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.”

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?

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 0.053m3/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.") 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."

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?

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 step 1 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?

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

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.

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

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.