Reviewer 1

General comments

The paper "A first attempt to model global hydrology at hyper-resolution" by van Jaarsveld and colleagues uses the recent global kilometer scale dataset CHELSA (Brun et al., 2022) as a reference to run the global hydrological model PCR-GLOBWB, using a new downscaling method of the forcing W5E5 (Lange et al., 2021) and introducing a new scheme for lateral transfers of snow.

It addresses important scientific questions in the current state of hydrological research and high-resolution simulation challenges. Their paper is well written, easy to read, with the method very well explained and detailed.

They point out the challenges linked to hyper-resolution modelling such as the calculation time and here develop a very clear and detailed spin-up and parallelisation approach. It allows them to keep the simulation time reasonable (about 17 days for the 30 arc-seconds resolution for the period 1985 - 2019).

Their downscaling method relies on the construction of monthly climatologies: the recently available CHELSA dataset is used to construct monthly kilometric scale climatologies as references, then the forcing data are downscaled by bilinear interpolation. Monthly climatologies are calculated from these interpolated fields and a correction factor is calculated comparing them to CHELSA climatologies, then temporally linearly interpolated to get a correction factor by Julian day for precipitation, temperature and potential evaporation.

Their lateral transport of snow scheme is relatively simple, introducing a frozen water threshold. When this threshold is reached, the lateral transport is activated, through a function of steepness towards neighbouring cells.

Their main results show that the hyper-resolution modelling allows for a better simulation of discharge, especially over smaller catchments that are not represented at coarser resolution. However, not all variables show a significant improvement, such as evaporation which is even worse at higher resolution when compared to MODIS data. The authors hypothesised that this is mainly due to how the land cover is handled at high resolution.

The results are clear but I feel the discussion would benefit from being expanded, with a more critical discussion on the limit of the downscaled climate variables and with more links between the different variables analysed. For instance, the river discharge tends to be underestimated for the 30 arc-second resolution: is it logical since the evaporation is overestimated?

Also, maybe some regional comparisons in performances and more discussion on the hypotheses behind some limitations would be interesting. To explain the poor performance of evaporation for instance.

We want to thank the reviewer for reviewing our manuscript, the reviewer's comments have provided us with valuable insights and suggestions that will improve the quality of this manuscript for future readers. Please note that when we refer to line numbers, we are referring the marked up version.

The reviewer suggests that the manuscript will benefit from an expanded discussion that places additional emphasis on the linkages between the variables we validated in section 4.2 & 4.3. The reviewer highlights that an overestimation of evaporation may be, in part, responsible for the better discharge predictions in the 30 arc-second simulation; where the underestimation in the 30 arc-seconds simulation corrects for an over estimation of evaporation in the 5 and 30 arc-minute simulations. We agree with the reviewer that additional emphasis on these issues will strengthen the manuscript. As such we have included the suggestions in the updated manuscript by elaborating on this point in section 4.2 (line 541 – 546) & 4.3 (line 570 – 584),

Line 541- 546: "*Reductions in bias as resolution increases also contribute to improvement of river discharge; for the 30 arc-minutes and 5 arc-minutes resolution the model tends to overestimate river discharge, whilst the 30 arc-seconds results are underestimated. When considering that, for the 30 arc-seconds resolution, discharge values are underestimated and in conjunction with the observation that evaporation is overestimated, the question arises whether this increased evaporation leads to better estimates of river discharge by correcting for overestimation in the coarser resolution. Indeed from the soil moisture and evaporation validations, we can conclude that an overestimation of evaporation may result in a better estimation of river discharge bias.*"

Line 570 – 584: "*To corroborate this claim we conducted an post-hoc analysis which was aimed understanding which proportion of the model domain consists of forests when using dominant land cover types compared to a fractional coverage. This analysis revealed that by expressing land cover as a single dominant class per grid cell leads to a 13 percentage points (Appendix B1) inflation in the total area covered by forests ($\sim$ 50\%) compared to when using the fractional cover ($\sim$ 37\%).*

To further evaluate the sensitivity of the water budget terms to changes in land cover parameterisation, for a small test region, we changed the land cover representation so that the entire region consisted of either forest, grasslands, or crops and compared the water budget terms to a 5 arc-minute simulation with unchanged land cover representation (see appendix B2). These simulations show that decreasing the relative abundance of forests within a domain will result in decreased rates of evaporation and increased rates of runoff. However although the results for this region are quite sensitive to land cover, it is unlikely that any combination of land covers will result in relative rates of evaporation and runoff similar to that of the coarser resolutions. Thus suggests that there are further opportunities, besides land cover representation, responsible for the difference in water budgets between resolutions. For instance it may be that neither downscaling approach is capable of reproducing meteorology accurately at the 30 arc-seconds resolution."

In addition, the reviewer suggests that by adding a regional comparison to strengthen support for the hypothesis that, overestimation of evaporation is attributed to the use of dominant landcover types in the 30 arc-seconds compared to fractional landcover types in the 5 and 30 arc-minute resolution. We agree that this could be a potential reason and would add a valuable sensitivity analysis to the manuscript. However, we are currently limited by the availability of high-resolution global land cover datasets which would be required to thoroughly conduct such an analysis. As an alternative to a regional analysis using observational data, we think that by further expanding on why the use of a single dominant landcover type at the 30 arc-seconds resolution results in an over estimation of evaporation can strengthen support for our hypothesis. Reviewer 2 rightfully pointed out that calculating and presenting differences in landcover representation when using dominant vs fractional aggregations would not be difficult and would add support to our hypothesis. In the updated manuscript we show that the area represented by forests are larger using a dominant landcover type compared to fractional. For this we have relied on the global 100m dynamic landcover dataset by the Copernicus Global Land Service (Buchhorn et al 2020).

A cursory analysis using the 100m landcover data reveals that at the 30 arc-seconds resolution the difference in total area represented by forests differ markedly when considering dominant verse fractional representations. When considering only a single dominant landcover class per grid cell the forests constitutes ~50% of the land surface; however, when we consider the fractional cover per grid cell this value decreases to ~37%, a difference of 13 percentage points (Please see appendix B1 & line 569 – 575).

In addition to the above analysis, we have also looked at how sensitive evaporation in the 30 arc seconds is to landcover changes, for a small test region (southern alps). We assumed that the entire domain is covered by forest, grassland or croplands.

These three simulations were compared to the original configuration 30 arc-seconds and 5 arc-minutes (as presented in the original manuscript).

From this sensitivity test we conclude that evaporation rates are indeed sensitive to changes in landcover representation, however none of these changes resulted in evaporation rates comparable to those of the 5 arc-minutes simulations nor were we able to obtain the same ratio between evaporation and precipitation. In addition, the evaporation estimates we present are at the higher end of what can be expected (94 – 130 $km^3.year^{-1}$). It also shows that currently no landcover configuration at the 30 arc-seconds resolution will be able to reproduce the values found at 5 arc-minutes, since the maximum possible runoff values at the 30 arc-seconds (125 $km^3.year^{-1}$) are still lower than what is simulated by the 5 arc-minutes resolution 126 $km^3.year^{-1}$.This is to be expected given that the 5 arc-minutes is known to underestimate rates of evaporation and overestimate rates of river discharge. This also means that even by changing all landcover to crops (which is unrealistic), we are still unable to obtain the same values as for the 5 arc-minutes model simulations. This has been added to the appendix (Please see B2) and further elaborated on in the discussion (line 576 – 584).

References:
Buchhorn, M., Smets, B., Bertels, L., Roo, B. D., Lesiv, M., Tsendbazar, N.-E., Herold, M., and Fritz, S.: Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe, https://doi.org/10.5281/zenodo.3939050, 2020.

Specific comments

The main questions and remarks I have related to the downscaling method and to the discussion regarding the poor performance of evaporation.

At such a resolution, climate variables will highly depend on altitude and topographic features, as underlined l. 415, and a bilinear interpolation will not be sufficient to maintain spatial coherence between all the climatic variables. I feel this should explain most of the poor performances on evaporation at high resolution since it is the climate variable most non-linearly dependent on a variety of climate factors. I understand you used the only global scale kilometric meteorological data available to you at such a scale and this is a good start but I feel this limitation should be introduced sooner and highlighted more.

I am not sure I fully understand what they did for the part "4.3 Untangling Model Scaling and Forcing Downscaling Affects" p23.

They conclude that since both downscaling methods lead to similar differences in simulations, these differences should be mostly due to parametrisation. Could neither downscaling method be sufficiently good to keep the spatial coherence of climate variables to calculate a correct evaporation? Could neither downscaling technique allow for an accurate disaggregation of the forcing variables?

It seems in Table 6 that there are smaller gaps between simulations than there are in Table 3 (I would add the ratios here to help the comparison). I would be interested in knowing the precipitation/evaporation ratio over Europe compared to the worldwide results, Europe may not be representative of the limitations occurring everywhere in the world. Since there is an issue with the representation of evaporation, if it is linked to the downscaling of potential evapotranspiration, the results should change depending on how much of the water cycle is driven by evapotranspiration. If it depends on the land cover types, comparing regions with very different land cover dominant types would be interesting.

We agree that by comparing regions of different landcover compositions could lead to a better understanding of how landcover relates to evaporation estimates, However, as stated in our general response, we are limited by the availability of landcover datasets with a global coverage to thoroughly conduct such an analysis. In other words, we do not believe that we have enough observational data to compare different regions in such a way that we can be sure the any differences are attributable to differences in the model as opposed to artifacts arising from variations in the availability of observation data between regions. We hope that by expanding on section 4.3, as stated in the response to the previous comment, we will have addressed the concerns highlighted by the reviewer.

We do agree with the reviewer that adding ratios to Table 3 (Pg 19) and 4 (Pg 31) will aid in comparison and the updated manuscript contains these edits. Last, the reviewer suggest that Europe may not be representative of the limitations occurring in the rest of the world. Ideally, we would have wanted the comparison presented in section 4.3 available at the global scale, however the 30 arc-second simulation is too computationally expensive to be completed for a multitude of model configurations. We have opted to focus on Europe due to these computational limitations, large degree of heterogeneity and availability of relatively large amounts of observational data. Acknowledging that Europe may not be representative of the global picture, we compared just how different the European analysis is to the global analysis and compared the ratios in Table 3 and 6. It can be seen that although small differences are evident between the the global and European analysis results are comparable which suggest that, at least for this analysis, Europe may provide a good reference for the rest of the world.

Here we address the previous three comments made by reviewer. In the first comment the reviewer correctly points out that bilinear interpolation will not maintain the spatial coherence between all the climatic variables. Section 4.3 serves to provide information on whether the differences we report for the 30 arc-second resolution are due to differences in landcover parameterisation or as a result of the new downscaling technique developed in this study. The reviewer correctly states that downscaling (irrespective of technique) may result in disaggregation's that can result in erroneous calculations in evaporation. However, another possibility is that downscaling may result in appropriate disaggregation's but that erroneous estimates of evaporation arise from the fact that we do not include sub-grid variability in landcover at the 30 arc-seconds resolution. Section 4.3 served to elucidate if the results are due to the downscaling technique (as the reviewer suggests) or due to the difference in landcover parameterisation. Considering the results in figure 11, we can see that the results are comparable when directly comparing downscaling techniques at the same resolution (30sec old vs 30sec new; 5min old vs 5min new); yet differ between resolutions (30sec new vs 5min new; 30sec old vs 30sec old). This suggest that that new changes in landcover representation could in part be responsible for the observed differences. A new section in the discussion (line 569 – 576) further investigates this by looking at the role of landcover on the evaporative fluxes. The link between section 4.2 and 4.3 has been expanded on in the updated manuscript to better highlight the motivation of section 4.3 (line 541 – 546). However, this does not negate the importance of incorporating high resolution meteorology in future attempts and we have expanded this point in the discussion (line 581 - 594, 605 - 608).

Here are some minor comments:

In the results, I would introduce the relative values analysed in the text in the table!

We agree with the reviewer's suggestion and this has been included in the revised manuscript. Please see table 3 and table 4.

And, as previously said, the discussion needs to be expanded a little.

We agree with the reviewer that by expanding the discussion the manuscript will be strengthened. We kindly refer the reviewer to our general response where we detail how the discussion has been expanded.

Maybe comment on the discrepancies in the validation over different regions? (Appendix maps)

We agree with the reviewer and have added the following sentence to the end of this section lines 307 – 309: *"It is important to note that for both observed evaporation*

and soil moisture are not uniformly represented across the modelling domain. Observations are denser over North America and the European continent compared to the rest of the world (Fig A1 & Fig A2).”

In the explanation of downscaling precipitation p5:

- Why use a multiplicative correction factor while an additive one was used for the other two other climatic variables? Is it to handle differently the issue with the variance conservation? I feel this should be at least justified.

We thank the reviewer for this comment. We have consulted the relevant sections and can report that text is not clear. The correction factors are multiplicative for precipitation and evaporation, whereas temperature it is additive. The reviewer is correct that a multiplicative correction factor was used to handle variance conservation and to avoid negative values. In the updated manuscript, this explanation has been added and formulas revised to reflect the context given in this response. Please see lines 165 – 166 and 182 - 183.

Paragraph on evaporation and soil moisture evaluation p10:

How did you use the simulation data there? Are the stations which have observations located and the nearest grid point value taken? How was handled the difference in resolution?
Thank you, this is a necessary addition. This has been made more clear in the updated manuscript by adding the following sentence (lines 293 - 294): *“To match the location of observation stations with the appropriate grid cells, we located the nearest grid cell relative to the coordinates of the observation station.”*

What are the optimum targets for rho, beta, and alpha there? It could help to understand the analysis later on.
This has added by adding the following sentence (lines 306 - 309): *“A perfect KGE score is 1, which arises when all components of the score ρ, α and β equal 1 (i.e, the observed and modelled values are identical).”*

l 233: Comparison with MODIS: Do the results differ if all the data are aggregated to the 30 arc-minutes resolution instead? Because the downscaling algorithm here may have an impact.

We agree with the reviewer that conducting the snow validation by aggregating all simulation results to the 30 arc-minute resolution may lead to different results. The rationale for validating all simulations at the 30 arc-seconds resolution is that we are interested in how well PCR-GLOBWB simulations are able to reproduce snow dynamics observed in reality. Given that 30 arc-seconds snow cover data from MODIS is the closest resemblance of real snow cover dynamics, we want to elucidate whether increased resolution results in better simulations at this fine

resolution. Validating all simulations at the 30 arc-minutes resolution would tell us how increased model resolution differs in reproducing 30 arc-minutes snow cover dynamics; although this is interesting in a statistical sense, it is not the focus of this current work.

Paragraph 3.2: Since the relative values are analysed, I would introduce them in Table 3 (and Table 6 later for comparison).
We agree with the reviewer that this is valuable and has been included in the updated manuscript, please see Table 3 (Pg 19) and Table 4 (Pg 31).

For the analysis of the snow cover p18, I had some questions:

- Are the differences significant? They seem very small.

- Are the results different in different regions?

Unfortunately we were not able to perform an analysis on whether the differences we report are statistically significant given that the metrics we present are not appropriate for such an analysis. A potential reason for lack of obvious differences in validation scores is that the validation was conducted at the global scale. As such in the updated manuscript we have confined the validation to regions that actually experience snow fall (see lines 329 - 333). We do however agree with the reviewer that by including a component on whether there may be regional differences would strengthen the manuscript. In the updated manuscript we have included a global map showing how the brier score differs between resolutions (please see fig 8 & lines 438 - 441). The global map of the brier score provides opportunity to discuss how regions differ in the discussion which was added too (lines 519 - 521). In additions we have updated the analysis for the snow cover component to allow for a more comprehensive comparison between the three resolutions. We believe this new figure (Figure 8 on page 25 and below for you reference) and the accompanying scores provide more insight in the performance of the different simulations and also highlight the potential benefits of using this 30 sec simulation for snow simulations.

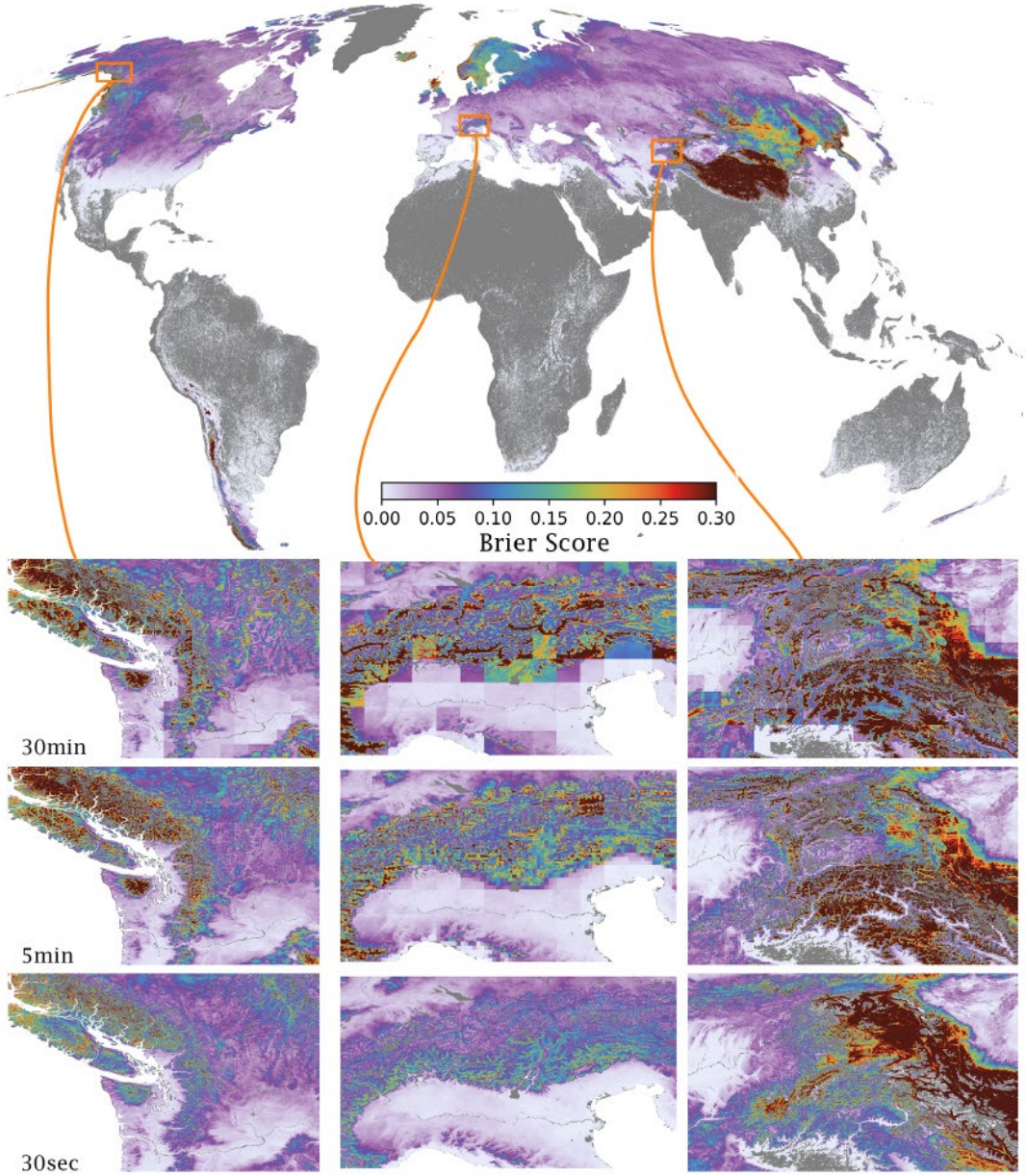

Figure 8. Brier score of simulated snow cover (-) simulated by PCR-GLOBWB at 30 arc-seconds resolution. Zoomed insets compare how the highlighted regions differ between the 30 arc-seconds, 5 arc-minutes and 30 arc-minutes resolution PCR-GLOBWB simulations.

Technical corrections

16: "cross-cutting"??

This has been corrected to crosscutting (see line 17).

48: "How ever valuable" and not "However, valuable…", it has a different meaning
This has been corrected, see line 60.

53: Missing verb? "lead to unrealistic accumulation…"
This has been corrected, see line 65.

126: “in each month of …”

This has been corrected, see line 173.

194: Issue with the structure of the paper: you have a 2.2.1 but no 2.2.2

This has be corrected by adding “2.2.1 Simulation” and “2.2.2 Evaluation”. Please see line 227 & 247

Equ. 11 and 13: you should explain the meaning of the terms that were not introduced before, even if most of them are classic notations.

This has been corrected by adding explanations for Eq 11 (lines 261) & Eq 13 (lines 285).

219: “were used” and not “we used”

Thank you, this has been corrected (line 293).

233-234: I don’t understand the link between the two part of this sentence. I believe the second part would go better with the next sentence. Maybe present the resolution of MODIS for comparison?
We agree with the reviewer, in the updated manuscript the first sentence is removed since this sentence is repeated information already present in the paragraph. The updated line reads as follows “*Given the mismatch in spatial resolution between the observation data (30 arc-seconds) and the 5 and 30 arc-minutes simulations were re-gridded to the 30 arc-seconds resolution”. Please see lines 317 – 320.*

240: “snow events to be correctly forecasted” ?

Thank you, this sentence has changed to: “*POD (perfect score = 1) indicates the probability of the observed snow events to be correctly forecasted…*” (line 324).

L 241: not “where” but “are incorrectly”; issue with “since where…”; “ the the”; “information” and not “info”; issue with the end of the last sentence, missing a verb?

Has been corrected to “*…whereas, FAR indicates which fractions of the simulated snow events incorrectly simulated the presence of snow when there was no snow in observed data.*” (lines 325 -328).

258: “relatively” and not “relative”
We agree with the reviewer that relative is not appropriate and have changed the sentence to: “*This allows for inference on whether a simulation improved compared to a benchmark simulation (Towner et al., 2019).*” Please see lines (lines 355-356).

259: “simulations”; “where” ?

This has been corrected, please see line 357.

274: “does, however, …”

This has been corrected, see line 374.

289: 4) ?? (Fig. 4)

This has been corrected, see line 395.

290: “The simulated total water storage changes are comparable between the three resolutions”, seems contradictory with the absolute values of storage changes presented in table 3. So where do you get that from?

Table 3 refers to delta storage from the global water budgets whereas line 290 refers the variable total water storage, we agree with the reviewer that this is not abundantly clear and have changed the section in the updated manuscript, please see lines 397 – 405.

290: “When comparing the total water storage the GRACE satellite observations.” is not a full sentence.
Thank you, this has been corrected. Please see line 498.

297: sentence not clear: “ 50% of 1 676 stations (Fig. A1) display a KGE greater than -0.41 for all simulations”
Thank you, this sentence has been corrected to “*Of the 1 676 stations used for validation (Fig. A1), 50% display a KGE greater than -0.41 (Fig. 6a) - this is true for all three simulations*” Please see line 409.

298: sentence also not clear: “ correlation and variability increases yet this is offset by a reduction in bias”, isn’t a correlation increase AND a reduction in bias what we are looking for? Offset by a variability increase?
Thanks, this has been corrected as follows “*As resolution increases the correlation increases and variability decreases (Fig. 6d).; yet this is offset by an increase in bias (Fig. 6c).”.* Please see lines 411 – 412.

316: “skillful”

Thank you, this has been corrected to “skillful”, please see line 444.

326: issue with “like asynchronous many tasks”
We have reviewed the sentence and reference to “like asynchronous many tasks “ is

appropriate but in response to the reviewers comments we realise that this lacks context. The updated manuscript contains more context please see lines 455 -466.

337: issue with "that is that"

The sentence has been corrected as follows: *"Although this would provide valuable information, it also means that hydrology is now faced with the same issues current GCM's face; namely, that while such simulations are possible, data storage becomes a limitation."* Please line 477.

419: Issue with second sentence, missing verb?
Thanks, it has been corrected to: *"On the other hand, moving to a higher resolution allows for a better match with in-situ observations and more recently released high resolution remote sensing products; the importance of scale commensurability between model outputs and that of in situ observations has been highlighted by Beven et al. (2022)." Please line 616.*

422: "As has been done for smaller scale studies." Not a sentence.

Has been corrected by combining with previous sentence: *"For instance, the caravan dataset which has 6 830 stations for small river catchments (Kratzert et al., 2023), could be used to better underpin the accuracy of simulated river discharge values at higher resolution - as has been done for smaller scale studies (Aerts et al., 2023)."* Please see lines 619.

434, last sentence of the conclusion, issue with the sentence. But important message.

This has been rephrased as *"Overall, the pursuit of hyper-resolution hydrological models is driven by the assumption that they will be able to provide stakeholders with more local estimates of water resources; one promising result reported in this study is that increased resolution is met with more accurate estimates of river discharge". please see lines 631-632.*

Reviewer 2:

OVERVIEW

The paper describes a first attempt to model global hydrology at 1 km resolution, daily, by using the well-developed PCR-GLOBWB model. Results are assessed by comparison with different data sources for soil moisture, groundwater storage, evaporation and river discharge. The comparison of model results at different resolutions is also discussed.

GENERAL COMMENTS

The paper is well written and clear, and the topic is undoubtedly relevant to the ESD readership. However, there are some major points that require further discussion and careful attention.

We want to thank the reviewer for reviewing our manuscript, the reviewer's comments have provided us with valuable insights and suggestions that will improve the quality of this manuscript for future readers. We have addressed the comments below in a pointwise fashion.

MAJOR: The authors should better discuss WHY we want to develop high-resolution simulations of the hydrological cycle. If we model hydrology at 1 km scale, we need to consider processes that are relevant at that scale, and if I well understood are not included in the simulations performed in the paper. I refer, for instance, to the human intervention on the water cycle through irrigation, reservoirs, water diversions. Additionally, at km-scale processes like surface runoff might be highly relevant. These are only some examples. I believe an improved discussion on these aspects should be added in the paper,

The thank the reviewer for this suggestion and agree that the manuscript will be better placed if given more context on why hyper-resolution is important. In the updated manuscript we have added additional emphasis regarding this in the introduction (lines 30 -40). As for the reviewer's suggestion to include more discussion on which processes should be included in hyper-resolution models, we agree that this will be an important addition. In section 4.4 of the original manuscript we provide some opportunities for adding fine-scale processes that are relevant when considering the cryosphere. In the updated manuscript we have extended on this information by elaborating on the importance of including fine-scale information related to landcover type and meteorological forcing data (lines 588 – 594 & 605 - 608).

We then also agree with the reviewer that including which processes this model represents at the 30 arc-seconds resolution is necessary. The parameterisation and input data used here do include high resolution information the reviewer mentions and arises from the work done by Hoch et al 2023. Since this information is attributable to that paper, we have added text that directs the reader towards this information and in addition provides an overview on which processes are represented at the 30-arc seconds resolution "*The model parameterisation and inputs used in the 30 arc-seconds implementation represent high resolution hydrological processes where possible and in the following sections, we provide a summary of these. For extensive details on the setup of the 30 arc-seconds PCR-GLOBWB implementation, we refer the reader to the original European implementation by Hoch et al (2023)"* Please lines 103-125.

MODERATE: A further crucial point for discussion is the manner in which such simulations should be evaluated. While a comparison with in situ data from across the globe is undoubtedly relevant, it is not a suitable approach for performing a reliable assessment. As previously stated, it is necessary to ascertain whether the model is genuinely capable of simulating the processes at high resolution in both space and time. The comparison with river discharge for small basins is useful, but it would be beneficial to conduct stress testing of such modelling simulations in highly disturbed basins, for instance, or in basins affected by processes acting at high resolution, e.g., basins characterised by complex topography. While this is not the focus of this paper (although I would be interested to see some examples), it should be highlighted in the discussion.

We agree with the reviewer that this incorporating this type of analysis would strengthen the manuscript. In the updates manuscript we have included an additional comparison focusing on the accuracy of river discharge for basins of similar size but with different catchment sizes and elevations Please see Fig 10 &

lines 448 - 450.

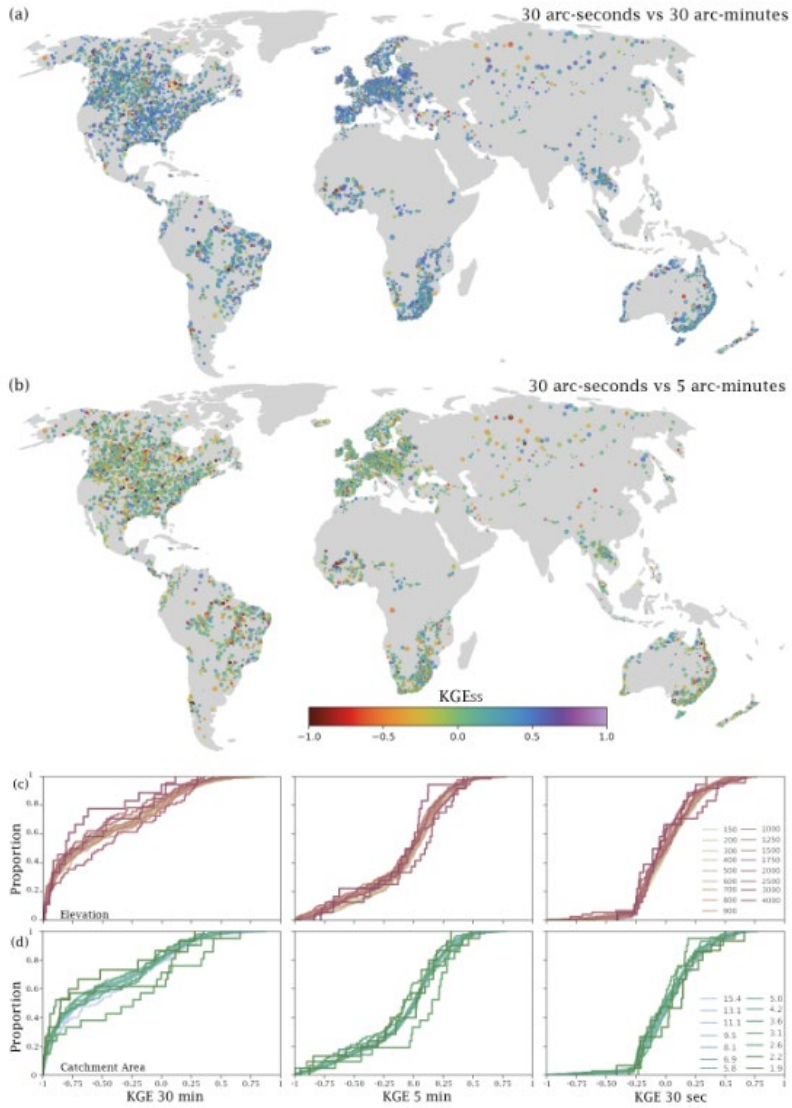

Figure 10. Spatial distribution of improvements in KGE skill score calculated for river discharge simulated using PCR-GLOBWB at the 30 arc-minutes, 5 arc-minutes and 30 arc-seconds resolution. (a) 30 arc-seconds vs 30 arc-minutes (b) 30 arc-seconds vs 5 arc-minutes. KGE cumulative distribution plots for catchments binned according to elevation (c) and catchment area (d).

MAJOR: High-resolution meteorological forcing is obtained by downscaling coarse resolution forcing with climatology (CHELSEA). While this approach is interesting, it is evident that it will not reproduce the high-resolution day-to-day variation of meteorological forcing, particularly precipitation, which is the crucial variable for hydrological modelling. This is briefly mentioned in the discussion but should be better analysed and potential way forward should be highlighted.

We agree with the reviewer that additional emphasis on the way forward regarding the climatology is relevant and should be included, in the updated manuscript a discussion of this has been added to section 4.3 and section 4.4 Please lines 590 – 594 & 605 - 608.

MAJOR: Several typos are present in the text, and some sentences seem to be quickly written, of course the simulation effort was huge, but the writing needs more attention. I spotted some of them in the text, but a carefully re-reading and check is highly necessary. For instance, tables and figures should be put after they are cited in the text, not before.

We thank the reviewer for highlighting these oversights, below we have responded to the suggestions and have proofread the updated manuscript so that it does not contain such oversights and have also corrected some typos that were spotted by other reviewers.

MODERATE: Figure 4 is reported in the text but not discussed and analysed. Please revise.

In the original submission, figure 4 was referenced on above the figure on line 290. However, in light of the reviewer's comments regarding the order of text in relation to figures, we have made sure that figures precede text in the updated manuscript.

MODERATE: To me, the only suitable metric to be used for the comparison between modelled soil moisture and in situ observations should be the correlation (see Figure 6). The bias and the variability ratio depend too much on the climatology that is not relevant. In terms of correlation, the 30 arc-seconds simulations provide much better results, and to me it is highly interesting and it should be better highlighted and discussed.

After reviewing the results in the original submission, we agree with the reviewer's assessment. There will be bias and differences in climatology that are more of a measure of the errors in forcing than the properties of the model. In the updated manuscript, we have included a discussion on this topic in section 4.2 under the "*Soil moisture and evaporation*" section. In addition, more emphasis has been placed on the fact that the correlation increases with resolution. Please see lines 504-516.

MODERATE: It is not clear whether the model has been calibrated against observations, typically in situ river discharge. Is there a calibration for each resolution of the model? I don't think so, and it's strange to see the much better results of the 30 arc-seconds simulation for simulating river discharge. It is not mentioned at all in the paper, which needs to be clarified.

Contrary to the reviewer's suggestion, PCR-GLOBWB is an uncalibrated global hydrological model, nor has the model been tuned for better estimates of river discharge. It is our philosophy that by calibrating a global hydrological model, it would bias the performance to regions with high-quality observations, create inconsistent parameterization between basins and lose the ability to transfer knowledge of model performance from a gauged basin to an ungauged basin. We agree with the reviewer that this is important information to add and as such have added the following text to section 2.2: "*We note that no calibration was performed*

for any of the simulations reported in this study." Please see lines 238.

Regarding the reviewer's statement that it is strange to see better results for the 30 arc-seconds simulation, we agree that the manuscript could be improved by elaborating on why the 30 arc-seconds discharge is better than the other coarser simulations. We have expanded on section 4.2 under the "River discharge" sub-heading as to why we think the estimates were better, in short this is due to the downscaling technique and high-resolution parameterisation contained within the model (please lines 541 – 546). In the updated manuscript we have included more information regarding the parameterisation of the 30 arc-seconds resolution (please see the response to the reviewer's comment starting with "MAJOR: The authors should better discuss WHY…" on page 12 of this document.

In the sequel, a number of specific comments to be addressed is reported, but not a comprehensive list.

SPECIFIC COMMENTS (L: line or lines)

L251: Please quantify how much similar should be, e.g., as percentage difference.

For this step we did not use a threshold percentage difference to assign the GRDC station to the correct grid cell. Instead, for each observation station we identified which cell, in a 5km*5km window, had the smallest difference in catchment area reported for the GRDC catchment and our modelling domain. We agree that this information needs to be added and the updated manuscript contains the following text "*The grid cell within a 5 km window of the station coordinate which had a catchment area closest to that reported by GRDC was selected as the representative point.*". Please see lines 346-347.

L256-257: The sentence is not clear and it should be revised.
This sentenced has been replaced with "*In addition, to obtain information on how the 30 arc-seconds simulation compared in relation to the 5 arc-minute and 30 arc-minute simulations, we calculated the KGE skill score (Eq. 19).*" Please see lines 353 – 355.

L274: Typo.

Thank you this has been corrected, please see line 369.

Table 2: What is NA?

NA is an acronym for not applicable. In the updated manuscript we have changed *"NA*" too *"not applicable."*

Table 3: Why is precipitation higher in the 30 arc-seconds simulation? The downscaling should not produce higher precipitation. Am I wrong? Please clarify.

The higher precipitation in the 30 arc-seconds resolution is due to the downscaling procedure and differences in total land area simulated. The correction factor ensures that precipitation volumes of the coarse scale W5E5 data are matched to the volumes presented in the CHELSA data set. In this case the coarse scale W5E5 data has less precipitation than what is reported in CHELSA and hence the precipitation volumes are inflated when applying the correction factor. This is also evident in Table 4 where the new downscaling technique results in higher precipitation values when applied to both the 5 arc-minute and 30 arc-seconds simulation.

L286: Typo.
In the updated manuscript the sentence reads as follows: *"For the 30 arc-minutes and 5 arc-minutes the results are comparable; however, when considering the 30 arc-seconds simulation, the relative evaporation rates are significantly higher and runoff significantly lower compared to the other two simulations". Please see line 392.*

L287: Should be Table 3 instead of Table 1.
Thank you, this has been corrected in the updated manuscript. Please see line 394.

L291: Typo-
This has been corrected, please see line 384.

L291: The R^2 of the comparison of modelling simulations with GRACE data is quite low, so any conclusion from this comparison is not reliable. Please revise.

We thank the reviewer for pointing this out. In the updated manuscript we have updated the validation which uses GRACE data. The current approach relies on upscaling all resolutions to the resolution of GRACE (i.e., 30 arc-minutes). In the updated manuscript, we have conducted a basin centric approach which involves aggregating basins until the area exceeds 400 000 km^2 (i.e., at least 4 GRACE grid cells). The validation score for the different resolutions are then compared at the basin level. Please see figure 5, lines 397 - 400.

Figure 6: The legend is missing.
In the original submission the legend is in between the plots, but for consistency we have placed the legend at the top as was done for the other similar figures (Please see fig 6 on Pg 22).

L326: The sentence is not clear and it should be revised.

We have reviewed the sentence and reference to “like asynchronous many tasks “ is appropriate but in response to the reviewers comments we realise that this lacks context. The updated manuscript contains more context, please see lines 454 -466.

L323-343: The high-resolution hydrology community should benefit from the remote sensing community who is working for making accessible large datasets (big data) through cloud computing and storage facilities. It can be included in the paper.
We agree that this is a good addition and have added the following sentence to this paragraph: *“To this end, the hyper-resolution hydrology community can benefit by emulating and drawing from the experiences of the remote sensing community who routinely depend on cloud computing and storage facilities to effectively disseminate large volumes of data to end users (Xu et al., 2022).” See lines 480-482.*

L362: Snow cover should be reported after soil moisture and evaporation as in the results.
This is a good suggestion and the order has been corrected in the updated manuscript.

L395-398: The percentage of forests for each spatial resolution can be easily computed. It should be quantified to test if it should be the reason for the higher evaporation in the 30 arc-seconds simulation.

The reviewer rightfully points out that calculating and presenting differences in landcover representation when using dominant vs fractional aggregations would not be difficult and would add support to our hypothesis. In the updated manuscript we aim to show that the area represented by forests are larger using a dominant landcover type compared to fractional. For this we rely on a global 100m landcover dataset by Copernicus (Buchhorn et al 2020).

A cursory analysis using the data set mentioned above reveals that at the 30 arc-seconds resolution the difference in total area represented by forests differ markedly. When considering only a single dominant landcover class per grid cell the forests constitute ~50% of the land surface area; where when we consider the fractional cover per grid cell this value decreases to ~37%, a difference of 13 percentage points (Please see appendix B1).

References:
Buchhorn, M., Smets, B., Bertels, L., Roo, B. D., Lesiv, M., Tsendbazar, N.-E., Herold, M., and Fritz, S.: Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe, https://doi.org/10.5281/zenodo.3939050, 2020.

L399: Typo.
Thank you, we have addressed this typo in our response below. Please see line 586.

L399-400: The sentence is not clear and it should be revised.
We agree that this sentence could be better phrased and have updated as follows: "*Our results suggest that global hydrological models need to incorporate landcover heterogeneity even at the kilometre scale in search of better predictive capacity. In congruence to this observation, landcover representation has previously been shown to be important in providing accurate predictions of hydrological states at the kilometre and even sub-kilometre resolution (Singh et al., 2015; Lazin et al., 2020). It is important to note that although needed for improving the accuracy of predictions, incorporating sub-kilometre landcover heterogeneity would further increase computation times.*" Please see lines 586 – 594.

L424: There is no Sentinel ESA-CCI high-resolution soil moisture product. For a list of high-resolution soil moisture products freely available check this preprint here: https://dx.doi.org/10.2139/ssrn.4737858

Thank you for pointing out this oversight, the updated manuscript has adapted text that allows for reference to the suggested paper "*In addition, recent advances in remotely sensed high-resolution soil moisture data will also be a valuable resource for evaluating hyper-resolution simulations once their time series are sufficiently long (Brocca et al., 2024).*" Please see lines, 621-623.

L434: Typo.

Thank you, this has been revised as follows: "*Overall, the pursuit of hyper-resolution hydrological models is driven by the assumption that they will be able to provide stakeholders with more local estimates of water resources; one promising result reported in this study is that increased resolution is met with more accurate estimates of river discharge.*" Please see lines 631 – 622.

RECOMMENDATION

Based on the above comments, I suggest the paper needs a major revision before its publication.

Reviewer 3:

Jaarsveld et al present 'a first attempt to model global hydrology at hyper-resolution' using a 30 arc-second version of the PCR-GLOBWB model. Overall I think this is an interesting proof of concept for scaling the global model and there are some interesting results. However I do have some concerns that need to be addressed:

We want to thank the reviewer for reviewing our manuscript, the reviewer's comments have provided us with valuable insights and suggestions that will improve the quality of this manuscript for future readers. We will address the comments below in a pointwise fashion.

1. **I take this paper to be a rough proof of concept in terms of the scaling abilities of this model. I think that is a very valuable exercise and is worthy of publication. However, the results themselves show a lot of concerning model behavior and low KGE scores which would be very concerning if this model were used to make any hydrologic evaluations. I understand that is not really the goal of this paper. However, I think it needs to be stated very clearly in both the abstract and conclusions that this model has not been validated for scientific use and should not be used to draw any scientific conclusions. Without these very clear usage notes I worry that this could be applied in very concerning ways.**

Below we respond to the reviewers comment #1 and #3 in union, since they are related. The KGE scores for PCR-GLOBWB are low when compared to calibrated continental or regional scale models. However, the majority of large-scale global models, like the one reported here, do not perform a calibration as this would be too computationally extensive, would bias the performance to regions with high-quality observations, create inconsistent parameterization between basins and lose the ability to transfer knowledge of model performance from a gauged basin to an ungauged basin. The performance of the PCR-GLOBWB2 model has always been on par with similar models from the large-scale modelling community (c.f. Gnann et al. 2023). The philosophy of these models is that they work towards capturing the right processes and do not rely on calibration to correct for errors or biases in process representation, parameterization and/or meteorological forcing data. So, in short, the KGE values, even though they are not high compared to calibrated models, are on par with what can be expected from global hydrological models. To convey this information to the reader we have included additional reference to other (a) global hydrological models and (b) continental hyper-resolution hydrological model. However, it is important to note that direct comparison between the results presented here and other studies are challenging since they differ in the observation data used, metrics calculated and validation approaches.

We kindly disagree with the reviewer that these models should not be used to draw any scientific conclusions, as they provide value about trends, global patterns and response of the hydrological cycle to changes and human-water interactions. This is also confirmed by the fact that these models are extensively used in policy decision or climate assessments like those from the WMO, WRI and IPCC. Global hydrological models have been used in many hundreds of peer-reviewed studies (over 400 with PCR-GLOBWB alone), which is testimony to the value it provides to scientist worldwide. Clearly, we should not overstate the accuracy or value of these

simulations, but they do serve a valuable function in policy making and science. Regarding the request that we include a statement in the abstract and discussion stating that this model not be used to draw any scientific conclusions; we are of the perspective that this would be in contradiction of the goals and outcomes of this study and undermines the contributions that this work makes towards the topic of truly global hyper-resolution hydrological models.

In the original submission we include validation of six variables that allow us to make inference on the quality of predictions and also provides valuable insights that pave the way for future work. We do however agree with the reviewer that this model, in its current form could be further improved to make our high-resolution outcomes more valuable and have included this in the discussion, which already partly tackled in section 4.4 (please see lines 595 – 623).

References:

Gnann, S., Reinecke, R., Stein, L. et al. Functional relationships reveal differences in the water cycle representation of global water models. Nat Water 1, 1079–1090 (2023). https://doi.org/10.1038/s44221-023-00160-y

2. **I would have liked to have seen some comparisons to existing high resolution products. The authors cite many continental scale products that exist for hydrology and snow at 1km resolution. Comparing to point observations is valuable but I would like to know if this model is also capturing the spatial heterogeneities that other well validated models show. I think this is especially important for snow since a new approach is being presented here.**

We agree with the reviewer that understanding how the model outputs may capture spatial heterogeneity is important. However, we are to a large degree limited by the availability of observation data that allows for this type of comparison at the global scale. The reviewer singles out snow as a variable that would benefit from understanding how well the model reproduces spatial heterogeneity. We agree with this notion and in the original manuscript we used MODIS daily 1km snow cover (and not point observations). In the updated manuscript we have included a global map that highlight how well the model reproduces snow cover at the 30 arc-seconds resolution and have included zoomed insets that allows for comparison across the resolutions. In addition to the summary statistics already presented, we added the Brier score to give a more balance performance score (Line 323, section 3.4 on page 23). We also note that for the validation of total water storage, we use GRACE which also allows us to investigate how well the model reproduces spatial heterogeneity and the updated manuscript contains this (please see reviewer 2's comment regarding line 291). The reason that we did not use spatially uniform data for validating the other variables is that (a) such data is not available at the

resolution we require (soil moisture & discharge) or (b) that the data that could be used for validation is based off another modelled estimate (evaporation).

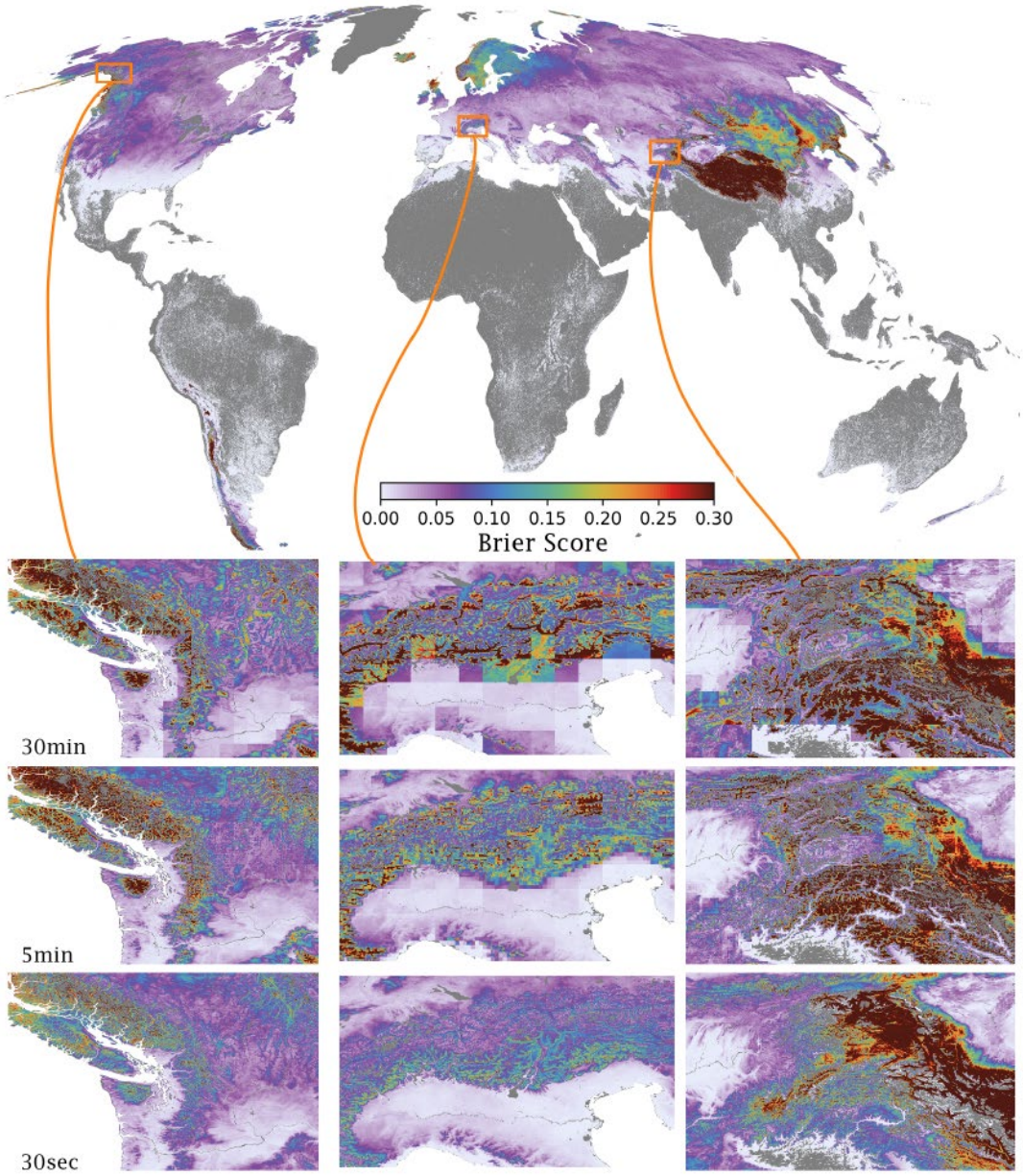

Figure 8. Brier score of simulated snow cover (-) simulated by PCR-GLOBWB at 30 arc-seconds resolution. Zoomed insets compare how the highlighted regions differ between the 30 arc-seconds, 5 arc-minutes and 30 arc-minutes resolution PCR-GLOBWB simulations.

3. **While I think the scalability of this model is a promising step I need to acknowledge that there are a lot of results that are concerning, for example the lack of sensitivity of soil moisture and the low KGE values across the board. To me these indicate that this model is not ready to be used at this resolution even if it is computationally possible (i.e. my first comment). Still though I wonder if the authors could provide some references for the performance of other high-resolution models to put this performance in context?**

Please see our response to the first comment, where we address the reviewers' concerns.

4. **No information is provided on how topography or the stream routing approach was scaled. The assumptions made here would have a big impact on the results I would think. Please provide more information on how the static geo-fabrics for the model were scaled.**

We agree with the reviewer that expanding on which processes this model represents at the 30 arc-seconds resolution is necessary. The parameterisation and input data used here do include high resolution information the reviewer mentions and arises from the work done by Hoch et al (2023). Since this information is attributable to that paper, we have added text that directs the reader towards this information and in addition provide an overview on which processes are represented at the 30-arc seconds. Please see lines 100-127.

Line 221: The KGE threshold of -0.41 seems arbitrary and also very low. Please explain more why this was chosen.
The threshold value of -0.41 is based of work done by Knoben et al. 2019 which shows that when KGE exceeds a value of -0.41 the model is a better predictor than using the mean value. We note that in the original text we included this citation on lines 222 & 254. Also in the figure 6,7,9 legend but without the citation. In the updated manuscript we have updated the text on line 300 & 351 to read as follows "*A perfect KGE score is 1.0, and values greater that -0.41 indicate that the model is a better predictor that using the variable mean value*". In addition, we have added the citation to the legends in Fig 6, 7, & 9 for clarity.

References:

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Figure 8 should list the variable that is being compared explicitly in the caption
We agree with the reviewer that this is necessary and have updated the caption to read as follows, "*Spatial distribution of improvements in KGE skill score calculated for river discharge simulated using PCR-GLOBWB at the 30 arc-minutes, 5 arc-minutes and 30 arc-seconds resolution. (a) 30 arc-seconds vs 30 arc-minutes (b) 30 arc-seconds vs 5 arc-minutes. Panel plots show the relationship between catchment size and KGE skill score; green to purple colours report a better KGE values for the 30 arc-seconds simulation compared to the (a) 30 arc-minute and (b) 5 arc-minute simulation.*"