

Comments on Improving Large-Scale River Routing Models with ESA long-term CCI discharge data assimilation: Advancing Accuracy for Climate Studies by Sadki et al.

This manuscript presents a study to evaluate the effectiveness of assimilating the long-term ESA CCI discharge and WSE data into two contrasting large-scale hydrological models based on model performance. The study seeks to conclude which type of satellite data (noisy flow data vs cleaner height), at which temporal resolution and spatial coverage, provides better model improvements. The study implemented DA into the uncalibrated CTRIP and calibrated MGB models, which allows the readers to evaluate if DA acts as a correction (or calibration) for a poor-physics or a refinement of already good parameters. The paper is well-organized and well-written, but the reviewers have some concerns before this manuscript is considered for acceptance for publication.

1. Although the authors explicitly state that they did not compare model absolute performance but rather the relative improvement from DA. However, two models are in different scales and driven by meteorological data on different scales. For instance, MGB uses satellite-based precipitation (GSMAP/IMERG) known for better tropical performance, while CTRIP uses ERA5 reanalysis. How do the authors consider the bias in precipitation? If CTRIP performs poorly, is this due to the model physics or because of input driving precipitation data? Why do the authors use the same consistent forcing data to isolate the impact from model structure vs. forcing error? The reviewer is curious if the authors can conduct a sensitivity test where both models use the same precipitation forcing for one sub-basin to see if the differences persist.

In the following section, the authors used different perturbation methods for each model. Does this also impact the conclusions the authors made in this paper? The better model performance might be due to a superior DA setup? It is good to see the same localization strategy used for both models.

2. The authors mentioned that the floodplain scheme was not activated in the CTRIP, but the Inner Niger Delta and the Congo Basin are described with floodplains. Does the deactivation lead to an over-reliance on DA to simulate water storage?
3. The discharge data from altimetry and multiple spectral images used for assimilation are derived from rating curves calibrated using in-situ discharge data. For the validation metrics presented (e.g., the NSE of 0.83 in the Niger), did the authors ensure that the in-situ gauges used for validation were independent of those used to calibrate the ESA CCI products? Could the author clarify this? If there is overlap, how can the authors distinguish between true model improvement and the assimilation of pre-calibrated information?

4. The open-loop analysis reveals a massive structural bias in CTRIP. Given that the authors assume the model is 'error-free' in their DA setup, isn't the filter essentially being used to compensate for missing physics (evaporation) rather than meteorological uncertainty? Could the authors discuss it? Is DA effectively acting as a bias-correction for missing physics?
5. Regarding Question 3, which aims to explore the differences between the Congo and Niger basins, the Results and Conclusion sections lack a dedicated synthesis of these findings. Could the authors provide a comparative Table that addresses: how did the inter-annual variability (high in Niger, lower/stable in Congo) affect DA stability? How did the gauge network density influence the localization's effectiveness? What is the generalized takeaway that applies to both, vs. what is basin-specific?"

In line 8, could the author briefly explain why discharge assimilation outperformed WSE?

In lines 12-15, it shows that DA works better for the calibrated model than for the uncalibrated model. Could we conclude that it is better to calibrate the model in the first place and then implement DA for further improvements?

In line 17, the author mentioned that the Congo Basin exhibits modest improvements. Could the author clarify the reason? Is it due to the CCI data quality in this region, as it covers wetlands?

In line 216, please complete the sentence.

In line 245, could the author mention the spin-up period which stabilizes deep soil moisture states before beginning the DA experiments?

In line 402, could the author clarify whether the temporal lag affects the DA performance?