Summary

This paper explores the application of Machine Learning (ML) algorithms to post-process physically-based hydrological models and find key predictors to improve daily streamflow predictions. The methodology is tested over 39 sub-catchments in Western Australia (Victoria), using the streamflow predictions of the GR4J physically-based model and a set of climatic and hydrological additional variables (rainfall, potential evapotranspiration, streamflow-derived variables).

Review

While the topic of the paper is definitely of interest for the scientific community and fits the scope of HESS journal, I have several comments that the authors should address to improve the overall quality and scientific relevance of this manuscript. I have both general and specific comments.

General comments:

1) 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.
2) 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.
3) 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.
4) 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.
5) 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.
6) 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.

7) 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.

Specific comments:

1) 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.
2) Line 32: space missing between “noteworthy example” and “(can Dijk et al., 2013)”.
3) Lines 34-36: I believe the referencing style is incorrect, the references should all be within the same brackets.
4) Line 47: there is a typo. It should be “non-stationarity” rather than “non-stationary”.
5) Line 95-96: I do not agree with this statement, as you need to code yourself also for ML models development. I suggest to revise.
6) Line 115: there is a mention to Section 2.2.2.8, but such section does not exist.
7) 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.
8) 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.
9) 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.
10) Section 2.2 and results: how is the rainfall change and streamflow change linked to the remaining of the investigations?
11) Line 194: what are the four parameters of the GR4J model?
12) Figure 2: there is no legend about the symbols used. What is En, Es, Pn, Ps, and so on?
13) 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?
14) 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.
15) Section 2.2.3: what is the target of the model? It is not really specified.
16) Line 256: how many steps back of rainfall and potential evapotranspiration are considered? What is the lag between the target and these predictors?
17) 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?

18) 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.
19) Lines 361-362: why NDVI, LAI, groundwater exchange and not other variables?
20) Lines 380-381: GR4J overestimates low flow, but is this in any period (pre and post-drought), or only after (during) millennium drought?
21) 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.
22) 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.
23) 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?