I find this paper to largely be well written and its work addresses an important gap in the literature. There are some points I would like addressed and some room to improve clarity but overall I think it should be accepted after revisions.

Some revisions to be considered:

1: One topic I would like to see expanded upon is in relation to how certain sources of error in the input data and method may be influencing the final results. Specifically, the fact that storage values are what is used to constrain model training. To discuss why I think this is important I refer to.

Line 745-747 "Therefore, we suggest reservoir operation models rely primarily on validation of storage in place of validation solely on streamflow as the available streamflow observations are rarely close to the release point of the reservoir and therefore not as sensitive to reservoir operations compared to storage. "

While this may be true if the modelers in question are primarily concerned with the reproduction of storage, it seems to me that there are plenty of other sources of error that could make this untrue for other metrics. For one example, errors in the other fluxes of ET, precip, and storage lost to recharge could easily be introducing errors to the actual releases. Given the model is being trained to reproduce observed storage values, any errors in these fluxes will be baked into the final release. Even if this does not result in large streamflow differences downstream, if one cares more about release sizes than storage levels I could easily see streamflow data providing additional information.

I do want to acknowledge the authors have already provided analyses of underlying error, an example of which can be seen in lines 740-742:

"We observe that the RF extrapolation is accurately able to depict the flood and conservation curves and that the main source of uncertainty is the errors associated with the storage estimations from satellite altimetry."

What I think could be made a little more clear though is that the core role of storage in the training process, combined with the need to integrate many sources of data all with their own error, could mean there could easily be difficulties in reproducing less related metrics. And in fact, the improvements to storage related results do strongly outperform the improvements other metrics such as streamflow.

Hopefully I have not fundamentally misunderstood the paper when providing this comment. I do see that, for example, the authors validated against streamflow data as well and made sure to select only locations with measurements directly downstream. But my interpretation is that at the larger scale, estimates of storage are the only available data and thus the only available constraint.

I do not want to add a lot of work and whole new analysis to an already well written paper. I think simply a paragraph or two more directly acknowledging the challenges presented by these underlying errors with an illustrative example of one way that might play out, as well as qualifying some of the stronger statements in the introduction and conclusion making recommendations to modelers, would be sufficient.

2:
My second point is in regards to lines 250-251

“This spatial resolution is the optimal balance between computational demand and model performance and has been extensively validated and benchmarked“

I feel that this statement is too strong given the particular problem context. It may be true that this resolution has been extensively validated, but optimality is always a question of “optimal for what metric” and it is not clear to me that this previous work was looking at optimality under the same set of tradeoffs. For example, given that this method can be used to produce datasets, an outcome with a lot of potential downstream consumers, it may be optimal to use a lot more compute for even marginal gains in accuracy.

Additionally, this specific problem has characteristics that may mean a finer resolution is actually optimal. One in particular is discussed on the next several lines, and it is the fact that at this resolution some groups of reservoirs need to be considered as one reservoir because they share a grid cell. This nonlinearity presents what to me is a clear difference in trade offs from a simple performance/compute analysis.

To be clear, I am not suggesting this work should have been done at a different resolution, but this statement seems too broad.

3:
My third point is in regards to the analysis of the various components of the KGE in figure 5. Because all of the models performed very similarly on KGE overall, I am suspicious of reading too much into the size of the various components. Even an individual model with enough degrees of freedom may have parameters tunings that all produce the same overall KGE but with quite different component values. In this case, which component the model appears to perform better on will depend entirely on where you start your gradient descent.

So what is not clear to me is if the different methods have different component values because they are actually better suited to handle that component, or if they have different component values because they both have similar levels of ability to fit the data and have both settled at a somewhat arbitrary local minima that weights the components slightly differently.

4:
My fourth point is in regards to the command area analysis. The authors state the command area does not matter much, but I think this could be an artifact of the particular method.

Particularly, it is not clear to me that downstream demand plays a large role after the release curves have already been so constrained based on historical data. What I would like to see is a more clear description of through which equations the command area plays a role in determining the releases in the section performing the analysis. While the equations are described in the methods, it's a bit hard to sort out the answer to this specific question given the breadth of material being covered.

5: It feels like both the abstract and introduction could be shorter. For the introduction, some of the context being provided might be better suited to the methods section.

6: While I find the illustrative examples used to examine storage dynamics improvement useful, I think additional analysis needs to be done given the small sample they provide. Particularly, I would like to know how the examples compare to the average to know they have not been cherry picked. Also, at least one of the selected examples should perform about average.

Minor suggestions for improved clarity

7: It took me a while to find that in the table 3 description it was specified that all RMSE values are in %/week. This made it very hard to interpret the results. I would suggest that these units be given wherever RMSE values are reported

Similarly, it seems the biases are being reported as percentages but that is not noted until many bias values have been reported. I similarly think the units should be specified at each location that biases are reported.

8: Figure 5.

Could the legend be made smaller to provide more room for wider graphs? They are narrow enough to be harder to interpret. Could also consider a 2x2 layout instead of 4x1. Also, it looks to me like the upper ylim was not set high enough for the KGE components and the distribution is being cut off at the top.

9: Lines 443-444: “Conversely, basins with a large amount of storage (Figure 2a) such as much of Central and South Eastern Asia, Central Africa, and Western Australia do not have a high degree of regulation”

I had to read this line a couple times to get it. I think changing to something like

“Conversely, **some** basins with a large amount of storage (Figure 2a) such as much of Central and South Eastern Asia, Central Africa, and Western Australia do not have a high degree of regulation, **which implies…**”

Would make what I believe to be the intended contrast to the previous lines more clear

10: Figure 3 y-axis just says %. It would be more immediately legible if it said % of what.