Response to Anonymous Referee #3

1. I consider that the results of the data assimilation are not conclusive. The model seems to lack the ability to reproduce the different type of available observations. The improvements of the simulations are marginally improved even for the assimilated variable (ET). However, there is not enough information in the manuscript to really have an idea on what were the settings used in the filter, and hence is hard to identify what was the main reason for the lack of improvement in the simulations.

To clarify and reinforce the value of the results, we performed more statistical analysis of the results by applying the Continuous Ranked Probability Score (CRPS; Hersbach, 2000), which measures the difference between the predicted and occurred cumulative distributions. This is specifically designed to assess probabilistic simulations. The CRPS intrinsically weighs errors by assigning a lower weight to the largest residuals (Schneider et al., 2020), thus accounting for observations that in other cases are defined as outliers. The CRPS is calculated, at a specific time step, for the *P*(x) cumulative distribution function given by the ensemble simulation for the variable of interest *x* (i.e. ET and WT levels) as follows:

$$CRPS_t = \int_{-\infty}^{+\infty} (P(x)_t - P_0(x)_t)^2 dx, \quad (1)$$

where P_0 is the observation distribution at the time step (t). As the observation (x_0) is usually a single value, P_0 is formulated as

$$P_0 = H(x - x_0), \quad (2)$$

with H being the Heaviside function

$$H(x) \begin{cases} 0 \rightarrow x < 0 \\ 1 \rightarrow x \geq 0 \end{cases}. \quad (3)$$

The expected value of zero is only possible in the case of a perfect deterministic forecast. The CRPS is usually calculated and averaged over a simulation period as follows:

$$\overline{CRPS} = \sum_{t=1}^{T} CRPS_t, \quad (4)$$

where T is the number of observations. Applying Equation 4 to WT levels and ET values of the two configurations yields the values presented in Table 1.

Table 1. $\overline{CRPS}$ calculated on ET and WT levels for the two configurations over the entire simulation period

	ET – Assimilation	ET – Open Loop	WT levels - Assimilation	WT levels – Open Loop
Configuration-1	0.557	0.570	0.134	0.161
Configuration-2	0.606	0.632	0.236	0.441

Figures 1 and 2 represent the evolution in time of the $CRPS_t$ for the WT levels of Configurations-1 and 2, respectively. These two figures show the temporal dynamics of the filter update effects and will be added to the results and discussion section of the manuscript.

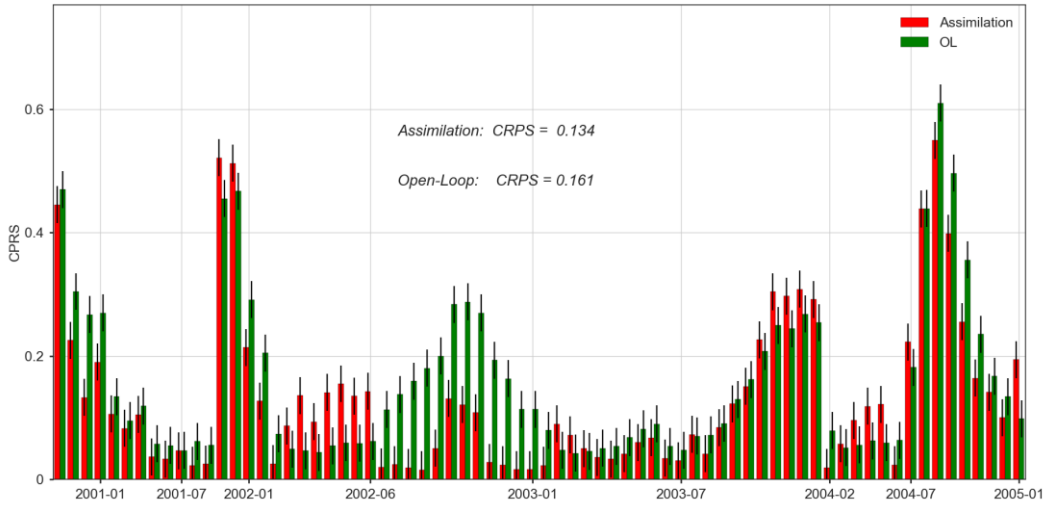

Figure 1- CRPS for WT levels of Configuration-1

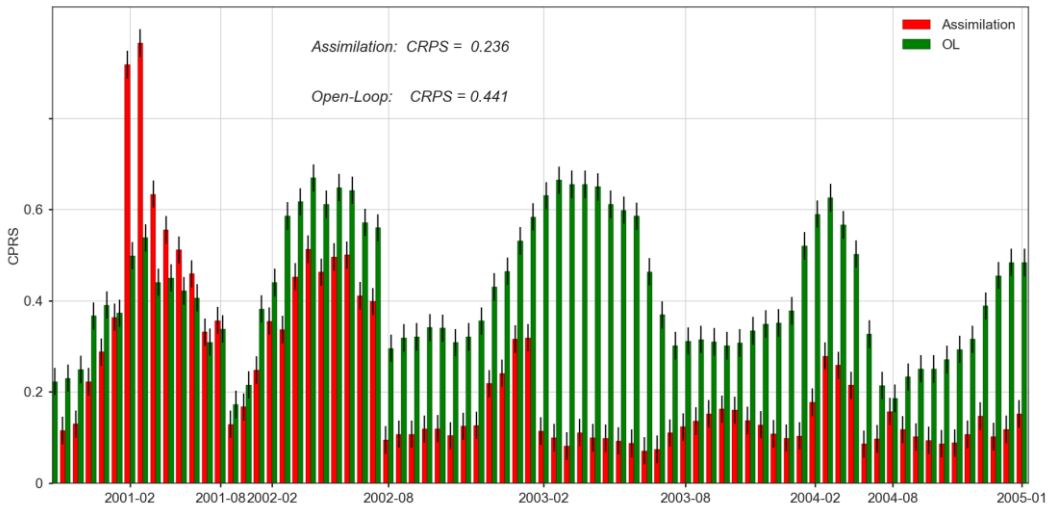

Figure 2- CRPS for WT levels of Configuration-2

For a better understanding of the filter application for this paper, we will expand Section 2.4 with a complete description of the EnKF, including the update equation.

Finally, to explain the effect of the filter on ET, it is to be noted that the assimilated value holds information about a period of 8 days before the application of the filter. The filter only modifies the states of the model, creating new initial conditions for the next simulation step. As actual ET is not a model state, the effects of the filter updates are shifted to the next time step. The main reason why ET do not show the same improvements, compared to those seen when assimilating quantities that are states of the model, is due to the evaluation metrics calculated on ET at the time step of the assimilation, i.e., when the effects of the updated states are not affecting the model output.

2. *In the same line, and without knowing the specific filter settings, it seems that the simulations are highly influenced by the boundary conditions. This is most likely a consequence of the model setup. Why did you choose to not further extend the model domain if it was somewhat hinted that the fixed boundary conditions are dominating the model behavior? What are the general run times of the coupled system? Given the number of cells in the model I suppose that MODFLOW does not take long, and I am not sure how compute-intensive are UnSAT and SWAP, however I am positive that modern computers would be able to handle numerical models with a better discretization.*

The runtime of the coupled system with the filter application for the simulation period of 5 years is about 4 hours; this is dominated by the UZMs and the filter I/O writing. The runtime of each MODFLOW instance usually takes less than a second using the MODFLOW solver CONJUGATE-GRADIENT SOLUTION PACKAGE, (VERSION 7, 5/2/2005).

The simple model domain is a result of numerical experiments showing that very similar water table dynamics could be obtained with both a fine (20 cells in the x-axis) and coarse (5 cells in the x-axis) model domain. Figure 1 shows the water table fluctuation, calculated with Configuration-1, in the central cell of the two domains. The run-time for the fine set-up is roughly 5 times larger than for the coarse set-up

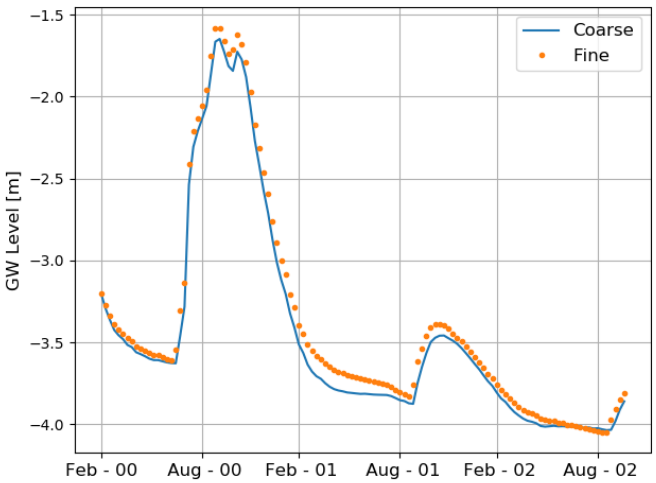

Figure 2- Sensitivity analysis of a fine and coarse model domains

In addition, with the intention of minimizing the influence of the boundary conditions (BC) on the simple domain conceptualization, we chose a location and an aquifer where the time variability of BC was low. The groundwater levels of the bore shown in Figure 2 were used to test our BC assumptions. The bore is located in the center of a forestry block, more than two kilometers from any groundwater extraction.

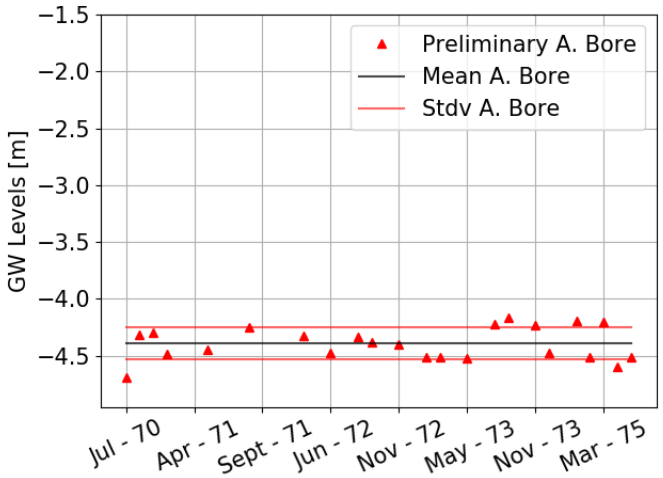

Figure 2- Groundwater levels of the bore used for the preliminary analysis on the BC.

In addition to the low variability of the BC, this location was also selected because of a previous study specifically looking at the impacts of forestry on groundwater (Benyon et al., 2006). Groundwater use by Pinus radiata forestry for this region (i.e., ET from a groundwater source), where the water table was less than 6 m below the surface (light to medium soil textured and not saline groundwater) was estimated by Benyon et al. (2006) to be on average 435 ML/yr (range 108–670 ML/yr) for a 1 km by 1 km fully forested cell. This value was seen to exceed by around 2 orders of magnitude the maximum groundwater extraction rate from a single bore for the region, albeit on a diffuse scale rather than a point scale. The localized effect of net recharge, and the water use by forestry, is recognized to be preponderant. For this reason, and from personal communication with water managers in the region, it is a valid assumption that the impact of localized recharge and transpiration from forestry on groundwater would far exceed the impacts of other temporal variations.

Additionally, this work represents a step toward the application of the EnKF to large-scale groundwater modelling, which, as stated in the manuscript, is our ultimate intention. However, because the spatial variability of the assimilation framework is not within the scope of this study, the model domain was not further extended.

3. *I do not agree with the authors that an ensemble of 32 members is appropriate for the study. The strong non-linearities of the system under study will be better addressed if a larger ensemble size is used. I do not think that compute time is the limitation, since the models being used are coarse and with very few cells. A larger ensemble size would also give the possibility to increase the ensemble spread, which I also consider to be too small in this work. The EnKF benefits strongly if the observations are within the ensemble spread. This might help to improve both the assimilation of ET and the additional model states updates.*

We understand the concern of the reviewer, but based on our experience with models of this level of complexity, we are convinced that an ensemble size of 32 is considered adequate. For example, for

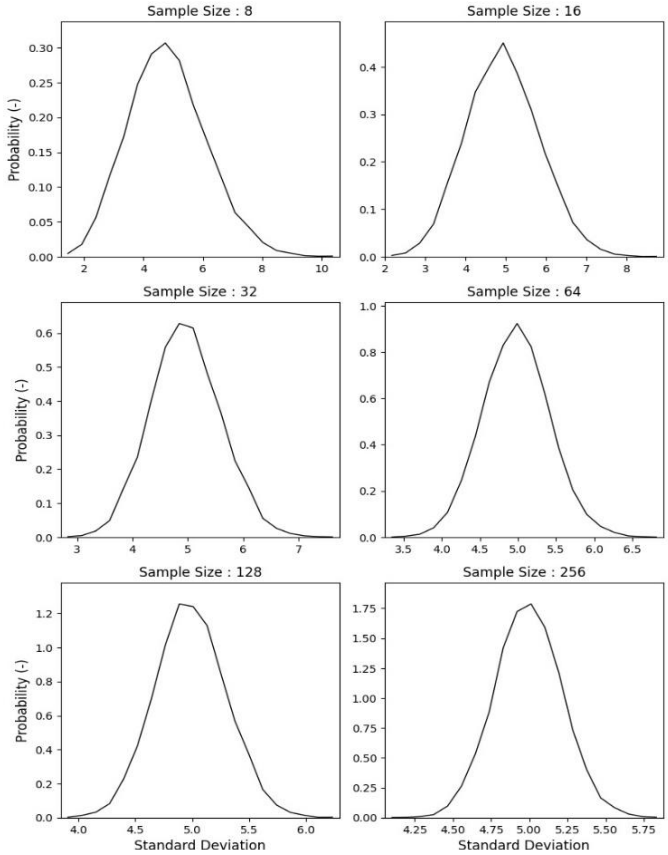

Figure 3- Distribution of 6 different sample sizes and their standard deviation. The figure shows how increasing the sample size does not increase the spread..

soil moisture assimilation, it has been shown that in land surface model data assimilation, an ensemble size of 12 (Yin et al., 2015) or even 10 (Kumar et al., 2008) is sufficient.
Furthermore, we show that assuming that a larger ensemble size increases the spread is not necessarily true. In Figure 3 we report a simple example where we draw 8, 16, 32, 64, 128, 256 random numbers, respectively and calculate the mean and std of 10,000 repetitions of an analysis of Gaussian numbers with mean 20 and standard deviation 5. As the ensemble size increases the spread in the standard deviations decreases. In conclusion, when considering the computational time versus the ensemble spread, this should make the case that there is no clear benefit by applying ensembles with a population greater than 32 members.

We also calculated the ensemble skill (ensk), ensemble spread (ensp), and mean squared error (mse) (Talagrand et al. 1997; De Lannoy et al., 2006) for ET values and applied them to verify the ensemble as in Gelsinari et al. (2020). We will clarify this in the revised version of the manuscript.

4. *I suggest to elaborate in the theory of the filter and the settings applied during the assimilation, I recognize the authors want to avoid a strong overlap with the related publication, however, important information is missing in the text and it does not stand by itself as it is presented.*

To clarify this manuscript, we will elaborate on the filter theory and set-up and expand section 2.4.1 as suggested by the reviewers.

5. *Have the authors considered using the ensemble Kalman filter for updating also model parameters? This is mentioned throughout the text, e.g. line 180 kind of hints that this was actually tried out, however is not clear. I consider that a large enough ensemble, and a large enough spread (which can be produce by perturbing the model parameters) would allow the filter to improve the model simulations not only by updating the model states, but also by calibrating the model parameters. If that is the case this would contradict the first conclusion of the work, which is that the calibration using a multi-objective function is needed prior to data assimilation*.

The statement of line 180 refers to the calibration phase of the models. Model parameters were perturbed a single time and kept constant throughout the entire simulation; this is common practice in data assimilation with the EnKF. To clarify, we will state this more clearly in section 2.4.1.

Consistent with the basics of the Kalman filter, we are updating the state variables and not the parameter values. Please note that our state vector, for one model grid, already contains ≥ 300 entries. Although interesting, updating the parameters in addition to the state variables is outside the scope of this study.

6. *I thought only ET was actually assimilated, and used to update all the model estates. However, in line 322 it is said ''the assimilation for actual ET, WT levels, and SM contents of the upper and lower soil layers''. Could you please clarify?*

We thank the reviewer for noting this. ET is the only quantity assimilated. We modified the sentence into "Table 3 summarizes the RMSE and *r* results applied to actual ET, WT levels, and SM contents (upper and lower soil layers) and compared the improvements between the open-loop and the assimilation"

Editorial Comments

1. *I suggest the authors to undergo a thorough proofread of the paper. While I am not a native speaker, and would argue that the work is not necessarily wrong from the language perspective, it is sometimes*
hard to read.

We thank the reviewer for pointing this out. The manuscript will be revised to improve fluency and readability.

2. *I recommend subtle modifications to the text to make it easier to read. I apologize the emphasis on this but I hope the authors can understand what I mean when I say the work is hard to read. In the following, I list a couple of examples in which I slightly modified the text. This might help the authors to understand my viewpoint:*

Line 91. The Morton equation (Donohue et al., 2010) and the Budyko curve (Donohue et al., 2007) classify the area as dominated by ET or water-limited (Jackson et al., 2009; Benyon et al., 2006).

We agree with this comment and we will reformulate the sentence as suggested.

Another example l104. We used remotely sensed data of actual ET from the CSIRO MODIS reflectance-based scaling evapotranspiration (CMRSET) algorithm (Guerschman et al., 2009).

In the manuscript, we prefer not to use the first-person form. To improve the fluency of this sentence we rephrased it to: "The actual ET data derives from the remotely sensed CSIRO MODIS reflectance-based scaling evapotranspiration (CMRSET) algorithm (Guerschman et al., 2009)."

Line 110. We tested two different configurations of coupled groundwater-unsaturated zone models. Figure 2 describes the UZMs conceptualization and the groundwater model coupling. In the following we detail in the description of the models used in this work as well as the coupling framework.

As per the previous point, we prefer not to use the first-person form. To improve the fluency of this sentence we rephrased it to "The tests presented in this study used two different configurations of coupled groundwater-unsaturated zone models, which are depicted in Figure 2. The following sections describes the models used in this work as well as the coupling framework."