Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-252-RC1, 2020

Hydrology and Earth System Sciences
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: 12 July 2020

1. Does the paper address relevant scientific questions within the scope of HESS? Yes

2. Does the paper present novel concepts, ideas, tools, or data? Partially

3. Are substantial conclusions reached? No (see pt. 5)

4. Are the scientific methods and assumptions valid and clearly outlined? Partially

5. Are the results sufficient to support the interpretations and conclusions? No

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? No

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? Yes

8. Does the title clearly reflect the contents of the paper? Partially

9. Does the abstract provide a concise and complete summary? Yes

10. Is the overall presentation well structured and clear? Structured: yes, clear: no

11. Is the language fluent and precise? Yes

12 .Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Yes

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Yes, see description below

14. Are the number and quality of references appropriate? Yes

15. Is the amount and quality of supplementary material appropriate? Partially (MODFLOW Input files would be an important addition)

In this study, the authors explore the coupling of different Unsaturated Zone Models (UZM) and a simple groundwater model (GWM), assimilating remotely sensed actual evapotranspiration (AET) data to improve the fidelity of their model predictions, and set out to gauge the degree of complexity required for a faithful representation of the relevant dynamics. This is a topic of high interest for hydrogeologists and hydrologists alike, particularly for those working in arid and semi-arid regions.

Unfortunately, I have severe reservations concerning the methodology, particularly the design choices made for the groundwater model and the data assimilation (DA) algorithm. Furthermore, the documentation and presentation of their methods and the corresponding theory is insufficiently detailed in some parts and simply incomplete in others. Several results of this study set off alarm bells and reinforce my doubts about the methodology: the inability of the DA algorithm to improve the mismatch of the as-

similated variable (AET) and the UZM states relative to an open loop run, the inability of the calibrated (!) coupled UZM-GWM to reproduce the groundwater dynamics with any degree of fidelity.

As a consequence, I am afraid I must recommend a rejection of the manuscript in its current form. However, as I find the fundamental concept very promising, I would encourage the authors to revisit their fundamental assumptions/choices and resubmit their revised work in the future.

The introduction of the study is well-written, informative, and supported with references. The only gripe I have with it is that in Line 78, the authors mention the validation of one their previous studies as an objective, and this objective is neither stated in the aspect nor revisited in the discussion section.

Section 2.1: This section is fine and does a good job at presenting the field site. It establishes AET as the assimilation variable in the following.

Unfortunately, Section 2.2 is where the problems begin. This section lacks critical information in several parts, and is simply incomplete in others. I will elaborate on both cases below. As a general comment, since this study attempts to explore the value of connecting groundwater models to UZMs, the study could profit tremendously from a comparison with an integrated hydrogeological model such as HydroGeoSphere (mentioned briefly in the introduction, then ignored), which already provides the coupling of both features naturally. Even MODFLOW has several unsaturated flow/ET modules (UZF, EVT). For a study which explores to which degree the fidelity of the UZ representation improves the representation of groundwater flow, failure to compare to what degree (if any) an external coupling to a UZM provides value over the already in-built features seems like a great oversight.

Section 2.2.1: 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).

Section 2.2.3: First off, normally FloPy generates the entire model input for MODFLOW (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. 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 time step 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.

Section 2.2.4:

Line 152-154: I cannot imagine that your groundwater model with only 3 (!) active cells and a single timestep between every forecast step (let alone in an ensemble with only 32 members) would have significantly greater computational cost than if you chose a finer temporal resolution. The MODFLOW simulation overhead alone should devour more computational resources than if you increased your temporal and spatial resolution a hundred-fold each. 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 list .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 by 3 km.

Line 160-163: 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?

Line 168-169: 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.

Figure 3: This figure is unclear and confusing. You attempt to show two different things at one: Your (uniform) initial conditions for Configuration-1, and the anticipated drawdown for Configuration-2, whereas the real distinction between the two configurations is the use of different UZMs. Choose one and stick with it, or separate the message into two figures. Also, since you chose an incredibly coarse spatial resolution of the groundwater model, the smooth drawdown on the right-hand side is deceptive: there is no way you can resolve that with only three cells.

Section 2.3: In this section, the authors made a number of highly questionable model design choices which ultimately call the validity of the entire study into question.

Line 173: 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)?

Line 175-176: 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.

Line 180-182: You glance over this part, but it is extremely important. While you never explicitly state anywhere what you mean by 'interdependence between WT and actual ET' (although you use the term several times), I can guess: 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). Since your groundwater model does not allow for any seasonal or regional hydraulic head fluctuations at all (you set its boundaries either to no-flow or a constant [never specified] prescribed head), your calibration must identify UZM and GWM parameters which couple both parts strongly enough together that your UZM can attempt to recreate the observed groundwater dynamics (which, as we shall see later, it does a very poor job at). I have yet to encounter a system where your fundamental assumption that the natural groundwater table is (a) uniform and (b) not seasonal (i.e., constant) would be valid, and I highly doubt that it is the case at your field site. 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.) can affect your predicted GW dynamics, and as a consequence you must inflate the degree (and, by extension, importance) of the UZM influence to have any GW dynamics at all. This assumption seems sloppy at best and fraudulent at worst.

Continuing 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. You can easily verify this in MODFLOW yourself. I have attached a simple example where I simulated a steady-

state, unconfined, 1D groundwater model with the a central 3 km wide strip on which we have a constant negative recharge, and fixed head boundary conditions on either side corresponding to the chosen grid spacing. In all three cases, the system is set up like you show in Figure 3 (a central, 3km strip of forest, a constant prescribed head boundary outside; the outermost two cells are the prescribed head boundaries). As you can see, the drawdown you obtain depends heavily on the chosen grid resolution. The finite volume approximation, just as the finite difference approximation, relies on a fine grid spacing. The coarser it is, the less valid the numerical approximation of the finite differences, the worse your representation of reality. Since you calibrated the model parameters to force a fit, I expect a heavy degree of parameter surrogacy in the optimized values.

(see the attached Figure1)

In summary: 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.

Table 1: 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.

Section 2.4: This section is unacceptable in its current form. It does not establish the theory behind the data assimilation algorithm. There is not even a single reference to Bayesian statistics (on which the EnKF is founded) in the entire manuscript. In many parts, information required to reproduce the authors' work is either just vague (e.g. Line 238 to 240: "adding a random number sampled from Gaussian distributions with different standard deviations"). The crowning part is the introduction of the EnKF – after showing how the state vectors are constructed, the authors simply refer to an

entirely different study (line 230) to complete the description of the methods. This is not acceptable. A manuscript should be able to stand on its own. You may refer to other publications for details, but not for the core of the method you are using.

Even more worrying are statements such as the one in Line 240, which suggests the authors may not have understood the theoretical foundations of the DA algorithm they are using: "The EnKF (Evensen, 1994) was used because of its ability to deal with highly non-linear systems." This should certainly NOT be the motivation for using the EnKF. I encourage the authors to revisit the derivation of the EnKF, and remember that the EnKF is a Monte Carlo approximation of the Kalman Filter. As such it is explicitly based on the assumption of linearity. This is a critical assumption of the algorithm because it guarantees that a Gaussian pdf remains Gaussian after the forecast, and the Kalman filter requires Gaussianity for its Bayesian inference. The EnKF can only use nonlinear models because it fits a Gaussian distribution to the ensemble after the forecast, but its assumption that mean and covariance are sufficient statistics for the resulting state pdf is only valid if (a) the ensemble was sampled from a multivariate Gaussian before the forecast, and (b) the forecast was sufficiently linear to preserve this property. The more nonlinear the system, the more unlikely the preservation of Gaussianity, the less justified the use of the EnKF.

Instead, the authors might consider alternative sequential Bayesian inference methods if their system is severely nonlinear. van Leeuwen et al. (2019), for example, provide an excellent overview of particle filters in geoscientific applications, which are significantly better suited for nonlinear Bayesian state inference, particularly in low-dimensional settings such as the one in this study:

https://doi.org/10.1002/qj.3551

Line 201-203: This also suggests that the authors may not have understood the EnKF. You shouldn't justify the designated ensemble size with choices in literature – the required ensemble size for any given problem depends on the dimensionality of the

inference problem, and is not just some algorithmic fudge value for which there are recommended best guesses.

Line 203-204: “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”.

Line 206-229: Here you specify aggregated state vectors composed of hydraulic head predictions and soil moisture predictions, but the data you assimilate are AET measurements – a variable type entirely absent from your aggregated state vectors. Common EnKF approaches extract the corresponding entries from the state vectors by multiplying it with a matrix of one and zeroes. If you follow this approach: how do you conduct the filtering step without the predicted AET as part of your state vector? The information in the manuscript does not allow the reconstruction of what exactly you are doing.

Line 238-240: Only after checking the referenced publication it became clear what approach the authors have used. Once again, do not force the reader to read a different publication in order to gather critical details on your methods in this study. In the previous study, the authors added temporally independent Gaussian noise to every single forcing data point. This is a well-known approach. However, the sentence in its current form suggests that the authors selected sampled a single (univariate) Gaussian distribution and added the resulting scalar as a constant offset to the entire forcing timeseries. Also, specify what you mean with ‘different standard deviations’. How large are those standard deviations? In which way are they differing? Are you using heteroscedasticity?

Line 240-241: This approach is poorly justified. You should derive statistically what exactly it is you are doing in terms of Bayesian inference when you simply ‘add noise

to the parameters’. Again, the EnKF is not some heuristic optimization routine but a statistical inference algorithm, which limits the amount of fudge you can introduce before you invalidate the method. Also, clarify how often you are adding noise: Just once in the beginning, or at every time step? Results suggest the former, but at least state so somewhere.

In general, there are three ways you could consider models with parameters in EnKF:

The parameters are considered deterministic constants. In this case, the posterior is p(x_t |y_(1:t),ΪŚ), where x_t are the latest state simulations, y_(1:t) is the time series of state observations, and ΪŚ are the static model parameters. In such a case, the parameters must remain the same for all ensemble members. Adding noise in such a case is invalid, as your ensemble members would be conditioned on different parameter sets, and you could no longer combine your perturbed particles into a common ensemble.

The parameters are considered part of an augmented state vector z_t=(x_t,ΪŚ_t), where the parameters are forecast with the identity function ΪŚ_t=ΪŚ_(t-1). In this case, the inferred posterior is p(z_t |y_(1:t)). This is the classic state vector augmentation approach, in which states and parameters are updated jointly through the cross-covariances. Adding noise in such a case is valid, but then the parameters must be updated jointly with the states. However, you mustn’t calibrate your parameters when creating the prior ensemble in this scenario, as this would be what statisticians like calling an ‘inverse crime’. Hendricks-Franssen & Kinzelbach (2008) have a nice paper explaining how to use this approach: doi:10.1029/2007WR006505. This could be a valid alternative to the calibration+inference setup you used in this study if you really want to stick with the EnKF.

You may represent model forecast error with parameter dynamics, which is a bit of an obscure but valid approach. In this case, you can (indeed: should) add errors to the parameters at every time step, but basing it on sequential applications of white noise would mean that your parameters drift uncontrollably. If this is the method you chose,

you should add noise through an Ornstein-Uhlenbeck process instead. See, e.g., this publication: doi:10.1029/2009WR007814

As it stands, it is unclear what you have done exactly, or how it is justified. It requires a lot more theory, and likely thought, to be justified.

Section 3:

Figure 4: The deterministic simulations once more cast doubt on the validity of the chosen approach. While the AET is reproduced somewhat faithfully (likely due to the strong dependence on observed climatic forcings), the hydraulic head fits are exceptionally bad, especially considering the fact that these are already calibrated results. This suggests that the groundwater model setup is inadequate to represent the local dynamics (unsurprising considering grid resolution and boundary conditions; see my criticism of the model setup above). As a consequence, the head predictions shouldn't be used gauge the validity of UZMs clearly unaffected by the hydrogeological boundary condition (compare the similar AET predictions of configuration-1 and 2 in light of their wildly different GW predictions).

Figure 5: Why do you not show the full resolution of the SM data? You have data in 30 cm increments down to 3 m – show it! It can be very interesting for the reader to see where the model reproduces the observed SM, and where it doesn't.

Section 3.2: As mentioned above, I have strong reservations that the arbitrary addition of noise to the initial particles (without explicitly representing them as random variables) is a valid approach for Bayesian inference. In essence, you are using a different forecast model for each ensemble member. Most hydraulic flow is dissipative, so prior state uncertainty should normally disappear with time – which it cannot, in your case, since the attractor is not shared between the particles (as they do not share the same parameters). It also seems not surprising that Configuration-2 has a smaller uncertainty bound than Configuration-1: Three of the parameters you perturb are MODFLOW variables, and the only UZM variable you perturb (the root depth) has

vastly different absolute values in both configurations – a choice I have questioned before – which affects the standard deviation of your perturbation and, by consequence, the forecast error. I would recommend testing what influence adding a fixed SD noise to this variable has, or setting the root depth to the same value for both configurations.

Line 312-313: This ensemble spread is entirely artificial. The difference in ensemble spread is based on a (possibly unjustified [lacking information], certainly biased) perturbation of the model parameters.

Figure 7: The assimilation of AET does not seem to improve the prediction of hydraulic head dynamics to any significant degree, likely due to the inadequacy of the GW model. The legends in subplot (c) and (d) are incomplete (red line: assimilation [mean?]). Furthermore, it looks as if the assimilation of AET measurements does not reduce the uncertainty in AET predictions in configuration-1, whereas it clearly seems to do so in configuration-1. Why is this the case?

Table 3: The improvements to the WT RMSE are unlikely to be significant. First and foremost, the reduction in RMSE of AET – the only data you actually assimilate – is virtually negligible in both cases. Since in the EnKF, WT predictions are updated only through the cross-covariances between AET and WT predictions (assuming it is only AET data you assimilate; your methods section is too incomplete to permit a clear reconstruction of your work), the resulting updates should be only minor. It certainly seems strange than the influence on WT RMSEs is so significant. A possible explanation might be that three out of four perturbed parameters are MODFLOW parameters simply result in a larger WT uncertainty, artificially amplifying the effect of even minor AET changes. This does not seem like a stable and particularly well-founded approach.

Line 373-375: You have not introduced what you did in Gelsinari et al. (2020) in this study, so stating that you 'consolidate[d] the synhetic approach' is something the reader has to believe without evidence.

Figure 10: This figure borders the comical, and provides further evidence that there

is something severely wrong with your model and/or DA algorithm. Even though AET is a product of your UZMs (and hence probably dependent on the other UZM states), the assimilation of AET data has virtually no influence on the fit of your UZM states. This can happen if your AEM uncertainty has almost completely collapsed, but at least Figure 7c clearly has residual uncertainty, which should be sufficient to significantly improve the AET RMSE during the update step.

4 Conclusions:

Line 384 to 386: The part under 'Calibration' lists things which you simply stated without proof (see Line 180-181). This conclusion does not result from the information you provided.

Line 387-401: In summary of the issues discussed above, I do not believe that improvements to the WT predictions are significant. The groundwater model assumptions are questionable, the different UZM differ seemingly arbitrarily (with the theoretical foundations of UnSAT-UZM not being sufficiently well explained). Furthermore, the DA algorithm failed to improve the fit of observed and unobserved UZM states to any significant degree. This is unsurprising in the SWAP-UZM, whose AET uncertainty collapsed (Figure 7d), and is inexplicable in the case of the UnSAT-UZM, whose AET uncertainty should have allowed significant Kalman updates (Figure 7c). All in all, it seems 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. The degree of fudging decisions in this study make it extremely hard to judge how reliable these results are.

Line 402-405: This is a strong claim, and one which has also not at all been backed by data from the results section. An easy way to check this claim would be to show the correlations of the ensemble covariance matrix, particularly those of all UZM and MODFLOW state variables with the AET. But even then, with the questionable methodology choices above, it is all but impossible to interpret the results with any degree of

confidence.

Line 406-412: It sounds like this conclusion was written before the data were obtained, as it is not well based on the results you obtained. SWAP-UZM clearly does a significant better job at reproducing SM observations (particularly in the deeper layers), and the AET effect on the WT predictions seems unrealistically amplified. I don't think that you can derive any generalizable conclusions from this study with the current, questionable methodology. As a consequence, I must recommend a rejection of the manuscript.

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

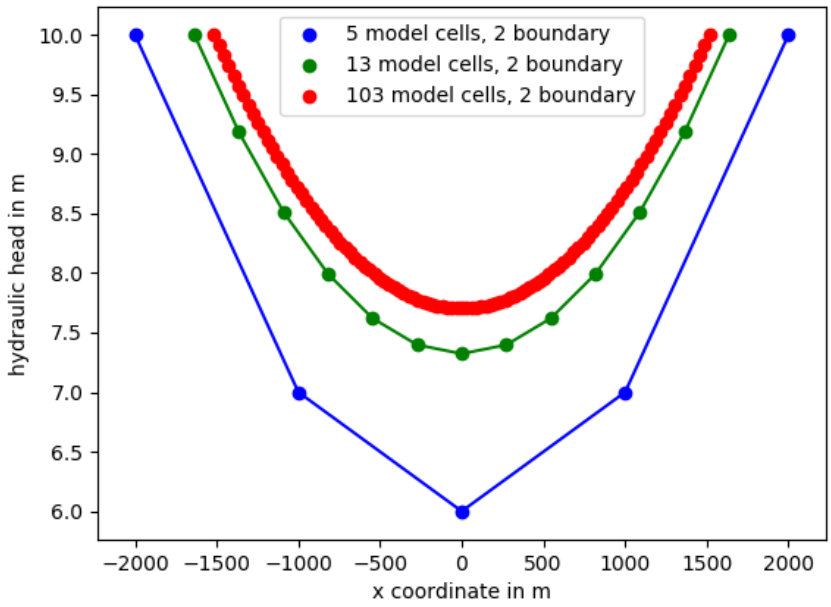

Fig. 1. Figure1