Uncertainty Analysis in the Decadal Survey Era: A Hydrologic Application using the Land Information System (LIS)

Ken Harrison, Sujay Kumar, Christa Peters-Lidard, and Joseph Santanello
NASA Goddard Space Flight Center, Greenbelt, Maryland 20771

***Abstract*—Computing and algorithmic advancements are making possible a more complete accounting of errors and uncertainties in earth science modeling. Knowledge of uncertainty can be critical in many application areas and can help to guide scientific research efforts. Here, we describe a plan and progress to date for a fuller accounting of hydrologic modeling uncertainties that addresses the challenges posed by decadal survey missions. These challenges include the need to account for a wide range of error sources (e.g., model error, stochastically varying inputs, observational error, downscaling) and uncertainties (model parameters, error parameters, model selection). In addition, there is a need to incorporate into an assessment all available data, which for decadal survey missions includes the wealth of data from ground, air and satellite observing systems. Our core tool is NASA's Land Information System (LIS), a high-resolution, high-performance, land surface modeling and data assimilation system that supports a wide range of land surface research and applications. Support for parameter and uncertainty estimation was recently incorporated into the software architecture, and to date three optimization algorithms (Levenberg-Marquardt, Genetic Algorithm, and SCE-UA) and two Markov chain Monte Carlo algorithms for Bayesian analysis (random walk, Differential Evolution-Monte Carlo) have been added. Results and discussion center on a case study that was the focus of Santanello et al. (2007) who demonstrated the use of remotely sensed soil moisture for hydrologic parameter estimation in the Walnut Gulch Experimental Watershed. We contrast results from uncertainty estimation to those from parameter estimation alone. We demonstrate considerable but not complete uncertainty reduction. From this analysis, we identify remaining challenges to a more complete accounting of uncertainties.**

***Index Terms*—parameter estimation, Bayesian analysis, optimization, value of information**

I. Introduction

Recently, novel statistical algorithms have been developed that enable a more complete accounting of modeling uncertainties. The Bayesian algorithms combine the strengths of conventional state and parameter estimation methods for exploiting remote sensing observations. They hold promise for improving the mission data products by providing estimates of the uncertainties therein. For scientists, knowledge of uncertainties helps to highlight areas that may benefit from further research. Decision-makers routinely weigh risks of alternate decisions and will benefit from the estimates of uncertainty.

State estimation methods admit modeling and observational errors within a probabilistic framework. However, these methods typically assume that model time invariant parameters (e.g., soil hydraulic properties) are perfectly known. But this is far from the case. By assuming that they are perfectly known, it is further assumed that remote sensing observations do not speak to the values of these underlying parameters but only to the states.

Parameter estimation methods, in contrast, can be applied to use remote sensing observations to improve the values of the unknown time invariant parameters. However, with the focus on identification of the best fit, in most applications there is no reporting of the uncertainty in the estimation. Implicitly it is assumed that there are a sufficient number of remote sensing observations to "identify" the model. In addition, overly simplistic error structures are often used. The familiar "least squares" estimation implicitly assumes a fairly restrictive error model (normally distributed, independent, zero mean and constant variance residual error).

Bayesian methods are more flexible, accounting for uncertainties in time invariant model parameters and capturing stochastic sources of error. The main challenge is computation time. Fortunately, advances in Markov chain Monte Carlo (MCMC) algorithms and in computing power, are expanding the application of Bayesian methods to increasingly complex models. These advancements are enabling a more complete accounting of uncertainties.

The uncertainty resolution in such parameters and in the resulting model simulations serves as an important measure of the worth of remote sensing observations. The uncertainty resolution is fully attributed to the information content of the remote sensing observations.

Here we use a new uncertainty estimation subsystem of the NASA Land Information System to account for uncertainty in the modeling of soil moisture.

II. BACKGROUND

A. Remote sensing of soil moisture

Soil moisture is the focus of the NASA's SMAP mission. Soil moisture plays a well-known role in the energy and water budgets for land-atmosphere exchange. Accurate prediction of soil moisture requires a combination of land surface modeling and remote sensing. Land surface models provide comprehensive spatial, temporal, and vertical resolution of soil moisture but are subject to considerable uncertainty. Remote sensing provides spatially comprehensive measurements but these measurements are infrequent and representative only of the near surface (~1 cm).

B. Land surface models

Most land surface models (LSMs) solve for the transport of moisture within the soil using Richards' (1931) formulations. Expressions known as soil moisture characteristic curves relate soil moisture (Θ) with matric potential (ψ), and soil moisture with hydraulic conductivity (K). The characteristic curves depend on a set of soil hydraulic properties (SHPs). One such set of characteristic curves includes as SHPs the saturated matric potential (ψ_s; aka "bubbling" or "air entry"), the saturated hydraulic conductivity (K_s), the saturated soil moisture content (porosity; Θ_s), the residual soil moisture content (Θ_r), and the pore size distribution index (b).

The Noah LSM used in this study was originally developed from the land component of the Oregon State University 1-D planetary boundary layer model (OSU; Troen and Mahrt, 1984). Noah is currently employed as the land surface scheme in NCEP's operational version of the Weather Research and Forecasting Nonhydrostatic Mesoscale Model (WRF-NMM).

C. Uncertainty in soil parameterizations

The uncertainty in SHPs is generally ignored in LSMs. However, as SHPs vary spatially and are scale-dependent, the soil texture-SHP datasets on which estimates rely are often unrepresentative of soils outside of the datasets. Further, the determination of SHPs from soil texture has associated uncertainties. Lookup tables based on soil texture class do not recognize the wide within-class variability of SHPs. The median (across soil texture classes and SHPs) coefficient of variation for a commonly used lookup table (Cosby et al., 1984) was 60%.

D. Accounting for uncertainty: Bayesian analysis via Markov chain Monte Carlo

The variable θ is used to denote a vector of uncertain and unobservable parameters such as the SHPs. The initial uncertainty in θ is described with the assignment of prior probability, p(θ). When new data y is made available, p(θ) needs to be updated to posterior probability, p(θ | y). This updating is achieved with application of Bayes' rule of probability:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \quad (1)$$

A standard numerical evaluation of Eq. 1 is not computationally tractable if there are many components of θ, as the denominator requires high dimensional integration (Gilks et al., 1995). Markov chain Monte Carlo (MCMC) is typically required. MCMC involves forming a carefully constructed chain—in Bayesian analysis, a chain of θ. The Metropolis algorithm is the basis of many algorithms:

Table 1. Prior distribution for SHPs and error variance

Parameter			Distribution		
Name	Symbol	Units	Type	Parameters	Values
Inverse of pore size distribution index	b	-	lognormal	median	6.77
				GSD	1.85
Saturated matric potential	Ψs	m	lognormal	median	0.263
				GSD	2.05
Saturated hydraulic conductivity	K_s	m/s	lognormal	median	2.45E-06
				GSD	1.80
Porosity	Θ_s	m^3m^{-3}	normal	mean	0.465
				std dev	0.0540
Standard deviation of residual error	σ	m^3m^{-3}	lognormal	median	1.50
				GSD	2.00

For k = 1 to K
  Sample θ_k^* from $j(\theta_k^*|\theta_{k-1})$
  R= TargetDensity(θ_k^*)/TargetDensity(θ_{k-1})
  AcceptProb = min(R, 1)
  If Rand() <= AcceptProb Then
    Set $\theta_k = \theta_k^*$
  Else
    Set $\theta_k = \theta_{k-1}$
  End If
Next k

where K is the sample size, θ0 is the starting point, θk* is the candidate point for inclusion as the k'th iteration of the Markov chain (θk)as drawn from symmetric proposal distribution j(θk*|θk-1,). In the case of Bayesian analysis, TargetDensity() is the posterior density as given by Eq. 1. AcceptProb is the acceptance probability for the candidate point θk*. If accepted (Rand() < AcceptProb), then the point is included else the previous point is repeated in the chain. For an infinitely long chain, the frequency (density) with which a state θ is visited is its posterior probability (Gelman et al., 1996).

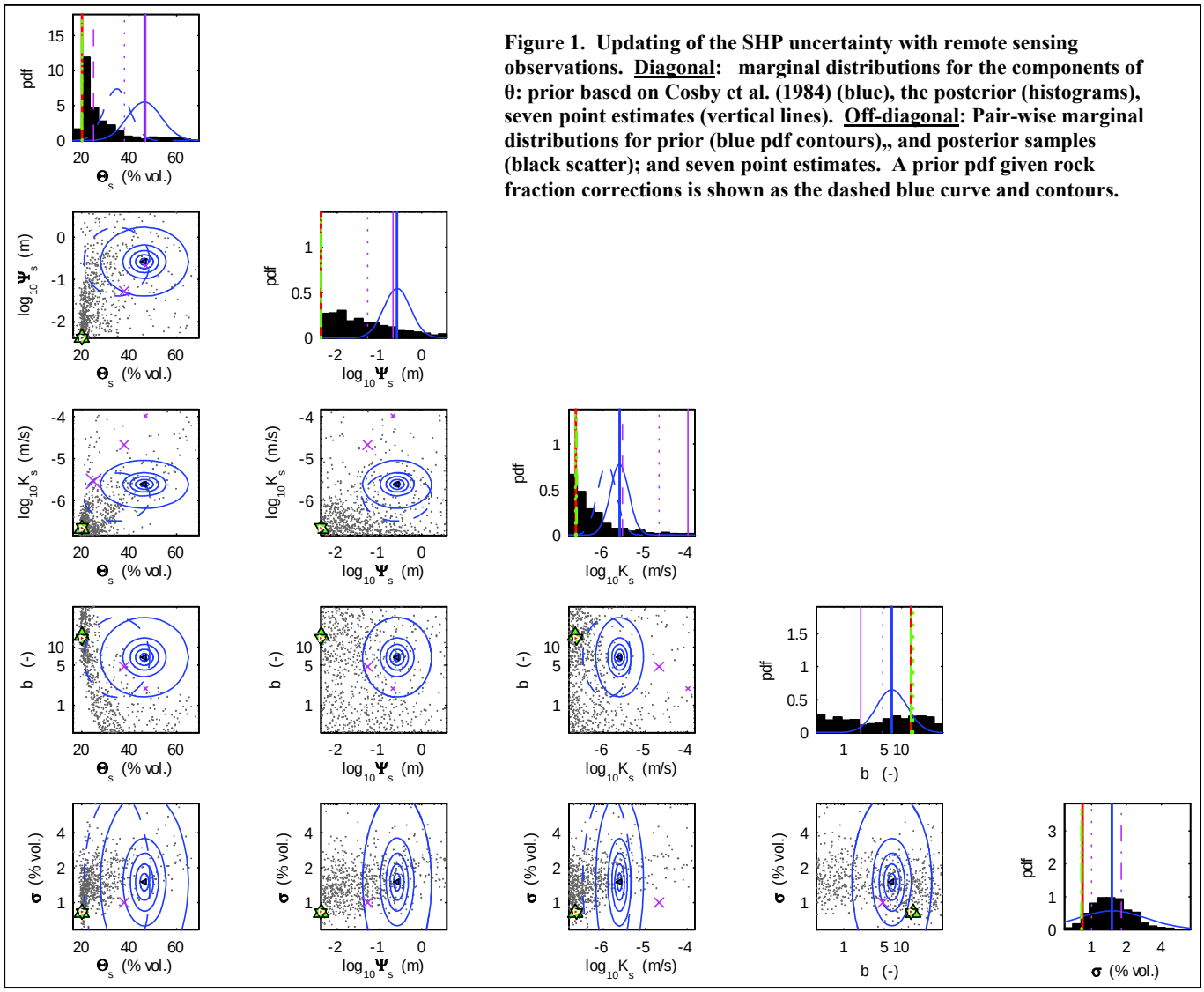


Figure 1. Updating of the SHP uncertainty with remote sensing observations. Diagonal: marginal distributions for the components of θ: prior based on Cosby et al. (1984) (blue), the posterior (histograms), seven point estimates (vertical lines). Off-diagonal: Pair-wise marginal distributions for prior (blue pdf contours),, and posterior samples (black scatter); and seven point estimates. A prior pdf given rock fraction corrections is shown as the dashed blue curve and contours.

III. Methods

A. *Land surface modeling*

Santanello et al. (2007) demonstrated for several sites in the Walnut Gulch Experimental Watershed (WGEW) the use remote sensing data in the estimation of SHPs. The SHPs were calibrated to PBMR observations of soil moisture collected during the Monsoon '90 experiment. The commonly applied sum of squares ("least squares") criterion was used to find the best fit to the PBMR observations. Here, we adopt the same case study time period, land surface modeling configuration, and remote sensing observations, but are interested in the uncertainty in the estimation of SHPs. We focus on the Kendall site (site 5 of WGEW), one of the WGEW supersites.

B. *Bayesian analysis*

For the likelihood model, $p(y|\theta)$, we use the normally distributed, independent, zero mean, constant variance residual error model that is implicit in least squares parameter estimation. This selection allows us to examine in a comparable manner the uncertainty around the least squares solution.

Each SHP's distribution within a soil texture class is assumed a reasonable prior for soils of that class. The prior (Table 1) is specified based on Cosby et al. (1984), with conversions to SI units. As Cosby did not report any, we assume zero covariance between SHPs, though there certainly are some correlations.

In addition to the SHPs previously mentioned, we include in θ the residual error standard deviation (σ). In Bayesian applications in hydrology, the standard deviation is often taken to be known and usually based on observational error (e.g., Vrugt et al., 2003).

C. *Uncertainty subsystem in the Land Information System (LIS) software*

As described in Kumar et al. (2006), LIS is designed using the principles of object-oriented frameworks, where all functional extensions (such as LSMs, DA algorithms, meteorological inputs, observational data, etc.) are implemented as abstract, extensible components. A large suite of modeling extensions has been incorporated in LIS using this design paradigm. The uncertainty and optimization subsystem in LIS defines three functional abstractions: (1)

variables (θ), (2) algorithm and (3) function of the variables (f(θ)). An algorithm iteratively adjusts θ based on feedback f(θ). A custom implementation of each of these three abstractions constitutes a specific instance of an optimization/uncertainty estimation problem.

D. *Differential Evolution Monte Carlo*

Two MCMC methods have been implemented to date within LIS: a random walk (RW) MCMC method and the Differential Evolution Monte Carlo (DE-MC) method. Both have been implemented in parallel fashion to take advantage of the efficiencies of the parallel ensemble run capability in LIS.

Both MCMC algorithms are special cases of the Metropolis algorithm. Differential Evolution Monte Carlo (DE-MC) is an example of adaptive MCMC in which the proposal distribution adapts to the scale and orientation of the posterior distribution as it is learned (ter Braak, 2006). In addition, DE-MC is an example of a population-based MCMC method in which a population $\theta = \{\theta_k^1, \theta_k^2, \ldots, \theta_k^N\}$ is advanced with each iteration (Jasra et al., 2007). The results shown here are for DE-MC.

IV. Results and Discussion

DE-MC was run for Site 5 (Kendall) of the Walnut Gulch Experimental Watershed for the Monsoon '90 experimental time period. The shift from the prior distribution $p(\theta)$ to the posterior distribution $p(\theta|y)$ is represented in Figure 1. The translation of the prior and posterior uncertainty to the soil moisture simulation is shown in Figure 2.

A. *Large uncertainty around the parameter estimation solutions*

In Figure 1, the triangles in the fairly dense regions of the pair-wise scatterplots represent the least squares solutions found through various algorithms. Clearly in this case, reliance on the least squares solutions ignores substantial remaining uncertainty.

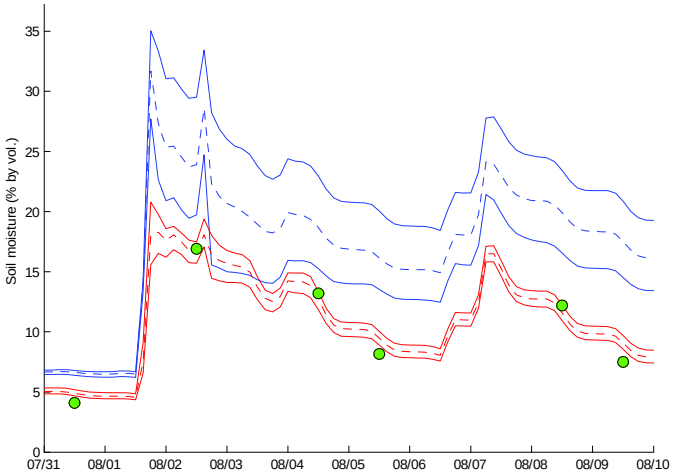
B. *Large uncertainty reduction from remote sensing*

In Figure 2, the upper and lower lines represent the 17th and 83rd percentiles of the distribution of soil moisture at each time step, i.e., an interval with two-thirds chance of containing the true soil moisture. There is an approximately five-fold uncertainty reduction for much of the soil moisture time series. This five-fold reduction is a direct measure of the benefits of the remote sensing observations.

C. *Limited influence of the prior*

Most striking is the shift away from the prior distribution. This shift can be seen in the parameter space and output (soil moisture) space. In the pairwise marginal plot for Θ_s-b (4th row, 1st column) the majority of the mass can be seen to curve around the prior. Were it possible to view the full five-dimensional space, the mass would be seen to lie nearly fully outside of the outside contour. In the soil moisture time series (Figure 1), the shift away from the prior is similarly dramatic. The soil moisture values shift markedly down from the prior

Figure 2. Remote sensing observations can reduce the uncertainty of land surface model simulations. The range of uncertainty in the soil moisture time series is shown prior to (blue) and after (red) consideration of remote sensing observations (green circles). Remote sensing enables a roughly five-fold reduction in uncertainty over the simulation period.



to the posterior, becoming much more in line with the PBMR observations (green circles).

The lack of the influence of the prior is attributed to the soils datasets not representing the soil under consideration. The soils of the region have an unusually high rock fraction. However, it is also likely that the systematic biases of the LSM are being absorbed by the SHPs. Given the zero-mean bias assumption of the error model, any systematic bias would be absorbed by the SHPs. Research into the development of more realistic error models would improve the accounting of uncertainty.

D. *Important to consider uncertainty in the error model itself*

In practice, the statistical parameters of the error model are assumed known even though they, too, are subject to considerable uncertainty. Least squares implicitly assumes σ to be known. Here, we have explicitly admitted uncertainty in σ by including it in θ. In reference to the marginal distribution for σ (row 5, col. 5 of Figure 1), it is clear that much uncertainty remains.

Acknowledgment

We gratefully acknowledge the financial support from the NASA Earth Science Technology Office (ESTO). Resources supporting this work were provided by the NASA High-End Computing (HEC) program through the NASA Center for Computational Sciences (NCCS) at Goddard Space Flight Center.

References

[1] Santanello, J.A., C.D. Peters-Lidard, M.E. Garcia, D.M. Mocko, M.A. Tischler, M.S. Moran, and D.P. Thoma. 2007. Using remotely-sensed estimates of soil moisture to infer soil texture and hydraulic properties across a semiarid watershed. *Remote Sens. Environ.* 110:79–97.

[2] Troen, I. B., & Mahrt, L. (1984). A simple model of the atmospheric boundary layer; sensitivity to surface evaporation. *Boundary - Layer Meteorology*, 37, 129−148.

[3] Cosby, B. J., Hornberger, G. M., Clapp, R. B., & Ginn, T. R. (1984). A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resources Research*, 20, pp. 682−690.

[4] Gilks, W.R., S. Richardson, and D. J. Spiegelhalter (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, Edited by W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, Chapman & Hall.

[5] Kumar, S., C. Peters-Lidard, T. Tian, P. Houser, J. Geiger, S. Olden, L. Lighty, J. Eastman, B. Doty, P. Dirmeyer, J. Adams, K. Mitchell, E. Wood, and J. Sheffield (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modeling and Software*, 21, 1402–1415.

[6] ter Braak, C.J.F. (2006) "A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces," *Stat Comput,* 16:239–249.

[7] Vrugt, J.A., W. Bouten, H.V. Gupta, and J.W. Hopmans (2003). Toward Improved Identifiability of Soil Hydraulic Parameters: On the Selection of a Suitable Parametric Model, Vadose Zone Journal, v.2, pp. 98–113.