We would like to thank the reviewer for their positive feedback and suggestions. All of the reviewer's comments have been addressed below. Please find the response in blue.

- Please provide a summary table for the input dataset used in this study and its specifications.

Observation Data	
Variables	- Streamflow/discharge (m^3/s) - Station IDs - Time
Source	United States Geological Survey: https://waterdata.usgs.gov/nwis
Number of Stations	107
Time period	September 1 - October 15, 2018
Frequency	Gauge dependent
Formats	Binary data gathered using dataRetrieval R package; converted to DART-style hourly observation sequence files

We constructed the above table in response to the reviewer's request. This exercise helped us improve and clarify some of the language used to describe the observations in Section 2.7 and also throughout the paper. However, we have chosen to not include the above table as part of the manuscript as the data are relatively simple to understand and we believe they are sufficiently summarized in the text.

- There are so many acronyms used in this paper; please try to avoid this.

The reviewer is right, we may have used a lot of acronyms in this paper but most of these refer to model or software names (e.g., WRF-Hydro, DART, Noah-MP), datasets (e.g., NHDPlus, RAP, HRRR), centers (e.g., NOAA, NWIS, USGS) or techniques and metrics (e.g., DA, M-C, RMSE). All of these acronyms are widely used in the literature and avoiding them makes the text quite long and cumbersome. Only few new acronyms were introduced in this study: ATS, PR-inf, PO-inf and PP-inf. We appreciate the reviewer's comment and we'll try to avoid excessive use of acronyms in future work.

- Why did the authors choose the Muskingum-Cunge Streamflow model to route the flows? For example, this method has limitations in backwater effects, flood plains storage and interaction of channel slope in hydrograph. Several dams and reservoirs are present in this region, and therefore, the backwater effect might affect the flow routing.

The choice of the Muskingum-Cunge streamflow model was made as part of the design of NOAA's National Water Model. In this paper, we investigate applying data assimilation to this particular model. A streamflow routing model with backwater effects and linkages to floodplains and lakes would indeed provide a more realistic simulation, but would be difficult to run at the scale of the NWM. The increased physical realism may provide additional challenges for instantaneous data assimilation which are not considered in our application. We focus on the streamflow routing model used in the NWM to be relevant to that model.

- If there is the streamflow's baseflow underestimation, how the NWM model solved this issue? Please clearly explain in the methodology section.

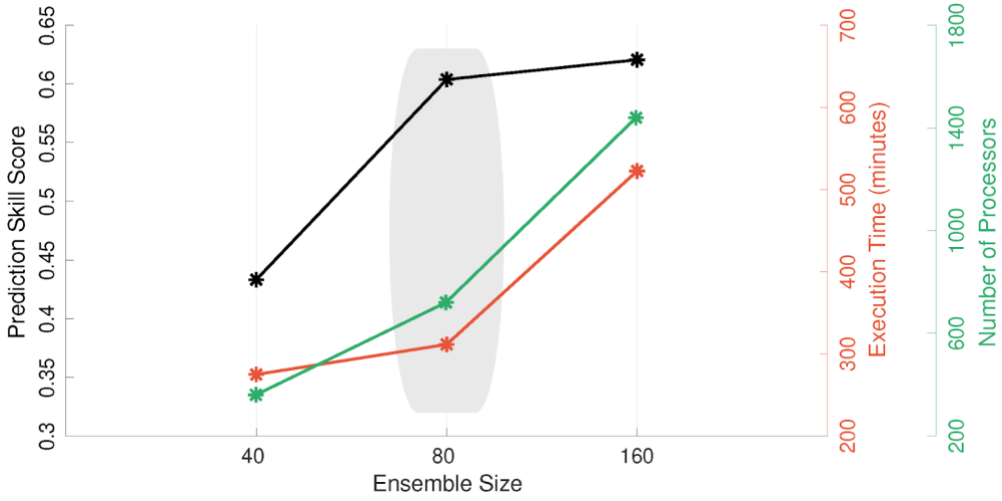
We believe that the answer to this question is present in the text already:

"The NWM employs a groundwater bucket model as a simple aquifer representation to mitigate this baseflow problem." and "The bucket scheme is simple and highly conceptual. For this reason, calibration of its parameters is critical for reasonable model simulations."

Which is to say that the model calibrates baseflow performance through the parameters of the bucket model. There is no additional mechanism in the model for adjusting baseflow. We hope this answers the reviewer's question.

- How is the ensemble of 80 members selected in this work? The author has written that this number is achieved based on the computational demand and statistical performance. I would expect at least a figure to justify this optimization.

We thank the reviewer for bringing this up. This is a great suggestion. In the revised manuscript, we added a new figure (also below) where we compare the performance (in terms of prediction skill score) and the computational demand (as function of both time and number of CPUs). We run additional experiments using 40 and 160 members and compare the results to the 80-member experiment we already had.



We argue that using 80 members produces estimates that are almost as good as those obtained using 160 members. Furthermore, the 80-member ensemble run cuts down on the computational demand (of the 160-member run) quite significantly. Lastly, one could further reduce the computational effort massively using 40 members however this sacrifices the accuracy of the streamflow. In short, we find an ensemble size of 80 an optimal way to balance the performance and the computational demand. A similar analysis was added to Appendix A of the revised manuscript.

- Why are multipliers sampled using a uniform distribution? If another distribution is used instead of the uniform distribution, how will it affect the final results?

The main reason for using uniform distribution to sample the channel parameters is because uniform pdfs offer an easy procedure to sample bounded quantities. The channel parameters are bounded from above and below and in this case the six of them are all positive. In addition, because the parameters are geometric ones, they are sampled under some physical constraints (for example, top width cannot be smaller than bottom width). Lastly, the channel parameters are unknown quantities and hence using bounded uninformative priors (such as uniform pdf) is a reasonable choice. Other distributions such as Gaussians can be utilized but one needs to make sure the sampled values do not fall outside the predefined physical bounds and more importantly the draws have to be positive. By doing so, we would be effectively utilizing biased and truncated Gaussians (not pure Gaussian pdfs). Other forms such as beta, lognormal or exponential distributions could be used but were not explored in this study. We don't believe the choice of the sampling distribution will have major effects on the streamflow results. In future studies, we could look into this in more detail. Thank you!