Editor Decision: Publish subject to revisions (further review by Editor and Referees) (27 Nov 2016) by Prof. Dr. Kurt Roth

Comments to the Author:

The manuscript improved significantly. This is also attested by the reviewers. Still, some important questions pointed out by the reviewers remain open and need to be addressed. Please go through them carefully. Since the raised issues are beyond the mere technical, I'll keep the reviewers involved also in the next, hopefully final iteration.

Reply: **Thanks. We revised the manuscript taking into account the comments of the reviewers. In addition, we reformulated in our new manuscript almost all the sentences. We believe that it strongly improves the readability of the manuscript. Our results and conclusions are however not modified.**

As the text was edited thoroughly we didn't submit a marked-up manuscript version because you would see that entire paper is red.

Reviewer I

I appreciate the detailed answers and the additionally performed simulations to investigate the raised issues. However, I still have few comments regarding answers, which were inconsistent or not sufficient to me. I would like to see them clarified:

Major comments:

1. The expected impact of heterogeneity is now explained in line 376-380: "Qu et al. (2014) described the statistics of soil properties for soil samples taken in the Rollesbroich catchment. Soil texture showed a relatively limited variation. In our work only vertical heterogeneity is considered. In this case, heterogeneity does not seem to be very strong and we do not face a challenging upscaling case for the land surface model." However, Qu et al. (2014) seem to state that there is considerable heterogeneity, e.g.: "Spatial variability of the measured soil water content was higher at the 50-cm depth than the 5- and 20-cm depths, as indicated by the temporal dynamics of the standard deviation of soil water content presented in Fig. 2 (bottom panel). We attribute this to the pedological situation (shallow soil above consolidated bedrock) in which the highly variable stone content in the subsoil leads to considerable spatial variability of soil water content at the 50-cm depth."

Furthermore, the revised discussion also states (line 622-624): "The fact that we replace heterogeneous soil properties and soil moisture content for a given area by spatially homogeneous values, also introduces temporal variability in the effective parameters that are estimated in this study." To me this sounds like heterogeneity is a rather important challenge for upscaling. Please clarify. In this context (if heterogeneity is important) I didn't apprehend why the representation of photosynthesis was considered the most noteworthy model structural error (line 451-454): "The model error was set to zero assuming that uncertainty was captured by uncertain parameters and model forcings. However, we agree that it can be expected that we have other model structural errors, for example in relation to the representation of photosynthesis."

<u>Reply:</u> **We thank the reviewer for pointing out the improvement of our manuscript. It is true that in spite of a limited spatial variability in soil texture characteristics the spatial variability in soil moisture content is not so small. This could be related to the influence of groundwater and the presence of a drainage system, but also to variations in soil hydraulic properties although texture is quite homogeneous. We decided therefore to reformulate the sentences cited to (line 541-551):**

"In this work, we conveniently assume the soil-land-surface domain of the Rollesbroich site to be homogeneous and characterized by areal average values of soil moisture content at 5, 20 and 50 cm depth. In other words, we consider only vertical variations in soil water storage. Common LSM data assimilation experiments published in the literature usually involve application to much larger spatial scales, especially when remote sensing data are used. Hence, it is important to evaluate the LSM performance for a site where heterogeneities are neglected. Qu et al. (2014) investigated the geostatistical properties of the soils of the Rollesbroich test site. This work demonstrated a rather small spatial variability of the soil texture. This does not suggest, however that we can ignore spatial variations in the measured soil moisture values. Indeed, the standard deviations of soil moisture vary between 0.04 and 0.07 cm^3/cm^3 depending on the actual soil layer. This spatial heterogeneity of the soil moisture data documents variability in the soil hydraulic properties, and complicates the application and upscaling of LSMs."

We take the simple or biased representation of photosynthesis as an example of model structure error, but it doesn't mean that the representation of photosynthesis is the only most noteworthy model structural error. Models are assemblies of assumptions and simplifications and thus inevitably imperfect approximations to the complex reality. We would say all these simple parameterizations and mathematical implementations (e.g., spatial and temporal discretization) can lead to model structure error. We reformulated the sentences in the new manuscript (line 578-585):

"*Also, we account crudely for errors in LSM model formulation via parameter uncertainty (discussed next) and the use of a stochastic description of the precipitation record of the Rollesbroich site. The hyetograph of each ensemble member is derived by multiplying the measured hourly precipitation rates of the tipping bucket with multipliers drawn from a unit-mean normal distribution with standard deviation of 0.10. This is equivalent to a heteroscedastic error of 10% of the observed precipitation (Hodgkinson et al., 2004). Forcing variables which govern evapotranspiration (incoming shortwave and longwave radiation, air temperature, relative humidity, and wind speed) were not corrupted. Of course, the prior parameter distribution and precipitation forcing do not account for errors in the photosynthesis module*."

2. I appreciate that the authors investigate the ensemble inflation method (line 549-552): "The effect of initial uncertainties on the performance of EnKF with the ensemble inflation method is also tested with the VIC-3L model. The forcing error was increased from 10% to 20%. Table 6 shows the RMSE values for soil moisture content characterization in the assimilation and verification periods. The difference between the results for 10% or 20% perturbation of the forcings is very limited, for both variants of the EnKF-method." However, I suspect that there is a misunderstanding. How can you assess the effect of the initial uncertainties of the parameters, by changing the forcing error and not the initial uncertainties? As a side note: The inflation method keeps the parameter uncertainties constant. However, you now also state that "parameter uncertainty decreased" (line 636). How could you attest this?

<u>Reply:</u> The reviewer asked about the impact of initial uncertainty (without further specification) on the performance of the ensemble inflation method, and we interpreted this as the role of the uncertainty of the model forcings, as this uncertainty was less well defined compared to the uncertainty of the soil texture. In the experiments also the uncertainty of texture was taken into account and we thought that this uncertainty was not underestimated and therefore did not consider increasing its perturbation. In addition, we thought the question was more about the inflation method itself. When Whitaker and Hamill (2012) proposed the inflation method we used in our study, experiments were conducted to account for background errors not accounted for by the first-guess ensemble, which in their study included both sampling error and model error. Their tests with large and small ensembles, with and without model error, suggested that this inflation method is well suited to account for unrepresented assimilation errors such as sampling error. In our study, we further found that the effect of forcing error on this method was limited. Therefore we think that this method is a well-built methodology and broadly used in data assimilation research. We admit that the "parameter uncertainty decreased" (line 638) is confusing. The inflation method kept the parameter spread constant. We deleted this statement in the revised version to avoid the misunderstanding.

Minor comments:

3. Table 5: In original manuscript Figure 8, there was basically no difference for the prediction of the water content of the first layer, whether there are parameters estimated or not. The authors clarified this: "Predictions of soil moisture content for layer 2 and layer 3 (in the verification period) improved significantly for the case of parameter estimation. Concerning the soil moisture content of layer 1, the RMSE value of the open loop run is 0.053m3/m3, which is already quite close to the observed values. In addition, the soil moisture content for the upper layer is strongly driven by single precipitation events. We extended the discussion of these results (line 533-536): "In the verification period, the RMSE values of the scenario noParamUpdate are close to the RMSE values of the open loop run. If soil parameters were updated during the assimilation period, the RMSE values for soil moisture characterization were reduced. More specifically, the four methods show a RMSE improvement of about 54% and 42% for the second and third model layer (compared with the open loop run)."

The part of the answer ("Concerning the soil moisture content of layer 1, the RMSE value of the open loop run is 0.053m^3/m^3, which is already quite close to the observed values. In addition, the soil moisture content for the upper layer is strongly driven by single precipitation events.") was added to the results for CLM instead of VIC-3L. Please correct. Furthermore, to me this statement doesn't seem entirely consistent with the discussion (line 675-682): "In the verification period soil moisture of the *top layer cannot be represented perfectly by the two LSM's, in spite of parameter updating with state of the art data assimilation methods. Table 5 and table 9 illustrate that the RMSE values of the four joint state and parameter assimilation methods are similar for both models, which means that both models have larger errors for the top layer. There is a number of reasons for the larger soil moisture mismatches for the upper layer: (i) the memory effect from initial conditions, very well identified at the beginning of the verification period, is smaller for the upper soil layer, as this layer is more affected by precipitation events and evaporation; (ii) these soil moisture changes make that it is also more affected by model structural errors, for example concerning evaporation processes."*

Reply: **We revised almost all the text in the revised version. We discussed this issue in the discussion (line 872-876):**

"Despite this improvement in model performance over an open-loop simulation, VIC-3L and CLM do not adequately characterize soil moisture dynamics of the top layer (5 cm measurement depth) during the evaluation period (RMSE values of about 0.05 cm^3/cm^3). We posit that these two models do not characterize adequately processes such as water infiltration, soil evaporation, and/or root water uptake (transpiration), which govern rapid variations in soil moisture storage in the top soil."

4. Figure 4: In original manuscript Figure 5, Parameter b estimated by MCMC showed a large difference to the other methods. But MCMC performed approximately as well as the other filters. The authors clarify this in their response: "Demaria et al. (2007) evaluated the sensitivity and identifiability of ten parameters which control surface and subsurface runoff in the VIC model for four U.S. watersheds along a hydroclimatic gradient. They found that parameter b is crucial in a dry environment, while its impact on model performance is not significant for wet sites. They concluded that parameter b plays a key role in partitioning rainfall into soil moisture and surface runoff in dry environments. [Liang and Guo, 2003] and [Atkinson et al., 2002] reached a similar conclusion. In our work, as the Rollesbroich catchment is very wet, even though parameter b estimated by MCMC shows a large difference with other methods, it shows small impact on the soil moisture content for layer 1 and layer 2. In the revised manuscript, all experiments of VIC-3L were done again. Evolution of parameter b estimated by MCMC was more reasonable (figure 4)." Figure 4 now shows a similar value for b estimated with MCMC. What was changed to achieve the new results?

Reply: **We redid all the experiments including the generation of the initial ensemble of parameters.**

5. Figure 5: In original manuscript Figure 6 parameter spreads at time 0 seemed different. The authors clarified this: "In original manuscript figure 6 shows the evolution of the parameter ensemble from time step1 but not time step 0. At time step 0, the ensemble spreads are the same, but at time step 1, the parameter ensemble is updated by PF or EnKF, and the ensemble spreads between EnKF and PF differ. We showed the evolution of the parameter ensemble from time step 0 onwards in the revised version of the manuscript (figure 5, figure 8 and figure 11)." Now MCMC, AUG and DUAL seem to have the same initial spread, but to me PF still seems to have a different spread?

Reply: **For figure 5, we checked the source datasets for the figure. PF has the same initial parameter ensemble at time 0, but when I plotted the figure, PF started still from time step 1. This has been modified in the new version of the manuscript.**

6. Equation A8: The authors missed to explain the inconsistent dimensions (you add [LT-1], [] and [L]).Please clarify or correct the equation.

Reply: **Thanks. We corrected the unit in the revised manuscript.**

7. Figure 11: I agree that the shown parameters are more meaningful than the estimated soil texture. I still think it is worth to mention (e.g. in the caption), that the shown parameters are not the directly estimated ones.

Reply: **The text was added in the caption of figure 11 in the revised manuscript:**

"Please note that $log_{10}k_s$ (in $log_{10}(m/s)$) and parameter B are derived from the sand, clay, and organic matter fractions of each soil layer, which are estimated during the assimilation period."

8. Line 768: If you do mean the soil matric potential and not the matric head, the dimension is not [L], but [EL-3].

<u>Reply:</u> We corrected this to "soil matric head". Confusing is that the technical description of CLM 3.5 [*Oleson et al.*, 2004] refers to it as soil water matric potential.

9. Concerning the reply about RMSE: "Thanks, we admit that for a prediction in the verification phase we cannot expect a better result than a RMSE equal to the measurement uncertainty but still an RMSE equal to 0 would be the best result." A RMSE equal to 0 is not the best result. It would mean, that the model perfectly describes the measurement noise, but not the actual state.

<u>Reply:</u> Thanks, you are right. But in reality we don't know the actual state values, so we still think that a smaller RMSE value indicates a better prediction (line 641-642):

"*Larger values of the NSE and smaller values of the RMSE are preferred as they indicate a better LSM performance.*"

Reference

Crow, W. T., D. Ryu, and J. S. Famiglietti (2005), Upscaling of field-scale soil moisture measurements using distributed land surface modeling, *Advances in Water Resources*, *28*(1), 1-14, doi:10.1016/j.advwatres.2004.10.004.

Oleson, K. W., et al. (2004), Technical Description of the Community Land Model (CLM), *NCAR Technical Note NCAR/TN-461+STR*, 186.

Qu, W., H. R. Bogena, J. A. Huisman, G. Martinez, Y. A. Pachepsky, and H. Vereecken (2014), Effects of soil hydraulic properties on the spatial variability of soil water content: evidence from sensor network data and inverse modeling, *Vadose Zone Journal*, 13(12), doi:10.2136/vzj2014.07.0099.

Vereecken, H., R. Kasteel, J. Vanderborght, and T. Harter (2007), Upscaling hydraulic properties and soil water flow processes in heterogeneous soils: A review, *Vadose Zone Journal*, *6*(1), 1-28, doi:10.2136/vzj2006.0055.

Whitaker, J. S., and T. M. Hamill (2012), Evaluating Methods to Account for System Errors in Ensemble Data Assimilation, *Monthly Weather Review*, *140*(9), 3078-3089, doi:10.1175/mwr-d-11-00276.1.

Xiao, J. F., K. J. Davis, N. M. Urban, K. Keller, and N. Z. Saliendra (2011), Upscaling carbon fluxes from towers to the regional scale: Influence of parameter variability and land cover representation on regional flux estimates, *Journal of Geophysical Research-Biogeosciences*, *116*, doi:10.1029/2010jg001568.

Reviewer II

As I already reviewed the first version of the discussion paper as reviewer #2, I will focus here on the reply of the authors and the changes made in this version of the manuscript.

I appreciate the efforts undertaken by the authors to revise this manuscript, which improved substantially at least from my point of view. All my comments were addressed properly, also the readability of the figures is much better now.

Reply: Thanks for pointing out the improvement of our manuscript.

I have only some words to say about the comment and discussion about temporal non-stationary parameters: The discussion introduced in ll.620-639 in the revised manuscript generally tackles this issue well, but does not really answer the question about the predictive power of such parameters. Similar with the editor, it is not quite clear for me, how parameters, which are meant to change in time, can be used for predictions. In case, that there is a theoretical reason for this, the temporal evolution should then enter the model equations to improve the predictions. It is reasonable, that the parameter may evolve in time as a consequence of using neglecting spatial heterogeneity, as described in ll. 622-625, but then we get the right answer for the wrong reasons, and predictive power is not enhanced. Although this issue is important at least from my perspective, I see that the authors follow a different focus in the paper, which is elaborated thoroughly and presented in a straightforward manner. I would appreciate when the authors could tackle this point as well, but I can accept their decision to work on that aspect not in great detail.

Reply: Thanks. Since our focus of this manuscript is to compare the four methods of joint assimilation of states and parameters in two land surface models and the manuscript is already long, we think that we will not work on this aspect in great detail. This is a difficult issue and we believe that estimated soil hydraulic parameters will also do a quite good job in other time periods, probably because the effective parameter values for the larger grid cell would not change so much (this is a guess). However, in case exhaustive high quality measurements would be available over a large area this could be further tested and it could for example be detected if over a year effective (estimated) parameters would show a typical yearly cycle which would be relatively similar for different years. Such a yearly cycle of effective soil hydraulic parameters could also be used then for other years. This remains however speculative and now we tend to think that estimated effective parameters for a larger grid cell are certainly not perfect under all conditions, but would in general lead to better predictions, also under different meteorological conditions. We included this small discussion on the quality of the effective parameters also in the discussion (line 917-925):

"*It is difficult to assess whether the inferred VIC-3L and CLM parameter values will do a good job at predicting soil moisture dynamics at the different measurement depths during a much longer evaluation period with wet and dry conditions. As the estimated parameters represent effective properties of the Rollesbroich site, one may expect their calibrated values not to change too much over time. We would need additional soil moisture data and/or other type of measurements to corroborate this. Nevertheless, the effective parameter values derived herein improve characterization of soil moisture dynamics at the Rollesbroich site compared to a separate state estimation run with VIC-3L and CLM using parameters drawn randomly from the prior distribution, or open loop simulation using the ensemble mean model output of a large cohort of parameter vectors drawn randomly from the prior parameter distribution (initial parameter ensemble).*"

ll.636-638: "For highly identifiable parameters, parameter uncertainty decreased and parameters converged fast. So we think that joint estimation of states and time-variant parameters with data assimilation still shows great potentials in terms of identifiability of parameters." I do not understand the point. When parameters were highly identifiable, then the method works well. That is clear to me, but what is the "great potentials in terms of identifiability"? Do you mean to distinguish between such parameters which can be constrained well and others which do not? It would be more interested to hear about the behaviour, when parameters are not well identifiable as it is typically the case for complex LSMs. Please clarify this.

<u>Reply:</u> We meant to emphasize the capability of joint estimation of states and time-variant parameters with data assimilation in parameter estimation. We revised almost all the text in the revised version and we discussed this issue in the new manuscript (line 898-916):

"In our implementation of the EnKF and KF, the VIC-3L and CLM parameters were assumed to be time-variant and their values updated jointly with the model states at each assimilation time step. The 5-month calibration period we used herein involves several large precipitation events, and as a consequence, the soil profile is rather wet. The resulting parameter estimates might therefore not be representative for dry periods with much lower moisture values of the soil profile. What is more, the assumption of spatial homogeneity might not characterize adequately the distributed soil properties of the Rollesbroich site and induce temporal variability in the VIC-3L and CLM parameters. Bias in model input and measurement errors of the forcing data also contribute to the temporal fluctuations of the estimated parameter values. These temporal parameter variations are meaningful in some cases as they can help diagnose structural model inadequacies and/or biases in model input and forcing data. Kurtz et al. (2012) successfully estimated a temporally-variant parameter with the EnKF, but these authors concluded that the algorithm needs a considerable spin-up period to "warm-up" to new parameter values. Vrugt et al. (2013) found considerable temporal non-stationarity in the parameters estimated by PMCMC as a result of the small time period used to calculate the acceptance probability of candidate particles. This finding is in agreement with the results of PMCMC in our paper. Of course, we could have assumed time-invariant parameters via a method such as SODA, yet this would have enhanced significantly computational requirements. Fortunately, parameters estimated via our implementation of the EnKF exhibit asymptotic properties during the assimilation period (e.g. see Shi et al. (2015)). This is particularly true for highly sensitive parameters. An example of this was parameter Dm of VIC-3L which quickly converged to values of around 1 – 2 mm after assimilating just a handful of soil moisture observations."