Editor

Dear authors,

The answers to the review comments are very solid, however, I do have a specific comment that I would like to see addressed before publishing: In one of your responses to referee 1 you write:

"To further disentangle the effect of spatial autocorrelation we have run a 10-fold spatial block cross-validation in the Random Forest analysis. This is, instead of randomly sampling catchments form the total set, we have split the samples according to contiguous spatial blocks (e.g. training set: catchments in south-west, testing: catchments in north-east of the domain). While training accuracies remained largely the same, testing accuracies decreased as compared to the original analysis (Table C2). These results suggests that spatial autocorrelation contributes to inflated performance under random cross-validation, and that spatial proximity has a considerable effect on model predictive skill. This, in turn, indicates that even the wide spectrum of climate and landscape attributes considered in the analysis, carries less information than what is suggested by a random analysis."

I disagree with this conclusion. Even if you have a extremely informative feature for your model (say, for example, precipitation for a rainfall-runoff model) and you only train a model on a non-random subset of values for that features (say, only when the precipitation is zero) your model skill will be good for the training part and bad for the validation/test where you evaluate on the other subset of values (say, when precipitation is non-zero. As much as I appreciate the effort and the conducted analysis, but I do believe your conclusions are wrong. Hence, I decided to give you minor revision in the hope that you can either clarify your position or rethink the argument for the manuscript.

We would first like to sincerely thank the editor for handling our manuscript review process, and for raising this specific comment. We agree that spatial block cross-validation introduces a distributional mismatch between training and test sets, since spatially neighboring catchments tend to share similar climate and landscape characteristics, the model may be evaluated on feature ranges it was not trained on, which can reduce predictive skill even for highly informative predictors. We have therefore toned down our original interpretation accordingly.

Nevertheless, we are of the opinion that the overarching goal of any model – data-driven or process-based – is to meaningfully represent real world processes (or combination/weighting of features) that are relevant to its objective. A “good” model, i.e. a model that captures these relevant processes, will also perform adequately when confronted with previously unseen data – no matter if they are from the same distribution or not. Clearly, we agree that in reality, there will in most cases be a performance drop,

particularly when the unseen test data are from a different distribution. Although there is no formal and quantitative consensus how much of a performance drop is acceptable, we interpret the results of our analysis in the way that the performance drop was too elevated to be exclusively attributed to differences in data-distributions. Instead, we think that it *also* reflects the models' inability to represent real-world processes either due to (1) insufficient discriminatory power in the data and/or (2) an unsuitable combination of feature variables (perhaps other, unknown and unavailable variables not used in our analysis may carry more discriminatory power).

As spatial block validation is merely a variation of the classic Differential Split Sample Tests as already described by Klemes (1986), which is widely accepted to be the gold-standard in model testing, we strongly believe that our conclusion is not generally invalid. Rather, we would argue that the substantial performance drop under spatial block cross-validation can also be viewed as plausible evidence that spatial autocorrelation contributes to *some degree* of performance inflation under random cross-validation, and that spatial proximity carries predictive information not fully captured by the measured attributes. We believe *both factors* contribute to the drop in testing accuracy and have revised the manuscript (lines 593-609) to reflect this more balanced interpretation, by also acknowledging the distributional mismatch as an explanation for the observed performance drop.

The spatial autocorrelation section is revised as follows (lines 593-609): **"These results suggest that the performance decrease under spatial block cross-validation likely reflects a combination of two effects. On the one hand, spatial blocking introduces a distributional mismatch between training and test sets, as spatially neighbouring catchments tend to share similar climate and landscape characteristics, meaning the model is evaluated on conditions that are underrepresented or absent during training. On the other hand, it remains plausible that spatial autocorrelation contributes to some degree of inflated performance under random cross-validation, given that spatially proximate catchments are more likely to have the same hydrological regime membership. Disentangling these two effects is non-trivial, and both mechanisms likely contribute to the observed performance drop. Nevertheless, the substantial decrease in test accuracy under spatial block cross-validation is consistent with the interpretation that spatial proximity carries predictive information not fully represented by the available climate and landscape attributes, suggesting that under random cross-validation, the model may partly rely on spatial covariance structures that are not fully resolved by the available attributes. This, in turn, highlights the challenge of disentangling hydrological drivers across regions with strong spatial structure, and underlines that currently available data are largely insufficient to resolve the "uniqueness of place" (Beven, 2000) in a meaningful way."**

References

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrol. Earth Syst. Sci., 4, 203–213, https://doi.org/10.5194/hess-4-203-2000, 2000.

KLEMEŠ, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024