Review of „EnKF with closed-eye period – towards a consistent aggregation of information in soil hydrology“

Summary: the manuscript presents a study on parameter estimation for an unsaturated zone model. The authors use real soil water content data to estimate parameters and forcings for a 4 layer model using the EnKF. A full iteration of the filter, divided into different steps is introduced to improve the estimation process. During a rain event it is observed that the estimation gets flawed by the models inability to model preferential flow and the observations corresponding to this rain event are excluded from the parameter estimation. This is called a closed-eye period and it is observed that this improves the stability of the estimated parameters.

General comments: The manuscript is interesting and is filling a missing space in the literatur as being a reasonably controlled test case using real data. However, the manuscript is sometimes difficult to overview and the proposed method feels very ad hoc and leaving a number of open questions. Suggestion is major revision.

(throughout the text references to the manuscript are given as: P9,L78 = page 9 Line 78)

Major concerns

1. **Data assimilation vs calibration**
   a. The manuscript introduces the concept of a fully iterative filter, which is an approach that I have never seen in the literature before (and obviously the authors neither since there are no citations on this part). Iterative filters are not uncommon, but they are usually local and not global, or restart versions where the model is restarted and ran until next observations. Some relevant references and comparisons to other iterative approaches would be needed as well as a good motivation to the current approach. The authors approach also shows clear similarities to the precalibration exercise by Hubert et al 2010, to the iterative Kalman Ensemble generator by Nowak 2009 (see bottom for references) and to the SODA by Vrugt et al 2005 (already cited, but not compared to the suggested approach).

   b. Further, the suggested approach seems to me a lot more similar to a batch calibration than a filtering, as the same data is used over and over again to calibrate the model towards one final parameter estimate. As it to me is unclear what effect the top boundary estimation has (see below), and as the truly dynamic parts of the data are removed, I have a problem seeing why a filter and not a proper calibration would provide a good solution, given that the filters after all are after all suboptimal for the unsaturated zone?

2. **Top boundary and its update**
    a. The top boundary leaves a few questions unanswered. It is claimed throughout the manuscript that the top boundary is updated and that this is largely also the reason for the use of a data assimilation method rather than a calibration. However, the authors neither show nor discuss the result of these updates, leaving a reader wondering about the necessity of this inclusion.

    b. Further, I do not understand what boundary was used to drive the model. On P8,L10 it is written that the temporal resolution is daily unless a change of mode occurs. On the other hand, all the figures show finer resolved top boundary, and also the scenario descriptions are dealing with temporal resolutions of 10 minutes when describing the rain intensity. If these events are smoothed across a full day (or if the event is shorter than a day, across a full event), it is a considerable smoothing, and it would be much more preferred if the authors also plotted the forcings they are using. Further, wouldn't this risk effecting the parameter estimation during the rain event, if reality has a much stronger peak than the model? Or is the first observation so far down that this has no effect? Please clarify!

3. **Purpose of model?**
    A nice and stable model is calibrated, that can predict water movement as long as nothing really happens. I find the result very interesting, as it properly questions the use of the standard Richards equation for modelling the unsaturated zone, given that the rainfalls that caused the problems here are quite moderate. The authors should take their space to consider the implications of what they show; can we use the Richards equation in the field? What purpose does a model have that is seemingly data driven (P11,L15) and that cannot predict during rainfall?

4. **Memory**
    Since the model used is disconnected from the groundwater, all information within the model has to travel from the top to the bottom in the 1D column. As the lower layers are more dependent on a longer memory to properly assess their behavior (e.g. it takes longer before we see an effect of a rain event in the second layer than in the first), there is a good risk that we smooth this information when continuously updating the first layer.

    a. Wouldn't a continues update of the top layer risk always smoothing the model such that the estimation in the bottom layers becomes difficult?
    b. Could this be helped by the closed-eye period (as here, the rain event that is too strong for the top layer, may give valuable insight in the lower ones and if the model is not updated during this period, the front reaches down)?

5. **Validation**
    Using data that has been used 10-20 times to calibrate a model with as a validation set is not a particularly strong case. More attention could be given to scenario D, where it is nicely shown that the CE-model has a much better performance than the Standard one. Even better would it

be if one of the three observations in the top layer could be taken out of the calibration and used for validating the model; this would be a strong case.

6. **All parameters needed?**
   The bottom ¾ of the model is only briefly discussed in the manuscript. That nothing happens in layer 4 is not so surprising, but how does it look for layers 2 and 3? Do the infiltration fronts reach down here and how is it with the parameter estimation? If nothing happens, do we need to estimate them?
   How is the effect of the closed-eye period on the second layer?

7. **Overview**
   The authors could consider including a nice block diagram to make their approach clearer (which time period is used for what and where is it iterated and what is feeded to where?). Two example of unclearness: 1) what is used from 2nd iteration onwards as initial condition for period C; the same output from B despite changed parameters? 2) for the top boundary, is it he finally updated value for each model that is also reused in the next iteration?

8. **How to select a closed-eye period?**
   One of the key findings of the manuscript it the improvement using a closed eye period. What is not so clear is how this is to be selected. How can we differentiate between a wanted changes of parameters away from a false prior, from the erroneous updates that the authors show in this work? This feels very ad hoc, and hence also leaves the full paper feeling very problem specific. I think the manuscript would give a more rigorous feeling if this issue was discussed in more depth.

Minor concerns

9. **Longer development plots of parameters**
   The parameter plots only show the last iteration, however, I would find it highly interesting to know how the development throughout the iterations also looks. Example: mean alpha is initially sampled at 4.8 and has in Figure 6 a value for the standard filter of maybe 4.7 with a jump of 0.05. Hence, it cannot have looked the same during the other 9 iteration, then the value would be different. Similarly, the closed eye filter has a stable value of around 5, but how did the way there from 4.7 look?

10. **Performance of org. model**
    Please include in the water content figures, also the performance of the mean of the original model, so that the reader has the possibility to assess the positive development of the model during the filtering.

11. **Why this damping?**
    As the selection of a damping parameter is anything but obvious, I think it is useful if authors

using it gives a one sentence motivation to why it was chosen this way!

12. **Resolution of model**
    The resolution is uniformed 1cm (P5,L20), which for a model with strong boundary fluxes may not be so small. Has the authors checked that the grid size is not also causing issues? Why is "the effect" minimized by 1 cm and not at 1mm?

13. **Why 100 members?**
    The model has 140 unknowns and is 1D, it cannot be a particular difficult to also run a larger ensemble size and reduce the ad hoc tunings, so why is this setup with a small quite small ensemble chosen?

14. **True parameters or heuristic model (was Wollschläger wrong?)**
    a. I'm a bit confused about this discussion. On P12L24 it is stated that the estimated parameters of the closed eye filter better resembles the believed true parameters and that the standard filter ones are more/only heuristic. However, on P14,L1 it is clearly stated that the estimated parameters are all heuristic and only valid in their estimated range, which hence suggests that there are no such things as true parameters.

    b. Also an elaboration on the different result presented here to that of Wollschläger 2009; you get quite different result for a plot quite close by. Are the result so local or where the previously published results not so reliable?

Technical stuff

1. Figure 6: negative alpha values?
2. Language: should be checked carefully. E.g. what does the sentence "The forcing or embedding in space is the initial condition" mean?
3. Consider splitting the "Conclusion" section into a "Summary and discussion" (where some of the discussion points taken up in this review would fit) and a short "Conclusion" section which only contains the actual conclusions.

References:

Huber, E., Hendricks-Franssen, H. J., Kaiser, H. P., Stauffer, F., 2011. The Role of Prior Model Calibration on Predictions with Ensemble Kalman Filter. Ground Water 49 (6), 845–858.

Nowak, W., 2009. Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. Water Resour. Res. 45 (4)