Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-252-RC2, 2020

Hydrology and
Earth System
Sciences
Open Access
Discussions

Interactive comment on "Unsaturated zone model complexity for the assimilation of evapotranspiration rates in groundwater modeling" *by* Simone Gelsinari et al.

Anonymous Referee #1

Received and published: 24 July 2020

article [utf8]inputenc xcolor [colorlinks=true]hyperref

We thank the Reviewer for their detailed analysis of the manuscript. We disagree with a number of the points raised in the review in relation to the presentation of the manuscript, the model conceptualization and the data assimilation algorithm. Details on these aspects of the manuscript are provided below, with the Reviewer's comments in italic.

I thank the reviewers for their response and their clarifications provided below. Unfortunately, no agreement was reached on the most fundamental issues I raised in the

initial response.

Critical among them remain the authors' specific implementation of the EnKF algorithm, and the boundary condition assumptions of the groundwater model, the latter of which I consider serious enough issue to invalidate the results obtained and the conclusions reached. In addition, several points of my initial review were unfortunately not addressed: The authors' precise implementation of the EnKF still remains unreported – with new details in the authors' response below confounding their implementation even further –, as do the governing equations of the authors' custom-made UZM.

Further examples follow below, with my responses being recorded in blue. As a consequence, I must stand by my initial recommendation for a rejection with invitation to resubmit. To keep this from ending up in a constant back-and-forth between the authors and myself - and to give other reviewers the time to form their own opinion -, I will leave the response at this, and potentially revisit this manuscript only towards the end of the review period.

Explaining what kind of parameters your model uses is not particularly helpful if you are not providing the governing equation(s) of your UZM. After skimming Gelsinari et al. (2020) (recommended reading 'for a detailed description') I could also not locate the governing equations anywhere in the original publication. If you develop a new method, at least show us exactly what you are doing mathematically. Section 2.2.2:The explanation of SWAP is great, but again, show us how your UnSAT-UZM functions to at least the same degree of detail (and this does not mean explaining what it does, but providing the equations).

The UnSAT unsaturated zone model (UZM) was presented in Gelsinari et al. (2020), with the equations reported in its supplemental material. We do not think that including the UnSAT model description in this manuscript or in its supporting information is needed. Please also be aware that the submission system does not allow manuscripts with too much overlap with published papers. However, we will revise the manuscript to be more specific on where to find these equations. We will also revise the description of the model to clarify the meaning of the parameters listed in the manuscript.

It is not necessary to reproduce the derivation of the algorithm in full detail, but remember that the reader is not as familiar with your work as you are. Your study sets out to compare two UZMs of differing complexity, but the reader cannot judge which is which (and in what ways they differ) without seeing what the model actually does. Simply provide the governing equations. A few lines of equations and an explanatory paragraph would likely be sufficient. For MODFLOW, for example, this would be a mass balance equation specifying that fluxes across the cell boundaries follow Darcy's Law, that is has various source and sink terms, and storage changes. The governing equations permit the reader to see at a glance which processes you considered and which ones you omitted.

If you are worried about the submission system flagging your paper due to significant similarities with your previous submission, reword your description of the model instead of copy-and-pasting.

First off, normally FloPy generates the entire model input for MOD-FLOW (not just Kh and Sy as well as the model discretization), so just say that you use the MODFLOW-Python interface FloPy to create the model – that's much clearer. Also, provide the MODFLOW input files, because the current information about the model parameterization is sparse at best, and choices made in the other packages (particularly the solver) can severely affect the fidelity of the results obtained.

We agree on the fact that FloPy generates the entire groundwater model. We will add the suggested sentence "FloPy was used to generate the model".

That's a step in the right direction, but you should still provide the model input files. They aren't large in size – particularly for a model as simple as yours – so there is no storage-based reason for omitting them, or simply add a table with all MODFLOW parameters to the supporting information. As it stands right now, the unspecified assumptions you make in the creation of the groundwater model remain undocumented. This makes it virtually impossible for a reader to reproduce your workflow and check the assumptions you didn't actively report.

In addition, choosing eight-day time steps for groundwater models is certainly not conventional. There is no real reason why you should reduce the MODFLOW timestep size to the resolution of the CMRSET data set. The CMRSET ET estimates do not serve as input for MODFLOW (your UZM predictions, available at a finer temporal resolution, already do this job). Even if you only had inputs on this temporal resolution, MODFLOW allows the definition of coarsely-resolved stress periods (during which forcings are either assumed blockwise-constant or are interpolated) subdivided by timesteps of a finer temporal resolution.

The 8-day frequency, albeit unusual for groundwater applications, is the most efficient time step for the assimilation framework, as the filter would update the state variable of MODFLOW (i.e. heads) with this frequency. This minimises the required input/output phase between the filter (to which the CMRSET is an input) and the model, which is one of the most time consuming tasks of the entire process.

Yes, but remember that there is a price to pay for coarser time discretization. If you only ever measure data on, say, Sundays, then you can still freely simulate forward on a daily timescale. Just use the last entry of your seven-day set of predictions for the data assimilation. If you jump ahead in weekly increments, you cannot resolve any dynamics on a finer time scale.

Why is this problematic? Because real groundwater dynamics operate on much finer time scales. Assume it rains during your WT measurement on Sunday. The water table will be temporarily (!) increased, but your model cannot reflect this adequately without having it rain for an entire week. This is not reflecting what's happening in nature. I get it that you want to save computational effort, but your UZM is significantly more finely discretized in space (per 1km cell) than the groundwater model AND runs on an hourly time step. Compared to the computational demand of the UZM, reducing the GWM (groundwater model) time step should come at a negligible extra cost.

I appreciate that coupling MODFLOW to an external UZM requires communication between the two modules, which comes at a computational cost if it occurs at a high frequency. This loops back to the question I asked in the first review: if you are interested in checking what UZMs add to GWM, why did you not compare the results to integrated GWMs with in-built UZMs. For MODFLOW or HGS, both modules are internally coupled and can simulate along the same time scales without major computational cost, and significantly more efficiently than your external coupling. This raises the question of what value your external coupling adds to pre-existing solutions. To answer this question, simply compare them. Does your coupling represent the vadose zone dynamics better than pre-existing solutions? Is it faster? What are the trade-offs?

The Reviewer further states that a finer temporal resolution for the groundwater model would not have greater computational time in the present configuration. Adding that:... If your model takes a suspiciously long time to solve, this might be evidence that the solution of MODFLOW's iterative solver is unstable, which should set off alarm bells concerning the model design. This is something you could check by analyzing the LST (list) file, looking at the residuals and the convergence. Also, you are definitely NOT on a regional scale for groundwater models: your model domain is only 1 km by3 km.

We acknowledge that, at this scale, refining the temporal resolution of the groundwater

model is possible. However, the model is set-up for application at the regional scale. This justifies the choice of the time step of the coupled models to match the frequency of the assimilated observation. Additionally, because observations of water table levels are available approximately every 3-4 weeks, the 8-days step can be considered a fine temporal resolution. The long time taken by the simulation is not due to the issues with the convergence of the iterative procedure to solve the equations. The MODFLOW solver (CONJUGATE-GRADIENT SOLUTION PACKAGE, VERSION 7, 5/2/2005) usually takes less than a second (with residuals below 0.05

This solution time seems fine, thank you for providing more details. I would advise the authors to stress their intention of application to a real regional model more strongly. At the moment, neither the title of the manuscript nor its abstract reflect this intention – only very few groundwater models are created for predictions on such large, truly regional scales.

Of course, this objective opens an entirely new can of worms: if you apply your model at a larger scale, you will suddenly have significantly more parameters and states to estimate. As a consequence, the dimensionality of the span of your state and/or parameter space will be commonly larger than the amoung of particles you can affort. As a consequence, the EnKF's covariance matrix will usually be rank-deficient, which makes the current DA performance unrepresentative for the application you have in mind (right now, you seem to have more particles than space dimensions, hence a good chance at full rank). On the positive side, however, the boundary effects I criticized would likely be less pronounced. Why not scale this model up to a small regional application? This would also address my criticism of the boundary effects I explain in greater detail below.

Are you using the MODFLOW-Sy also as the specific yield for Configuration-2? If so,why do you call it MODFLOW-Sy in the table? Is Sy the same for both soil layers?

Also, the reference to section 2.3 is nonsense – there is not a single mention of Sy in this entire section.

Specific yield (Sy) is a MODFLOW parameter that is independently used and calibrated in Configuration-1 and Configuration-2. The second question is unclear, please reformulate. There are not two soil layers for the MODFLOW model, as explained inline 171-172 "...applied to domain of 1 x 5 cells of 1 km2 each, and a single vertical unconfined layer" Sy is then mentioned in the table of section 2.3; however, we agree that is not further explained in the section. We will amend this.

I assumed that the authors did not calibrate soil parameters which have a corresponding meaning in both modules (UZM and MODFLOW) separately. Specific yield (used by MODFLOW) and porosity (used by the UZM), for example, are related variables: specific yield can never be higher than porosity. As such, I assumed the authors calibrated all model parameters for all layers, with some form of sanity checks in place. I guess that it is unlikely that the water table will reach the upper soil layer, the assumption of keeping the groundwater model one-layered is fine. Carry on.

The reviewer makes a number of remarks on Figure 3

We agree with the reviewer on better presenting Figure 3, and we will modify it accordingly in the revised version of the manuscript.

That's good to hear!

You never mention at what tables you fixed your prescribed head boundaries. Also, you should not reference section 3.1 (part of the results section). You have to explain the model setup in the Methods section, not refer to the results section to complete the picture. Clarify what you mean by saying you changed the boundaries to obtain 'more

shallow water tables': Shallower in terms of distance from the surface (=increased WT) or a shallower water body (= decreased WT)?

We thank the Reviewer for noting this. We described why we selected the constant head boundaries, but missed to report the actual value (i.e -3.5 m below the surface);this will be added in the revised version of the manuscript. We will also modify the text to avoid referencing to following sections in the methods. In the entire manuscript, we always referred to shallow water table being a water table closer to the surface. We will make this clear in the revised manuscript.

A good decision to improve the future manuscript, but the fundamental issue of the boundary conditions remains. More comments on this below.

If your UnSAT-UZM is discretized into layers anyways, why can't you represent the sediment heterogeneity the same way as the SWAP-UZM? This limitation seems poorly justified and reduces the comparability between the two approaches. Elaborate on why you had to make this distinction

The conceptual UnSAT UZM becomes unstable when accounting for sharp vertical changes in the soil parameterizations. Therefore, we accounted for the decrease in the hydraulic conductivity across the soil column. These details were not clearly stated and will be added in the revised manuscript.

This is a bit worrying. Did you use a temporally explicit or implicit solution scheme? If your system is explicit, I recommend converting it to an implicit solution scheme - those are usually a lot more stable. Python has a few elegant solvers for systems of nonlinear equations. The fact that you have stability issues might prevent you from roughening the temporal scale of your UZM without also roughening the spatial resolution (see CFL number: https://en.wikipedia.org/wiki/Courant-Friedrichs-

Lewy$_{c}ondition), which may conflict with your intention to use this for highly computationally-intensive regional models.$

You should definitely report this in the manuscript, as it can be a significant factor for the decision for or against a certain framework. Be honest about the limitations of your work - you should only omit information because it is unnecessary. The fact that your inability to reproduce the sharp discontinuity in the UnSAT UZM was not a design choice but a numerical stability issue (one the SWAP UZM does not seem to share) is critical information to make an informed decision about which framework to use.

The reviewer requests that we clarify what we mean by "interdependence between WT and actual ET".

In many ecosystems where the roots can reach the capillary fringe, transpiration rates depend on groundwater levels; in turn, water table depths in these situations are related to the transpiration rates. This means that the link between the water table dynamics and transpiration must be correctly modeled. This relationship has been shown in many studies, both experimentally (Vincke and Thiry, 2008; Benyonet al., 2006, Marchionni et al., 2019) and using numerical models (Loheide et al.,WRR 2005). In many areas in the southern part of Australia, trees can grow roots to several meters to tap the capillary fringe at large depths. An often used reference for eucalyptus plantations, for example, is that the roots can use groundwater when the water table is about 6 m (Benyon et al., 2006); however, trees have been observed to use groundwater at very different depths, reaching more than 20 m below ground (Eamus et al, 2015). We will provide more details on this in the revised version of the manuscript to strengthen the justification for the application of the models and the assimilation procedure that we used.

Throughout the manuscript we deal with this concept in:

Abstract: "..ET rates are assimilated because they have been shown to be related to the groundwater table dynamics. "

Introduction: "...Because ET is a function of the soil water content within the rootzone, as the root water uptake is distributed along the entire root system, improving ET estimates, by means of a detailed modeling of the soil water transport, can lead to better 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..."

Study area description: "Because in the area more than these plantations have been shown to have direct access to groundwater (Benyon and Doody, 2004)."

Calibration section: "...For both configurations, attempting to assimilate ET fluxes, without reproducing the interdependence between WT and actual ET, yielded poor filter performances." and again by introducing a multi objective function which calibrates specifically on the two quantities. And exposed in the “Results and Discussion” and “Conclusion” section as a fundamental step for the ET data assimilation.

We are convinced this should be enough explanation for the concept. However, we are available to provide more details

I thank the authors for the detailed response. Unfortunately, I’m afraid you might have misunderstood my comment. I have no doubt that evapotranspiration can have an influence on groundwater tables, and that tree can access groundwater through their roots. If that were the case, my criticism concerning the exaggerated effect of ET on the WT dynamics would be nonsensical.

To me, the term ‘interdependence’ implies causation between the two variables. The only thing you have, however, is correlation between the two variables. And as the authors certainly know, correlation does not imply causation. As an example blatantly stolen from “Statistics explained to my cat”: ‘going to bed in your shoes’ and ‘having a headache the next morning’ correlate with each other. However, to correctly reproduce the *cause* for this correlation would require reproducing the confounding factor: getting extremely drunk the night before, thus being unable to take off your shoes, thus waking up with a hangover.

In your case, the situation is a bit more dangerous, as ET can indeed cause WT fluctuations. The critical point is that you have no way of distinguishing which parts of the water table fluctuations are a consequence of *local* evapotranspiration and which are not. You solve this conundrum simply by assuming that there are no regional and seasonal fluctuations of the water table whatsoever. This assumption is incorrect. I have demonstrated this below in Figure 1 and Figure 2 using freely available WT data near your field site. As a consequence, your model must reproduce the observed water table fluctuations exclusively as a direct result of evapotranspiration on the small plot of land which you model.

In short: you reproduce the correlation between AET and WT by neglecting natural WT fluctuations and (at least implicitly) assume it is causation. Since there are additional effects with influence water table dynamics, your model must exaggerate the influence of ET in order to achieve a fit. Your simulated evapotranspiration is partially real evapotranspiration, and partially a surrogate for omitted dynamics.

To verify the spread and accuracy of the ensemble, a number of statistical variables, originally developed for numerical weather prediction by Talagrand et al. (1997), were calculated on the ensemble population.” What kind of ‘statistical variables’ did you calculate? Why did you calculate them? Right now this sentence provides as much information as writing “and then we did some unspecified math.

We agree that more detail are needed on this aspect. We calculated the ensemble skill(esnk), 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.

That’s a step in the right direction, but you have not reported the results of your ensemble verification either. If you go into details about what metrics you use, also report what values you obtained and how they can be interpreted.

2 - Model Choice, Conceptualization, and Set-up

The Reviewer asks why we did not use HydroGeoSphere, or the MODFLOW EVT or UZF packages.

Despite its limitations, MODFLOW is a very commonly used model, particularly in groundwater management. This project was developed to underpin and support existing MODFLOW groundwater management models in the South East and many other parts of Australia. We therefore focused on which conceptualisation of the unsaturated zone is most appropriate to couple with MODFLOW. We will add details on this in the revised manuscript to justify the selection of the model. We did not use the MODFLOW EVT or the ETS packages as they do not account for more complex relationships between depth to groundwater and both evapotranspiration and recharge (Doble and Crosbie, 2017).

This is a valid decision, and it’s good you want to add details. But your manuscript sets out to explore which degree of complexity is required in order to reproduce the UZ dynamics – and their influence on WT – adequately (see your manuscript title and abstract). This is exactly the reason why you should compare it to pre-existing, inbuilt solutions! Using the simplified integrated UZ modules of MODFLOW would be a perfect background to contrast what value (if any) apparently more sophisticated UZM such as SWAP or UnSAT add to the groundwater models, and at what expense they come. This would be a great opportunity to show what your approach can provide, and at what cost it comes (of course provided you fix the GWM first, else the results are unrepresentative of realistic conditions).

As it stands, you essentially present two different variations of a highly expensive SwissArmyHammer$^{\text{TM}}$ ("the high-powered tool for all your hammering needs") – the 1000 dollar and the 2000 dollar variants – but fail to show which value they would provide over the standard run-of-the-mill hammer we already have in our workshops. That's not very good marketing.

The UZF1 package, which is already coupled to MODFLOW, would be an alternative to SWAP. However, SWAP solves the full Richards equation, while UZF1 solves a kinematic wave approximation of the Richards equation. Also, SWAP is considered a better model for the soil-water-vegetation interaction (Vereecken et al., 2016), an important aspect when modelling vegetated areas. SWAP combines detailed modules on soil water flow, root water extraction, vegetation development and evapotranspiration. The model has been extensively applied in scientific research (https://swap.wur.nl/).

Certainly, but as I already said above: your manuscript sets out to gauge what degree of complexity is required. In groundwater flow, you could theoretically attempt to resolve flow on the pore scale, but in most applications little value is gained by doing so. More realism in the UZM always sounds nice, but in practice we have to make trade-offs. The more complex the model, the fewer ensemble members and the less fine resolution (temporal or spatial) you can afford. In the end, the advantages gained by coupling to a more sophisticated UZM may not be worth it.

To see how much SWAP really adds, you might compare it to HGS instead of MOD-

FLOW, which also solves the full Richard's equation and simulates both the vadose and phreatic zone, requiring no external coupling. Again, you are attempting to improve on something which might already have a more efficient pre-existing solution. Show the reader what value your coupling adds over what already exists (in terms of prediction fidelity, versatility, or computational effort) so that she or he may make a more informed decision.

Your groundwater model is highly unrealistic, its coupling with the UZM likely extremely exaggerated (as hinted by the need to calibrate both parts to avoid 'poor filter performance'), which invalidates the results you obtain and the conclusions you draw from it.

We disagree with this statement and we address all the remarks the Reviewer made in relation to this issue below.

I thank the reviewers for addressing these points and will also reply below.

...with the groundwater model, I take issue with the decision to resolve a 3 km by 1 km model domain with only three numerical cells (plus two prescribed-head boundary cells). This has a number of highly questionable consequences: If you wish to resolve a hydrogeological depression, choosing such a coarse spatial resolution severely affects which drawdown curve you obtain

The scope of this manuscript is not to resolve a hydrogeological depression in space. The observation of water table levels presented are at a single location. We recognize that the model conceptualization phase is a fundamental aspect of any numerical simulation. The simple model domain is a result of numerical experiments which have shown us how it is possible to obtain similar water table dynamics with both a fine (20cells in the x-axis) and coarse (5 cells in the x-axis) model domain. Figure 1 at the end of this document shows the water table fluctuation, calculated with Configuration-1,in the cell of the two domains simulating a similar location in the same study area. Itis worth to say that the run-time for the fine set-up is roughly 5 times larger than for the coarse set-up. If we were to resolve the cone of depression induced by the root transpiration we would have certainly decided for a different conceptualization.

But you do resolve a cone of depression induced by root transpiration, whether you set out to do so or not. This is literally the only thing your model is doing. You assign two constant fixed head boundary conditions, so any water table fluctuations you simulate in the centre of your observed model domain is the direct result of a water table depression induced exclusively by your UZM.

Your manuscript even explicitly states that you also have to consider the water tables as part of the calibration objective function in order to obtain reasonable results (more on this below). Since you calibrated both models jointly, any conceptual errors from grid resolution and inadequate boundary assumptions do not remain confined to the GWM. During calibration, these errors propagate to the UZM as well – a phenomenon known as parameter surrogacy (although in this case “model surrogacy” would be more fitting).

Furthermore, your statement here seems to conflict with the objective you set out earlier: to apply the resulting coupling to a regional flow field. In this setting you will have to reproduce a hydrogeological depression in space – else your UZM would not be required for the GWM at all.

It is, however, good to see that the coarse spatial resolution does not seem to have a major impact on the predictions.

We purposely did not use a very fine resolution because (1) the MODIS-CMRSET

data are already at a 1 km resolution and (2) the intention is to apply this work at a regional scale, thus prohibiting very small cell sizes. The work is the first step toward the application of the EnKF to large-scale groundwater modelling.

As mentioned before, I recommend stressing the intention to use this in a regional-scale model more strongly in the manuscript. And if that is the case, why did you not test the model on a limited regional scale with multiple observation wells and a limited number of interacting cells (Limiting yourself to, say, a 20x20 grid)? The advantages you reap by confining your study to a small, well-defined region (namely somewhat detailed knowledge of the UZM-relevant parameters) is squandered by calibrating your system to a groundwater model with poorly justified, unrealistic boundary conditions.

The Reviewer says that with the given boundary conditions, "In essence, you artificially prevent that any other external influence (say, regional irrigation water extraction during summer, regional water table fluctuations due to seasonal variations in precipitation and ET, etc.)"

The location of the study is in the centre of a forestry block, more than two kilometres from any groundwater extraction, and was originally selected for a previous study to specifically look at the impacts of forestry on groundwater (Benyon et al., 2006). Groundwater use (not total ET, but ET from a groundwater source) by forestry (Pinus radiata and Eucalyptus globulus) for this region, where the water table was less than6 m below the surface (and soils were light to medium textured and groundwater was not saline) was estimated by Benyon et al. (2006) to be on average 435 ML/yr (range108–670 ML/yr) for a 1 km by 1 km fully forested cell. This exceeds 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 point scale. Working on the assumption that there is not likely to be more than one or two high extraction bores in a 1 km square area, water use by forestry is seen to be significant.

For this reason, and from personal communication with water managers in the region, it is a valid assumption that the impact of recharge and transpiration from forestry on groundwater would far exceed the impacts of extraction through pumping and other temporal variations.

Again, the fact that the forest is the exclusive cause of WT fluctuations is a very strong claim which you have not sufficiently supported with data. This shouldn't be particularly hard to do. Australia has a very dense and for the most part publicly accessible network of groundwater wells:

http://www.bom.gov.au/water/groundwater/explorer/map.shtml.

I have taken the liberty to select a few wells near your field site (see attached images at the end of the document, Figure 1 and Figure 2) and took a look at the water table dynamics. Unsurprisingly, even though the well I show in the Figures is about 2 km away from the forest, it shows a clear seasonal water table trend.

The problem is that the water table dynamics you observe are not exclusively local. They depend on local and regional recharge, local and regional water extraction (plant-based or anthropogenic), evaporation, exchange with rivers, and simple physical groundwater flow – effects which occur and accumulate over hundreds of kilometres.

If your background is in vadose zone modelling, it may seem natural to neglect lateral flow (thus permitting one-dimensional, pixel-based approaches), as they are more or less justified in the unsaturated zone, but this assumption is simply not valid for groundwater flow. I wish it were – it would allow hydrogeologists to employ the same elegant localization schemes meteorologists are using. Unfortunately, local changes to the groundwater levels propagate and accumulate non-trivially in all directions.

In your case, the assumption of a constant head boundary condition is demonstrably not supported by nearby GW observations. I recommend revising the groundwater model significantly: make it properly 2-D (add cells in each horizontal direction, so you can consider regional flow in 2-D) and assign time-varying prescribed head boundary conditions according to the regional water table dynamics in all directions from the forest. Only then can the reader appreciate how much value the UZMs add to the GW dynamics.

In an uncalibrated, coupled model the influence of actual ET on the water table would be likely negligible (there is a reason why hydrogeologists often neglect the UZ).

Where groundwater is deep, this is a reasonable assumption. ET is exclusively from infiltrated rainfall in the unsaturated zone, and not supplied by the groundwater. Most of the surface processes are damped by the large lag time between infiltration and recharge. However, where groundwater is within plant rooting zones, more specifically lower than 10 m from the surface with a range of 6 m 20 m (Crosbie et al., 2015,Benyon et al., 2006), particularly in semi-arid areas, ET can have a significant effect on the water table, as demonstrated in Vincke and Thiry (2008), Benyon et al. (2006)and Eamus et al. (2015).

Yes, but what I meant in this context is far more specific. The critical word in your response is 'CAN have a significant effect on the water table'. It doesn't have to. If local (!) AET is indeed the driving force for the WT fluctuations you observe, you would not have to calibrate the groundwater model as well. You could simply calibrate your UZMs, and apply the resulting AEM estimations as negative recharge forcing to a groundwater model using prior (i.e., uncalibrated) hydraulic parameters.

The water tables would of course fluctuate a bit, maybe even significantly, but likely not enough to reproduce the correct WT dynamics. I assume this is what happened in your unreported preliminary experiments, before you had to add WT dynamics as a part of your multi-objective function. As a consequence, you must calibrate both the UZM and the GWM – because, as I stressed too many times already, local AET is not the only source of WT fluctuations. Since you assume that it is, the GWM must meet the UZM halfway to produce the right response for the what are likely largely the wrong reasons. And, considering your WT results, even this does not seem to be enough. Again, the fix is simple: create a more representative model of regional dynamics (e.g., 20x20, which is still very small) and assign time-variable head boundaries at the edges. Maybe your UZM wouldn't have to fling the WT as strongly around as it does at the moment, and small nudges might be enough.

Since you dynamically readjust the resolution of your UZM: Do you account for hysteresis effects? How thick are these layers? Are you assuming the water tables as blockwise-constant or do you interpolate between them?

We do not account for soil water hysteresis effects, as this has a minor effect on regional groundwater recharge simulation. The layer thickness for the UnSAT model is fixed to 10 cm. We assume that the water table is measured at the centre of a model cell and blockwise-constant.

Fair enough, thank you for the information. This might warrant a small comment in the manuscript, though.

why is the root depth so dramatically different between the two configurations? This should be a system property, and as such be constant between the two configurations. It makes no sense to assume two different root depths if you want to compare the two models.

As the two UZM are conceptually different, they dissimilarly account for this parameter.

Specifically, as UnSAT does not simulate the capillary fringe, the direct extraction from groundwater is possible with a root depth that allows roots to reach the water table. Therefore the root depth in UnSAT is larger than for SWAP. We will add a comment on this in the revised version of the manuscript.

This sounds more like both models account for the parameter the *same* way, but (in yet another instance of parameter surrogacy) the UnSAT UZM uses this parameter to compensate for an omitted effect – in this case, capillary rise. If we cannot trust the parameters of the UZM to represent the processes they are supposed to represent, what do we stand to gain from choosing a potentially more 'realistic' representation of the UZ dynamics? This could warrant greater discussion.

3 - Data Assimilation

The Reviewer makes a number of remarks regarding the data assimilation algorithm, which can be summarized as follows

The Reviewer states that we may not have understood the theoretical foundations of the DA algorithm because we state that the EnKF was used because of its ability to deal with highly non-linear systems, and encourages us to revisit the derivation of the EnKF.

Evensen (1994) developed the EnKF to deal with nonlinear systems, by replacing the propagation of the error covariance in the Extended Kalman Filter by an ensemble-based calculation. He stated in the abstract of his paper: "The proposed method can therefore be used with realistic nonlinear ocean models on large domains on existing computers". The difference between the Discrete Kalman Filter (as originally derived for linear systems by Kalman), the Extended Kalman Filter and the Ensemble Kalman Filter is the calculation of the error covariance. The only one in this family of filters that does have a derivation is the original DKF.

As I stated in my original response, being able to deal with nonlinear systems, and using the algorithm *because* a system is nonlinear are two different things entirely. Simple example: In a pinch, a cooking pan can be used to dig a hole, but choosing a pan (over a spade) *because* you want to dig a hole is still the wrong decision. Choosing a pan because you cannot afford a spade, however, can be a valid choice. At the moment, it sounds like you extol the pan for its great digging properties.

(Replace all instances of ‘pan’ with EnKF and ‘spade’ with particle filter/variational approaches of choice)

And, as you correctly stated, the key is the calculation of the error covariance, which is only sufficiently representative of the underlying probability distribution if the prior is sufficiently Gaussian and the forecast sufficiently linear. Otherwise, these properties aren’t retained and eventually lost.

There are too many publications already which are vague on the properties and limitations of the EnKF, resulting in a lot of derivative studies from other authors which learned the algorithm from such studies without fully understanding the (admittedly not very intuitive) source material. It pays to be precise. I stand by my initial request that you should report the exact implementation of your EnKF application as well as the assumptions and simplifications you introduced. Both are still missing.

The Reviewer states that the EnKF can only use nonlinear models because it fits a Gaussian distribution to the ensemble after the forecast.

The EnKF does not fit any distributions, it updates state variables. Also, a filter does not use a model, it is the model that uses the filter.

Yes, the EnKF updates state variables under the assumption of Gaussianity (or at least the assumption that mean and covariance [as estimated from the ensemble] are sufficient statistics for the Bayesian update). Per definition, this assumption is only true for Gaussian distributions.

No, the model is part of the filter, not the other way around. This is easy to check. A model can provide predictions without a filter. A filter, on the other hand, does not work without some kind of model (or, more precisely: a forecast distribution). This is why I recommended revisiting the underlying theory. Despite how it is often treated in some disciplines, the EnKF is decidedly *not* some heuristic optimization algorithm. Fundamentally, the EnKF is solving a filtering problem, the sequential Bayesian inference of state uncertainty. In a Bayesian framework, uncertainty represents belief or knowledge about a system variable.

https://en.wikipedia.org/wiki/Recursive$_{Bayesian_e}stimation$

The model is what establishes a causal relationship between successive time steps, allowing you to salvage some (but, without a perfect model, never all) information from the previous time step to inform the states at the next time step. A model can be as simple as “do nothing (deterministic identity), add noise” to a complex “coupled GWM+UZM, add noise”. Without a model, none of the information at previous time steps could be salvaged, and the filter would essentially just solve completely disjointed Bayesian inference problems. As a consequence, it would not constitute a filter in the classic sense.

In essence, a filter progresses like this:

state prior at time zero stochastic state forecast from t=0 to t=1 (increase entropy) state assimilation step at time t=1 (decrease entropy) stochastic state forecast from t=1 to t=2 (increase entropy) state assimilation step at time t=2 (decrease entropy) etc.

This is essentially an ebb and flow of information. What makes the entire thing a little bit more complex is that the deterministic model can also affect entropy depending on whether its dynamics are dissipative, chaotic, or conservative. Part of the reasons why the EnKF performs a lot better in the atmospheric sciences is that their systems are chaotic, and hence already naturally add entropy to the system during the deterministic forecast. Most hydrologic systems are dissipative and hence necessitate large forecast errors or fudging the assimilation step (e.g., through localization, damping, etc.) in order to retain ensemble spread against the combination of Bayesian inference and deterministic dynamics both removing entropy.

The Reviewer suggests using the particle filter.

We have experience with Particle Filters, and our experience is that a much larger ensemble size is required than for the EnKF, which is one of the reasons why we decided to use the latter for this study.

Computational limitations are an acceptable pragmatic reason for using the EnKF instead of theoretically more justified approaches. As mentioned above, I would recommend that you motivate your choice of the DA algorithm this way, instead of justifying it with the nonlinearity of your system. This makes little sense.

The Reviewer argues that we should not justify the ensemble size with examples from literature.

Enough papers have been published on this issue. 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. Based on our experience with

models of this level of complexity, an ensemble size of 32 is more than sufficient.

Then justify the ensemble size through the allegedly comparable complexity of the related studies you cite, not just the fact that someone else used this ensemble size for some other model. It's a simple fix.

The Reviewer argues that the assimilated variable should be in the state vector and that the observation matrix should contain zeros and ones.

The main advantage of the EnKF is the possibility of assimilating proxies of state variables that are nonlinearly related to the state variables. Examples are the assimilation of brightness temperatures or radar backscatter values into land surface models to update the soil moisture simulations, or the assimilation of streamflow into flood forecasting models to update the modeled soil water content.

Yes, you can use nonlinear observation operators to relate your observed variables deterministically or stochastically to the states you are predicting.

But that's not what you are doing, is it? As far as I can see, your UZMs predict AET *jointly with* soil moisture, as AET is required to close the deterministic mass balance. Similarly, the remote sensing data you assimilate has already been converted to AET estimates.

You don't require a proxy, you are simulating the state of interest directly. At least you state so in line 115-116 for the UnSAT-UZM, and according to Figure 2 for the SWAP-UZM. As such, AET should be part of the state vector and can be extracted with a matrix of zeros and ones (or any other linear combination, unlikely to be relevant in your case). For the purpose of data assimilation, it does not matter that AET at time t may not directly predict AET at time t-1, as it does depend on other state variables which connect the two.

The examples provided by the authors may reveal some confusion on their part: An example of a simple nonlinear observation operator would be

$$f(x) = x^2$$

The examples you mentioned seem to describe two different systems: converting radar backscatter values (which no model I am aware of predicts directly) deterministically into a physical variable the model *is* predicting would be an example of a (possibly nonlinear) observation operator. In this case, the physical variable would be a proxy for radar backscatter, or vice versa. Another example would be rise in a mercury thermometer, which can be (more or less linearly) transformed into a temperature estimate.

However, all flood forecasting models I am aware of predict a state corresponding to observed streamflow directly (the clue is in the name). This does *not* mean that streamflow is a proxy for soil moisture. In the example you mentioned, soil moisture would be a hidden (i.e., non-observed) state, which in the EnKF can be updated through the cross-covariances between streamflow (the assimilated state) and soil moisture (the hidden state).

Again, let me re-iterate what I have said before: please show explicitly what your EnKF algorithm is doing. To meet you halfway, and better illustrate my bewilderment about your (as yet unreported) implementation, I have attached a small Python script for a simple, standard EnKF with this response, which demonstrates the properties I describe here. If your algorithm differs significantly mechanistically from the one I have shown, show us in what ways it differs and how this is justified.

The Reviewer states that parameters should not be disturbed when generating the ensemble.

The essence of an ensemble is that all members have different properties. Reichle et al. (2002), who introduced the EnKF to the hydrological sciences, explicitly state that parameters can be disturbed or that even different models can be used to estimate the error covariance

You failed to understand my comment, and simplified my response significantly. First and foremost, I did *not* simply state that the parameters should not be disturbed. In my initial review of the manuscript, I described three different scenarios in which model parameters are explicitly considered, two of which explicitly permit or even require initial variance in the parameters, and invited the authors to clarify which of these configurations (if any) they ascribe to.

As I mentioned in the original review, and repeated here, the EnKF solves a filtering problem. At its core is a sequential Bayesian inference which seeks to infer a specific probability density function. If you make any changes to the EnKF, show us what this means in terms of the Bayesian inference. In your case, I do not see what meaning the resulting system could have. To aid the authors in future attempts to clarify it, I have provided a simple figure (Figure 3) illustrating what a standard state-only EnKF and an augmented state vector EnKF are inferring, and how the authors' approach does not seem comparable.

The key difference is that in both the classic and state-vector augmentation EnKFs particles share the same space (in the former case: full state space, conditioned on a specific slice of parameter space; in the latter case: a combined space with all uncertain state and parameter dimensions). In your case, each ensemble member is located in its own space (due to conditioning each sample on different parameter values), yet you somehow infer a joint pdf. Clarify this, please.

I would also like to note that the Reichle et al. publications only refers to this possibility in their outlook and did neither use nor detail this approach in the manuscript provided, or whether they would have implemented this the same way you did. If you use this method, I encourage you to include a short derivation of what pdfs you are exactly approximating with your ensemble. It is dangerous to make changes to a Bayesian inference algorithm without being completely certain what they would entail. Help the reader understand what the consequences of your proposed solution are.

The Reviewer refers to Hendricks-Franssen and Kinzelbach (2008) for an explanation on how to update parameters and states.

The main point of the paper that is referred to is error covariance inflation, which is necessary because the ensemble is inadequately generated and over time be comes too narrow. Our approach explicitly avoids this problem.

Yes, the paper focuses on error covariance inflation, but it provides a very nice, concise overview on how to infer states and parameters jointly. This might keep you from having to initialize your parameters with a calibration using the states you want to predict later, which is already quite a dark grey area as far as Bayesian inverse crimes go. As a nice side-effect, it gives you estimates of parameter uncertainty.

No, error covariance inflation is not only required because the ensemble is inadequately generated. The issues the authors describe in this paper are a lot more fundamental than that: you will almost always have an insufficient ensemble size, and as a consequence you will get spurious correlations. ‘Adequately generating’ your ensemble can at best help to ensure you do not end up with a geometrically degenerate ensemble (i.e., dimensionality of the support of your ensemble $\leq$ ensemble size-1 or dimensionality of state space, whichever is smaller). Further issues can arise from the fact that the assumptions of linearity and Gaussianity may not be met.

Yes, your approach avoids this issue, but this alone is not worth much if you are not showing what this means in terms of Bayesian inference. What consequences does it have for the underlying pdf?. Forecast errors are supposed to compensate for the imperfection of the numerical model, and that’s the reason why people add independent noise onto the forecasts: so that the stochastic forecast can obtain states the deterministic model could never predict.

As far as I can see, you are not really making a stochastic forecast, your ensemble merely spreads because each particle has its own deterministic attractor due to the fact that you condition each particle on a different parameter set. This is not the same as a stochastic forecast. All other approaches I have listed above add noise to the state (and, optionally, parameter) variables for a reason. This can upset mass balances (and has to: in case the mass balances are wrong). If that’s an issue, the approach I mentioned by Reichert and Mieleitner, for example, circumvents this issue by randomizing the parameters (and thus the mass fluxes between different state variables) instead. Your approach is the only one which does not seem to be stochastic at all. Explore this in detail, or revisit the fundamental assumptions to see if they make sense.

The Reviewer states that figure 10 is comical, and that there is something wrong with the DA algorithm, ecause the soil moisture does not change while the AET does change, and that this means that the ensemble has collapsed

The AET is the integrated response of the root system. The DA system has lowered or increased the water table level, and thus layers are added to or removed from the unsaturated zone. This will modify the modeled AET. This lowering or raising of the WT means that the ensemble has not collapsed. This result does not mean that the DA system doesn’t work, but, conversely, it proves that it does work.

Furthermore, here the Reviewer states that the ensemble has collapsed. Towards the end of the review, the statement is made that the settings (through selective parameter perturbation) seem to have been deliberately designed to inflate the WT uncertainty disproportionally, amplifying the Kalman update through the cross-covariances. These comments are contradictory.

No, these comments are not contradictory, and I can show you why. As I mentioned above already, the EnKF updates hidden states (soil moisture) through the cross-covariances with observed states (AET). This has two consequences:

a) the only time the EnKF does not update hidden state variables is if the correlation between both variables is zero, or if the uncertainty of both state variables has (virtually) collapsed. Since I heavily doubt that soil moisture is uncorrelated with AET*, the ensemble must have collapsed in the corresponding state space dimensions.

b) if the uncertainty of a hidden variable is larger than the observed one and both variables correlate (a pre-requisite for updating hidden states), then the magnitude of the update step is amplified for the hidden variable. Think of a lever, with the length of each side corresponding to the uncertainty in the observed and hidden state dimensions.

Since seeing is believing, I have attached a Python script showing a single EnKF analysis step for a primitive system with only two states. State 1 is observed, State 2 is hidden. The uncertainty of State 1 has practically collapsed (i.e., the observed data point lies significantly outside the ensemble). To reproduce the issue I described in point (a), simply lower the standard deviation of the hidden variable to a value similar to the observed variable (thereby reducing the magnitude of the update of the hidden state), or set the off-diagonal entries of the covariance matrix to zero (thus preventing an update of the hidden state altogether). You will see that hidden state updates are negligible if the uncertainy has collapsed to a similar degree than the observed variable, and that the hidden state updates are zero if there is no correlation.

To see what I describe in point (b), play around with the uncertainty of the hidden state: the larger its uncertainty, the larger the hidden state update. Your setup seems to always ensure that there is a large ensemble spread of the WT, hence augmenting the Kalman update effect on the WT predictions. Increase the standard deviation of the hidden variable, and the magnitude of the hidden state update will increase accordingly, while the update for the observed state remains unchanged.

If neither of these effects above is responsible for the failure to update the soil moisture predictions, then I encourage the authors to demonstrate what exactly is happening.

*even if soil moisture and AET were uncorrelated, it is unlikely that you would reproduce this with an ensemble of only 32 particles. That's the problem with spurious correlations, and the motivation behind localization approaches.

4 - Final remarks

With this document we are expressing our disagreement with the reviewer's remarks about the methods adopted in the manuscript, reinforcing the appropriateness of our choices and the theory underpinning the experiments. However, in this reply we are not going to provide a response to the remarks on the "Results and Discussion" and "Conclusion" sections, as we have shown the Reviewer is basing these comments on inexact assumptions.

I thank the reviewers for their comments. Unfortunately, simply stating disagreement is not a productive way to address criticism raised during a review. Several points I raised were not addressed at all, and those were not confined to the "Results and Discussion" and "Conclusion" sections. A detailed description of how exactly the authors implement the EnKF, and what the novel assumptions made therein entail, is still entirely missing. The governing equations of the UZM should be reported. A comparison to pre-existing methods would be a valuable addition. And showing the ensemble correlation matrices can aid the troubleshooting of your DA algorithm.

In a bid to help you improve the manuscript for a future re-submission, I have gone to great lengths to clarify my comments in as simple terms as I can, and demonstrated where the authors seem to err in several aspects of this study.

In general, the tone of the review and the use of words such as "fraudulent", "fudging", "comical", etc., are not appropriate for a review for a scientific journal. Certainly not for a review in the public domain. It is also not appropriate to state that the authors do not understand the methods that they are using, while is the reviewer that appears not to have understood the applied methods.

It was certainly not my intention to disparage the authors. It is, however, important to stress how certain sections of this manuscript can be received during an initial reading. The reader cannot know the intention of the authors, so when the authors make poorly justified assumptions which have the potential to bias the results in preferential directions, and fail to draw attention to this bias, alarm bells must be set off. Similarly, declaring that data assimilation improved the simulation of soil moisture and then proceeding to show a figure which shows virtually no difference between the open loop and the DA run [Figure 10, "However, the assimilation improved the 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."] is a quite serious case of over-selling what are not great results. This can be amusing during first contact: You read that DA improves soil moisture predictions (not entirely unexpected), then scroll down to a figure which shows that virtually nothing changed relative to the open loop run. Since I trust the authors did not intend their manuscript to be received this way, it is worth pointing out this probably unintended effect. I sincerely hope that the

code snippet I provided will aid the authors in gaining a better understanding of how the (standard) EnKF updates work and support them in revising the setup of their DA algorithm. If for whatever reason their approach functions differently, I hope the example provided alerted the authors to the existence of an alternative interpretation of the EnKF, and stresses the importance of reporting exactly what they were doing.

Don't take criticism of your work as an insult, use it as an opportunity to improve and clarify your manuscript. If I questions whether you understood the DA algorithm, take this as an opportunity to show the world that you did: report your EnKF implementation, or write a small synthetic example (as I did) to demonstrate where my assumptions are wrong. This is a review, not trench warfare. Don't just defend your points, engage with the criticism raised.

Case in point: the boundary condition issue I mentioned here a few times. It can be readily demonstrated to be unrealistic, with data freely available online – again, something that would have been your task which I had to do for you. Basing a publication about the beneficial use of UZMs for GWMs on a model setup which can only create WT fluctuations through local effects of a UZM is the hydrological equivalent of locking a cat in a box for a week with nothing but a lettuce and 200 slugs. Sure, at the end of the week the cat will have eaten the slugs to stave off starvation, but subsequently declaring that "cats are the new secret super-weapon against garden pests" would still be deceiving. The experiment may have yielded results which could be interpreted in result of the headline, but the experiment itself was not representative of reality, where cats make other choices of food. A similar effect might apply here. Refuting this criticism would be pretty straightforward: simply create a new groundwater model, set time-variable prescribed head boundaries from nearby wells, revise the DA assumptions, and see if the results are still significant and how much the UZM contributes. Your work will be all the better and more relevant for it.

All in all, I encourage the authors to adopt a less antagonistic stance in the future. I understand that you may identify strongly with your work - it's always good if scientists are passionate about their work -, but criticism voiced during a review is not a personal attack. Be honest about the limitations of your work and don't force the reader to read between the lines. Don't omit critical information. Other scientists may want to base their own work on your contributions, so make sure that the results you have obtained are significant and that your assumptions well-documented.

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2020-252/hess-2020-252-RC2-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-252, 2020.

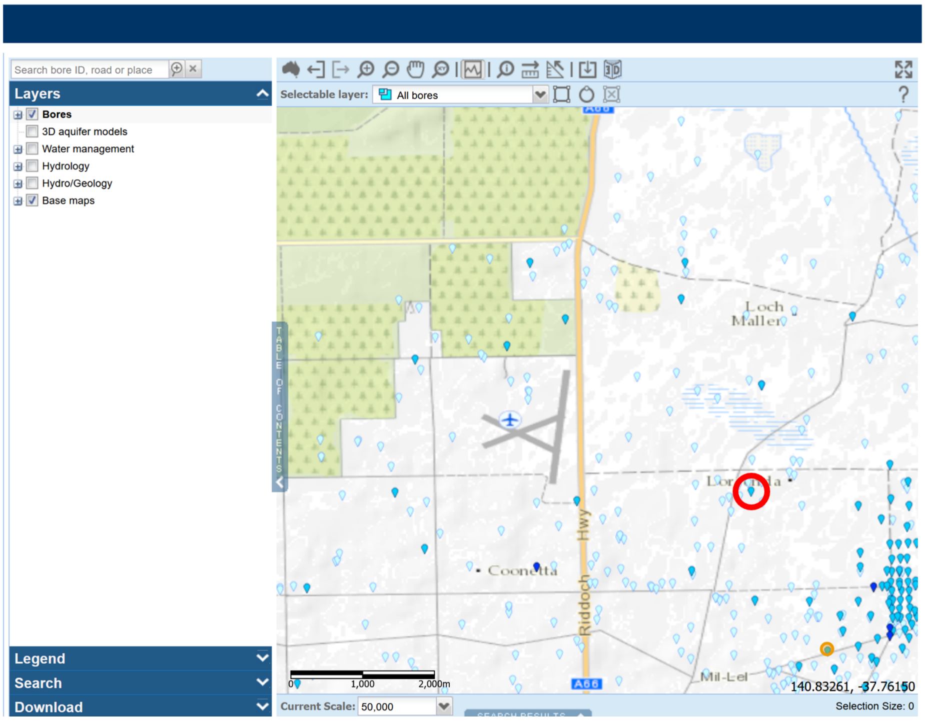
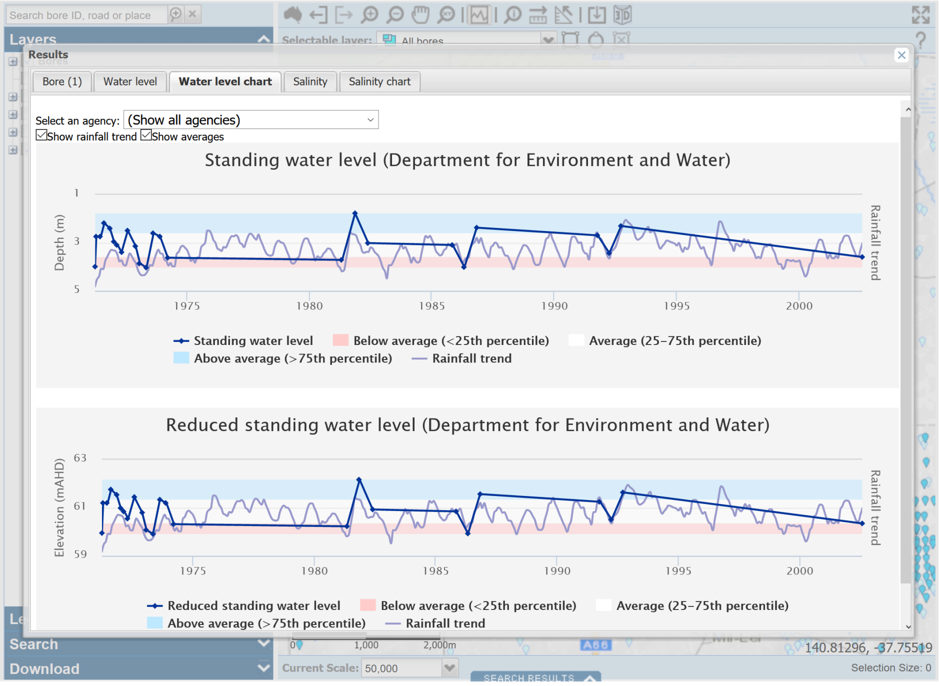


Fig. 1. Figure 1

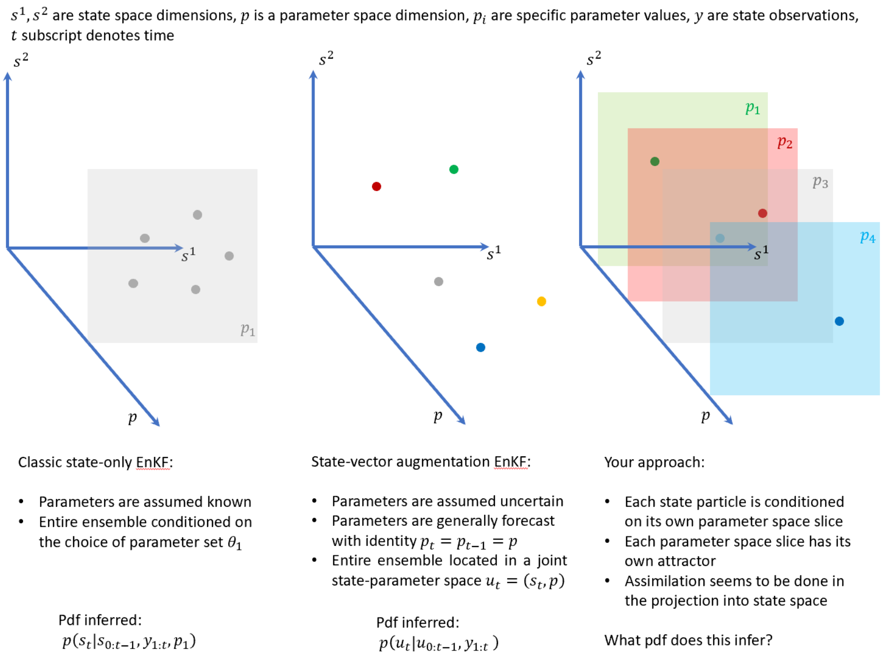

Search bore ID, road or place
Layers
Selectable layer: All bores
Results
Bore (1)
Water level
Water level chart
Salinity
Salinity chart
Standing water level (Department for Environment and Water)
Depth (m)
1
3
5
1975
1980
1985
1990
1995
2000
Rainfall trend
Standing water level
Below average (<25th percentile)
Average (25–75th percentile)
Above average (>75th percentile)
Rainfall trend
Reduced standing water level (Department for Environment and Water)
Elevation (mAHD)
63
61
59
1975
1980
1985
1990
1995
2000
Rainfall trend
Reduced standing water level
Below average (<25th percentile)
Average (25–75th percentile)
Above average (>75th percentile)
Rainfall trend
Search
Download
0
1,000
2,000m
A66
140.81296, -37.75519
SEARCH RESULTS
Selection Size: 0

Select an agency: (Show all agencies)

☑Show rainfall trend ☑Show averages

Current Scale: 50,000

Fig. 2. Figure 2



s^1, s^2 are state space dimensions, p is a parameter space dimension, p_i are specific parameter values, y are state observations, t subscript denotes time

Classic state-only EnKF:

- Parameters are assumed known
- Entire ensemble conditioned on the choice of parameter set θ_1

Pdf inferred:
$p(s_t|s_{0:t-1}, y_{1:t}, p_1)$

State-vector augmentation EnKF:

- Parameters are assumed uncertain
- Parameters are generally forecast with identity $p_t = p_{t-1} = p$
- Entire ensemble located in a joint state-parameter space $u_t = (s_t, p)$

Pdf inferred:
$p(u_t|u_{0:t-1}, y_{1:t})$

Your approach:

- Each state particle is conditioned on its own parameter space slice
- Each parameter space slice has its own attractor
- Assimilation seems to be done in the projection into state space

What pdf does this infer?

Fig. 3. Figure 3