Authors' Response to Editors/Reviewers of

A Distributed Hybrid Physics-AI Framework for Learning Corrections of Internal Hydrological Fluxes and Enhancing High-Resolution Regionalized Flood Modeling

Huynh et al.
HESS,

ED: ***Editor Decision***, **RC:** ***Reviewers' Comment***, AR: Authors' Response, □ Manuscript Text

Dear Editors and Reviewers,

We greatly appreciate your time and effort in handling our manuscript. We extend special thanks to the reviewers for their thorough and constructive comments, which have significantly improved our work.

We have revised our paper in accordance with the comments and recommendations from both reviewers. In addition, we carefully re-read the entire manuscript to double-check and correct any typographical errors.

We hope these revisions have strengthened our paper and made it ready for publication in HESS.

Ngo Nghi Truyen Huynh, on behalf of the authors.

ED: ***This manuscript presents a robust and novel hybrid modeling framework that integrates physically based and neural network components within a spatially distributed hydrological model. Both reviewers found the study well-written and scientifically sound, commending its methodological rigor, physical interpretability, and clarity in model comparisons. Reviewer 1 raised thoughtful suggestions on model benchmarking and statistical significance, which the authors addressed thoroughly in their planned revisions. Reviewer 2 endorsed the manuscript with minor technical corrections. Given the reviewers' positive assessments and the authors' comprehensive and constructive responses, I recommend acceptance pending minor revisions.***

AR: Thank you for handling our manuscript. We acknowledge your positive decision and have made the necessary minor revisions as suggested by the reviewers. Below, we provide a point-by-point response addressing the reviewers' comments. Additionally, we have made minor corrections to address typographical and formatting errors throughout the paper.

1. Reviewer 1

1.1. Major comments

RC: ***Model comparison: The authors compare a stand-alone GRU model and their hybrid model approaches. I think the comparison between these two models is necessary, valuable, and the results are presented clearly. However, the fact that the hybrid performs better than the stand-alone process-based model is expected. With the hybrid approach, you have a model with more degrees of freedom, and the embedded NN can compensate for structural deficiencies in the process-based part, which will increase performance.***

AR: We appreciate your feedback regarding the model comparison. Indeed, we have clarified this point by adding a paragraph in the discussion to address your concerns.

> Although hybrid models have more degrees of freedom compared to classical GR models, it is important to note that the inputs and outputs of the flux correction model are physically consistent and of the same dimension as the original model. This design allows the hybrid model to learn nonlinearities in

the internal flux laws, which we analyze thoroughly in the flux correction analysis in both time and space throughout the paper. The hybrid models do not necessarily have more conceptual parameters (maintaining the same number of reservoirs and connections here), they do introduce more nonlinearity in the internal flux laws corrections with neural network ϕ_1. This added complexity effectively increases the model's degrees of freedom while maintaining robustness in both spatial and temporal validations, as demonstrated by numerical results.

RC: ***What I think is missing, to have a better idea of where the hybrid stands, is a comparison with a purely data-driven approach. For example, having a stand-alone LSTM, trained regionally with lumped meteorological inputs (e.g., catchment average values) would be a good benchmark. Or use as inputs, not only the basin-averaged values of precipitation, temperature, etc,... but also include other basin-averaged statistics (mean, std, max and min) that you can compute from the gridded products. This way we can see how the hybrid approach performs against purely data-driven methods, and if the extra effort of going distributed is worth it.***

AR: Thank you for your suggestion. This work focuses on a spatially distributed conceptual model based on physics and its physical hybridization. Our emphasis is on the rigorous presentation and analysis of this framework over a large sample, its performance in calibration and spatio-temporal extrapolation, and the interpretability of internal fluxes corrected with the hybrid approach. You mentioned that the comparison between the GR and the hybrid GRNN models is "necessary, valuable, and the results are presented clearly." First, analyzing the question of lumped versus spatialized models is not within the scope of this study. A spatially distributed approach is essential given the high spatial variabilities involved within the flash flood-prone catchments of this dataset.

Second, unlike traditional and hybrid process-based models, which rely on physical conservation equations of mass, momentum, energy, and empirical closures, pure LSTM or ML models do not inherently impose physical constraints, leading to reduced physical interpretability and generalizability, especially under extreme or unseen hydrological conditions (Beven, 2020; Sit et al., 2020; Shen, 2018).

Given this context, we believe that a comparison with a stand-alone LSTM is beyond the scope of this study, which focuses on the hybridization and physical interpretability of spatially distributed process-based models. Our goal is not to benchmark the hybrid approach against purely data-driven methods but to demonstrate the improvements achieved through the hybridization of a well-established, spatially distributed, differentiable numerical hydrological model. A benchmark represents the scope of another full study.

Moreover, building a stand-alone LSTM, operating at an hourly time step, trained regionally, and capable of accounting for basin-average forcings/descriptors and spatial information, would be interesting but would represent a significant undertaking based on our previous experience with daily LSTM models (Hashemi et al., 2022). This is the scope of another full study. Such pure deep learning models, with even more degrees of freedom than our relatively parsimonious hybrid model, are hardly interpretable or extrapolable beyond the training set - as other neural networks. Therefore, we believe that a comparison with our proposed method is not useful. The present paper features a rigorous presentation and detailed analysis, especially of optimized quantities and internal fluxes, over a large sample and in regionalization. We appreciate your suggestion and will keep it in mind for future benchmarking studies that compare a broader range of modeling approaches.

RC: ***Section 3: Here you present two datasets, with which you run two sets of experiments. The first dataset includes 235 non-nested catchments in France with 13 years of data. In this dataset, you test the effect of having a NN for process parameterization. In the second dataset, you have 21 catchments in the Mediterranean region, both nested and independent, with 7 years of data. This one you use to test the model regionalization. Is there a reason why this last test cannot be made in the first dataset? One can evaluate regionalization from catchment to catchment, and not only inside the same catchment. Moreover, having results in 235 catchments for the second experiment will give more robust tests. Also, you can mix everything in a single dataset with 256 catchments. I was just wondering why did you make this division?***

AR: Thank you for your valuable feedback. Here is a detailed explanation for the use of both datasets:

- **Dataset with 235 catchments:** test the performance of ϕ_1 (NN for flux correction) using local calibrations only at downstream gauges. This setup demonstrates the efficiency of the ϕ_1 NN in improving model performance; *Reason for separate testing:* conducting a multi-catchment setup across the entire mesh of France is computationally challenging, given the high spatio-temporal resolution of the data and model (see Huynh et al. (2024), for details on computational costs of the adjoint model). Moreover, this would require investigating cost functions adapted for meaningful information selection/weighting over a set of catchments with contrasted area and physics (see issues for regionalization with downstream catchments in Huynh et al. (2024)).
- **MedEst dataset:** Evaluate regional calibration in a multi-catchment setup. This dataset tests the performance of both ϕ_1 (flux correction NN) and ϕ_2 (regionalization NN); *Reason for focused regionalization:* analyzing physical interpretability in a national multi-catchment setup is complex. For this initial study, we focused on regionalization performance within a specific and well-known study zone. It is worth noting that regionalization performance over a larger zone, covering approximately 1/4 of France, has already been studied in Huynh et al. (2024). Future studies can certainly explore a national multi-catchment setup, as you suggested.

By separating the datasets, we aimed to provide a clear and focused analysis of both the flux correction and regionalization capabilities of our hybrid model. We have provided these clarifications in the revised version of our manuscript, which should address your points:

> While the first dataset aims to test the performance of the neural network ϕ_1 for flux correction using local calibrations only at downstream gauges over Metropolitan France, the second aims to evaluate regional calibration in a multi-catchment setup, testing the performance of both ϕ_1 (for process-parameterization) and ϕ_2 (for regionalization). Note that performing a global multi-catchment calibration (e.g., at the national scale across the entire mesh of France) for process-parameterization and regionalization with neural networks is computationally challenging, given the high spatio-temporal resolution of the data and model. In this study, we focus on a specific and well-known study zone. It is worth mentioning that regionalization performance (without process-parameterization network) over a larger zone, covering approximately one quarter of France, has already been investigated in Huynh et al. (2024). Future work could extend this study to a national-scale multi-catchment setup.

RC: ***Line 282-294: The differences between the models are quite small. For example, the difference shown in Figure 4 between the median NSE for the GR.U and the GRNN.U is 0.008 and between the GRD and the GRNN.D is 0.014. Are the differences between the reported distributions statistically significant? I think this point should be further discussed. Because the hybrid approach has a higher flexibility than the process-based model. The embedded neural networks produce flux-correction parameters for each pixel and timestep, and if the differences between the hybrid and the stand-alone process-based model are small, it would be interesting to find out why. Maybe the physical dissipation of the basins makes it unnecessary to have so much detail, if one is just interested in the simulated discharge at a specific point. Or maybe the meteorological data is restricting further increases in quality.***

AR: We appreciate your insightful comments regarding the differences between the models and the significance of the reported distributions. We have revised the text to better explain why the differences, though small, are still valuable and promising.

> In temporal validation, GRNN.U achieves a median NSE of 0.73 compared to 0.76 for GR.D, and both models reach a median KGE of 0.75, while GRNN.U shows a lower median RMSE of 1.38 compared to 1.42 for GR.D. Although median improvements may appear small, it is important to consider the entire distribution. In addition to the median values, we observe notable enhancements in other statistical measures, such as the interquartile range (0.25 and 0.75 quantiles) and whiskers in the boxplots. For

> catchments that already exhibit satisfactory performance, the effect of hybridization is relatively small, leading to similar median, 0.75 and 0.95 quantile values. However, for poorly performing basins, the hybrid models provide substantial improvements, as evidenced by enhanced performance in the lower quartiles. Notably, the hybrid model with spatially uniform hydrological parameters (GRNN.U) performs comparably to, and in some cases surpasses, the classic GR model with spatially distributed parameters (GR.D). This result is promising, as it demonstrates that while the original model with spatially uniform conceptual parameters (GR.U) inherently leads to under-parameterization compared to GR.D, this limitation can be compensated for by the spatially distributed flux correction in GRNN.U.

RC: ***As an additional question, do the flux correction parameters allow the model to artificially increase/decrease the amount of water (violate the mass-conservation principle) in the control volume?***

AR: Thank you for this valuable comment. The simulated water balance is influenced by the correction of k_{exc}, which represents the exchange flux and can result in either gains or losses relative to the original model (which is already non-conservative). We added this clarification in section 4.2.2 in the revised version.

> The simulated water balance is influenced by the correction of k_{exc}, which represents the exchange flux and can result in either gains or losses relative to the original model (which is already non-conservative). In this case, the exchange flux is moderately affected by hybridization, with a median trend of reduced exchange from -23.5 mm to -13.2 mm. This reduction is compensated, in terms of water balance, by an increase in evaporation from the production reservoir (from 254.5 mm to 265.1 mm in median). Notably, both fluxes exhibit a larger interquartile range across basins compared to the classic model structure. Therefore, the proposed ϕ_1-hybridization enables the learning of spatio-temporal corrections of internal model dynamics, resulting in physically interpretable fluxes that remain within imposed ranges and lead to overall model improvement.

RC: ***Line 295-300: You indicate that you are evaluating the performance of the model in flash floods, and then you evaluate it in 2700 events during the 6-year validation period. Are these 2700 events flash floods or just regular floods? How did you classified them?***

AR: Thank you for this comment. The selection of flood events is performed using an automatic segmentation algorithm based on Huynh et al. (2023), which identifies events where peak flows exceed a certain quantile threshold. Therefore, we acknowledge that these 2700 events are not necessarily flash floods but rather general flood events that include flash floods. To ensure clarity, we removed the term "flash" from the text.

RC: ***Line 327-333: In these lines (and Figure 7) you compare the NSE for 143 flood events, indicating that the hybrid models perform better. Even if this is true, all the models performed quite badly. For the GR.U and GRNN.U the median NSEs are -0.48 and 0.09, which is a clear indication that the models do not work at all. Just taking the mean of the observed data would yield to a NSE of 0. For the other two models, the NSE did improve, but was still quite low (0.19 and 0.37). You should expand the discussion here and try to understand why all models are performing so badly.***

AR: Thank you for this comment. We acknowledge that the NSE values computed for the 143 flood events are relatively low across all models. We have expanded the discussion here to explain why the models did not perform as expected in this case.

> Although the NSE values computed for the 143 flood events are relatively low across all models, it is important to note that NSE for flood events–which are short time series with high values–is highly sensitive to small timing errors. Even slight discrepancies in peak timing can lead to substantial decreases in NSE. Additionally, data and modeling uncertainties may vary between events, making the accurate prediction of highly contrasted events particularly challenging. The models are calibrated on the entire time series, and we evaluate the validation results specifically for flood events, where

classical approaches often struggle to accurately estimate water dynamics. This discrepancy highlights the difficulty in capturing the rapid and intense nature of flood events, even with advanced hybrid models. This underscores the need to investigate potential sources of error, including input data quality and model structural limitations, as well as the impact of using a calibration metric based solely on flood events. These factors could explain the overall challenges in flood event simulation.

1.2. Minor comments

RC: ***Line 61: Clarify "This study".***

AR: This is clarified.

RC: ***Line 72: Replace "have to be advanced" to "should advance"***

AR: This is corrected.

RC: ***Line 74: What do you mean by "earth critical zone"?***

AR: Earth critical zone refers to the near-surface environment where complex interactions between water, soil, rock, etc. regulate the Earth's surface dynamics. We added this clarification in the text.

RC: ***Line 136: The purple color of the parameters is almost red. I would suggest choosing another color scheme, more colorblind-friendly.***

AR: We agree and have changed the purple color to the brown color to make it more color-blind friendly.

RC: ***Line 183: What do you mean by neutralized atmospheric inputs?***

AR: We refer to net rainfall or evapotranspiration, which is neutralized by the interception reservoir. This terminology is specific to GR models, as referenced in Perrin et al. (2003) and Santos et al. (2018). We have added this clarification to both the figure and the text.

RC: ***Line 278-279: You indicate about Figure 3 "The results demonstrate the superior accuracy of hybrid methods compared to the classic models…" but it is not clear from the Figure, because one cannot see any details. There are certain peaks in which the hybrid is better, but you have 6 subplots, each with 5 years of hourly discharges, so you cannot really appreciate much. I imagine that if one looks at specific events, sometimes the hybrid is better, sometimes both are similar, and sometimes the process-based is better. Maybe print only a subset of the testing period, or specific events where the differences are significant. Then, with general metrics you can make the point on which model tends to perform better.***

AR: We agree and thank you for this comment. The comparison between methods does not need to be emphasized at this stage. We have revised the sentence to make it more precise.

RC: ***Line 285-290: I would separate more clearly (in different paragraphs) the results reported in calibration and validation. It is not a usual practice to compare models using results in the calibration period, as any meaningful comparison should be made in validation. If you want to report the results in calibration that is perfectly ok, but a more clear distinction should be made.***

AR: Thank you for your comment. Yes, we have separated and revised the paragraphs to ensure better clarity and understanding.

RC: ***Line 290: The RMSE for GRNN.U, according to Figure 4, is 1.38 not 1.30. You should correct this in the text.***

AR: Thank you for pointing this out. The value of RMSE for GRNN.U has been corrected.

RC: ***Figure 5. Is the Ebf metric (baseflow) a good/necessary indicator for performance during flood events?***

AR: Thank you for this comment. We agree that the baseflow may not be the most direct indicator for evaluating performance during flood events. However, we would like to clarify that we have already included other metrics specifically dedicated to flood performance, such as peak flow and flood volume. Additionally, the baseflow is computed during flood events, not the entire series, providing valuable insights into low-flow dynamics during high-flow periods. Therefore, we believe retaining the Ebf metric enhances the overall evaluation of flood performance.

RC: ***Line 323-326: It is not clear what you want to say.***

AR: We want to highlight that the hybrid structure with uniform parameters (GRNN.U), which does not rely on physical descriptors to estimate conceptual parameters, already achieves acceptable results. By avoiding the use of additional data such as physical descriptors (which may contain uncertainties), this approach can mitigate the problem of data error in modeling. Therefore, the development of a semi-lumped hybrid structure appears promising, alongside rationalization approaches that use physical descriptors. We have lightly modified the text for clarification.

RC: ***Line 352: You indicate that "Some spatial patterns in these corrections seem to emerge across France, and although analyzing trends in corrections as a function of physical explanatory factors may yield insights, it is beyond the scope of this study focusing on detailed quantitative analysis of those spatio-temporal corrections". What are the spatial patterns showing in Figure 8? Because for me they are not so clear. Also, why analysing the correction factors as a function of physical characteristics is out of the scope? I think this is one of the most interesting parts you should focus on. If one of the advantages of hybrid models is that they produce physical interpretability, then one should interpret what the models are doing.***

AR: Thank you for this comment. We agree that the term "spatial pattern" in the original paragraph may have been confusing. We have rewritten the paragraph to clarify the observations.

> These maps reveal different trends of flux corrections across France. Several regions exhibit strong corrections (either negative or positive) for P_s and E_s, while others show near-zero corrections (white points with values close to 1). However, the exchange flux is generally the most influenced by the corrections, as indicated by the dark colors across the maps, playing a crucial role in refining the model's state dynamic.

RC: ***Line 361: You indicate that "a majority of exchange flux corrections fq,4 that share the same sign as fq,1." Can you quantify this with a metric? Because from the figure is not obvious. fq4 shows more red in the bottom, but I am not sure if the majority of cases are in accordance.***

AR: Thank you for this comment. We agree that the statement lacks quantitative evidence, and since this interpretation is not central to our findings, we have removed it from the text to avoid any confusion.

RC: ***Line 265: You indicate that "periodic behaviors are observed over time in all four heatmaps". For fq4 I cannot distinguished clear periodic behaviors. Could be useful to plot the timeseries of some basins. Maybe include them in the appendix.***

AR: Thank you for this comment. We agree that the heatmaps are not easily interpretable, as also noted by another reviewer. To address this, we have replaced this figure by time series plots of typical basins, as you suggested. This change makes the findings clearer. We have also clarified and revised the text accordingly. However, we would like to retain the heatmap figure and move it to the appendix.

RC: ***Line 385 and Figure 10b: You indicate that "Interestingly, these maps also reveal spatial variability in internal flux corrections." It would be interesting to analyse why these patterns emerge.***

AR: Thank you for this comment. We have added a sentence to interpret the spatio-temporal patterns found in the flux corrections based on the inputs of the neural network ϕ_1. Specifically, we explain that temporal patterns arise from the periodic behaviors of the model states, while spatial patterns emerge due to the spatial variability of atmospheric data.

RC: ***Line 424: Rephrase "Also, one could also..."***

AR: This has been corrected.

2. Reviewer 2

2.1. Major comments

RC: ***Figure 9: While it's important to understand the internal fluxes of your system, Figure 9 is confusing. It appears there is no temporal differences in flux per catchment, so the X axis is distracting. Further, the ordering of the fluxes is not clear. I don't think this figure is required, as a spatial understanding of flux, as shown in Figure 10b is clearer to the reader***

AR: Thank you for this comment. We agree that the heatmaps in Fig. 9 are not easily interpretable, as also noted by reviewer 1. To address this, we have replaced the heatmaps with time series plots of typical basins, as the suggestion of reviewer 1. We have also clarified and revised the text accordingly. However, we believe the heatmap of flux correction is an interesting result since it provide a global view of the results over large sample datasets and thus we would like to retain the figure of the heatmaps in the appendix.

2.2. Minor comments

RC: ***Some of the latex math functions regarding TanH in section 4.2 were messed up when translating to PDF***

AR: This has been corrected.

References

Beven, K., 2020. Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes 34, 3608–3613. doi:.

Hashemi, R., Brigode, P., Garambois, P.A., Javelle, P., 2022. How can we benefit from regime information to make more effective use of long short-term memory (lstm) runoff models? Hydrology and Earth System Sciences 26, 5793–5816.

Huynh, N.N.T., Garambois, P.A., Colleoni, F., Javelle, P., 2023. Signatures-and-sensitivity-based multi-criteria variational calibration for distributed hydrological modeling applied to mediterranean floods. Journal of Hydrology 625, 129992. doi:.

Huynh, N.N.T., Garambois, P.A., Colleoni, F., Renard, B., Roux, H., Demargne, J., Jay-Allemand, M., Javelle, P., 2024. Learning regionalization using accurate spatial cost gradients within a differentiable high-resolution hydrological model: Application to the french mediterranean region. Water Resources Research 60, e2024WR037544. doi:.

Perrin, C., Michel, C., Andrèassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. Journal of hydrology 279, 275–289.

Santos, L., Thirel, G., Perrin, C., 2018. Continuous state-space representation of a bucket-type rainfall-runoff model: a case study with the gr4 model using state-space gr4 (version 1.0). Geoscientific Model Development 11, 1591–1605.

Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research 54, 8558–8593. doi:.

Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I., 2020. A comprehensive review of deep learning applications in hydrology and water resources. Water Science and Technology 82, 2635–2670. doi:.