Answer to reviewer 03

General response

We thank the reviewer for the thorough and constructive assessment of our manuscript. We agree that the original manuscript did not sufficiently distinguish the scientific contribution from the technical benchmarking of ERA5-Land and ISIMIP3a, and that the experimental design does not permit attribution of their differences to individual modelling choices.

In the revised manuscript, we will therefore make clear that the objective is not simply to determine which runoff product performs better, but to examine how two contrasting runoff-production workflows translate into river discharge when evaluated within a common routing framework, and whether their relative performance depends on the aspect of streamflow being considered. In particular, we will more directly connect historical daily streamflow skill with long-term trends in low, mean, and high flows, since these represent different dimensions of hydrological performance that may be relevant to different applications.

We will also add targeted diagnostics of meteorological forcing, pre-routing runoff, and routed discharge to identify where the principal differences between the two workflows emerge or are modified along the modelling chain. These analyses will support a more process-based interpretation while avoiding formal causal attribution to individual factors. In parallel, we will narrow several conclusions, clarify the treatment of anthropogenic influences and ensemble uncertainty, expand the performance and trend diagnostics, and improve the methodological description and presentation.

Reviewer comment 01: *The main conclusion is that ERA5-Land performs better for historical daily streamflow dynamics, whereas ISIMIP3a provides more conservative estimates of long-term trends and a broader representation of uncertainty. However, ERA5-Land and ISIMIP3a differ simultaneously in meteorological bias correction, land-surface or hydrological model structure, calibration, treatment of human influences, ensemble design, and spatial resolution. As the authors acknowledge, these simultaneous differences prevent attribution of the reported performance contrasts to any specific process or modeling choice. Consequently, the analysis remains largely descriptive and provides limited general insight beyond the two products evaluated here. Moreover, the principal result largely confirms expectations already established in the Introduction "Reanalysis-based datasets such as ERA5-Land provide comparatively high spatial resolution and near-real-time coverage, which may better capture fine-scale hydrological patterns (Muñoz-Sabater et al., 2021). In contrast, ensemble hydrological products, such as ISIMIP3, employ bias-adjusted meteorological forcing and multi-model hydrological ensembles designed to reduce systematic forcing biases while preserving large-scale climatic trends and representing structural uncertainty (Hempel et al., 2013; Lange, 2021)". The reported findings are broadly consistent with these expected product characteristics. The manuscript does not yet demonstrate what new hydrological understanding emerges beyond this confirmation.*

Response 01: We thank the reviewer for this important and constructive comment. We agree that the differences between ERA5-Land and ISIMIP3a cannot be attributed to any individual modelling component because the two products differ simultaneously in meteorological forcing treatment, runoff-generation models, ensemble design, spatial resolution, and other aspects of their respective modelling frameworks. We therefore do not interpret our results as a causal attribution of the observed performance differences to any particular process or modelling choice.

We also agree that some of the original wording made the conclusions appear more general than supported by the experiments. We will therefore further moderate our claims and emphasize that the study provides a systematic comparison of two contrasting runoff forcing frameworks rather than a general ranking of reanalysis-based and ensemble runoff products.

At the same time, we believe that the analysis provides insights beyond the general characteristics of the two products described in the Introduction. In particular, our results show that historical daily streamflow skill and long-term trend fidelity represent related but distinct dimensions of hydrological performance, and their relationship warrants explicit evaluation. The ensemble member or forcing that provides the highest KGE for daily streamflow is not necessarily the one that most accurately reproduces the observed long-term trend. This indicates that model performance based on historical discharge dynamics and performance in reproducing long-term changes represent distinct aspects of hydrological skill and should not necessarily be used interchangeably.

More broadly, the study quantifies how these contrasting runoff forcing frameworks translate into river discharge when passed through a common routing model and evaluated against observations from 5,071 gauging stations. We document systematic differences in historical daily streamflow performance, long-term trends in low, mean, and high flows, and the representation of uncertainty across diverse hydroclimatic regions. To strengthen the interpretation beyond a product-level comparison, we will add targeted diagnostics of meteorological forcing, pre-routing runoff, and routed discharge, together with a station-level analysis relating daily streamflow skill to long-term trend error. These analyses will help identify where the principal differences between the two runoff-production workflows emerge or are modified along the modelling chain, while avoiding formal attribution to individual uncertainty sources.

We acknowledge that these results should not be interpreted as demonstrating that any single characteristic of ERA5-Land or ISIMIP3a is responsible for the observed differences. Rather, the contribution of the study is to characterize the consequences of selecting these two widely used runoff forcing frameworks for global streamflow simulation within a common routing framework.

Planned revision 01: We will revise the Introduction and objectives to more clearly define the scope and contribution of the study. Specifically, we will state that our objective is to evaluate how the choice between ERA5-Land and ISIMIP3a runoff forcing affects large-scale streamflow simulations within a common CaMa-Flood routing framework, rather than to identify which individual component of the two frameworks is responsible for the observed differences.

We will also revise the Discussion and Conclusions to distinguish clearly between what is demonstrated by the comparison and what cannot be attributed from the present experimental design. In particular, we will emphasize the finding that streamflow skill assessed from historical daily dynamics does not necessarily indicate skill in reproducing long-term streamflow trends, and that evaluation of runoff forcing should therefore consider both short-term discharge variability and long-term changes.

Finally, we will clarify the role of the spatial-resolution sensitivity experiment. Rather than presenting it as a test of the general importance of forcing resolution, we will describe it more narrowly as an assessment of whether post-hoc aggregation of ERA5-Land runoff from 0.1° to 0.5° can explain the differences observed between the two forcing frameworks. Together, these revisions will sharpen the scientific scope of the study while making its contribution and limitations more explicit.

Reviewer comment 02: *The manuscript does not clearly describe at which stage the ISIMIP3a multi-model mean is calculated. The methods suggest that runoff from the nine hydrological models is routed separately through CaMa-Flood, after which multi-model means are calculated during the regional assessment. However, it remains unclear whether the authors average routed daily streamflow series, station-level performance metrics, or regional statistics.*

Response 02: We thank the reviewer for pointing out this ambiguity.

We clarify that the ISIMIP3a simulations are not merged at the daily discharge time-series level at any stage of the analysis. Each GCM–GHM combination is routed independently through CaMa-Flood, producing an individual daily discharge time series. Performance metrics are then calculated separately for each GCM–GHM combination at each gauging station. Ensemble aggregation is subsequently performed on these station-level performance metrics, rather than on the routed daily discharge series. The resulting station-level metrics are then used for the regional assessment and spatially aggregated as described in the Methods.

Planned revision 02: We will revise the Methods section to explicitly describe this workflow and clarify the different levels of analysis. In particular, we will state explicitly that daily discharge time series are never averaged across ISIMIP3a members; instead, performance metrics are calculated independently for each GCM–GHM combination and subsequently aggregated across ensemble members for the regional assessment. We will also avoid referring to this aggregation simply as a "multi-model mean" where this terminology could imply averaging of the underlying discharge time series.

Reviewer comment 03: *The best ISIMIP3a model is selected using observed streamflow performance within each region and is then evaluated using the same general evidence base. This is an in-sample selection and does not show that the best member can be identified in an ungauged basin, a data-poor region, or an independent period. It does not provide an operational model selection method and has limited practical meaning IMO.*

Response 03: We thank the reviewer for this important point. We agree that selecting the best-performing ISIMIP3a member based on observed streamflow within each region constitutes an in-sample assessment and should not be interpreted as an operational model-selection framework. The purpose of this analysis was to characterize the range of performance represented within the ISIMIP3a ensemble and, specifically, to illustrate an observationally informed upper bound on the performance represented within the tested ISIMIP3a ensemble. It does not demonstrate that the best-performing member can be identified a priori in ungauged or data-sparse regions, nor does it constitute a transferable model-selection method.

Planned revision 03: We will revise the interpretation of this analysis throughout the manuscript and more clearly acknowledge its limitations. In particular, we will describe the analysis as an assessment of the in-sample upper-bound performance within the tested ISIMIP3a ensemble under the available observations, rather than as a practical or operational model-selection approach. We will also reduce any implications regarding the practical applicability of this analysis and explicitly acknowledge that evaluating the transferability of model selection would require an independent validation period, spatial cross-validation, or evaluation in data-sparse or ungauged regions, which is beyond the scope of the present study.

Reviewer comment 04: *If I understand correctly, the ISIMIP3a simulations already include time-varying human influences such as water abstraction, dams, and reservoirs. At the same time, the dam module is enabled in CaMa-Flood for the routing simulations. The manuscript does not clearly explain how reservoir and regulation effects represented in the ISIMIP3a runoff interact with the CaMa-Flood dam module. It is also unclear whether ERA5-Land and ISIMIP3a are treated consistently in this respect.*

Response 04: We thank the reviewer for raising this important point. The apparent inconsistency results from the distinction between human influences on runoff generation and human influences on river routing and discharge regulation. In our simulations, the runoff provided by both ERA5-Land and ISIMIP3a is used as the forcing input to CaMa-Flood. The CaMa-Flood dam/reservoir module is then applied during the routing stage to represent the effects of reservoirs and river regulation on the resulting discharge.

The ISIMIP3a runoff should therefore not be interpreted as a regulated discharge product. In this study, we use only the pre-routing runoff output from the ISIMIP3a models, rather than their routed river discharge. Although the ISIMIP3a hydrological models account for various human water-use processes, which can have indirect effects on simulated runoff generation, for example through groundwater abstraction and associated changes in water storage, the direct effects of dam and reservoir operations on river-flow storage, timing, and release are not carried into the CaMa-Flood input. These effects are represented separately through the CaMa-Flood routing and dam module.

The same CaMa-Flood dam/reservoir module and routing framework are applied to both the ERA5-Land and ISIMIP3a runoff forcings. Thus, the two forcing datasets are treated consistently with respect to reservoir regulation during the routing stage.

Planned revision 04: We will revise the Methods section to clearly distinguish between human influences on runoff generation and the representation of reservoir regulation during river routing, and to explicitly describe how these processes are treated for the ERA5-Land and ISIMIP3a simulations.

In addition, we will conduct a dedicated sensitivity analysis comparing simulations performed with and without the CaMa-Flood dam module. The results will be presented in an additional supplementary figure and summarized in the manuscript. This analysis will assess the extent to which the study results depend on the representation of reservoir regulation and provide evidence regarding the robustness of the principal findings to the inclusion or exclusion of the dam module.

Reviewer comment 05: *The manuscript states that station-level trends are bootstrapped for all datasets, but the ISIMIP3a analysis additionally includes variability across nine hydrological models. It's not clear how these two dimensions are combined. Are model members pooled within the bootstrap, resampled separately, or averaged before station resampling?*

Response 05: We thank the reviewer for pointing out this ambiguity. For the ISIMIP3a simulations, the bootstrap is performed independently for each GCM–GHM model member. Within each model member, station-level trend estimates are resampled with replacement, and the median across the resampled stations is calculated for each bootstrap replicate. Thus, model members are neither pooled before station resampling nor averaged before the bootstrap. After the station-level bootstrap is performed independently for each model member, the resulting model-specific bootstrap estimates are combined across the ISIMIP3a ensemble, and the 2.5^{th}, 50^{th}, and 97.5^{th} percentiles of this combined distribution are used to characterize the uncertainty. Accordingly, this interval reflects both station-sampling variability and inter-model spread, rather than a bootstrap confidence interval associated with station sampling alone.

Planned revision 05: We will revise the Methods section to explicitly describe the two levels of variability in the ISIMIP3a analysis and clarify the order of operations: station-level bootstrap resampling is first performed independently within each GCM–GHM model member, after which the resulting bootstrap estimates are combined across the ensemble to characterize the overall uncertainty.

Reviewer comment 06: *The sensitivity test remapped an existing runoff product, but it does not test how runoff generation would change if the land-surface model and meteorological inputs operated natively at 0.5 degrees. The experiment therefore supports only the narrower conclusion that CaMa-Flood results are relatively insensitive to post hoc aggregation of this particular runoff field. It does not mean that forcing resolution is generally less important than runoff realism.*

Response 06: We thank the reviewer for this important clarification. A similar concern was also raised by Reviewer 1, and we agree that our original interpretation of the sensitivity experiment was too broad. The experiment does not represent a hydrological simulation performed natively at 0.5°, because the ERA5-Land runoff field was first generated at 0.1° and subsequently spatially aggregated to 0.5° before being used to force CaMa-Flood. Thus, the experiment specifically evaluates the sensitivity of CaMa-Flood discharge simulations to the post hoc spatial aggregation of the ERA5-Land runoff forcing, rather than the combined effects of hydrological-model resolution and meteorological forcing resolution on runoff generation.

Planned revision 06: We will revise the relevant sections of the Results, Discussion, and Conclusions to clarify that the observed sensitivity is specific to the tested resolutions, evaluated streamflow metrics, and CaMa-Flood configuration. We will replace broad statements regarding the influence of spatial resolution with more precise language indicating that the aggregation of ERA5-Land runoff from 0.1° to 0.5° had a limited effect on simulated streamflow performance within the modelling framework considered in this study.

Reviewer comment 07: *The presentation quality needs substantial improvement. Figures 1 and 2 are difficult to read at standard journal-page size. The Methods section contains substantial duplicate text. The Introduction is longer and broader than needed for the research question. The opening discussion covers global flood losses, population exposure, the Paris Agreement, future warming, and demographic projections before reaching the specific issue of runoff forcing. Much of this material is generic and does not help define the comparison between ERA5-Land and ISIMIP3a.*

Response 07: We thank the reviewer for this constructive assessment and agree that the presentation and organization of the manuscript require substantially improvement.

Planned revision 07: We will substantially revise and streamline the Introduction by reducing the broader discussion of global flood risk and placing greater emphasis on runoff forcing uncertainty, global streamflow simulation, and the challenges of reproducing both historical discharge dynamics and long-term streamflow trends. We will also revise the latter part of the Introduction to more clearly establish the research gap, highlight the novelty of the study, and explicitly state the scientific objectives and hypotheses being tested.

We will remove the duplicated description of the CaMa-Flood model currently appearing at the beginning of Section 2.2 and retain a single, consolidated description of the model in Section 2.1.2.

We will also reformat Figures 1 and 2 by increasing the size of the global maps and legends and improving their overall layout to ensure that the figures are clearly readable at standard journal-page size.