Multi-site learning for hydrological uncertainty prediction: the case of quantile random forests

Taha-Abderrahman El Ouahabi, François Bourgin, Charles Perrin, and Vazken Andréassian

https://doi.org/10.5194/egusphere-2025-3586

Reply to Derek Karssenberg's comments

Thank you very much for your comments on our manuscript, which will help improve our work. Please find below our replies (in blue) to your comments (in black), as well as how we intend to modify the paper to account for your suggestions and recommendations.

The manuscript proposes and evaluates the use of quantile random forests for correction of streamflow predicted with a process-based model. The main innovation compared to previous studies on streamflow error correction is the use of quantile random forests as it provides a means to estimate uncertainty in corrected streamflow. Also, unlike previous studies, this study extensively compares results for approaches that use local, regional, or national (France) data for training the error correction model. In my opinion this is a very interesting study. The methodology is state of the art, and the manuscript is relevant to development of error correction methodology (also in other domains).

My comments mainly refer to how the study is presented while I suggest a number of relatively small additions. Please find below my main comments followed by a list of minor comments.

Introduction

The introduction needs some revision to further increase the impact of the study and to make it more accessible. The problem definition needs to be defined more completely and more precisely. It remains unclear what the 'simulation context' (used in the methods section, line 125) is. In my opinion it is important to clearly state that this paper is about error correction of process-based model predictions, in the situation/context where predictions are made without relying on extrapolation of past observations of streamflow (for short range (small lead time) forecasts this would be more powerful). Also, the paper is not about prediction for ungauged catchments as all models are trained on historical streamflow at the location for which predictions are made. The simulation context thus is, I think, mainly reconstructing or projecting (e.g. under scenarios of climate change) streamflow for catchments that have streamflow available.

Also, the second contribution (line 62, spatial catchment descriptions) does not come with a clearly substantiated problem addressed by this contribution.

Please clearly describe what is meant by ‘regional’ in ‘regional learning’, ‘regional approaches’, ‘regional bias’, ‘regional post-processing’, etc. It is central to the study but it is not clearly defined. Is ‘multi-site learning’ (line 70) the same (please explain in manuscript).

Please clearly describe what is error corrected (i.e. streamflow from a process-based model). This is not clearly stated in the introduction (e.g. line 64 ‘model states’, model states from what?).

What are ‘spatially varying catchment characteristics? Line 64. Please explain or rephrase.

By ‘simulation context’, we mean the reconstruction/projection of streamflow for catchments that have streamflow available. We will clarify the manuscript in order to precisely introduce the problem definition. Used terminology and syntax will be introduced more clearly.

Hyper parameter tuning – metrics used

I suggest moving information from Appendix A (line 398 – 403) to the main text (Methods), in particular the fact that hyperparameter tuning is done on metrics that refer to probability distributions (instead of deterministic ones). To my knowledge this study is ***quite unique in doing so*** (but I may be wrong but even then I would still move it to the main text). It is also suggested to state in the main text that in the local modelling, hyperparameters are different between catchments (which should further improve the results for the local model compared to an approach fixing hyperparameters across catchments).

We have moved the section. We have also added the section that used to be in the Appendix to highlight that QRF-local hyperparameters are fitted for each catchment individually.

Assessment criteria

The assessment criteria are well chosen. However, I think it can be presented better in the Methods and Results section.

First, I suggest giving additional explanation on the terminology. If I am correct ‘sharpness’ refers to the magnitude of the uncertainty of the prediction, i.e. the lower the better. Please try to explain this more extensively as not all readers will be familiar with this term. The term ‘reliability’ in the context of your manuscript refers to whether the modelling approach is capable of providing correct estimates of the uncertainty (preferably the complete distribution should be correct).

Second, I suggest then to somewhat more clearly explain to what (sharpness or reliability) each metric refers. For instance, both the alpha score and coverage refer to ‘reliability’. Connecting these metrics could also be done in the Results section; e.g. one would expect similar results (relative performance between local, regional, national) for alpha and coverage ratio as these both refer to reliability. This is not stated at all in the Results and the reader has to make it up by herself.

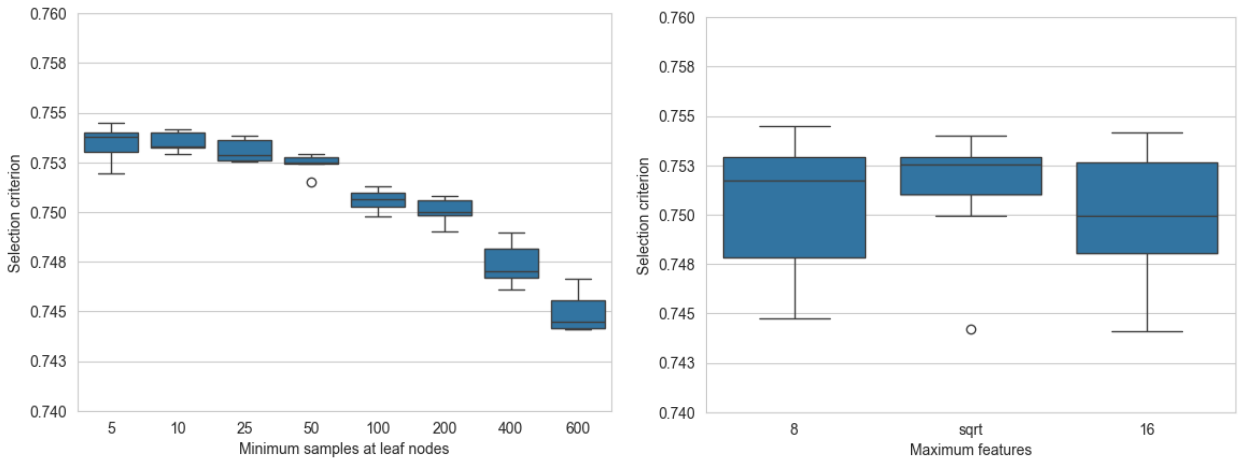
Third, please be precise in the explanation. For instance ‘It calculates the closeness of predicted uncertainty distributions to the statistical distribution of observed streamflows’ (line 221) reads like you compare the distribution of the error term (from the model) with distributions of streamflow (over time?). This is not at all the case! Instead alpha is a metric summarizing the QQ plot, which is really a probabilistic property (as the authors will certainly be aware of). Also, please use correct units (for instance, CRPSS is given as percentage in the figure while in the main text it is in the range of 0-1 it seems).

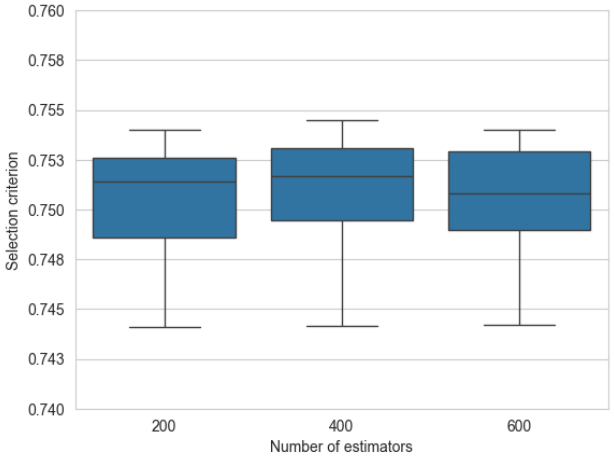
We agree the original wording was unclear. Indeed, sharpness refers to the magnitude of the uncertainty estimates, and normally, lower values of sharpness refer to better uncertainty estimates. The used metric however was transformed to a skill score, which made values of the Dispersion score closer to 1 better than lower score values (similarly to CRPSS). We have revised the text to better explain the definitions of reliability and sharpness, in addition to clearly highlighting the better performances direction for the used metrics. The link between the alpha score and coverage ratio will also be clearly highlighted.

Results

Please summarize the results of the hyper parameter tuning.

We conducted a hyperparameter grid search for each QRF variant and used the aforementioned probabilistic metric combining both Alpha score and CRPSS. These hyper-parameters include the size of the forest (number of trees), the minimum number of samples at child nodes, and the maximum number of candidate variables to use for splitting at each tree node. Each hyperparameter search was repeated with three seeds and the median score was calculated to obtain a more robust selection. Overall, the performances of QRF were most sensitive to the minimum number of samples at child nodes. Using the dataset described above, QRF was trained with minimum number of samples at child nodes ranging from 5 to 600 data points. The following figure shows the impact of hyperparameters on the selection metric for QRF-national:





***New figure 1**: Hyper-parameters optimization results for QRF-national. The selection criterion is the median hyperparameter criterion across the catchments of the study.*

It is notable that best results were recorded for lower values of the hyper-parameter 'minimum samples at leaf nodes'. The improvement slows for values lower than 25, and the mean scores were within one standard deviation of the mean scores for the other values. It is worth mentioning that very low values of minimum samples at leaf nodes might result in overly complex QRF models. As such, a minimum samples at leaf nodes of 10 is selected. Overall, QRF was found to be fairly insensitive to the number of candidate predictors used for splitting at each node. By default, the quantile-forest library uses the integer value of the square root (sqrt) of the total number of predictors for this parameter. With 31 total predictors for QRF-national, 6 would be the default and the previous figure showed that using the default value of square root was slightly better. For the number of trees parameter, a forest with more trees will generally be more skillful than one with fewer trees, as it can fit more on the nuances of the training set, and there is a point when the rate of improvement with more trees is negligible, as noted in [1, 2]. Most of the boxplots ranges overlap, and it seems that the results are not overly sensitive to this QRF parameter. For the experimented values, the above Figure shows that a number of 400 trees allows for slightly better performances.

We followed a more automatized approach for the selection of the hyperparameters of the other QRF-variants. And the distribution of the selected hyper parameters can be found in the appendix.

Magnitude of uncertainty (sharpness)

It seems the modelling approach underestimates uncertainty for all scenarios. This is an important outcome. Please add this information to the Results (it is not mentioned at all) and provide possible explanations in the Discussion section. One possible cause is the fact that the approach neglects uncertainty in the streamflow prediction of the process-based model.

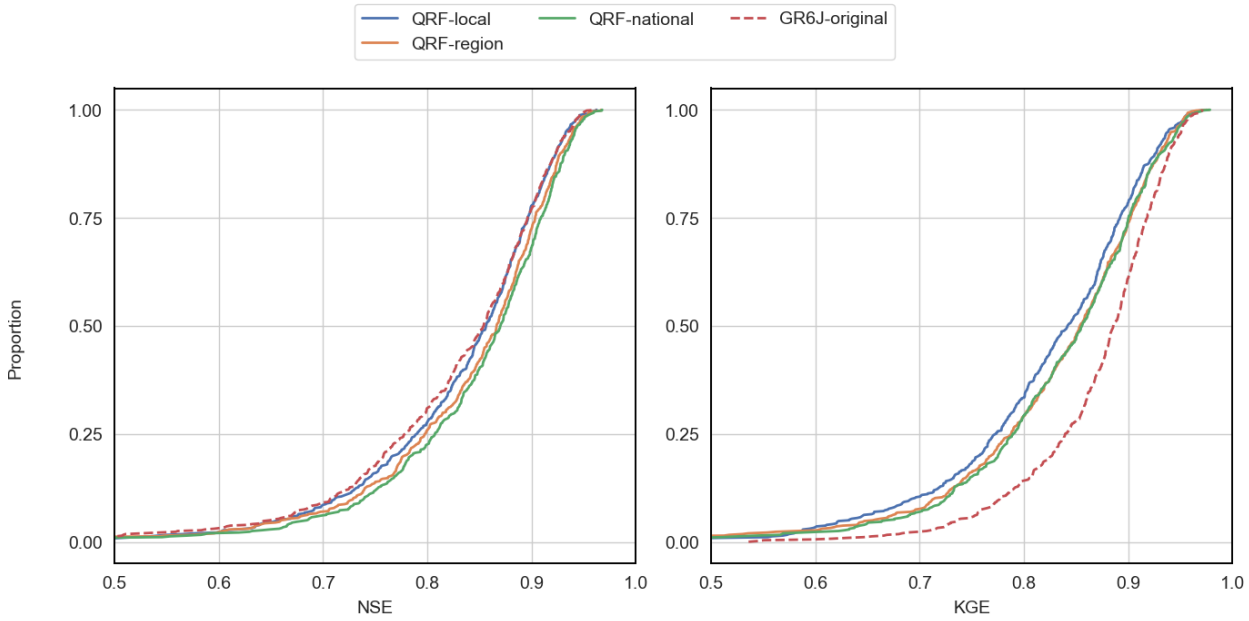
It is true, there are certain catchments where uncertainty is severely underestimated. It is worth mentioning that negative values rather denote overestimation. This is linked to the

existence of 0 values for certain catchments which makes the estimation of uncertainties more difficult.

Process-based model as benchmark

I am aware the main question of the manuscript is not on how much error correction contributes to improved streamflow prediction compared to using streamflow from the process-based model (without error correction). However I am in the opinion that it is still extremely interesting to add information (and if possible a short discussion) on the performance of the process-based model before adding the error correction. This could be done by adding curves for this process-based only benchmark to figures, or values in tables, or values in the main text. This would also allow you to compare the results regarding the improvement of streamflow prediction after error correction with those in Magni (2023), Shen (2022) and possibly others. It the improvement in your study comparable to other studies?

We agree with this element and we propose to add the following paragraph to section 4.3 (Deterministic metrics). It compares mean QRF estimates with the original GR6J predictions:



New Figure 2. *CDF of deterministic metrics across the 564 catchments for the QRF variants in the study during the testing period. The blue line represents the performance of QRF-local, orange represents QRF-region, green represents QRF-national, while red the performance of raw GR6J predictions.*

Figure 1 also provides deterministic metrics for the raw GR6J predictions. The figure highlights that the proposed QRF methods can improve hydrological deterministic predictions, especially for NSE. For example, QRF-national produced better NSE

performances compared to GR6J predictions (0.87 vs 0.86 in median NSE) and for 75% of the study's catchments. Overall, QRF variants had better NSE for the majority of the catchments. For KGE, the raw GR6J estimates outperforms all tested QRF approaches. We expected that the proposed variants would also improve KGE as was observed in the previous studies of [3, 4, 5]. For example, [4] used a closely related deterministic RF framework for hydrological error correction and found that the hybrid RF approach boosted streamflow predictions from a median of -0.03 to 0.51. Here, the post-processing was not beneficial for KGE performances. We argue that this can be in part attributed to how QRF hyper-parameters were selected. The used statistical criterion in this study aims to maximize the probabilistic performances of reliability and sharpness of the uncertainty estimates. While for the aforementioned studies, the RF post-processor was optimized specifically for the KGE criterion.

Title

Consider revising such that it also covers the fact that this manuscript is about error-correction (or combining process-based modelling and machine learning – sometimes referred to as hybrid modelling). I agree that the case is quantile random forests but the case is also error correction (maybe more so).

Thank you for highlighting this point. We propose to modify the title as follows:

Multi-site learning for hybrid error-correction: using quantile random forests for hydrological uncertainty prediction

Minor comments

Line 64 What does 'For this..' refer to?

Figure 1 Please add a scale bar.

These point will be addressed in the revised version.

Line 105 Please state what parameters were calibrated.

The following parameters of GR6J were calibrated:

- X1 – Production store capacity [mm]

  Controls the maximum soil moisture storage capacity of the production reservoir.

- X2 – Groundwater exchange coefficient [mm/day]

  Governs the rate of exchange between groundwater and the streamflow system.

- X3 – Routing store capacity [mm]

Represents the capacity of the non-linear routing reservoir responsible for routing.

- X4 – Time constant of unit hydrograph [days]

  Controls how quickly water is routed through the hydrograph

- X5 – Groundwater exchange threshold [–]

  Influences the sign of groundwater exchange.

- X6 – Exponential store coefficient [mm]

  Relates to the capacity/depletion of the exponential store, which improves low-flow simulation.

These parameters will be included in the revised version of the manuscript.

Line 105 'prior transformations' What is meant by this?

The two power transformations are applied separately to streamflow values – both observed and simulated values – before calculating the two associated KGE criteria. We will add a reference to the recent study of [6], where the use of streamflow transformations for model calibration is investigated.

$$KGE(Q^{obs}, Q^{prd}) = 1 - \sqrt{(r(Q^{obs}, Q^{prd}) - 1)^2 + (\alpha(Q^{obs}, Q^{prd}) - 1)^2 + (\beta(Q^{obs}, Q^{prd}) - 1)^2}$$

The power transformation is applied on Q^{obs} and Q^{prd}. This will be added to the appendix.

Line 125

It is stated here that in the simulation context of this study, streamflow is not available. This is not really true. The manuscript describes a methodology that only applies to locations where streamflow is available (for training, validation). For testing (or projections/reconstruction) I agree it can be done without measured streamflow (for the timesteps for which testing is done) but in this manuscript, results/testing metrics are only presented for locations where streamflow was used for training (i.e. this is not an ungauged catchment study). This is in my opinion not an important limitation, but it has to be clearly stated what this study is about (please refer to my comments related to the introduction).

Indeed, we were not clear with this element. The context of the proposed manuscript does not target the prediction at ungauged catchments context -but this will be specifically treated in an additional paragraph in the discussion section. We will clear out any mis-understanding regarding this point and highlight that the proposed methodology necessitates observed flows (to calibrate the hydrological model and calculate errors) and is geared towards cases of projections and reconstruction.

Line 130, ‘production’

What is meant here?

We intended to refer to the production store of the hydrological model. This point will be clarified in the revised version of the manuscript.

Line 131, ‘moving averages’

What was the filter size?

Moving average filter size is equal to the catchment response time, which was obtained from the X4 parameter of the used hydrological model. This hydrological parameter is related to the timing of the catchment’s response to rainfall. We will add this information in the revised version of the manuscript.

Line 194

Refer to Figure 1

Line 236

Number -> proportion

The remaining points will be addressed and corrected in the revised version. Thank you very much for your comments.

Bibliography

[1] Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas, 2012. How many trees in a random forest? In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-31537-4_13

[2] Breiman, Leo, 2001. Random forests. Machine Learning, 45(1), pp.5–32. https://doi.org/10.1023/A:1010933404324

[3] Zhang, Yuhang, Aizhong Ye, Bita Analui, Phu Nguyen, Soroosh Sorooshian, Kuolin Hsu, and Yuxuan Wang, 2023. Comparing quantile regression forest and mixture density long short-term memory models for probabilistic post-processing of satellite precipitation-driven streamflow simulations. Hydrology and Earth System Sciences, 27(24), pp.4529–4550. https://doi.org/10.5194/hess-27-4529-2023

[4] Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E.H. and Karssenberg, D., 2022. Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. Computers & Geosciences, 159, p.105019. https://doi.org/10.1016/j.cageo.2021.105019

[5] Magni, M., Sutanudjaja, E.H., Shen, Y. and Karssenberg, D., 2023. Global streamflow modelling using process-informed machine learning. Journal of Hydroinformatics, 25(5), pp.1648–1666. https://doi.org/10.2166/hydro.2023.217

[6] Thirel, Guillaume, et al. "On the use of streamflow transformations for hydrological model calibration." Hydrology and Earth System Sciences 28.21 (2024): 4837-4860. https://doi.org/10.5194/hess-28-4837-2024