Reviewer #2

	General comment	Response
	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.	We thank the Reviewer for a detailed review, valuable suggestions and comments. It helped us improve the quality of the manuscript. We have tried our best to implement the changes in regard to your comments. Responses to reviewer's comments are in blue. *In-text changes in the manuscript are in red (italic).*

	Major comments	Response
1	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)?	We thank the review for the comment. Please find a more detailed response to this comment in Reviewer #1 responses Comment 1. The additional proposal of how the workflow would be generalized on the large sample hydrology modelling is added in the Conclusion (L451–457): *For a large-sample, multi-basin extensions of this framework, a more economical design is suggested, given the size of the dataset. The tuning factor and repeated hyperparameter optimization of each configuration (replicates) can be collapsed, with a small number of random-seed replicates to capture training stochasticity only. In exchange, several optimum hyperparameter configurations can be trained to completion, replacing the stochasticity of tuning and representing model equifinality. The remaining factors can be explored through a sampled (fractional) design rather than exhaustive full-sweep combinations, chosen so that each factor level is well replicated across basins without training every configuration to completion. This decreases the computational efforts and keeps the framework usable on large datasets.*
2	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	The part where exact framework was explained was firstly omitted from the main text. It still is present in the Figure 1

	uncertainty given by the loss function (I am not suggesting to add it to the pipelines, but to at least mention about it)?	caption. We have now placed it in the main text (L93–97): *A systematically varied ensemble of forecasting pipelines is generated by combining alternative forcing products (4 NWP model runs), feature combinations (4 different feature combinations), neural architectures (4 different ML model architectures), tuning algorithms (4 HPO algorithms) and by accounting for variability introduced by tuning and training stochasticity (4 replicates of the same configuration). This sums up to 1024 different pipeline realizations (Fig. 1b).* The loss function uncertainty is also added to the text as limitation in the conclusions: *Several sources of uncertainty were held fixed rather than factorized, including the meteorological grid and resolution, re-gridding and interpolation, and the use of a single deterministic NWP run, as well as the choice of loss function, for example. Each could be incorporated as an additional factor in future applications of the framework.* Additionally, we have now included the hyperparameter search space in Supplementary Materials to show constraints used for HPO algorithms.
3	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?	We have included the additional information in the manuscript (L130–131; L133–134): *NWP models are initiated at different frequency, which is why only the first run of the day was taken into consideration (at 00:00 for all models).* *Only the control member of the ensemble is used for the forecasting task.*
4	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?	The NWP runs used are hindcast, with lead time corresponding to the lead time of the streamflow prediction. We have expanded on this to explain it more specifically (L123–124):

		Meteorological predictors were acquired from archived Numerical Weather Prediction (NWP) model runs, enabling retrospective forecasts (hindcast) up to five days ahead, matching the target forecast horizon. More specifically, the hindcast set with a lead time of h is used in the the h-th spot in the windowed input.
	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.	This issue is not directly addressed in the manuscript, but we agree that it perhaps should be. Namely, the literature that uses large context windows are non-endogenous forecasting models. More specifically, the models do not use previous streamflow as a feature. This is, of course, done intentionally, for the purpose of trying to generalize across domains, and use the forecasting models where there are no records of flow data. However, this is not the point of the models we have developed. Since we need good precision and accuracy for a specific profile for operational usage (e.g. hydropower plant operations), we include the previous flow data, which indicated the previous state of the system, dropping the need for large context window. Regarding the snowmelt process, yes, they are very slow. However, once a certain volume of water melts (calculated via difference in SWE), it should travel to the profile relatively closely to the time of concentration. We have provided additional explanations regarding this comment (L176–178): *As previous discharge is used as input in the prediction task, which stores the system state information implicitly, dropping the need for large context windows (e.g. 365 days for forcing-only-based prediction in Kratzert et al., 2018).*
5	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.	We have modified the Table 5 right column to contain the reason for selection and key reference.
6	About the metrics used: the probabilistic metrics are they applied to the meteorological ensemble	The probabilistic metrics are applied to the pipeline ensembles created using the

	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?	framework. The added value is that probabilistic metrics evaluate ensemble predictions as probabilistic forecast, while deterministic only measure the mean forecast accuracy/error. We hope this provides a satisfactory response to the Reviewer's comment.
7	About metrics again: I feel too many metrics are used but not all of them are discussed. Maybe some can be dropped.	MAPE and r can be dropped (r is left in the methods table, as it is used in KGE).
8	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.	We agree with the Reviewer. The Introduction section, indeed, did not appropriately cover the cost-effectiveness angle. The Introduction is now widened to include this (L46-49): *Hyperparameter-optimization (HPO) strategy adds a further layer, as different search algorithms and budgets can converge to different configurations of the same architecture, so the tuning stage is itself a source of variability in modelling outcomes. HPO is computationally expensive and allocating a finite computing budget to find an optimal hyperparameter configuration remains a significant problem to tackle (L. Li et al., 2018)* The effects of the tuner are discussed in Section 3.3. We have now clarified it wasn't discussed in Section 3.2 as to not overburden the text: *HPO algorithms performed similarly and are for that reason omitted from the Figure 3.*
9	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.	Thank you for this comment. We agree and have now clarified this in the manuscript (L264–265). *The Pareto front was not constructed using a multi-objective optimization but was rather built using the results of all independently trained pipelines.*

	Minor comments	Response
L80	what are “idealized observed inputs”?	We thank both Reviewers for pointing out that this phrasing was unclear. We have rephrased it to make this explicit (L80-81): *(1) development of a workflow-level multi-model UQ framework for operational streamflow forecasting trained on past NWP runs rather than the observed forcings, so the model learns under the same forecast errors it faces operationally*
L82	claims that statistically robust pipelines are identified. Where is this shown?	We thank the Reviewer for pointing this out. The statement in L82 was imprecise in regard to the current text in Section 3.2. There were no actual pipelines highlighted. The robust configurations are now identified in Section 3.2 (where they intended to be). In the revised version we make the identification explicit: *Taken together, these results identify the configurations that perform most reliable across horizons: IFS- and UKMO-driven pipelines, discharge- and precipitation-centred feature sets (Qp–Qpts), with TCNs generally performing best, as consistently shown in recent literature*
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.	We have clarified this in the revised passage, where we now mention the gap filling as a possible source of uncertainty. Missing values were rare and short (several instances with few consecutive days each), all outside rainy and extreme-flow periods. We agree that gap filling (and re-gridding, as raised by the Reviewer #1) could contribute to the uncertainty of the modeling chain, and the revised text now states that this step could be treated as a separate factor in future applications of the framework. *Isolated missing values in the time series were infrequent and short (at most several consecutive days) and occurred outside rainy and extreme-flow periods; these gaps were filled using simple interpolation. Two elements of the data cleaning procedure (re-gridding and gap filling) could in principle contribute to the overall prediction uncertainty and could be incorporated as an additional factor within*

		the framework. This was, however, not implemented, as their expected contribution is small: only a single NWP product required re-gridding, and gap filling affected only a handful of values across the full dataset, none during high-flow conditions.
Fig.2	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?	It seems as the values are not the same, because of the different scale. This is, however, not the case. Additionally, these two subplots show different things, so it is deemed to look a bit different. Fig.2b shows successively 1-day-ahead predictions, while Fig.2c shows a fan chart (so only 1 value is the same, the rest of them are not).
L273	L273 mentions a "skill", but there is no benchmark model here?	This is true, no benchmark model in this study. However, in this particular phrasing, the comparison is made with the results of published work. In addition, we have now added the naïve persistence prediction as a baseline in Section 3.2 with which we compare the model predictions with: *UKMO and IFS show slightly stronger performance than GFS and GEM at short lead time, more consistently outperforming the persistence baseline, while at longer horizons GFS-driven pipelines show a larger decline and wider dispersion compared to the rest, indicating reduced robustness.* *Architectural differences are clearest at short lead time. LSTM and TCN show higher skill (with the average TCN model prediction crossing the persistence baseline threshold in almost all cases) and narrower performance distributions than MLP, consistent with improved representation of temporal dependence and hydrologic memory.*
L303	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.	We agree this could very well be the case. We have now addressed this in the discussion (L339–341): *This could also be the consequence of a short input window, where model did not capture snowmelt and, more importantly, flow generated by it.*

L384	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).	We agree that the improving the NWP forecasts yields the largest gains at larger lead-times, as stated in the discussion (now underlined in conclusion). However, the feature exploration actually stayed consistently important across all lead times. This is exactly the reason why we have suggested this kind of future work. We have now adjusted this part a bit to be more conclusive. *Looking ahead, this UQ framework naturally supports both richer NWPs (especially important for longer lead-times) and feature exploration, moving beyond incremental "add-one-variable" designs to evaluate targeted feature subsets that reflect hydrologic process relevance and reduce redundancy.*