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 2:

General comments:

The introduction section is far too long and sometimes addressing points that are not mentioned in the paper any longer. This is the case of the non-stationarity issue for physically-based hydrological models. The authors reserved quite some space in the introduction for this topic, which is then no longer touched upon in the analysis nor in the discussion. I suggest to shorten it or change something in the analysis (see my general comment #6). In general, the introduction could be shortened and focused towards the two research questions.

We thank the reviewer for the helpful suggestion. We will shorten the introduction to improve its focus and ensure alignment with the two main research questions. In particular, we will reduce the emphasis on non-stationarity in physically-based models, as this topic is not explored further in the analysis or discussion. This revision will help streamline the manuscript and improve clarity.

The introduction is focusing very much on drought conditions, while this is not reflected in the two research questions mentioned. As also part of the results is focused in analysing the performances of all models for low flow conditions, I suggest to either specify the current research questions for the low flow conditions or add a third question about it.

We thank the reviewer for the valuable observation. We will revise the introduction to ensure consistency with the stated research questions. Specifically, we will either refine the current research questions to better reflect the focus on low flow conditions or include a third research question explicitly addressing model performance during drought or low flow periods. This will help maintain coherence between the introduction, research objectives, and analysis presented.

Results and discussion section: I believe the results section could be divided in sub-paragraph addressing different research questions, making the readability easier. As they are now, it is difficult to focus on the two research questions.

We appreciate the reviewer’s suggestion regarding the structure of the Results and Discussion section. We will revise this section by introducing sub-paragraphs that are clearly aligned with each research question. This restructuring will improve the readability and allow the reader to follow the flow of the analysis more easily in relation to the stated objectives of the study.

Results and discussion section: there is no real discussion of the results in the context of literature, neither for the GR4J nor for the Machine Learning (ML) models. In literature there is plenty of evidence of the role of past streamflow in improving streamflow predictions (because of the high autocorrelation), but this is never mentioned in the paper. The relevance of past streamflow should have not come as a surprise. Also, there is no mention of the limitations of this study.

We thank the reviewer for highlighting this important point. We will revise the Results and Discussion section to include a more thorough discussion of our findings in the context of the existing literature, particularly regarding the predictive value of past streamflow and its well-documented role in improving model performance due to streamflow autocorrelation. Additionally, we will incorporate a clearer reflection on the limitations of our study to provide a balanced interpretation of the results.

The comparison of GR4J with ML models trained with streamflow is not entirely fair, as GR4J is not using the same input variables. While I understand that part of the purpose of this paper was to indeed find out which other predictors not used by GR4J should instead be considered, I believe that the authors should add a sentence acknowledging the difference in the models, also mentioning the well-known importance of past streamflow for ML models.

We appreciate the reviewer’s observation. In the revised manuscript, we will explicitly acknowledge the differences in input variables between GR4J and the machine learning models, particularly noting that GR4J does not use past streamflow as input. We will also add a sentence recognising the well-established importance of past streamflow in enhancing predictive performance in ML models, as reported in other studies.

I do not agree in the way performance are presented, i.e. showing the performance of ML and GR4J models in the pre-drought, drought and post-drought conditions all together. I believe that it would be more interesting to first show the comparison of performances in the testing set (during the pre-drought period), which are supposed to be the hydrological “regular” conditions. This would allow identifying (potential) deficiencies that are specific of the GR4J model (for instance the lack of long-term memory) and relevant input variables for streamflow prediction. Then, repeating the same comparison in the period of millennium drought and post-drought, would show if the deficiencies and relevant input features are the same, or if the millennium drought indeed brought a change in the hydrological characteristics of the catchments analysed.

In case ML models show better performances in both periods, then it would validate the thesis that physically-based models are not adequate for hydrological studies in non-stationarity conditions, while ML are, also justifying the introduction on non-stationarity of climate.

In the revised manuscript, we will reorganise the presentation of the performance results by first comparing the models' performances during the pre-drought period, which represents more typical hydrological conditions. This approach will help clearly identify specific limitations of the GR4J model, such as its capacity to capture long-term memory, and highlight the relevant input variables for streamflow prediction under regular conditions.

We will present the performance comparisons for the drought and post-drought periods. This will allow us to examine whether the deficiencies and key predictors remain consistent or if changes in hydrological characteristics during the Millennium Drought influence model behavior. Such a structure will also better support the discussion regarding the adequacy of physically-based versus ML models under non-stationary climatic conditions, as initially motivated in the introduction.

The ML models used exclude the Long Short-Term Memory (LSTM) model, which is nowadays considered the state of the art in terms of ML algorithms for hydrology. While I understand that adding yet another model in the analysis would now require quite some time, I believe that the authors should acknowledge the use of these models in the literature introduction, explain why LSTM was not used, and potentially add this in the limitations of the work.

We chose to focus on simpler ML algorithms in this study because our primary aim was to identify key predictors that influence streamflow rather than to optimise predictive accuracy with complex models. We will also clearly state this rationale in the limitations section and acknowledge that incorporating LSTM models represents an important path for future research to further enhance hybrid modeling approaches.

Specific comments:

Line 25: climate change is mentioned in the key words, but there is nothing about it in the paper, only the mention in the introduction. I suggest the authors to change the keyword or update the manuscript with additional considerations about climate change.

We will update the manuscript to better reflect the role of climate variability and change, particularly in relation to the Millennium Drought and its impacts on streamflow dynamics. Alternatively, if this is not feasible within the current scope, we will revise the keywords to better align with the core content of the paper.

Line 32: space missing between “noteworthy example” and “(can Dijk et al., 2013)”.

Will be corrected

Lines 34-36: I believe the referencing style is incorrect, the references should all be within the same brackets.

Will be corrected

Line 47: there is a typo. It should be “non-stationarity” rather than “non-stationary”.

Will be corrected

Lines 95-96: I do not agree with this statement, as you need to code yourself also for ML models development. I suggest to revise.

Will be corrected

Line 115: there is a mention to Section 2.2.2.8, but such section does not exist.

Will be corrected

Lines 130-131: I do not agree about this literature gap, as there are several attempts to use ML in a hybrid configuration to improve streamflow predictions.

We acknowledge that there are several studies exploring the use of machine learning in hybrid configurations for streamflow prediction. We will revise the manuscript to better reflect this body of literature and clarify the specific contributions and focus of our study within this context.

Section 2.1: as part of the focus of the paper is related to finding relevant input features, I would already specify here which variables are used in the ML models, or at least add a sentence guiding the reader to read section # xx to have more information about it.

Will be added

Section 2.2: it would have been easier to go through the methodology if a general workflow or information about the overall methodology is given before going into the details of each modelling part.

We will add an overview or general workflow diagram at the beginning of Section 2.2 to provide readers with a clearer understanding of the overall methodology before presenting the detailed descriptions of each modelling component.

Section 2.2 and results: how is the rainfall change and streamflow change linked to the remaining of the investigations?

We will revise the manuscript to better clarify the role of observed rainfall and streamflow changes in structuring the analysis. Specifically, we can add an overview in the methodology section explaining that these changes are used to define distinct hydrological periods, pre-drought, drought, and post-drought, which frame the evaluation of model performances. This will make clear how climatic variability informs the design and interpretation of the modeling experiments.

Line 194: what are the four parameters of the GR4J model?
Will be added

Figure 2: there is no legend about the symbols used. What is En, Es, Pn, Ps, and so on?

We will add a clear legend

Lines 205-208: the fact that random forest, Gradient Boosting, MLP have the best performance is coming from the results of this work or from the papers referenced? And also, if the other models are discarded, why presenting them as part of the methodology in the first place?

The better performance of Random Forest, Gradient Boosting, and MLP is based on the results obtained in our study, where these algorithms consistently outperformed others across the evaluated catchments and conditions. However, we initially included other models such as Linear Regression, Ridge Regression, SVR, and Decision Tree in the methodology to provide a comprehensive comparison and to assess a range of machine learning approaches of varying complexity. Unnecessary algorithms will be removed.

Line 223: it is mentioned that RF can be used to check importance of features, but why is this characteristic not leveraged, since part of the part focuses on finding the most relevant predictors to improve streamflow predictions? If not used, I believe it should be justified in the paper, especially after adding this line.

You are correct that Random Forest's ability to assess feature importance is valuable, especially given our focus on identifying key predictors for improving streamflow predictions. In our study, we did explore feature importance from the Random Forest models; however, due to space constraints and the complexity of presenting these results across multiple catchments and models, we did not include a detailed analysis in the manuscript.

In the revised version, we will explicitly mention that feature importance was examined using Random Forest, and we will provide a brief justification for the level of detail presented. Additionally, we will consider adding a summary or example of the most influential predictors identified, to strengthen the discussion on relevant features and their role in model performance.

Section 2.2.3: what is the target of the model? It is not really specified.

Will be added

Line 256: how many steps back of rainfall and potential evapotranspiration are considered? What is the lag between the target and these predictors?

It is mentioned in lines 305 to 308

Lines 278-289: the explanation of the variables used in the predictive mode and the training mode is not clear. If real observations are used to compute the runoff coefficient and short/long-term memory in the predictive mode only, which variables are used in the training mode?

Will be clarified in the next version

Line 332: it is mentioned that negative values are omitted, while previously (line 328, figure 5 caption), it is mentioned that negative values are shown as 0. I believe the same approach should be followed everywhere, to avoid inconsistent evaluation across the paper.

Will be modified

Lines 361-362: why NDVI, LAI, groundwater exchange and not other variables?

Thank you for this question. NDVI and LAI were selected as vegetation-related indices because they are widely used and well-established indicators of vegetation health and density, which influence evapotranspiration and soil moisture. Groundwater exchange was included as it plays a critical role in sustaining streamflow, especially during low flow periods.

While other variables could also be relevant, these were chosen based on their availability, relevance to hydrological processes, and support from previous literature. We acknowledge that incorporating additional variables could provide further insights, and this will be considered in future work.

Lines 380-381: GR4J overestimates low flow, but is this in any period (pre and post-drought), or only after (during) millennium drought?

Will be investigated.

390-391: it is mentioned that it is not evident which predictor has the greatest overall impact. As this is part of the research questions of the paper, this result should be addressed again when discussing limitations of this work.

Thank you for the comment. Indeed, assessing the relative importance of predictors could be further explored by leveraging Random Forest feature importance measures.

Lines 403-406: ranges of runoff conditions are given. However, it would be interesting for the reader to know to which hydrological conditions or regime these coefficients refer to? For instance, catchments with flashy response, long recession limbs, high/low interannual variability.

We agree that providing information on the hydrological regimes associated with the runoff coefficients would add valuable context. However, due to space limitations, we will include a concise summary to briefly characterise the catchments' hydrological behaviors related to these coefficients without substantially increasing the manuscript length.

Lines 464-466: these lines are not clear. How is it possible that the results of the post-drought period show that there is memory effects in the pre-drought conditions?

Thank you for pointing this out. We acknowledge that the wording in lines 464-466 could be misleading. What we intended to convey is that the memory effects identified during the post-drought period reflect underlying catchment characteristics that were also present in the pre-drought period but became more apparent or quantifiable after the drought. We will revise the text to clarify this point and avoid confusion.