Published in the Proceedings of the XIV International Conference on Computational Methods in Water Resources, June 2002, Delft, Netherlands, by Elsevier Science, New York, NY, USA.

Land data assimilation with the Ensemble Kalman filter: Assessing model error parameters using innovations

Rolf H. Reichle * and Randal D. Koster

NASA/GSFC, Code 974, Greenbelt, MD 20771, USA

Abstract

Successful soil moisture data assimilation requires information about the uncertainties in the land surface model and the observations. While the errors in the observations are relatively easy to quantify, it is very difficult to specify the errors in the model, including approximations in the dynamical equations, wrong model parameters, or inaccurate forcing inputs. In synthetic experiments, it is straightforward to derive optimal model error covariances by minimizing the actual estimation errors (estimates minus synthetic "true" fields) as a function of the model error parameters. When satellite observations are used, however, we cannot easily compute the actual estimation errors. By contrast, the innovations (observations minus model forecast) are always readily available. In a twin experiment we generate synthetic "true" land surface fields and an open loop model trajectory using different model parameters and forcing fields. We also generate synthetic observations from the true fields which are then assimilated into the open loop model many times using different model error covariances. The resulting innovations sequences are analyzed and their statistics are contrasted to the actual estimation errors. We show that for our nonlinear problem the innovations sequence is a valuable tool for assessing the actual estimation errors and thereby for identifying appropriate model error covariances.

Key words:
data assimilation, Ensemble Kalman filter, soil moisture

* Corresponding author. Also at the University of Maryland, Baltimore County, USA.
Email address: `reichle@janus.gsfc.nasa.gov` (Rolf H. Reichle).
URL: `http://janus.gsfc.nasa.gov/~reichle` (Rolf H. Reichle).

1 INTRODUCTION

The seasonal predictability of summer precipitation over land at mid-latitudes depends in part on our ability to accurately specify the soil moisture initial condition. This initial condition can be derived by assimilating satellite observations into a land model up to the start time of the prediction, where the goal is to extract the information that is contained in the model (including forcing inputs) as well as in the observations. The information gain from the assimilation can be maximized when the respective uncertainties of the model and the observations are known. The problem is that we are uncertain about the uncertainties. While the errors in the observations are relatively easy to quantify, it is very difficult to specify the errors in the model. Such errors can be caused by approximations in the dynamical equations, wrong model parameters, or inaccurate forcing inputs. We refer to these errors collectively as model errors.

In synthetic experiments, it is straightforward to derive optimal model error parameters simply by determining which error parameters minimize the actual estimation errors (estimates minus synthetic "true" fields). When satellite observations are used, however, we cannot easily derive the model error parameters, because we cannot easily compute the actual estimation errors. High-quality independent measurements that would allow us to approximate the actual errors are typically unavailable or not suitable. For example, ground-based point measurements of soil moisture are difficult to compare to estimates that are representative of catchment-scale soil moisture. In contrast, the innovations (observations minus model forecast) are readily available, and their properties indicate whether or not our choice of model error parameters is appropriate.

2 ENSEMBLE KALMAN FILTER AND INNOVATIONS

The standard Kalman filter is the optimal sequential data assimilation method for linear dynamics and measurement processes with Gaussian error statistics. The Ensemble Kalman filter (EnKF) is a Monte-Carlo approach to the filtering problem (Evensen, 1994) and can be used with nonlinear land surface models. The EnKF is based on the approximation of the conditional probability densities of interest by a finite number of randomly generated model trajectories. Applications of the EnKF to land data assimilation can be found in Reichle et al. (2002a) and Reichle et al. (2002b), along with the filter equations.

Figure 1 highlights the key features of the EnKF. The filter works sequentially from one measurement time to the next, applying in turn a forecast step and an update step. The EnKF propagates an ensemble of state vectors $\mathbf{x}^i$ in parallel, each state vector representing a particular realization of the possible model trajectories (e.g., with certain random errors in model parameters and/or a particular set of errors in forcing). The EnKF state estimate is given by the mean of the ensemble members. The state error covariance P_k provides information about the uncertainty of the model forecast and is needed to determine the relative weights of the model forecast and the observation at the update time. The EnKF does not explicitly integrate the state error covariance, but computes it in-

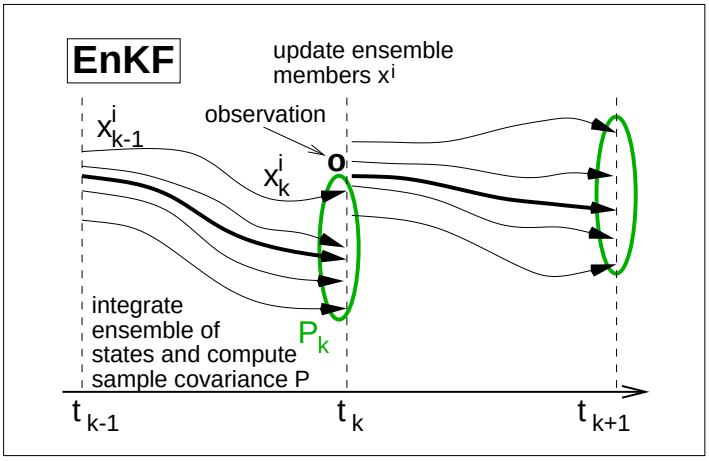

Fig. 1. Schematic of the Ensemble Kalman filter (EnKF).

stead diagnostically from the distribution of the model states across the ensemble. Each ensemble member is updated separately. The reduction of the uncertainty is reflected in the reduction of the ensemble spread.

Kalman filtering concepts were originally developed for linear problems. If (i) the model is linear, (ii) the model and measurement error distributions are Gaussian, and (iii) their mean values and covariances are correctly specified, then the filter will be operating optimally. Since state-of-the-art land surface models are nonlinear, we cannot expect fully optimal behavior. Fortunately, Reichle et al. (2002b) find that, from an assimilation perspective, the nonlinearities of a typical land surface model are not severe and are treated adequately by the EnKF.

The optimal operation of the filter is closely related to the statistical properties of the innovations, which are given by the difference between the observations and the corresponding model forecast. For linear problems with Gaussian errors and correctly specified model and measurement error covariances, the innovations are white (temporally uncorrelated) and have covariance $(\mathsf{H}_k \mathsf{P}_k \mathsf{H}_k^T + \mathsf{R}_k)$, where H is the measurement operator and R the measurement error covariance (Reichle et al., 2002b). The innovations covariance is internally calculated by the filter and can be used to normalize the innovations. The normalized innovations $\boldsymbol{\nu}$ must then follow a standard normal distribution $N(0,1)$.

3 EXPERIMENTS

In a twin experiment for a region of the south-eastern United States, we first generate synthetic "true" land surface fields by forcing the Catchment Model (Koster et al., 2000) with 1987 data from the International Satellite Land Surface Climatology Project (ISLSCP) (Sellers et al., 1996), starting from a spin-up initial condition on January 1. The Catchment Model has non-traditional prognostic variables, namely the surface excess, the root zone excess, and the catchment deficit, from which the soil moisture content in the surface

layer (2 cm depth), the root zone layer (1 m depth), and the total profile down to the water table can be diagnosed.

Next, we integrate the model again over the same time period but with an intentionally poor initial condition and different forcing data and model parameters (Table 1). Instead of the ISLSCP data we use the re-analysis data of the European Centre for Medium-Range Weather Forecasts (ECMWF) for the model forcing (Gibson et al., 1997). Moreover, we change the time scale parameters of the catchment model to those for a 5 cm surface layer and a vertical decay factor $\gamma = 2.17$ for the saturated hydraulic conductivity with depth (rather than for the 2 cm layer and $\gamma = 3.26$ that we use in the true integration). Collectively, these "wrong" inputs and parameters represent our imperfect knowledge of the true land processes. The resulting fields constitute our best guess prior to assimilating the remote sensing data and will be referred to as the "prior" or "open loop" solution.

The synthetic observations used in the assimilation are derived from the true fields by adding random measurement noise. We generate synthetic observations of the surface soil moisture with an error of 2 % (volumetric) once every three days for all catchments. These data are subsequently assimilated into the model using the "wrong" forcing and model parameters described above. The resulting fields are referred to as the "estimates".

The setup of the twin experiment implies that we do not know the exact statistics of the model errors. Besides perturbing the forcing inputs, we add synthetic model error fields to each ensemble member. These synthetic error fields are generated from a specified covariance matrix assuming a normal probability distribution. We assume that the standard deviation of each type of model error is identical for all catchments and that all model errors are uncorrelated. We also impose a correlation time of 3 days on the model error time series (autoregressive process of order one). An analysis of the innovations sequence will allow us to test the adequacy of the specified model error parameters.

4 RESULTS

With all inputs fixed except the magnitude of the model error variances for the surface excess, root zone excess, and catchment deficit, we calibrate these remaining parameters. Since the twin experiment is designed such that the true solution is known, a convenient measure of estimation performance is the actual error, which is the difference between the true soil moisture and its filter estimate. As an aggregate measure of filter performance we sum up the average actual errors in the surface excess, root zone excess and catchment deficit, where the average is taken in the root-mean-square sense over all catchments from Feb to Dec 1987. The first month is excluded to avoid initialization effects.

We have conducted a total of 235 assimilation integrations with different sets of model error variances (and with 10 ensemble members each). The model error standard devia-

Table 1
Inputs to the true, prior, and assimilation integrations (see text).

	True	Prior	EnKF
Initial condition	spin-up	perturbed moisture excess/deficit	
Forcing	ISLSCP	ECMWF	perturbed ECMWF
Parameters	2 cm surface layer, $\gamma = 3.26$	time scales for 5 cm surface layer, $\gamma = 2.17$	

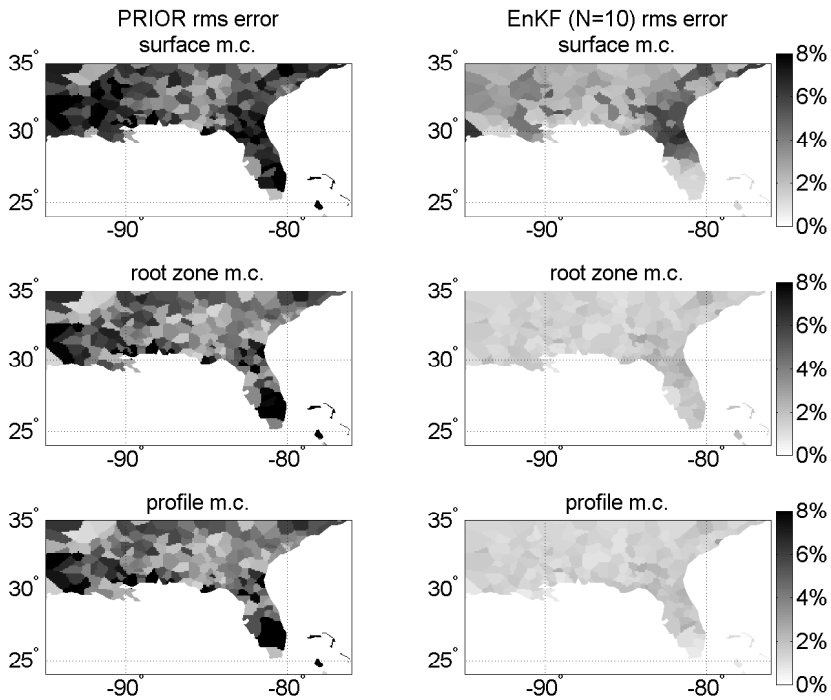

Fig. 2. Time-average error of the moisture content (m.c.) prior to the assimilation and for the EnKF with $N = 10$ ensemble members. Units are volumetric moisture percent.

tions that yield the smallest aggregate estimation error are $\sigma_{\text{surface excess}} = 0.026$ mm/d, $\sigma_{\text{root zone excess}} = 3.5$ mm/d, and $\sigma_{\text{catchment deficit}} = 0.036$ mm/d (corresponds to Experiment C below). For these parameters Figure 2 shows the time average actual errors of the moisture content variables. Obviously, the estimation errors are higher for the surface moisture content than for the root zone and profile moisture contents. This is because the surface moisture content varies on time scales of a day or less, while we assimilate observations only once every three days. Between observation times, errors in the model time scales and in the forcing (notably in precipitation) degrade the surface estimates significantly. The situation is different for the root zone and profile moisture contents. These lower layers exhibit greater memory and variations in their moisture content occur over longer time scales. Consequently, short-term errors in the forcing do not significantly impact the root zone and profile estimates.

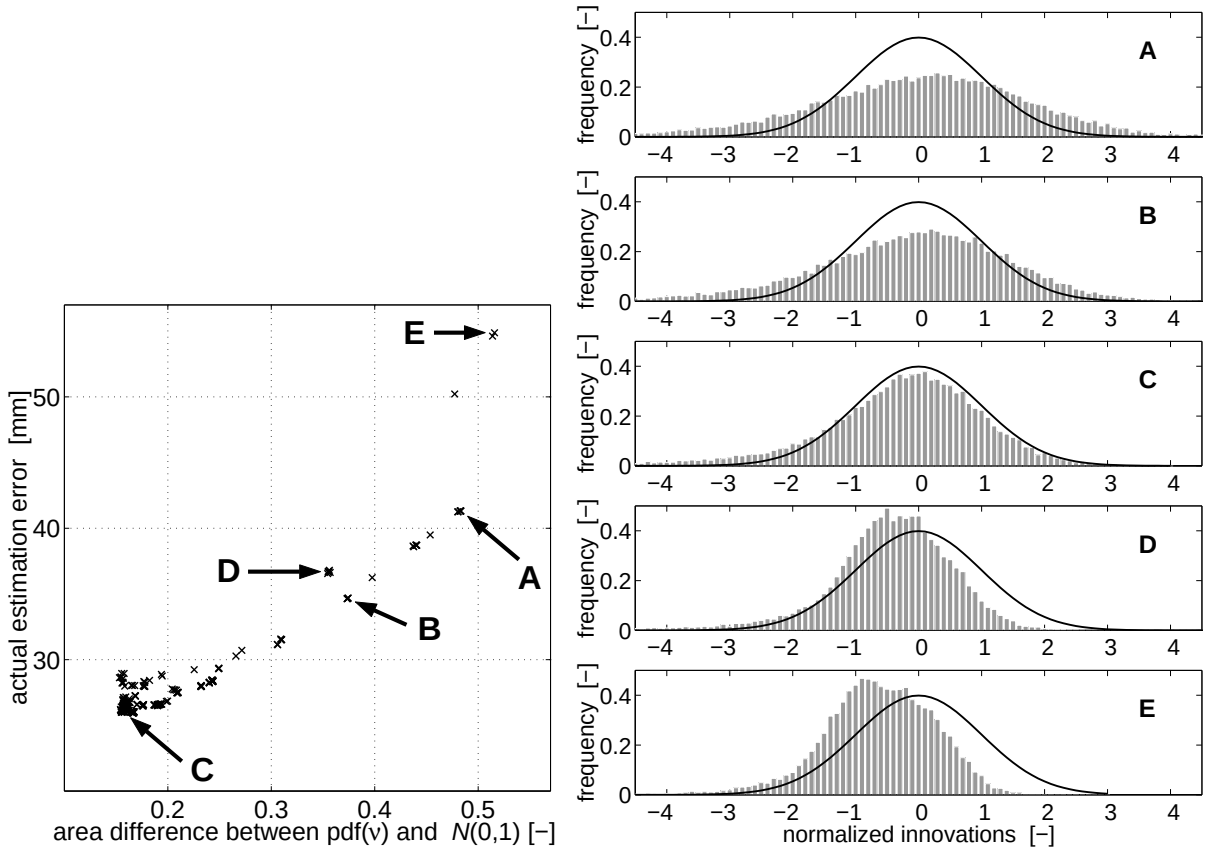

Fig. 3. Left: Aggregate actual estimation error versus deviation of the innovations histogram from the standard normal distribution $N(0,1)$. Right: Relative histograms of normalized innovations for five representative assimilation experiments. For comparison, $N(0,1)$ is shown as a solid line.

Next, we analyze the innovations sequences of all 235 assimilation integrations and contrast their statistics to the actual estimation errors. The key variable for our analysis is the relative histogram of the normalized innovations $\boldsymbol{\nu}$, which is an approximation of their probability density function, or pdf($\boldsymbol{\nu}$). We measure the deviation of pdf($\boldsymbol{\nu}$) from the standard normal distribution $N(0,1)$ by the total area where the two densities do not overlap. This aggregate measure, which must lie between 0 and 2, is plotted against the actual estimation error in the left-hand panel of Figure 3. There is clearly a strong correlation between the actual estimation errors and the statistics of the innovations. While the innovations of Experiment C (smallest actual error) do not strictly match $N(0,1)$ most closely, a larger deviation of the innovations histogram from $N(0,1)$ typically indicates larger actual estimation errors.

A closer look at five representative experiments reveals more details of the intimate connection between the innovations and the actual estimation errors. The model error parameters of these five experiments are summarized in Table 2, along with the actual estimation errors and the statistics of the normalized innovations. The right-hand panel of Figure 3 shows the corresponding innovations histograms. Recall that Experiment C produced the

minimum actual estimation errors. The model error variances in Experiments A and B are smaller than for Experiment C, which results in unrealistically small forecast state error variances. Consequently, the innovations covariances that are computed by the filter and used to normalize the innovations are smaller. Moreover, the small forecast error variances imply that the model forecast receives more weight in the update, which tends to increase the distance between the next forecast and observation. Both effects lead to larger values of the normalized innovations and a generally broader histogram. Conversely, Experiments D and E overestimate the model error variances, resulting in a narrower histogram.

There is also a negative bias in the innovations, which implies that the model forecast of the surface soil moisture is consistently too wet. This can be traced back partially to the fact that the ECMWF forcing data have consistently larger rainfall, leading to wetter surface soil moisture in the open loop integration as compared to the true fields. In addition, the forcing perturbations for precipitation are asymmetric, because negative rain is not allowed. Furthermore, the different time scales that are used in the true and the assimilation integrations (Table 1) generate additional bias in the soil moisture forecast. Finally, the absolute value of the bias increases with increasing model error variances. When model errors are added to each ensemble member, care must be taken that each state variable stays within its physically meaningful range. The boundedness of soil moisture implies that symmetric (Gaussian) model errors result in asymmetric perturbations of the soil moisture content. This effect is more pronounced for the larger perturbations associated with larger model error variances.

While the bias grows with increasing model error variances, the variance of the innovations decreases. It is instructive to note the compromise that is struck by the model error variances of Experiment C, which produced the minimum estimation error. Although Experiment C neither has the smallest bias nor the variance closest to unity, its innovations histogram is among the closest matches to the standard normal distribution when measured in terms of area differences (Figure 3). This indicates that the area difference measure conveniently integrates the bias and variance criteria and provides a useful indicator of the actual estimation error.

5 CONCLUSIONS

Although the innovations sequence is a powerful tool, the output from the assimilation system can be assessed in various other ways. If available, independent observations can assist with the verification of the estimated fields. Another option for evaluating the soil moisture estimates is to use them as the initial condition in a (retrospective) seasonal prediction. Provided the prediction is sensitive to the soil moisture initial condition, its skill can provide information on the value of the assimilation product. Finally, the soil moisture estimates must be checked for physical realism. If the model error variance of the catchment deficit is too large, for example, the update will put too much trust in the (uncertain) surface observations, and the estimates of the catchment deficit will exhibit unphysically large jumps whenever an observation is assimilated.

Table 2
Model error standard deviations σ, actual estimation errors, and statistics of the normalized innovations $\boldsymbol{\nu}$ (pdf = probability density function, $N(0,1)$ = standard normal distribution).

Experiment label		A	B	C	D	E
$\sigma_{\text{surface excess}}$	[mm/d]	0.0	0.006	0.026	0.052	0.207
$\sigma_{\text{root zone excess}}$	[mm/d]	0.0	0.7	3.5	13.8	27.6
$\sigma_{\text{catchment deficit}}$	[mm/d]	0.0	0.0	0.036	0.072	0.144
Aggregate actual estimation error	[mm]	41.3	34.7	25.9	36.5	54.9
Mean value of $\boldsymbol{\nu}$	[-]	-0.02	-0.13	-0.28	-0.46	-0.65
Variance of $\boldsymbol{\nu}$	[-]	3.03	2.49	1.72	1.04	0.84
Area difference of pdf($\boldsymbol{\nu}$) and $N(0,1)$	[-]	0.48	0.37	0.17	0.35	0.52

In a synthetic experiment we have shown that for the nonlinear soil moisture problem the analysis of the innovations sequence is a valuable tool. When satellite observations are assimilated, we can use the innovations to identify appropriate model error variances, provided that (i) the relationship between the estimation accuracy and the innovations in Figure 3 holds and (ii) the errors are close to Gaussian. Alternatively, the information from the innovations sequence can be used to improve the model and the forcing inputs.

References

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.* **99**(C5): 10143–10162.

Gibson, J., Kallberg, P., Uppala, S., Hernandez, A., Nomura, A. and Serrano, E. (1997). ERA description, *ECMWF Re-Analysis Project Report Series 1*, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, United Kingdom.

Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M. and Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model, 1: Model structure, *J. Geophys. Res.* **105**(20): 24809.

Reichle, R., McLaughlin, D. and Entekhabi, D. (2002a). Hydrologic data assimilation with the Ensemble Kalman filter, *Mon. Weather Rev.* **130**(1): 103–114.

Reichle, R., Walker, J. P., Koster, R. D. and Houser, P. R. (2002b). Extended vs. Ensemble Kalman filtering for land data assimilation, *J. Hydrometeorol.*, submitted Sep. 2001.

Sellers, P. J., Meeson, B. W., Closs, J., Collatz, J., Corprew, F., Dazlich, D., Hall, F. G., Kerr, Y., Koster, R., Los, S., Mitchell, K., McManus, J., Myers, D., Sun, K.-J. and Try, P. (1996). The ISLSCP Initiative I global datasets: Surface boundary conditions and atmospheric forcings for land-atmosphere studies, *Bull. Amer. Meteorol. Soc.* **77**(9): 1987–2005.