraska-Lincoln, and the European Commission's Joint Research Centre, as shown in Figure 5-2. This global land cover data set is based on 1-km Advanced Very High Resolution Radiometer (AVHRR) data spanning April 1992 through March 1993. The derived NDVI dataset during the same period is also used in remote sensing forward model.

The soil texture map comes from the USDA State Soil Geographic Database (STATSGO) as shown in Figure 5-3. At each pixel of the computational domain, the uncertainties in soil properties enter CLM in the form of random sand and clay fractions, which are generated by using MCMC mehods as discussed in Appendix D. There exists spatial error correlation between the sand/clay fractions at different pixels. A sample of the clay fraction is shown in Figure 5-4

The leaf area index (LAI), stem area index (SAI), vegetation canopy top and

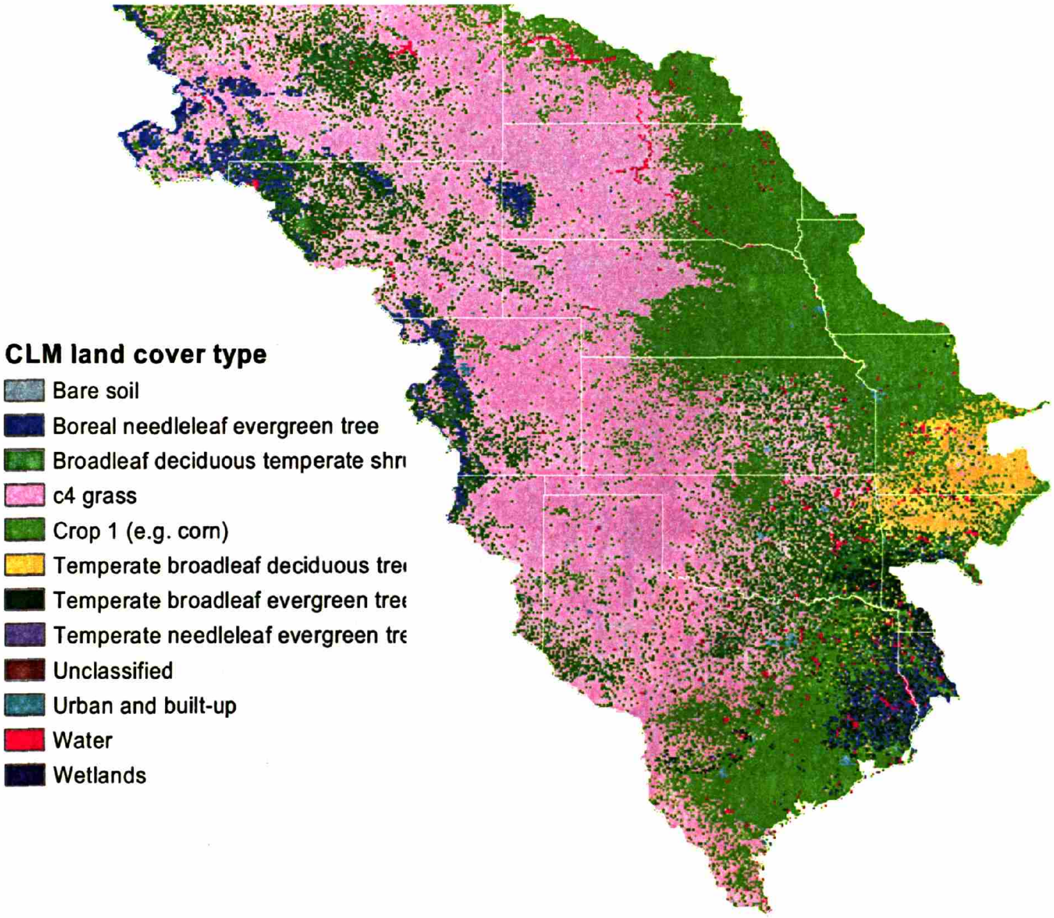

CLM land cover type
Bare soil
Boreal needleleaf evergreen tree
Broadleaf deciduous temperate shrı
c4 grass
Crop 1 (e.g. corn)
Temperate broadleaf deciduous tre
Temperate broadleaf evergreen tre
Temperate needleleaf evergreen tre
Unclassified
Urban and built-up
Water
Wetlands

Figure 5-2: The land cover type over the domain

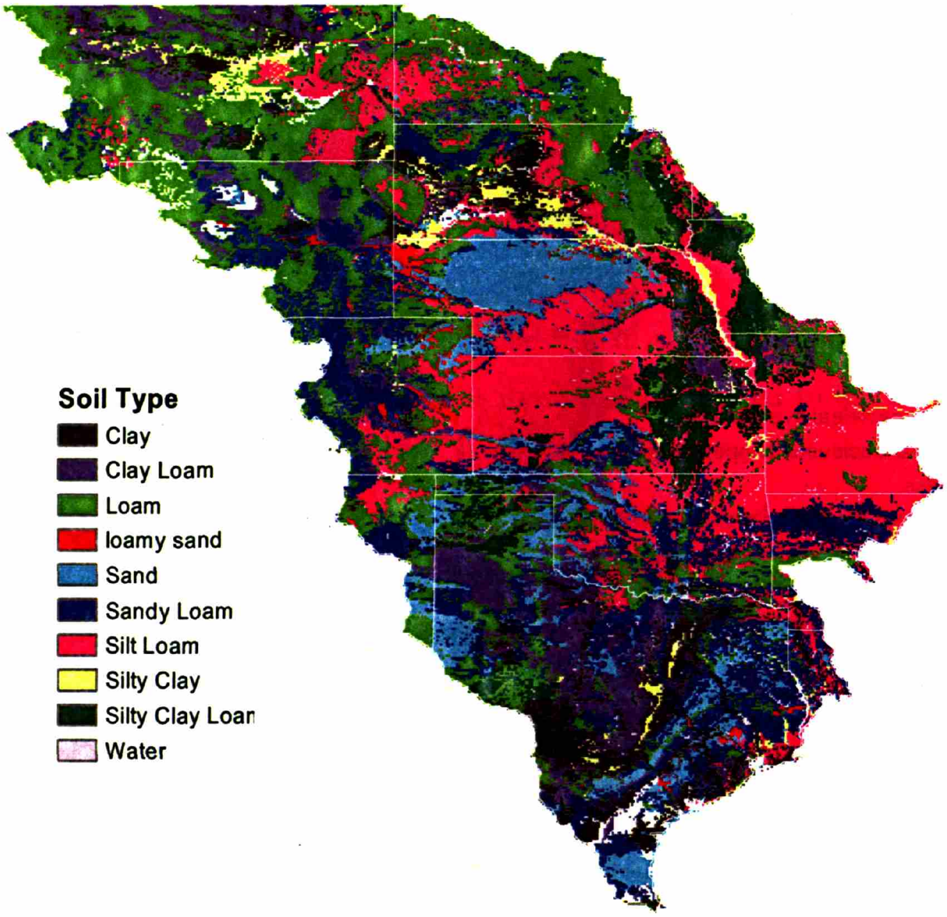


Figure 5-3: The soil type over the domain

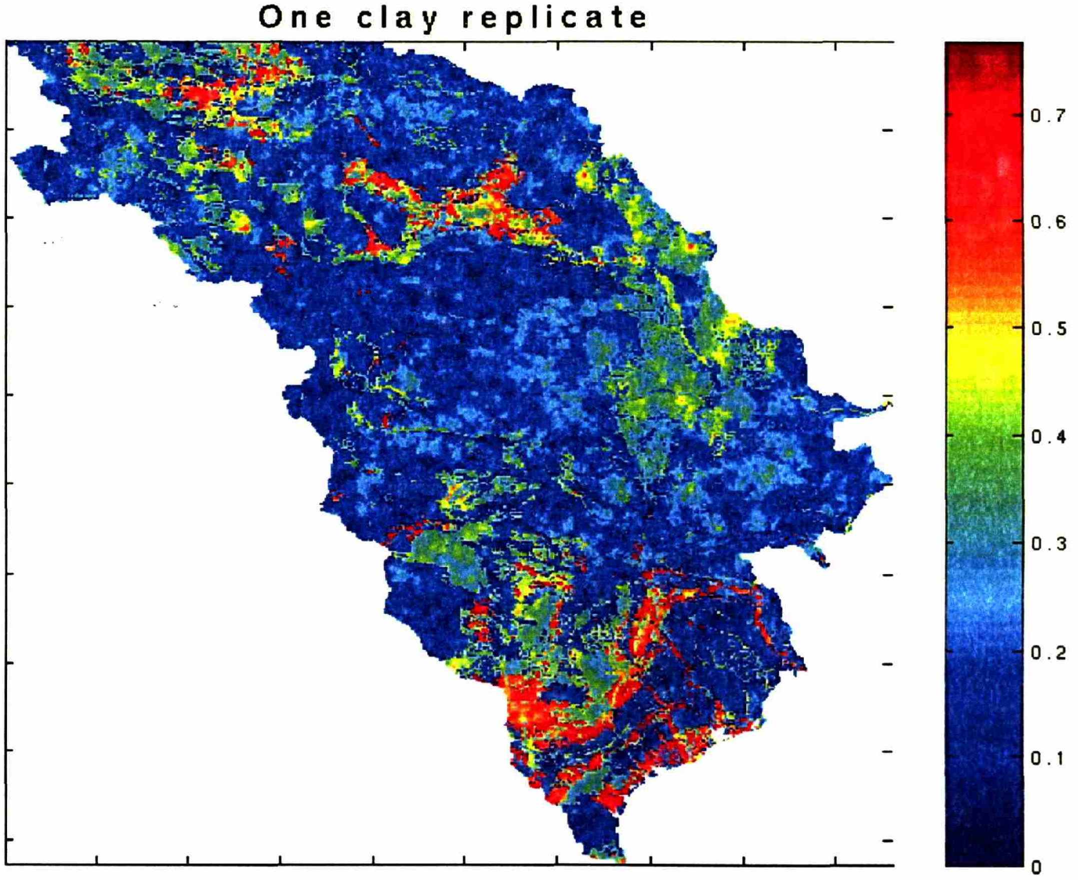


Figure 5-4: A replicate of clay fraction field generated by MCMC method.

bottom height required by CLM are obtained by looking up the tables used in NCAR Land Surface Model [5]. Each land cover has such a set of corresponding parameters. These variables are not perturbed. All the parameter tables are listed in Table E.2-E.4

5.2 Observation System Simulation Experiment

5.2.1 Synthetic Truth Run

To assess the performance of the EnMSF in large scale land surface applications, it's better to have a truth data to compare against. The synthetic experiment approach serves well for this purpose. The truth run provides all the true soil moisture, evapotranspiration, ground temperature, and so on. The time step size is set to 15 minutes for both the spin-up run and the truth run, and the spatial resolution of the grid pixel is chosen as 2km. The finer spatial and temporal resolution is regarded as the "true" physical resolution. The 5km land surface model resolution for the data assimilation experiments is coarser than the truth resolution so it can't capture certain fine scale processes, which is always the case for any land surface model. The scale disparity between model resolution and real physical processes would cause model error, which is hard to be quantified as simple additive error. In this chapter, the scale error is not explicitly treated. So in some sense this chapter is also a test of the tolerance of the EnMSF to error in the model errors, which is ubiquitous in any data assimilation experiment.

To make the data assimilation more realistic, the true rainfall is also carefully chosen so that it is not a direct rainfall replicate given by Chatdarong's model in [15]. This is equivalent to assuming the rainfall input error in the assimilation experiments is not perfectly known. WSI-NOWRAD dataset between June 1 and June 30 2004 [36] is used as rainfall input for the truth. NOWRAD is a radar image product generated by Weather Services International Corporation (WSI) of Andover, M.A., U.S.A. Using the WSR-88D Next-Generation radars that form the NEXRAD mosaic, NOWRAD algorithms remove ground clutter, anomalous propagation and other radar-induced artifacts that mar data validity to produce a reflectivity product of 15-min temporal and roughly 2 km spatial resolution covering almost the entire continental U.S. A combination of automated modification and manual adjustments by trained radar meteorologists at WSI serve to enhance the raw NEXRAD imagery [36]. Further discussion of WSI-NOWRAD data can be also found in [47]. All the other forcing variables are interpolated from NCEP reanalysis dataset [1]. Land cover data is resampled from the 1km USGS global dataset using nearest neighbor approach. Soil texture is also interpolated from STATSGO dataset. No perturbation is added to all these data. To generate the initial condition, the same set of forcing and land surface data is used to spin up CLM started with volumetric soil moisture profile [0.35 0.35 0.35 0.35 0.35 0.35 0.31 0.31 0.31 0.31], soil temperature profile [304 302 293 290 290 288 288 288]K, vegetation canopy temperature 305K for 3 months. The resulting soil moisture and temperature and canopy temperature are then used as the initial condition for the truth run.

5.2.2 Measurement Models

As a well studied means of observing large scale soil moisture, microwave remote sensing has many advantages. Microwave measurement is not affected by cloud coverage and variable surface solar illumination. In the absence of dense vegetation cover, soil moisture has the dominant effect on the received signal [76]. For soil moisture measurement, the best microwave frequency is between 1-3Ghz, since it has reduced cloud attenuation and larger vegetation penetration capacity.

L band remote sensing instruments have already been used to collect data in field experiments such as SGP97, SGP99, SMEX01, and the followups and been proved promising in retrieving soil moisture. In the near future, L band soil moisture remote sensing data from HYDROS [32] will be available. That will be the best data available for large scale soil moisture estimation.

Categorizing by sensor type, there are two kinds of microwave instruments: one is passive sensors and the other is active sensors. Passive soil moisture sensor (Radiometer) usually has a coarse resolution for space-borne systems, and it has greater sensitivity to soil moisture in vegetated regions. While active sensor (Radar) has higher measurement resolution but lower sensitivity to soil moisture in vegetated regions due to vegetation volume scattering. It is likely to combine these two sets of data to a moderate resolution appropriate for weather forecasting and hydrological application with low uncertainties. Both sensor will be used in this study to demonstrate the advantage of using the multiscale filter to fuse the multiresolution data.

1). Radiometer Models

The radiometer models are slightly different from those in chapter 2 and 3. Here the models obey the HYDROS forward model [22]. The nadir vegetation opacity is related to the columnar vegetation water content $W(kgm^{-2})$ by

$$\tau_o = b_o W \tag{5.1}$$

The coefficient b_o depends on vegetation type. The vegetation water content W is considered as an average value over a computational pixel, i.e., there is no attempt to model fractional vegetation cover within a pixel. A bare soil surface is thus represented by $W = 0$. At L-band the roughness effect is:

$$r_p = r_{sp} exp(-h) \tag{5.2}$$

The parameter h is related empirically to the RMS surface height s. The reflectivity of a smooth soil surface, r_{sp} is determined by the soil dielectric constant e using the Fresnel equations. The dielectric constant is related to the soil moisture content through empirical relationships that have a parametric dependence on soil texture (sand fraction, and clay fraction) and other characteristics, including bulk density and specific particle density (Dobson et al., 1985 [26]).

The canopy (foliar) vegetation water content $Wc(kgm^{-2})$ was obtained from the $NDVI$ database, as shown in Figure 5-5 using the relationship suggested for grassland

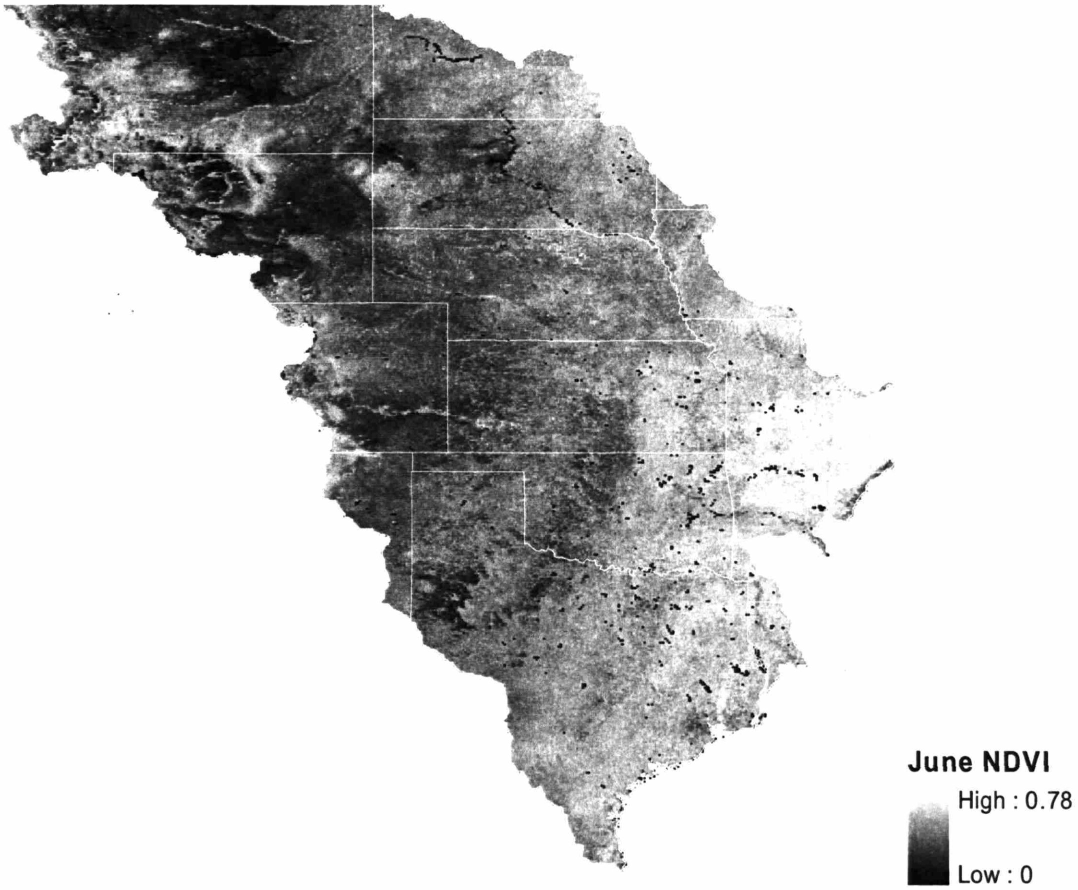


Figure 5-5: The June NDVI

by Jackson et al. (1999)

$$Wc = 0.3215NDVI + 1.9134NDVI^2 \tag{5.3}$$

The computed Wc using (5.3) gives slightly negative values for $NDVI < 0.168$. In this case Wc is set to zero (assumed bare soil). Since the $NDVI$ is sensitive to vegetation greenness it responds primarily to the foliar canopy and not the woody components of the vegetation. Hence, a woody component fraction fT is used to scale the foliar water content derived from (5.3) to a total vegetation water content W

$$W = Wc/(1 - fT) \tag{5.4}$$

(5.4) assumes values for fT are given in Table E.1.

Also included in Table E.1 are the other static parameters used for simulation: surface roughness (h), single scattering albedo (w), vegetation opacity coefficient (b_0).

2). Radar Backscatter Models

The fundamental basis of active remote sensing is similar to the passive remote sensing. For active remote sensing, the received electromagnetic signal is also a function of the dielectric constant of soil which is affected by soil moisture. The signal is quantified as the backscattering coefficient σ^o, which is a unitless quantity representing the radar cross-section of a given pixel on the ground per unit physical area of that pixel. Due to its dramatic dynamic range, it is usually presented in decibels (dB). According to Hoeben et al., (1997) [54], a variation of relative dielectric constant between 3 and 30 (a shift in volumetric moisture content between approximately 2.5% and 50%, depending on frequency and soil texture) causes an 8 to 9 dB rise in backscatter coefficient for vv (vertical transmit vertical receive) polarization.

Besides soil dielectric constant, the surface scattering behavior is also governed by topography, vegetation cover, surface roughness, observation frequency, wave polarization and incidence angle. The relationship between backscattering coefficient and dielectric constant is non-linear, having a higher sensitivity at low dielectric values. Currently available backscattering models include the empirical model (EM) of Oh et al. (1992) [78]; the theoretical integral equation model (IEM) of Fung et al. (1992) [39]; and the semi-empirical model (SEM) of Oh et al. (1994), model of Dubois et al. (1995) [28], Chen et al. (1995) [16] and Shi et al. (1997) [90]. Walker [100] provides a comprehensive review of the first 3 active remote sensing models.

In this thesis, the soil-vegetation radar backscatter model also follows the HYDROS specification and is expressed as the sum of three components:

$$\sigma^t = \sigma^s exp(-2\tau_0/cos\theta) + \sigma^v + \sigma^{sv} \tag{5.5}$$

In this expression, σ^t represents the total radar scattering cross-section, σ^s represents the scattering cross-section of the soil surface modified by the two-way vegetation attenuation, σ^v is the scattering cross-section of the vegetation volume, and σ^{sv} represents the scattering interaction between the soil and vegetation. Empirical models are used to relate the scattering components σ^s,σ^v, and σ^{sv} to the soil moisture and vegetation characteristics.

For soil, the Dubois et al [28] model is used for the co-polarized backscatter, expressed as:

$$\sigma^s_{hh} = 10^{-2.75}\frac{cos^{1.5}\theta}{sin^5\theta}10^{0.028\epsilon' tan\theta}(ks \cdot sin\theta)^{1.4}\lambda^{0.7} \tag{5.6}$$

where, λ (cm) is the wavelength, $k = 2\pi/\lambda$ (cm^{-1}) is the wave number, s (cm) is the surface RMS height, and ϵ' is the real part of the dielectric constant.

For vegetation, it's assumed that over the spatial extent of the HYDROS 3-km radar footprints, the scattering is dominated by randomly-oriented components. Models for the co- and cross- polarized backscatter from vegetation represented as randomly orientated disks are given by Ulaby et al. (1986) [97]. The expressions for the volume scattering and surface-volume interaction terms are shown below.

Volume contribution:

$$\sigma_{hh}^{v} = \ 0.74\omega cos\theta[1 + 0.54\omega\tau_o - 0.24(\omega\tau_o)^2] \ \ [1 - exp(-2.12\tau_o sec\theta)] \quad (5.7)$$

Surface-volume interaction contribution:

$$\begin{aligned}\sigma_{hh}^{sv} =&1.9\omega cos\theta[1 + 0.9\omega\tau_o + 0.4(\omega\tau_o)^2][1 - exp(-1.93\tau_o sec\theta)] \\ &exp(-1.37\tau_o^{1.12} sec\theta) exp(-0.84(ks)^2 cos\theta) r_{sp}\end{aligned} \quad (5.8)$$

3). Synthetic Measurement and the Measurement Errors

As for the measurement error, it includes instrument noise (that limits the measurement precision) and calibration relative error. It is assumed that the absolute calibration errors (static or slowly-varying bias components) can be removed in the HYDROS post-launch processing and are not considered.

For passive sensor, the two other key factors contributing the relative calibration error include the signal attenuation by vegetation and the interplay between nonlinear retrieval physics and the relatively poor spatial resolution of microwave space-borne sensors for heterogenous surface and subsurface conditions. To generate radiometer measurement error, according to [32] spatially independent Gaussian random variables with zero mean and standard deviation 1 K is added to the 40km radiometer measurements aggregated from the 2km true radiometer measurement. A sample of the synthetic radiometer measurement is shown in Figure 5-6.

The radar measurement precision depends on the signal-to-noise ratio (SNR) and the number of independent samples or 'looks' averaged in each measurement. Assuming that in most situations the error is determined primarily by the number of looks and is not limited by the SNR, the average error across the hi-res swath is 0.15. To generate radar measurement error, Gaussian random variables with mean 1 and standard deviation 0.15 is multiplied to the 5-km radar backscatter coefficients (in units of linear power ratio not dB) aggregated form the 2km true radar measurement.

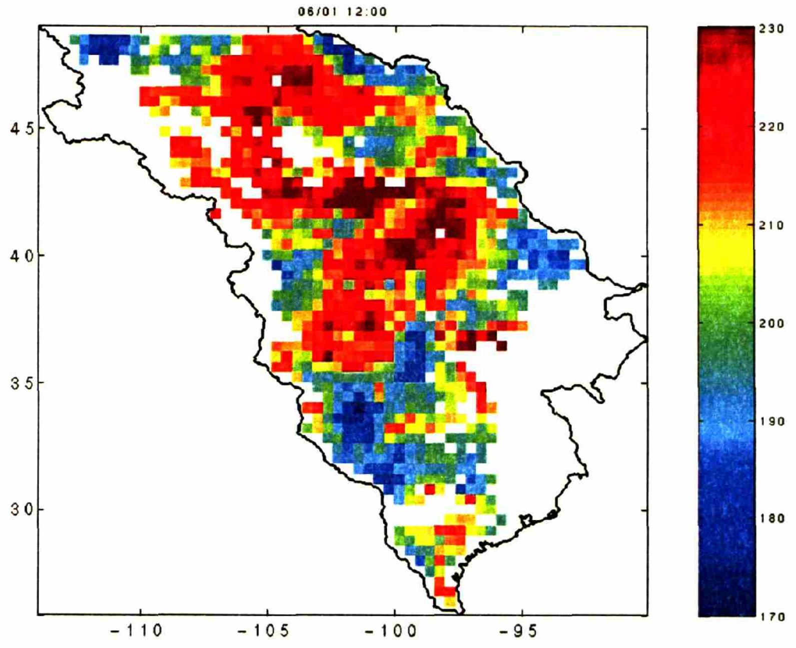


Figure 5-6: A sample of synthetic passive measurement (K)

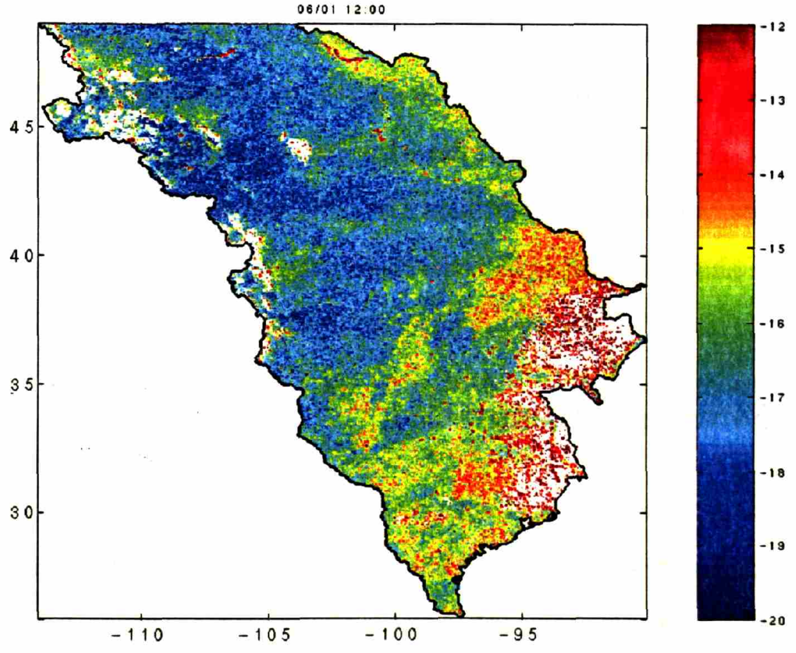


Figure 5-7: A sample of synthetic active measurement (dB)

5.3 EnMSF Experiments with CLM and L-Band Microwave Soil Moisture Measurements

In the following experiments, to assess the ability of EnMSF to assimilate microwave measurements, CLM is used. Initial condition for the assimilation is obtained from a one month spin-up run with the same set of June NCEP reanalysis forcing as in the truth run, June rainfall replicates from Chatdarong [15], spatially uniformly distributed random initial soil moisture and temperature condition, the mean of which are the same as that in the spin-up run for the truth generation. Standard deviation for soil moisture and temperature profile is 0.1 and 4K respectively. For all the assimilation experiments here, a quad-tree is used. A node at the finest scale represents a 2×2 group of pixels. For each pixel, only the first 3 layers of soil moisture are updated. So the total number of state variables is 3×463×479=665331. Coarse scale state dimension $d = 20$ if at the second finest scale, $d = 4$ if above the second finest scale. Cutoff region size $h = 2$. The active and passive measurement resolution is 5km and 40km respectively, so the active measurements are associated with the finest scale nodes, and passive measurements are associated with the third finest scale nodes. At measurement times, both measurements are placed on the corresponding nodes on the tree, which forms a multiresolution estimation problem. Because in real applications the microwave measurements over forests are problematic, only pixels without forest land coverage are observed and updated. For all the runs only 50 replicates are used.

5.3.1 Soil Moisture Estimation

To examine the value of passive and active measurements for estimating soil moisture, four different scenarios are run. First is the soil moisture estimation unconditioned on microwave measurements, the second is the assimilation with passive data only, the third is with active data only, the fourth is with both passive and active measurements. The same tree structure is used for all the scenarios. The spatial RMSE of the ensemble mean of soil moisture with respect to the true soil moisture at 5km resolution is plotted in Figure 5-8, in which the "N" on x-axis represents UTC noon time .

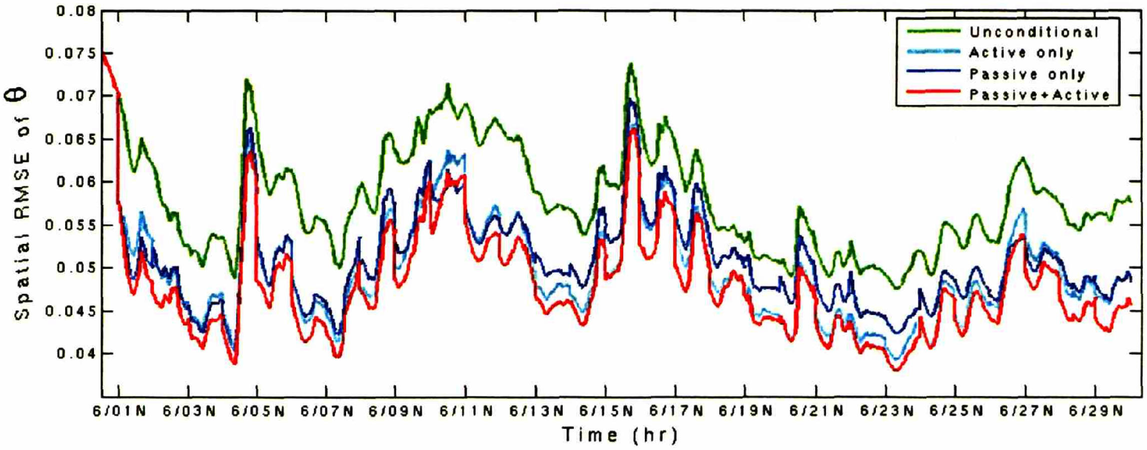


Figure 5-8: Top layer soil moisture RMSE of ensemble mean with respect to the truth over the whole domain. "N" on x-axis represents UTC noon time.

The RMSE figure shows that using both active and passive measurement is consistently better than using either of them. This is because coarse passive measurements provides better sensitivity to soil moisture and fine active measurements provide more spatial details. Also, using any measurement is better than the unconditional estimate. The reduction of RMSE is more obvious right after rainfall when the uncertainties introduced by the rainfall replicates are significant.

The effects of assimilating microwave measurements can also be assessed from the soil moisture time series at some pixels. The updated ensemble of soil moisture at some pixels shown in Figure 5-11 are plotted in Figure 5-9 and Figure 5-10. The yellow color in Figure 5-11 represents the pixels with both passive and active measurements, while the cyan color represents the those with active measurements only.

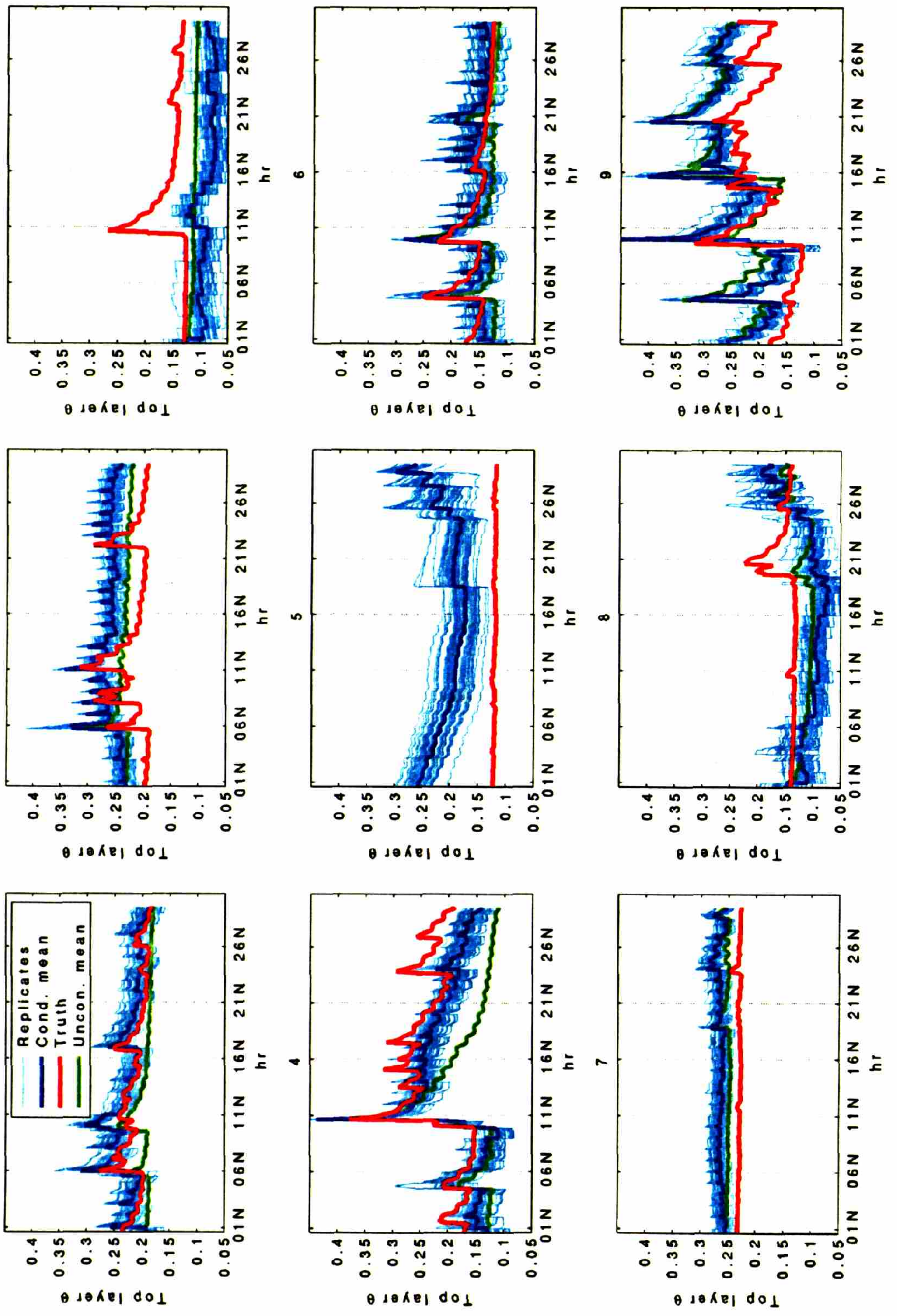


Figure 5-9: Soil moisture ensemble of pixel 1-9 from EnMSF update. "N" on x-axis represents UTC noon time. Panel title is the pixel number in Figure 5-11.

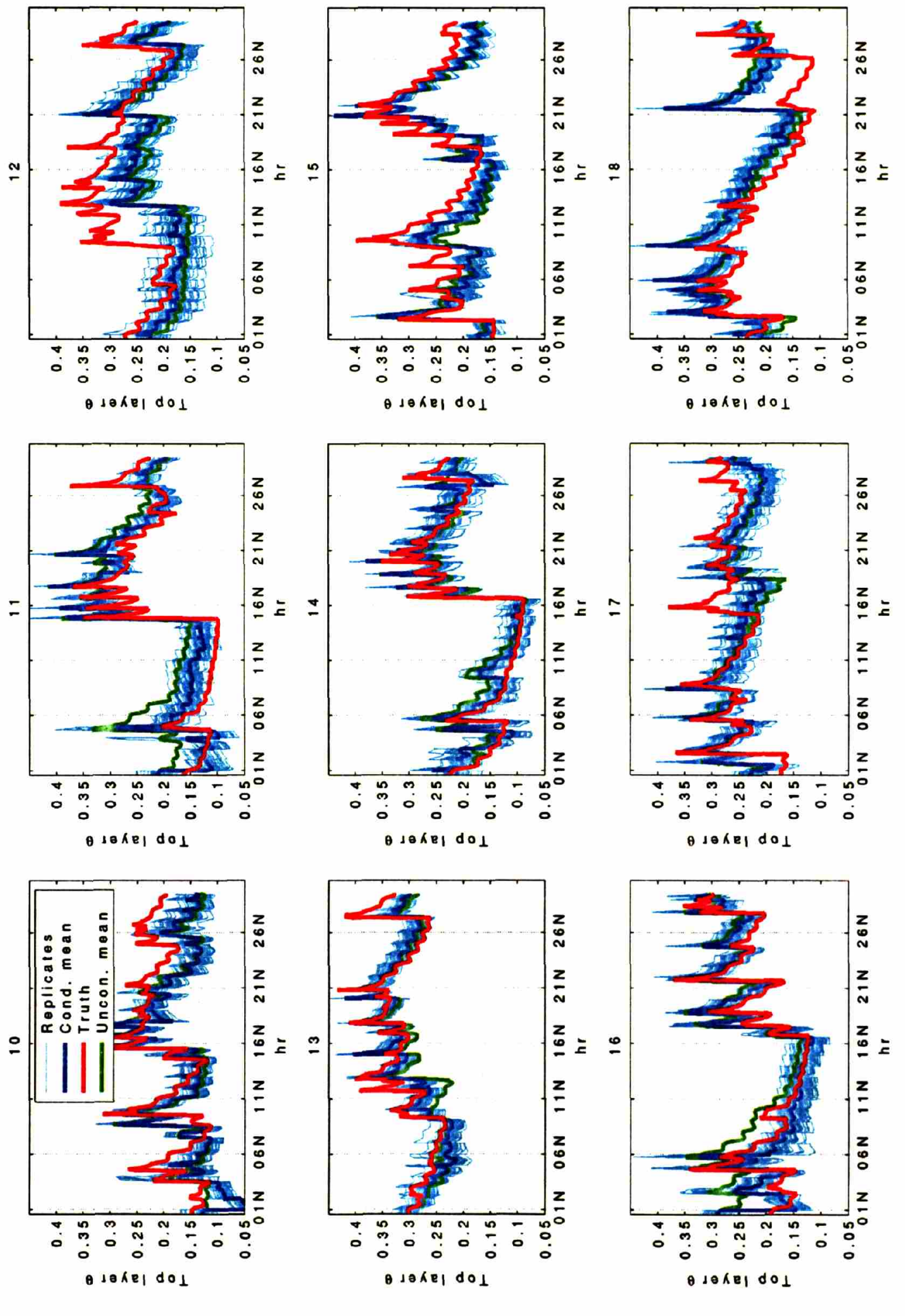


Figure 5-10: Soil moisture ensemble of pixel 10-18 from EnMSF update. "N" on x-axis represents UTC noon time. Panel title is the pixel number in Figure 5-11.

In general soil moisture estimation of the pixels without radiometer measurements are inferior to those with. The unobserved pixels are even worse compared to those measured ones. At some measurement times, the soil moisture field estimate after update are shown in Figure 5-12. In Figure 5-12 the error is defined as the ensemble mean of the estimate minus the truth. The improvements of the spatial pattern from assimilating microwave measurement is obvious at all the tree times.

Figure 5-13 shows the standard deviation of top layer soil moisture estimate with both sensors before and after update at three typical measurement times. The uncertainty reduction from measurements largely reside in the regions with precedent rainfall. For those dry regions, the reduction of standard deviation is very limited due to the narrow ensemble spread.

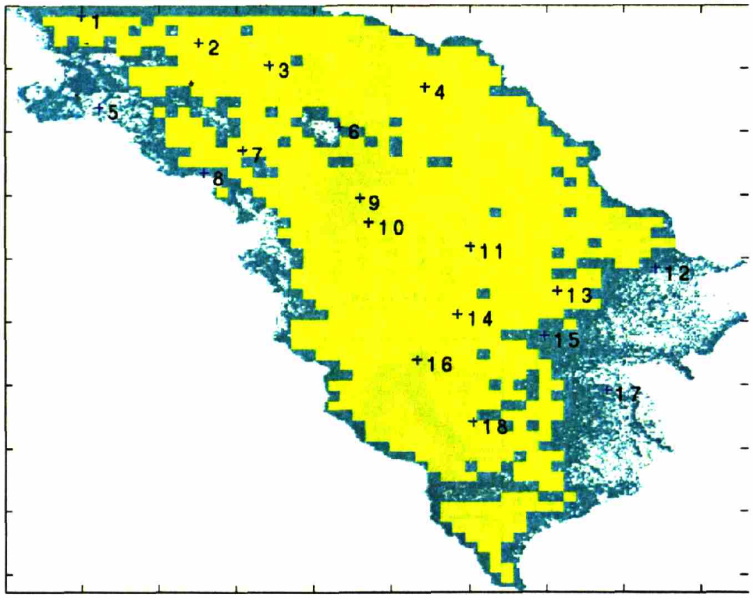


Figure 5-11: Locations of pixels 1-18 with ensemble soil moisture and evaporative fraction output

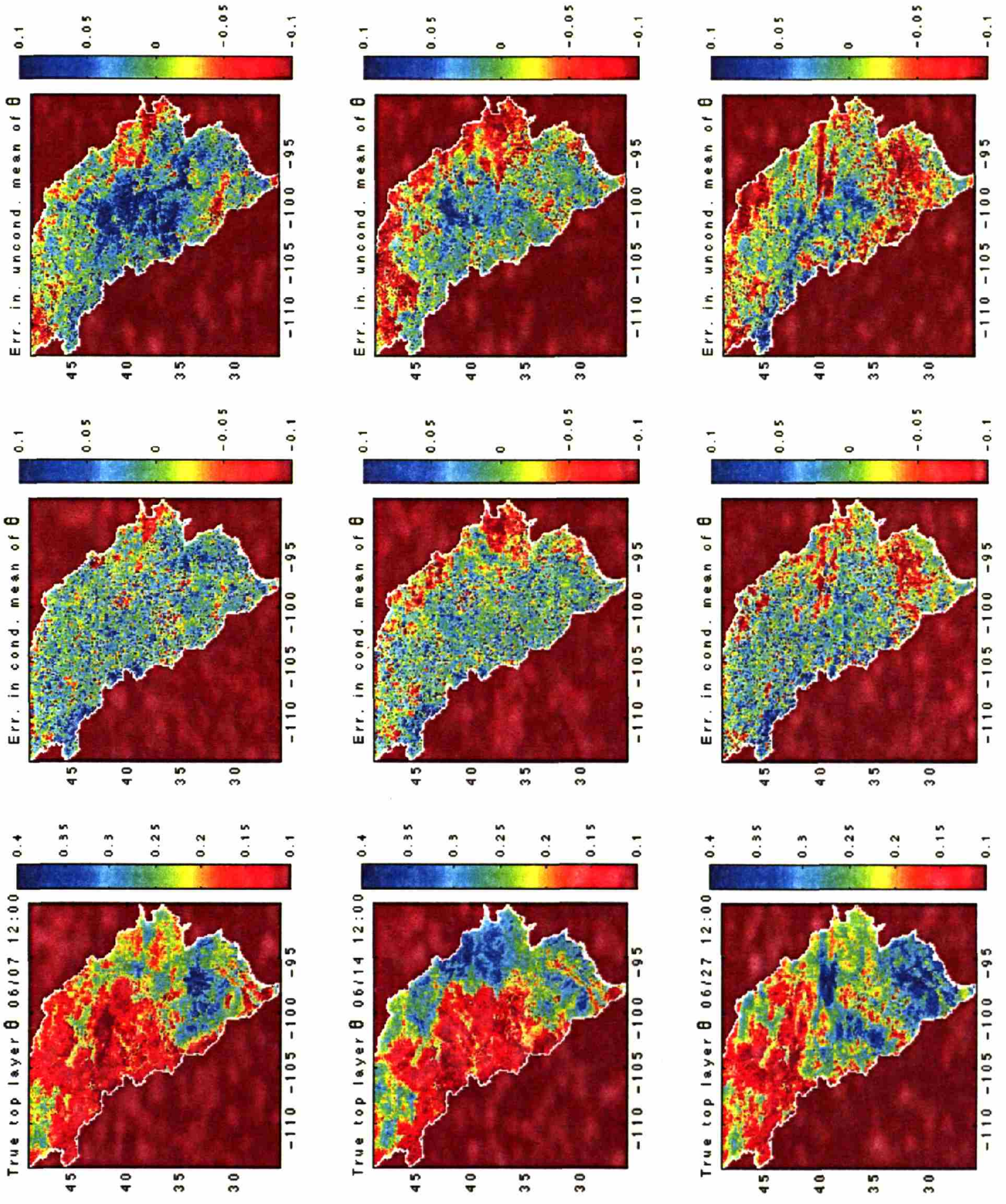

Err. in. uncond. mean of θ
Err. in. uncond. mean of θ
Err. in. uncond. mean of θ
Err. in cond. mean of θ
Err. in cond. mean of θ
Err. in cond. mean of θ
True top layer θ 06/07 12:00
True top layer θ 06/14 12:00
True top layer θ 06/27 12:00
0.1
0.05
0
-0.05
-0.1
0.4
0.35
0.3
0.25
0.2
0.15
0.1
45
40
35
30
-110
-105
-100
-95

Figure 5-12: Top layer soil moisture field estimation compared with the truth and unconditional estimation at three measurement times. The error is defined as the ensemble mean of the estimate minus the truth.

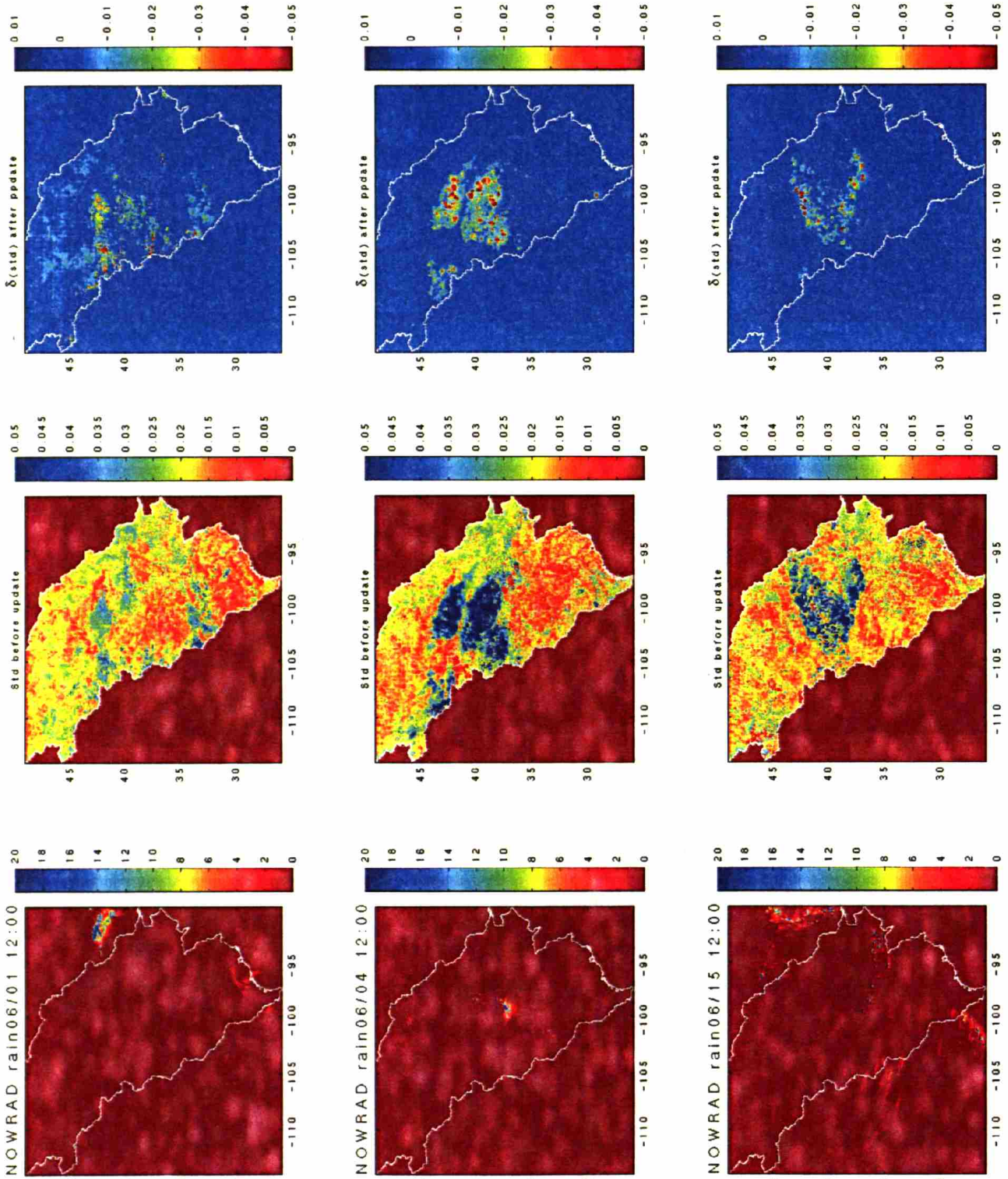


Figure 5-13: Top layer soil moisture standard deviation field before and after update at three measurement times

5.3.2 Evapotranspiration Estimation

ET is an inferred quantity from soil moisture by the means of CLM prediction. The utility of microwave soil moisture measurements for ET estimation is also examined for the same four scenarios as in the previous section. The spatial RMSE of the ensemble mean of ET with respect to the true ET at 5km resolution is shown in 5-14. The diurnal pattern of ET is dominant in Figure 5-14. The update time is 12:00 UTC, but the difference of ET between the unconditional estimate and conditional estimate is most significant around noon local time. At night times, there is very little difference in the ET RMSE. Also, for ET estimation, either passive or active measurement is adequate compared to the passive+active case.

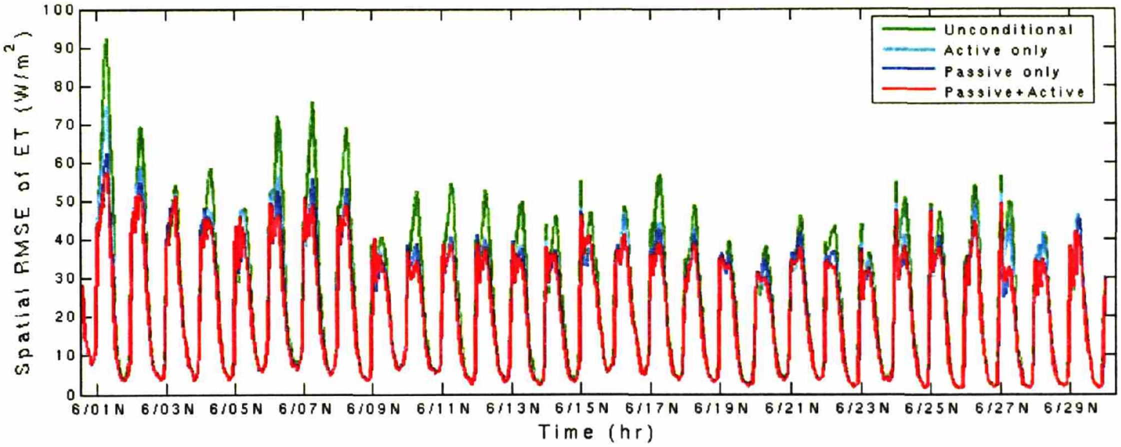


Figure 5-14: RMSE of evapotranspiration ensemble mean with respect to the truth over the whole domain. "N" on x-axis represents UTC noon time.

To better present ET estimate, the diurnal pattern of it can be eliminated by using daytime evaporative fraction (EF) rather than ET itself. The daytime (between 13UTC and 23UTC) EF is defined as

$$EF = \frac{\text{daytime latent heat flux}}{\text{daytime latent heat flux+daytime sensible heat flux}}$$

The RMSE between EF ensemble mean and the true EF is plotted in Figure 5-15. In general, the EF estimate with soil moisture measurement is better than the unconditional one, which is consistent with the ET estimate. EF from the active

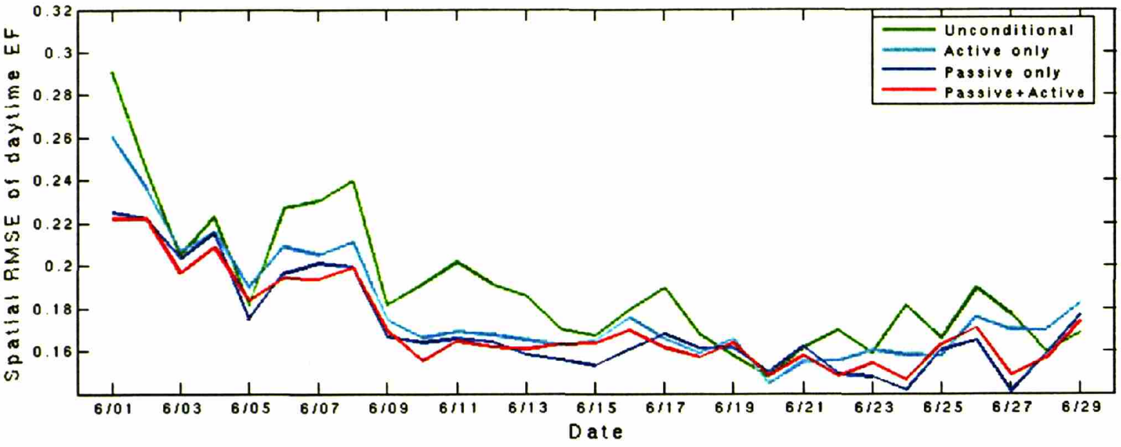


Figure 5-15: RMSE of evaporative fraction ensemble mean with respect to the truth over the whole domain

only case is worse than either passive or passive+active case. The difference between passive and passive+active is not that much.

The ensemble of EF at pixels in Figure 5-10 updated with the EnMSF and passive+active data are also plotted in Figure 5-16 and Figure 5-17. During the wet periods, available energy is mostly used for evaporation, so EF is close to 1. The difference in the EF ensemble mean and the truth is not much if there is no uncertainties associated with radiation input to CLM. While in dry periods, when ET is controlled by soil moisture availability, the microwave measurements helps a lot. EF of the pixels observed are generally much better than those unobserved, which is consistent with the soil moisture estimate. EF estimates at noon local time on the same tree days as in Figure 5-12 are shown in Figure 5-18. The error is defined as the ensemble mean of the estimate minus the truth. The spatial pattern improvements of the conditional EF estimate can be easily observed in Figure 5-18. The improved spatial pattern can be completely attributed to the improved soil moisture estimate.

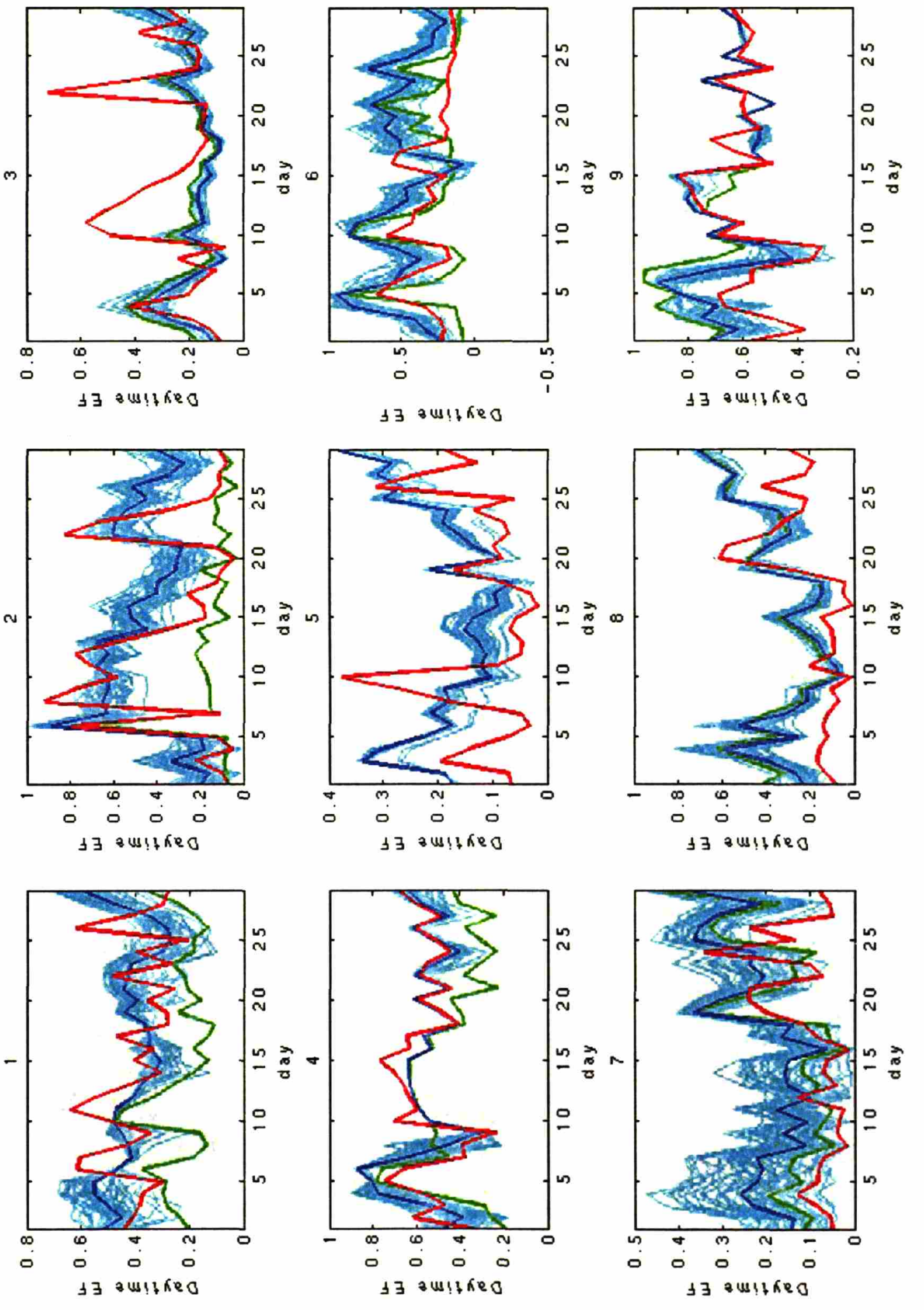


Figure 5-16: Daily EF ensemble of pixel 1-9 from EnMSF update, Legend: cyan-replicates from EnMSF; blue-conditional ensemble mean; red-truth; green-uncondition mean

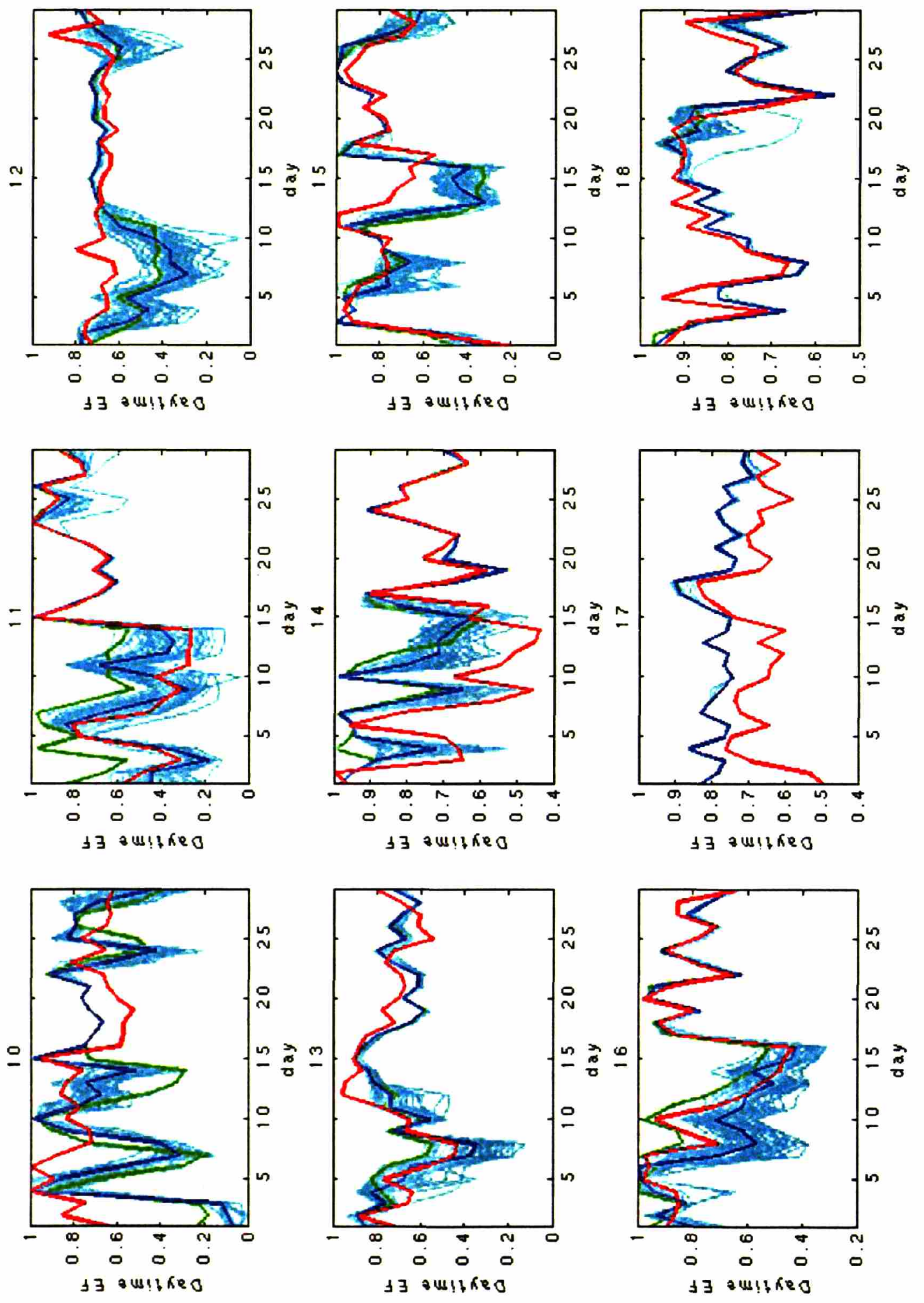


Figure 5-17: Daily EF ensemble of pixel 10-18 from EnMSF update, Legend: cyan-replicates from EnMSF; blue-conditional ensemble mean; red-truth; green-unconditional mean

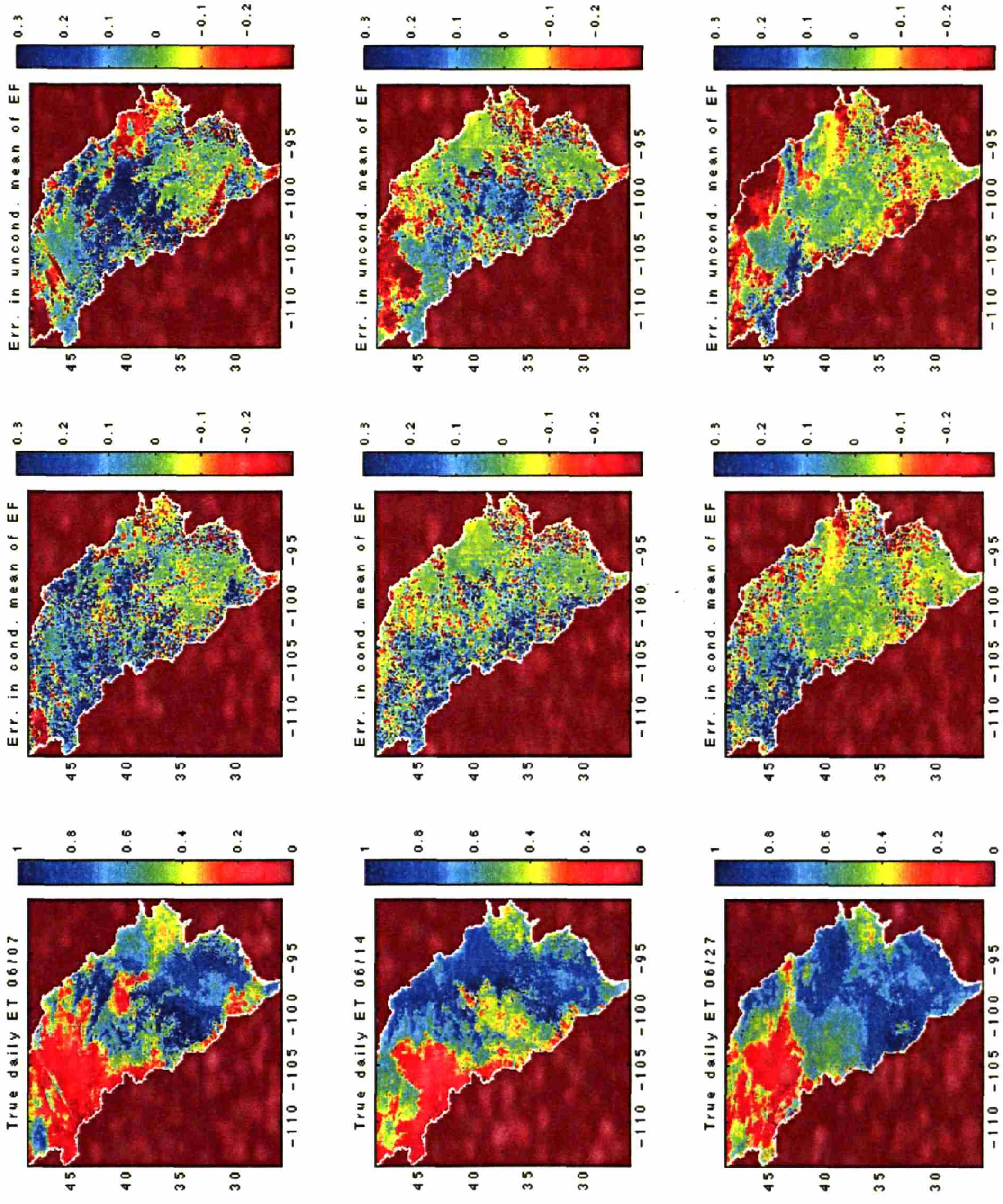


Figure 5-18: EF field estimation compared with the truth and unconditional estimation at three times. The error is defined as the ensemble mean of the estimate minus the truth.

5.3.3 Root Zone Soil Moisture Estimation

Root zone soil moisture over vegetated areas is an important variable for many applications such as short and medium term meteorological, climate modelling, and hydrological studies. It is very hard to observe over large scale domain. However, it is directly related to surface soil moisture which can be observed. The effects of the observed surface soil moisture on root zone soil moisture estimation is examined in this section. Here the root zone soil moisture is defined as the soil water depth (mm) above 1.06m deep (top 7 layers in CLM). Figure 5-19 shows the RMSE of the root zone soil moisture ensemble mean with respect to the truth. The conditional estimation is consistently better than the unconditional mean. Also, the passive+active data helps the most. The downward trend of the RMSE reflects the stabilization of the error in the initial condition.

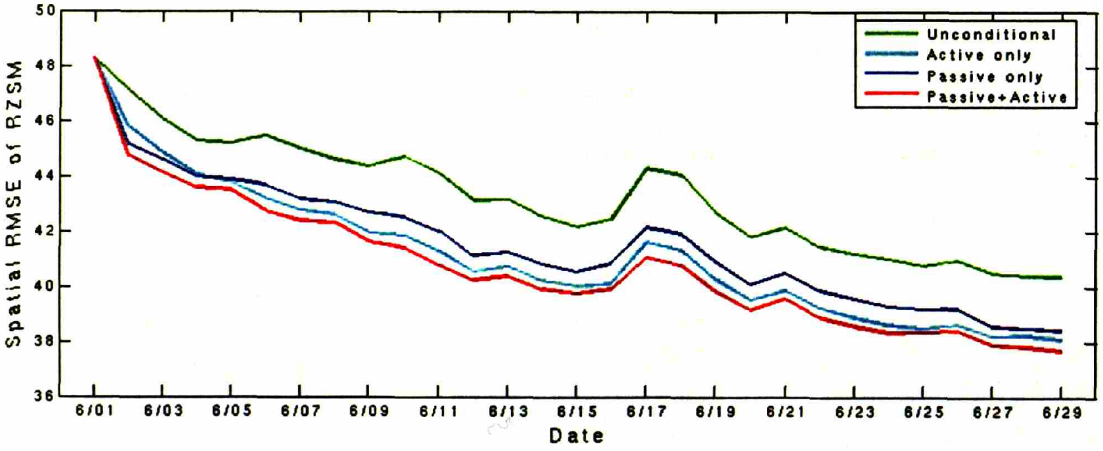


Figure 5-19: Root zone soil moisture RMSE of ensemble mean with respect to the truth over the whole domain.

The root zone soil moisture field estimation at three different times are shown in Figure 5-20. The error is defined as the ensemble mean of the estimate minus the truth. It can be seen that the error in the spatial distribution pattern decreased as time increases. The difference between the conditional and unconditional estimate is also obvious.

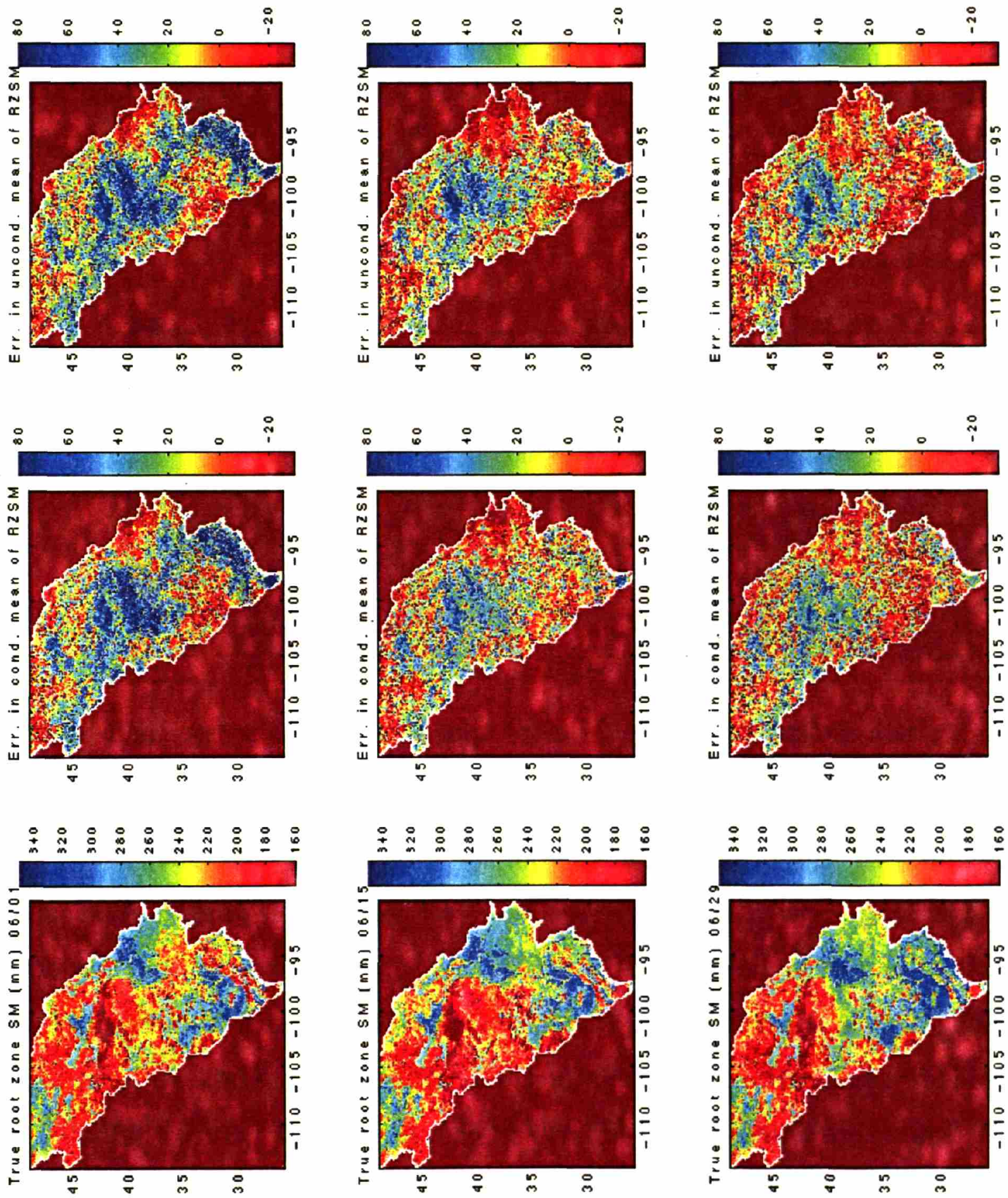

Err. in uncond. mean of RZSM
Err. in cond. mean of RZSM
True root zone SM (mm) 06/01
True root zone SM (mm) 06/15
True root zone SM (mm) 06/29
80
60
40
20
0
-20
340
320
300
280
260
240
220
200
180
160
45
40
35
30
-110
-105
-100
-95

Figure 5-20: Root zone soil moisture field estimation compared with the truth and unconditional estimation at three times. The error is defined as the ensemble mean of the estimate minus the truth.

5.3.4 Tree Performance Assessment

The performance of the EnMSF for estimation largely depends on how well the tree structure keeps error correlation structure and discards the sampling error. To assess if the selected tree structure works, the top two layers of soil moisture error correlation realized from the tree model and sampling covariance at pixel (148,264) are plotted at two typical times right before EnMSF update. The soil moisture at these two times are shown in Figure 5-21.

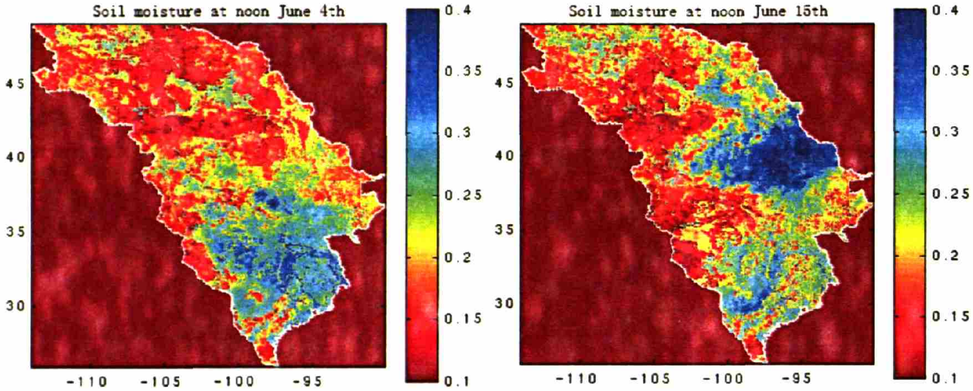


Figure 5-21: Soil moisture mean field before update with both sensors and EnMSF at 12:00 on June 4th and 15th

In Figure 5-22, the pixels far away from the target pixel (148, 264) nearly has no error correlation with it in the tree realized error correlation, which is reasonable. It indicates the tree model is capable of deleting spurious error in sampling error correlation. This is consistent with the dry soil moisture mean field in Figure 5-21, which leads to local error correlation structure. While in Figure 5-23 with the same tree structure, there is some realized large error correlation structure attributed to the precedent rainfall. Figure 5-21 shows that at 12:00 June 15th the soil moisture error correlation is also consistent with the soil moisture mean field due to the large scale rainfall input. However, in the original sampling error correlation map, this structure is not so clear. In other words, the tree model has the ability of extracting information from the noisy data in some sense. The corresponding standard deviation of top layer soil moisture change at this two times are shown in Figure 5-13.

Also, the error correlation structures shown in Figure 5-22 and 5-23 are different

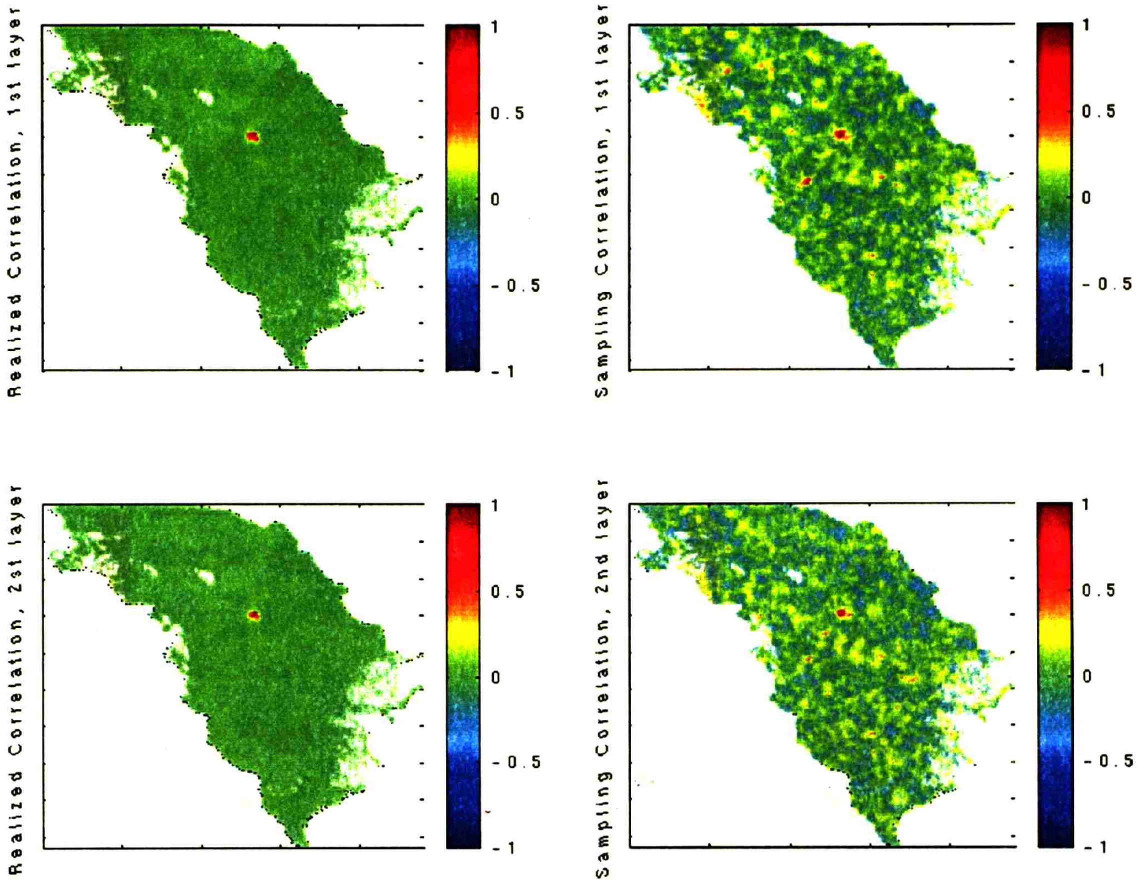

Realized Correlation, 1st layer
1
0.5
0
-0.5
-1
Sampling Correlation, 1st layer
1
0.5
0
-0.5
-1
Realized Correlation, 2st layer
1
0.5
0
-0.5
-1
Sampling Correlation, 2nd layer
1
0.5
0
-0.5
-1

Figure 5-22: The sampling error correlation and the realized error correlation for a pixel at 12:00 on June 4th

than Figure 4-3 in chapter 4. The anisotropic error correlation structure shown in chapter 4 is not so obvious here. Most of the error correlation structure are local for the Great Plains case. This is because the scale of domain in chapter 4 is rather small compared to the rainfall scale. The soil moisture error correlation largely reflects rainfall error correlation. Here the rainfall scale is much smaller than the domain size. The dissipation of soil moisture in the interstorm periods dominate the majority of the time, which implies soil properties control most of the uncertainties and thus leading to local error correlation structure.

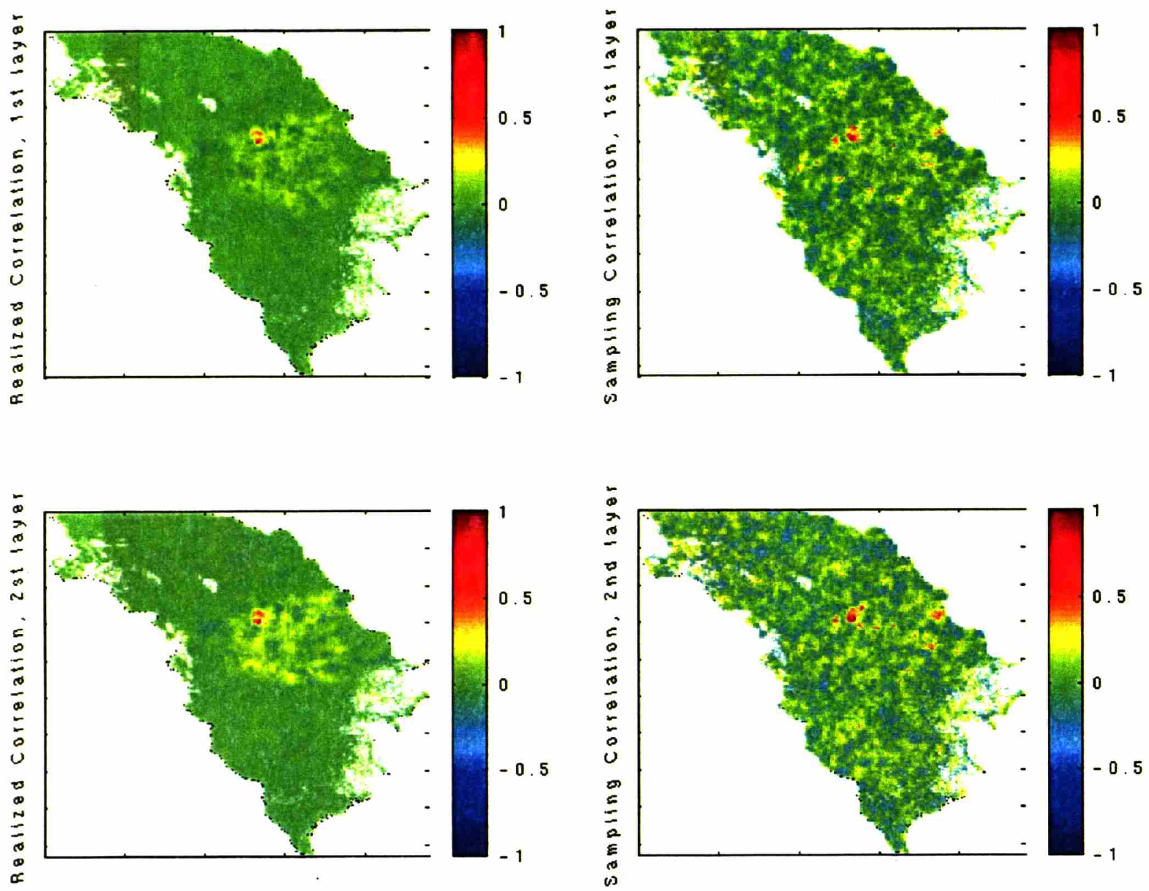

Realized Correlation, 1st layer
1
0.5
0
-0.5
-1
Sampling Correlation, 1st layer
1
0.5
0
-0.5
-1
Realized Correlation, 2st layer
1
0.5
0
-0.5
-1
Sampling Correlation, 2nd layer
1
0.5
0
-0.5
-1

Figure 5-23: The sampling error correlation and the realized error correlation for a pixel at 12:00 on June 15th

5.4 Discussion and Conclusions

In this chapter, the EnMSF is successfully applied to a large scale land surface system CLM to estimate soil moisture and evapotranspiration using synthetic microwave sensor measurements. Compared to the Navier-Stokes equations in the last chapter, the system dynamics is significantly different. The land surface system is mainly driven by exogenous rainfall and radiation. The uncertainties in the soil properties controls the uncertainties of soil moisture especially in the interstorm periods. The difference in the dynamics is evident in the sampling spatial error correlation structures. In the previous chapter, there exists significant large scale and small scale features that are created by the vortices. But for the large scale land surface problem, the error correlation structure is largely local, resulting from the complicated effects of atmospheric forcing, soil, and vegetation properties. For both type of systems, the tree algorithm can handle the error correlation structure quite well while the tree structure is very different.

For soil moisture estimation, the results show that using both active and passive measurement is better than using either of them and using any measurement is better than the unconditional update. The reduction of RMSE is more obvious right after rainfall when the uncertainties introduced by the rainfall replicates are significant. The results also proves that the improved soil moisture estimation by using microwave measurements helps with the ET and root zone soil moisture estimation.

By checking the error correlation structures at most of times (not shown in figures), we can find the error correlation structure is like the error correlation in Figure 5-22. So they can be roughly approximated with local isotropic error correlation structure if using the current random replicates generation approaches for soil and rainfall. The nearly isotropic error correlation structure implies Schur product method as for the traditional EnKF assuming local isotropic error correlations might work well. However, it might be problematic for a general application when error correlation structure can't be assumed isotropic and hard to know the error correlation scale. The tree models can identify any error correlation structure, regardless whether the field is highly correlated or independent. It is not a tailored algorithm to any particular application but a general filtering technique. Although the tree structure assumption is independent of any structure assumption but any prior knowledge of the error correlation structure would definitely help to design a more efficient tree structure.

There are several other facts about the land data assimilation that deserve attention: (1) For land surface system, once the rainfall is overwhelming, it can help reduce the difference in the estimates from different filters. (2) The performance of evapotranspiration estimate is different for wet or dry soil moisture regime. During wet period the evapotranspiration is basically controlled by available energy rather than soil moisture. The benefits from microwave soil moisture measurements can't immediately appear. But it can help shorten the time to get to the transition period from wet to dry during or after which updating soil moisture would have great impact on evapotranspiration. (3) For different soil properties, a filter might have different performance. For instance, model prediction of soil moisture for sandy soil usually is very confident compared to loam or clay soil. Microwave measurements can't help much. So is any filer. Also, the response function of latent heat flux to soil moisture is different. For sandy soil the update in soil moisture has less impact on evapotranspiration. (4) For a large scale land surface problem, the state vector size would be as large as few millions. Because soil temperature is not a key factor, only including soil moisture in the state vector can reduce the dimension of state vector size and save computational cost.

Chapter 6

Future Research Directions

6.1 Simple Dynamic Model

As discussed in the first chapter, the basic dynamic system in land surface data assimilation problem is composed of nonlinear water and heat transfer equations. The CLM model used in this thesis directly solves Richard's equation at coarse resolution vertically. The empirical parameterizations in this model require a number of land cover parameters, soil parameters for the empirical models and forcing inputs. For example, it includes canopy radiative model, surface albedo models, sensible and latent heat models, soil thermal and hydraulic parameter models, stomatal resistance and photosynthesis models, etc. Such a huge collection of the parameterizations make the model structure rather complex. The model complexity effects on data assimilation are twofold. First of all, a complex model would increase the computational burden especially for Monte Carlo simulations. Although the land surface simulation essentially is pixel by pixel, a complex model is still rather expensive to compute even for one pixel due to the iterative approach for solving the latent and sensible fluxes. On the other hand, the current land surface data assimilation techniques usually takes a LSM and perturb the parameters and inputs while the model structure is fixed. When observation becomes available, the state variables are updated based on the new measurement. Sometimes updating state variables is not adequate. If the model structure or the parameters are wrong, the estimates would quickly deviate from the true state during the propagation steps, which would give incorrect prior ensemble for the next update. For example, if the soil type is sand, but in the model it's specified as clay, one will never get the correct soil moisture trajectory. Since this is very common in land surface data assimilations, an reasonable way to deal with this is to allow the model parameters to be updated with measurement. Then data assimilation becomes a state AND parameter estimation problem, or a system identification problem. Since the land surface system is a nonlinear dissipative system with many inputs and parameters and spatially and temporally sparse measurements, it is almost impossible to use the traditional system identification techniques to solve for all these parameters. A tangible way to identify the system is to use some simply structured model relatively easier to identify. Another advantage of using a simple model is that it is easier for variational approaches to solve for model parameters.

Another system model related issue is the bias and model error estimation. As stated above, model structure error might lead to wrong prior information in the ensemble. A direct consequence of this is bias in the states would arise. An easy way to deal with the bias estimation is to model it as an additive term in the right hand side of the state equation. Then formulate the problem in a variational framework and solve for the bias as the control variable. Or treat the bias as a state variable in the state space framework, and use the filtering techniques to solve for it.

6.2 Error Models for Soil and Vegetation

According to equation (1.25), given $x^i_{t-1|t-1}$, *i=1, 2, ... N*, the prediction ensemble $x^i_{t|t-1}$ only depends on u^i_t and the land surface model $f(\cdot)$. In real problems u_t can be correlated in space and time. It might also be state dependent. The "closeness" of the updated ensemble to the true uncertainties of the physical states relies on the ability of the predicted ensemble $x^i_{t|t-1}$ to represent the true uncertainties in model parameters and inputs. To obtain a physically realistic ensemble, firstly a proper uncertainty model of u_t should be able to describe its mean and variance. Secondly, a proper uncertainty model of u_t should also be able to describe spatial and temporal correlation structure. Such u_t not only can help get a realistic prior ensemble but also can help the filter transfer the measurement information in one pixel to other correlated pixels, thus improving the accuracy of the estimation. An advantage of the ensemble filtering approaches is that any reasonable distributions for u_t can be utilized. In this thesis only temporally and spatially correlated rainfall replicated are used, while other forcing variables, soil, and vegetation parameters are simply treated as deterministic variables. Ideally, other forcing variables and vegetation parameters should be temporally and spatially correlated random variables as well, while soil parameters should be spatially correlated. However, generating these correlated random field is not an easy task, especially soil and vegetation parameters. In order to generate these random fields, some constraints should be satisfied. For example, the sum of sand and clay fraction can not exceed 1; vegetation canopy top height should be greater than bottom height, etc. For soil, the spatial correlation structure should be a function of soil type or topography due to geomorphological reasons. Further more, vegetation and soil might also be correlated through the link of soil nutrition. A possible approach to deal with these issues is to use Markov Monte Carlo Chain (MCMC) method for the random field generation. However, MCMC method is iterative and also relies on the solution to filtering problems, which is hard for any non-Gaussian random field. This is because if use filters with Gaussian assumptions the error due the approximation to non-Gaussian problems would accumulate in the iteration procedure.

6.3 Sampling Issues in Ensemble Filters

For all the Monte Carlo type methods, a fundamental problem is to reduce the sampling error. Since the sampling standard deviation is proportional to the inverse of the square root of the sampling size, to reduce the standard deviation measures have to be taken. To obtain correct sampling correlation and high order moments, the required sample size would dramatically increase. Traditional variance reduction methods include control variate, antithetic methods, quasi random sampling, importance sampling, etc. There are already some applications of these methods, such as [52, 27]. These belong to the control variate method. For the quasi random sampling and antithetic methods, the generated samples are correlated, so they are problematic for all the variance based methods. But they should be helpful for simulation based methods. In the multiscale tree approach described in Chapter 4, the basic idea is to reduce the high frequency noise and pick up as much the large scale correlation as possible which require less sampling size. This results from the conditional decorrelation requirements in the tree identification step. The high frequency noise are simply discarded. In this case, if the true signal is in the high frequency range, and the noise is in the low frequency then this tree approach would also be problematic. In this case, to separate the noise and signal, it is better to have some sort of prior knowledge of signal and noise. In order for the tree to satisfy the scale Markovianity, a new approach should be investigated.

6.4 Nonlinear Measurement Model and Nonnormal State Effects in the EnMSF

In the theoretical multiscale filter development, the measurement model is linear with additive Gaussian error and the prior distribution of the state variables is assumed Gaussian for the derivation of the optimality. For the land assimilation applications, the measurement model is nonlinear and the measurement error might be multiplicative, such as the Radar measurement of soil moisture. Also, the state variable is rarely Gaussian. To what extend the nonlinearity in the measurement model and additive Gaussian assumption would affect the result is still unknown. Heuristically, the state variable distribution tends to be more Gaussian as one goes upward the tree due to central limit theorem, which makes the multiscale tree model more favorable at coarse scales. At fine scales, the effects of the nonnormality in the prior ensemble on tree model performance should be closely related to the case in Chapter 2 since the tree update is derived from the EnKF. To study these problems, a large ensemble size experiment must be done.

6.5 Graphical Models for Large Scale Nonlinear Estimation

Although tree models are powerful for large scale estimation, its ability to represent the true covariance is impaired by the induced structure error (blockiness as seen in Figure 4-8). Some approaches have been developed to resolve this problem. In [58], this blockiness is eliminated by discarding the standard assumption that distinct nodes on a given level of the multiscale process correspond to disjoint portions of the domain; instead, overlapping portions of the domain are allowed. In [92] additional edges are added to the tree to make the tree a graph with cycles. This complicates the model structure and estimation, but the blockiness from tree can be drastically reduced to the additional edges. For the graph with cycles, embedded trees approaches are employed to iteratively solve the estimation problem. For a tree derived from samples, the overlapping tree approach seems more promising than the graph models. This is because the sampling error due to small ensemble size in a large scale problem might accumulate in the iterations. Further study in this area is needed.

Appendix A

Convergence of the Upward Sweep of the Ensemble Multiscale Filter Update

A.1 Introduction

The upward sweep of the ensemble multiscale filter update is designed to produce a set of replicates $\chi^j(s|s)$ that approximates the mean and covariance of the Gaussian random state $\chi(s)$, given all measurements at or below s. The error in the sample mean and covariance estimates should converge to zero in the limit as the number of replicates approaches infinity.

A.2 Exact Ensemble Mean and Covariance for the Upward Sweep

Define $\hat{\chi}(s|s) = E[\chi(s)|y(\mathcal{V}_s)] = E[\chi(s)|s]$, where $\mathcal{V}_s$ is the set of nodes consisting of s and all its descendants. So $\hat{\chi}(s|s)$ is the mean of $\chi(s)$ conditioned on all measurements at or below s. The corresponding conditional covariance is $Cov[\chi(s)|y(\mathcal{V}_s)] = Cov[\chi(s)|s]$. We wish to approximate these two moments, for the Gaussian case, with sample estimates from an updated ensemble $\chi^j(s|s)$ derived from (4.40).

We adopt a recursive definition of an augmented measurement vector $Y(s)$ as follows:

$$Y(s) = y(s) \quad ; m(s) = M \tag{A.1}$$

$$Y(s) = \begin{bmatrix} \hat{\chi}(s\alpha_1|s\alpha_1) \\ \vdots \\ \hat{\chi}(s\alpha_q|s\alpha_q) \\ y(s) \end{bmatrix} \quad ; m(s) < M \tag{A.2}$$

where the state estimates $\hat{\chi}(s\alpha_i|s\alpha_i)$ at the chidren of s are defined by (A.6) below. The vector $y(s)$ is the local measurement at s, defined as in (4.31):

$$y(s) \;=\; h(s)\chi_M(s) + e(s) \tag{A.3}$$

The linear structure of the internal model enables us to write a similar measurement equation for the augmented vector $Y(s)$:

$$Y(s) = H(s)\chi_M(s) + \epsilon(s) \tag{A.4}$$

where $H(s)$ is a global measurement matrix. The augmented measurement error vector $\epsilon(s)$ has a covariance $R(s)$ which is defined by the recursion given in (A.12) and (A.13). We do not need an explicit expression for $H(s)$ for this derivation.

The estimates $\hat{\chi}(s\alpha_i|s\alpha_i)$ appearing in the $Y(s)$ definition are derived from measurements $y(\mathcal{V}_{s\alpha_i})$ at or below $s\alpha_i$. It follows that the conditional mean $E[\chi(s)|y(\mathcal{V}_s)]$ is equal to the conditional mean $E[\chi(s)|Y(s)]$, since $Y(s)$ is completely determined from $y(\mathcal{V}_s)$. That is, conditioning on $Y(s)$ is the same as conditioning on $y(\mathcal{V}_s)$.

If all the states $\chi(s)$ and measurement errors $e(s)$ are jointly Gaussian and the

$e(s)$ are uncorrelated with one another the desired conditional mean is given by [89]:

$$\hat{\chi}(s|s) \;=\; E[\chi(s)|Y(s)] = K(s)Y(s) = K(s)y(s) \quad ; m(s) = M \tag{A.5}$$

$$\hat{\chi}(s|s) = E[\chi(s)|Y(s)] = K(s)Y(s) = K(s)\begin{bmatrix} \hat{\chi}(s\alpha_1|s\alpha_1) \\ \vdots \\ \hat{\chi}(s\alpha_q|s\alpha_q) \\ y(s) \end{bmatrix} \quad ; m(s) < M \tag{A.6}$$

where $K(s)$ is an augmented Kalman gain matrix defined as:

$$\begin{aligned} K(s) &= Cov[\chi(s), Y(s)]Cov^{-1}[Y(s)] = \\ &Cov[\chi(s), H(s)\chi_M(s)]\left[H(s)Cov[\chi_M(s)]H^T(s) + R(s)\right]^{-1} \\ &\quad ; m(s) \leq M \end{aligned} \tag{A.7}$$

Here $Cov[\chi(s), Y(s)]$, $Cov[Y(s)]$, $Cov[\chi(s), H(s)\chi_M(s)]$, and $Cov[\chi_M(s)]$ are all prior covariances that may be derived from the covariance of the propagated state $\chi(t|t-1)$ and the specified error covariances for the measurements in $\mathcal{V}_s$. Note that $\hat{\chi}(s|s)$ is the minimum variance linear estimate, even when $\chi(s)$ and $e(s)$ are not Gaussian.

In order to evaluate $R(s)$ use (A.4), (A.5) and (A.6) to obtain the following expression for $Y(s)$:

$$Y(s) \;=\; y(s) \quad ; m(s) = M \tag{A.8}$$

$$\begin{aligned} Y(s) &= \begin{bmatrix} K(s\alpha_1)Y(s\alpha_1) \\ \vdots \\ K(s\alpha_q)Y(s\alpha_q) \\ y(s) \end{bmatrix} \\ &= \begin{bmatrix} K(s\alpha_1)H(s\alpha_1)\chi_M(s\alpha_1) \\ \vdots \\ K(s\alpha_q)H(s\alpha_q)\chi_M(s\alpha_q) \\ h(s)\chi(s) \end{bmatrix} + \begin{bmatrix} K(s\alpha_1)\epsilon(s\alpha_1) \\ \vdots \\ K(s\alpha_q)\epsilon(s\alpha_q) \\ e(s) \end{bmatrix} \\ &= H(s)\chi_M(s) + \epsilon(s) \qquad ; m(s) < M \qquad \text{(A.9)} \end{aligned}$$

The first part of this expression defines an alternative $Y(s)$ recursion and confirms that the augmented measurement vector at s is a linear function of all the measurements in $\mathcal{V}_s$. The second part defines the following recursion for $\epsilon(s)$.

$$\epsilon(s) = e(s) \quad ; m(s) = M \qquad \text{(A.10)}$$

$$\epsilon(s) = \begin{bmatrix} K(s\alpha_1)\epsilon(s\alpha_1) \\ \vdots \\ K(s\alpha_q)\epsilon(s\alpha_q) \\ e(s) \end{bmatrix} \quad ; m(s) < M \qquad \text{(A.11)}$$

This gives a recursion for the covariance $R(s)$ of the augmented mesurement error $\epsilon(s)$:

$$R(s) = Cov[e(s)] = r(s) \quad ; m(s) = M \qquad \text{(A.12)}$$

$$\begin{aligned} R(s) = diag[&K(s\alpha_1)R(s\alpha_1)K^T(s\alpha_1), \ldots, \\ &K(s\alpha_q)R(s\alpha_q)K^T(s\alpha_q), r(s)] \quad ; m(s) < M \qquad \text{(A.13)} \end{aligned}$$

where $diag[\cdot]$ represents a square matrix with $q+1$ by $q+1$ square blocks. Diagonal blocks $i = 1, \ldots, q$ have dimension $n(s\alpha_i)$ and diagonal block $q+1$ has dimension $n[Y^j(s)]$. All off-diagonal blocks are zero.

The updated state covariance is obtained from the estimation error, which may be written as:

$$\begin{aligned}
&\chi(s) - \hat{\chi}(s|s) = \chi(s) - K(s)Y(s) = \\
&\quad \chi(s) - K(s)[H(s)\chi_M(s) + \epsilon(s)] = \\
&\quad [W(s) - K(S)H(s)]\chi_M(s) - K(s)\epsilon(s) \quad ; m(s) \leq M
\end{aligned} \tag{A.14}$$

This expression yields the following error covariance, which is also the updated covariance for the Gaussian case:

$$\begin{aligned}
&Cov[\chi(s)|s] = E\left[[\chi(s) - \hat{\chi}(s|s)][\chi(s) - \hat{\chi}(s|s)]^T\right] = \\
&\quad [W(s) - K(S)H(s)]Cov[\chi_M(s)][W(s) - K(S)H(s)]^T + \\
&\quad K(s)R(s)K^T(s) \qquad\qquad ; m(s) \leq M
\end{aligned} \tag{A.15}$$

A.3 Sample Mean and Covariance for the Upward Sweep

The proof that the sample multiscale mean and covariance from the upward sweep converge to the exact expressions obtained above follows the analogous proof for the classical ensemble Kalman filter with perturbed measurements. We begin by noting that the sample prior covariance estimates $\widehat{Cov}[\chi(s), Y(s)]$ and $\widehat{Cov}[Y(s)]$ used to compute the sample Kalman gain $K'(s)$ in (4.37) and (4.39) are consistent and converge to the corresponding exact prior covariances as $N \to \infty$. It follows that $K'(s)$ converges to the exact Kalman gain $K(s)$ of (A.7) as $N \to \infty$. With this in mind we set $K'(s) = K(s)$ and can consider the convergence properties of the updated sample mean and covariance.

The expression given in (4.40) for the update of the replicates at node s is:

$$\chi^j(s|s) = \chi^j(s) + K(s)\left[Y^j(s) - \hat{Y}^j(s)\right] \quad ; m(s) \leq M \tag{A.16}$$

We seek the sample mean $\overline{\chi^j(s|s)}$ of $\chi^j(s|s)$, where the overline indicates the arithmetic average. In order to evaluate $\overline{\chi^j(s|s))}$ we define an ensemble $\epsilon^j(s)$ of augmented measurement error replicates, by analogy with (A.10) and (A.11):

$$\epsilon^j(s) = e^j(s) \quad ; m(s) = M \tag{A.17}$$

$$\epsilon^j(s) = \begin{bmatrix} K(s\alpha_1)\epsilon^j(s\alpha_1) \\ \vdots \\ K(s\alpha_q)\epsilon^j(s\alpha_q) \\ e^j(s) \end{bmatrix} \quad ; m(s) < M \tag{A.18}$$

Note that the $\epsilon^j(s)$ have sample means of zero at all nodes.

The perturbed measurement ensemble constructed in (4.32) and (4.34) has the following property:

$$Y^j(s) = Y(s) + \epsilon^j(s) \quad ; m(s) \leq M \tag{A.19}$$

We prove this by induction, noting that if (A.19) holds for the children of s, starting with the finest scale, then the definitions of $Y(s)$ from (A.8) and (A.9) and of $\epsilon^j(s)$ from (A.17) and (A.18) imply:

$$Y^j(s) = y(s) + e^j(s) = Y(s) + \epsilon^j(s) \quad ; m(s) = M \tag{A.20}$$

$$
\begin{aligned}
Y^j(s) &= \begin{bmatrix} K(s\alpha_1)Y^j(s\alpha_1) \\ \vdots \\ K(s\alpha_q)Y^j(s\alpha_q) \\ y(s) + e^j(s) \end{bmatrix} \\
&= \begin{bmatrix} K(s\alpha_1)Y(s\alpha_1) \\ \vdots \\ K(s\alpha_q)Y(s\alpha_q) \\ y(s) \end{bmatrix} + \begin{bmatrix} K(s\alpha_1)\epsilon^j(s\alpha_1) \\ \vdots \\ K(s\alpha_q)\epsilon^j(s\alpha_q) \\ e^j(s) \end{bmatrix} \\
&= Y(s) + \epsilon^j(s) \qquad\qquad ; m(s) < M
\end{aligned}
\tag{A.21}
$$

Using the same approach it also can be shown that:

$$
\hat{Y}^j(s) = h(s)\chi^j(s) \quad ; m(s) = M \tag{A.22}
$$

$$
\begin{aligned}
\hat{Y}^j(s) &= \begin{bmatrix} K(s\alpha_1)\hat{Y}^j(s\alpha_1) \\ \vdots \\ K(s\alpha_q)\hat{Y}^j(s\alpha_q) \\ h(s)\chi^j(s) \end{bmatrix} \\
&= \begin{bmatrix} K(s\alpha_1)H(s\alpha_1)\chi^j_M(s\alpha_1) \\ \vdots \\ K(s\alpha_q)H(s\alpha_q)\chi^j_M(s\alpha_q) \\ h(s)\chi^j_M(s) \end{bmatrix} \\
&= H(s)\chi^j_M(s) \qquad\qquad ; m(s) < M
\end{aligned}
\tag{A.23}
$$

It follows from (A.19) that:

$$
\overline{Y^j(s)} = Y(s) + \overline{\epsilon^j(s)} = Y(s) \quad ; m(s) \leq M \tag{A.24}
$$

Also, (A.22) and (A.23) imply that:

$$\overline{\hat{Y}^j(s)} = H(s)\overline{\chi^j_M(s)} = 0 \quad ; m(s) \leq M \tag{A.25}$$

The desired sample mean of $\chi^j(s|s)$ is then:

$$\overline{\chi^j(s|s)} = K(s)Y(s) = \hat{\chi}(s|s) \quad ; m(s) \leq M \tag{A.26}$$

So the sample mean of the updated ensemble converges to the exact condition mean $\hat{\chi}(s|s)$.

In order to examine convergence to the conditional covariance we compute the deviation of the updated replicate from its sample mean:

$$\begin{aligned}
&\chi^j(s|s) - \overline{\chi^j(s|s)} = \\
&\quad \chi^j(s) + K(s)\left[Y^j(s) - \hat{Y}^j(s)\right] - K(s)\overline{\hat{Y}(s)} = \\
&\quad \chi^j(s) + K(s)[Y(s) + \epsilon^j(s) - H(s)\chi^j_M(s)] - K(s)Y(s) = \\
&\quad \chi^j(s) + K(s)\epsilon^j(s) - K(s)H(s)\chi^j_M(s) = \\
&\quad [W(s) - K(s)H(s)]\chi^j_M(s) + K(s)\epsilon^j(s) \quad ; m(s) \leq M
\end{aligned} \tag{A.27}$$

The sample covariance is then:

$$\begin{aligned}
&\widehat{Cov}[\chi^j(s|s)] = \frac{N}{N-1}\overline{[\chi^j(s|s) - \overline{\chi^j(s|s)}][\chi^j(s|s) - \overline{\chi^j(s|s)}]^T} = \\
&\frac{N}{N-1}[W(s) - K(s)H(s)]\overline{\chi^j_M\chi^{jT}_M}[W(s) - K(s)H(s)]^T + \\
&\frac{N}{N-1}K(s)\overline{\epsilon^j(s)\epsilon^{jT}(s)}K^T(s) \quad ; m(s) \leq M
\end{aligned} \tag{A.28}$$

When the number of replicates approach infinity the arithmetic averages in this expression can be replaced with expectations so (A.28) becomes:

$$
\begin{aligned}
&Cov[\chi(s) - \hat{\chi}(s|s)] = \\
&\qquad [W(s) - K(s)H(s)]Cov[\chi_M(s)[W(s) - K(s)H(s)]^T + K(s)R(S)K^T(s) \\
&\qquad\qquad ; m(s) \leq M \qquad \text{(A.29)}
\end{aligned}
$$

This is the same as the updated Gaussian covariance expression given in (A.15). From these proofs we conclude that the first two sample moments of the ensemble of replicates $\chi^j(s|s)$ generated by the upward sweep of the multiscale update algorithm converge to the corresponding exact updated moments of $\chi(s)$.

Appendix B

Convergence of the Downward Sweep of the Ensemble Multiscale Filter Update

B.1 Introduction

The downward sweep of the ensemble multiscale filter update is designed to produce a set of smoothed replicates $\chi^j(s|S)$ that approximates the Gaussian conditional mean and covariance of $\chi(s)$, given all measurements on the tree. The error in the sample mean and covariance estimates should converge to zero in the limit as the number of replicates approaches infinity.

B.2 Exact Ensemble Mean and Covariance for the Downward Sweep

The exact smoothed conditional Gaussian mean $\hat{\chi}(s|S) = E[\chi(s)|y(\mathcal{V})]$ and covariance $Cov[\chi(s)|S] = Cov[\chi(s)|y(\mathcal{V})]$ are given in [101]:

$$\hat{\chi}(s|S) \;=\; \hat{\chi}(s|s) \qquad ; m(s) = 0 \qquad \text{(B.1)}$$

$$\hat{\chi}(s|S) \quad = \quad \hat{\chi}(s|s) + J(s)[\hat{\chi}(s\gamma|S) - \hat{\chi}(s\gamma|s)] \ ; m(s) > 0 \qquad \text{(B.2)}$$

$$Cov[\chi(s)|S] \quad = \quad Cov[\chi(s)|s] \qquad ; m(s) = 0 \qquad \text{(B.3)}$$

$$\begin{aligned} Cov[\chi(s)|S] = Cov[\chi(s)|s] + \\ J(s)\,[Cov[\chi(s\gamma)|S)] - Cov[\chi(s\gamma)|s)]]\, J^T(s) \ ; m(s) > 0 \end{aligned} \qquad \text{(B.4)}$$

where:

$$J(s) \quad = \quad Cov[\chi(s)|s]F^T(s)Cov^{-1}[\chi(s\gamma|s)] \qquad ; m(s) > 0 \qquad \text{(B.5)}$$

The matrix $J(s)$ is a smoothing gain analogous to the Kalman gain $K(s)$.

The updated estimate $\hat{\chi}(s|s)$ and covariance $Cov[\chi(s)|s]$ appearing in (B.2) and (B.4) are defined by the upward sweeep. The smoothed mean $\hat{\chi}(s|S)$ and covariance $Cov[\chi(s\gamma)|S]$ are defined by the previous iteration of the downward sweep. The fine to coarse projection mean $\hat{\chi}(s\gamma|s)$ and conditional covariance matrix $Cov[\chi(s\gamma)|s]$ are defined by:

$$\hat{\chi}(s\gamma|s) \quad = \quad F(s)\hat{\chi}(s|s) \qquad ; m(s) > 0 \qquad \text{(B.6)}$$

$$\begin{aligned} Cov[\chi(s\gamma)|s] &= F(s)Cov[\chi(s)|s]F(s)^T + Cov[\chi(s\gamma)] - \\ &\quad F(s)A(s)Cov[\chi(s\gamma)] \\ &= F(s)Cov[\chi(s)|s]F(s)^T + Q'(s) \ ; m(s) > 0 \end{aligned} \qquad \text{(B.7)}$$

B.3 Sample Mean and Covariance for the Downward Sweep

First we consider the value of the gain $J'(s)$ computed from sample covariances. Unless otherwise noted all expressioins apply for $m(s) > 0$:

$$J'(s) \quad = \quad \widehat{Cov}[\chi(s)|s]F^T(s)\widehat{Cov}^{-1}[\chi(s\gamma)|s] \tag{B.8}$$

Appendix A shows that the sample covariance $\widehat{Cov}[\chi(s)|s]$ obtained during the upward sweep converges to the corresponding exact covariance $Cov[\chi(s)|s]$ in the limit as $N \to \infty$. The upward projection sample covariance $\widehat{Cov}[\chi(s\gamma)|s]$ can be written as:

$$\widehat{Cov}[\chi(s\gamma|s)] \quad = \quad F(s)\widehat{Cov}[\chi(s|s)]F^T(s) + \widehat{Cov}[w'(s)] \tag{B.9}$$

Since Appendix A shows that $\widehat{Cov}[\chi(s|s)]$ converges to $Cov[\chi(s)|s]$ and since $w'(s)$ is generated to insure that $\widehat{Cov}[w'(s)]$ converges to $Q'(s)$ it follows that $\widehat{Cov}[\chi(s\gamma)|s]$ converges to the value of $Cov[\chi(s\gamma)|s]$ given in (B.7). Therefore, $J'(s)$ converges to the exact gain $J(s)$ of (B.5) as $N \to \infty$. With this in mind we set $J'(s) = J(s)$ and consider the convergence properties of the smoothed sample mean and covariance. For completeness, note that $\widehat{Cov}[\chi(s\gamma))|S]$ converges to $Cov[\chi(s\gamma))|S]$. This follows by applying induction to (B.4), using the fact that $\widehat{Cov}[\chi(s)|S]$ converges to $Cov[\chi(s)|S]$ at the root node $s = 0$ and all the other sample covariances appearing in (B.4) converge to their exact counterparts.

Equation (4.44) describes the replicate update on the downward sweep:

$$\chi^j(s|S) \quad = \quad \chi^j(s|s) + J'(s)[\chi^j(s\gamma|S) - \chi^j(s\gamma|s)] \tag{B.10}$$

We seek the sample mean $\overline{\chi^j(s|S)}$ of $\chi^j(s|S)$, where the overline indicates the arithmetic average. Taking the sample mean of both sides of (B.10) gives

$$\begin{aligned} \overline{\chi^j(s|S)} &= \overline{\chi^j(s|s)} + J'(s)[\overline{\chi^j(s\gamma|S)} - \overline{\chi^j(s\gamma|s)} \\ &= \hat{\chi}(s|s) + J(s)[\overline{\chi^j(s\gamma|S)} - \overline{\chi^j(s\gamma|s)}] \qquad \text{(B.11)} \end{aligned}$$

where the second equality relies on the proof from Apppendix A that shows that the updated sample mean on the upward sweep converges to the exact updated mean.

The projected replicate $\chi^j(s\gamma|s)$ appearing in (B.11) is given by (4.41):

$$\chi^j(s\gamma|s) = F(s)\chi^j(s|s) + w'^j(s) \qquad \text{(B.12)}$$

The sample mean of this expression is:

$$\begin{aligned} \overline{\chi^j(s\gamma|s)} &= F(s)\overline{\chi^j(s|s)} + w'^j(s) \\ &= F(s)\hat{\chi}(s|s) \\ &= \hat{\chi}(s\gamma|s) \qquad \text{(B.13)} \end{aligned}$$

where the second equality follows from Appendix A and the third equality follows from (B.6).

Since Appendix A confirms that $\chi^j(s|S)$ converges to $\hat{\chi}(s|S) = \hat{\chi}(s|s)$ at $s = 0$ it follows from induction that (B.11) can be written:

$$\begin{aligned} \overline{\chi^j(s|S)} &= \hat{\chi}(s|s) + J'(s)[\hat{\chi}(s\gamma|S) - \hat{\chi}(s\gamma|s)] \\ &= \hat{\chi}(s|S) \qquad \text{(B.14)} \end{aligned}$$

So the sample mean of the smoothed ensemble generated on the downward sweep converges to the corresponding exact mean.

In order to examine convergence of the sample covariance rearrange (4.44) as follows:

$$\chi^j(s|S) - J(s)\chi^j(s\gamma|S) = \chi^j(s|s) - J(s)\chi^j(s\gamma|s) \qquad \text{(B.15)}$$

Subtract the sample mean of (B.15) from each side, multiply the result by its transpose, and take the sample mean of the product expression to get:

$$\begin{aligned} &\widehat{Cov}[\chi(s)|S] + J(s)\widehat{Cov}[\chi(s\gamma)|S]J(s)^T - D_1 - D_1^T = \\ &\quad \widehat{Cov}[\chi(s)|s] - D_2 - D_2^T + J(s)\widehat{Cov}[\chi(s\gamma)|s]J(s)^T \end{aligned} \tag{B.16}$$

where the temporary matrices D_1 and D_2 account for the cross-covariances generated by the multiplication:

$$D_1 = J(s)\overline{\tilde{\chi}^j(s\gamma|S)\tilde{\chi}^{jT}(s|S)} \tag{B.17}$$

$$D_2 = J(s)\overline{\tilde{\chi}^j(s\gamma|s)\tilde{\chi}^{jT}(s|s)} \tag{B.18}$$

Here the symbol $\tilde{}$ represents the deviation of the indicated replicate from its sample mean.

The first step in evaluating the terms D_1 and D_2 is to use (B.10) to write an expression for the deviation of the smoothed replicate from its mean:

$$\tilde{\chi}^j(s|S) = \tilde{\chi}^j(s|s) + J(s)(\tilde{\chi}^j(s\gamma|S) - \tilde{\chi}^j\chi(s\gamma|s)) \tag{B.19}$$

Then substitute (B.19) into (B.17) to get:

$$\begin{aligned} D_1 = \quad & J(s)\left[\overline{\tilde{\chi}^j(s\gamma|S)\tilde{\chi}^{jT}(s|s)} - \overline{\tilde{\chi}^j(s\gamma|S)\tilde{\chi}^{jT}(s\gamma|s)J^T(s)}\right] \\ & + J(s)\widehat{Cov}[\chi(s\gamma)|S]J^{jT}(s) \end{aligned} \tag{B.20}$$

The next step is to evaluate the individual overbarred terms in this expression.

To do this, note that the exact smoothed conditional mean $\hat{\chi}(s\gamma|S)$ depends on all the measurements on the tree. Consequently, it may be written in a batch form. The batch expression for $\hat{\chi}(s\gamma|S)$ is just the conditional expectation of the variable $\chi(s\gamma)$ given the vector of all measurements $y(\mathcal{V})$. This conditional expectation has the form:

$$\hat{\chi}(s\gamma|S) = \hat{\chi}(s\gamma|y(\mathcal{V})) = K(s\gamma)\begin{bmatrix} v_1(s\gamma) \\ v_2(s\gamma) \\ \vdots \\ v_m(s\gamma) \end{bmatrix} \tag{B.21}$$

where $v_i(s\gamma)$, $i = 1, 2, \ldots, m$ are the deviations (or innovations) of the m measurements in $\mathcal{V}$ from the corresponding measurement predictions and $K(s\gamma)$ is an $n(s\gamma)$ by m dimensional block diagonal gain matrix. The measurement predictions used to obtain the $v_i(s\gamma)$'s are based on the updated state estimate obtained from the upward sweep.

With suitable rearrangement and an orthogonalization transformation of the innovations elements (B.21) may be written as:

$$\hat{\chi}(s\gamma|S) = \tilde{K}(s\gamma)\begin{bmatrix} v_s \\ v_{sc} \end{bmatrix} = \tilde{K}_s v_s + \tilde{K}_{sc} v_{sc} \tag{B.22}$$

where the innovations elements in v_s are linear functions of the measurements at nodes below $s\gamma$ while those in v_{sc} are linear functions of measurements at all other nodes, including $s\gamma$. These two subvectors are orthogonal by construction so $E[v_s v_{sc}^T] = 0$. The matrix $\tilde{K}(s\gamma)$ is the corresponding transformed gain matrix, constructed from the elements of $K(s\gamma)$. From the definitions of $v(s\gamma|s)$ it follows that:

$$\hat{\chi}(s\gamma|s) = \tilde{K}_s v_s \tag{B.23}$$

$$\hat{\chi}(s\gamma|S) = \hat{\chi}(s\gamma|s) + \tilde{K}_{sc} v_s \tag{B.24}$$

The ortogonalization operation leading to (B.23) and (B.24) is discussed in more detail in [35].

The exact smoothed conditional mean can be approximated with an ensemble Kalman filter update that has exactly the same form as (B.23) and (B.24), but is applied to individual replicates:

$$\begin{aligned} \chi^j(s\gamma|s) &= \tilde{K'}_s v_s^j & \text{(B.25)} \\ \chi^j(s\gamma|S) &= \chi^j(s\gamma|s) + \tilde{K'}_{sc} v_{sc}^j & \text{(B.26)} \end{aligned}$$

where $\tilde{K'}_s$ and $\tilde{K'}_{sc}$ are sample approximations of $\tilde{K}_s$ and $\tilde{K}_{sc}$. Using arguments similar to those used for the Kalman and smoothing gains on the upward and downward sweeps we can show that these batch sample Kalman gains converge to their exact counterparts as $N \to \infty$. So in the following we set $\tilde{K'}_s = \tilde{K}_s$ and $\tilde{K'}_{sc} = \tilde{K}_{sc}$.

Now subtract the sample mean of (B.25) from (B.25)to get:

$$\tilde{\chi}^j(s\gamma|S) = \tilde{\chi}^j(s\gamma|s) + \tilde{K}_{sc}\tilde{v}_{sc}^j \qquad \text{(B.27)}$$

Substitute (B.27) into (B.20)to get:

$$\begin{aligned} D_1 &= J(s)\overline{\tilde{\chi}^j(s\gamma|s)\tilde{\chi}^{jT}(s|s)} + J(s)\tilde{K}_{sc}\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s|s)} - \\ &\quad J(s)\overline{\tilde{\chi}^j(s\gamma|s)\tilde{\chi}^{jT}(s\gamma|s)}J^T(s) - J(s)\tilde{K}_{sc}\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s\gamma|s)}J^T(s) \\ &\quad +J(s)\widehat{Cov}[\chi(s\gamma)|S]J(s)^T & \text{(B.28)} \end{aligned}$$

Then apply the definition of D_2 from (B.18):

$$\begin{aligned} D_1 &= D_2 + J(s)\tilde{K}_{sc}\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s|s)} - \\ &\quad J(s)\widehat{Cov}[\chi(s\gamma)|s]J^T(s) - J(s)\tilde{K}_{sc}\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s\gamma|s)}J^T(s) \\ &\quad +J(s)\widehat{Cov}[\chi(s\gamma)|S]J(s)^T & \text{(B.29)} \end{aligned}$$

Since v_s and v_{sc} are orthogonal by construction and $\hat{\chi}(s\gamma|s)$ is proportional to both v_s and to $\hat{\chi}(s|s)$ the vectors v_{sc} and $\hat{\chi}(s|s)$ are also orthogonal. That is, $E[v_{sc}\hat{\chi}^T(s|s)] = 0$. Since the $\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s|s)}$ term appearing in (B.29) converges to $E[v_{sc}\hat{\chi}^T(s|s)]$ it is zero in the limit.

The term $\overline{\tilde{v}_{sc}^j\tilde{\chi}^{jT}(s\gamma|s)}$ also appearing in (B.29) involves the deviation of the projected replicate $\chi^j(s\gamma|s)$ from its mean. Subtract the first line of (B.13) from (B.12) to obtain the following equation for this deviation:

$$\tilde{\chi}^j(s\gamma|s) \quad = \quad F(s)\tilde{\chi}^j(s|s) + w'^j(s) \tag{B.30}$$

Multiply the transpose of this equation from the left by $\tilde{v}^j_{sc}$ and take the sample mean to get:

$$\overline{\tilde{v}^j_{sc}\tilde{\chi}^j(s\gamma|s)} = \overline{\tilde{v}^j_{sc}\tilde{\chi}^{jT}(s|s)}F^T(s) + \overline{\tilde{v}^j_{sc}w'^j(s)} \tag{B.31}$$

As $N \rightarrow \infty$ the sample mean in the first term approaches zero because its limit $E[\tilde{v}_{sc}\tilde{\chi}^T(s|s)]$ is zero by by

construction of v_{sc}. The second term approaches zero because the random perturbation at s is uncorrelated with the states and measurements that $\tilde{v}_{sc}$ depends upon.

If we substitute (B.29) in (B.16) and take the limit, noting that the terms containing $\tilde{v}^j_{sc}$ go to zero and the sample covariances approach the true covariances, as discussed earlier, we obtain the exact smooting covariance from (B.4) :

$$\begin{aligned} Cov[\chi(s)|S] = Cov[\chi(s)|s] + \\ J(s)\left[Cov[\chi(s\gamma)|S)] - Cov[\chi(s\gamma)|s)]\right] J^T(s) \end{aligned} \tag{B.32}$$

Appendix C

Proof of the universal downscaling matrix

Denote $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, the prior states of two nodes at fine scale, as X, the coarse resolution state x_0 is a linear transform of its children X: $x_0 = AX$. The observation equation of the coarse scale state is $y = Hx_0 + v$, where v is uncorrelated with any state with error covariance R. Suppose the actual measurement is y^o, we need to estimate fine scale states X. According to linear estimation theory the linear least square estimate of the fine scale states follows

$$\hat{X} \;=\; X + K(y^o - Hx_0) \tag{C.1}$$

where

$$K = P_X A^T H^T (H P_{x_0} H^T + R)^{-1} \tag{C.2}$$

Rewrite K in (C.2) as

$$\begin{aligned} K &= P_X A^T (P_{x_0}^{-1} P_{x_0}) H^T (H P_{x_0} H^T + R)^{-1} && \text{(C.3)} \\ &= P_X A^T P_{x_0}^{-1} [P_{x_0} H^T (H P_{x_0} H^T + R)^{-1}] && \text{(C.4)} \end{aligned}$$

Since at coarse scale the optimal estimate is

$$\hat{x_0} \;=\; x_0 + k_0(y^o - Hx_0) \tag{C.5}$$

where

$$k_0 = P_{x_0}H^T(H^TP_{x_0}H^T + R)^{-1} \tag{C.6}$$

Using C.6, C.5, C.4, and C.1 we have

$$\hat{X} \;=\; X + P_XA^TP_{x_0}^{-1}k_0(y^o - Hx_0) \tag{C.7}$$

$$\;=\; X + J(\hat{x_0} - x_0) \tag{C.8}$$

where the optimal downscaling matrix J is defined as

$$J \;=\; P_XA^TP_{x_0}^{-1} \tag{C.9}$$

Hence, as long as the optimal estimate of the coarse scale states are available, the optimal fine scale states can be obtained using (C.8)

Appendix D

Conditional Simulation of Soil Properties Using MCMC Method

D.1 Gibbs Sampler

Markov chain Monte Carlo (MCMC) method [42] is a useful tool for the simulation of any distribution on a finite-dimensional state space specified by any unnormalized density. It draws samples from approximate distributions and then corrects those draws to better approximate the target distribution. The samples are drawn sequentially, with the distribution of the sampled draws only dependent on the last draw. Hence, the draws form a Markov Chain. The markovianity helps prove the convergence of the sampling distribution to the true target distribution. The key to the success of the simulation is that every time the samples are corrected. A sufficient condition for a Markov chain having a unique stationary distribution is that the *detailed balance equation (reversibility condition)* holds:

$$T(x_j, x_k)\pi_j = T(x_k, x_j)\pi_k$$

where $T(x_j, x_k)$ is the transition probability from x_j to x_k, π_k is the event probability of event x_k.

Among many MCMC techniques, the Gibbs sampler [43], the Metropolis and

Metropolis-Hastings algorithms [50] are popular.

Suppose we want to draw sample from any pdf $p(x)$. One can construct a transition function $T(x_{t-1}, x_t)$ for a Markov process as follows:

- Start with any reasonable initial guess x_0.

- For t=1:T Sample a new x_t^* from a proposal density $J(x_t|x_{t-1})$, accept x_t^* as a new sample x^t from p(x) with probability min(r,1)

  $$r = \frac{p(x_t^*)/J(x_t^*|x_{t-1})}{p(x_{t-1})/J(x_{t-1}|x_t^*)}$$

  Otherwise, $x_t = x_{t-1}$

This is the Metropolis-Hastings algorithm. If we use a special proposal distribution $J(x_t|x_{t-1})$, the algorithm is called Gibbs sampler. Then

$$J_{j,t}(x_t^*|x_{t-1}) = \begin{cases} p(x_{j,t}^*|x_{-j,t-1}) & if\ x_{-j,t}^* = x_{-j,t-1} \\ 0\ otherwise & \end{cases}$$

where subscript j is the jth component of x,$-j$ are the rest components, t is the iteration number, each component has its own proposal density $J_{j,t}$, which essentially only updates jth component conditioned on all the rest states. In this case, acceptance ratio r is

$$\begin{aligned} r &= \frac{p(x_t^*)/J_{j,t}(x_t^*|x_{t-1})}{p(x_{t-1})/J_{j,t}(x_{t-1}|x_t^*)} = \frac{p(x_t^*)/p(x_{j,t}^*|x_{-j,t-1})}{p(x_{t-1})/p(x_{j,t-1}|x_{-j,t}^*)} \quad &\text{(D.1)} \\ &= \frac{p(x_{j,t}^*, x_{-j,t}^*)/p(x_{j,t}^*|x_{-j,t-1})}{p(x_{j,t-1}, x_{-j,t-1})/p(x_{j,t-1}|x_{-j,t}^*)} \quad &\text{(D.2)} \\ &= \frac{p(x_{j,t}^*, x_{-j,t-1})/p(x_{j,t}^*|x_{-j,t-1})}{p(x_{j,t-1}, x_{-j,t-1})/p(x_{j,t-1}|x_{-j,t-1})} \quad &\text{(D.3)} \\ &= \frac{p(x_{-j,t-1})}{p(x_{-j,t-1})} = 1 \quad &\text{(D.4)} \end{aligned}$$

D.2 Gaussian Markov Random Field

A Markov random field is a field which satisfies the conditional independence property:

$$p(x_{i,j}|x_{k,l},(k,l)\in\Lambda\backslash\{(i,j)\}) = p(x_{i,j}|x_{k,l},(k,l)\in\partial_{i,j})$$

where $\Lambda\backslash\{(i,j)$ represents all the pixels in the full domain Λ excluding pixel (i,j), $\partial_{i,j}$ is the neighbor of pixel (i,j). In Figure one can assume a particular pixel α only has neighbor size of 2, pixel β has neighbor size of 1.

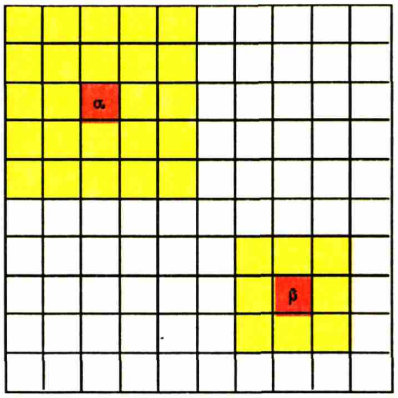


Figure D-1: Markov random field

A Gaussian field is a random field which can be described using a multivariate Gaussian distribution. Mean and covariance completely determine the distribution. If the field is also Markov, it is called a Gaussian Markov random field. For a GRMF, the PDF of x follow

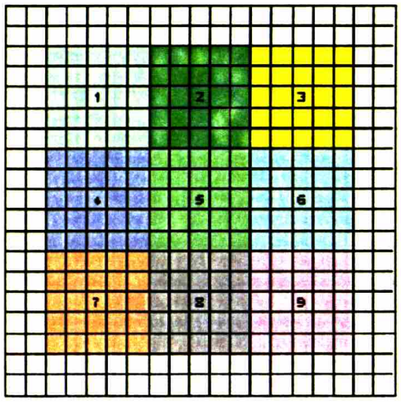
$$x \sim \exp(-(x-u)^T C^{-1}(x-u))$$

where C is the covariance matrix of the random field. Given the value of a part of the field, say B, the distribution of the rest of the field, say A, follows

$$x_{A|B} \sim N(u_A - C_{AB}C_{BB}^{-1}(x_B - u_B), C_{AA} - C_{AB}C_{BB}^{-1}C_{BA}) \tag{D.5}$$

To generate a Gaussian Markov random field, using Gibbs sampler one can do blockwise update. The field shown in Figure D.2 is first broken into 9 blocks, and use the neighbor size of 2. Then draw a random sample for all the pixels from any distribution to initialize the field.

```
For iteration $i = 1 : T$
    for b=1:9
        Do conditional simulation for block b based on its latest
        neighbor using the conditional distribution (D.5) ,
    end
End
```



D.3 Simulation of Spatially Correlated Random Field of Soil Properties

To generate spatially correlated soil random field, we can use the Gibbs sampler and GMRF approximation. In this thesis only sand and clay fraction need to be generated. Suppose we want a Gaussian random field for the fraction. Since the fraction must be greater than zero and the sum of clay and sand must be less than one, the joint PDF of sand and clay fraction to be sampled from is a truncated Gaussian distribution. In this case, the conditional distribution in (D.5) is only an approximation. Assume a spatial correlation function and calculate $C_{AB}C_{BB}^{-1}$ and $C_{AA} - C_{AB}C_{BB}^{-1}C_{BA}$ beforehand based on the spatial correlation function. The mean of the random field is determined by the specific STATSGO soil type of that particular pixel. For a particular soil type, the sand and clay fraction of the center point of the portion in the soil triangle is used. The standard deviation of clay and sand fraction is set to 0.1 and 0.2 respectively. The reason of using different standard deviation for sand and clay is because this helps convergence of the simulation. A spherical correlation function is used to model the correlation of sand or clay fraction. Here is the procedure for doing the conditional simulation for one replicate:

```
For iteration $i = 1 : T$
    for b=1:9
        while(stop==0)
            Do conditional simulation for block b based on its
            latest neighbor using the calculated coefficients,
            If(sand fraction>0 and clay fraction>0 and sand fraction +clay fraction<1)
                stop=1;
            end
        end
    end
end
```

Since this approach is an iterative approach, and there is an approximation in the conditional update, as the total number of iteration increases, the error would accumulate. So T is set to a small number 200. Usually this is not the converged solution, whatever the random field is like at this stage would be used as the final sample. Thus, the sampling statistics would deviate away from the specified variance and correlation structure. However, the random field looks still reasonable. A further study of this is still needed.

Appendix E

Vegetation Parameters

Type	Category name	$s(cm)$	h	ω	b	b_v	b_h	f_T
0	Bare soil	-	-	-	-	-	-	
1	Temperate needleleaf evergreen tree	1	0.1	0.12	0.1	0.12	0.08	0.8
2	Boreal needleleaf evergreen tree	1	0.1	0.12	0.1	0.12	0.08	0.8
3	Boreal needleleaf deciduous tree	1	0.1	0.12	0.1	0.12	0.08	0.8
4	Tropical broadleaf evergreen tree	1	0.1	0.12	0.12	0.144	0.096	0.8
5	Temperate broadleaf evergreen tree	1	0.1	0.12	0.12	0.144	0.096	0.8
6	Tropical broadleaf deciduous tree	1	0.1	0.12	0.12	0.144	0.096	0.8
7	Temperate broadleaf deciduous tree	1	0.1	0.12	0.12	0.144	0.096	0.8
8	Boreal broadleaf deciduous tree	1	0.1	0.12	0.12	0.144	0.096	0.8
9	Broadleaf evergreen temperate shrub	1	0.1	0.12	0.11	0.121	0.099	0
10	Broadleaf deciduous temperate shrub	1	0.1	0.12	0.11	0.121	0.099	0
11	Broadleaf deciduous boreal shrub	1	0.1	0.12	0.11	0.121	0.099	0
12	c3 grass	1.2	0.12	0.05	0.13	0.143	0.117	0.1
13	c3 arctic grass	1.2	0.12	0.05	0.13	0.143	0.117	0.1
14	c4 grass	1.2	0.12	0.05	0.13	0.143	0.117	0.1
15	Crop 1 (e.g. corn)	1.2	0.12	0.05	0.13	0.143	0.117	0.1
16	crop 2 (e.g. wheat)	1.2	0.12	0.05	0.13	0.143	0.117	0.1
21	Water	-	-	-	-	-	-	
22	Wetlands	-	-	-	-	-	-	
23	Urban and built-up	-	-	-	-	-	-	
24	Snow and ice	-	-	-	-	-	-	

Table E.1: CLM land cover classifications and representative roughness and vegetation parameters

Type	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
1	4.1	4.2	4.6	4.8	4.9	5.0	4.8	4.7	4.6	4.2	4.0	4.0
2	4.1	4.2	4.6	4.8	4.9	5.0	4.8	4.7	4.6	4.2	4.0	4.0
3	0.0	0.0	0.0	0.6	1.2	2.0	2.6	1.7	1.0	0.5	0.2	0.0
4	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
6	0.8	0.7	0.4	0.5	0.5	0.7	1.7	3.0	2.5	1.6	1.0	1.0
7	0.0	0.0	0.3	1.2	3.0	4.7	4.5	3.4	1.2	0.3	0.0	0.0
8	0.0	0.0	0.3	1.2	3.0	4.7	4.5	3.4	1.2	0.3	0.0	0.0
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10	0.9	0.8	0.2	0.2	0.0	0.0	0.0	0.2	0.4	0.5	0.6	0.8
11	0.0	0.0	0.0	0.0	0.0	0.2	1.4	1.2	0.0	0.0	0.0	0.0
12	0.4	0.5	0.6	0.7	1.2	3.0	3.5	1.5	0.7	0.6	0.5	0.4
13	0.4	0.5	0.6	0.7	1.2	3.0	3.5	1.5	0.7	0.6	0.5	0.4
14	0.4	0.5	0.6	0.7	1.2	3.0	3.5	1.5	0.7	0.6	0.5	0.4
15	0.0	0.0	0.0	0.0	1.0	2.0	3.0	3.0	1.5	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	1.0	2.0	3.0	3.0	1.5	0.0	0.0	0.0

Table E.2: Monthly Leaf Area Index (LAI) for CLM land cover type

Type	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.4	0.5	0.4	0.3	0.4	0.5	0.5	0.6	0.6	0.7	0.6	0.5
2	0.4	0.5	0.4	0.3	0.4	0.5	0.5	0.6	0.6	0.7	0.6	0.5
3	0.3	0.3	0.3	0.4	0.4	0.4	1.7	1.2	1.0	0.8	0.6	0.5
4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
6	0.4	0.3	0.5	0.3	0.3	0.3	0.3	0.7	0.7	1.1	0.9	0.2
7	0.4	0.4	0.4	0.4	0.5	0.4	0.9	1.4	2.6	1.4	0.6	0.4
8	0.4	0.4	0.4	0.4	0.5	0.4	0.9	1.4	2.6	1.4	0.6	0.4
9	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
10	0.1	0.2	0.6	0.1	0.6	0.0	0.1	0.1	0.1	0.1	0.1	0.1
11	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	1.4	0.1	0.1	0.1
12	0.3	0.3	0.3	0.3	0.3	0.4	0.8	2.3	1.1	0.4	0.4	0.4
13	0.3	0.3	0.3	0.3	0.3	0.4	0.8	2.3	1.1	0.4	0.4	0.4
14	0.3	0.3	0.3	0.3	0.3	0.4	0.8	2.3	1.1	0.4	0.4	0.4
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table E.3: Monthly Stem Area Index (SAI) for CLM land cover type

Type	Top (m)	Bottom(m)
1	17.0	8.5
2	17.0	8.5
3	14.0	7.0
4	25.0	1.0
5	25.0	1.0
6	18.0	10.0
7	20.0	11.5
8	20.0	11.5
9	0.5	0.1
10	0.5	0.1
11	0.5	0.1
12	0.5	0.0
13	0.5	0.0
14	0.5	0.01
15	0.5	0.0
16	0.5	0.0

Table E.4: Vegetation canopy top and bottom height (m) for CLM land cover type

Bibliography

[1] Ncep/ncar reanalysis 1 data. *http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.html.*

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Ieee Transactions on Signal Processing*, 50(2):174–188, 2002.

[3] L Auger and AV Tangborn. A wavelet-based reduced rank kalman filter for assimilation of stratospheric chemical tracer observations. *MONTHLY WEATHER REVIEW*, 132(5):1220–1237, 2004.

[4] L. M. Berliner, R. F. Milliff, and C. K. Wikle. Bayesian hierarchical modeling of air-sea interaction. *Journal of Geophysical Research-Oceans*, 108(C4):–, 2003.

[5] G. B. Bonan. A land surface model (lsm version 1.0) for ecological, hydrological, and atmospheric studies: Technical description and user's guide." ncar tech. note ncar/tn-4171str. *http://www.cgd.ucar.edu/tss/lsm/availability/technote.tar.Z.*

[6] G. Boni, D. Entekhabi, and F. Castelli. Land data assimilation with satellite measurements for the estimation of surface energy balance components and surface control on evaporation. *Water Resources Research*, 37(6):1713–1722, 2001.

[7] R. J. Bouchet. Evapotranspiration reelle et potentielle, signification climatique. *General Assembly Berkeley, Int. Assoc. of Sci. Hydrol. , Gentbrugge, Belgium*, 62:134C142, 1963.

[8] G. Boulet, Y. Kerr, A. Chehbouni, and J. D. Kalma. Deriving catchment-scale water and energy balance parameters using data assimilation based on extended kalman filtering. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 47(3):449–467, 2002.

[9] A Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31(7):333–390, 1977.

[10] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[11] U. Callies, A. Rhodin, and D. P. Eppel. A case study on variational soil moisture analysis from atmospheric observations. *Journal of Hydroly*, 212:95–108, 1998.

[12] M. A. Cane, A. Kaplan, R. N. Miller, B. Tang, E. C. Hackert, and A. J. Busaslacchi. Mapping tropical pacific sea level: Data assimilation via a reduced state space kalman filter. *J. Geophys. Res.*, 101C(22):59922 617, 1996.

[13] F. Caparrini, F. Castelli, and D. Entekhabi. Estimation of surface turbulent fluxes through assimilation of radiometric surface temperature sequences. *Journal of Hydrometeorology*, 5(1):145–159, 2004.

[14] F. Castelli, D. Entekhabi, and E. Caporali. Estimation of surface heat flux and an index of soil moisture using adjoint-state surface energy balance. *Water Resoures Research*, 35(10):3115–3125, 1999.

[15] Virat Chatdarong. *Multi-Sensor Rainfall Data Assimilation using Ensemble Kalman Filter.* MIT Ph.D. Thesis, 2005.

[16] K. S. Chen, S. K. Yen, and W. P. Huang. A simple-model for retrieving bare soil-moisture from radar-scattering coefficients. *Remote Sensing of Environment*, 54(2):121–126, 1995.

[17] T. M. Chin, A. J. Mariano, and E. P. Chassignet. Spatial regression and multiscale approximations for sequential data assimilation in ocean models. *J. Geophys. Res.*, 104:79918014, 1999.

[18] K. C. Chou, A. S. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization. *Ieee Transactions on Automatic Control*, 39(3):464–478, 1994.

[19] B. J. Choudhury, T. J. Schmugge, A. Chang, and R. W. Newton. Effect of surface-roughness on the microwave emission from soils. *Journal of Geophysical Research-Oceans and Atmospheres*, 84:5699–5706, 1979.

[20] S. E. Cohn and R. Todling. Approximate data assimilation schemes for stable and unstable dynamics. *J. Meteor. Soc. Japan*, 74:6375, 1996.

[21] D. Crommelin and A. Majda. Strategies for model reduction: comparing different optimal bases. *Journal of the Atmospheric Sciences*, 61:2206–2217, 2004.

[22] W. T. Crow and W. P. Kustas. Utility of assimilating surface radiometric temperature observations for evaporative fraction and heat transfer coefficient retrieval. *Boundary-Layer Meteorology*, 115(1):105–130, 2005.

[23] W. T. Crow and E. F. Wood. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble kalman filtering: a case study based on estar measurements during sgp97. *Advances in Water Resources*, 26(2):137–149, 2003.

[24] R. Daley. *Atmospheric Data Analysis.* Cambridge University Press, New York, 1992.

[25] M. M. Daniel and A. S. Willsky. A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing. *Proceedings of the Ieee*, 85(1):164–180, 1997.

[26] M. C. Dobson, F. T. Ulaby, M. T. Hallikainen, and M. A. El-Rayes. Microwave dielectric behavior of wet soil - part ii: dielectric mixing models. *IEEE Trans. Geosci. Remote Sens*, GE-23:35–46, 1985.

[27] Arnaud Doucet, Nando de Freitas, Kevin Murphy, and Stuart Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence,http://citeseer.ist.psu.edu/article/doucet00raoblackwellised.html*, pages 176–183, 2000.

[28] P. C. Dubois, J. vanZyl, and T. Engman. Measuring soil moisture with imaging radars (vol 33, pg 915, 1995). *Ieee Transactions on Geoscience and Remote Sensing*, 33(6):1340–1340, 1995.

[29] S. Dunne and D. Entekhabi. An ensemble-based reanalysis approach to land data assimilation. *Water Resources Research*, 41(2):–, 2005.

[30] M. Ehrendorfer and J. Tribbia. Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.*, 54:286313, 1997.

[31] D. Entekhabi, H. Nakamura, and E. G. Njoku. Solving the inverse problem for soil moisture and temperature profiles by sequential assimilation of multifrequency remotely sensed observations. *IEEE Trans. Geosci. Remote Sens.*, 32(2):438–448, 1994.

[32] D. Entekhabi, E. G. Njoku, P. Houser, M. Spencer, T. Doiron, Y. J. Kim, J. Smith, R. Girard, S. Belair, W. Crow, T. J. Jackson, Y. H. Kerr, J. S. Kimball, R. Koster, K. C. McDonald, P. E. O'Neill, T. Pultz, S. W. Running, J. C. Shi, E. Wood, and J. van Zyl. The hydrosphere state (hydros) satellite mission: An earth system pathfinder for global mapping of soil moisture and land freeze/thaw. *Ieee Transactions on Geoscience and Remote Sensing*, 42(10):2184–2195, 2004.

[33] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte-carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*, 99(C5):10143–10162, 1994.

[34] G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

[35] P. Fieguth, D. Menemenlis, T. Ho, A. Willsky, and C. Wunsch. Mapping mediterranean altimeter data with a multiresolution optimal interpolation algorithm. *Journal of Atmospheric and Oceanic Technology*, 15(2):535–546, 1998.

[36] David Flagg. *Great Plains Project Data Documentation.* MIT Civil and Environmental Engineering Project Report, Cambridge, MA, 2005.

[37] A. B. Frakt and A. S. Willsky. Computationally efficient stochastic realization for internal multiscale autoregressive models. *Multidimensional Systems and Signal Processing*, 12(2):109–142, 2001.

[38] I. Fukumori and P. Malanotte-Rizzoli. An approximate kalman filter for ocean data assimilation: An example with an idealized gulf stream model. *J. Geophys. Res.*, 100C:67776794, 1995.

[39] A. K. Fung, Z. Q. Li, and K. S. Chen. Backscattering from a randomly rough dielectric surface. *Ieee Transactions on Geoscience and Remote Sensing*, 30(2):356–369, 1992.

[40] G. Gaspari and S. E. Cohn. Construction of correlation func- tions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, 125:723–757, 1999.

[41] Arthur Gelb. *Applied Optimal Estimation.* MIT Press, Cambridge, MA, 1974.

[42] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, London, 2004.

[43] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(11):721–741, 1984.

[44] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in Geophysics*, 33:141–266, 1991.

[45] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel-approach to nonlinear non-gaussian bayesian state estimation. *Iee Proceedings-F Radar and Signal Processing*, 140(2):107–113, 1993.

[46] I. P. Gorenburg, D. McLaughlin, and D. Entekhabi. Scale-recursive assimilation of precipitation data. *Advances in Water Resources*, 24(9-10):941–953, 2001.

[47] C. Grassotti, R. N. Hoffman, E. R. Vivoni, and D. Entekhabi. Multiple-timescale intercomparison of two radar products and rain gauge observations over the arkansas-red river basin. *Weather and Forecasting*, 18(6):1207–1229, 2003.

[48] Waymire E.C. Gupta, V.K. A statistical-analysis of mesoscale rainfall as a random cascade. *J. Appl. Meteor.*, 32:251–267, 1993.

[49] J. Hammersley and D. Handscomb. *Monte Carlo Methods.* Methuen, London, 1964.

[50] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[51] K.L. Hawk and P. Eagleson. Climatology of station storm rainfall in the continental u.s.: Parameters of the bartlett-lewis and poisson rectangular pulses models. *Technical Report 336, Massachusetts Institute of Technology, Department of Civil Engineering*, 1992.

[52] AW Heemink, M Verlaan, and AJ Segers. Variance reduced ensemble kalman filtering. *Monthly Weather Review*, 129(7):1718–1728, 2001.

[53] T. T. Ho, P. W. Fieguth, and A. S. Willsky. Computationally efficient steady-state multiscale estimation for 1-d diffusion processes. *Automatica*, 37(3):325–340, 2001.

[54] R Hoeben, PA Troch, Z Su, M Mancini, and K Chen. Sensitivity of radar backscattering to soil surface parameters: a comparison between theoretical

analysis and experimental evidence. *Proceedings, International Geoscience and Remote Sensing Symposium (IGARSS), Singapore*, page 13681370, 1997.

[55] P. R. Houser, W. J. Shuttleworth, J. S. Famiglietti, H. V. Gupta, and D. C. Syed, K. H. andGoodrich. Integration of soil moisture remote sensing and hydrologicmodeling using data assimilation. *Water Resoures Research*, 34(12):3405–3420, 1998.

[56] P. L. Houtekamer and H. L. Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001.

[57] W. W. Irving. *Multiscale stochatic realization and model identification with applications to large-sclae estimation problems.* MIT Ph.D. Thesis, 1995.

[58] W. W. Irving, P. W. Fieguth, and A. S. Willsky. An overlapping tree approach to multiscale stochastic modeling and estimation. *Ieee Transactions on Image Processing*, 6(11):1517–1529, 1997.

[59] T. J. Jackson and T. J. Schmugge. Vegetation effects on the microwave emission of soils. *Remote Sensing of Environment*, 36:203–212, 1991.

[60] A.H. Jazwinsky. *Stochastic processes and filtering theory.* Academic, San Diego, CA, 1970.

[61] M. Jordan. Graphical models. *Statistical Science*, 19(1):140155, 2004.

[62] S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the Ieee*, 92(3):401–422, 2004.

[63] C. P. Kim and D. Entekhabi. Examination of two methods for estimating regional evaporation using a coupled mixed layer and land surface model. *Water Resources Research*, 33(9):2109–2116, 1997.

[64] P. Kumar and A. L. Kaleita. Assimilation of near-surface temperature using extended kalman filter. *Advances in Water Resources*, 26(1):79–93, 2003.

[65] J. K. Li and S. Islam. Estimation of root zone soil moisture and surface fluxes partitioning using near surface soil moisture measurements. *Journal of Hydrology*, 259(1-4):1–14, 2002.

[66] S. P. P. Mahanama and R. D. Koster. Intercomparison of soil moisture memory in two land surface models. *Journal of Hydrometeorology*, 4(6):1134–1146, 2003.

[67] S. A. Margulis and D. Entekhabi. Temporal disaggregation of satellite-derived monthly precipitation estimates and the resulting propagation of error in partitioning of water at the land surface. *Hydrology and Earth System Sciences*, 5:27–38, 2001.

[68] S. A. Margulis and D. Entekhabi. Boundary-layer entrainment estimation through assimilation of radiosonde and micrometeorological data into a mixed-layer model. *Boundary-Layer Meteorology*, 110(3):405–433, 2004.

[69] S. A. Margulis, D. McLaughlin, D. Entekhabi, and S. Dunne. Land data assimilation and estimation of soil moisture using measurements from the southern great plains 1997 field experiment. *Water Resources Research*, 38(12), 2002.

[70] S. A. Margulis, D. McLaughlin, D. Entekhabi, and S. Dunne. Land data assimilation and estimation of soil moisture using measurements from the southern great plains 1997 field experiment. *Water Resources Research*, 38(12):–, 2002.

[71] D. Mclaughlin. Recent developments in hydrologic data assimilation. *Reviews of Geophysics*, 33:977–984, 1995.

[72] D. McLaughlin. An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering. *Advances in Water Resources*, 25(8-12):1275–1286, 2002.

[73] D. McLaughlin, Y. Zhou, D. Entekhabi, and V. Chatdarong. Computational issues for large-scale land surface data assimilation problems. *Journal of Hydrometeorology*, In Press.

[74] K. E. Mitchell, D. Lohmann, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, B. A. Cosgrove, J. Sheffield, Q. Y. Duan, L. F. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, D. P. Lettenmaier, C. H. Marshall, J. K. Entin, M. Pan, W. Shi, V. Koren, J. Meng, B. H. Ramsay, and A. A. Bailey. The multi-institution north american land data assimilation system (nldas): Utilizing multiple gcip products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research-Atmospheres*, 109(D7):–, 2004.

[75] F. I. Morton. Potential evaporation and river basin evaporation. *J. Hydraul. Div. Am. Soc. Civ. Eng.*, 91:67–79, 1965.

[76] E. G. Njoku and Entekhabim D. Passive microwave remote sensing of soil moisture. *J. Hydrology*, 184:101129, 1996.

[77] E. G. Njoku, W. J. Wilson, S. H. Yueh, S. J. Dinardo, F. K. Li, T. J. Jackson, V. Lakshmi, and J. Bolten. Observations of soil moisture using a passive and active low- frequency microwave airborne sensor during sgp99. *Ieee Transactions on Geoscience and Remote Sensing*, 40:2659–2673, 2002.

[78] Y. Oh, K. Sarabandi, and F. T. Ulaby. An empirical-model and an inversion technique for radar scattering from bare soil surfaces. *Ieee Transactions on Geoscience and Remote Sensing*, 30(2):370–381, 1992.

[79] L. M. Parada and X. Liang. Optimal multiscale kalman filter for assimilation of near-surface soil moisture into land surface models. *Journal of Geophysical Research-Atmospheres*, 109(D24):–, 2004.

[80] S. Popinet. Gerris: a tree-based adaptive solver for the incompressible euler equations in complex geometries. *J. Comput. Phys.*, 190(2):572–600, 2003.

[81] C. H. B. Priestley and R. J. Taylor. On the assesment of surface heat flux and evaporation using large-scale parameters. *Mon. Weather Rev.*, 100:81–92, 1972.

[82] R. H. Reichle, R. D. Koster, J. R. Dong, and A. A. Berg. Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. *Journal of Hydrometeorology*, 5(3):430–442, 2004.

[83] R. H. Reichle, D. B. McLaughlin, and D. Entekhabi. Variational data assimilation of microwave radiobrightness observations for land surface hydrology applications. *Ieee Transactions on Geoscience and Remote Sensing*, 39(8):1708–1718, 2001.

[84] R. H. Reichle, J. P. Walker, R. D. Koster, and P. R. Houser. Extended versus ensemble kalman filtering for land data assimilation. *Journal of Hydrometeorology*, 3(6):728–740, 2002.

[85] L. P. Riishjgaard. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus*, 50A:4257, 1998.

[86] M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J. K. Entin, J. P. Walker, D. Lohmann, and D. Toll. The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381–+, 2004.

[87] I. Rodriguez-Iturbe and P.S. Eagleson. Mathematical models of rainstorm events in space and time. *Water Resources Research*, 23:181–190, 1987.

[88] I. Rodriguez-lturbe, D. Entekhabi, and .L. Bras R. Nonlinear dynamics of soil moisture at climate scales, 1. stochastic analysis. *Water Resources Research*, 27:1899–1906, 1991.

[89] Fred C. Schweppe. *Uncertain Dynamic Systems.* Prentice-Hall, 1973.

[90] J. C. Shi, J. Wang, A. Y. Hsu, P. E. ONeill, and E. T. Engman. Estimation of bare surface soil moisture and surface roughness parameter using l-band sar image data. *Ieee Transactions on Geoscience and Remote Sensing*, 35(5):1254–1266, 1997.

[91] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.

[92] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky. Embedded trees: Estimation of gaussian processes on graphs with cycles. *Ieee Transactions on Signal Processing*, 52(11):3136–3150, 2004.

[93] O. Talagrand. Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan*, 75(1B):191–209, 1997.

[94] A. Tangborn and S. Q. Zhang. Wavelet transform adapted to an approximate kalman filter system. *Appl. Numer. Math.*, 33:307316, 2000.

[95] M. K. Tippet, S. E. Cohn, R. Todling, and D. Marchesin. Lowdimensional representation of error covariance. *Appl. Numer. Math.*, 52A:533553, 2000.

[96] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131:1485–1490, 2003.

[97] FT Ulaby, RK Moore, and AK Fung. *Microwave Remote Sensing, Active and Passive; Volume III: from Theory to Applications.* Artech House, Norwood, MA, 1986.

[98] P. J. van Leeuwen. A variance-minimizing filter for large-scale applications. *Monthly Weather Review*, 131:2071–2084, 2003.

[99] M Verlaan and AW Heemink. Tidal flow forecasting using reduced rank square root filters. *STOCHASTIC HYDROLOGY AND HYDRAULICS*, 11(5):349–368, 1997.

[100] J. P. Walker, P. R. Houser, and G. R. Willgoose. Active microwave remote sensing for soil moisture measurement: a field evaluation using ers-2. *Hydrological Processes*, 18(11):1975–1997, 2004.

[101] A. S. Willsky. Multiresolution markov models for signal and image processing. *Proceedings of the Ieee*, 90(8):1396–1458, 2002.

[102] Y. Zhou, D. McLaughlin, D. Entekhabi, and V. Chatdarong. Assessing the performance of the ensemble kalman filter for land. *Monthely eather Review*, In Press.