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 the problem of applying significance tests to model simulations is that 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. However, because of the reasons mentioned above, it was preferred not to discuss these tests in the article.

To reinforce the value of the results, the Continuous Ranked Probability Score (CRPS; Hersbach, 2000), which measures the difference between the predicted and occurred cumulative distributions, is calculated. 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

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

H being 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:

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

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.

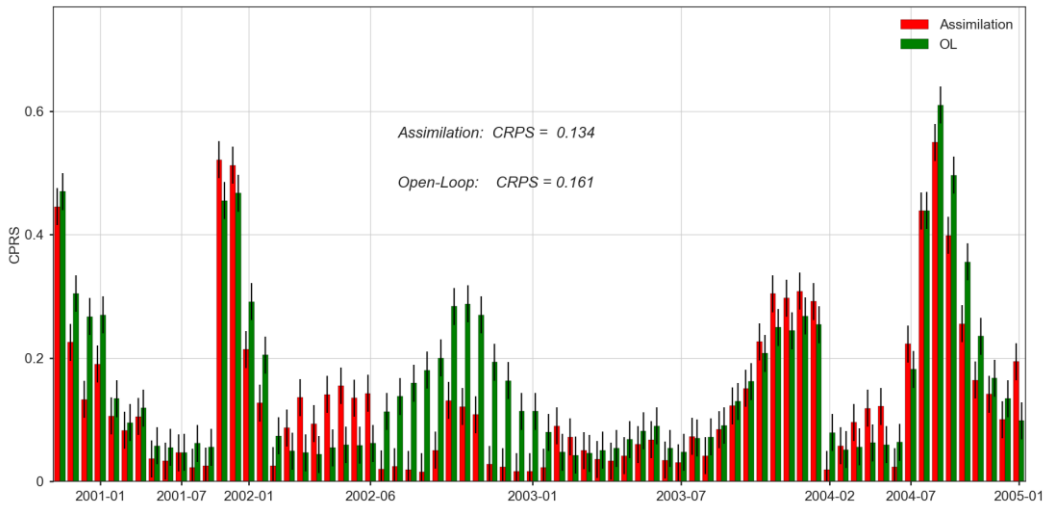

Figure 1- CRPS for WT levels of Configuration-1

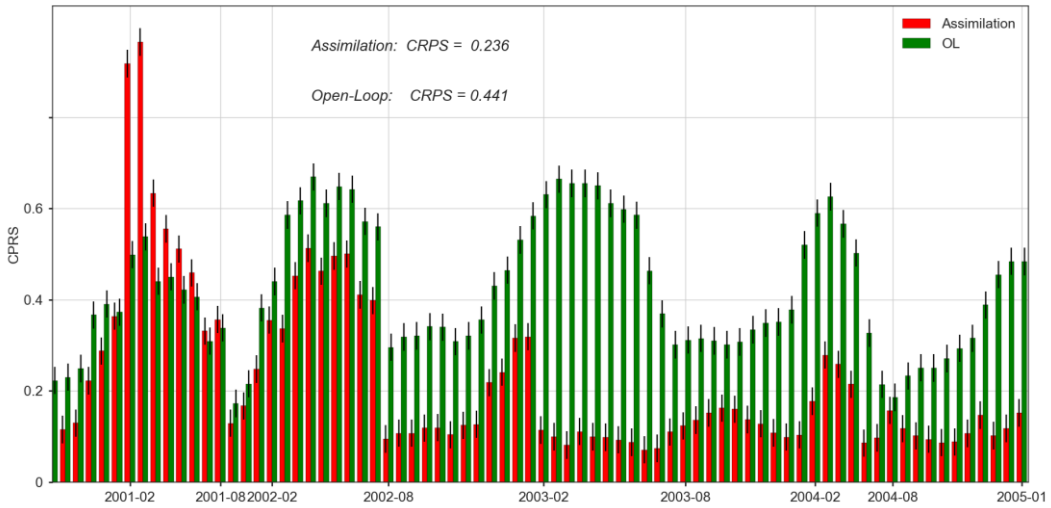

Figure 2- CRPS for WT levels of Configuration-2

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

Line 373 to 376 will be modified into:

"Although marginally, this filter set-up showed the potential to improve the simulated SM content of the bottom part (Figures 10[g] and [h]), for which the best results are obtained (i.e. 0.015). The updating of the entire soil column is a positive result of the assimilation of ET rates, as opposed to the assimilation of remotely sensed SM values. Because of the reduced correlation between the SM contents in the upper and deeper parts of the soil column, assimilating SM values may result in stronger updates in the upper parts of the soil. These updates are not seen within the assimilation framework presented in this paper."

Line 402 to 405 will be modified into:

"The ability of this filter set up to update the entire soil column is an advantage of the assimilation of remotely sensed ET over satellite SM retrievals. However, the anticipated hypothesis, based on ET being influenced by the moisture status of the entire root zone, could not be entirely de monstrated. It was shown that assimilating ET overcomes 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, but the updates of the deep SM content were marginal."

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 ET are calculated against CMRSET. The observation assimilated is originated from the same dataset, to which an observation error is added. 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 at the end of this period, 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 metrics calculation which is performed on ET outputs at the time of the assimilation.

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

We will reword this sentence to "The original forestry plantation dates back to.. "

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 will add "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 ET and WT. In particular, in line 13-16 it is written:

"Because ET is a function of the soil water content within the root zone, as the root water uptake is distributed along with the entire root system (Grinevskii, 2011; Neumann and Cardon, 2012), improving ET estimates, by means of a detailed modeling of the soil water transport, can lead to better simulation of recharge and WT dynamics"

or line 76:

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

Improving the simulation of actual ET 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 ET and WT. To better explain this concept, we will add the following paragraph to the conclusions:

''This study explored the high value of ET data for constraining unobservable estimates (i.e. net recharge) calculated by hydrogeological models. Improving the actual ET fluxes led to better recharge estimates. Thus, as recharge is a key quantity driving the WT dynamics, the link between actual ET and WT 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 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 ET 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 will replace this with the coefficient of variation.

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 will elaborate on the filter theory and set-up and expand this section, as also suggested by other reviewers.

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 will be modified to read as:

"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.*

We will elaborate on the filter theory and set-up and expand this section as suggested by the reviewers.

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

By "limited" we intended that the constrains applied to the value updated by the filter. This is further specified for SM at line 232. For WT the limitation applied was set to 300 mm (i.e. 3 times the layer thickness). We will change 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 detail is 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). We will clarify this in the revised version of the manuscript.

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

We agree with this comment and we will replace the terms as suggested.

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

We agree that is not always lower and we will add this in the text. 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.*

We will indicate the blurred area in the figure.

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

We agree with this comment and we will replace the terms as suggested.

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

We will modulate the contrast to improve readability.

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 will replace the terms 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:

"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 ET 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. ET)"

22- *Figure 8 and Figure 9: what is the cloud of points in the openloop 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?*

It is correct, they are all the ensemble members at the given time-step. This is a choice about the visual representation of the violin plots. The x-axis does not have a dimension, as we understand this could be misleading, we will modify the visualization of these insets.