We thank the reviewers and the editor for their thorough and constructive review. In this reply, we have copied the comments in black. Our responses are entered in red.

Reviewer 1

Major comments

Section 2.2.3:

In this section, the authors explain the different models used in their study. In general, the output from the GR4J model, and several additional variables, are used as input of different machine learning algorithms, that act as postprocessors (Figure 3). However, the performance comparison of the different models has errors.

As a benchmark, the authors are using the GR4J, which is a rainfall runoff model that receives meteorological input and predicts discharge. However, models 2c, 3, 4, 5 and 6, besides the predictions made by the GR4J, also use observed discharges as input to the ML algorithm. One cannot compare a model that receives observed discharge as an input with a model that does not, it is expected that the former one will be better. Discharge is a highly temporally correlated variable, so the discharge from time t-1 is an extremely good predictor for the discharge at time t. This is why in the results, they report that "Model 2c demonstrates improvement with respect to GR4J, which emphasises the importance of taking into account short-term streamflow memory." This is not a surprising finding and makes the model comparison invalid between models that receive discharges and models that do not.

We thank the reviewer for their insightful comment regarding the comparison of models that include observed discharge as input versus those that do not.

Indeed, Models 2c, 3, 4, 5, and 6 incorporate observed streamflow data as predictors in the machine learning (ML) post-processing step, whereas the baseline GR4J model does not. As the reviewer correctly points out, this means that these ML models leverage short-term streamflow memory, which is a strong predictor due to the high temporal autocorrelation in discharge data. The significant improvement observed in models such as 2c is therefore expected and consistent with established hydrological understanding.

Our intention in structuring the study this way was to explore how incorporating various hydrological memory components, both short-term (recent streamflow) and long-term (runoff coefficients), could address the known limitations of the GR4J model, particularly under changing climatic conditions like droughts. The ML models are applied as

post-processors precisely to diagnose and quantify these potential improvements, and to help identify specific structural weaknesses in the GR4J model.

We acknowledge that direct comparison of model predictive skill between the baseline GR4J and ML models that use observed discharge input should be interpreted with caution. The key contribution of this work is not to claim superiority of the ML models as standalone predictive tools, but rather to demonstrate:

1. The extent to which GR4J predictions can be improved by augmenting them with hydrologically meaningful predictors, including short-term discharge memory.

2. The insights gained from ML models about the importance of such predictors, which suggest directions for future model development and refinement of physically based rainfall-runoff models.

Regarding short-term prediction, it is well recognised in hydrology that recent streamflow (e.g., discharge at time t-1) provides valuable information for forecasting at time t due to strong temporal correlation. Our results confirm this, as models incorporating short-term discharge memory achieve substantial gains in prediction accuracy. This outcome validates the use of short-term streamflow as a critical component in improving model responsiveness to recent hydrological conditions, especially in dynamic drought and post-drought periods.

To emphasise this, we will clarify in the revised manuscript that the performance gains of models using observed discharge input are indicative of the potential to incorporate these memory effects within GR4J or similar models to improve robustness across varied hydrological conditions.

We appreciate the reviewer's comment, which has helped us improve the clarity and framing of our analysis.

Section 3.

In line 333, the authors indicate that "None of the models 1, 2a, and 2b show any improvement over the GR4J predictions". This is contrary to what has been shown in literature, where using ML models as postprocessors of process-based models improves performance (Frame et al, 2021) because of the enhanced flexibility of the

resulting hybrid model. Nevertheless, the results shown by the authors are contrary to that. Further explanation of why this is the case is required.

Moreover, in the case reported by the authors, the models are performing badly. Based on Figure 7b and 7c, models 1, 2a and 2b reported a negative NSE for 60% (or more) of the basins. This indicates that just taking the average flow is better than the model, and consequently, the models are not working at all. Why is this the case?

We thank the reviewer for raising this important point about the unexpectedly poor performance of Models 1, 2a, and 2b compared to the GR4J baseline, which contrasts with findings in the literature (e.g., Frame et al., 2021) where ML post-processing typically improves hydrological model performance.

There are several factors that likely contribute to the observed results in our study:

1. **Predictor Selection and Model Structure:**
Models 1, 2a, and 2b rely on a limited set of predictors that do not include observed streamflow (discharge) data as inputs, unlike other models in our study. Their predictors mainly consist of meteorological variables and GR4J simulated discharge without additional memory terms. This limits their ability to capture important temporal dependencies and hydrological memory effects that are crucial for accurate daily streamflow predictions, especially under highly variable climatic conditions such as drought.

2. **Hydrological Complexity and Variable Climate Conditions:**
The catchments studied exhibit complex rainfall-runoff dynamics, strongly influenced by drought and post-drought periods where streamflow patterns deviate markedly from pre-drought conditions used for model training. Models 1, 2a, and 2b are less flexible in capturing such non-stationarities because they lack critical hydrological memory predictors (e.g., recent discharge), resulting in models that are not robust to these changes. This leads to poor generalisation and, consequently, negative NSE values in many sub-catchments.

3. **Inadequate Representation of Nonlinearities:**
The combination of predictor variables and model formulations in Models 1, 2a, and 2b may not sufficiently address the non-linear relationships inherent in rainfall-runoff processes under drought conditions. Without streamflow memory and long-term runoff coefficient terms, these models effectively perform worse than even a naïve average flow predictor in some basins.

4. **Training Data and Period Differences:**

Our models were trained exclusively on pre-drought data, which is wetter and hydrologically different from the drought and post-drought periods used for evaluation. This temporal mismatch exacerbates the poor performance of simpler models, particularly those lacking additional hydrological memory inputs.

In lines 370-372, the authors indicate that GR4J model lacks the capacity to represent low flow context, because the other ML algorithms performed better. However, this is again an unfair comparison because all the ML algorithms that you are using in this comparison receive discharge as input, which will be an extremely informative predictor of the discharge in the next time step, especially during low flow periods. Therefore, this is not a valid comparison.

We appreciate the reviewer's insightful observation regarding the comparison of low-flow predictions between the GR4J model and the ML algorithms.

Indeed, the ML models in our study incorporate observed discharge as an input predictor, which provides a strong temporal correlation and valuable information for predicting streamflow at the next time step, particularly during low-flow conditions. This inherently gives the ML models an advantage over the GR4J model, which does not use observed discharge as input.

We acknowledge that this difference in input data means the comparison is not fully "apples-to-apples." However, the intent of our study was to evaluate the potential for machine learning to act as a post-processor that can correct and improve GR4J predictions by leveraging additional hydrological memory effects (including short-term streamflow memory), which are difficult for conceptual models like GR4J to represent.

The improved performance of the ML algorithms during low-flow periods highlights the value of incorporating such short-term memory terms and more flexible, data-driven approaches for capturing dynamics that are challenging for traditional rainfall-runoff models under drought conditions.

We will revise the manuscript to explicitly clarify this point, emphasising that the ML models serve as post-processing tools that exploit additional input information (including lagged discharge) to enhance predictions, and that direct performance comparisons with GR4J should be interpreted within this context.

General comment:

Even though the authors present an interesting study, the ML methods used are far from current state-of-the-art. It has been shown in multiple studies that LSTMs perform well as purely ML methods (Kratzert2019b, Kratzert2021 and Feng2020 for CAMELS US, Less2021 for CAMELS GB, Loritz2024 for CAMELS DE) and as postprocessors of process-based models (Frame et al, 2021). The overall poor performance of the hybrid models presented in this study (when they did not receive discharge as input) indicates that the general pipeline could be improved, as the ML postprocessor is not doing its job.

Moreover, it should be noted that other strategies for constructing hybrid models, like using ML methods to parameterize a process-based model (Kraft2022, Feng2022, AcuñaEspinoza2024) or using ML methods to replace process-based model parts (Höge2022, Li2023, Li2024) have shown improved performance with respect to the stand-alone conceptual model, which would be worth considering given that in this case, the hybrid models that did not receive discharge as input are not able to outperform the stand-alone GR4J model.

Therefore, I believe the models presented in the paper are not up to standard with current state-of-the-art, and further improvement is necessary.

We thank the reviewer for the constructive feedback and for highlighting recent advances in hydrological modeling using advanced ML techniques such as LSTMs and novel hybrid modeling strategies.

We acknowledge that state-of-the-art ML methods, including LSTM-based architectures and hybrid approaches that integrate ML within process-based model components, have demonstrated promising improvements in streamflow prediction across multiple recent studies.

However, the primary aim of our study was not to develop the most advanced or optimised ML model for streamflow forecasting, but rather to investigate the potential of specific hydrological predictors, such as long-term runoff coefficient and short-term streamflow memory, in improving daily streamflow predictions during drought conditions. To achieve this, we deliberately chose simpler and widely-used ML algorithms (e.g., MLP, Random Forest, Gradient Boosting) which offer interpretability and robustness, facilitating clearer analysis of predictor importance and behavior.

By focusing on simpler ML algorithms as postprocessors, we were able to isolate and highlight key predictor contributions and understand structural weaknesses in the GR4J model's representation of rainfall-runoff processes, especially under variable climatic conditions.

We will clarify this rationale in the revised manuscript to better position our work within the broader field and to acknowledge opportunities for advancement using more sophisticated ML techniques.

Minor comments:

Line 34: Use proper citation format.

Will be modified in the final version.

Line 43: This sentence does not read well. Please improve the phrasing.

Will be modified in the final version.

Line 47: Should be: hydrologic non-stationarity, use the noun and not the adjective.

Will be modified in the final version.

Line 60: What are you referring to here as validation data? Is it the data that you use to evaluate your model after calibration (but this will also include the forcings)? or the target variable that you are predicting? It just seems that the word validation here is out of context because errors can be found in other types of data too.

Thank you for pointing this out. To clarify, in Line 60, the term “validation data” refers specifically to the observed streamflow data used to evaluate the model’s predictive performance after calibration, rather than the input forcing data. We recognise that the phrasing may cause confusion and agree that “validation data” is not always the best descriptor in this context.

To improve clarity, we will revise the manuscript to explicitly state that the validation refers to the observed target variable (streamflow) against which model predictions are compared, rather than the input forcing data or other datasets. This distinction helps avoid ambiguity regarding the source of errors discussed.

Line 62: Are you using conceptual hydrological models as a synonym of process-based models, as a subcategory or as a different category? The connection with the previous idea could be improved.

Thank you for this insightful question. In our manuscript, we use conceptual hydrological models as a subcategory within the broader class of process-based models. Conceptual

models represent hydrological processes through simplified storage components and flux relationships, rather than detailed physical equations, but still aim to capture the key processes driving streamflow.

We agree that the connection between these terms could be better clarified. To improve readability and conceptual flow, we will revise the manuscript to explicitly state this hierarchical relationship and clarify the terminology when these models are introduced.

Line 95: I do not agree this phrase. There is code development by the user, because you are still using a model. Machine learning methods are models, and they need to be coded. It would be better to indicate that during the training, the model learns to map the input-output relationships using less prior constrains on how this mapping should be done.

Thank you for this insightful comment. We agree that machine learning methods involve code development and that ML models require training to learn input-output relationships. We will revise the sentence to better reflect this by emphasising that during training, the ML model learns to map inputs to outputs.

Lines 98-101: It would be good to cite the studies that use these types of models.

Will be modified in the next version.

Line 124: “or with data containing irrelevant or redundant information.” Do you have a source or examples that justify this? Because in principle, if data is not relevant for a ML model, the model could just ignore it.

Lima, A. R., Cannon, A. J., and Hsieh, W. W.: Forecasting daily streamflow using online sequential extreme learning machines, J Hydrol (Amst), 537, https://doi.org/10.1016/j.jhydrol.2016.03.017, 598 2016.

“ML algorithms may produce less accurate and less understandable results if the data are inadequate or contain irrelevant or redundant information (Hall and Smith, 1996).”

Line 128: Which published studies?

Studies will be added.

Line 130: I disagree that there is a “literature gap on how machine learning can be used to improve hydrological models”. There are a lot of studies published in this area. Of course, there are things that can be improved, but what you mentioned here is too general.

We agree that there is a substantial body of literature exploring the integration of machine learning with hydrological models. To clarify, our intention was to highlight specific gaps related to the identification and use of key predictors, such as long-term runoff coefficients and short-term streamflow memory, in hybrid modeling frameworks, especially under drought conditions. We will revise the manuscript to better reflect it.

Line 205-207: You should only mention the models that you will present the results for. You are saying that multiple algorithms were assessed in the study, you are naming them, and then saying that some of them are not going to be discussed. So why mention them at all? To make the study cleaner, I suggest you should talk only about the results you are presenting. Also, why is Less et al 2021 cited in this part? He used an LSTM model, which you are not using. Moreover, please clarify what the other citations are referring to.

Thank you for your insightful suggestion. We agree that focusing the discussion on the machine learning algorithms for which results are presented will make the manuscript clearer and more concise. Accordingly, we will revise manuscript to mention only the algorithms included in the results and remove references to models not discussed further.

Regarding the citation of Less et al. (2021), we acknowledge that this study uses LSTM models, which are not included in our analysis. We will relocate this citation to the introduction or discussion sections where it better fits the context of state-of-the-art methods. Additionally, we will clarify the purpose of each cited work to ensure their relevance and connection to the content.

Line 214: I would suggest avoiding this kind of phrase. Saying that random forest is one of the most powerful statistical learning methods is subjective. This would depend on the application you have, the metric you are using, and many other factors.

We will revise the phrasing to present Random Forest more objectively, highlighting its widespread use and robustness in various applications without making a generalised claim about its overall power.

Line 227: Same here, avoid saying that gradient boosting is widely recognised as one of the most powerful algorithms. This is again subjective, case-dependent and not related to the main point you are trying to make.

We will revise the phrasing to present gradient boosting more objectively, highlighting its widespread use and robustness in various applications without making a generalised claim about its overall power.

Line 238: This is not true. MLP is not the most popular type of neural network in hydrology. The current state-of-the-art has been achieved with LSTMs (Kratzert2019b, Kratzert2021 and Feng2020 for CAMELS US, Less2021 for CAMELS GB, Loritz2024 for CAMELS DE). Transformers have also shown good results in CAMELS US. Both of these methods considerably outperform MLP.

It will be corrected to: MLP is one of the most popular type of neural network …

Line 252: What do you mean by calibration and optimisation were conducted for the training and test period only? You should not calibrate for the test period. The test period is used to evaluate the model that was calibrated during the training period. I think there is a misunderstanding on the names you are using.

Will be rephrased in the next version.

Line 266: Improve phrasing of “an intentional effort”.

Will be rephrased in the next version.

Line 272: Are you referring to mean-squared error or sum-squared error?

We are referring to the mean squared error (MSE) as the loss function used for training the models. We will clarify this in the manuscript to avoid any confusion. However, in the SKlearn description it is written squared error.

Line 270-277: What you are referring to here as a testing period is what it is normally referred to as validation.

Will be modified in the next version.