Reviewer 2:

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.

Thank you for your kind words and comments.

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.

Thank you for your comments. We do agree that there could be issues that are propogated from other soures of error. We suggest changing Lines 745 – 747 to read as follows: “emphasize model validation on reservoir storage in addition to validation based on streamflow,” in place of validation solely on streamflow as available streamflow observations are rarely close to the release point of the reservoir and therefore not as sensitive to reservoir operations compared to storage.

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.

Thank you for your comprehensive comment. We do agree that adding a section on the propogation of errors would enhance the paper. We suggest this paragraph falls under Section 4.2 and contains the following:

- **Sources of potential error using PCR-GLOBWB 2 inputs, which potentially cancel out the associated errors in the PCR-GLOBWB 2 model**
- **Sources of error in using remotely sensed data**
- **Sources of error in only looking at storage as a validation method.**
- **The impact these errors have on the storage and release of the reservoir**

We will include the following paragraph regarding this error propogation:

"Apart from errors accruing from above assumptions, the accuracy of our results is also limited by the errors that are propogated through our workflow. Specifically, PCR-GLOWBWB 2 underestimates the flashiness of streamflow regimes. It is also less accurate in specific regions such as the Niger, the Rocky Mountains and portions of continental Eastern Europe due to errors in the snow dynamics, estimation of the groundwater responses and data limitations (Sutanudjaja et al., 2018). Additionally, the estimation of the operational STARFIT rules from the remotely sensed storage data of GloLAKES is limited by the revisit time of satellites, the influence of cloud cover and atmospheric interference as well as the statistical models that back calculate storage that are limited by the digitial elevation model resolution (Hou et al., 2024; Chen et al., 2022). As storage is typically not a measured value and, even in the case of in-situ observed water levels observations, is back calculated from storage/area or storage/elevation relationships, validation primarily on storage alone is inherent to uncertainty. Primarily, these limitations can affect the actual storage value as they rely on storage elevation charts that are only periodically updated (Steyaert et al., 2022) While the single errors are propogated through our system, the results of the independent validation with ResOpsUS (Figure 6 in the manuscript) and GloLAKES (Figure 6 and Figure 7 in the manuscript) show improved performance for storage values in PCR-GLOBWB 2 and suggest similar improvements for other global hydrologic models with the caveat that errors may propagate through the modelling system."

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.

This is a really good point. We intially meant this statement to refer to the computational time for running our model on the global scale. By moving to a higher resolution of the PCRGLOBWB 2 model (such as the 30 second resolution), we introduce more potential errors in land cover type and snow dynamics that further complicate the results due to increased evapotranspiration from crop types and lack of lateral transport for snow (van Jaarsveld et al 2025). Additionally, running the PCRGLOBWB 2 model globally on the 30 second resolution takes 401 computational hours accoding to van Jaarsveld et al., 2024. Ultimately, we agree that this is a broad statement and suggest the following change: "We opt for the 5 minute resolution in order to capitalize on the extensive validation and benchmarking done by Sutandujaja et al., 2018 and to limit excessive calculation times that occur at higher resolutions (van Jaarsveld et al., 2025)."

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.

This is a really good point. To expand on this point, we created scatters plots of the different KGE components between the Turn250 and vanBeekGeo models (Figure 1). While the scatter for the R component makes this component appear to be the most important, we find that both the R and beta components have almost equal values above and below the 1:1 line suggesting that these two components are muting the KGE differences. Comparatively, alpha has 1196 points above the 1:1 line and 779 points below the 1:1 line which suggets that alpha is the most sensitive to the operational changes and contributes the most to the KGE changes (1210 above the 1:1 line and 1158 below the 1:1 line). To show this in the analysis, we propose to add this figure and analysis to the supplementary.

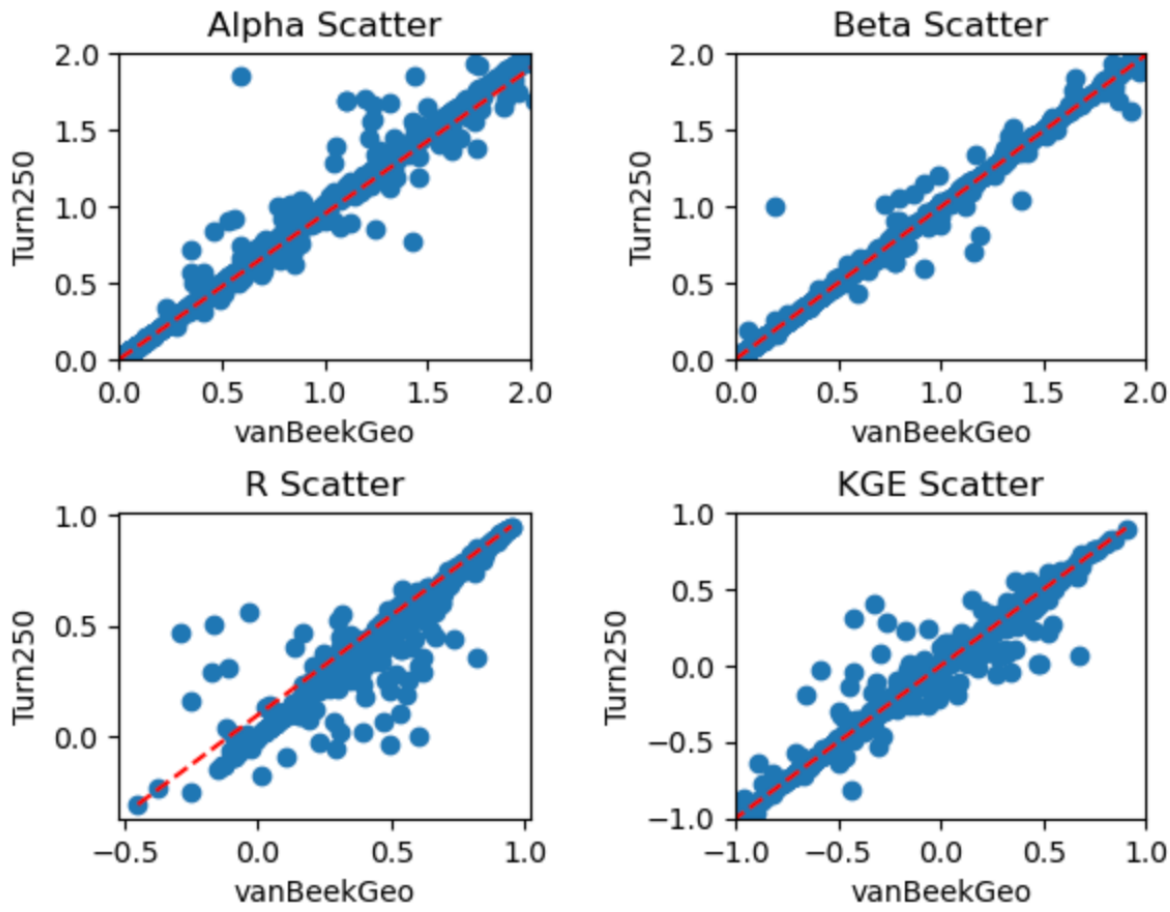

Figure 1: Scatter plots of the streamflow KGE components between each model and observations. We plot the KGE components (alpha, beta, and R) for the Turn250 model on the y axis and the KGE components for the vanBeekGeo model on the x axis. The dashed red line is the one to one line.

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.

Thank you for your comments. The command areas are taken into account in equations 7 and 9. We calculate the command area as the downstream region that the reservoir could supply water to. Therefore, D in equations 7 and 9 refers to the maximum downstream demand that is aggregated over the specified command area per model (i.e. 250, 600, and 1100). We suggest adding the following to clarify this on line 356: "where D refers to the maximum demand aggregated at the specified downstream area (250, 650, 1100)."

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.

We agree that shortening the introduction and the abstract would be useful and will shorten the abstract to less than 300 words. We will also shorten the introduction from 13 paragraphs to 7. We do think some of the context is quite lengthy and we can still cover the main components in a simplier fashion.

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.

Thank you for your comment. We thought the single point location was a nice way to illustrate the potentiall differences in operational dynamics and their impacts. We agree that a point location does not tell the full story. To better tell this story, we have opted to include the climatology of the storage fraction and the storage integral to show the average changes between the different model scenarios. From this figure, we observe on average that the storage fractions in Figure 4 in the manuscript align with the general trends we see in the average storage fraction climatology (Figure 2). That said, the average storage values are lower in modelled values in Figure 4 in the manuscript compared to all the dams in our analysis. To compliment this qualitative analysis, we also calculated the correlation and the KGE for the three models between the longterm monthly storage of all the dams and the Clinton and Koelnbrein dams (below table). We do observe that the Koelnbrein dam has high correlations and slightly positive KGE values that suggest that this dam is fairly representative of the dynamics we observe when taking the average of all the dams in the longterm storage. The Clinton dam, however, has a varied performance depending on the model suggesting that this dam has different dynamics than the longterm monthly storage values.

Model	Clinton KGE (storage)	Clinton R (storage)	Clinton KGE (storage integral)	Clinton R (storage integral)
Baseline	0.068	0.05	-7.95e17	-0.88
vanBeekGeo	-0.45	0.43	-3.37e18	-0.69
Turn250	-0.219	-0.032	-2.01e18	-0.911

Model	Koelnbrein KGE (storage)	Koelnbrein R (storage)	Koelnbrein KGE (storage integral)	Koelnbrein R (storage integral)
Baseline	0.13	0.96	-4.47e17	0.41
vanBeekGeo	0.25	0.85	-5.62e17	0.65
Turn250	0.13	0.85	-7.56e17	0.56

When looking at the storage integral average climatology, we see varied dynamics in the summer months that align with the average of the two examples in Figure 4 in the manuscript. However, the KGE and correlation values show that the Clinton dam is

not well represented by the longterm average plots and the Koelnbrein dam on the other hand is representative of the average dynamics. Therefore, the two examples shown in Figure 4 in the manuscript both show an example of a dam that aligns with the expected average values (Koelnbrein dam) and an example of a dam that is not indicative of the average trends (Clinton dam). We plan to include this figure (Figure 2) and the two tables in the supplementary and keep figure 4 in the manuscript as is. We will also include plots of the longterm monthly discharge at major basin outlets so the regional differences can be better seen and add this plot to the supplementary.

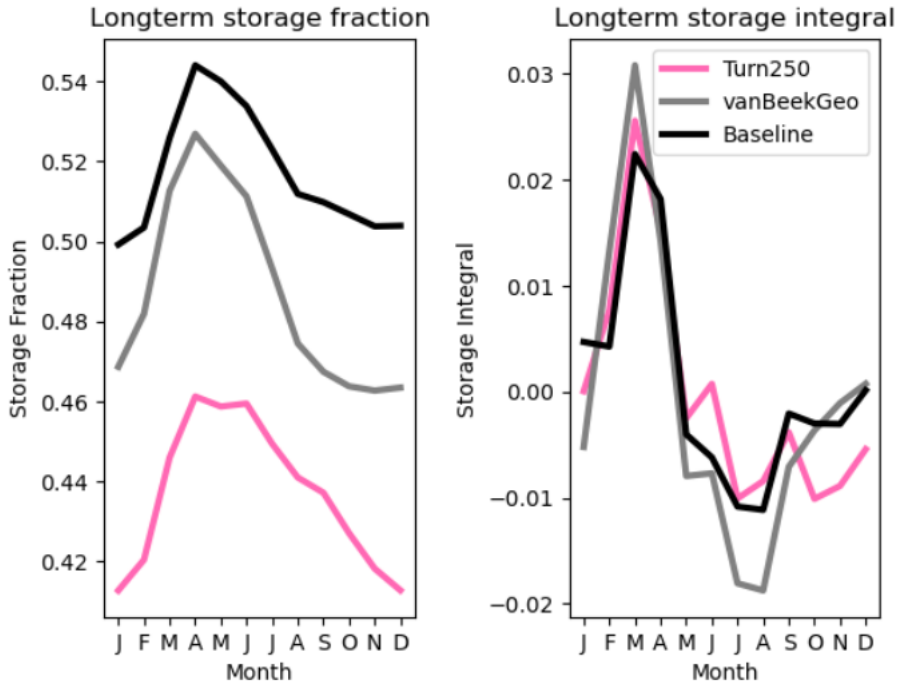

Figure 2: Plots of the longterm storage fraction (left) and the longterm storage integral (right) for the three models: Baseline (black), vanBeekGeo (grey) and Turn250 (pink).

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.

Thank you for your comment. We will update the table to include the units. We also will go through the manuscript and make sure the units are stated when we first mention the metric.

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.

We agree that a 2x2 layout would align better with this figure. We specifically set the ylim in order to see the small differences in the distribution as the majority of values were at 0 for alpha, beta and and R. We will update the original figure to include the 2x2 panel (Figure 4) and will put the plot without the zoom (Figure 3) in the supplementary.

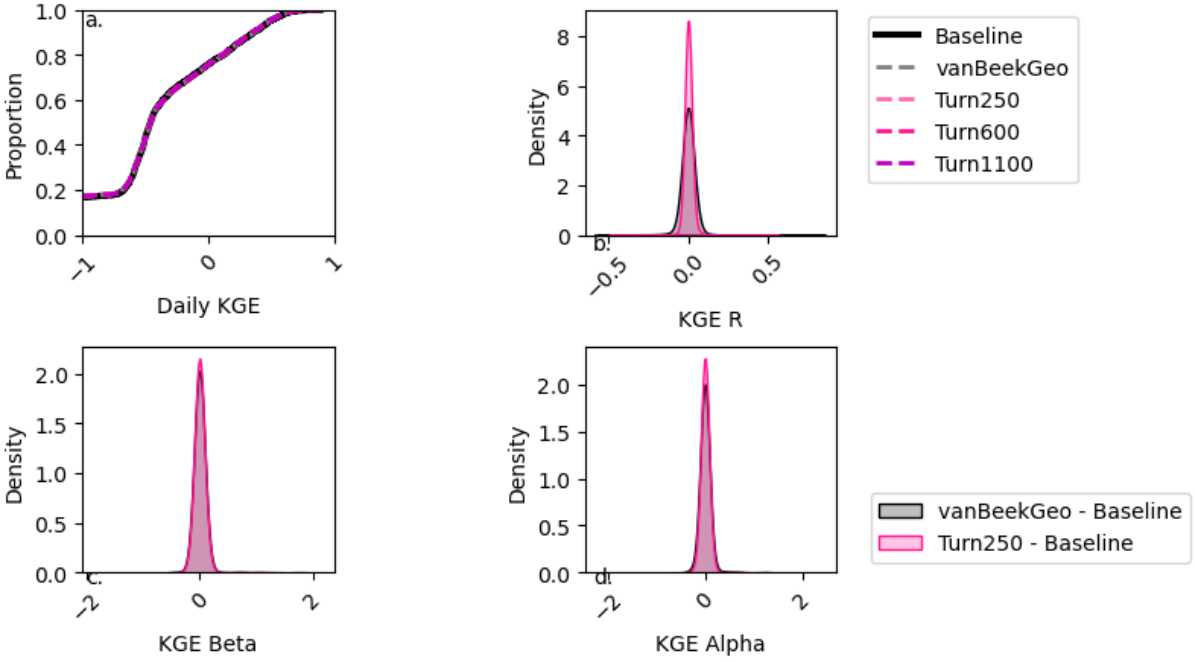

Figure 3: Show the KGE, and KGE components of the three models we used in our analysis for Figure 5 in the manuscript without any zoom.

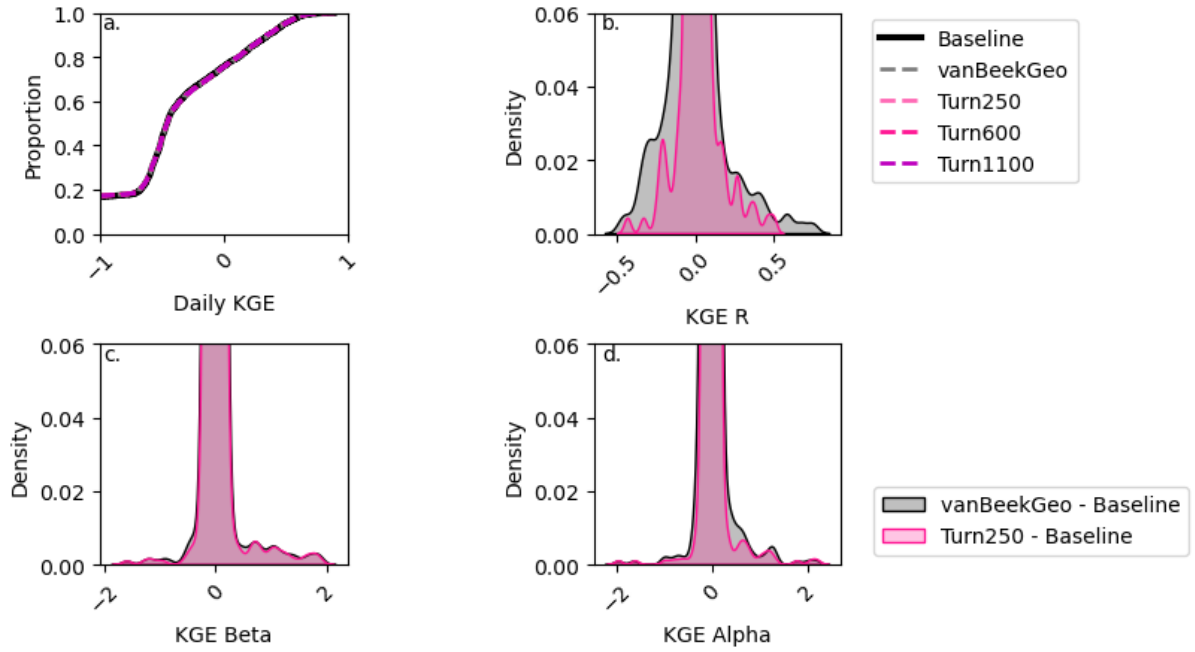

Figure 4 Show the KGE, and KGE components of the three models we used in our analysis for Figure 5 in the manuscript as a 2x2 panel plot.

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...**”

Thank you for the comment. We think this is a really nice change and will ammend the manuscript as follows: “Conversely, some basins with a large amount of storage (Figure 2a) such as much of Central and Southeastern Asia, Central Africa, and Western Australia do not have a high degree of regulation, which implies that there is not a direct relationship between total storage and a high degree of regulation (Figure 2b)”

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.

We agree that this could be more clear. We will change the axis to include Storage Percent (%).