Dear Reviewer,

Many thanks for your critical yet constructive review of the submitted manuscript. Before the manuscript is resubmitted, we will put additional emphasis on language and grammar.

On a more general note, and answering your question whether 'hyper-resolution is dead before it really started?', we would like to say that we do not find the presented results to be sobering, but very much meeting our expectations. Even though the topic of hyper-resolution hydrological modelling has now been discussed for a decade already, there have only been a few actual attempts to move hydrological models to a (sub-)kilometer resolution over large scales. As such, we think it would be illusive to assume that a first take (and that is exactly what the study presents) would yield major improvements across all evaluated variables, especially against the backdrop of observation data which indeed is not yet entirely commensurate as you rightly state. In a revised version of the manuscript, we will ensure that the 'expectation management' is improved with respect to what a first-generation hyper-resolution hydrological model can achieve.

Below, we address your further comments (in blue) and will outline how we plan to improve a revised manuscript.

In a way, the study has a conceptual problem, because upscaling and re-classification of soil texture and land cover (and water management/reservoirs?) was used to go from fine to coarse resolution. Thus the models are different not only in terms of spatial resolution and atmospheric forcing but also in terms of structure (i.e. different models at different resolution). Thus, comparability is not necessarily guaranteed, as claimed in the methods section. That's OK, but needs to be made transparent to the reader and discussed in detail. Perhaps it's one of the reasons why resolution does not do the trick in case of soil moisture and evaporation.

Many thanks for this comment. As also mentioned below, it is indeed true that model schematizations at different spatial resolutions are not identical, and that comparability is hampered. Schematizations of runs where only the forcing resolutions is changed are, however, identical. By using identical input data and parameters together with consistent scaling approaches, we aimed at minimizing differences across schematizations and their impact on comparability. In that sense, we do not compare model resolution in an isolated way, but rather model schematizations at different spatial resolutions (hyper-resolution and coarser) including their indirect impacts on how input data is processed internally. As these aspects are key, **we will add more transparency and explanation to the manuscript in general and especially its method section**.

The introduction is prominently missing a discussion of the recent relevant paper by Condon et al. (2022) on global (hpyer-resolution) groundwater modeling.

We kindly thank the reviewer for this literature suggestion. However, we could not find an article lead by (Laura) Condon from 2022 on hyper-resolution groundwater

modelling. We assume that this article from 2021 is meant instead, which indeed is a great addition to the manuscript and **will therefore be included in a revised version**.

2, 39: This statement is misleading. PFCONUS is just a naming convention (just as naming the setup of PCR-GLOBW over Europe PGEU). Of course ParFlow can be applied at the global scale, in principle; it's a generic simulation tool like many others.

Many thanks for pointing out this ambiguity. We are aware that both ParFlow and PCR-GLOBWB can be applied at any spatial resolution and spatial extent provided appropriate data is available to be fed into the model. What we failed to describe properly was that PFCONUS is a ParFlow model tailor-made for the CONUS region, whereas the 1k PGEU model – if you like – is using only data that could also be used for a global application. Being aware of this difference is crucial as more bespoke national or regional data sets will very likely be more accurate than data sets with global extend, which in turn will be reflected in the outcome of the evaluation. In a revised version of the manuscript, **we will rewrite the section under consideration such that this ambiguity is removed**.

4, 5: Here, additional information is required in the main text. From the appendix it follows that upscaling was used for soil texture and special classification for land cover was used to move from high to low resolution (how are reservoirs upscaled/downscaled?). Thus, the models are not identical in addition to the resolution of the forcing.

Thank you for your remark. Indeed, the model schematization are not identical across runs and resolutions. However, we never made this claim but only stated that the input data and parameters are identical. When creating schematizations at different spatial resolution, it cannot be avoided that these schematizations differ where data had to be down- or upscaled. Due to that, we did our best to keep the schematization as aligned as possible. In the revised manuscript, **we will put extra emphasis on explaining this properly**.

With respect to your specific questions on reservoirs, which we base on the GRanD data base (version 1.3) providing shapefiles of reservoirs, including information about their surface areas and capacities: for every resolution, we "rasterize" these shapefiles. In fact, it is more than just a "simple" rasterization process as we need to consider many factors, such as their locations to the drainage networks (at different resolutions), number of reservoirs within pixels, and so forth. If a pixel contains more than one reservoir (which is very likely in coarser resolution), we merged their surface areas and capacities and treated them as one reservoir. As such, the overall physical properties of reservoirs should not be overly different when moving from finer to coarser resolutions and thus their impact of flow estimates should be small. **We will append the appendix with this information and point towards it more prominently in the main body of the manuscript**.

Figure 3: remove 50k_50k from plot.

We thank the reviewer for proposing this improvement. While we initially kept the 50k_50k simulation in there as a reference point for the other simulations, we now agree that it does not add information to the plot and **will therefore be removed in the revised version of the manuscript**.

Why not applying the relative KGE to all variables (also soil moisture, ET)

Many thanks for this suggestion. Applying the KGE to all variables is not possible as particularly soil moisture evaluation needs to be treated carefully. Since satellite-based evaporation estimates are typically based on the first few centimeters of the top soil, PCR-GLOBWB uses the first 30 cm. Absolute values of soil moisture simulation and observations are therefore not directly comparable and we only can assess the correlation, as also mentioned in the manuscript on page 5. Consequently, we decided to use consistent metrics (RRMSE, R2) across all variables which we evaluated in space and time (i.e. soil moisture, evaporation, terrestrial water storage anomaly) and KGE for all variables which we only evaluated in time (i.e. discharge). Additionally, it is worth mentioning that by using the RRMSE we were able to account for differences in spatial variability of the signal we like to predict that exist between different areas, which is not possible when using KGE. Therefore, we are confident that the choice of metrics is well defined. Nevertheless, we may want to **add a better explanation of our reasoning to the revised manuscript**.

Figure 6: Replace "other" with correct information. Plot 1:1 line correctly everywhere. The plot almost suggests the 50k_50k is also doing better than 1k.

We thank the reviewer for pointing this out. **We will replace "other" with the actual run names** for improved comprehensibility. The 1:1 line in each plot, however, is plotted correctly, as is your observation that the 50k_50k run is (slightly) better than the 1k_1k run. This is also quantified in Table 3.

A couple of questions for the discussion and conclusions: Perhaps the observation data is not scale commensurate and can not be used to assess hyper-resolution modeling results? Perhaps PCR-GLOBWB is not scale commensurate and can not be used at hyper-resolution?

We thank the reviewer for these questions. In our opinion, it is rather the observation data that is not (yet) scale commensurate, at least when evaluating the simulations over a longer period (>5 years). The model itself should be applicable at the 1 km scale. Nevertheless, we need to acknowledge that there is still room for improvement which is not surprising as the study presents a real first-of-its-kind application and analysis of a hyper-resolution hydrological model at the continental scale. While we already discuss these questions in the manuscript, **we will ensure that they are answered more clearly in a revised version of the manuscript**.