Response to Anonymous Referee #1

Main points:

The manuscript is much improved, and the authors mostly addressed my 3 main comments. I suggest Minor Revisions.

First, adding the HD results as a baseline improves the paper. The poor HD performance shows a real need for model development, and justifies the study contribution and why it fits in a Model Development journal. It would be good to highlight this more in the manuscript. In the results, I suggest including a figure from one of the locations in the main manuscript (i.e., add one location from Figure S4 to the manuscript).

Second, in their response, the authors make a compelling argument for keeping the WRF-Hydro calibration results as a separate section (rather than integrated as I had asked for). However, I didn't see the author's points reflected in the manuscript, which might help the reader see this more clearly. (I am referring to the points in their response, copied/pasted below; I suggest integrating some of these comments in the manuscript later in my line-by-line comments).

- ***We thank the reviewer for their positive assessment of the revised manuscript and for the constructive suggestions. We appreciate that the reviewer acknowledges the added value of including the HD results as a baseline and the contribution this brings to the model development context of the journal.***

- ***As recommended, we have now included one representative location from Figure S4 in the main manuscript, while retaining the full set of locations in the Supplementary Information.***

- ***Regarding the structure of the manuscript, we note the reviewer's request for a clearer justification of why the WRF-Hydro calibration results are presented as a separate section. We have now incorporated this rationale—previously provided in our response letter—directly into the manuscript to ensure that readers understand the motivation behind this organizational choice.***

- ***Additional details are provided in the line-by-line comments below.***

Third, the paper is improved in clarity, organization, and flow, and I only offer a few suggestions in my line-by-line comments below.

Introduction

Line 106. I would remove "interpreted as a baseline reference". Now that you have added the HD results (which could also be interpreted as a baseline reference) I think this is confusing.

- ***We agree and have removed the phrase "interpreted as a baseline reference" to avoid confusion with the newly added HD results.***

Line 111-112 – replace "with a particular focus on WRF-Hydro" with "focusing on WRF-Hydro".

- ***This change has been implemented as suggested.***

2. Materials and methods.

123. Suggest replacing sentence starting with "Considering", by saying: "The ENEA-REG model, developed withing the MED-CORDEX framework, (Anav et al., 2021)…"

- We have replaced the original sentence with the suggested formulation.

Line 128. Are these the same 10 rivers you mention in line 118? Can you reorganize this paragraph to better connect which rivers are the largest versus which are in ENEA-REG versus the rivers that you actually validate in this study? This would help the flow.

- We have modified both this paragraph and the previous one to improve clarity and flow. The revised text now explicitly distinguishes between: (i) the 18 rivers routed in ENEA-REG, (ii) the largest Mediterranean rivers, and (iii) the 10 rivers used for validation in this study, including the Danube. The updated paragraph also clarifies the rationale for river selection and highlights which basins are representative of major versus medium-sized rivers, as well as different climatic and hydrological regimes.

The revised text now reads:

"The ENEA-REG model (Anav et al., 2021), developed within the Med-CORDEX framework, incorporates into its ocean component river discharge from 18 major Mediterranean rivers simulated by the river routing component. Among these are several of the largest Mediterranean rivers—Rhone, Po, Ebro, Adige, Tiber, and Ceyhan—along with additional basins such as Drin, Maritsa, Goeksu, Vjosa, Jucar, Buyuk Menderes, Arno, Kopru, and Struma. For the Nile, a climatological monthly mean is prescribed, as suggested by the Med-CORDEX protocol (https://zenodo.org/records/11659642), due to the atmospheric model's limited domain coverage of the basin and the significant anthropogenic modifications to its natural discharge.

In this study, the validation focuses on 10 of the ENEA-REG routed rivers, in addition to the Danube River (Fig. 1), chosen based on the availability of at least five consecutive years of daily observations after 1990. These rivers include: Maritsa, Goeksu, Arno, Kopru, Rhone, Po, Ebro, Ceyhan, Adige, and Tiber. Notably, six of these (Rhone, Po, Ebro, Tiber, Adige, and Ceyhan) are also among the ten largest Mediterranean rivers listed earlier, providing a representative mix of major and medium-sized basins. The selected rivers span different climatic and morphologic conditions, ranging from mountainous alpine regions with pluvio-nival hydrological regimes (e.g., Rhone, Po, Adige) to the semi-arid climate of southern Turkey's Ceyhan River. Other river basins were excluded due to insufficient daily observational data or records shorter than five years. The Júcar and Nile rivers were specifically excluded because their flow is heavily influenced by human interventions, such as reservoirs and water diversions, which are not explicitly represented in the modelling setup."

Line 153. It might be useful here to organize the paragraph to introduce the reader systematically to your subsequent experiments for simulating daily discharge. Maybe draw upon your response points (see above). You could also note you are moving from more simple to more complex. You start with simple, CaMaFlood – which only requires runoff and doesn't support parameter calibration. Then you do more complex – WRF-Hydro where the inputs are the set of atmospheric variables, but you do this using default parameters (uncalibrated) so it's a fair comparison. Then, you add complexity by calibrating the WRF-Hydro model parameters.

- ***We have revised this paragraph to clarify the experimental design and the progression from simpler to more complex model setups. The new text now reads as following:***

"To guide the reader through the experimental design, we structured the modelling experiments from the simplest to the most complex configuration. First, we evaluate the ENEA-REG–driven CaMa-Flood and WRF-Hydro in their default setups, ensuring a fair comparison between the two models. These default configurations, described in Sections 2.2.1 and 2.2.2, constitute the core model intercomparison of this study. In a second step (Section 2.2.3), we introduce an additional level of complexity by calibrating key hydrological parameters of WRF-Hydro. This supplementary analysis assesses how much calibration can further enhance discharge performance, providing insight into potential improvements when computational resources permit.

Both CaMa-Flood and WRF-Hydro hydrological models are run for the period 1990-2014, after five years of spin-up. None of the simulations considers reservoir operations and lakes, due to the lack of consistent and comprehensive information on reservoir management and characteristics across all Mediterranean basins. This choice ensures a fair spatial evaluation across the study domain, but we acknowledge it as a limitation that may contribute to reduced performance in some rivers."

Line 200. Seems you need to add something for the reader, either earlier or here or in the next paragraph, referring to your response points (see above) for why you added this section as a stand alone.

- ***We have added a paragraph to clarify the rationale for presenting WRF-Hydro calibration as a separate section:***

"To maintain a fair and balanced comparison between the two models evaluated in this study, the WRF-Hydro calibration is presented as a distinct and complementary analysis rather than as part of the core intercomparison. This decision reflects several methodological and practical constraints. Unlike WRF-Hydro, neither CaMa-Flood nor the Noah-MP LSM within ENEA-REG currently allows basin-specific or spatially variable parameter calibration, and Noah-MP is not two-way coupled with CaMa-Flood in our framework. Additionally, calibration of WRF-Hydro was feasible due to the availability of the PyWRFHydroCalib package, whereas no equivalent calibration tool exists for CaMa-Flood. Given the substantial computational cost of regional-scale calibration, we therefore present the WRF-Hydro calibration results as an "add-on" analysis, illustrating the potential benefits where resources permit while keeping the main model comparison consistent and unbiased."

Line 285. Thanks for adding the HD results, which clearly show how the models you are presenting are a clear improvement. If you have space in the ms, I would suggest putting one of the locations from Figure S4 in the manuscript to more clearly make the point. If you need to swap out a figure to make space, I suggest putting Figure 7, the scatterplot of lowflow and high flow bias, into the Supplemental.

- ***As suggested, we have added the first location from Figure S4 to the main manuscript results to better illustrate model improvements, while retaining the full set of locations in the Supplementary Information.***

Conclusions

Line 541. Suggest explicitly saying that both models evaluated in this paper showed improvements over the HD baseline.

- ***We have updated the Conclusions section to explicitly state that both models evaluated in this study show improvements over the HD baseline.***

Response to Anonymous Referee #2

The manuscript addresses an important topic and has the potential to be a valuable contribution. However, addressing these major technical concerns is essential for its scientific soundness, reproducibility, and suitability for publication in Geoscientific Model Development. The reviewer therefore recommends that this manuscript be returned to the authors for major revisions, as suggested below, before a revised manuscript may be resubmitted for publication.

My comments are given below:

The core of this study is a comparison between CaMa-Flood and WRF-Hydro. However, they are not driven by identical inputs, which invalidates a direct, controlled comparison. CaMa-Flood is forced with daily runoff from ENEA-REG, while WRF-Hydro is forced with 6-hourly atmospheric variables and runs its own Noah-MP LSM internally. Therefore, differences in discharge are a convolution of differences in routing schemes and in generated runoff. Hence, the authors must explicitly state this limitation in the discussion and conclusion. The title and narrative should be reframed from a pure "model comparison" to a comparison of "modelling systems" or "workflows." Ideally, to isolate the routing component, both models should be run with the same runoff input (e.g., the daily runoff from ENEA-REG, which is also fed to CaMa-Flood). If this is not feasible, the manuscript must heavily caveat all comparative conclusions.

- ***We acknowledge the reviewer's concern regarding the use of different inputs for CaMa-Flood and WRF-Hydro and fully agree that this prevents a strict, controlled hydrological benchmark in which both models are driven by identical runoff fields. However, the purpose of this study is not to perform a routing-only comparison under artificially harmonized inputs, but rather to assess the performance and suitability of each hydrological model with the same configuration as it was coupled in a regional model.***

  CaMa-Flood is designed to be coupled downstream of a land surface model, receiving total runoff as its primary input, whereas WRF-Hydro integrates its own Noah-MP LSM, driven directly by atmospheric variables. Forcing both models with the same runoff (e.g., ENEA-REG daily runoff) would produce a technically "fair" benchmark, but would no longer reflect their intended operational roles in regional Earth system modelling. Such an experiment would fall outside the scope and practical purpose of this work, which is to evaluate how these two systems behave as they would be used in a Euro-Mediterranean coupled modelling framework, not under artificially standardized conditions.

  For this reason, we emphasise that the comparison should not be interpreted as a purely hydrological model intercomparison based on identical inputs, but rather as a comparison of modelling systems, evaluating their respective performances, strengths, and limitations within a realistic coupling environment.

  To address the reviewer's request, we have now explicitly clarified this point in both the Discussion and the Conclusions:

  In the Discussion (end of Section 3.1):

  "While this study provides a systematic comparison between CaMa-Flood and WRF-Hydro, it is important to underscore that the two models are not driven by identical types of inputs. CaMa-Flood receives total runoff from Noah-MP within the ENEA-

REG framework, which in our configuration is provided at daily resolution, although the model can also operate with higher-frequency inputs such as hourly runoff. In contrast, WRF-Hydro uses atmospheric fields generated by ENEA-REG at 6-hourly resolution to internally compute runoff through its embedded Noah-MP land-surface model, even though it is likewise capable of operating with finer-scale atmospheric forcing. Consequently, differences in simulated discharge reflect the combined influence of both runoff generation and routing processes. For this reason, the comparison should not be interpreted as a purely hydrological benchmark in which both systems operate under identical inputs. Instead, it reflects how each model performs within the coupling configuration in which it would realistically operate in a regional climate or Earth system modelling environment. In such frameworks, CaMa-Flood is typically coupled downstream of a land-surface model, while WRF-Hydro employs its own land-surface component driven directly by meteorological fields. Forcing both models with identical runoff inputs would not represent their intended operational use and would fall outside the scope of this study. Our aim was therefore to assess their relative suitability and behaviour in configurations consistent with their implementation within regional coupled systems such as ENEA-REG."

In the Conclusions:

"Importantly, this model comparison should be interpreted within the context of their integration into a regional coupled modelling system, rather than as a pure hydrological benchmark with identical runoff inputs or meteorological forcings."

Additionally, to ensure the narrative fully reflects this framing, we have updated the manuscript title to:

"Towards Improved Euro-Mediterranean Discharge Simulations in Regional Coupled Climate Models: A Comparative Assessment of Hydrologic Performance."

This revised title clearly anchors the study within the context of regional coupled modelling workflows, consistent with our objectives and the reviewer's feedback.

Finally, we have also made several reminders throughout the manuscript—in different sections—stating that the comparison concerns model configurations driven by or within regional coupled systems.

The calibrated parameters are evaluated over the entire 1990-2014 period. However, a standard practice is to use a separate, independent validation period (e.g., calibrate on 1995-2004, validate on 2005-2014). Using the entire validation period, including the calibration sub-periods, can lead to overly optimistic performance metrics. Include the parameter bounds. Justify the 1-year spin-up decision or test its adequacy. If possible, perform a proper split-sample validation to strengthen the findings. At a minimum, clearly state that the validation is not fully independent.

- ***We thank the reviewer for this valuable comment. We agree that using a fully independent validation period is the standard recommendation. However, long-term retrospective evaluation is also an established practice in hydrological modeling (e.g., Cosgrove et al., 2024). In our case, the calibration period represents only 5 years out of the 25-year evaluation window, meaning that the vast majority of the assessment period is temporally independent of the calibration.***

Additionally, the purpose of the calibration in this study is not to perform a standalone hydrological calibration experiment, but rather to quantify the improvement of the calibrated WRF-Hydro configuration relative to the default version, which was originally evaluated over the full 1990–2014 period. Using the same evaluation window for both configurations ensures methodological consistency and a like-for-like comparison, fully aligned with the objectives of the study.

We clarify this rationale and clearly state at the end of Section 2.2.3 that the validation is not fully independent.

"The optimized parameters from the calibration were then used to evaluate the model over the entire available period within 1990–2014, focusing on ensuring consistent performance across the selected basins. Although this evaluation is not fully independent—since the 5-year calibration window is included—it remains largely independent, given the 25-year evaluation window, and was chosen, in line with previous literature (e.g., Cosgrove et al., 2024), to ensure methodological consistency with the default (non-calibrated) WRF-Hydro evaluation performed over the same full period. This alignment enables a direct, like-for-like assessment of the improvements attributable solely to calibration. Station-observed streamflow data served as the reference for both calibration and validation, providing a robust basis for model evaluation. Figure S3 in the supplement summarizes the different calibration steps."

Regarding parameter bounds: the parameter bounds are already provided in the "Range" column of Table S3 in the Supplement.

We also thank the reviewer for raising the question of the 1-year spin-up. This is an acclimation phase that was applied after each candidate parameter set to allow the land surface and routing states to adjust to the modified parameters and avoid non-physical transients at the beginning of each simulation. We have added literature that supports the adequacy of such spin-up durations in regional WRF-Hydro applications:

"Each calibration iteration included a one-year spin-up as an acclimation phase to align the model state with current conditions and suppress instabilities from parameter changes (RafieeiNasab et al., 2024; RafieeiNasab et al., 2025)."

The study acknowledges excluding the Nile and Júcar due to heavy human influence, but proceeds to evaluate other highly regulated rivers (e.g., Po, Ebro, Rhone) without any representation of reservoirs or water withdrawals in the models. This is a significant source of bias, particularly for flow timing, low flows, and high-flow attenuation. The discussion must explicitly acknowledge this as a primary limitation. The authors should qualify their conclusions by stating that the model's performance reflects a "naturalized" scenario and that human interventions are a key driver of the observed discrepancies.

- *We thank the reviewer for highlighting this important point. Although several Mediterranean rivers are regulated to varying degrees, the exclusion of the Júcar and Nile basins requires specific justification.*
  For the Júcar, its relatively small drainage area (~22,200 km²) combined with an exceptionally complex and dense regulation system makes it fundamentally different from the other rivers considered. The basin contains multiple major reservoirs—Alarcón, Contreras, Tous, Bellús, and Forata, among others—providing both flood-control and water-supply functions, as well as a cascade of reservoirs for hydropower generation (Molinar, Cortes, Naranjero, and La Muela) and several smaller reservoirs, groundwater–river interactions, and inter-basin water transfers. This extensive degree of regulation dramatically alters both the magnitude and temporal variability of discharge, to the point that the naturalized flow signal is almost fully masked. As a result, the models—configured without explicit representation of reservoirs or managed releases—are unable to reproduce even the basic variability of observed discharge, making meaningful validation impossible.
  For the Nile, in addition to the strong anthropogenic influence, the Med-CORDEX atmospheric domain does not cover the full catchment, preventing a physically consistent simulation of its hydrological cycle. For this reason, the Med-CORDEX protocol prescribes a climatological monthly discharge, and the river cannot be included in the validation of dynamically simulated flows.
  In contrast, while rivers such as the Po, Ebro, and Rhone are also regulated, their basin scale is substantially larger and the natural hydrological signal remains dominant enough for the models to capture seasonal and interannual variability in a meaningful way. Therefore, these rivers can still provide useful insights into model performance.

  This clarification has now been explicitly added in Section 2.1 of the revised manuscript, immediately following the mention of the exclusion of the Nile and Júcar rivers.

  "The Júcar and Nile rivers were specifically excluded because their flow is heavily influenced by human interventions, such as reservoirs and water diversions, which are not explicitly represented in the modelling setup. In the case of the Júcar, its relatively small drainage area (22,200 km²) combined with an exceptionally dense and complex regulation system—including major reservoirs for flood control and water supply (e.g., Alarcón, Contreras, Tous, Bellús, Forata), hydropower reservoirs (e.g., Molinar, Cortes, Naranjero, La Muela), additional smaller reservoirs, river–aquifer connections, and inter-basin water transfers (Momblanch et al., 2014; Suárez-Almiñana et al., 2017)—strongly alters both the magnitude and timing of discharge. These extensive modifications suppress the natural hydrological signal, making it difficult for hydrological models to reproduce even the basic flow variability when compared with observations; meaningful validation is therefore not feasible under the present modelling configuration. For the Nile, in addition to the substantial anthropogenic modifications along its course, the Med-CORDEX atmospheric domain does not cover the full basin, which is why the protocol prescribes a monthly climatological discharge. As a result, the Nile cannot be included in the validation of dynamically simulated river flows."

- ***We agree that the absence of human regulation in both hydrological modelling systems is a major source of uncertainty, especially for Mediterranean rivers where reservoir operations, hydropower management, and irrigation withdrawals strongly affect streamflow timing, low flows, and peak attenuation.***

  To address this comment, we have explicitly incorporated this limitation into the manuscript:

  In Section 3.2.1:

  "Post-calibration improvement was actually expected, particularly because several important sources of structural and input uncertainty are not explicitly represented in our modelling framework.

  First, none of the models account for human interventions such as reservoirs, hydropower operations, and water withdrawals, which can substantially modify the timing, magnitude, and variability of streamflow in many Mediterranean rivers. This omission is primarily due to the lack of consistent and comprehensive information on reservoir characteristics and management practices across all Mediterranean basins, which span multiple countries and governance systems. Consequently, the simulated discharge reflects a naturalized flow regime, whereas the observations often reflect regulated conditions, a well-documented issue for most Mediterranean rivers (Grill et al., 2019). This inherent mismatch contributes to discrepancies in low-flow periods, peak attenuation, and seasonal timing."

  We also stated this limitation in the Conclusions Section.

The authors should provide a deeper analysis, perhaps using the decomposition of KGE (r, β, γ) for specific basins where trade-offs occur, to explain the underlying dynamics. They should discuss the implications of these trade-offs for model application (e.g., is a model with better KGE but worse helpful bias better suited to water resources planning?).

- ***We thank the reviewer for the suggestion. We have added a discussion in Section 3.1 that explicitly considers the decomposition of KGE and the high and low-flow performances for Mediterranean basins. This analysis clarifies the trade-offs in model performance and their implications for different applications, such as mean freshwater fluxes, hydrological extremes, floodplain dynamics, and long-term climate impact projections.***

  "In summary, both models provide added value over the HD baseline and can reproduce Mediterranean discharge dynamics to an intermediate level of skill. Their major applications include: improving freshwater fluxes into the ocean component to reduce biases in sea surface salinity; predicting hydrological extremes; supporting long-term water availability for planning; and projecting future hydrological impacts of climate change. Decomposing KGE into its components and examining high- and low-flow biases helps to explain the underlying trade-offs in model performance across Mediterranean basins. CaMa-Flood tends to produce smoother hydrographs with smaller bias in some basins but underestimates

variability and extremes, whereas WRF-Hydro more accurately captures flow timing and variability but occasionally overestimates peak flows.

These trade-offs have practical implications for different applications. For ocean–land coupling and improving mean freshwater fluxes into the Mediterranean, CaMa-Flood may be sufficient due to its stable estimates of flow volumes. For extremes prediction, which focuses on high- and low-flow events relevant to short-term forecasting and risk assessment, WRF-Hydro is generally better suited due to its richer process representation, more dynamic flow regime, and ability to capture peak flows. CaMa-Flood retains the advantage of simulating floodplain dynamics, including flood depth and extent, which is valuable for flood hazard assessment in smaller or steep catchments. For projections of future hydrological impacts under climate change, which rely on climate model forcings and often assume stationarity of biases over time, both models are appropriate, as moderate differences in extremes do not strongly affect long-term water balance. Additionally, WRF-Hydro is particularly suitable for studies investigating precipitation–soil moisture feedbacks, as it allows full coupling between soil hydrology and the atmosphere, capturing critical land–atmosphere interactions.

Basin characteristics further influence model suitability: WRF-Hydro performs well in large temperate basins, while CaMa-Flood provides stable, conservative estimates in smaller or steep catchments. Overall, model choice should be guided by the intended application, basin characteristics, and the balance between process detail and computational efficiency."

Section 3.1 is very dense and descriptive, repeatedly listing metric values for each model in each basin. It becomes difficult for the reader to extract the main message. The section would benefit from a stronger narrative structure. Start with a high-level summary paragraph stating the main findings (e.g., "Overall, the two models showed comparable timing but systematic differences in variability and bias..."). Then use the figures (Taylor diagrams, etc.) to support these overarching conclusions rather than describing each data point in the text. The text should interpret the statistics, not duplicate them.

- ***We thank the reviewer for this suggestion. Section 3.1 has been revised to provide a clearer narrative structure. A high-level summary paragraph now highlights the main findings across basins, focusing on systematic differences in timing, variability, and bias between the models. The detailed metrics are now presented in the figures, and the text emphasizes interpretation and discussion of the results rather than repeating individual values.***

I feel the authors need to provide some discussion on the limitations of the study in the discussion section and what needs to be done to address those. For example, a comprehensive analysis of climate sensitivity, given the datasets (that are used in the study), has a significant uncertainty amongst their models. Although I see that authors have mentioned this in their manuscript, it still needs a more detailed discussion with the help of recent literature.

- ***We thank the reviewer for this suggestion. In the revised manuscript, we have added a discussion at the end of Section 3.2.1 addressing the main limitations of our study and potential avenues for improvement.***

"Post-calibration improvement was actually expected, particularly because several important sources of structural and input uncertainty are not explicitly represented in our modelling framework.

First, none of the models account for human interventions such as reservoirs, hydropower operations, and water withdrawals, which can substantially modify the timing, magnitude, and variability of streamflow in many Mediterranean rivers. This omission is primarily due to the lack of consistent and comprehensive information on reservoir characteristics and management practices across all Mediterranean basins, which span multiple countries and governance systems. Consequently, the simulated discharge reflects a naturalized flow regime, whereas the observations often reflect regulated conditions, a well-documented issue for most Mediterranean rivers (Grill et al., 2019). This inherent mismatch contributes to discrepancies in low-flow periods, peak attenuation, and seasonal timing.

Second, locally calibrated scores help compensate for errors in catchment boundaries (Kauffeldt et al., 2013; Lehner, 2012), primarily due to the coarse resolution of elevation data, as exemplified by the Adige basin. From a purely hydrological perspective, this resolution is inadequate, and a higher resolution on the order of hundreds of meters would be preferable to better represent small basins. However, this is constrained by the very high computational resources required for simulations covering a large domain, such as Med-CORDEX.

Third, calibration also mitigates errors in meteorological forcing, particularly in precipitation data (Beck et al., 2017, 2019), as shown in Sect. S1 and Fig. S8 in the supplement. The regional coupled models generate meteorological forcing at relatively coarse spatiotemporal resolution and rely on parameterized convection, known source of uncertainty and errors. In contrast, convection-permitting models, km-scale models that explicitly resolve deep convection, can more realistically represent sub-daily statistics and extreme events (Fosser et al., 2024; Struglia et al., 2025), highlighting an additional limitation of regional models in capturing high-flow and short-duration extremes. Further efforts in bias correction or data assimilation of climate forcing could improve hydrological model performance, pending the extension of convection-permitting models to large domains such as Med-CORDEX, given the substantial computational resources required.

Readers should interpret the findings within this context, and future work incorporating anthropogenic controls, higher-resolution runs, and refined meteorological forcing could further improve model performance."

Conclusions are not clarified. Besides, these findings are not well supported by the data or the results.

- ***We thank the reviewer for this observation. The Conclusions section has been revised to provide clearer statements that are directly supported by the results presented in the manuscript.***