Reviewer #1

	General comment	Response
	The paper addresses a current and very active area of research regarding data-driven models to forecast discharge. It is well written and the results are supported by the evidence shown. However, I suggest major revisions before publication.	We thank the Reviewer for detailed review, suggestions and insightful 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	I cannot help but feel that the experimental setting for this experiment is quite limited. Just over a year for a testing period for a single basin really doesn't tell me much about how the results of this study are generalizable or if the overall framework is useful to adopt. For this particular basin, adopting a data-splitting strategy of time-series cross-validation would be a step toward obtaining results that are more representative of the study area in general, rather than simply the periods chosen. Furthermore, it would be highly beneficial for the authors to comment on if and how this experimental framework could be adapted to a setting of large sample hydrology.	We thank the review for the comment. We are aware of the limitations in the experimental setup. The testing period is indeed short. This is a direct consequence of the limited record over which all NWP products share synchronous coverage (March 2019–December 2024). Within this constraint, we chose the split so that the testing period spans at least one full seasonal cycle and includes a significant flood peak (November 2023), exercising the conditions most relevant to the framework. This is now more clearly addressed in the text (L149–150): *The testing period of just over a year is relatively short, which is a constraint imposed by the limited record length. However, it is carefully chosen to span across at least one full seasonal cycle and include a significant peak flow from November 2023.* Time-series cross-validation, such as moving or expanding window, is preferable to relying on a single train–test split. However, this kind of training can be significantly more computationally expensive, especially across the whole factorial of 1024 pipelines. A single pass already requires ~2 months, and expanding-window evaluation would increase this roughly threefold. We therefore retained the chronological data-splitting strategy, which is the standard hydrological validation framework. We have clarified this choice now in the manuscript (L144–147): *A chronological data-splitting strategy (train-validation-test split) is used. Applying the expanding-window validation instead would likely*

		provide a more robust, less split-dependent performance estimate, but applying it across all 1024 pipelines would increase the computational cost significantly. For this reason, the standard testing framework in hydrology is adopted. As this comment primarly has the small-sample robustness concern, we have improved the issue through a bootstrap variance decomposition on resampled data (added in Methods and Supplementary Materials). A resampling approach well suited to limited records. This quantifies the sampling variability of the attributed uncertainty components. Regarding the comment about implementation of the framework on large sample hydrology, which was raised with both Reviewers: this was briefly mentioned in the original submission, but we have expanded on that more clearly in conclusions (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	I have the feeling that there is a lot of information left out of the paper regarding the model setups. Particularly, the authors refer to 1024 pipeline realizations, yet within each pipeline, there is independent testing between different weather products, combinations of features, and hyperparameters of models. For the last one especially, combinations of hyperparameters can be infinite; therefore,	The part where exact framework was explained was firstly omitted from the main text. It still is present in the Figure 1 caption. The 1024 pipeline realizations represent all of the combinations of NWPs, features, models and HPO algorithms (with 4 replicates for each pipeline). We have now explained it in more detail and placed it in the main text (L93-97):

	what were the constraints used for the HPO algorithms? I can see them in the code, but they are not mentioned in the text. Mentioning the number of training replicates obtained would also be useful. What is the total number of model configurations tested?	*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).* We did not explore different combinations of hyperparameters but rather used the existing tools for HPO, tested them and incorporated in this framework. We have now, however, included the hyperparameter search space to show constraints used for HPO algorithms in Supplementary.
3	This study doesn't have a benchmark. The authors only report the performance of their ensembles, and a comparison against simply predicting the last measured discharge value for the deterministic performance metrics would be very useful. For the probabilistic models, perhaps a single model configuration performing quantile regression across all quantiles could be used as a benchmark too.	The study doesn't have a standard benchmark, as the main focus of the paper is uncertainty decomposition. Also, we did briefly mention and compare to the results of the previous published work that was on the same case study. However, we agree that incorporating a naïve persistence forecasting provides some valuable additional information when defining the robust pipeline configurations. We added the persistence baseline in Fig. 3 and mention the performance compared to the baseline on several occasions in the text (L330–L331; L345–346): *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*
4	Particularly, the fan for Peak 1 in Figure 2c is worrying because the model wouldn't be able to give the correct signal in a clear decision-making scenario of issuing a	We agree with the Reviewer that Peak 1 is not well captured by the ensemble mean at short lead time. We would note, however, that point accuracy is not the most decision-relevant

	flood warning. Even at t-1, simply forecasting the last measured discharge value would be a better predictor than the ensemble.	property here, but rather the prediction bounds. The ensemble bounds widens its interval sharply during Peak 1 (containing the observed peak), appropriately signalling low forecast confidence at precisely the point where caution is warranted. We have done additional analysis to further understand the cause of this issue. The single-prediction variance decomposition attributes a large share to the NWP forcing (>30% at one day ahead, above the typical level), indicating genuine disagreement among weather inputs for this event, that yielded different outcomes in predictions. The models also contribute significant uncertainty, as expected for a one day ahead prediction. This is however, not within the main goal of the paper, as we do not claim these models are ideal. We are, rather, proposing an uncertainty quantification framework for hydrologic prediction. Moreover, using this framework, we can see exactly what failed in this instance, which tells us what improvements are required for this kind of event to be captured more adequately. Results of the additional analysis for this specific prediction was added to the main text (L285–288): *Even though the mean one-day-ahead prediction performed worse than simple naïve persistence forecast, a single-prediction variance decomposition showed a large portion of the variance (more than 30%) attributed to NWPs, indicating that there were significant differences in weather prediction that yielded different outcomes in prediction.*
5	Following the previous point, the manuscript introduces this framework as a tool to support operational management and decision-making, however, the evaluation relies entirely on standard goodness-of-fit metrics. If the purpose of the framework is to inform operations, there is a disconnect between the stated goals and the evaluation. It would be highly beneficial for the authors to clarify the exact scope of the framework and consider evaluating it, or discussing its	The framework is not proposed as a standalone decision tool but as a means of producing decision-relevant information (uncertainty attribution across the workflow and predictive intervals). How that information is turned into a specific operating action (e.g., a warning threshold) is application-dependent and outside this study's scope, which is why the evaluation targets forecast reliability rather than a particular decision rule. We have clarified the scope of the framework more clearly:

	limitations, using decision-centric metrics (e.g. derived from a confusion matrix like false alarm rate).	*The framework is intended to attribute predictive uncertainty and characterize forecast reliability across the modelling chain, providing decision-relevant information rather than a specific operating rule.* We agree decision-centric metrics are valuable, and note the limiting factor is the number of events, not the framework. The test period contains only several peaks, which supports an event-capture (coverage) check but not rate-based metrics such as false-alarm rate, whose estimates would be unstable with so few events. We therefore now report whether each observed peak falls within the ensemble bounds across all test-set peaks in Supplementary Materials (Fig. S2–2, Table S2): *This is confirmed by the Figure S2–2, showing fan charts for all of the peaks in the test set. To quantify if the ensemble spread captured the peak, a hit or miss is given to the ensemble forecast for each of the peaks, shown in Table S2. This is strongly lead-time dependent: at one day ahead the 95% interval brackets all eight peaks (8/8) and the 50% interval four (4/8), whereas at three-days-ahead prediction falls to 3/8 and 1/8, consistent with the growing underestimation of peak magnitude with lead time shown in Table 6.*
6	Section 3.3: For an uncertainty quantification paper, leaving 40% of the variance unexplained seems quite high. It would be useful if the authors commented on the unexplained sources of uncertainty or how they could be mapped within their framework.	Thank you very much for this comment. It initiated a deeper analysis, with a new insight. We found that a large portion of the unexplained variance, corresponding to the residual of variance decomposition analysis, is contributed by the stochasticity of the tuning and training of the same pipeline configuration. This adds another angle to the conclusions of the paper. We consider these uncertainties mostly aleatoric, meaning that they cannot be reduced. This part is expanded and further explained in Section 3.3 (L382-397). *The residual in the main-effects variance decomposition remains substantial (~40%). Figure S3-2 shows higher order interactions among workflow factors account for roughly 10% to 20% of total variance, while the remainder is the within-configuration term (replicates), contributing to 20% to 30% of total variance. This significant spread is produced by stochastic tuning and training under an otherwise identical*

		workflow configuration. Under a fixed modelling protocol this component is effectively aleatoric, originating through randomness and cannot be reasoned away by additional knowledge, only characterized (Kiureghian & Ditlevsen, 2009). This is, in principle, an equifinality problem, that remains no matter which model or optimization technique are used. It is not, however, strictly irreducible, since the stochasticity manifest as different fitted (hyper)parameters, overlapping to a degree with parameter and optimization uncertainty. This highlights that aleatoric/epistemic boundary as a matter of framing rather than a fixed property of the source (Hüllermeier & Waegeman, 2021). We have also updated the abstract and conclusions to reflect these findings: *Aleatoric uncertainty, driven with the randomness of tuning and training processes, corresponds to a significant piece of the uncertainty, meaning that even a fully optimized, well-chosen pipeline retains an irreducible run-to-run spread that must be represented in the predictive distribution rather than reported as a single deterministic value.*

	Minor comments	**Response**
L27	The same could be said of underconfident predictions.	We have revised this to acknowledge that underconfident predictions can also carry consequences, while retaining the paper's focus on overconfidence, since ML-based predictions are mainly overconfident. *Forecasts support early flood warning, reservoir operation, and other risk-sensitive decisions, where the consequences of overconfident predictions (or, less commonly, underconfident) can be substantial.*
L80	At this point, I'm not sure what the authors refer to as 'idealized observed inputs'; a forecasting hydrological model necessarily has to be driven by NWP.	We thank both Reviewers for pointing out that this phrasing was unclear. We agree that an operational forecasting model must be driven by NWP. However, the intended contrast was with the common practice of training and evaluating ML streamflow models on observed (or reanalysis) forcing, which is unavailable in operational

		settings. We have reworded the passage to make this explicit: *(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*
L124, L131	Not that I'm asking you to do this, but could the proposed framework be used to assess the uncertainties introduced by re-gridding and interpolating?	We agree that re-gridding can contribute to modeling-chain uncertainty, and note that the framework is in principle capable of treating it as a separate factor. Using different re-gridding and interpolation techniques, the framework supports variance decomposition across that factor as well. We did not do so here because only a single NWP product required re-gridding to the common grid, with only a few missing values, so its expected contribution is small. The revised text now states this and flags re-gridding as a factor that could be incorporated in future applications. *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.* We have also added the several other possible uncertainties we have not incorporated in the study (L446–448): *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.*

L139	Using only sine encodes different days with the same value, e.g., days 45 and 137 in a 365-day year. Normally, this is addressed using both sine and cosine encoding.	The Reviewer is correct that a paired sine/cosine encoding is standard, and we thank them for the careful reading. We note two points. First, temperature, which has a strong seasonal cycle and is included among the input features, provides a complementary seasonal signal that partially offsets the sine ambiguity. Second, and more importantly, day-of-year contributed little incremental skill in our experiments (Fig. 3), consistent with past discharge already capturing most of the relevant seasonal information. A feature adding little signal under an imperfect encoding would not be expected to change the results materially under a corrected one. We have clarified the encoding and noted paired sine/cosine as a straightforward refinement, though still unlikely to affect the present results given the limited contribution of day-of-year at these horizons. *Generally, a paired sine/cosine encoding is used, as a single sinusoid assigns the same value to two days per year. Here, only the sine function is used, alongside temperature, which has a strong seasonal component of its own and provides a complementary signal that partially distinguishes the two dates mapped to the same sine value. In any case, day-of-year contributed little incremental skill at the horizons studied (Section 3.2), which we attribute mainly to past discharge already carrying most of the seasonal signal. Its encoding is therefore unlikely to have significantly affected the results.*
183	100 trials sounds low. For example, in https://doi.org/10.1016/j.jhydrol.2023.129414, even if the reference is for conceptual models, trials for SCE are set to 1000.	We appreciate this point. We note, however, that the cited 1000-trial budget refers to a global optimiser applied to conceptual-model calibration over a continuous parameter space. However, for HPO, optimization usually hits a plateau within the 100 trials (e.g. random search Bergstra & Bengio, 2012). In addition, increasing the budget tenfold would raise tuning cost roughly proportionally across all 1024 pipelines, which is computationally impractical. The variance decomposition also shows that tuning strategy contributes only a small fraction of

		total uncertainty, so a larger budget would be unlikely to change the study's conclusions. We have addressed this in the manuscript: *The fixed number of trials is chosen where the optimization usually plateaus (Bergstra & Bengio, 2012)*
Fig.4	I don't think these colors work for a person with color-blindness (monochromacy/achromatopsia).	We agree with the Reviewer. New color palette is chosen for the Figure 4.