Reviewer #1

MAIN COMMENT

The paper discusses the interesting issue of the difficulty to identify simultaneously hydraulic conductivity and recharge from piezometric head observations. For this purpose the authors use the ensemble Kalman filter and build a synthetic experiment to conclude that unless the prior information used to generate the initial set of realizations is "correct" it is impossible to identify simultaneously both hydraulic conductivity and recharge.

Although the authors prove nicely their point, the arguments given to justify the final results are flawed: the emphasis should not be put in the prior information but rather in the need of extra information to be able to single out the combination of hydraulic conductivity and recharge that is correct out of the many combinations that are coherent with the observed piezometric head.

It is wrong to say that the prior information used to generate the initial set of ensemble realizations determines what the final estimates will be (after data assimilation) and that you cannot generate realizations outside the prior random function model. The final ensemble of realizations can largely depart from the initial one, when the observations are inconsistent with the prior model. The sequential set of ensembles that are obtained after each assimilation step can be interpreted as a Markov chain that will "forget" the structure built in the initial ensemble after some assimilation steps. More so, the updating of the ensemble realizations is solely based on the covariance structure of the ensemble, and for this reason, there is a tendency for the final ensemble of estimates to converge towards realizations drawn from a multiGaussian function, even when the initial ensemble is far from being multiGaussian. Enough observation data can change completely the random function of the final ensemble with regard to the one used to generate the initial ensemble.

What happens in the example presented by the authors is that the observations are consistent with all prior models used. In fact, the reference could be the one used, or it could be a realization generated with the "wrong" model, the results would be the same. Therefore, you need some additional piece of information, to discriminate among the different prior models which one is the one consistent with your unknown reality. It is not that the EnKF does not work when the prior model is incorrect. Knowing which is the orientation of the deposition will draw to prefer a prior model over another.

My request is that the paper be rewritten removing all these comments about the importance about the prior information, and the influence that this prior information has in the final ensemble, and replacing it by talking about the importance of having additional information that would allow you to discriminate among alternative models. I contend that assimilating data on fluxes or concentrations would also alleviate the problem of the prior model. And I insist that the prior model information will fade away as time passes and data are assimilated, and could vanish if the prior model is inconsistent with the observational data.

ADDITIONAL COMMENT

I do not think there is any need to present the extended Kalman filter equations, especially when they are not used to justify the use of the ensemble Kalman filter. In this respect, there are some conceptual misunderstandings about the ensemble Kalman filter that must be corrected. First of all, there is no need to make any multi-Gaussian assumption to get to the ensemble Kalman filter equations, like there is no need to make any multi-Gaussian assumption to get to the cokriging equations; it is true that under the multi-Gaussian assumption the ensemble mean and ensemble covariance would be the mean and covariance of the conditional distribution given the obsevations; however, from the point of view of optimal estimate in the a least-square sense, the Kalman filter equations do not need any multi-Gaussian assumption.

We would like to thank the reviewer for his/her extended review, and we are happy that she/he overall has a positive image of it. In this reply we will start by addressing the major comment together with the first of the additional comments, which we believe are strongly related to each other.

The reviewer's points may be summarized as follows: 1) The reviewer is convinced that formal priors, as considered in the manuscript, are not essential, they are only used to initialize the filter and given enough informative data, the prior knowledge will fade away, 2) the reviewer points out that Gaussian assumptions are not needed to use the EnKF, there are enough informative data, 3) following 1 and 2, (s)he concludes that we obviously have too little data to perform our EnKF analysis and should thus focus on evaluating which additional data we would need to overcome the issues exemplified in the manuscript rather than focusing on the importance of prior knowledge.

Obviously, we and the reviewer have a fundamentally different view on parameter estimation in general, which roots in a decades-old debate on the validity of the Bayesian paradigm, known as 'Bayesian vs Frequentist debate'. Honestly, we don't think it would be beneficial for the manuscript to continue this debate. By this we don't want to imply that the frequentists' view is necessarily wrong, but we clearly take the Bayesian standpoint, and will not leave it. We find, among other reasons, the Bayesian framework to be refreshingly transparent when it comes to which assumptions are being made throughout the estimation process (in difference to making more subjective assumptions, such as selecting a specific objective function). As our consideration in this manuscript is the importance of relevant prior, the Bayesian framework is elegant, as it offers a way to compensate for non-unique data by the use of a strong prior.

The major point to be made clear is: groundwater-head data, which we consider in this study, are non-unique with respect to unknown conductivity and unknown recharge. This is derived and exemplified in Section 2 of the manuscript and this non-uniqueness is also the core of our manuscript. We do not disagree with the reviewer that different types of data would help improve the parameter estimation. On the contrary, we share those views. However, in subsurface hydrology the data scarcity is usually a great problem. Fluxes, as suggested by the reviewer, cannot be measured as such; conductivity measurements are, if existing and trusted, very local; tracer tests are time consuming; age tracers may be costly and require very long simulation times. Head measurements are common and trustworthy measurement, and also in the literature (as cited in the Introduction) the observation setup used by us is not uncommon. Hence, our example is rather realistic for a real world scenario in which we want to estimate aquifer parameters! If anything, we have unrealistically many observation points. The main difference between the Bayesian standpoint and the standpoint suggested by the reviewer is that without a formal inclusion of the prior, we require fully informative data to perform parameter estimation. This is, as rightly pointed out by the reviewer, not what we are doing.

The reviewer points out that “need of extra information to be able to single out the combination of hydraulic conductivity and recharge that is correct” is important. However, this is also a form of prior. If we have this extra information we are, in the view of our manuscript, in the Good (correct) prior case. As such, we do not see that there is a large difference, in practice, between the suggestions of the reviewer and the approach and aims stated in the manuscript. The bottom line of the manuscript is that we need to be careful to have correct prior information. How to do this is not the topic of the current manuscript, nor do we wish it to be.

Many practitioners would argue that we should simplify the model, that is, assume fixed zones of uniform hydraulic conductivity and recharge. While we agree that this can “fix” the problem of non-uniqueness, restricting the solution space by hard-wiring the structure of the subsurface is the most extreme prior to be thought of, as it does not allow any uncertainty in the identified structure itself. That is, the problem does not really vanish, it’s just hidden.

Concerning the statement that the prior caries over, we again agree with the reviewer. If the data is fully informative, we can also estimate values outside of our initial sample. Similarly, the prior could in such cases vanish as more data come in. As pointed out above, the data is rarely fully informative in the subsurface-flow applications. However, the original manuscript was probably not clearly enough formulated that we throughout consider the case of non-unique data (hence, using head observations). In the revised manuscript this will be revised and made clearer.

To summarize, we see the requests made by the reviewer as a wish to alter our philosophy of parameter estimation. We believe, by contrast, that the Bayesian framework is a legitimate way of introducing and interpreting Ensemble-Kalman approaches. Within this framework, studying the importance of the prior is highly relevant as the prior is at the very heart of Bayesian analysis. Thus, we will not rewrite the manuscript to coincide with philosophical viewings of the reviewer, who obviously thinks otherwise. However, in the revised manuscript we will take great care to make clear that the data considered are non-unique, and will surely point out that other types of data could help resolving the issue of the incorrect prior.

I do not quite understand the last paragraph in page 5576 when the authors say that the EnKF is a linearized estimate that is alleviated by the repeated application over many time steps. The authors should understand that the EnKF captures the linear relationship that there is between the parameter and state variables through the experimental covariance –that’s all–, the fact that you apply the updating equations over many time steps does not "alleviate the effects of non-linearity". The reason why the EnKF works and the extended Kalman filter did not is because the covariances are computed on parameters and states which have been obtained by solving the state equation through an ensemble of realizations, and therefore are much closer to the "true" covariance than the one obtained by propagating the initial covariance in time through a linearization of the state equation.

The section referred to was not particularly well formulated. What we really mean is that when we have many observations in time we make good use of the filter dampening. This slows the filter down (hence we may need more temporal observations to reach a stable result) but also avoids making too large jumps in the parameter space. This can be beneficial when the non-linear relations between parameters and observations cases erroneous updates, and as such it is helping to alleviate the effects of the linearizations. We see that this was not clear within the text and in the revised manuscript the full section will be thoroughly revised.

Please revise your presentation and discussion of the EnKF.

While we are willing to explain steps more clearly, we will keep the EnKF within a Bayesian framework, with all the consequences.

OTHER COMMENTS

Since the state (piezometric heads) is not updated, you should explain how the state is computed after each assimilation step. Is the model rerun from time zero with the updated parameters?

We do not understand why the reviewer believes, that we do not update heads, which, in fact, we do.

Page 5577, line 2, the original prior knowledge is smeared out after the first assimilation step by the Kalman gain, it carries over at the beginning but it will eventually disappear.

See answer to general comments.

Page 5577, line 8, if the prior knowledge is erroneous and there are sufficient observation data inconsistent with that prior model the estimates will converge to the "truth".

With enough informative data we agree with the reviewer. This was maybe not clear in the original submission and will be improved in the revision. However, for the data in question (groundwater heads), we cannot expect to recover any pattern that are not part of the prior. For a better reasoning on this question, please consider the reply to the major comment.

Page 5578, why is NRMSE only computed at the observation wells and not over all the aquifer, since you have the reference information?

The NRMSE could, as the reviewer suggests, also be calculated for all grid cells in the model. The approach to calculate the error given the observation locations available is chosen to represent the type of information available in a real case, for which one would of course not know the head value in each cell. In the end we also conclude that the NRMSE on its own is not a sufficient metric to judge which EnKF setups are good or bad, which we find an illustrative example of a problem that can occur in any real parameter estimation setup.

Page 5578, it is unclear what is the denominator of the equation, is it the ensemble variance? or is it the prior measurement uncertainty (in which case it is a constant value)?

Here we mean the uncertainty of the measurement (which indeed is considered a constant). This missing information will be added to the revised manuscript.

It would have been nice to see some variance maps along the ensemble mean maps.

This is a useful suggestion that we intend to pick up in the revision.

Top of 5583, in real applications you need information to discern among alternative combinations of conductivity and recharge, this information could help you in choosing the prior model, or it could be other types of data (such as fluxes).

Yes, we do agree with the reviewer. This will be better highlighted to the revised manuscript.

End of 5583. No, multiple-point statistics will not help here, that is, it will not allow you to discriminate between a good and a wrong model as long as the observation data are consistent with those models. Besides, it has been proven that the EnKF will filter out the non multi-Gaussian characteristics of the initial ensemble of realizations.

To a certain extent this is right: if we have the wrong training images, we will be back to square one again (wrong initial sample). However, what we wanted to express was that the use of training images is a slightly more advanced way of introducing prior information than generating multi-Gaussian fields.

SUMMARY

I liked the paper and I think it should be published, but only after the emphasis in the conclusions is shifted.

We are glad that the reviewer likes our work; the focus of the paper has been discussed above.