Dear Authors and Editors,

I believe the manuscript “Exploring uncertainties across the modelling chain in machine-learning-based streamflow forecasting” is of relevance to Machine Learning and forecasting community and fits the aim and scope of Hydrology and Earth Systems Sciences (HESS) journal. However, I believe the manuscript needs major reviews before being published. My comments are below.

Major comments:

- The manuscript employs a single catchment to attribute uncertainty to the different phases of the Machine Learning (ML) modelling chain. How can the results be generalised and the same workflow be applied to large sample hydrology modelling (e.g. all the regional LSTM models developed with the different versions of CAMELS dataset)?
- The possible combinations in the models architectures could be endless, considering that infinite layers could be added. Are there external constraints applied? And what about the uncertainty given by the loss function (I am not suggesting to add it to the pipelines, but to at least mention about it)?
- Numerical Weather Prediction (NWP) model runs were used as input for the different ML pipelines but important information is missing: how often are the models initialised and, hence, how many samples are available for training? Is the full NWP ensemble used for training the models or is only the control member used (or ensemble mean) used?
- In the building of the input features sequences, the NWP data used from t-L up to t, are they hindcast or reanalysis? If they are hindcast, which is their lead time?
- Still in the building of the input features sequences: I believe using only 3 days of past historical information to forecast the next 5 days might be very little, even if the concentration time is around ~1 day; there is a well known literature of LSTM models using one year (365 days) of past observations to predict the next day streamflow, how do you justify using only 3 days to forecast next 5 days streamflow? Especially in snow-driven catchments, where the snowmelt process is very slow, longer information is needed.
- Tables are used to summarise the choice of the selection of a specific model or metric, but they could be omitted in my view, or at least modified to retain the reason for selection within the main text. In table 6, the column of the Key reference is mixed with references and reasons for selection of the metric.
- About the metrics used: the probabilistic metrics are they applied to the meteorological ensemble of NWP or to the ensemble created by all the pipelines j tested? If it is applied to the ensemble of the pipelines, what is the added value of using these probabilistic metrics?
- About metrics again: I feel too many metrics are used but not all of them are discussed. Maybe some can be dropped.
- Most of the manuscript, including the title, is focused on the uncertainty quantification and very little is said in the introduction about the computation cost-forecast accuracy analysis. I think more should be said about this analysis, or it can be dropped and keep the manuscript within the uncertainty quantification scope. Additionally, the effects of the tuner into the uncertainty quantification are not discussed, while they are presented as part of the UQ pipeline.
- In the computational costs-accuracy analysis, in L226 it is mentioned that the aim is to minimize both costs and RMSE. It should be clarified, however, that there is no multi-objective optimization leading to the construction of the Pareto front, as the front is artificially built later by using the results of all independently trained pipelines.

Minor comments:

- L80: what are “idealized observed inputs”?
- L82 claims that statistically robust pipelines are identified. Where is this shown?
- L131: gaps are filled with interpolation. I think this part should be clarified in terms of how many gaps were present and mention that this could also contribute to the uncertainty of the modelling chain.
- Charts in Fig 2c in the last row do not correspond to those of Fig 2b, which should be referred to the same lead time. Can you clarify why?
- L273 mentions a “skill”, but there is no benchmark model here?
- L303 mentions that snow water equivalent is not relevant as input, despite the catchment being snow-driven. This is because only little days are used in the input sequence, compared to the speed of the snowmelt process.
- In the conclusion (from L384) it is mentioned that richer feature exploration could be done. However, the conclusions do not mention that this would help only for the shorter lead times, while for the longer it would make more sense to invest in the NWP forecasts quality (as rightfully stated in the abstract).