Simone Gelsinari, Phd
Room 106, 18 Alliance Lane
Monash University, Clayton Campus
Clayton, VIC 3800, Australia
Email: simone.gelsinari@gmail.com

17th January 2021

Re: hess-2020-252 (Editor: Harrie-Jan Hendricks Franssen) - Editor Decision

Dear Prof. Hendricks Franssen,

Please find attached the revised manuscript **"Unsaturated zone model complexity for the assimilation of evapotranspiration rates in groundwater modeling"**. On behalf of all the authors, I thank the editorial team for the detailed and constructive comments, which have significantly improved our original manuscript. We have thoroughly revised the manuscript by adding sections and performing more analysis, with the goal of addressing all the questions and concerns raised by the Referees.

Thank you for taking the time to help us to improve our manuscript.

We look forward to hearing from you.

Sincerely,

Simone

Response to Referee comments

We thank the editorial team for the thoughtful and constructive review. In this document we address each concern that was raised. We have quoted the remarks made by the Editor and Referee in boldface, and have listed our replies in normal font. Relevant edited lines are hereafter listed in red, italic font. Line numbers refer to the final non-annotated manuscript.

Editor

Your manuscript "Unsaturated zone model complexity for the assimilation of evapotranspiration rates in groundwater modeling" has been subjected now to review by three reviewers and I apologize for the delay. Two of the reviewers recommend major revision and one reviewer recommends rejection. The main points to be handled are, in my opinion:

(i) lack of improvement by DA and the significance of the results (addressed by all reviewers). The significance of the results should be quantified.

The point about the significance of results led to new analyses, which we are confident have improved the paper. We agree that knowing the confidence intervals of the RMSE or other quantities would be useful to quantify the significance of the results, as it was suggested by one Referee. Unfortunately, analytical solutions to calculate only exist for the unbiased RMSE and the bias (Gruber et al., 2020). These quantities are not the focus of this study. It can also be said that the distribution of the square of the errors, which are the components of the RMSE, presents outliers, and thus can cause skewness, for which a chi-squared statistic to form a confidence interval is not accurate. This means a correction would be needed to account for the bias caused by the skewness and a more robust measure, such as a probability rank score measure, is more appropriate. This motivates the further analysis to consider an additional measure of Continuous Ranked Probability Score (CRPS) to determine a more representative error. The CRPS uses a cumulative probability measuring the difference between the two models and obtaining probabilistically an error similar to the RMSE. As we deal with stochastic models, we adopted the CRPS to reinforce the significance of the results, alongside with the RMSE that is widely accepted and understood in the modeling community. By applying the CPRS, which calculates the average error based on the probability distribution at each time step, a more representative and robust quantitative measure for the errors is provided.

Using both RMSE and CRPS offers a better quantification of the data assimilation improvements, by providing detailed and more accurate quantifiable results. The RMSE is a reliable relative quantifiable measure for the overall average error; however, it excessively captures error from outliers and is not as robust as CRPS. From this approach, the comparison between the values of the CPRS, as a relative percentage decrease, is considered a sufficient quantification of the improvement. By further analyzing results with this metric, we reduced the previously claimed benefits of the data assimilation on the soil water content.

We further discuss this concept in the detail answers to A/Prof Manuela Girotto and Referee #3.

To explain the apparent lack of improvement on the quantity assimilated (actual evapotranspiration - AET), 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 AET is not a model state, the effects of the filter updates are shifted to the next time step. The main reason why AET does not show the same improvements, compared to those seen when assimilating quantities that are states of the model, is due to the evaluation metrics. These are calculated on AET at the time step of the assimilation, i.e., when the effects of the updated states are not yet affecting the model output.

We report this discussion in the detail answer to Referee #3 as well.

Finally, in the manuscript (Line 414-416) we state:

"*However, these are non-trivial results as the data assimilation, through the EnKF, is designed to only improve the model states. Therefore, the observed reduction in AET errors suggests that the model states (i.e.WT, SM) updated by the filter are contributing to better modeling of other hydrological quantities (e.g. AET)* "

(ii) the conceptual groundwater model including the imposed boundary conditions with lack of transient dynamics. A realistic groundwater system would show transient variations related to varying lateral fluxes, pumping and meteorology. Meteorology is considered now, but not for the boundary conditions. These additional sources of uncertainty could affect the impact of assimilation of evapotranspiration data. This should be assessed.

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 a bore (shown in Figure 5 at the end of this document) 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. 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 impacts of localized recharge and transpiration from forestry on groundwater would far exceed the impacts of other temporal variations.

We report this discussion also in the detail answer to Referee #1 and Referee #3.

The justification for the model domain conceptualization and BC was also added to the manuscript at lines 186 to 196:

"The domain discretisation was chosen as a result of a sensitivity analysis conducted on a range of model domains varying from fine (1 x 20 cells) to coarse (1 x 5 cells). The boundary cells were set to a constant head obtained via calibration (i.e. 3.5 m below the surface). The location chosen allows the model configuration to be kept simple imposing the boundary conditions of the saturated model as constant head. This is due to the site being in the centre of a forestry block, more than two kilometres from any groundwater extraction. For this region, where WT are 6 m deep or shallower, it has been shown that forestry transpiration from groundwater is around 2 orders of magnitude (i.e. 435 ML/yr for a 1 km by 1 km fully forested cell) larger than the maximum groundwater extraction rate from a single bore. To further reinforce the selection of constant head boundary conditions, an analysis of the WT fluctuations was conducted on bores in proximity of the study area but outside of the forest, showing a mean of the WT level of 4.4 m below the surface with standard deviation equal to 0.12 m. This supports the assumption that the WT table fluctuation at the observed site is highly dependent on the local net recharge."

(iii) the small ensemble size. Typically, in groundwater DA studies larger ensemble sizes are used. It should be shown that 32 is enough, or the study repeated with larger ensemble sizes.

We understand the concern of the referee, 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 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.

Following the Editor suggestion, we performed a repeated study applying 32 and 64 members of the ensemble respectively. Figure 7 of this document shows the mean of the two ensembles, where the difference between "Ens_32" and "Ens_64" is hardly distinguishable. As expected, the simulation time for "Ens_64" is about twice what simulation "Ens_32" requires. Hence, we believe that an ensemble with 32 members is enough for this experiment.

We report and expand this discussion in the detailed answer to Referee #3.

(iv) the coarse representation of the groundwater system. It should be clarified why this would not affect the study outcomes.

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 4 of this document, 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 the coarse set-up.

(v) in almost all groundwater DA studies also parameters are updated as these are the main source of uncertainty. This could be considered; assimilation of ET data could have more impact and result in more improvement if also parameters are updated. If parameters are also updated, a larger ensemble size will be needed.

We understand that in many groundwater studies model parameters are updated along with the state variables, but in system theory, a parameter is defined as a factor that remains unchanged. The Kalman filter is derived to estimate the state of a system, and we already have ≥ 300 entries, for one model grid, in the state vector. Adding the parameters would make this even larger, and would make the data assimilation system more complicated. For these reasons we decided to focus on estimating the state of the system and not the parameters.

(vi) more details should be provided in parts of the manuscript as suggested by the reviewers.

We have addressed this by providing more details in sections:

2.2.1 - UnSAT – UZM (Line 121 – 142)
2.3 - Model domain and calibration (Line 187 to 199)
2.4 - Assimilation (Line 247 to 279)
2.4.1 - Ensemble Generation (Line 284-287; 290-295)
2.5 - Verification Skills (Section rewritten)
3.2 - Ensemble Results (401 – 414; 429 – 440; 443 – 448; 454 – 456; 459-461)
4.0 - Conclusion (470 – 474; 500-512)

For the sake of readability of this document, we do not report here all these additions, which can be found in the annotated (track-change) version of the manuscript.

Response to A/Prof Manuela Girotto

1- I think the paper is well written and of interest to the HEES readership. My main concern is related to the robustness of the main conclusion related to the assimilation part. I am a bit doubtful about the significance of the authors results. Yes, the results show evidence that the assimilation of ET improves WT dynamics, but the authors should test for the significance of these results. In fact, the improvements reported in table 2 and 3 seem very marginal and small. I would like to see confidence intervals added to the calculated RMSE and r so that the authors can conclude whether their approach lead to significant improvements or not.

We received similar comments during the review of Gelsinari et al. (2020). On that occasion, we argued that applying significance tests to model simulations with an artificially large sample size will lead to very high test power. In other words, the test would have been set up to conclude that the simulations are different, after which the test concludes that they are different. This was confirmed by applying the student t-test to the correlation metric of our results. Instead of discussing these tests in the manuscript, we provide an additional analysis, which we are confident has improved the paper.

We agree that knowing the confidence intervals of the RMSE or other quantities would be useful to quantify the significance of the results. Unfortunately, analytical formulas to calculate confidence intervals only exist for the unbiased RMSE and the bias (Gruber et al., 2020, which are not the focus of this study. It can also be argued that the distribution of the square of the errors, which are the components to the RMSE, might present outliers, and thus can cause skewness, to which a chi-squared statistic to form a confidence interval is not accurate. This means that a correction would be needed to account for the bias caused by the skewness and a more robust measure, such as a probability rank score measure, is more appropriate. This motivates the use of the Continuous Ranked Probability Score (CRPS) to determine a more representative error. The CRPS uses a cumulative probability measuring the difference between the two models and obtaining probabilistically an error similar to the RMSE. As we deal with stochastic models, we adopted the CRPS to reinforce the significance of the results, alongside with the RMSE that is widely accepted and understood in the modeling community. By applying the CPRS, which calculates the average error based on the probability distribution at each time step, a more representative and robust quantitative measure for the errors is provided.

The use of both RMSE and CRPS offers a better quantification of the data assimilation improvements, by providing detailed and more accurate quantifiable results. This additional analysis explains that the RMSE is an appropriate measure for the overall average error, but it excessively captures error from outliers and is not as robust as CRPS. The comparison between the values of the CPRS, as a relative percentage decrease, is considered a sufficient quantification of the improvement associated with the data assimilation. By further analyzing results with this metric, we reduced the previously claimed benefits of the data assimilation on the soil water content.

The CRPS is calculated, at a specific time step, for the cumulative distribution function P(x) given by the ensemble simulation for the variable of interest x (i.e. AET 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

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

where H is the Heaviside function defined as

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

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:

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

where T is the number of observations.

Applying the Equation 4 to WT levels and AET values of the two configurations yields the values presented in Table 1. These results have been added to Table 3 in the manuscript.

Table 1. $\overline{CRPS}$ calculated on AET, WT levels and shallow and deep soil moisture for the two configurations over the entire simulation period

	AET – Assimilation	AET – Open Loop	WT levels - Assimilation	WT levels – Open Loop	SMS - Ass	SMS - OL	SMD - Ass	SMD - OL
Configuration-1	0.452	0.564	0.134	0.161	0.206	0.204	0.077	0.078
Configuration-2	0.606	0.632	0.236	0.441	0.036	0.037	0.013	0.013

In this document, 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 have been added to the results and discussion section of the manuscript (Fig. 8 in the manuscript).

2- Please reduce the strengths of statements like those in lines 373-376 or lines 402-405

Line 373 to 376 has been modified into (Line 467-469):

"In addition, albeit marginally, the filter improves the unsaturated zone state variables regardless of the manner in which the SM content is calculated (volumetric SM or pressure head)."

Line 402 to 405 were modified into (Line 494-499):

"The updating of the entire soil column is an advantage of the assimilation of remotely sensed AET over satellite SM retrievals. AET rates express the moisture status of the entire root zone. Thus, assimilating AET has the potential to overcome the SM assimilation tendency to produce stronger updates in the most superficial part of the soil because of the reduced correlation between the upper and lower SM contents. This experiment only showed the feasibility of the proposed assimilation framework to improve SM contents. Preliminary results indicated that Configuration-2 is preferred to conduct more experiments in order to quantify the significance of the SM updates."

3- Also, if I understand correctly, the RMSE and r statistics for ET are calculated against the same data that are assimilated, correct? If so, I would have excepted the verification statistics of the assimilation to improve much more, but the improvements are marginal. Can the author comment on this?

The RMSE and r metrics for AET are calculated against CMRSET. The observation assimilated is coming from the same dataset, to which an observation error is added. 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 at the end of this period, creating new initial conditions for the next simulation step. As AET is not a model state, the effects of the filter updates are shifted to the next time step. The main reason why AET does not show the same improvements, compared to those seen when assimilating quantities that are states of the model, is due to the metrics calculation which is performed on AET outputs at the time of the assimilation.

4- Line 94: "The area was originally planted". What was planted? The area or trees? Please reword.

The sentence was reworded as (Line 96-97)

"The trees were originally planted in July 1996 with a density of 1225 trees/ha and there was no thinning of the plantation during the observations."

5- What is the influence of the sea level to the groundwater level? (Figure 1 indicates that the test domain is located near the coast)

The test location is about 40 km from the coast, with an elevation of about 54 mAHD (Australian Height Datum). For these reasons, we did not explore the effect of sea levels on groundwater.

6- Line 138: Add reference to the section where you explain the coupling.

We added *"MODFLOW 2005 (Harbaugh, 2005)."*

7- What do you mean by "ET WT link"? Please explain.

In the introduction and throughout the paper we often mention the relation between AET and WT.

In particular, at Line 16-22 it is written:

"Actual evapotranspiration (AET) and groundwater recharge are two related major components of the water cycle. This is because AET is a function of the soil water content within the root zone, as the root water uptake is distributed along the entire root system (Grinevskii, 2011; Neumann and Cardon, 2012). Improving AET estimates, by means of a detailed modeling of the soil water transport, can enhance the simulation of recharge and WT dynamics. This is particularly important when the WT is within the reach of the roots, as it is common in Australian semi-arid catchments (Banks et al., 2011), because the root water uptake from groundwater and the capillary fringe largely contribute to AET (Mensforth et al., 1994; Orellana et al. 2012)."

or Line 78-79:

"…yield different AET estimates, producing distinct recharge values and, in turn, diverse dynamics of the WT"

Improving the simulation of AET leads to the improvement of net-recharge estimates. Thus, because net-recharge is the quantity that drives the WT dynamics, this creates a link between AET and WT.

We also added to the conclusions (Line 510-513):

"This study explored the use of AET information for constraining unobservable estimates (i.e. net recharge) calculated by hydrogeological models. Improving the AET fluxes led to better recharge estimates. Thus, as recharge is a key quantity driving the WT dynamics, the link between AET and WT in the model is strengthened."

8- Further, how do we see that the "link is reproduced" in figure 4. I have a hard time to see a clear relationship between ET and WT in figure 4?

The description of Figure 4 (of the manuscript) Line (350-352) was modified, and it now reads:

"With the calibration technique proposed in Section 2.3, the coupled models were able to simultaneously reproduce the dynamics of both the WT and AET for the two configurations."

9- Table 1 (and in text) is perturbation fraction referred to the coefficient of variation? If so, I'd replace it with coefficient of variation which is more commonly used term in statistics?

The perturbation fraction was calculated as the ratio of the standard deviation of the probability function to the mean. Thus, it is the equivalent of the coefficient of variation. We have replaced this with the coefficient of variation in the table caption and in the text.

10- Section 2.4.1. Table 1 reports some perturbation numbers, but the reader is referred to Gelsinari et al., 2020 for the ensemble generation. I recommend adding a list/table of all perturbed parameters/meteorological inputs/prognostic states to this article too.

We expanded Section 2.4.1 which mentions the parameters perturbed and forcing inputs perturbed. Furthermore, Line 386 – 389 reads:

"For the meteorological data, the best candidates are obtained by perturbing the input with a random number sampled from a Gaussian distribution having a standard deviation proportional to the value of the forcing inputs (i.e. 50% for Configuration-1 and 10% for Configuration-2). For parameters, the last column of Table 1 lists the coefficient of variation. Additionally, for Configuration-2, S_y has a lower limit of 0.1 to preserve numerical stability of the coupled models. "

11- Line 200: I think there are other algorithms that work better in highly non linear system (e.g. particle filters) so I'd remove this as a reason for choosing the EnKF.

The sentence was modified to read as (Line 221):

"The EnKF (Evensen, 1994) was used because of its reduced computational burden when dealing with highly non-linear systems."

12- Line 230. Please add the update equation to this article so that the reader does not have to go back to Gelsinari et al., 2020 to see it.

The section has undergone a major revision with an elaboration on the filter theory and set-up. Among other additions, the updated equation is now displayed at line 273 (Eq. 17).

13- Line 231: What do you mean "limited". Please reword and clarify in the article.

By "limited" we intended the constraints applied to the value updated by the filter. This is further specified for SM at line 74. We changed the word "*limited*" with *"constrained".*

14- Line 249: What ensemble verification skills do you use? I think it is important to expand this part, especially since you refer to it later in the article (line 305)

We agree that more details were needed on this aspect. We calculated the ensemble skill (ensk), ensemble spread (ensp), and mean squared error (mse) (Talagrand et al. 1997; De Lannoy et al., 2006) for the ET values and applied them to verify the ensemble as in Gelsinari et al. (2020). This was added to section 2.4.1 and reads: (Line 284 – 288)

"*The average over the verification period of the ratios between ensemble skill and ensemble spread, which should tend to 1, and between ensemble skill and mean squared error, which should tend to* $\frac{\sqrt{(M+1)}}{2M}$ *(Talagrand et al., 1997; De Lannoy et al., 2006), were calculated on the modeled AET values. First, a simple perturbation of forcing inputs, by adding a random number sampled from Gaussian distributions with different standard deviations, as performed by Gelsinari et al. (2020), was tested."*

(Line 294 – 298)

The Talagrand et al. (1997) verification skills were applied to the ensembles generated with the aforementioned approach, and the most adequate ensembles for the two configurations were retained. The scores obtained for the two ratios were comparable to others found in the literature (e.g. De Lannoy et al. (2006); Pauwels and De Lannoy (2009); Gelsinari et al. (2020)). These ensembles are defined as the open loop, which represents the "prior" distribution. After applying the filter, the resulting distribution is called the assimilation run and represents the 'posterior'."

15- Line 257: . . . assimilation results and to the respective . . ..

This section has been entirely rewritten and the remark was considered.

16- Line 271: config. 1 temporal dynamics is not always lower. What happen in 2005?

We agree that is not always lower. In 2005, there was an intense precipitation event that led to saturation of part of the unsaturated zone in Configuration-1, which, in turn, produced an elevated value of recharge. This suddenly increased WT levels.

17- Line 278: indicate the blurred area in the figure too so that the reader knows what you are referring to.

As part of the general readability improvement, this line has been removed. We do not refer to the blurred area anymore to avoid confusion.

18- Line 288-289: replace seasons with months.

The term was replaced by

"*Southern hemisphere later summer/early autumn*" (Line 370-371).

19- Figure 6, 7. Please darken the ensemble replicates. I can barely see them on my screen.

These figures were regenerated to improve the contrast.

20- Line 318: add figure reference: e.g.: "(see panel b in Figure 7)" or "(see Figure 7b)".

We agree with this comment and we added the panel references as suggested.

21- Line 327-328: Can you be more explicit in explaining why the reduction in ET errors suggests improved state variables? From your table 3, some of these states degrade even if ET improves.

We rephrased this paragraph to improve the cohesion with the previous one. It reads (Lines 410-413):

"*However, these are non-trivial results as the data assimilation, through the EnKF, is designed to improve the model states. Therefore, the observed reduction in AET errors suggests that the model states (i.e. WT, SM) updated by the filter are contributing to better modeling of other hydrological quantities (e.g. AET)"*

22- Figure 8 and Figure 9: what is the cloud of points in the open loop and assimilation? I assume these are all the ensemble member at the given time step. If so, why do they have a x-axis dimension on the bottom plot?

As part of the paper revision, due to the introduction of the new CRPS figures which convey similar information in a clearer manner, we removed the mentioned figures.

Response to Anonymous Referee #1

Because of the exchange of posts during the public discussion phase of the manuscript and the long series of points provided by the Referee, we focus the reply to the main concerns that were raised. We summarized the response to Referee #1 to two main points: 1) the Ensemble Kalman filter applied to forward modelling and 2) the assumptions behind the choice of boundary conditions.

Ensemble Kalman filter applied to forward modelling

We have no choice but to explain how nonlinear operation systems work. Enough papers have been written on this (a few listed at the end of this item), but clearly the message is not getting through. We will not explain the variables here, assuming the referee knows what they are, but we cannot explain how nonlinear operation systems work without the equations. We have to start from the state update equation:

$$x_k^{i,a} = x_k^{i,f} + \boldsymbol{K_k}\,[y_k - y_{k(i,f)} + v_k^i] \qquad (1)$$

The gain is calculated as:

$$\mathbf{K}_k = \frac{\mathbf{PH}^{\mathrm{T}}}{\mathbf{HP}\boldsymbol{H}^{\mathrm{T}}+\mathbf{R}_k} \qquad (2)$$

The two matrix products in the Kalman gain are:

$$\begin{cases} \mathbf{PH}^T = \frac{1}{M-1}\boldsymbol{X}_k^f \boldsymbol{Y}_k^{f^T} \\ \mathbf{HPH}^T = \frac{1}{M-1}\boldsymbol{Y}_k^f \boldsymbol{Y}_k^{f^T} \end{cases} \quad (3)$$

The state and observation-simulation deviation matrices are written as:

$$\begin{cases} \mathbf{X}_k^f = \left[x_k^{1,f} - \bar{x}_k^f \quad x_k^{2,f} - \bar{x}_k^f \quad \dots \quad x_k^{M,f} - \bar{x}_k^f \right] \\ \mathbf{Y}_k^f = \left[y_k^{1,f} - \bar{y}_k^f \quad y_k^{2,f} - \bar{y}_k^f \quad \dots \quad y_k^{M,f} - \bar{y}_k^f \right] \end{cases} \quad (4)$$

Pauwels and De Lannoy (2009) provide an in-depth analysis of what this means, which we will summarize here. Assuming that x is one state variable (e.g. catchment wetness) and y is one observation (e.g. streamflow), Eq. 2 becomes:

$$\mathbf{K}_k = \frac{\sigma_{xy}}{\sigma_y^2 + v_k^i}\,. \qquad (5)$$

If the observation error is zero. The Kalman gain becomes:

$$\mathbf{K}_k = \frac{\sigma_{xy}}{\sigma_y^2}\,. \qquad (6)$$

This is a linear regression, across all ensemble members, between the observation and the state. Thus, if the model predicts different streamflow than the observation, the state update becomes:

$$x_k^{i,a} = x_k^{i,f} + \frac{\sigma_{xy}}{\sigma_y^2} \left[y_k - y_{k(i,f)} \right]. \qquad (7)$$

In other words, the gain maps the difference in observation space to state space. If the model underestimates the streamflow, Equation 7 will increase the modeled catchment wetness (if the covariance between catchment wetness and streamflow is positive, which it usually is). If the model overestimates streamflow, Eq. 7 will reduce the modeled wetness. Therefore, streamflow becomes a proxy for the wetness, because the catchment wetness is updated without observing it. If the observation error is nonzero, the update will be reduced. For multiple observations and state variables, similar reasoning can be made.

This is the way assimilation of brightness temperatures or backscatter values or streamflow into hydrologic models works. The observation system:

$$y_k = h(x_k) + v_k \qquad (8)$$

in this case is a radiative transfer model (for brightness temperature assimilation), a backscatter model (for backscatter data assimilation), or the hydrologic model (for streamflow assimilation, as just explained). Thus the model does not have to directly predict the variable it assimilates. They can be calculated through the observation system (Equation 8).

What one cannot do, as is unfortunately done frequently in streamflow assimilation papers (and also suggested by the referee), is to enter these values in the state vector, because they are not state variables. One of the two prerequisites of a correct system description is that the system must be controllable. This means that an external input needs to be able to move the internal state of a system from any initial state to any other final state in a finite time interval. If discharge is in the state vector, at the beginning of the time step, one can assign it any value, regardless of the forcing, it will have no impact on the soil moisture content and streamflow at the end of the time step. We know no hydrologic models for which streamflow is an initial condition. Another prerequisite is that the system must be observable, meaning that all state variables must be able to be inferred from the observations. The two prerequisites are very closely linked to each other.

Because flood forecasting models (especially the ones that are used for streamflow assimilation) usually have a very limited amount of state variables, when streamflow is entered into the state vector, problems with these two prerequisites usually do not occur, even though they are not met. But in a more complicated system such as ours, working with a system that is not correctly described can lead to significant problems, most likely excessive state updates.

In the case of assimilating backscatter values, which the referee correctly points out no hydrologic models directly compute, how would this approach work, using an observation matrix with ones and zeros? How can a model enter a variable it doesn't compute in the state vector? Same question for assimilating brightness temperatures. The reason for the confusion when assimilating discharge is that, in this case, the model does directly compute the variable that is assimilated. And the poor system description does not lead to problems because of the limited amount of state variables.

Reading the second review, we believe that the referee assumes that we are updating the parameters (which we do not and did not state), or that we disturb them at every time step (which we also do not). This is not clear. The pdf we infer is simply $P(x_k | x_{k-1}, y_k)$. Disturbing initial parameter values is common practice in

data assimilation with the EnKF. There is no parameter space in the Kalman filter equations, only state and observation space. This makes the last slide in the second review very confusing.

The assumptions behind the choice of boundary conditions.

In order to justify the assumptions behind the choice of the fixed boundary conditions (BC), we compared the water table (WT) dynamics of the bore used by the Referee to another bore, which is more representative of the conditions we have in the experiment. To minimize the BC influence on the simple domain conceptualization, we have purposely chosen a location and an aquifer where the time variability of BC was low. Figure 3 compares the WT dynamics of the bore selected by the Referee (Blue), which is monitoring the levels of a different aquifer, and the bore we used to test our assumptions (Red), which is located in the center of a forestry block, more than two kilometers from any groundwater extraction. Constrained by the availability of the observations at the Referee's bore, the comparative analysis of the WT dynamics is only possible during the first part of the 70s, but a similar trend has been reported for other bores with more recent observations.

As the WT dynamics at the Referee's bore is driven by conditions that are not representative of our simulation a large "seasonal" fluctuation of the WT levels (in the order of 2 meters) is observable.
It is worth specifying that the aquifer monitored by the incorrect bore is classified as "Qpcb", part of the Pleistocene Bridgewater Formation. This aquifer is described as formed by Aeolianitic calcarenite, a partially calcretised ancient dunal system. Further description of the aquifer characteristics reports that it contains minor quantities of groundwater, unconfined to semiconfined aquifers, often containing small fresh lenses that provide small stock/domestic supplies. Finally, it is to be noted that this incorrect bore is close to anthropogenic activity.

In Figure 3, the WT dynamics shown in red refers to a bore screened in correspondence to the aquifer reported in our study. The aquifer observed at this bore is classified as "Thgg", consisting of grey marl with coarse bioclastic presenting frequent chert band. In the figure, the substantial differences in the WT fluctuations of the two bores are also indicated by the solid lines that mark the interval within one standard deviation (σ) from the means. For a large part of the simulation, the seasonal fluctuation of the correct bore is in the order of 10-30 cm, making it hard to distinguish regional patterns from the effects of localized net-recharge. Therefore, we decided to maintain the BC head elevation constant over the simulated period.

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.**

As reported in the response to the Editor and the detailed response to A/Prof Manuela Girotto, the remark made by the referee led to new analyses, which we are confident have improved the paper. 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 observed 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 use of both RMSE and CRPS offers a better quantification of the data assimilation improvements, by providing detailed and more accurate quantifiable results. This additional analysis explains that the RMSE is an appropriate measure for the overall average error, but it excessively captures error from outliers and is not as robust as CRPS. The comparison between the values of the CPRS, as a relative percentage decrease, is considered a sufficient quantification of the improvement associated with the data assimilation. By further analyzing results with this metric, we reduced the previously claimed benefits of the data assimilation on the soil water content.

The CRPS is calculated, at a specific time step, for the cumulative distribution function P(x) given by the ensemble simulation for the variable of interest x (i.e. AET 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

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

where H is the Heaviside function defined as

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

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:

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

where T is the number of observations.

Applying the Equation 4 to WT levels and AET values of the two configurations yields the values presented in Table 1. These results have been added to Table 3 in the manuscript.

Table 1. $\overline{CRPS}$ calculated on AET, WT levels and shallow and deep soil moisture for the two configurations over the entire simulation period

	AET – Assimilation	AET – Open Loop	WT levels - Assimilation	WT levels – Open Loop	SMS - Ass	SMS - OL	SMD - Ass	SMD - OL
Configuration-1	0.452	0.564	0.134	0.161	0.206	0.204	0.077	0.078
Configuration-2	0.606	0.632	0.236	0.441	0.036	0.037	0.013	0.013

In this document, 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 have been added to the results and discussion section of the manuscript (Fig. 8 in the manuscript).

For a better understanding of the filter application for this paper, we have expanded Section 2.4 with a complete description of the EnKF, including the update equation (Line 245 - 270). Finally, to explain the effect of the filter on the AET quantity 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 AET is not a model state, the effects of the filter updates are shifted to the next time step. The main reason why AET does 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 AET at the time step of the assimilation. Hence, 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 4 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

As reported in the Editor and Referee#1 responses, 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 a bore (shown in Figure 5 at the end of this document) 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. 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.

We report this discussion also in the detail answer to Referee #1 and Referee #3.

The justification for the model domain conceptualization and BC was also added to the manuscript at lines 186 to 196:

"The domain discretisation was chosen as a result of a sensitivity analysis conducted on a range of model domains varying from fine (1 x 20 cells) to coarse (1 x 5 cells). The boundary cells were set to a constant head obtained via calibration (i.e. 3.5 m below the surface). The location chosen allows the model configuration to be kept simple imposing the boundary conditions of the saturated model as constant head. This is due to the site being in the centre of a forestry block, more than two kilometres from any groundwater extraction. For this region, where WT are 6 m deep or shallower, it has been shown that forestry transpiration from groundwater is around 2 orders of magnitude (i.e. 435 ML/yr for a 1 km by 1 km fully forested cell) larger than the maximum groundwater extraction rate from a single bore. To further reinforce the selection of constant head boundary conditions, an analysis of the WT fluctuations was conducted on bores in proximity of the study area but outside of the forest, showing a mean of the WT level of 4.4 m below the surface with standard deviation equal to 0.12 m. This supports the assumption that the WT table fluctuation at the observed site is highly dependent on the local net recharge."

We would also like to add that this work represents a step toward the application of the EnKF to large-scale groundwater modelling and we stated that as our ultimate intention. However, at this stage, the spatial variability of the assimilation framework was not fitting into the scope of the study, thus 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 Referee, 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 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 do not believe that a larger ensemble size necessarily increases the spread. In Figure 6 we report an example where we draw 8, 16, 32, 64, 128, 256 random numbers, 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.

To further address this concern, Following the Editor suggestion, we performed a repeated study applying 32 and 64 members of the ensemble respectively. Figure 7 of this document shows the mean of the two ensembles, where the difference between "Ens_32" and "Ens_64" is hardly distinguishable. As expected, the simulation time for "Ens_64" is about twice what simulation "Ens_32" requires. Hence, we believe that an ensemble with 32 members is enough for this experiment.

We finally calculated the ensemble skill (ensk), ensemble spread (ensp), and mean squared error (mse) (Talagrand et al. 1997; De Lannoy et al., 2006) for the AET values and applied them to verify the ensemble as in Gelsinari et al. (2020). This was added to section 2.4.1 and reads: (Line 284 – 288)

"*The average over the verification period of the ratios between ensemble skill and ensemble spread, which should tend to 1, and between ensemble skill and mean squared error, which should tend to* $\frac{\sqrt{(M+1)}}{2M}$ *(Talagrand et al., 1997; De Lannoy et al., 2006), were calculated on the modeled AET values. First, a simple perturbation of forcing inputs, by adding a random number sampled from Gaussian distributions with different standard deviations, as performed by Gelsinari et al. (2020), was tested."*

(Line 294 – 298)

The Talagrand et al. (1997) verification skills were applied to the ensembles generated with the aforementioned approach, and the most adequate ensembles for the two configurations were retained. The scores obtained for the two ratios were comparable to others found in the literature (e.g. De Lannoy et al. (2006); Pauwels and De Lannoy (2009); Gelsinari et al. (2020)). These ensembles are defined as the open loop, which represents the "prior" distribution. After applying the filter, the resulting distribution is called the assimilation run and represents the 'posterior'."

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.

For a better understanding of the filter application for this paper, we have expanded Section 2.4 with a complete description of the EnKF, including the update equation. (Line 245 - 270) as suggested by the Referee.

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 (Line 206 in the updated manuscript) 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 have stated this more clearly in section 2.4.1 (Line 291-295).

"Thus, a mixed method involving the perturbation of both inputs and parameters, with the latter perturbed by adding a random number proportionally to the calibrated value, was applied. For the UZMs, the parameters selected for the perturbation were K_s and root depth, and for MODFLOW the saturated K_h and S_y. Initial conditions of WT levels were also perturbed to induce a good spread in the ensemble from the early stages of the simulation."

We understand that in many groundwater studies model parameters are updated along with the state variables, but in system theory a parameter is defined as a factor that remains unchanged. The Kalman filter is derived to estimate the state of a system, and we already have ≥ 300 entries, for one model grid, in the state vector. Adding the parameters would make this even larger, and would make the data assimilation system more complicated. For these reasons we decided to focus on estimating the state of the system and not the parameters.

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 Referee for noting this. We modified the sentence into (Line 402-404)

"Table 3 summarizes the RMSE, r and CRPS values for AET, WT levels, and SM contents (upper and lower soil layers), and compares the results of the assimilation run to the open-loop."

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 Referee for pointing this out. The manuscript has been entirely 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 reformulated the sentence as suggested (Line 94).

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

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 (Line 106-107):

"AET data are derived 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 (Line 111-112):

"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 describe the models as well as the coupling framework."

References:

De Lannoy, G.J.M., and R.H. Reichle, Assimilation of SMOS bright-ness temperatures or soil moisture retrievals into a land surface model, Hydrology and Earth System Sciences, 20, 4895-4911, 2016.

Lievens, H., et al. (17 co-authors), Assimilation of SMOS soil moisture and brightness temperature products into a land surface model, Remote Sensing of Environment, 180, 292-304, 2016.

Margulis, S.A., D. McLaughlin, D. Entekhabi, and S. Dunne, Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment, Water Resources Research, 38(12), 1299, DOI:10.1029/2001WR001114, 2002.

Pauwels, V.R.N., and G.J.M. De Lannoy, Ensemble-based assimilation of discharge into rainfall-runoff models: A comparison of approaches to mapping observational information to state space, Water Resources Research, 45, W08428, DOI:10.1029/2008WR007590, 2009.

Reichle, R.H., D.B. McLaughlin, and D. Entekhabi, Hydrologic Data Assimilation with the Ensemble Kalman Filter, Monthly Weather Review, 130, 103-114, 2002.

Gruber, A, G. De Lannoy, C. Albergel, A. Al-Yaari, L. Brocca, J.-C. Calvet, A. Colliander, M. Cosh, W. Crow, W. Dorigo, C. Draper, M. Hirschi, Y. Kerr, A. Konings, W. Lahoz, K. McColl, C. Montzka, J. Muñoz-Sabater, J. Peng, R. Reichle, P. Richaume, C. Rüdiger, T. Scanlon, R. van der Schalie, J.-P. Wigneron, W. Wagner, Validation practices for satellite soil moisture retrievals: What are (the) errors? , Remote Sensing of Environment, Volume 244, 2020.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of Probabilistic Prediction Systems, Tech. rep., Meteo-France, Illkirch, France, 1997.

Schneider, R., Henriksen, H. J., and Stisen, S.: A robust objective function for calibration of groundwater models in light of deficiencies ofmodel structure and observations, Hydrol. Earth Syst. Sci. Discuss., pp. 1–26, https://doi.org/10.5194/hess-2019-685, 2020.

Figures

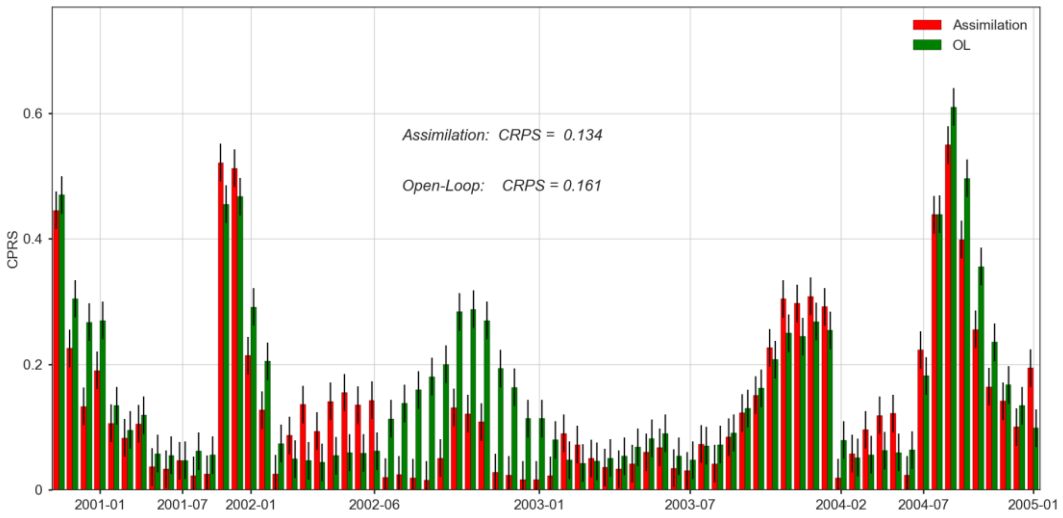

Figure 1- CRPS for WT levels of Configuration-1.

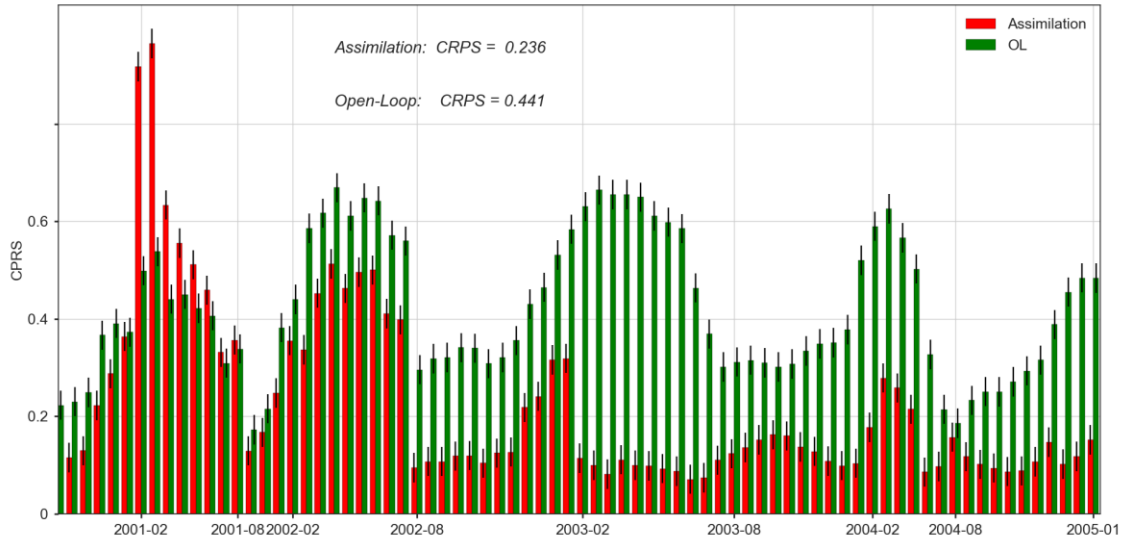

Figure 2- CRPS for WT levels of Configuration-2.

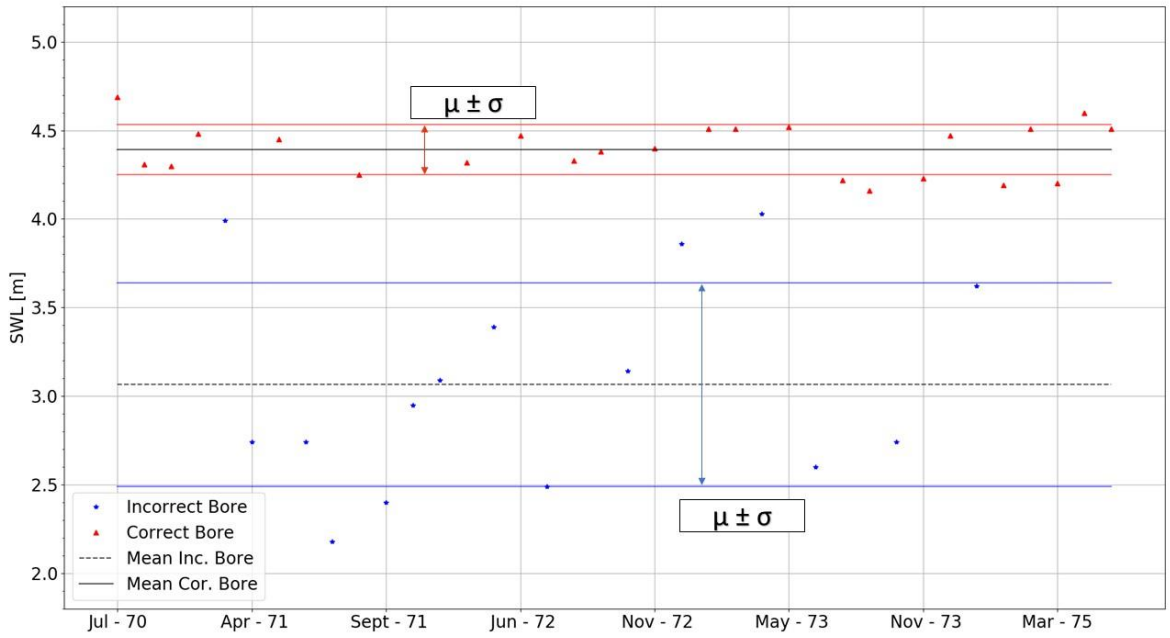

Figure 3 - Water level dynamics for two bores. SWL is distance from the surface.

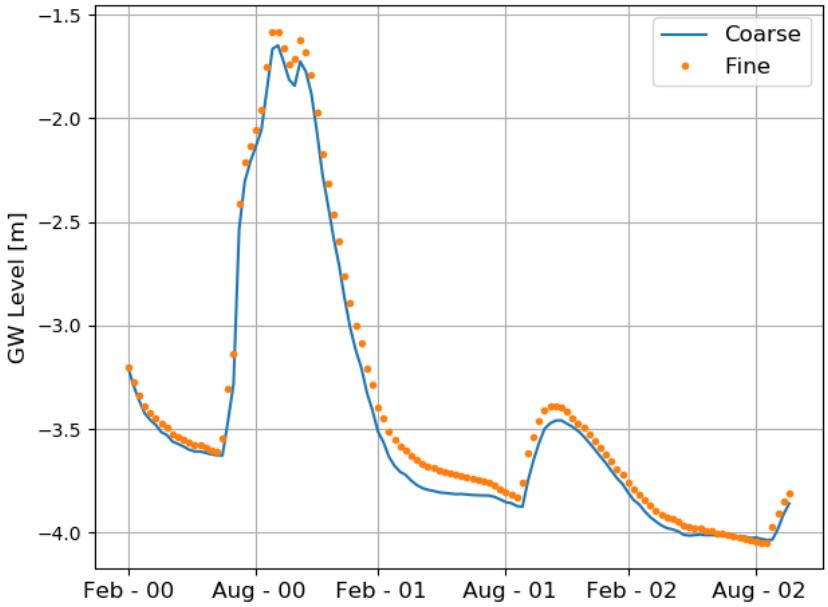

Figure 4- Comparison between modeled groundwater levels using a fine and coarse model domains

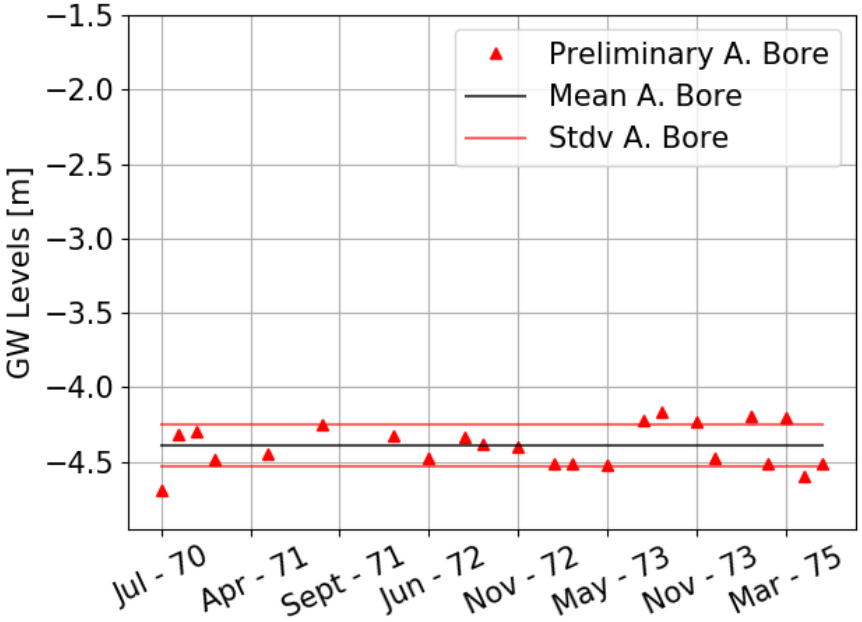

Figure 5- Groundwater levels of the bore used for the preliminary analysis on the BC.

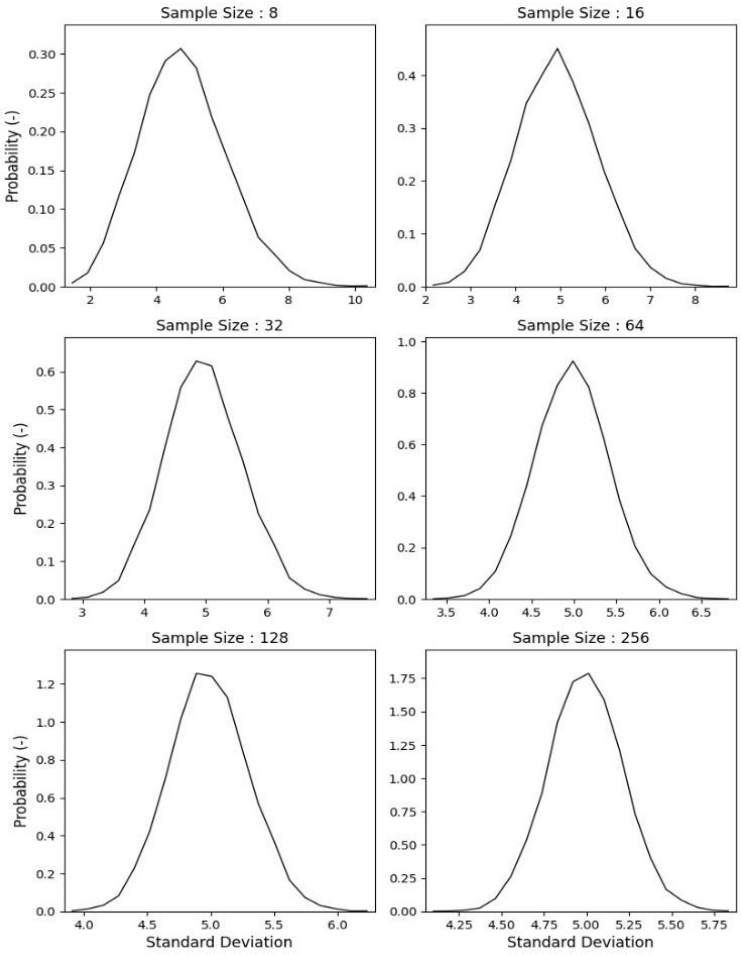

Figure 6- Distribution of different sample sizes.

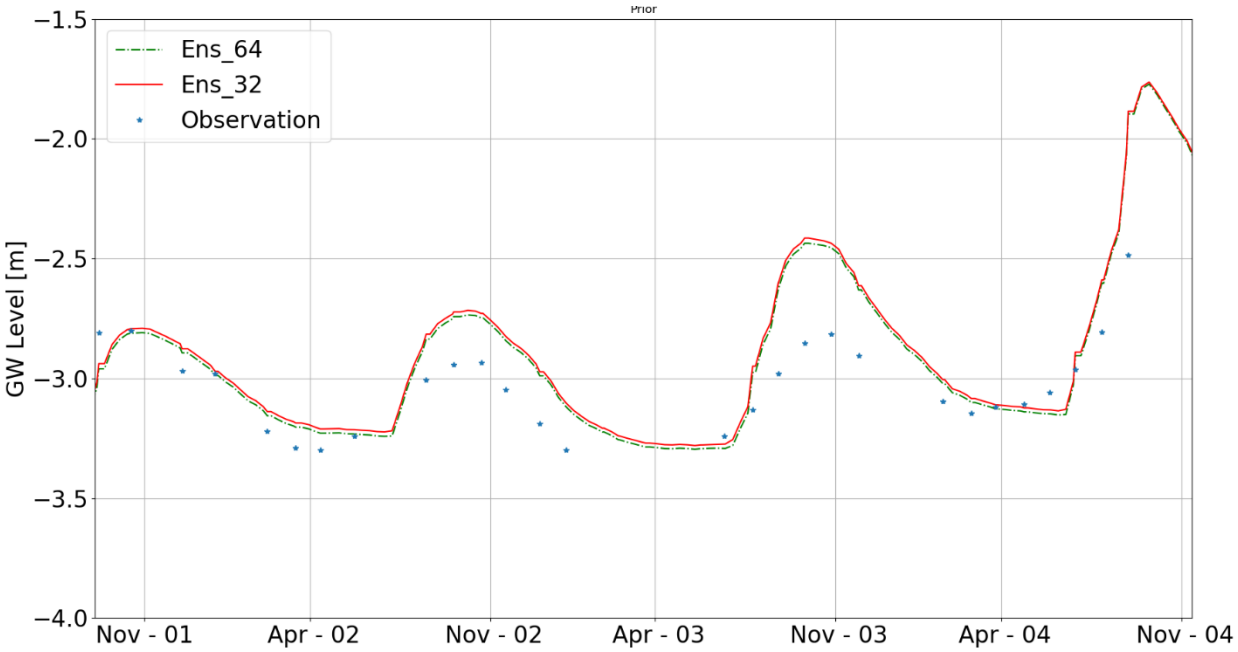

Figure 7- GW level means of the 32 and 64 members Ensembles.