On the Representation of Spatial Uncertainties with Stochastic Simulation in Land Data Assimilation

Xujun Han* and Xin Li

Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, Gansu, 730000, China

Abstract. We use the random simulation and geostatistical sequential simulation to represent the model uncertainties by generating the uncertain input ensembles in land data assimilation, as the proper representation of the spatial uncertainties is very important to ensembles-based filter land assimilation methods and the efficiency of the filter depends on the proper representation of the model noise statistics. The method outlined in this paper explicitly acknowledges three sources of uncertainty and takes the spatial structure of the variables into consideration. To restrict the simulation interval of the uncertain inputs, the geostatistical interpolation technique, the geostatistical extrapolation technique and the truncated normal random number generator were applied. Simulation results from these uncertain inputs indicated that this method was sufficient to guarantee the separation of the soil moisture ensembles and easy to introduce the non additive noise. We proved the applicability of the stochastic simulation in representing the model spatial uncertainties in the land data assimilation.

Keywords: random simulation, geostatistics, sequential Gaussian simulation, data assimilation.

1. Introduction

Generally two approaches can be used to study the water and energy cycles: modeling and observation. Both of the methods have their advantages and disadvantages. The model simulation and the observation are both uncertain and unbiased to each other. It is important to fuse these two uncertain sources to get a better result. Data assimilation is one data fusion method that estimates the states of a dynamic system as a sequence of noisy observations become available. This dynamic system can be a land surface model or a distributed hydrologic model. The objective of land data assimilation is to merge multi-source observations into the dynamics of land surface model for improving the estimation of land surface states (Li et al., 2007). Bayesian filtering provides a rigorous general framework for the sequential data assimilation, such as ensemble Kalman filter (EnKF) and particle filter (PF) (Evensen, 2003). The optimality of the Bayesian filtering depends on the proper representation of the system and the observation noise statistics.

The main sources of the model uncertainty in the hydrological model include the inaccurate soil and vegetation properties, the time-dependant forcing data, the uncertainties in the model physics and the uncertain initial conditions. In the ensemble-based filters (EnKF or PF), the model states are represented by the uncertain ensembles, so it is easy to introduce these input uncertainties to the filters. Many methods for representing the model uncertainty based on some stochastic process in generating the input ensembles have been proposed in recent years. Wit and Bruin (2006), Almeida and Lopes (2006) have applied the geostatistical sequential simulation to simulate the uncertainty in precipitation fields. McLaughlin et al. (2006) and Turner et al. (2007) utilized the random simulation methods to construct the ensemble replicates in EnKF. These methods possess their own feature in describing the uncertainties. The geostatistical sequential simulation can keep the spatial structure of the simulated variables, but it is complex and...
computational demanding. On the other hand the random simulation is easy to use, but it can only be applied in a point scale.

We propose a new way to generate the uncertain input ensembles in the land data assimilation by taking advantage of the random simulation and the geostatistical sequential simulation. The spatial structures of the precipitation data and the initial soil moisture data have been taken into consideration and the independence of the ensemble members is guaranteed. Simulation results of the soil moisture and soil temperature from these uncertain inputs are presented. The outline of the paper is as follows. In section 2, we describe the random simulation theory and the geostatistical sequential simulation. Section 3 illustrates the experiment setup and the detailed ensemble generation method. In section 4, we showed the stochastic simulation results and summarize the paper.

2. Random Simulation and Geostatistics

The random simulation is often employed to draw random numbers from a predefined probability distribution; these numbers reproduce the underlying distribution. The distribution is specified according to the type of the variables such as a normal distribution or a uniform distribution.

The geostatistics is a collection of statistical methods to describe spatial correlations based on the sample data. It aims to use a series of correlation function models to provide a quantitative description of the spatial-temporal distribution of the natural variables. The ordinary kriging (Ok) interpolation and the sequential Gaussian simulation (Sgsim) are two popular algorithms in the spatial data representation. The Ok method allows for the local variability of the means by restricting the stationary of the variables to a local neighborhood centered on the location being estimated. But it tends to smooth out local details of the spatial variability of variables. The Sgsim method can model the spatial uncertainty by generating multiple realizations of the joint distribution of attribute values in space. These realizations usually serve as input to complex transfer function to assess the spatial uncertainty (Goovaerts, 1997).

2.1. Sequential Gaussian Simulation

Because the Sgsim assumes the shape of all conditional distributions is Gaussian and many variables in the earth sciences show an asymmetric distribution. So the Sgsim approach starts with an identification of the standard normal distribution and transforms the original conditional data into values with a standard normal histogram. Such a transform is referred to as a normal score transform, by which the Sgsim can be applied into the simulation of the non-Gaussian variables. This Sgsim is very fast and straightforward because the modeling of the Gaussian cumulative distribution function at each location requires the solution of only a single system at that location.

The steps in Sgsim are as follows:

1) Transform the original conditional data to the normal space
2) Establish the grid network and the coordinate system
3) Decide whether to assign data to the nearest grid node or keep separate
4) Determine a random path through all of the grid nodes
   a. Search for nearby data and previously simulated grid nodes
   b. Construct the conditional distribution by kriging
   c. Draw simulated value from conditional distribution
5) Back transform and check results

2.2. Interpolation and Extrapolation Methods

In the Sgsim algorithm, the cumulative distribution function of the unknown location should be modeled before simulation. Since the data are sometimes too sparse to allow a reliable inference of the sample cdf, so we should increase the resolution of the cdf. The approach consists of interpolating cdf values within each class of thresholds interval and extrapolating these cdf values beyond the smallest sample value (lower tail) and the largest sample value (upper tail). The power cdf model and the hyperbolic cdf model can be used, depending on the part of the cdf that is modeled: the lower tail, middle interval, or the upper tail. Because the power model for an upper tail calls for the sometimes arbitrary choice of a maximum sample value. The hyperbolic model allows one to extrapolate the upper tail of a positively skewed distribution toward an infinite upper bound (Goovaerts, 1997).
3. Experiment Setup

3.1. Study Area and Hydrological Model

Our simulation experiment is conducted over the Little Washita watershed, a critical research field in the SGP97 experiment (Jackson et al, 1999), which is located in south central Oklahoma and covers an area of 611 km². The SGP97 is a collaborative soil moisture mapping experiment in the central plains of Oklahoma from 18 June to 18 July, 1997. There are 42 meteorological stations distributed across the watershed at 5 km spacing; these provide the precipitation data used in the simulation. The advantage of the EnKF data assimilation technique is that it allows the flexible representation of the input uncertainties. The main sources of the model uncertainty in the hydrological model include the inaccurate soil and vegetation parameters, the time-dependant forcing data, the uncertainties in the model physics and the uncertain initial conditions. The random perturbations of the each uncertain input are provided to the distributed hydrological model GEOtop to generate random ensemble replicates of the model states (Rigon et al., 2006). GEOtop is a distributed hydrological model with coupled water and energy balance, which simulates the complete hydrological balance in a catchment by combining the main features of the land surface models and the distributed rainfall-runoff models.

3.2. Stochastic Simulation Methods

We used the random simulation and geostatistical sequential simulation to represent the model uncertainties by generating the uncertain input ensembles. The uncertain inputs of the initial soil moisture, the initial soil temperature, the soil sand and clay fractions, the vegetation LAI and the meteorological data (precipitation, air temperature, humidity, solar radiation and wind speed) were generated. Depending on the variables the generation methods were different. The multiple realizations of the initial soil moisture and the precipitation were generated by using the sequential Gaussian simulation method. To control the simulated interval of the soil moisture and the precipitation, we used the power model to model the lower tail and the upper tail of the initial soil moisture and used the power model to model the lower tail of the precipitation distribution and use the hyperbolic model to model its upper tail, respectively. These two models can make sure the positiveness and the upper limit of the simulated value. Because distributions of the precipitation and the soil moisture are not Gaussian, so the normal score transforms were applied during the Sgsim.

The spatially uncorrelated additive Gaussian noise was added to the initial soil temperature and the meteorological data. The spatially uncorrelated multiplicative uniform noise was added to the LAI, the soil sand and clay fractions. The random perturbations were generated within a physically reasonable range by the random number generator. The truncated normal distribution was applied to generate the Gaussian noise and it can assure the simulated random number of falling into a specified interval.

4. Results

Firstly we made use of the Ok method to interpolate the precipitation across the whole watershed. Secondly the Sgsim method was applied to simulate the random realizations of the precipitation map. All the parameters used in Ok and Sgsim are the same and derived from the analysis of the spatial pattern of the precipitation observation. Figure 1 shows the precipitation map at 8 p.m., July 10, 1997. The first one is the Ok interpolation, and the rest are the Sgsim results. From the figures it is obviously that the Ok result is more smoothed than the Sgsim results. Although the three Sgsim realizations show small variations, the spatial patterns of these four results are very similar. This proved that the Sgsim method can not only simulate the uncertainties, but also keep the spatial distribution characteristics of the precipitation observation.

The Quantile-Quantile (Q-Q) plot is a method for diagnosing the differences between the probability distribution of a statistical population from which a random sample has been taken and a comparison distribution. In order to validate the simulation results we showed the Q-Q plot results in Figure 2. The precipitation results were selected from four different rain time. The X-axis is the kriging quantile and the Y-axis is the Sgsim quantile. The Q-Q plots of Figure 2 show that the distributions are nearly identical, but the occurrences of the high value of the precipitation in Sgsim are greater than that of the Ok. This explains that the sequential simulation and the kriging interpolation come from a population with the same distribution approximately.
After the stochastic simulation, the uncertain inputs such as the meteorological data and the initial soil data were input into the hydrological model GEOtop and the surface soil moisture in the watershed were
modeled. Figure 3 shows the surface soil moisture map at 11 a.m., July 12, 1997. The first one is based on precipitation from Ok interpolation, and the rest are Sgsim results. From Figure 3, we can see the response of the soil moisture value to the uncertain inputs. As the precipitation map in Figure 1, the surface soil moisture map shows some uncertainties and keeps the spatial distribution structure meanwhile.

For the purpose of checking the uncertainties contained in the Figure 3, we extracted the surface soil moisture value from one simulated grid of the whole watershed. Figure 4 shows the time series result of the surface soil moisture in that specific grid. The Ok result and five Sgsim results during one month period are illustrated. One independent ensemble curve corresponds with one uncertain input. The separation of the ensembles reflects the uncertainties contained in the model input data.

Fig. 3: the surface soil moisture map

Fig. 4: the soil moisture ensembles
5. Summary

The land data assimilation is a new emerging research field in the earth science. During the assimilation we should specify the uncertainties or errors contained in the model and the observation, and then the land data assimilation system determines the corresponding weight of each part to get a better estimation of the land surface state. The ensemble-based Bayesian filtering algorithms such as ensemble Kalman filter and the particle filter become to the main assimilation methods gradually. The advantage of these data assimilation technique is that it allows the flexible representation of the input uncertainties and the way to construct the ensembles plays an important role in the filters.

In this paper we proposed a new way to construct the uncertain inputs for the hydrological model to generate the ensemble replicates for the ensemble Kalman filter in a land assimilation system. The sequential Gaussian simulation method was used to simulate the uncertain precipitation and initial soil moisture data, and the geostatistical interpolation and extrapolation techniques were employed to control the interval of the simulated value. The random number generators are used to simulate the noise contained in other input data, and the truncated normal random number generator was applied to fix the simulation interval of the random noise.

Afterward we used these simulated data as the uncertain inputs in the GEOtop to get the ensemble forecast of the land surface variables. The simulation results of the surface soil moisture indicate that this new perturbation method is sufficient to guarantee the separation of the ensembles. The precipitation map proved the advantages of the sequential Gaussian simulation in keeping the spatial pattern.

The uncertainties contained in the state ensembles such as the surface soil moisture ensembles could represent the model error in a certain degree, if we neglect the parameterization error of the model physics. So it is an alternative convenient way to introduce the model error information to the land data assimilation system, and easy to introduce the non additive noise. Then we examined the ensemble results in a land data assimilation system. The assimilation results showed that this new proposed ensemble generation method was sufficient to simulate the model spatial uncertainties and improve the performance of the land data assimilation system. The detail information of this assimilation experiment can be found in Han and Li, 2008.

In a word we have proved the applicability of the stochastic simulation in representing the model spatial uncertainties in the land data assimilation. It is an efficient way to construct the ensemble members for the ensemble Kalman filter.

6. Acknowledgements

This work is supported by the CAS (Chinese Academy of Sciences) Action Plan for West Development Project "Watershed Airborne Telemetry Experimental Research (WATER)" (grant: KZCX2-XB2-09) and the NSFC (National Science Foundation of China) project "Bayesian Filters in Land Data Assimilation" (grant: 40771036) and CAS International Partnership Project "The Basic Research for Water Issues of Inland River Basin in Arid Region" (grant: CXTD-Z2005-2).

7. References

