Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-431-AC4, 2018

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Hybridizing Bayesian and variational data assimilation for robust high-resolution hydrologic forecasting" *by* Felipe Hernández and Xu Liang

Felipe Hernández and Xu Liang

feh17@pitt.edu

Received and published: 19 January 2018

We would first like to express our appreciation to both Referee 1 and Referee 2 for their careful and thorough review of our manuscript. Their comments will certainly help us improve the quality and clarity of our manuscript. Below we offer our response to Referee 2.

Anonymous Referee 2 on 21 December 2017: *"Before beginning review of this manuscript, although not mentioned in the text and reference, this should be considered a re-submission of the previous HESSD manuscript entitled "Hybridizing sequential and variational data assimilation for robust high-resolution hydrologic forecasting*

(https://doi.org/10.5194/hess-2016-454)" by the same authors, which was rejected in 2016. I suggest the editorial board compare the final revision of the previous HESSD manuscript with the current one if the track record was not screened yet. I cannot examine whether the authors submitted the final revision in the previous submission in 2016 or not. However, if so, improvement and uniqueness of the current manuscript over the rejected final manuscript should be carefully evaluated. In addition, since HESSD is independent publication, the previous manuscript in 2016 should be cited and discussed in this manuscript."

This point had been discussed with the editor, to whom we expressed that we will adhere to the guidelines required by the journal and the editorial board. This manuscript is a revised version of the cited one in 2016, which takes into account the comments made by the referees back then and by the previous editor. A detailed account of the changes made was submitted to the journal. The 2016 manuscript was rejected by the editor because he considered that the required modifications deserved a more careful timeframe than the one available for the special issue it was submitted to. No final version of the 2016 manuscript was submitted.

"This manuscript proposed a hybrid DA method, OPTIMISTS, combining sequential and variational methods, and compared performance of developed methodology over PF and VAR using distributed hydrologic models. The topic is of interest to a wide range of hydrologic modelling community. The strategy of the proposed methodology to leverage different DA approaches, sequential and variational DA, is one of the important trends in recent studies. However, there are major gaps in experimental setup and evaluation, and incomplete reasoning in new methodology which require significant changes before publication. I hope the followings would be helpful to improve the quality of manuscript. 1) Evaluation period and methods In this manuscript, the total evaluation period is 10 weeks (5 cases with a 2-weeks period each): 3 scenarios for 2-weeks forecasts in the Blue River and 2 scenarios for 2-weeks forecasts in the Indiantown Run, not including assimilation period. The evaluation period for hydrologic

modelling and data assimilation is usually longer than at least 6-8 months and up to multiple decades. The total 10-weeks forecasts (2-weeks piecewise each) and associated metrics cannot be accepted as a rigorous evaluation. Given that the selected events in the Blue River in the 2016 manuscript are different from those in the current one, there seems to a potential to further increase evaluation period. In Table 3, the authors also mentioned calibration periods are 85 and 20 months, respectively. Considering the availability of observation data, what is the maximum evaluation period for two catchments? Why don't you use the whole or most calibration period for DA evaluation? Was there any reason to use the limited period for evaluation? For the larger domain, the Blue River catchment, is there just one streamflow observation gage over 3,000 square kilometer area? Why don't you assimilate observations in multiple locations to reduce equifinality and overfitting? In this study, evaluation metrics were estimated for the whole 2-weeks forecast period. However, it is more common to evaluate metrics for varying forecast lead times because the impact of updating varies and disappears over time. I highly suggest the evaluation period and method should be reconsidered to qualify a kind of general standard shown in many forecast and DA-related papers: simulating more than several months for each catchment and evaluating metrics for varying forecast lead times."

Thank you for the good suggestions. We will evaluate our method based on the suggestions here in the revised manuscript. We already extended our scripts to allow running an extended-time data assimilation experiment where assimilation is performed with OPTIMISTS continuously to allow producing multiple time series of forecasts with a fixed lead time for the Blue River. We should be able to produce these forecasts for multiple months in order to analyze the performance of the algorithm. We will similarly develop a script to run the particle filter in the same fashion and to be able to compare its forecasts with those of OPTIMISTS.

On the other hand, this new experimental setup will preclude the comparison with the 4D evolutionary variational algorithm for the reason that we were using the same prior

"particle" ensemble to seed its population that the one used for the other methods. However, given that 4DVar is inherently a deterministic approach, such ensemble will not be able to be updated to be the seed of continuous assimilation periods. In practice, variational methods compensate the lack of an ensemble by performing a guided search of the initial state solution space until convergence, but in this case we consider that the model simulation quota that we are allowing each of the methods will not suffice to reach an optimal solution. Moreover, evolutionary variational methods are rare in the literature (more so in operational settings) and therefore we now consider that the comparison would not be of enough significance. While ideally we would like to compare OPTIMISTS with proven 4DVar methods, these require the linearization of the model's dynamicsâĂŤwhich is rare in hydrology, would require an enormous amount of additional work, and it is outside the scope of this paper.

Also, while we will implement this new evaluation scheme for the comparison of OPTIMISTS with the particle filter, we will still maintain the scenario-based design for the analysis of the parameters of the assimilator. Even though not "rigorous," we are confident that the variability in the scenarios considered is enough to differentiate relatively adequate configurations of the algorithm. As discussed in the manuscript, we will conduct future tests for cases where there was not enough statistical evidence to conclude one configuration was better than the other. Extended-time evaluations would, in this case, require an extensive computational budget given the large number of parameter combinations selected.

We agree that assimilating data from additional streamflow gages in the Blue River would allow for improved forecasts but, for the purposes of this manuscript, the assimilation of a single measurement provides a balanced challenge that enables the analysis of the strengths and weaknesses of OPTIMISTS in contrast with other methods, and the determination of an adequate set of parameters. That said, we have actually already worked on testing the algorithm using distributed high-resolution observations in a watershed and look forward to include such analyses in a future publication at its due

time.

"2) Probabilistic evaluation Although the proposed method is a stochastic approach, probabilistic metrics were not measured and analyzed. At least, basic metrics such as reliability, CRPS, predictive QQ plot and Brier score should be compared over the conventional method such as PF. Without such evaluation, improvements and features of the hybrid ensemble method cannot be understood in terms of stochastic perspectives. In addition, Figs 5 and 8 (streamflow hydrographs) should include traces or spreads of ensemble for visual inspection."

Indeed, probabilistic evaluation is very important to determine if forecasts are over-confident or under-confident. We will include an evaluation metric that allows comparing the confidence of forecasts between OPTIMISTS and the particle filter. We will also modify the plots to show the temporal evolution of the distributions in the revised manuscript.

"3) uncertainty specification on hydrologic DA In order to apply DA for hydrologic modelling, uncertainties for states and observations should be carefully taken care of. Sometimes, not surprisingly, noise configuration or specification may significantly affect DA performance. However, there is no description on how uncertainties of different state variables and observations such as interception, snow, soil moisture and streamflows were formulated and implemented for hydrologic ensemble modelling, which should impact DA process to generate ensemble, optimize state variables and estimate likelihood or weight. A detailed description is required for reproducibility of this study. Regarding this issue, for example, how different particles of distributed hydrologic models are generated in "the sampling step" of this DA algorithm? More specifically, how high-dimensional model states are being perturbed to avoid sample impoverishment in this step?"

As explained in subsubsection 2.1.4 in the original manuscript, any numerical objective can be used to judge candidate particles in OPTIMISTS. The likelihood of simulated

outputs given the distributions of the corresponding observations is cited as an example. In such a case, the user would require specifying how the likelihood is computed based on how the error of the observations is being modeled. However, the error metric used in our tests (the mean absolute error, page 12, line 4) is a deterministic one. While this constitutes a departure from the Bayesian theoretical framework, the estimation performed in OPTIMISTS retains its probabilistic character due to the way in which samples are generated and, especially, due to the proposed probabilistic interpretation of the resulting Pareto front (page 7).

Uncertainties in the state variables are all captured by the use of kernel density probability distributions, which is the whole focus of subsection 2.2. The details of the implementation are not introduced in subsection 2.1 but saved for subsection 2.2 because OPTIMISTS offers a modular design in which any type of non-parametric (ensemble-based) probabilistic representation could be used. How new samples are generated from the prior distributions (the core mechanism to "perturb" the ensemble) and how the likelihood of samples given these priors is computed is all explained in this part of the manuscript. While this arrangement was announced in page 4, lines 12-15, we will add reminders on subsubsections 2.1.2 and 2.1.4 in the revised manuscript to make the presentation clearer.

"4) Under-simulation or filter degeneracy in assimilation step In the analysis or assimilation step which corresponds the first 2-weeks in Figs 5 and 8, under-simulation or filter degeneracy (scenario 3 in Fig. 5 and scenario 2 in Fig. 8) is found. Usually, whatever filter is used, traces of simulated states (here streamflow) overlap observations in the assimilation step since uncertainty of observation is set smaller than that of state variables. It is common that NSE values of the assimilation step or the first forecast step are higher than 0.9 – 0.95. However, a large gap between simulation and observation exists even in the assimilation step, which should be clearly diagnosed and discussed."

There are several reasons that might explain the relatively low level of agreement seen

between the observations and the adjusted ensemble during the assimilation period. In the first place, it must be noted that the models do indeed have considerable errors, probably mainly in their structures, that prevent them from faithfully replicating the observations at every time step precisely. This is especially apparent in Scenario 3 for the Blue River and both scenarios for Indiantown Run, in which there appears to be conflict between fitting the peaks and fitting the drier inter-peak periods. While both models underwent parameter calibration processes, as documented in subsection 3.1 and in Table 3, no attempt was made to optimize the models' structures (e.g., equations, missing phenomena, resolution, connectivity, etc.). The calibration process, similar to the assimilation, was based on multiple objectives and not only on the maximization of the NSE: we also used the relative error which is more sensitive to errors during dry periods than those during peaks. There is also a telescopic effect of the NSE, according to which, computing it over long periods of time yields higher values than when computed over short ones: for example, if the Indiantown Run model had an overall NSE of 0.81 during the entire calibration period, zooming in on a specific month would result, in average, on a reduced rating. This effect is compounded with the relatively short period of time used for assimilating data and performing forecasts. Finally, a comparably "poor" performance during the assimilation period was also observed for the particle filter and the variational algorithm. With these, we do not find the results to be especially concerning in this regard and, on the other hand, consider that all the provided contrasts are valid given that these conditions were uniform in all cases. In fact, these "defects" reflect the current state-of-the-art challenges in the operational forecasts and it is one of the objectives that we all try to improve from different aspects/angles. We will, however, include a few words in the manuscript regarding these low fitting scores.

"5) Comparison of posteriors of state variables What potential readers want to see in the result section may be not only comparison of NSE at the outlet location. The authors need to address why and how their DA method can improve over the conventional ones in hydrologic forecasting from perspectives of distributed modelling. A compar-

ison of posterior distributions of state variables updated by the new and conventional methods may be useful to show how and why the new DA works for high dimensional applications. Especially, given that the authors urged OPTIMISTS employed essential features from but outperformed particle filters, a comparison of posteriors between two methods is also required to demonstrate whether non-Gaussian and multi-modal distributions are preserved or not."

We will include probabilistic time series of average soil moisture for forecasts produced both by OPTIMISTS and the particle filter and perform the corresponding analysis. However, we plan to perform detailed analyses of OPTIMISTS' capability of estimating soil moisture, and not only aggregated outputs like streamflow, in a later investigation (when such observations are available). For this study, due to the data limitations at the test watersheds and the length of the manuscript, the distributed comparisons won't be carried.

"6) Evaluation and optimization steps for hydrologic modelling It is not clear how the cost function is formulated for distributed hydrologic models. The authors need to show explicitly how multiple spatially-distributed state variables and associated uncertainties are taken into account to formulate the cost function in evaluation and optimization steps."

As explained in page 12, lines 4-7 in the original manuscript, one or two objective functions were used for our experiments: the mean absolute error given the streamflow observations and the likelihood of the particle given the prior state distribution. These objectives can be seen as analogous to the "cost function" used in variational data assimilation, and their equivalence is established in subsubsection 2.1.4. The likelihood is computed using either equation 8 or 9 depending on which type of kernels are used for the state variable distribution. These distributions encode the spatial variability and relationships between state variables in all cells of the model, so the likelihood is thus a measure of how well a candidate particle conforms to the values and (spatial) patterns in the prior distribution. Again due to the limitations of spatial data availability, such

evaluations are not directly carried out in this study in a spatially distributed fashion, but indirectly evaluated through the integrated quantity of streamflow.

"7) Tuning hyper-parameters There are numerous hyper-parameters such as time step, objectives, no. particles, optimization, Wroot and Kf-class, Psamp and g, related to this DA method which may increase uncertainty and subjectivity of forecasting. However, analysis methods and results on hyper-parameters shown in Figs. 3, 6 and7 are still confusing and do not provide well-organized understandings. A summary or guideline is required for proper range or values of hyper-parameters."

We acknowledge that using factorial experiments is not a common practice when evaluating the hyper-parameters of these kind of methods. We will revise our presentation of the results to attempt to convey their significance in a more understandable and clearer way. This will include the suggested summary of guidelines for potential OPTIMISTS users to parameterize the algorithm to better fit the needs of their specific application. We will possibly remove Figure 4 which introduces a format different from the other boxplots but that does not provide many significant insights.

"8) Terms and units In Table 3, use of two different units for one variable (m3/s and l/s) is not recommended."

We will change the table to use unified units to m3/s.

"Throughout the manuscript, the term 'time step' is used to represent 'assimilation time step' or 'assimilation window'. Since the time step usually stands for a temporal increment for numerical schemes, 'assimilation window' or 'analysis window' may be more appropriate to avoid possible confusion."

It is true that the term "assimilation time step" might be confused with the model time step, but it is still necessary to use it because the "assimilation window" or the "analysis window" would refer to the entire period of time in which data is assimilated prior to performing a forecast. As explained in the original manuscript, OPTIMISTS allows

dividing this time window into "time steps" of arbitrary length at which the main loop of the algorithm is executed. These time steps can be as short as the model time step ("sequential," like with particle filters) or as long as the assimilation window (like in 4DVar). The assimilation or analysis window would correspond to two weeks in most of our examples, and the assimilation time step varies from one day to two weeks in the Blue River case or from one hour to four weeks in the Indiantown Run case. We will review the entire manuscript in search for places in which the distinction between model and assimilation time steps could be made more apparent to avoid confusion.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-431, 2017.