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 Anonymous Referee #1

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.

Overall comments

This is a consistently interesting and well-considered study on the benefits of using multiple sites to train a probabilistic machine learning method (QRF) to predict hydrological model errors. The study shows that the inclusion of multiple sites indeed does improve the performance of QRF. The study has clear aims, and its conclusions are well supported by rigorous cross-validation and forecast verification. This is a minor point, but I was quite taken with their innovative way of measuring sharpness using the CRPS, which neatly sidesteps the issue of having to focus on one or two intervals with average width of prediction intervals, the conventional method for assessing of sharpness (which can give contradictory results for different intervals, and of course tells you nothing about intervals that are omitted).

The finding that the use of multiple sites helps the QRF may in some ways seem obvious in retrospect, but this is the thing with the best studies: the findings often look obvious after the authors have presented them! Accordingly, I think the study makes a significant contribution to the literature on hydrological error modelling. Their further investigation of the use of regional/national methods of including sites provides practical guidance to anyone wishing to implement their methods, of which I am one.

For my own interests I would have liked to have seen a comparison with a more conventional error modelling technique - e.g. a simple AR1 model assuming Gaussian errors after transformation - but I understand that this would have considerably lengthened the study, and is not strictly within the aims of what the authors set out to do. So I am happy for this to be omitted. I have a few questions about methods in the specific comments, the most notable of which is whether static climatic/hydrologic predictors are cross-validated. Assuming the answer is 'yes', I recommend this study be published essentially in its present form, subject to technical corrections.

Thank you for the positive feedback on our work. We agree that a comparison with other and simpler methods would be interesting, but as mentioned we feel it is beyond the scope of this study. We will add a call for more comparative studies in our conclusion. Regarding the cross-validation of static predictors, we have provided some arguments in our response

below, but we remain open to perform additional experiments. This will involve re-computing the climatic predictors from the CAMELS-FR database and checking the impacts on our results.

Specific comments

L53 "They found that larger LSTM models trained on all available basins outperform smaller models trained on a limited set of catchments. This is because, for some ML approaches, models calibrated on larger training datasets can outperform smaller and more specialized models" This is effectively saying "LSTMs perform better on larger datasets because LSTMs perform better on larger datasets". Please avoid instances of circular reasoning like this. An additional point is that as far as I understand it LSTMs significantly outperform conventional rainfall runoff models for predictions in ungauged basins. This differs from applications where models are calibrated and used on the same catchment, where conventional rainfall-runoff models can perform similarly well to LSTMs. This may be worth mentioning.

Thank you for this feedback, we will modify the manuscript to improve clarity. We agree that LSTMs outperform conventional rainfall-runoff models for predictions in ungauged basins, but that the current situation is more balanced for predictions in gauged basins. This will be highlighted in the modified manuscript.

L83 "Potential evaporation (PET) is calculated using the formula proposed by Oudin et al. (2005)." Could the authors briefly list the inputs used in this formula?

Yes, PET formula is given by the following equation:

$$PET = 0.408 \times Ra \times (T + 5) / \lambda$$

If $T \leq 5°C$, then $PET = 0$

Where:

- PET = potential evapotranspiration (mm/day)
- Ra = extraterrestrial radiation (MJ/m²/day)
- T = mean daily air temperature (°C)
- λ = latent heat of vaporization (≈ 2.45 MJ kg^{-1})

Ra depends on the localization of the basin and Julian day values and the temperature is the only dynamical meteorological input used to estimate PET. We will include this in the revised version of the manuscript.

L84 "Since our interest is in developing a multi-site QRF post-processor, we used several static basin-averaged attributes describing climate, topography and geology." Would be good to foreshadow that these are listed in Table 1.

Thank you for this suggestion, we will add this information.

L108 "with a power of 0.5 and -0.5 prior transformations on streamflow" It's not clear to me where the power is being applied and what this transformation is. Please specify - in an appendix is fine.

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 [1], 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.

L154 Table 1. As is common, many of the static descriptors are based on climatic/hydrologic predictions (mean precip, PE, temp, etc.). I just wanted to confirm that these were cross-validated for this study: i.e., that they were computed separately for each of the training, validation and testing periods. As I'm sure the authors are aware, rigorously cross-validating predictors is an important aspect of testing any prediction system. I realise this kind of cross-validation is sometimes not done in ML studies, but it should be.

We understand the concerns and we agree that rigorously cross-validating predictors is important. As rightly pointed out, this is indeed not done in most of the ML studies we are aware of (e.g. [2]), and we followed this debatable common practice. We can understand both arguments: the need to cross-validating predictors and, on the other hand, the use of long-term climatic values taken as “static” descriptors. Our understanding of the role of the static descriptors in multi-site training is that they allow for creating some similarity-based relationship between catchments and that this relationship should be “static”. We understand that some catchments have a non-stationarity behavior, but we feel that looking at the implications of non-stationarity is beyond the scope of this study.

L225 "The sharpness metric is the continuous ranked probability score (CRPS)" This is a really nifty way of measuring sharpness!

Thank you for your comment. It is quite unusual indeed, but we found in our previous works that it was an interesting alternative to other ways of measuring sharpness. It is related to an interesting decomposition of the CRPS proposed in the PhD thesis of [3] (in French). We will add the reference to the work of Bontron.

L258 "In terms of sharpness, the different QRF variants performed similarly, which is interesting given that multi-site setups significantly improve CRPSS values". Might be worth saying why it's interesting: it's quite possible (even unsurprising) that sharpness (a property of the prediction only) is the same but CRPSS (which considers the joint distribution of obs and predictions) differs.

Thank you for pointing this out: we intended to say that this shows that regional QRFs give more importance to reliability (as they consider the joint distribution of observations and predictions).

L272 Figure 4. Might be worth stating that curves that track closer to the right of the plots indicate better performance in the caption.

We will clarify this in the revised manuscript.

L286 "Table 3 summarizes the average values of the alpha, dispersion, CRPSS, and interval scores for three flow groups: high (> 67%Qsim), medium

(> 34%Qsim and < 66%Qsim), and low flows (< 33%Qsim)." Please confirm that performance scores are stratified based on when predictions exceed these thresholds, not when observations exceed them.

Yes, performances were stratified based on the median values of the probabilistic predictions. This will be clearly highlighted in the revised version of the manuscript.

L314 "Furthermore, the aforementioned scale discrepancies occurred specifically for catchments characterized by frequent zero values in simulated and observed streamflows." I wondered about this. Normalising errors with a log transformation is one thing, but maintaining normality in the presence of zeros in observations - and potentially also in the the simulations after the QRF is applied - is quite another. While this isn't a total solution, might it be helpful to consider the proportion of zero flows as a static predictor?

Yes, indeed. We will further discuss this in the following paragraph.

L320 Discussion. I would have liked to see a paragraph or two added that briefly discusses the following topics. However, I understand that my interests are not necessarily the authors' and also may not be interesting to a more general audience, so I leave it to the authors to decide which of these issues (if any) they may wish to discuss:

1. The weaknesses of the method (discussed throughout the manuscript) - e.g. application to ephemeral catchments - and how the authors might improve these.
2. The sensitivity of the method to data availability. The authors used the astonishingly comprehensive CAMELS-FR dataset, but many of us work in regions with only a fraction of this gauge coverage. e.g. What would have happened if they only had access to 50 gauges in their dataset? What might have happened if observations are concentrated on a particular hydrological type, but applied outside this type?
3. Are there prospects for applying this method to produce reliable probabilistic predictions in ungauged basins?

We plan to include the following paragraphs in the revised manuscript (section 4.6), in order to discuss these issues:

The provided analyses highlighted some limitations of the multi-site QRFs. The first concerns ephemeral catchments which are characterized by zero flow values. Modelling

ephemeral catchment dynamics is generally challenging [4, 5], and this is especially the case in our study with the use of the log-based error transformation. Figure 9 of the results section clearly showed that multi-site learning can degrade the predictions for ephemeral catchments as they often overestimate uncertainty. Although we have added the δ offset parameter, the use of an alternative transformation that is less sensitive to zero flow values (e.g., a Box–Cox transformation) could better stabilize hydrological errors used to train the QRF models. Another pragmatic solution could be the treatment of such catchments separately when training multi-site QRF. As showed in Figure. 9, QRF-local better managed the case of zero flow values. In the literature, other approaches [6, 7] use adapted catchment groupings based on other attributes (climatic, etc.) or a statistical clustering approach classifying homogenous catchments together. Similarly, the use of the number of zero flow values as input feature could also help QRF to better distinguish catchments characterized by this issue, and help QRF to find adequate analogous events.

We used a large sample dataset (CAMELS-FR) for models training, but many practical hydrological applications only have access to a limited number of gauges for training purposes. The generalizability of uncertainty estimates of QRF to catchments outside its training region was treated in various studies. We did not explicitly test this, but we believe that the generalizability of regional QRF variants depends on the similarity between hydrological errors for catchments where the training occurred and regions to which extrapolation of the uncertainty estimates will be carried out. Finally, the proposed framework can be adapted for a prediction at ungauged basins, and a proper spatio-temporal cross-validation experiment [7] would be needed to verify this. The main practical difficulty lies in obtaining consistent hydrological model states for ungauged catchments, and to adapt the significant additional uncertainty usually associated with such settings [8, 9, 10]. If this can be properly handled, the proposed QRF multi-site variants could provide meaningful uncertainty estimates for the context of uncertainty estimation for ungauged catchments.

L362 "However, because of memory issues, we trained QRF-national on Jean-Zay HPC, where a single node with two CPUs (at 2.5 GHz) and 128 GBs of memory was sufficient." I'd be interested to know how long the parameters estimation took on this hardware. Cloud computing means that many now have access to large computers, but the run-time can still make these resources expensive.

For each parameter, it takes around 25 minutes for QRF-national.

L141 "But, it is important to note that these scale features are not available" suggest deleting 'But,'

Grammar and terminology will be checked in the final version of the manuscript. Thank you very much for your comments.

Bibliography

[1] 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

[2] Auer, A., Gauch, M., Kratzert, F., Nearing, G., Hochreiter, S., and Klotz, D.: A data-centric perspective on the information needed for hydrological uncertainty predictions, Hydrol. Earth Syst. Sci., 28, 4099–4126, https://doi.org/10.5194/hess-28-4099-2024, 2024.

[3] Bontron, G., 2004. Prévision quantitative des précipitations: Adaptation probabiliste par recherche d'analogues. Utilisation des Réanalyses NCEP/NCAR et application aux précipitations du Sud-Est de la France (Doctoral dissertation, Institut National Polytechnique Grenoble (INPG)). https://tel.archives-ouvertes.fr/tel-01090969

[4] McInerney, D., Kavetski, D., Thyer, M., Lerat, J. and Kuczera, G., 2019. Benefits of explicit treatment of zero flows in probabilistic hydrological modeling of ephemeral catchments. Water Resources Research, 55(12), pp.11035–11060. https://doi.org/10.1029/2018WR024148

[5] Li, M., Wang, Q.J., Bennett, J.C. and Robertson, D.E., 2016. Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. Hydrology and Earth System Sciences, 20(9), pp.3561–3579. https://doi.org/10.5194/hess-20-3561-2016

[6] Hashemi, R., Brigode, P., Garambois, P.A. and Javelle, P., 2022. How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models? Hydrology and Earth System Sciences, 26(22), pp.5793–5816. https://doi.org/10.5194/hess-26-5793-2022

[7] Fang, S., Johnson, J.M., Yeghiazarian, L. and Sankarasubramanian, A., 2024. Improved national-scale above-normal flow prediction for gauged and ungauged basins using a spatio-temporal hierarchical model. Water Resources Research, 60(1), e2023WR034557. https://doi.org/10.1029/2023WR034557

[8] Razavi, T. and Coulibaly, P., 2013. Streamflow prediction in ungauged basins: review of regionalization methods. Journal of Hydrologic Engineering, 18(8), pp.958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

[9] Oudin, L., Andréassian, V., Perrin, C., Michel, C. and Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44(3). https://doi.org/10.1029/2007WR006240

[10] Bourgin, F., Andréassian, V., Perrin, C. and Oudin, L., 2015. Transferring global uncertainty estimates from gauged to ungauged catchments. Hydrology and Earth System Sciences, 19(5), pp.2535–2546. https://doi.org/10.5194/hess-19-2535-2015