Response to Anonymous Referee #2

The paper by Hamitouche et al. presents the performance of two hydrological models, namely CaMa-Flood and WRF-Hydro, driven by the ENEA-REG atmosphere-land-ocean coupled model run at a 12km resolution over the Med-Cordex domain. The performances analyzed concern default (uncalibrated) versions of both models, as well as a calibrated version of WRF-Hydro. While the authors highlight the higher performance of the calibrated version of WRF-Hydro, they present this work (to my understanding) as a preliminary step for further application "for offline atmosphere-hydrological simulations to close the water cycle at the land-ocean interface" (L526).

In brief, the key findings presented in this paper are the performances of two uncalibrated models and one calibrated model, which utilize data from an offline climate model as input, on a series of Mediterranean basins. These findings can be, of course, useful, but presented alone, rather than well inserted in a workflow providing much more advanced results, are no more than calibration exercises or, in the case of the uncalibrated models, simple applications of already existing hydrological models, not providing any novelty or relevant contribution to the field. Indeed, model calibration is a preliminary and unavoidable necessity for any hydrological model. The two research questions presented (LL95-97) have quite obvious answers, especially the second.

- ***We respectfully disagree with the reviewer's assessment that the study lacks novelty. The selection of an appropriate hydrological model is a critical step for accurately representing rivers and discharge simulations into global or regional Earth System Models. In this context, our work provides the first model-to-model comparison at the Mediterranean scale, offering an alternative to the pre-existing and less compatible HD model. This represents a key contribution by identifying solutions to improve river representation and freshwater fluxes into ocean models, which are central to coupled Earth system modelling efforts in the Med-CORDEX framework.***

  Regarding the research questions, we acknowledge that their answers may appear generally intuitive. However, the results demonstrate that the extent of improvement following calibration is not trivial. For example, in three basins, the default parameters already yielded optimal performance, and calibration did not produce further improvement. This underscores that calibration outcomes are basin-dependent and not universally beneficial, which is an important and non-obvious finding for guiding future modelling applications in this region.

I strongly suggest that the research presented be strengthened by associating this initial evaluation/calibration step with a more meaningful analysis (e.g., long-range hydrological reanalysis, sensitivity analysis, or exploring issues related to hydrological extremes). The results achieved so far seem too preliminary.

- ***We thank the reviewer for this suggestion. We agree that more advanced analyses—such as long-range hydrological reanalyses, sensitivity experiments, or the study of hydrological extremes—are highly relevant and form a natural continuation of this work. Indeed, we are currently developing a follow-up study that substitutes the HD model with the more advanced hydrological model assessed here, in order to investigate future projections of hydrological extremes over the Euro-Mediterranean***

region. However, the scope of such an analysis is too broad to be integrated into the present manuscript.

The current study was designed as a preliminary and essential step: evaluating and calibrating hydrological models for their suitability within Euro-Mediterranean coupled modelling systems. This type of contribution is directly aligned with model evaluation papers in Geoscientific Model Development (GMD), which explicitly include the comparison of the performance of different model configurations or parameterizations as a recognized manuscript category (see GMD guidelines). We therefore consider the presented analysis to be sufficiently substantive for the scope of this journal, while laying the foundation for more advanced applications in subsequent work.

Furthermore, in reviewing the study, I suggest considering the following comments.

1) Introduction: It is unclear why these two hydrological models were chosen and not (also) others. In addition, especially concerning WRF-Hydro, there is a vast amount of references describing significantly more advanced research, in which the calibration issue has been overcome with interesting strategies to be taken into consideration, simulating even multiple basins simultaneously for a large number of years, both in one-way and fully-coupled modes. A basic search on Scopus reveals approximately 300 documents containing "WRF-Hydro" in the Article Title, Abstract, or keywords.

- ***We thank the reviewer for this observation. The choice of CaMa-Flood and WRF-Hydro is motivated by both scientific and technical reasons, including the modularity of the model which allows to easily couple them with atmospheric and ocean components. Besides, our selection builds on a previous study in which we analysed the impact of different Noah-MP runoff schemes on discharge simulations using CaMa-Flood. That study showed overall good performance against observations but also revealed important limitations, such as delays in capturing seasonal peak flows due to inherent constraints in CaMa-Flood. These limitations were successfully resolved by WRF-Hydro when tested in the same framework (Gochis et al., 2021), as stated in LL66–75. This provided a strong rationale for the present model-to-model comparison focusing on the most important Mediterranean rivers.***

  From a scientific perspective, both models are highly relevant for the Euro-Mediterranean regional coupled modelling framework. WRF-Hydro is the hydrological extension of WRF and uses Noah-MP as its land surface scheme. Since WRF and Noah-MP are the default atmospheric and land surface components of Med-CORDEX regional coupled models such as ENEA-REG, WRF-Hydro represents a natural candidate for improved river representation. CaMa-Flood, on the other hand, has been extensively validated for large-scale discharge simulations and provides a robust benchmark for comparison, as stated in LL82–85.

  Regarding calibration, our objective was not to identify the best calibration strategy, but rather to assess the extent to which calibration can improve discharge simulations in this regional framework. While advanced calibration studies on WRF-Hydro exist, these are often constrained to smaller basins, limited numbers of parameters, or short

simulation periods due to computational demands. For example, Sofokleous et al. (2023) calibrated 31 small mountainous catchments (5–115 km²) in Cyprus, while Tijerina-Kreuzer et al. (2021) compared WRF-Hydro and ParFlow over the CONUS domain without calibration, simulating only one water year at 1 km resolution at the cost of a full year of supercomputing. In our context, the basins of interest are much larger, reaching hundreds of thousands of square kilometres, and many Mediterranean sub-catchments remain ungauged, making detailed sub-basin calibration infeasible. In addition, we would underline that this study represents a first attempt to use WRF-Hydro over the whole continental Europe.

Finally, we acknowledge promising recent approaches such as the iterative Ensemble Smoother (iES) applied to WRF-Hydro by RafieeiNasab et al. (2025), which provides ensemble calibration with uncertainty bounds. While powerful, such methods are computationally very expensive and beyond the scope of the present study. Our focus is therefore on demonstrating the potential of calibration within a tractable and regionally relevant setup.

2) One model requires daily runoff as input, while the other uses 6-hour meteorological data. One model is calibrated, the other not. Any comparison between these two models looks quite unbalanced. I would propose that the results of the analysis be presented as a sequence of steps to further improve the hydrological output (even though I don't consider the uncalibrated analysis interesting).

- ***We thank the reviewer for this important observation. To ensure fairness in the comparison, we treated the WRF-Hydro calibration as a separate analysis, presented in its own subsection. This structure avoids mixing calibrated and uncalibrated results and makes the rationale for the calibration exercise clearer, as reflected in the revised section headings and flow of the paper.***

  Regarding model inputs, we clarify that CaMa-Flood and WRF-Hydro operate differently by design. CaMa-Flood is a hydrodynamical model that only relies on runoff input, whereas WRF-Hydro is a full hydrological framework embedding Noah-MP to simulate runoff and route it to discharge. Within ENEA-REG, CaMa-Flood is coupled to WRF, where Noah-MP generates runoff from the same 6-hourly meteorological forcing used by WRF-Hydro. This ensures that both models rely on Noah-MP runoff driven by the same atmospheric inputs, thereby making the comparison consistent.

3) The calibration and validation methodology is rather unclear. Reading LL231-235, the reader cannot understand for each basin analysed how many years were used for calibration and how many for validation. The sentence "The optimized parameters from the calibration were then used to evaluate the model over the entire 1990–2014 period" suggests that the years used for calibration were also used for validation. A simple Table would have helped. No hydrographs are shown in the entire manuscript, nor in the Supplementary Material. In addition, a detailed comparison between ENEA-REG and WRF-Hydro runoff is lacking (I can only see Fig. S4, which shows the mean seasonal cycle for only one river).

- *We thank the reviewer for pointing this out. The calibration was performed for 5 years (as stated in L229), plus a one-year spin-up at each iteration (L230). The specific calibration periods varied across basins depending on the availability of continuous observations. Validation was then carried out over the entire available period within 1990–2014 (excluding missing values), using the calibrated parameters. For example, in the Danube basin, the calibration period was 01/10/1999–01/10/2005 (including one-year spin-up), while the evaluation extended over 1990–2014.*

  To clarify this aspect, we added a new Supplementary Table (Table S4) reporting the spin-up and calibration periods for each basin, and we revised LL231–234 accordingly.

  Revised LL231–234:
  "The specific spin-up and calibration periods varied across basins and were selected based on the availability of reliable and continuous observational records (i.e. at least 5-year all falling within the 1990–2014 period), and are provided in Table S4 in the supplement. 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."

Table S1: Summary table of spinu-up and calibrations periods for each river basin

Basin	Spin-up period	Calibration period (including one-year spin-up)
Danube	01/10/1994 – 01/10/1999	01/10/1999 – 01/10/2005
Rhone	01/10/1985 – 01/10/1990	01/10/1990 – 01/10/1996
Po	01/10/2000 – 01/10/2005	01/10/2005 – 01/10/2011
Ceyhan	01/10/1995 – 01/10/2000	01/10/2000 – 01/10/2006
Adige	01/10/2000 – 01/10/2005	01/10/2005 – 01/10/2011
Tiber	01/10/1992 – 01/10/1997	01/10/1997 – 01/10/2003
Maritsa	01/10/1994 – 01/10/1999	01/10/1999 – 01/10/2005
Goeksu	01/10/1995 – 01/10/2000	01/10/2000 – 01/10/2006
Arno	01/10/2000 – 01/10/2005	01/10/2005 – 01/10/2011
Kopru	01/10/1995 – 01/10/2000	01/10/2000 – 01/10/2006
Ebro	01/10/1994 – 01/10/1999	01/10/1999 – 01/10/2005

Additionally, discharge hydrographs have been included in the Supplement (as Fig. S4), showing both CaMa-Flood and WRF-Hydro simulations compared to observations and to the HD model (as a benchmark).

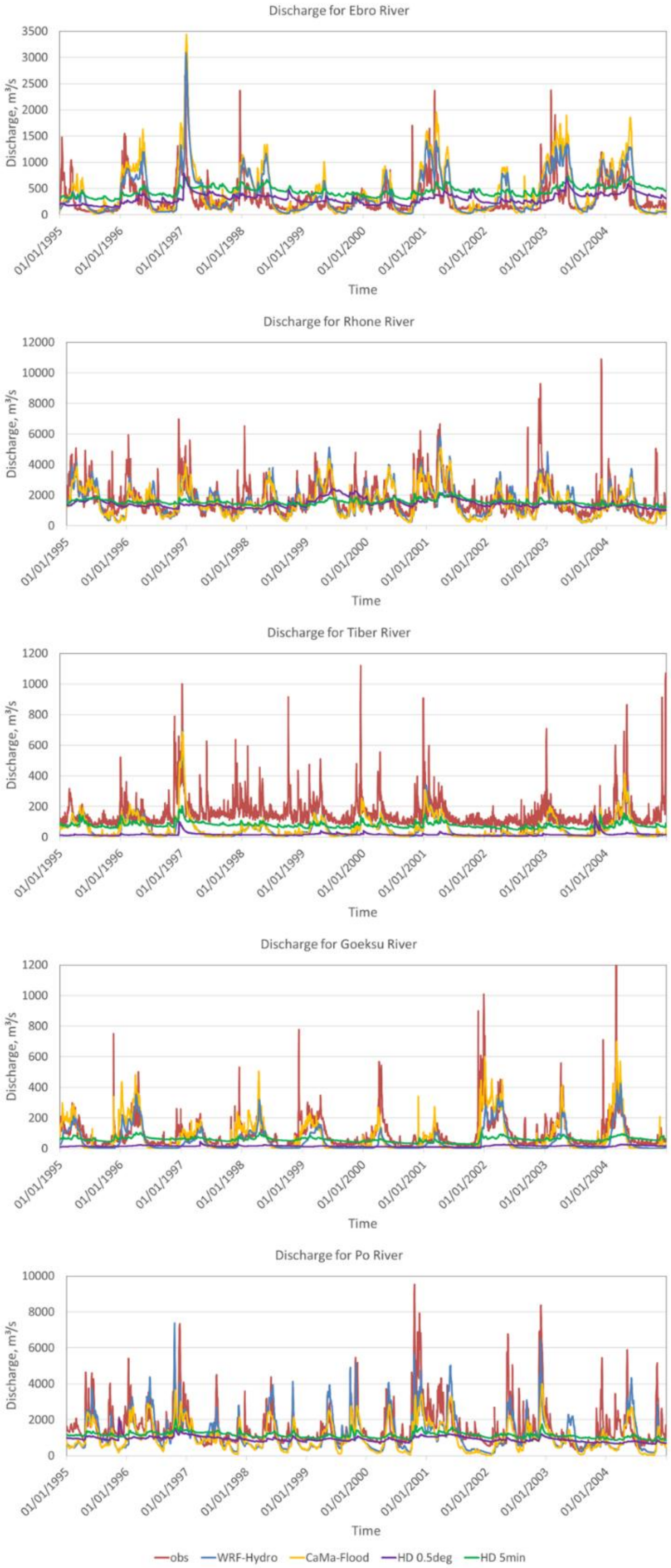
Discharge for Ebro River
Discharge, m³/s
Time
Discharge for Rhone River
Discharge, m³/s
Time
Discharge for Tiber River
Discharge, m³/s
Time
Discharge for Goeksu River
Discharge, m³/s
Time
Discharge for Po River
Discharge, m³/s
Time
obs
WRF-Hydro
CaMa-Flood
HD 0.5deg
HD 5min

Figure S4: Observed and simulated daily discharge for the Ebro, Rhone, Tiber, Goeksu and Po rivers for common 10 years from 1995 to 2004. Simulations include ENEA-REG–driven WRF-Hydro and CaMa-Flood, as well as the HD model at 0.5° and 5-minute spatial resolutions, evaluated near the corresponding gauge stations.

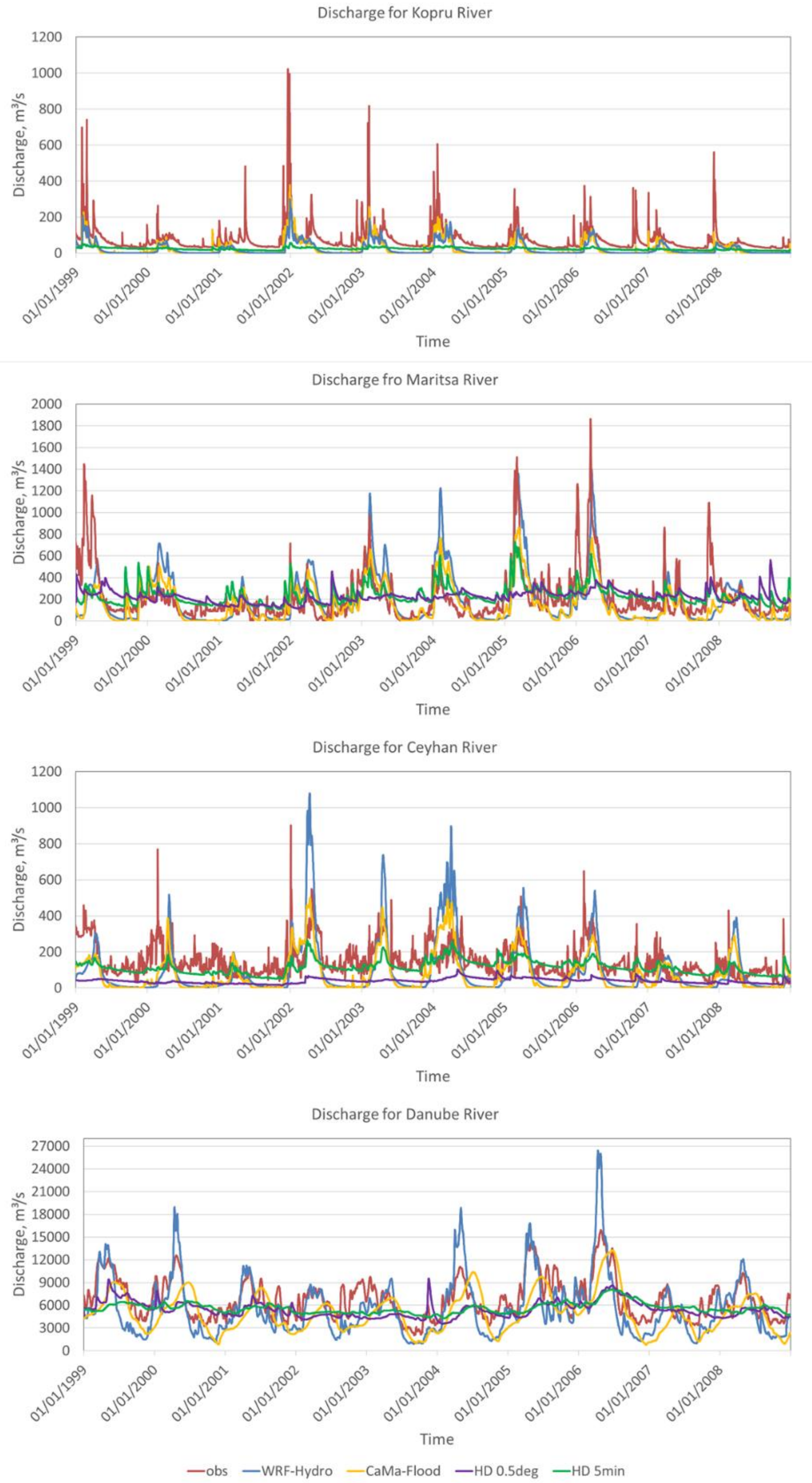

Figure S4 (continuity): Observed and simulated daily discharge for the Kopru, Maritsa, Ceyhan and Danube rivers for common 10 years from 1999 to 2008. Simulations include ENEA-REG–driven WRF-Hydro and CaMa-Flood, as well as the HD model at 0.5° and 5-minute spatial resolutions, evaluated near the corresponding gauge stations.

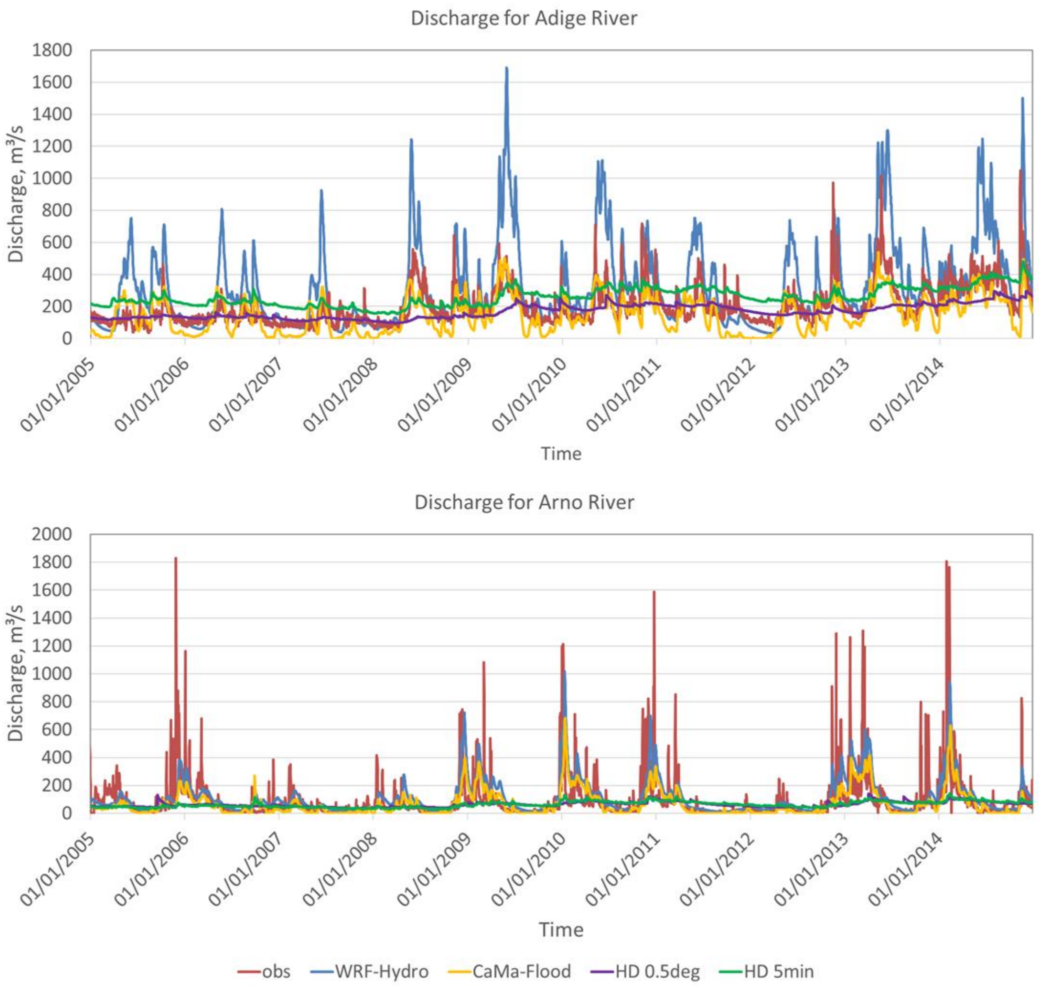

Figure S4 (continuity): Observed and simulated daily discharge for the Adige and Arno rivers for common 10 years from 2005 to 2014. Simulations include ENEA-REG–driven WRF-Hydro and CaMa-Flood, as well as the HD model at 0.5° and 5-minute spatial resolutions, evaluated near the corresponding gauge stations.

- ***Finally, the comparison of ENEA-REG and WRF-Hydro runoff was previously presented in Supplementary Table S5 and discussed in LL466–489. Now it has been revised and is provided in a dedicated subsection ("3.2.2 Effects of calibration on runoff generation and partitioning"). Instead of Supplementary Table S5, we now include a clearer grouped bar chart (as Fig. 9 in the main manuscript), consisting of two subplots: (a) average total runoff and (b) subsurface-to-total runoff ratio, with each group of bars corresponding to a river basin.***

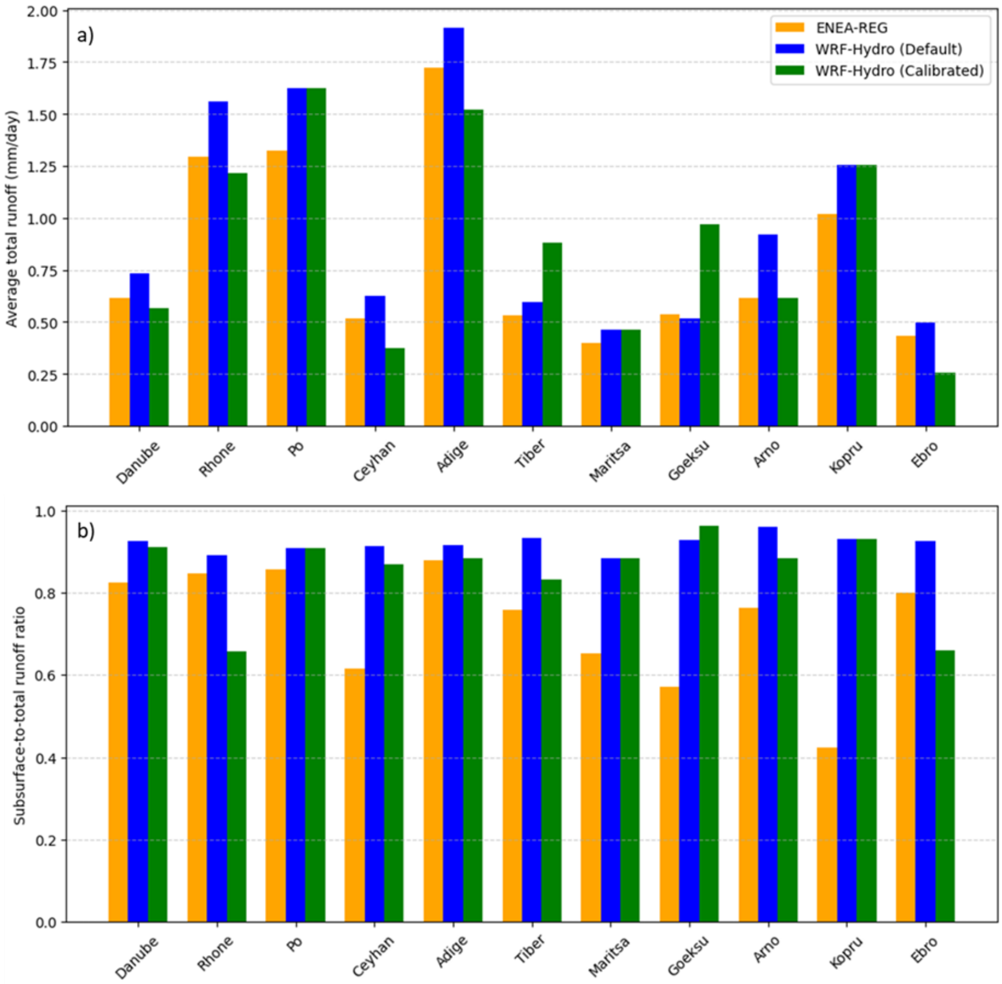

Figure 9: Comparison of a) average daily total runoff and b) the ratio of subsurface to total runoff across each river basin as simulated by ENEA-REG (orange), WRF-Hydro with default parameters (blue), and WRF-Hydro with calibrated parameters (green).

4) In addition, concerning the calibration strategy, the values of the calibrated parameters should be duly analysed and discussed. For example, the parameter smcmax represents the maximum soil moisture content for each soil type. Why is it higher than 1 in most basins? Please consider the issue of equifinality seriously while dealing with a multi-parameter calibration.

- ***We thank the reviewer for this comment. The values of the calibrated parameters, including smcmax, and their influence on hydrological processes are already analysed and discussed in detail in the Supplement (Sect. S1 (LL15-27), Table S4). As stated there, smcmax is a scalar multiplier of soil porosity with a valid range of 0.8–1.2, so values above 1 remain consistent with the model formulation and correspond to increased infiltration and soil evaporation. In addition, our discussion explicitly addresses the issue of equifinality, noting that parameter interactions can compensate or dominate each other's effects, ultimately determining the hydrological outcome.***

LL15-27 in the supplement:

"The calibration process revealed some notable overall trends across all basins. There

was a consistent increase in refkdt, which controls runoff partitioning and results in a higher proportion of subsurface runoff compared to surface runoff. Additionally, the slope parameter consistently decreased, reducing deep drainage from the soil column to the groundwater reservoir. The zmax parameter showed a general increase across most basins, except for Göksu, leading to reduced baseflow contributions to river discharge.

While these three parameters exhibited relatively uniform behaviour across all basins, others displayed more variable trends. For instance, parameters such as bexp, smcmax, dksat, and rsurfexp showed both increases and decreases depending on the basin, reflecting local hydrological conditions.

This variability underscores the complexity of the calibration process. The interplay between parameters often determines the final hydrological outcome, as the effect of one parameter can be moderated or even compensated by another influencing the same physical process. In some cases, one parameter may dominate over others, masking their effects. These findings highlight the importance of considering the combined influence of multiple parameters to ensure accurate and balanced calibration for hydrological modelling."

5) L146: None of the simulations considers reservoir operations and lakes. That's probably one reason why some of the rivers are simulated very poorly. The rivers' flow should be preliminary naturalized.

- ***We thank the reviewer for raising this important point. Indeed, none of the simulations include reservoir operations or lakes. This decision was taken primarily because of the lack of consistent and comprehensive information on reservoir management and characteristics across all Mediterranean basins, which span multiple countries and governance levels. To ensure a fair spatial evaluation across the study domain, we chose not to incorporate reservoirs. We agree that this is a limitation and acknowledge that the absence of reservoir regulation is likely one reason for poor performance in some rivers.***

  Revised LL146-147:

  "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."

6) L178: Hydrological routing model components of WRF-Hydro were run on the same 6 km spatial resolution grid. Such a resolution is very low for most of the Mediterranean catchments. Additionally, this approach does not utilize one of the main features of WRF-Hydro, namely subgrid disaggregation and aggregation.

- ***We agree with the reviewer that a 6 km spatial resolution is coarse from a hydrological perspective and does not exploit one of the main features of WRF-Hydro, namely subgrid disaggregation and aggregation. However, in the context of Euro-Mediterranean coupled models, even coarser resolutions are typically used (e.g., 0.5° ≈ 55 km or 5′ ≈ 11 km). Thus, the 6 km grid still represents a step forward in improving river representation and discharge simulations. While higher resolutions are certainly preferable, they are strongly constrained by the availability of computational resources. For our purposes, the chosen resolution was a compromise that allowed us to meet the objective of closing the water cycle at the land–ocean interface across the regional scale.***

7) L200-202: "snowmelt parameters were left at their default settings due to the limited relevance of snowmelt in most of the study region and for potential future regionalization over other snow-free basins." This sentence sounds very strange for basins for which the Alps or the Pyrenees are fundamental. Probably, this is another reason why some of the rivers are simulated very poorly.

- ***We thank the reviewer for raising this important point. Snowmelt processes indeed play a significant role in several Mediterranean basins, particularly the Ebro, Po, Rhone, and Danube. In this study, snowmelt parameters were left at their default settings to ensure consistency across basins and fairness in the intercomparison, and for possible regionalization over other snow-free basins. We acknowledge that this choice may have limited performance in snow-influenced basins. However, it is worth noting that even without explicit calibration of snow-related parameter, the performance in these basins was satisfactory, with Kling–Gupta Efficiency (KGE) scores improving from intermediate to good after calibration (e.g., Danube: 0.46 → 0.66; Rhone: 0.54 → 0.65; Ebro: 0.46 → 0.65). This suggests that while snow processes are important, the calibrated parameters affecting other hydrological processes (e.g., infiltration, runoff partitioning) also played a major role in improving performance. We now explicitly acknowledge in the text that future work should investigate snowmelt parameter calibration to further refine simulations in these catchments.***

  Revised LL199-202:
  "In this study, calibration focused on the first four groups, covering 16 parameters (Table S3), while snowmelt parameters were left at their default settings to ensure consistency across basins and fairness in the intercomparison, and for potential future regionalization over other snow-free basins. We acknowledge that this choice may limit performance in snow-influenced basins and that future work could further refine results by including these parameters."

8) Conclusions: Regarding calibration for fully coupled atmosphere-hydrological models, please note that many papers demonstrate that it is based on different concepts compared to offline calibration, and it should not be directly based on observed meteorological inputs.

- ***We thank the reviewer for this comment. Following your suggestion, we have revised the text to remove the phrase "based on observed meteorological inputs" and now refer only to a tailored calibration approach.***

References:

Gochis, D. J., Barlage, M., Cabell, R., Casali, M., Dugger, A., Fitzgerald, K., Mcallister, M., McCreight, J., Rafieeinasab, A., Read, L., Sampson, K., Yates, D., and Zhang, Y.: The WRF-Hydro Modeling System Technical Description. (Version 5.2), NCAR Technical Note, 108 pp, https://ral.ucar.edu/sites/default/files/docs/water/wrf-hydro-v511-technical-description.pdf (last access: 27 February 2025), 2021.

Sofokleous, I., Bruggeman, A., Camera, C., and Eliades, M.: Grid-based calibration of the WRF-Hydro with Noah-MP model with improved groundwater and transpiration process equations, J Hydrol (Amst), 617, 128991, https://doi.org/10.1016/j.jhydrol.2022.128991, 2023.

Tijerina-Kreuzer, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M. M., Sampson, K., Gochis, D., and Maxwell, R.: Continental Hydrologic Intercomparison Project, Phase 1: A Large-Scale Hydrologic Model Comparison Over the Continental United States, Water Resour Res, 57, e2020WR028931, https://doi.org/10.1029/2020WR028931, 2021.

RafieeiNasab, A., Fienen, M. N., Omani, N., Srivastava, I., and Dugger, A. L.: Ensemble Methods for Parameter Estimation of WRF-Hydro, Water Resour Res, 61, e2024WR038048, https://doi.org/10.1029/2024WR038048, 2025.