Authors' reply to Short Comment by Reviewer 1

We thank the reviewer for their time and helpful suggestions in reviewing our paper. We provide responses to each individual point below. For clarity, comments are given in italics, our responses in plain text.

This paper has applied the Retropsective Ensemble Kalman Filter (REnKF) to assimilate hourly stremflow into the TopNet model and compared its performance to the Ensemble Kalman Filter (EnKF). Modeling results were presented to illustrate the impact of incorporating a Lag Time parameter to account for the time of concentration, the time taken for the watershed to respond to precipitation event. As a result, this paper presents no new methods but an assessment of existing methods in several regional watersheds and findings on the rationale for different accuracy levels in the REnKF and the EnKF. The authors have provided detailed information to reproduce their experiment, but the grammar and coherence of sentences should be improved in some sections.

We are glad that the reviewer felt we provided sufficient data to reproduce our method. Our aim in the paper was to provide information and guidance for other groups wishing to apply a Kalman Filter approach that explicitly takes into account the catchment lag time. This is in common with many other papers on the subject of data assimilating and/or Kalman filters in hydrology, that assess the suitability of different methods in different hydrological contexts. In fact, there are differences between our iterative implementation of the Retrospective filter and the set-up of Pauwels & De Lannoy (2006) which we will explain clearly in our revised paper, in response to the Short Comment of Yuan Li.

a). The comparisons made between REnKF and EnKF are based on flow estimations at the update time step; but it is important to examine the accuracy of the estimations for different lead times, e.g., 6-hour, 12-hour, 24-hour ahead, etc. In other words, how will the two methods perform for future time steps? This is important for operational streamflow forecasts.

Thank you to the reviewer for this suggestion. In our revised paper we will analyse forecast lead time as a factor in forecast performance.

b). The evaluation of the estimated streamflow should include the percentage bias, which gives information about the proportion of the observed streamflow that is in error or not predicted. It is widely known that large streamflows have a distortional impact on Nash Sutcliffe score, the inclusion of percentage bias is an important volumetric measure.

Thanks also for this suggestion. In our revised paper we will include alternatives to the Nash Sutcliffe score to present a more rounded view on model performance.

c). The conclusions drawn in 'sensitivity to error parameters' section is inadequate. To test for sensitivity for your chosen parameters, you need to evaluate for each parameter the distribution of its values using the ensemble members for the entire assimilation time steps and across the various watersheds. That way, you can determine if there are differences or commonalities between watersheds for each parameter. Additionally, the authors try to isolate the spread in the ensemble predictions to individual model states. But the evaluation procedure is not comprehensive: these

evaluations require extended time periods across the various watersheds than as shown in the results.

The error parameters that we referred to in the section on 'sensitivity to error parameters' were the fractional error parameters which determine the size of the perturbations applied to the ensemble members at each time step. As such, they are pre-specified and do not vary across time steps. We agree with the reviewer that we can improve the evaluation procedure of the effect of these parameters on ensemble spread. We are in the process of running additional sensitivity analyses to study the spread over extended time periods, and will include the results in our revised paper.

d). Page 9540, lines 8-10: it is okay that you use an ensemble size of 50 but the reason is vague. The ensemble size is usually chosen to adequately represent the number of parameters, and to sufficient for your problem.

We will make the reason clearer, as the Clark et al., 2008 study cited found that 50 parameters was sufficient when using the same model structure and the same number of parameters.

e). Page 9542, lines 13-16: did you mean the M was chosen to be 1, which is same as the observation time step?

Yes we did mean that, although we will remove the section from the revised paper as the option to specify a stride parameter was not used and adds confusion.

f). Page 9544, lines 3-5: In other words, model parameters were not modified between assimilation time steps?

Yes, the reviewer is correct to say we did not modify the parameters between time steps. In our study, we allow for model error by perturbing the model state variables rather than the parameter values. We will state this explicitly in the revised paper.

g). Page 9547, lines 4-5 and Figure 6: the 'spikes' it was referred to in the EnKF mostly occur when model states for distorted by the observation data from previous assimilation time step. This distortion occurs when the observation data overwhelms the update time step where the observation drives the update towards itself. The REnKF in this case, provides a modulating/balancing effect to minimize the error.

Thank you to the reviewer for providing additional interpretation of the spikes/instabilities found when using the EnKF. We will use some of this wording in our revised paper.