Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-430-AC3, 2021

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Resampling and ensemble techniques for improving ANN-based high streamflow forecast accuracy" *by* Everett Snieder et al.

Everett Snieder et al.

usman.khan@lassonde.yorku.ca

Received and published: 1 January 2021

RC1-1. This manuscript compares different resampling methods and different ensemble- building techniques to improve ANN-based flood forecasts (high streamflow). Those resampling methods and ensemble techniques are also combined, resulting in a total of 16 variants, that are compared to a base model. The base model is a classic Mul- tilayer Perceptron with 10 hidden neurons and 25 input variables, trained using a stop training approach and the Levenberg-Marquardt algorithm. All the 16 variants and the base model are applied to simulate (I think) the streamflow for two rivers in Canada.

The manuscript is very well written, well organized and very clear. However, I am sorry to say that I find the originality and contribution to be very low, too low in my opinion for a publication in HESS. The research was certainly conducted with great care, but ensemble techniques have already been used for quite a while in hydrology (both in the hydro-informatics community and beyond).

AC1-1. Thank you for the positive comments on our manuscript. However, we would like to clarify the novelty and contribution of the research presented in our manuscript:
+ Extensive comparisons of resampling and ensemble methods (independently and combined) to address the data imbalance problem in data-driven hydrological models. The authors restate the novelty of embedding resampling methods in ensemble methods (as opposed to resampling as preprocessing), which has not previously been studied for hydrological stage forecasting.

+ Many of the combination methods proposed in our work encourage diversity-in-learning, which distinguishes the algorithms from previous work on simple Bagging or boosting methods.

+ Comparison of two watersheds that have different dominant hydrological processes, spatial and temporal scales.

+ The heteroskedastic nature of flow forecasting models, which our work attempts to address, is a persistent issue in the field of hydroinformatics.

+ A particular focus on high stage (rather than the entire timeseries) to assist in early warning systems or flood forecasting.

+ While some previous use in the broader machine learning literature, some methods, adapted and implemented in the manuscript, are a first in hydrology for flood forecasting.

+ To our knowledge, the variations of SMOTER, SMOTER-AdaBoost, and LSBoost with resampling developed in our research are novel implementations.

+ Even though some methods, and resampling or ensemble techniques, have been used in hydrological studies, a systematic comparison, as presented in our manuscript, is still needed to properly evaluate their efficacy, particularly for high flows.

RC1-2. While the authors mention that most of the resampling and ensemble techniques they used are not common in hydrology, this is still just an additional application of existing techniques, some of which have already been compared. Further, there is very little discussion and analysis of the results.

AC1-2. We apologise that the depth of analysis in the original submission was not found to be thorough enough. We have expanded the analysis and discussion in addressing specific comments made by the other reviewers. The revised manuscript will have a longer discussion on the more important points. In summary, these include: + A formalised grid-search of ensemble size for base ensemble methods (with no resampling), which is discussed in greater detail in our response to AC1.13.

+ Additional analysis of the relative effects of the selection of the high stage threshold value (ranging from 50 to 90th percentile) for Bagging and SMOTER-Bagging.

+ Added an evaluation of the relative effects of ensemble size (ranging from 2-100) for Bagging and SMOTER-Bagging.

+ Added additional statements elaborating on the concept of diversity-in-learning and the role and effects of diversity attained using combined resampling - ensemble methods.

RC1-3. The author conclude that boosting methods provide only marginal improvements, they offer very little explanation.

AC1-3. The main finding of this research is that common ensemble methods, when combined with methods with resampling methods that increase the representation of high flow samples in the training distribution, only offer marginal improvements to model performance on high flows. Basic ensemble methods (with no resampling) are included

as a reference point, to quantify the improvements produced by the added resampling. The improvements of ensemble methods over the single model scenario is not discussed in detail, as this is already well established in existing research.

Additional analysis and discussion, as requested by Reviewer 2, has now been added; please refer to Figs 6, 11, and 12, and the associated text in the attached supplement.

RC1-4. In addition, I am not convinced that improving only specifically high flows, sometimes at the expense of what the authors call "typical flows", is the way to go. I also fail to see it as a major contribution to the hydrological sciences, worthy of publication in an international journal.

AC1-4. As stated in text, many studies have specifically noted poor performance of data-driven flow forecasting models on underrepresented flows (i.e., when the dataset is imbalanced, and in practice this may be high flows, as explored in this present research, or low flows, which may suffer from similar imbalance problems, and hence, the same methods may be applied). This issue of poor performance of flow forecasting models on underrepresented flows is the main thrust for the present research - a limitation that has been identified in existing research, and resampling or ensemble techniques (independently, or combined) is one way to address this problem.

The objective of this research is to obtain homoscedastic residuals across different flow values. However our results indicate a notable trade-off between typical and high flow performance. We believe that in certain circumstances such as a high flow warning system, such a trade-off would be desirable; in such a scenario, the discrete values of typical flows are of little importance, so long as they do not result in false positive warnings.

RC1-5. I will still detail some major and minor comments below, but my main concern regarding this paper is the level of the contribution, that I unfortunately find too low.

AC1-5. As mentioned above, we think there is a need for a systematic comparison

of resampling and ensemble methods within hydrology to address the data imbalance problem, particularly for high flows, since they are important for accurate flood forecasting models. While resampling is commonly used in flow forecasting studies, the SMOTER algorithm has only been used in a few instances and never for regression-based hydrological forecasting. Moreover, the resampling in such studies typically takes the form of preprocessing; in our studies, resampling methods are embedded within the ensemble algorithms. Such combinations appear in machine learning literature (mainly in classification studies) but, to our knowledge, ours is the first study to assess the effects of these algorithms on high flow performance and among the first applications in hydroinformatics. Combined resampling - ensemble learning algorithms achieve greater diversity-in-learning than either respective standalone method, hence have the potential to produce better performing models. The specific resampling algorithms chosen increase the influence of high flows in the model training process, thus addressing a common weakness of data-driven flow forecasting models: poor accuracy on high flows.

RC1-6. Detailed comments (apart from the contribution/originality issue): Abstract, lines 1-2: The affirmation "(. . .) are increasingly used for operational flood warning systems" should be supported by references to operational flood forecasting systems (not journal articles but documentation from company web- sites or government websites). At the moment, the affirmation is not only un- supported, but also opposite to my experience. To the best of my knowledge, there are very, very few operational agencies who use ANNs for flood forecasting, despite their use in research for more than 25 years. Also, this affirmation in the abstract somewhat contradicts page 10 lines 157-158 ("Consequently, such models may not be suitable for flood related applications such as flood warning systems").

AC1-6. We apologise for incorrectly inferring high commercial use of data-driven flood forecasting. The text has been modified to emphasise that such models have increasingly been featured in hydrological research, not necessarily used in practice: “Data-

driven flow forecasting models, such as Artificial Neural Networks (ANNs), are increasingly featured in research for their potential use for operational flood warning systems.”

The statement on page 10 aims to address the common claim that data-driven flow forecasts have high potential for flow forecasting applications; however, often such claims do not explicitly evaluate the performance of such models on high flows. When studies explicitly evaluate high flow performance for these models, they are often found to be lacking, especially relative to the performance on low or typical flows. Hence, there is a need to evaluate the ability of preprocessing and/or ensemble methods to improve model performance for high flows (which are important for flood forecasting and early-warning systems).

RC1-7. Page 2 lines 24-30: One of the causes of errors when simulating high flows in northern locations such as Canada is the occurence of ice jams. Ice jams are very common and not accounted for in any way by typical hydrological models. Maybe it is possible to account for them using ANNs, but I’m not sure. I think this is one major aspect that should have been present in the manuscript, both on page 2 but also when presenting the Bow River and Don River (it would be important that those rivers be ice jams free, otherwise you have to account for that). Another issue regarding high streamflow that is not discussed by the authors is the fact that those readings are extrapolations from the rating curve. The rating curve of a gauging station is typically constructed with very few (if any) observations of very high streamflow. Therefore, this part of the rating curve is very uncertain, and this is what we use to obtain "observations".

AC1-7. We apologise for the lack of clarity in the text about ice jams. Indeed ice jams were unaccounted for in this work: periods during which ice jams would occur in either watershed were removed from the data. Specifically, data from November to April and November to December were removed from the Bow and Don River models, respectively. This has been clarified in text: “Data from November to April and November to December were removed from the Bow and Don, respectively, datasets prior to any

analysis; these periods are associated with ice conditions."

The authors recognise the uncertainty associated with stage-discharge transformations. While we utilise language of 'typical' and 'high flows', only stage observations are used as model input or target features. Predicting stage directly (rather than flow) is sufficient for application flood early warning systems; thus the uncertainty associated with the stage-discharge transformation does not need to be considered. If discharge is required, the subject models could be reconfigured, thus the ANN would implicitly model stage-discharge uncertainty, or discharge could be calculated in post-processing, in which case it would be recommended that uncertainty be estimated and communicated with discharge forecasts. We have also changed the terminology in the manuscript, replacing all occurrences of "flows" with "stage", where appropriate.

RC1-8. Page 6, Table 2 and also Page 7 line 128: Why did you use the Levenberg-Marquardt algorithm? Although it is a popular algorithm, it was shown to have oscillation problems around local minima. I think the use of this algorithm should be better justified. See for instance Kwak et al. (2011)

AC1-8. The LM algorithm was selected because of its popularity, speed of convergence, and reliability [1]. Its suitability was confirmed based on a manual comparison with other available training algorithms in MATLAB for the baseline model. Existing literature has favourably described the LM algorithm, specifically for its ability to escape local minima in the error surface [2, 3]. We have added justification for using the LM algorithm in text as follows: "The Levenberg–Marquardt algorithm was used to train the base models, because of its speed of convergence and reliability [1-3]."

RC1-9. Page 6 line 116: What is the forecasting horizon? You mention the word "predict" here, but from reading the manuscript it seems like you perform simulations. If they are really forecasts, I think the forecasting horizon should be specified.

AC1-9. We apologise for the lack of clarity surrounding the forecast horizon. The forecast horizon is specified in text: "For both rivers, the input variables are used to

forecast the target variable 4 timesteps in advance, i.e., for the Bow River, the model forecasts 24 hours in the future, whereas for the Don River, the model forecasts 4 hours in the future."

Our models use previous timesteps to predict future data (e.g., Qt+4 is predicted using Qt, Qt-1, Qt-2). All models are calibrated using 80% of the available data while 20% of the data is isolated from model calibration and reserved for testing. Using a 10-fold cross-validation scheme, all of the data is included for testing exactly 3 times, across 10 different ensembles. The performance of the 10 ensembles are averaged or visualised in boxplots. In other words, when the models are "predicting" the testing data, they are doing so without "seeing" the test data - historical (i.e., previous) data is used to forecast future data. These predictions are compared to the observations for performance evaluation. Our models can easily be deployed in a real-world scenario as stage, precipitation, and temperature data are all instantaneously available online, and lagged versions of these data are used to forecast the future state of the system.

RC1-10. Page 10 line 170: I think there is a mistake here "(. . .) studies that featuring each (. . .)"

AC1-10. Thank you for identifying this mistake; it has been corrected.

RC1-11. Page 11: the distinction between RUS and ROS seems extremely thin to me.

AC1-11. You are correct - the distinction between RUS and ROS is minor. As stated in-text, RUS undersamples data as to achieve a balanced training set with no duplicates whereas ROS creates includes all available data and creates duplicates. Some studies have compared the two sampling techniques, however neither one of these two methods consistently outperforms the other, thus both are included in our study [4-6]. As requested by Reviewer 2, we will include pseudo-algorithms of each method in the revised manuscript to clarify the difference. The algorithms are attached in the supplement.

RC1-12. Page 13 lines 256-257: The definition of ESP that you provide here corresponds to Extended Streamflow Prediction, as per Day (1985), not Ensemble streamflow predictions. Ensemble streamflow forecasts (or predictions) can be obtained by a variety of manners, including feeding a hydrological model with dynamical meteorological ensemble forecasts.

AC1-12. We apologise for the confusion caused by this statement. This definition has been removed from the text.

RC1-13. Page 14: Why 20 members?

AC1-13. The ensemble size of 20 used in the original manuscript was determined based on an informal trial-and-error search. We deliberately used the same ensemble size for each method for the sake of comparison. In the revised manuscript, we have included a formalised ensemble size grid-search for ensemble sizes ranging from 2-100 for each of the base ensemble methods (i.e., with no added resampling). This analysis reveals that each ensemble method has a different optimum ensemble size; however, we decided to maintain a fixed ensemble size of 20 between all methods. Please refer to Fig. 6 and associated text in the attached supplement.

RC1-14. Results: Why do you aggregate the ensemble into a deterministic simula- tion and therefore evacuate the information about the uncertainty? Why not use ensemble-based performance assessment criterion such as the Continuous Ranked Probability Score, logarithmic score, etc.?

AC1-14. Thank you for this recommendation. The ensembles were combined for two reasons. Firstly, it allows for comparison against the single model scenario. Next, boosting methods are designed to be aggregated; AdaBoost ensemble predictions must be made using the weighted mean of the ensemble predictions and LSBoost uses a weighted sum combiner. Thus, ensemble performance criteria are only valid for uniformly weighted ensembles (RWD and Bagging). Your comment is absolutely correct in that the capability of generating a spread of predictions is an advantage

of these methods. However, for consistency, ensembles were combined into discrete predictions for all cases. That is, ensemble-based metrics are not applicable for all methods used in this research, and the only way for direct comparison is to use a "combined" metric.

RC1-15. Page 19 lines 457-459: How can overtraining happen if you have used stop training? I think this has to be explained more.

AC1-15. Thank you for identifying the point of confusion. Stop-training ensures that a model trained on a 'training' data partition achieves good generalisation on the 'validation' partition. However, this does not guarantee that the generalisation will carry to the independent 'test' partition. For example, if small and similar 'training' and 'validation' subsets are used and a large 'test' set is used, the model could very well be overfitted, despite the use of a stop-training criterion for determining the number of training epochs.

RC1-16. Page 21, Figure 7: From a general perspective, I don't see how decreasing the quality of simulation for typical flow values could be positive, even if the simulation of high flows is improved. Typical and especially very low flows are also important.

AC1-16. We believe that our statement is true for very specific applications, such as data-driven early flood warning systems (where the primary interest is on high flows that may lead to floods, rather than on typical or low flows). Moreover, the objective of this research was to reduce forecast error on high flows, not necessarily at the expense of typical flow performance; the trade-off between high and typical flow performance is simply the observed effect of the methods under study and not the objective. Finally, it is important to note that the methods for improving high flows examined in this research are transferable to other rare observations, such as low flows.

RC1-17. Page 26 lines 531-536: The analysis and discussion are very thin.

AC1-17. Thank you for this feedback. We have added a citation [7] to support the

tendency for boosted models to overfit. The brief discussion here is simply intended to allude that the improvements produced by boosting in this study being relatively lower than what is observed in other studies could be owed to the use of ANNs as the base learner. The vast majority of studies that use boosting utilise decision trees as the base learner, thus the outcome of these studies may not provide a reliable comparison. A formal comparison between different base learners for ensemble methods is not a goal of this study.

RC1-18. Page 26, section 4.2: I disagree with the idea of fine tuning the hyper parameters differently and specifically for each model, as it would violate the ceteris paribus principle, making it difficult to isolate and compare the influence of ensemble techniques and resampling techniques.

AC1-18. We did not intend to violate the ceteris paribus principle with the referenced statement. We made an effort to keep overlapping hyperparameters equal between different ensemble models. However, for the same base learner (a simple ANN) a standalone model and boosted model will have different complexities, hence different effective degrees of freedom (DOF). So, the question becomes should the individual model topology be kept equal (e.g., number of hidden neurons) or should the DOF be made equal? Moreover, some hyperparameters (e.g., Learning Rate in LSBoost) are specific to the method and have no counterpart among the other methods. Ideally, the comparison in this paper would be carried out for varying hyperparameter values and at varying degrees of model complexity; however, such a comparison would have a large computational cost and was considered beyond the established scope.

RC1-19. Page 26-27 lines 562-563: Ensembles are already quite common in ANN-based flow forecasting model, so this is not a very useful recommandation.

AC1-19. Thank you for this feedback. We acknowledge that ensembles are widely used in flowcasting, however we believe that they are worthy of investigation because of (1) the amount and variety of different ensemble methods and (2) the potential for

their combination with resampling methods. The outcome of the comparison presented in this work reveals that the combinations of ensemble and resampling methods under study do not outperform simple, proven ensemble methods with no resampling. In some cases, resampling can be used to produce marginal improvements in high flow performance, at a disproportionate trade-off with typical or low flow performance. Thus, our recommendation of using ensemble methods does not suggest that their superiority over single models is a novel finding; rather, that the combination of resampling and ensemble methods, which are widely used in machine learning literature, do not result in meaningful improvements when compared to the same ensemble methods with no resampling methods.

Additionally, we believe that our recommendation to the employ of simple ensemble methods such as Bagging are warranted, as despite their benefits being well established in this field of research, their use should be considered a minimum requirement for data-driven hydrological modelling.

References

[1] N. Lauzon, F. Anctil, and C. W. Baxter, “Clustering of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts,” Hydrol. Earth Syst. Sci., vol. 10, no. 4, pp. 485–494, Jul. 2006.

[2] H. R. Maier and G. C. Dandy, “Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications,” Environ. Model. Softw., vol. 15, no. 1, pp. 101–124, Apr. 2000.

[3] H. Tongal and M. J. Booij, “Simulation and forecasting of streamflows using machine learning models coupled with base flow separation,” J. Hydrol., vol. 564, pp. 266–282, Sep. 2018.

[4] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah,

“An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets,” Lect. Notes Electr. Eng., vol. 285 LNEE, pp. 13–22, 2014.

[5] M. Bach, A. Werner, J. Åżywiec, and W. Pluskiewicz, “The study of under- and over-sampling methods’ utility in analysis of highly imbalanced data on osteoporosis,” Inf. Sci. (Ny)., vol. 384, pp. 174–190, Apr. 2017.

[6] J. F. Díez-Pastor, J. J. Rodríguez, C. I. García-Osorio, and L. I. Kuncheva, “Diversity techniques improve the performance of the best imbalance learning ensembles,” Inf. Sci. (Ny)., vol. 325, pp. 98–117, Dec. 2015.

[7] A. Vezhnevets and O. Barinova, “Avoiding Boosting Overfitting by Removing Confusing Samples,” in Machine Learning: ECML 2007, vol. 4701 LNAI, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 430–441.

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2020-430/hess-2020-430-AC3-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-430, 2020.