

Reply letter to Anonymous Referee #2

1) Synthetic and real data assimilation experiments are not performed or presented in a consistent way. This prevents results to be meaningfully compared:

1.a) Synthetic studies show “localization degrade analysis for the univariate case when either SM or GW observations are assimilated (Figure 3)” while the multivariate synthetic case and the real data assimilation case shows “improvements in the analysis (Table 3 and Figure 4)”. Synthetic studies always reflect the ideal/perfect conditions as they can be fully controlled (particularly the observation errors). I don't see the utility of an application that does not give satisfactory results using synthetic simulations. On the other hand, I am completely puzzled how real data scenario improves the analysis (Table 3) despite synthetic studies fail (Figure 3). Frankly, I would have been more convinced if synthetic results showed improvements while real DA case showed problems (after all real life is not perfect), but not vice versa. This inconsistency should be explained in detail.

We believe the misunderstanding here is due to our misleading sentence ‘As we can see from the above, the experiments with assimilating head or soil moisture show degraded results using localization.’(P10, L31 in original manuscript). In univariate assimilation result shown in Figure 4, the assimilation experiments with localization (DA_HLoc, DA_SM5Loc and DA_SMBotLoc) all improve the assimilated variable compared to the NoDA experiment. However, we did see a larger improvement if localization is not used (DA_H and DA_HLoc, DA_SM5 and DA_SM5Loc). The sentence is revised in the updated manuscript (P11, L17-19). The reasons are also explained in the text (P11, L17-24). Figure 5, however, shows that localization is preferred when assimilating both variables (DA_HSM and DA_HSMLoc_DV).

We did not attempt to demonstrate ‘the synthetic studies fail but the real data studies work’. Actually, from Figure 4, the assimilation results in all experiments improve either groundwater head or soil moisture. The synthetic studies (on Karup catchment) work fine with respect to univariate assimilation and multivariate assimilation with the proposed localization scheme. The selected settings are used for the Ahlrigaarde catchment with real data assimilated, and also work fine. Overall, the synthetic studies and real studies are consistent.

1.b) Only the total ET values are given for the synthetic simulations, but not the error statistics of ET (Table 2). They should be given similar to real data assimilation scenario.

In the real data assimilation case, we have the observed ET data at Voulund station, which allows us to calculate and evaluate the error statistics. In the synthetic study, we believe it would be very arbitrary if we examine ET at one random location. Therefore, in the revised manuscript, we calculate the averaged RMSE with respect to the true model for actual ET at 35 stations (same locations as SM and H observations) and show the error statistics in Table 3 instead of accumulated values.

1.c) Pick an accuracy statistics (R²; RMSE; or Nash-Sutcliffe efficiency, NSE) and just present it in a consistent way through out the study. Presentation of mixed statistics is very confusing: Table 2 shows R² and NSE for discharge, Table 3 shows RMSE for SM, Table 4 shows R² and RMSE for ET while the same table shows R² and NSE for discharge, Figures 3,4,5,9 shows RMSE. To start with, just give the equations of the error chosen statistics. In general, showing both NSE and RMSE is redundant. $NSE = (1 - \sum(X-Y)^2 / \sum(X-\mu_x)^2)$ can be shown to be equal to $(1 - RMSE^2 / \text{variance}_x)$, implying there is a directly relationship between NSE and RMSE: if NSE is high we can expect RMSE to be low and vice versa.. This analytical expectation is also supported by the results given in Tables 2 and 4: higher R² experiments have higher NSE and lower RMSE. Briefly, these three statistics are consistent with each other in representing the accuracy of the variable of interest. Just pick one out of these three statistics and just show it consistently for all cases (I recommend RMSE in this case, it is up to authors though).

We agree that the presented accuracy measures are related. In the revised manuscript, RMSE is used for head, soil moisture and actual ET, while NSE is used for discharge.

1.d) Table 2 (synthetic data assimilation experiments) is missing “DA_H” and “DA_SM5” scenarios (no localization scenarios). These scenarios are necessary to see whether or not Localization improves or

degrades the analysis accuracy. Given one of the major conclusions the authors are making is “Localization does not only provide better results . . .”, presentation of these results is very important.

We apologize for the flaw and have now included all experiments in the table (Table 3).

1.e) Figure 3 (synthetic data assimilation experiments) is missing “DA_SMBot” (no localization scenario). Similar to above, these scenarios are necessary to see whether or not Localization improves or degrades the analysis accuracy.

As above all experiments are now included in the figure (Figure 4).

1.f) Tables 3 and 4 (real data assimilation experiments) are missing “DA_H”, “DA_SM”, and “DA_HSM” scenarios (no localization scenarios). Similar to above, these scenarios are necessary to see whether or not Localization improves or degrades the analysis accuracy.

Different experiment settings are investigated in details using the synthetic data in Karup catchment as the experiments are controllable and computationally efficient. More specifically, both univariate and multivariate assimilations are fully studied including different localization schemes and different ensemble sizes.

In general, the synthetic DA study is used to evaluate and compare the different methods while the real DA study is used to evaluate the applicability of selected methods. Therefore we found that it is not required to repeat all experiment scenarios for the real data experiment. Moreover, the computational time for the Ahlrigaarde catchment is excessive and to carry out additional three scenarios would take 20+ days.

2) I see SM simulations/assimilation experiments are not realistic.

2.a) SM observations often have inconsistencies with models (see Reichle and Koster, 2004). For this reason, SM observations are matched to model before they are assimilated. Have authors done something similar? I see authors have matched the means of GW before observations are matched (page 14, line 9), but not the standard deviations or any other statistical property. To start with, do model and observations have similar statistical property? Did authors check the innovations of ETKF? Is it white?

We agree that it is very important to eliminate inconsistencies in SM and other observations before assimilating. For the ground observations used in this study, a quality check was carried out for both GW and SM to filter out the unrealistic values (e.g., negative values) and to match the means. We made sure that the processed observations were basically in the range of the model ensemble space, and that no significant model/observation biases were present, which could degrade the DA performance. Examples can be seen in Figure 7 and 8. Overall, we see that the SM observations show larger dynamics compared to the model simulations. The observation error standard deviations are defined empirically in ‘trial and error’ manner. By evaluating the assimilation results and the innovation statistics, we chose the current settings. A systemic study examining the assimilation results with respect to different model uncertainty and observation uncertainty can be found in Zhang et al. (2015).

2.b) Many satellite missions (e.g., SMAP) have the goal of retrieving SM observations with 4% error and consistently Mike She model using real data has similar error magnitude. However, the synthetic Mike She model runs do not seem to be realistic with SM errors of 2%.

An observation noise for SM (volumetric water content) corresponding to a standard deviation of 5% is assumed for both synthetic and real data. This observation error is assumed based considering equipment accuracy specifications and experience. Please also see our reply in 4.c)

2.c) Root-zone SM varies much slower than surface SM. As a result, actual root-zone SM errors have smaller magnitude than surface SM errors (i.e., root-zone RMSE is expected to be smaller, while RMSE values normalized with actual variability of the variable could be different for two layers). In the current study, root-zone SM errors are higher than surface SM errors (Fig. 4). Deeper layer (22.5cm) SM seems more

noisy than surface (2.5cm) as shown in Figure 9. Explanation is needed.

For both observations and model simulations SM shows much higher dynamics near the surface. Both Figure 4 and 9 show the SM error between the assimilation result and the observation, but do not indicate the SM dynamics.

2.d) Above points can be clarified very easily using a table showing: mean and standard deviations of observations and model (no assimilation) for both SM (at 5, 22.5, 50cm depths) and GW. Such a table seems necessary to clarify all of above points.

Overall, we found the SM data shows general higher dynamics compared to the model simulation. For the GW, model simulation has similar dynamics as observations in most wells, with the averaged RMSE of around 1 meter after calibration (mentioned in P9, L23). One can also see the example from Figure 7 and 8, to have an idea about the different dynamics between model and data. Therefore, we think readers should have a better overview of the assimilation and relevant data and we hope to convince the readers that the experiments presented are realistic.

3) Why select ETKF but not EnKF? In atmospheric sciences ETKF could be a viable option because it does not perturb observations (atmospheric models are chaotic, hence such perturbations could be problematic; but hydrological models are not chaotic). Avoiding the perturbation of observations does not seem to be a sufficient justification to use ETKF. Many studies use EnKF, hence it is more relevant to general audience. Briefly, it is better to say "ETKF selection is arbitrary, EnKF could have been selected as well" rather than justifying the ETKF selection via "it avoids the additional perturbation step". In case authors would like to support the selection of ETKF with "it avoids the additional perturbation step" argument, then they should support this claim with a reference (i.e., a study shows this additional perturbation could be problematic in hydrological sciences).

We agree and revised this sentence accordingly (P6,L3-4). We actually tested EnKF and found out that it gave similar result as ETKF for this hydrological application.

4) Many details are not given in the paper and currently the experiments cannot be replicated by another researcher. Besides many experiment set up decisions looks very arbitrary:

4.a) which datasets are available for the warm up period over Ahlrigaarde (i.e., for forcing/parameter)?

The data used driving the model is now described in section 5 (P14, L6-8).

4.b) forcing and parameter perturbation statistics for real data assimilation case are completely missing (which forcing variables/parameters are perturbed?, additional/multiplicative noise? mean?, standard deviation?, daily/weekly perturbation?, etc). How did authors decide about these statistics, justification? The overall principles behind forcing and parameter perturbation are described in section 3.1 (Page 6, line 13-17). The forcing perturbation is defined empirically and the parameter perturbation is defined based on the calibration results. The choice and the uncertainty of model parameters for the Karup model has been reported before (Zhang et al., 2015), and therefore is not fully described here. In the revised manuscript, we provided details on the perturbation including a new table (Table 2) where details on calibrated parameters and associated uncertainties are given for the Ahlrigaarde model.

4.c) for the synthetic experiments observations are perturbed using noise with 0.15m and 5% error standard deviation. Why these numbers? Do authors know Decagon 5TE SM sensors have errors of 5% at daily time step? (i.e., SM observations are assimilated at daily time steps even though they are observed every 30min; implying the errors at daily time steps could be much lower than 30 min time step).

According to the specification of the Decagon 5TE, the accuracy is $\pm 3\%$ VWC. However, in reality, the uncertainty can be much larger due to other factors. Averaging of data can indeed reduce the error to some extent. Overall, we assume the standard deviation of SM observation noise to be 5% (volumetric water content) for both synthetic and real data. The standard deviation of the observation error for GW is also assumed empirically.

4.d) For the real data assimilation scenario, the soil moisture error standard deviation is assumed constant 5%. Why 5%? Is it something the company who produces these Decagon 5TE sensors suggests? Besides, these half hourly observations are averaged into daily values, implying the

observation noise if further reduced. Justification for the observation error standard deviation is needed.

Please see above.

4.e) perturbation details in generation of ensembles is missing as well (the study of Zhang et al., 2015 is referred for these perturbations but the experiment details should be given specifically).

Perturbation details are elaborated in the revised manuscript. Please see the reply in 4.b) and Table 2

4.f) what is the temporal resolution of the model? Daily? It is forced by daily precipitation and reference evaporation (page 4 line 26), bi-hourly GW observations are obtained for real DA case (page 14, line 2), and GW observations interpolated to weekly time steps are assimilated (page 14, line 28) but the temporal resolution of the model (i.e., at what time step the model is run) has never been explicitly mentioned.

The temporal resolutions of the model are added in the revised manuscript (P4,L22-24 and P4,L31 to P5,L2)

4.g) if the model is run at daily time step, then it is not clear how weekly GW observations are assimilated.

The model is running with a dynamic time step less than half day, which is smaller than SM observation frequency (daily at midnight) and GW observation (weekly at midnight). The assimilation is carried out at the moment when available observations (SM or H or Both) are available. The procedure is also described in P6 L22-25. We do not aggregate observations and the asynchronous assimilation approach is not used in this study.

4.h) How ET observations are obtained at the stations? What kind of observations are these (eddy covariance? lysimeter? pan evaporation? Reference ET?)? How frequently obtained? No details are given.

Data on daily ET are available from an eddy covariances flux station established in the catchment (P15 L15-16). Details on data retrieval and processing can be found in Ringgaard et al. (2011) .

4.i) Which data are used as validation? Do authors use the same stations where observations are collected? Perhaps some stations should be reserved and observations obtained over these stations should not be used in any ways (e.g., assimilation) and should be left purely for validation. After all, if the station has soil moisture probe, then why bother with estimating the SM over that location (this comment is only relevant with direct validation of SM, should not be thought in the framework of SM assimilation to estimate other parameters such as runoff, ET, etc).

In both catchments, the stations used for assimilation are also used for validation. We agree that cross-validation ideally should be done. In the Ahlgaard catchment the number of measurement stations is limited and thus we decided to use all of them for assimilation. However, in the synthetic DA experiment in Karup catchment, we validated the assimilation in the entire catchment including at the unobserved locations.

4.j) Have you considered assimilating remote sensing-based SM data? Why not?

Yes remote sensing-based SM data was also considered for assimilation e.g. products from the SMOS satellite. However, the spatial resolution of this product is coarse (about 44 km) and the catchment is therefore covered by only few SMOS grids. The differences in resolution, value and uncertainty between in-situ data and SMOS data will significantly increase the assimilation complexity. As the first step, we constrained the test to assimilation of in-situ measurements of soil moisture. Ridler et al. (2014) already tested assimilation of SMOS-derived soil moisture for the same catchment. Please also note that we list assimilation of remote sensing based products as a suggestion of future work.

4.h) "The assimilation performance is evaluated by comparing model output with the actual observations using average RMSE." Now I am puzzled, which observations are assimilated and which observations are used as validation? Are they the same?

Yes they are same. This sentence is a bit misleading and is revised (P16, L6-7).

5) The study could present more literature review.

5.a) For example Franssen & Kinzelbach (2008) did assimilate GW observations into a hydrological model. Mike She model background/physics (page 4, lines 14-20) could be supported with references as well. ETKF is first introduced by Bishop et al., (2001) not by Sakov et al., (2010); the latter study used EnKF they did not even use ETKF (Page 3, line 16).

We corrected and improved the reference in the revised manuscript. The Mike She model

background/physics is referenced with MIKE SHE technical reference manual (P4,L20). ETKF reference is also corrected.

5.b) SM and GW time scales are often very different; the impact of precipitation/SM variations could translate to GW variations only after days/weeks/months, depending on the catchment/GW levels/conductivity/etc. On the other hand, the Mike She model simulations have daily time steps (my understanding daily, I could be wrong → authors should clarify this info). If the time scale is very long, it is not immediately clear how SM anomaly for today will give meaningful information about GW anomaly of today. Perhaps the filter (ETKF) should be changed to accommodate past observations? So, sufficient motivation about the utility of SM observations to improve GW simulations using ETKF should be given.

We also don't know if SM anomaly can give meaningful information to GW anomaly or the other way around by using ETKF given that the time steps are quite different. In MIKE SHE the saturated and unsaturated zones are explicitly coupled. This is done to optimize modelling time steps used in the unsaturated zone (minutes to hours) and saturated zone (hours to days), respectively. The flux between the unsaturated and saturated zones is calculated by an iterative procedure that conserves mass for the entire column. This means that assimilation of SM may have an effect on GW and vice versa through this explicit coupling. Therefore, to study their joint assimilation in an intergraded model is very meaningful.

6) "Although the improvements are relatively small, we nevertheless see the benefits in other model process results when improving groundwater head and soil moisture." It is not clear what is meant with this sentence; do authors imply "we did not find improvement but we think it is still likely in other applications, hence GW and SM should be assimilated together"? After improvements were not found in this study, I think the only comment can be said is "these results should be verified using other models". The interpretation of "seeing the benefits in other model process results" is not can not be made using the results of this study alone.

By saying the 'other model process', we mean the other modelling processes within MIKE SHE (e.g. streamflow and ET). We do not refer to other models. We revised this sentence to avoid the misunderstanding (P17,L26-27).

7) "Localization does not only provide better results . . ." (Page 16, line 25), synthetic results do not support this comment (Figure 3).

Actually, we will argue that the synthetic results support this comment if we use the deterministic model (NoDA) as the reference (P11,L17).

8) Abstract requires slight modification: the assimilation method used here "ETKF" has been around for almost 15 years, so this is not the first time it is introduced. Assimilation is not a new method either, has been around for a long time too. So, this current study is just a case study (GW and SM observations are assimilated in a hydrological model). Abstract should be changed to something similar to below:

"Observed groundwater head and soil moisture profiles are assimilated into an integrated hydrological model. The study . . . model code. Experiments were firstly performed using synthetic data in a catchment of less complexity (the Karup catchment in Denmark), and later performed using real data in a larger and more complex catchment (the Ahlergaarde catchment in Denmark)." ~C~

We agree and have revised accordingly.

MINOR:

Page 3, line 15, ". . . performance of assimilating soil moisture" → ". . .performance of a filter assimilating soil moisture" .

Page 4, line 4, "Karup catchment is well-studied catchment", in terms of what parameters/variables?

Page 10, line 17, "assimilating soil moisture at 5 cm depth . . . and 1 m depth". 5 cm refers to the depth of the assimilated observations, how about 1m? Soil moisture states up to 1m depth are updated? Is it what is meant?

Page 10, line 18-19, similar to above (Page 10, line 17). Page 13, line 21, "spilt" → "split"

Page 16, line 14, "EnKF" → "ETKF" ? I believe authors imply "ETKF is a flavor of EnKF" here, but

still use of EnKF is confusing since EnKF is not used in this study.

We considered all the above minor comments and revised accordingly.

Craig H. Bishop, Brian J. Etherton, and Sharanya J. Majumdar (2001). Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, 129:3, 420-436.

Reichle R. H., Koster R. D. (2004). Bias reduction in short records of satellite soil moisture. *Geophysical Research Letters*, 31, L19501, pp 1-4.

Hendricks Franssen, H. J., and W. Kinzelbach (2008). Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem, *Water Resour. Res.*, 44, W09408, doi:10.1029/2007WR006505.

Ridler, M. E., Madsen, H., Stisen, S., Bircher, S., and Fensholt, R.: Assimilation of SMOS-derived soil moisture in a fully integrated hydrological and soil-vegetation-atmosphere transfer model in Western Denmark, *Water Resour. Res.*, 50, 8962-8981, Doi 10.1002/2014wr015392, 2014.

Ringgaard, R., Herbst, M., Friborg, T., Schelde, K., Thomsen, A. G., and Soegaard, H.: Energy Fluxes above Three Disparate Surfaces in a Temperate Mesoscale Coastal Catchment, *Vadose Zone J.*, 10, 54-66, 10.2136/vzj2009.0181, 2011.

Zhang, D., Madsen, H., Ridler, M. E., Refsgaard, J. C., and Jensen, K. H.: Impact of uncertainty description on assimilating hydraulic head in the MIKE SHE distributed hydrological model, *Adv. Water Resour.*, 86, Part B, 400-413, <http://dx.doi.org/10.1016/j.advwatres.2015.07.018>, 2015.