Responses to referee #1' comments on

Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model

Y Sun, Z Hou, M Huang, F Tian, and LR Leung

We greatly appreciate the constructive comments and suggestions from the anonymous referee, which help us improve the quality of this paper. Our responses to the comments are provided item-by-item as follows.

Overview

1. The authors explore a method for the retrieval of hydrological parameters of a land surface scheme, using observations of runoff and latent heat flux observations. They use a Monte Carlo Markov chain method to retrieve parameters of CLM4, and explore several alternative scenarios within this framework. They also use Bayesian model averaging to combine the results from several of the scenarios.

   Response: *Thank you for the summary.*

2. I think that the topic area is of considerable interest, since land surface schemes are moving towards increasingly complex representations of subgrid hydrological processes, for which reliable parameter estimates are currently difficult to obtain. Inversion methods are a potentially useful approach to obtaining parameter estimates.

   Response: *Thank you very much for the encouraging comments.*

3. I was not aware of the previous work by the authors showing that LH flux was sensitive to some of the hydrological parameters in the CLM4 model, and I was intrigued because it seems a little surprising that parameters which are intended to control runoff generation are also having a significant impact on latent heat flux at the monthly time scale. It is plausible that changes in parameters which affect runoff generation can also lead to consequential changes in soil moisture, which in turn affect latent heat fluxes. However, the relatively weak physical basis for the hydrological parameters means that some of them have extremely wide prior distributions and very complex joint distributions with other parameters, so perhaps any reliable information on the water cycle might be helpful in narrowing those very wide priors.

   Response: *Agreed. This is exactly our motivation to design a UQ and inversion framework that could be applied to CLM4 by fusing observations/information from various aspects of the hydrological cycle, to better estimate model parameters, by constraining them with observations. Our previous papers in this series, Hou et al. (2012) and Huang et al. (2013), have demonstrated clearly that runoff generation parameters, especially those relevant to subsurface runoff generation, have extremely wide prior distributions and complex joint distribution with other parameters, partially due to the lack of information in the deep subsurface, and therefore need more attention. Since streamflow and ET are mutually constrained by responses of the land surface and subsurface systems, measurements of these fluxes could provide constraints for both surface and subsurface parameters through inversion. We agree that the choices of prior distributions has important impacts on the relative* *contributions of the corresponding input parameters to the total variability of output responses (e.g., LH, SH, runoff); for example, a parameter with weak local sensitivity can have a significant contribution in the output variability if its prior range is very large. Therefore, we put significant effort collecting information about various field site conditions and carefully decide the prior ranges of the input parameters.*

4. I would like to see some discussion on whether the SIMTOP concepts used in CLM4 are meaningful at these specific sites; what are the dominant pathways for runoff generation in Walnut River? TOPMODEL-style concepts are not very relevant in some physical settings, for example where deep groundwater flows comprise a substantial component of the water balance, or where parts of the catchment are disconnected from the river system for extended periods of time, or where infiltration excess is the dominant runoff generation mechanism.

   Response: *Agreed. TOPMODEL has its limitations, which has been well-documented and discussed in the hydrologic and land surface modeling literature (Beven, 1997; Duan and Miller, 1997; Iorgulescu and Musy, 1997; Woods et al., 1997; Huang et al., 2008; Li et al., 2011).*

   The site characteristics of Walnut River watershed are presented in detail in LeMone et al. (2000). Measuring 100km from north to south and 60km from east to west, the watershed includes headwaters of the Whitewater and Walnut rivers; it has minimal leakage into the substrata, and the shapes of the streambeds are reasonably stable. The watershed is located in a region with a strong east-west gradient of annual precipitation (760 mm/yr – 860 mm/yr), geology, and vegetation. Therefore, the runoff generation mechanism of the watershed could be highly heterogeneous, both spatially and temporally.

   On the other hand, we would like to point out that the implementation of TOPMODEL parameterizations in CLM4 is rather ad hoc and inconsistent with the original formulations. For example, f_{over} and f_{drai} should be the same parameter in the TOPMODEL framework, but were separated into two parameters when the parameterizations in Niu et al. (2005; 2007) were ported from CLM3.5 to CLM4 (i.e., Oleson et al., 2010). Therefore these parameters lost their real physical meanings consistent with the TOPMODEL framework. Discussions on this discrepancy are documented in Huang et al. (2013) and Li et al. (2011).

   Specifically, f_{drai} serves as an important parameter determining subsurface runoff generation by scaling the water table depth (i.e., equation A5 in Huang et al., 2013), but is now decoupled from the surface runoff generation formulations. Therefore, now f_{drai} is the reciprocal of the effective subsurface storage capacity scaled by the water *table depth for the subsurface runoff generation calculation, which is still consistent with the TOPMODEL concept in some sense but the assumptions embedded in TOPMODEL that "the distribution of down slope transmissivity with depth is an exponential function of storage deficit of depth to the water table" is greatly relaxed in CLM4.*

We have added discussions on site characteristics of the watershed and CLM4 runoff generation formulations in the revised manuscript.

5. I found the authors' discussion of the posterior distributions rather subjective, and I reached quite different conclusions to the authors. To me, it seemed that the posterior distributions were very sensitive to small changes in the reference acceptance probability, and the sensitivity seemed random at times. If the method is working correctly, I don't understand why the distributions of all the parameters do not change gradually as the p_ra value is varied from 1.0 to 0.95 to 0.90. My observation is that as p_ra is varied gradually, the posterior distributions jump around randomly sometimes, and this leads me to doubt the reliability of the results; I would like to be reassured that the method is in fact working correctly.

   Response: *Thanks for bringing up this point. This is exactly one of points we want to convey in this study. In practice, most of the inverse problems are ill-posed and also due to nonlinearity, many parameters, especially these insignificant parameters, cannot be uniquely determined. Those parameters can be reduced through sensitivity analysis or parameter screening. Theoretically, when there are infinite numbers of runs, the inverse results should be the same, no matter what inversion approach we are using, which in practice cannot be achieved due to computational limitations and the convergence criteria that are used. In our study, we found that some posterior distributions were sensitive to small changes in the reference acceptance probability. This issue can be reduced by reducing the parameter dimensionality with only the most identifiable parameter kept in the inversion parameterization. The most significant/identifiable parameters can be easily captured and their posterior distribution conform more to expectations, while the insignificant parameter may vary widely within the feasible range and appear random due to non-uniqueness issues which are common to inverse problems. Another method to deal with non-uniqueness is to replace the nonlinear numerical CLM simulator with simplified surrogates, but some information about the complexity in input-output relationships will be lost. In this study, we demonstrated that although the inversion results are subject to non-uniqueness (multi-modality), with reduced parameters, the calibration converges much faster and yield much more consistent results. Our study, together with our previous studies on parameter screening across flux tower sites and MOPEX basins, can provide useful guidance on similar practices of calibrating hydrological* *model parameters.*

Main points

6. 5079L10 "It is also important for an inverse approach to be capable of quantifying and evaluating the prediction uncertainty" Please briefly explain why this is important.

   Response: *Traditional approaches, such as least-square fitting, are designed to estimate a single optimal parameter set, which may not exist in practice, particularly for complicated systems, where unknown parameters and observable variables have nonlinear and non-unique relationships, or observations are limited, or the forward models (e.g., CLM) are not perfect. Therefore, the parameter estimates are subject to great uncertainty, which will introduce uncertainties in predictions made using the models. Hence uncertainty in model parameters, and hence model predictions, needs to be well quantified.*

7. 5080L1-11 This paragraph seems to address uncertainty in parameters, and in measurements of the model output variables, but it does not mention uncertainty in the forcing data, or in the model structure, both of which should be addressed.

   Response: *Agreed. Uncertainty of a model can stem from the model input data, data used to evaluate model outputs, model parameters, and model structure. Some model intercomparison experiments have been designed to specifically address uncertainty in forcing data and model structures by take advantages of multiple observational datasets or providing multiple forcing data, and developing common experimental protocols to isolate uncertainty due to model structural differences. Our study focuses primarily on uncertainty related to model parameters. In the revised manuscript, we acknowledge the other types of uncertainty to clarify the broader issues.*

8. 5082L1-6 What is the uncertainty in the forcing data which you derived from NLDAS? This is relevant because errors in forcing would affect the inversion process.

   Response: *It is true that forcing data quality affects the inversion accuracy. In typical inversion setup, we do not add further noises/errors on top of the existing data and assume the corresponding measurement and modeling errors are represented by the residuals. We used the NLDAS data because it is widely used in land surface modeling studies including model intercomparison experiments so our results can be compared to other studies in the broader context. In addition the NLDAS data was* *compiled from multiple data sources to take advantage of different spatial/temporal characteristics and has been quite extensively evaluated. We have also performed a comparison of the NLDAS and MOPEX forcing data (See the response to comment #30), which shows that NLDAS dataset agrees with the MOPEX dataset very well.*

9. 5083L8 "However, simulations of heat flux and runoff using the calibrated parameters show only small improvements compared to simulations using the default parameter values." This seems like a result, and belongs later in the paper. In any case, it deserves more discussion. Why do you think PEST was unable to find better parameter sets than the default? What PEST options/features did you use? Is PEST a less efficient optimiser than MCMC? What if you had used a different optimiser with the same objective function? Is the least squares objective function really very different to the log-likelihood function?

   Response: *We used PEST to do the inversion with the default parameters as initial values, and the results showed that simulations of heat flux and runoff using the calibrated parameters show small improvements. There is a set of tuning parameters in the PEST inverse setup, and we adopted the widely used default settings as given in the user guide. PEST, using the least-square fitting approach, is aimed to find single optimum set of parameter values, and works well for more linear systems with monotonic relationships between unknown parameters and observable variables and weak interactions between the unknowns. For a complicated system, it is very likely that a single optimum set of parameter values does not exist, and a probabilistic description of all possible solutions is more reasonable. MCMC-Bayesian allows all ranges of possible solutions to be explored and avoids convergence to local minima. There is a single objective function in PEST, but in MCMC-Bayesian, we assign proper weights (i.e., likelihood) to all possible solutions of parameter sets.*

10. 5083L8 The discussion on the PEST application is too brief to be useful to readers. I suggest you either expand it or remove it altogether (since PEST is not central to the paper).

    Response: *Please see the discussion in our response to comment #9. We have added the above discussion in the text.*

11. 5084L6 "assumptions that ε_{ij} are normally distributed with variance σ_{ij}, and the distributions are independent" Do you have any information to support these assumptions? Did you make any transformations of the outputs to ensure these assumptions were approximately satisfied? Would your study have reached different conclusions if your assumptions were incorrect?

Response: *Thanks for the comments. It is true that people can use different probability distributions to describe the errors, although normal distribution is the most widely adopted given the probability theory (e.g., central limit theorem), which is also what we used in the study. In practice, it is helpful to check the normality of the error distributions. Our previous study (Hou et al., 2012, Huang et al., 2013) show that by fully exploring the input parameter space, the resulting output responses (heat fluxes, runoff) without further transformation are rather symmetrically distributed without much tailing events, which indicate that normal assumptions is applicable. The nonzero mean is not an issue – as MCMC sampling enables us to explore local ranges, and parameter realizations in those ranges corresponding to large mean errors will be assigned very low probability/likelihood.*

12. 5084L6 What assumptions did you make about the variances for runoff and LH flux?

    Response: *In our MCMC-Bayesian inverse approach, we do not make any special assumptions for observational variables such as runoff and LH flux. The residuals include both modeling and measurement errors, and their variances will be updated as MCMC sampling explores the full ranges of input parameters. Of course, if the model is wrong, or forcing/flux data are of very bad quality, the error variances would be big.*

13. 5087L14 "the US-ARM site and one MOPEX basin (07147800), which are located in close proximity with similar climate and land surface conditions." Since the basin has an area of over 4800 sq km, is it meaningful to say that these two are close together? And is this 4800 sq km basin really that homogeneous in climate and land cover? You stated earlier that the US-ARM site is in croplands, but the basin is only 22% croplands.

    Response: *This is an important point. In this manuscript, we want to provide a comparative study between different responses. The LH flux can be measured by flux tower which represents a small typical area, while the runoff is a composite response of a drainage basin which is a large area. We acknowledged that a basin cannot be homogeneous in climate and land cover, but this is the closest pair that we could find out of the Ameriflux first priority and MOPEX sites. The basin has roughly a diameter of 100km, and with relatively small variations in topography, we assume that the basin and the Ameriflux have similar climate conditions. To be consistent with the MOPEX site characterization, we used 'corpland' to represent the dominant land cover. We have modified it and added a few clarifications.*

14. 5088L10 "Posterior distributions with different reference acceptance probabilities generally are consistent, except for fdrai, Qdm and Sy when the rejection rate is very

low with a reference acceptance probability pra of 0.5" In my view, the posterior distributions are NOT generally consistent across the various values of p_ra. To meet my criterion for consistency, I would expect the distributions to substantially overlap. For example, for f_over, the distributions for p_ra=1.0 and p_ra=0.95 hardly overlap at all. And for Ks, the higher P_ra values lead to posterior distributions lying mainly between -1 and 0, but the posterior distribution for p_ra=0.5 lies mainly between -2 and -1. I think the degree of consistency needs to be quantified if the authors wish to pursue this point.

Response: *We accept your point and agree that "consistency" is not general, particularly for those parameters that are less identifiable due to nonlinearity and non-uniqueness issues. Our previous study (Hou et al., 2012) showed that fover, fmax, Ks, and θs are insignificant at the US-ARM site. Therefore, the posterior distributions of these parameters perform relatively poor. This demonstrates the necessity of performing parameter screening such that the inverse problems can be less ill-posed.*

15. Figure 1: I was surprised that the posterior distributions did not change in a more systematic manner as p_ra varied, and so it is not clear to me whether the sampler has converged. What were the stopping criteria?

    Response: *Each p_ra is a different inverse setup and a completely different set of posterior samples are selected and assigned weights. We don't expect linear correspondence between the posteriors (e.g., means, modes) and p_ra. For high-dimensional parameter space, we expect that the weakly identifiable parameters to show some inconsistency even with a huge number of samples. We used the summary statistics of the posterior samples during the burn-in period to evaluate the convergence with a tolerance relative error, e.g., 0.001. We have clarified more in the text.*

16. 5089L6 "which might be due to errors in the observed heat fluxes, errors in the CLM forcing data, and/or under-representation of the complicated physical processes using the current parameterization schemes." It would greatly aid the reader if the authors could provide uncertainty estimates for the heat fluxes (could the true mean LH flux for January really be 10 W/m2 lower than the measured value?) and in the CLM forcing data. It would also be helpful for the authors to point out any features of the land surface processes at this site which they consider are not well parameterized in CLM4.

    Response: *This is an interesting and indeed an important question – what if the observational data used for calibration is wrong? Our answer is that the calibration would also be wrong; and there is no way to avoid the issue using any calibration* *approach. If we know clearly that the LH observations are lower than the 'true' values, we can do the adjustment accordingly. This is about reducing uncertainty associated with data or improving data accuracy, which however cannot be done through calibration that aims to reduce uncertainty associated with model parameters. Beyond the data quality issue, the reviewer also pointed out the parameterization issue, which for sure affects the modeling errors. Efforts to address parameter uncertainty have to assume perfect models with no systematic simulation errors and that the major parameters have been included in the unknowns based on our previous study. While this approach is limited and must be performed for each model with different structures and parameter sets, an important goal is that through inverse modeling and sensitivity analysis, more insights will be gained about the model limitations to provide guidance for improvements. Hence inverse modeling and uncertainty quantification is not a one-time process but rather should be considered part of a model development cycle to provide models with practical skills as well as understanding of behaviors of the physical systems and the models. We added the discussion that inversion can be affected if there are data quality issues and model structural errors.*

17. 5089L3 "However the estimates with reference acceptance probability of 0.5 noticeably deviate from other inversion estimates" Whether the differences between simulations are considered large or not must depend to some extent on the uncertainties in those predictions. How large are the uncertainties (due to parameter uncertainty) in the simulations of LH flux for each p_ra?

    Response: *The rejection standard is the major reason of the differences. Each different p_ra is a different inverse setup and a completely different set of posterior samples are selected and assigned weights. A p_ra of 0.5 has a largely relaxed rejection standard, and would yield wide uncertainty ranges in predictions. As shown in Figure R1, when the reference acceptance probability is 1.0, the uncertainty range is small compared to the uncertainty range when the reference acceptance probability is 0.5.*

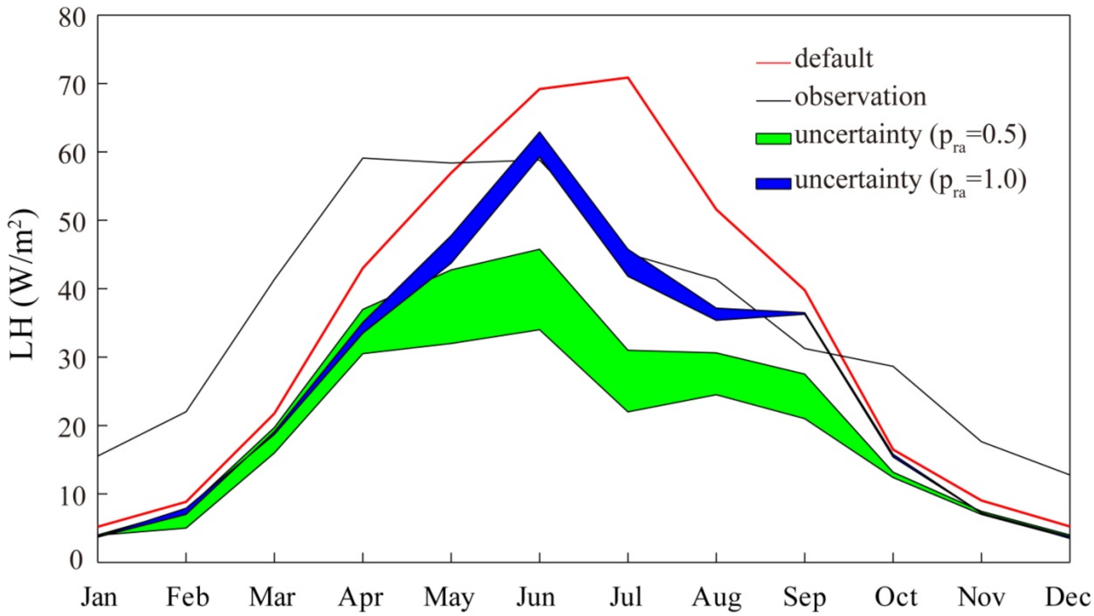


Figure R1 the uncertainty in the simulations of LH flux due to the uncertainty in the parameters at the US-ARM site.

18. 5089L12 "They show consistent patterns for different reference acceptance probabilities, except for the parameter b." Again, in my view, the posterior distributions are NOT generally consistent across the various values of p_ra. For f_over, the p_ra distribution does not overlap with the others, the distributions of C_s occupy most of the feasible space, and there seem to be two distinct Ks distributions. The authors must be using other criteria to decide on consistency; these criteria need to be made explicit in the paper.

    Response: *Again, we accept your point and agree that "consistency" is not general, which is as expected, particularly for those parameters with low identifiability. We have modified the text accordingly.*

19. Figure 6: To understand the relatively poorer model performance at US-ARM for LH flux, it would be helpful to have some basic information about the comparative climates at the two flux sites. Why does the measured LH flux at US-ARM have a seasonal peak in April (and similar values in May-June, while the US-MOz has a clear peak in June? Is this an effect of moisture limitation at US-ARM, or plant development/harvesting, or something else?

    Response: *This is a nice point. We studied the two sites with different field and climate conditions to test the generality of the inversion approach, and we agree that* *the patterns of the observational data help interpret the different performances of our calibrations. Precipitation has an important control on soil moisture, which plays a role in the seasonality of LH. At US-ARM, precipitation has a larger seasonal cycle with relatively dry winter and wet summer. Because annual rainfall is lower, LH is more limited by soil moisture availability. In contrast, US-MOz receives more rainfall throughout the year so LH is more controlled by solar radiation and hence peaks in July. Because US-ARM is more moisture limited, LH is more sensitive to model parameterization of soil hydrology. The inability of the model to capture the correct timing of the peak LH despite parameter calibration suggests that there may be structural limitations of the parameterizations used in CLM4. We have added more discussions on the results in the broader context of the various sources of model uncertainty and how their relative contributions may vary depending on the climate/hydrologic regimes in the revised manuscript.*

20. Figure 6: It seems that simulations of LH flux using some of the posterior parameter distributions (especially p_ra=0.9) are worse than the default set of parameters. This is especially so in winter, when none of the posterior distributions are better than the default, and most are worse. This should be commented on, since it is at variance with the authors' later claim that "Inversion results at the flux tower and MOPEX sites using monthly and daily surface flux and runoff observations show that the MCMC Bayesian inversion approach effectively and reliably improves the simulation of CLM under different climates and environmental conditions". The use of the adjective "reliably" does not seem justified.

    Response: *We agree and have added a few clarifications accordingly. We want to mention that for both Bayesian and deterministic (e.g., least-square fitting) approaches, likelihood is used to select posterior samples or optimum parameter values are sought to minimize an objective function. However even the most likely sample set, or the optimum sample set, are not expected to minimize the distances between observations and model predictions at every single observational point (e.g., month).*

21. 5091L6 "It is interesting to see that fmax is identically estimated by inversions with different reference acceptance probabilities. When the rejection standards are relaxed, the bounds of posterior distributions of most parameters become wider, and multi-modal patterns occur" It would be helpful if the authors could explore the reasons why the distribution of fmax might be insensitive to p_ra, especially when most other parameters are more sensitive.

    Response: *In fact, according to our previous study (Huang et al., 2013), fmax has a secondary impact on runoff in CLM. A possible reason for the consistency in its* *posterior bounds is that although fmax is not the major contributor of output variability, its relationship with the output response variables is relatively robust as determined by the parameterizations in CLM.*

22. Figure 7: These results conform more with my expectations (compared to Figures 1, 3, and 5, which did not). The posterior distributions using p_ra=0.5 (for Cs, f_over, f_drai, Q_dm, Sy, Psi_s) tend to be quite distinct from those obtained using other p_ra values (1, .95, .9). I would conclude from this that, using the authors methods, inversion of several hydrological parameters from runoff data can be achieved, but inversion of hydrological parameters from LH data cannot be reliably achieved.

    Response: *Yes, the simulated hydrological processes are apparently more sensitive to hydrological parameters. Therefore, the results of inverse approach using runoff data are better. But since these parameters also indirectly affect the energy budgets, observations of heat fluxes can provide useful or supplementary information for inferring model parameters and improving model performance.*

23. Figure 7: Do the authors agree that the results in Figure 7 are more in line with their expectations than those of Figures 1, 3, and 5? If yes, what do they think are the implications of that result?

    Response: *We agree that the results of Figure 7 are more in line with our expectations. It means that runoff data are more directly related to the unknown parameters and can be more reliably used for model calibration; for calibration using heat flux data, reduction of parameter dimensionality might be helpful to make the inverse problem less ill-posed, as we demonstrated later in the text.*

24. 5091L13 "larger variability than observations is noted from July to October" the modelled variability seems to arise from a high modelled streamflow in August. Why is that? Was there a single very large rainfall event one August for which the model runoff greatly exceeds the measured runoff?

    Response: *We can see from Figure 15 that the simulated runoff has more daily variability than the observed runoff in late summer and early fall. This leads to larger simulated monthly runoff than observed and is not related to a single rainfall event. Generally the model does not capture the large daily runoff peaks in the summer from rainfall events. Hence the excess soil moisture leads to larger runoff in later summer and early fall when rainfall events are less frequent and the observed runoff is basically very low. This suggests some systematic biases in the model parameters that cannot be fully removed by parameter calibration. We cannot exclude the possibility of measurement errors in either external forcing or observational heat fluxes.*

25. 5091L15 "Among the four sets of simulations based on inversion, more stringent sample rejection criterion results in a better match between the simulated responses with observations." This is as expected, and is good to see. Did this also happen for the LH simulations? If not, why not?

    Response: *Because hydrological processes are more sensitive to hydrological parameters, this conclusion is* not surprising for *inversion with runoff. But when heat fluxes are used for calibration, we also found that the RMSEs of flux simulations using the posterior estimates with the reference acceptance probability of 1.0 are better than others. It is worth mentioning though, that for more directly related observational data, a more stringent rejection criterion can help with calibration accuracy and precision, but for indirect information, a relaxed criterion is helpful to avoid biased estimates.*

26. 5092L2 "simulated LH and runoff are most sensitive to three subsurface parameters." Which three parameters? Is this sensitivity result reflected in the present study - were their posterior distributions narrower than those of other parameters?

    Response: *The three most sensitive subsurface parameters are fdrai, Qdm, Sy. In most cases, the posterior distributions of these three parameters with reference acceptance probability of 1.0, 0.95, and 0.9 have more similar pattern and narrower range than other parameters. We clarified this in the revised manuscript.*

27. 5092L20 "Using posterior estimates of the reduced parameter set can significantly improve the latent heat flux simulations compared to the results using the full-set of parameters, especially from October to December, and from January to May" Given these improvements, Figures 9 and 10 would be more interesting if they contained results from US-ARM, rather than US-MOz. Did the authors choose US-MOz for some other reason?

    Response: *Results from US-ARM and US-MOz deliver similar messages, so we kept only one set of figures for brevity. The following two figures show the results at the US-ARM site. The simulations using posterior estimates of the reduced parameter set can also significantly improve the heat flux simulation over all seasons. Especially in winter, simulations using reduced parameter set perform better than using the full-set. But the improvement only occurs at the US-ARM site.*

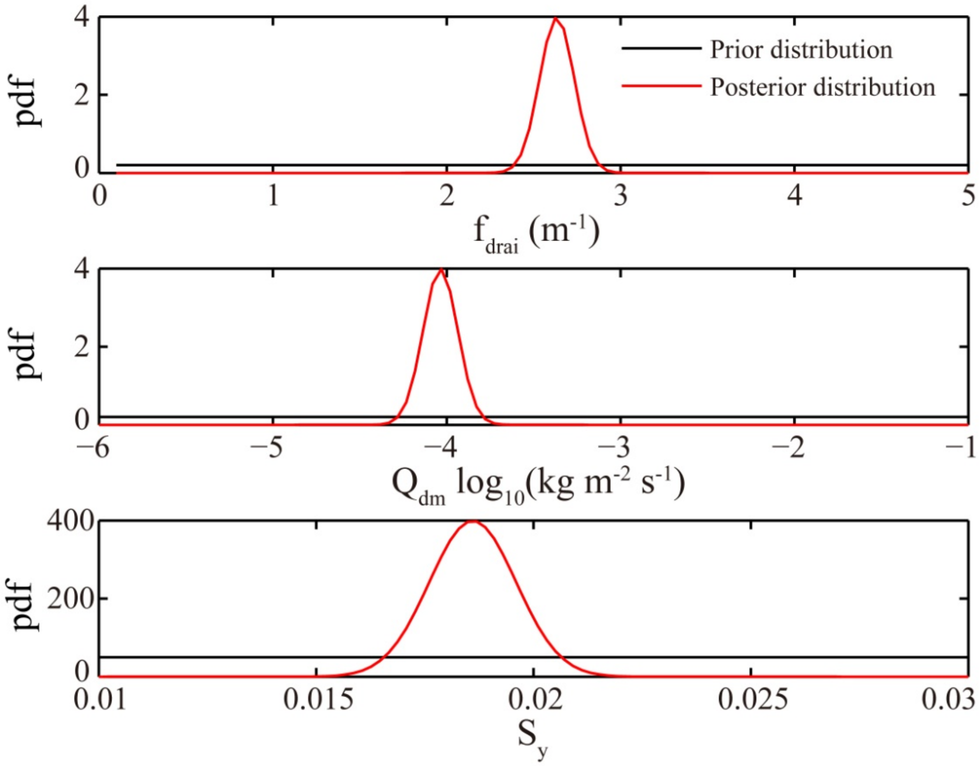


Figure R2 Posterior distribution of the reduced parameter set from previous sensitivity analysis at the US-ARM site.

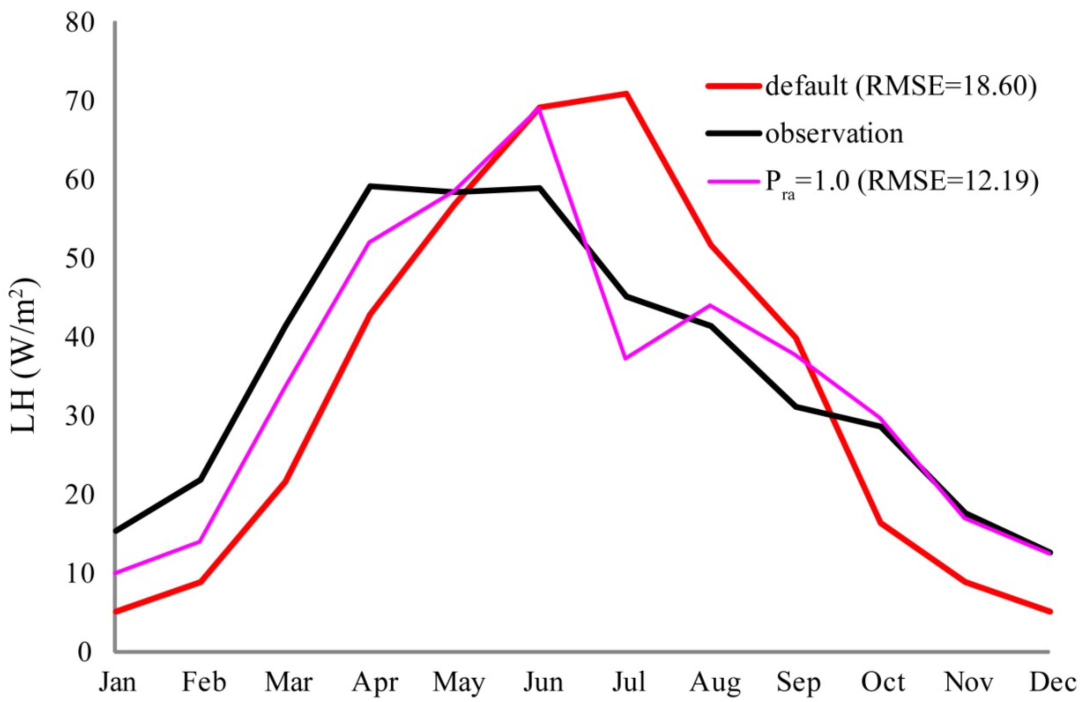


Figure R3 Simulated heat fluxes using the posterior estimates of parameters at the US-ARM site.

28. 5095L10 “Inverse modeling using heat flux at US-ARM and runoff at the MOPEX basin, which is located close to US-ARM, provides an opportunity to assess the impacts of data type on inverse modeling.” I think it is more than just data type that differs between these two model assessments! There are substantial differences in spatial scale, and hence in the dominant land surface processes and the spatial heterogeneity thereof, between the two cases.

    Response: *Please see the response to comment #13.We added more discussions on the scale discrepancy in the revised manuscript.*

29. 5095L14 “model inversion leads to more significant improvements in runoff (Fig. 8) than heat flux (Fig. 2) compared to simulations that use the default parameter values.” This might also be caused by having rather poor default estimates of the parameters which control hydrological processes, and rather better default estimates of the parameters that control LH fluxes.

    Response: *Thanks for the insightful suggestion. The default values, although yield poor simulations particularly in runoff, are currently being widely used in the CLM community, and this is the motivation of this study. Compared to runoff generations parameters, the parameters for LH simulations, such as LAI, are better constrained by satellite observations. We added this point to the revised manuscript.*

30. 5095L25 “may require structural changes in the hydrologic parameterizations combined with parameter calibration to improve model skill.” It could also be a problem with the forcing data for the MOPEX catchment. Does the NLDAS precipitation data agree with the MOPEX precipitation data (which is based on a relatively large number of rain gauges)?

    Response: *It could be related to the quality of forcing data, although not necessarily the precipitation data. We have the MOPEX precipitation data during 2002-2003, and Figure R4 shows the comparison between NLDAS and MOPEX data. These two datasets are consistent. The total precipitation of these two years is 1841 mm from MOPEX and 1901 mm from NLDAS respectively.*

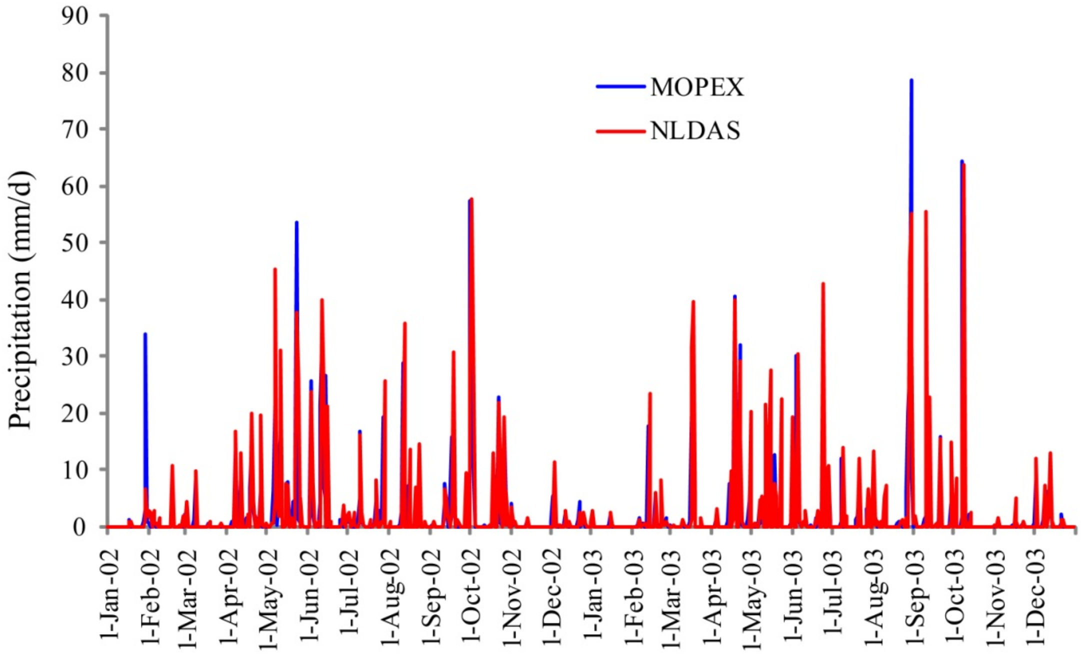


Figure R4 the comparison between NLDAS and MOPEX precipitation data.

31. 5097 I would have liked to see some discussion on the potential benefits of (i) using other observations (such as soil moisture), which more tightly link the water and energy components of the CLM model (ii) doing inversions which simultaneously consider runoff, LH flux (and soil moisture).

    Response: *Thanks for the suggestion. Your point that observations on different aspects of the hydrologic cycle could be used to better constraint the model parameters is well taken and it is among our top interest to explore in our subsequent studies. We added discussions on this in the revised manuscript.*

32. 5098L25 "The improvement is more significant for runoff than heat flux because the calibrated parameters are more directly related to runoff processes." I do not agree. I think the main reason that the improvement is larger for runoff is that the default parameters produce very poor simulations of runoff.

    Response: *Yes, the poor default is also an important reason. This is relevant to the response to point #27. We added discussions on this in the revised manuscript.*

Minor points

33. 5078L13 Unclear meaning: "the predictive intervals of the calibrated parameters

become narrower”.

Response: *The predictive interval means the range of posterior distribution range. We’ve clarified it in the text.*

34. 5079L08 “However, as the conditions are usually violated in practice, some regularization is generally needed to introduce mild assumptions on the solution and prevent parametric over-fitting.” Which of the 3 conditions is usually violated? All of them?

Response: *Yes, all of them, particularly uniqueness and stability.*

35. 5080L1 “the input and output uncertainties” By input do you mean the external forcing data, or the parameters or both?

Response: *the parameters.*

36. 5081L26 “covered by 6 % C3 grass, 22 % C4 grass, and 20 % croplands” What about the other 52%?

Response: *There is a typo in the manuscript. The Walnut River basin is covered by about 56% C3 grass, 22% C4 grass, and 20% croplands.*

37. 5083L16 “In practice, it is critical to evaluate and quantify the uncertainty associated with parameter estimation; therefore, we should consider stochastic inversion/calibration approaches (e.g. Bayesian inference) and describe the input/output uncertainties in a probabilistic manner.” This text belongs more in an introduction.

Response: *We have reduced the sentences.*

38. 5084L20 the symbol n is not defined.
39. 5085L2 the symbols p_ra and p are not defined.
40. 5085L6 the symbol q is not defined

Response: *We modified this part as follows:*

“The procedure is as follows:

a. *Initialize a random vector* ***m*** *from the prior distributions* $\{m_i^{(0)}, i = 1, ..., p\}$, *where p is the number of parameters;*

b. *Generate a random variable* $m_i^*, i=1,...,p$ *from the proposal distributions, and calculate the following ratio (note the probabilities in the formula are calculated using equation 2):*

$$\alpha = \min\left(p_{ra}, \frac{prob(m_i^* \mid m_1^{(1)}, m_2^{(1)}, ..., m_{i-1}^{(1)}, m_{i+1}^{(0)}, m_{i+2}^{(0)}, ..., m_p^{(0)})}{prob(m_i^{(0)} \mid m_1^{(1)}, m_2^{(1)}, ..., m_{i-1}^{(1)}, m_{i+1}^{(0)}, m_{i+2}^{(0)}, ..., m_p^{(0)})} \right), \qquad (3)$$

where p_{ra} *is reference acceptance probability;*

c. *Generate a random value* u *uniformly from interval (0,1);*

d. *If* $\alpha > u$, *let* $m_i^{(1)} = m_i^*$; *otherwise, let* $m_i^{(1)} = m_i^{(0)}$.

Repeating steps (b) to (d) by replacing index (k) with index (k+1), we can obtain many samples as follows: $\left\{ (m_i^{(k)}) : i=1,...,p,\ \ k=0,1,\cdots,n \right\}$, *where n is the number of sample sets. From the procedure, we can see that the value* $m_i^{(k)}$ *only depends on the current state of **m**, but not the previous states; therefore, these samples form a Markov Chain."*

41. 5092L1 "Our global sensitivity analyses across 13 flux towers and 20 MOPEX basins," Are you citing earlier work? If yes, please reference it.

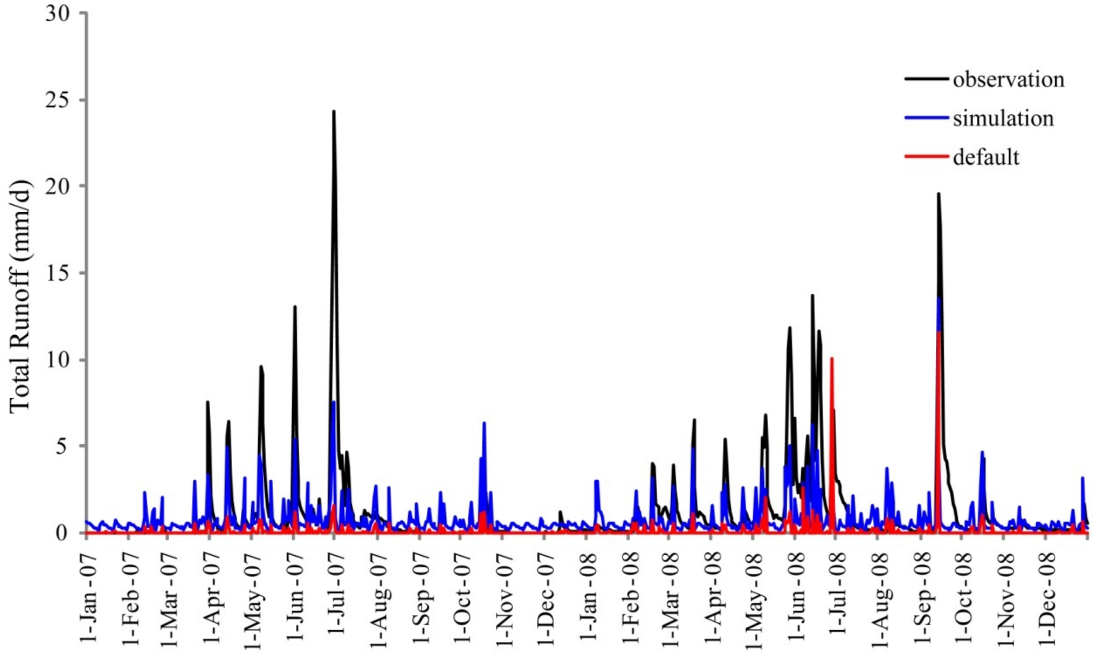
    Response: *We modified it as follows*:
    "Our global sensitivity analyses across 13 flux towers and 20 MOPEX basins (Hou et al., 2012; Huang et al., 2013),"

42. 5092L8 "a reduced set of parameters" How did you choose the values of the fixed parameters?

    Response: *The fixed parameters are the same as defaults.*

43. Figure 15B: there is too much temporal information compressed into this graph – the authors need to find an alternative presentation. For example, just show a single year of the validation, or present the daily data in summary form (e.g. flow duration curves).

    Response: *We replaced it with the runoff simulations during 2007-2008 as follows.*



44. 5097L19 "We found that RMSEs are reduced more for monthly data than for daily data" This was not clear to me. Which reduction are you referring to? What causes it?

    Response: *Sub-monthly variations involve more internal/external factors and processes that are more complicated. It is harder to capture the higher frequency components of the output variability.*

References

Beven, K.J., 1997. TOPMODEL: A critique. Hydrological processes, 11, 1069-1085.

Duan, J., Miller, N.L., 1997. A generalized power function for the subsurface transmissivity profile in the TOPMODEL. Water Resour. Res., 33(11). 2559-2562.

Hou, Z., Huang, M., Leung, L. R., Lin, G., and Ricciuto, D. M.: Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model, J. Geophys. Res., 117, D15108, 10.1029/2012jd017521, 2012.

Huang, M, Hou, Z., Leung, L. R., Ke, Y., Liu, Y., Fang, Z., and Sun, Y., 2013. Uncertainty Analysis of Runoff Simulations and Parameter Identifiability in the Community Land Model – Evidence from MOPEX Basins, J. Hydromet., 10.1175/JHM-D-12-0138.1.

Huang M, X Liang, and LR Leung. 2008. A Generalized Subsurface Flow Parameterization Considering Subgrid Spatial Variability of Topography and Recharge. J Hydrometeor., doi:10.1175/2008JHM936.1.

Iorgulescu, I, Musy, A., 1997. Generalization of TOPMODEL for a power law *transimissivity profile. Hydrological Processes, 11, 1353-1355.*

LeMone, Margaret A., and Coauthors, 2000: Land–Atmosphere Interaction Research, Early Results, and Opportunities in the Walnut River Watershed in Southeast Kansas: CASES and ABLE. Bull. Amer. Meteor. Soc., 81, 757–779.

Li H, M Huang, M Wigmosta, Y Ke, A Coleman, and LR Leung. 2011. Evaluating Runoff Simulations from the Community Land Model 4.0 Using Observations from Flux Towers and a Mountainous Watershed. J Geophys. Res., Vol. 116, D24120, 19 PP, doi:10.1029/2011JD016276.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, and L. E. Gulden (2005), A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models, J. Geophys. Res., 110(D21), D21106.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, L. E. Gulden, and H. Su (2007), Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data, J. Geophys. Res., 112(D7), D07103.

Oleson, K. W., D. M. Lawrence, G. B. Bonan, M. G. Flanner, E. Kluzek, P. J. Lawrence, S. Levis, S. C. Swenson, and P. E. Thornton (2010), Technical Description of version 4.0 of the Community Land Model (CLM), NCAR/TN-478+STR, National Center for Atmospheric Research, Boulder.

Woods, R.A., Sivapalan, M., Robinson, J.S., 1997. Modeling the spatial variability of subsurface runoff using a topographic index. Water Resour. Res., 33(5), 1061-1073.