Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-430-RC2, 2020

Hydrology and
Earth System
Sciences
Discussions
Open Access

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

Anonymous Referee #2

Received and published: 6 November 2020

The authors study the effect of resampling techniques, when integrated with ensemble learning frameworks, on the ability of the ANN based regression ensemble learners to improve prediction of high steam flow events. Two case studies are presented, with different temporal resolution and, essentially, hydrologic topology. One individual learner, that is MLP-ANN, is utilized in this study along with two ensemble models (Bagging and Boosting) as well as a randomized set of members (i.e. RWB model). Three resampling plans are examined, RUS, ROS, and SMOTER, to serve as the preprocessing re-sampler stage for the ensemble models. a combination of the latter is used with the ensemble models and all configurations are evaluated.

This paper attempts to answer an important question which is usually overlooked; can we diminish the heteroscedastic nature of stream flow predictions, which is inevitable when dealing with limited intelligence about the system dynamics (drivers to the instantaneous change in streamflow). The authors are concerned with the most volatile aspect in this setting, high flows, and whether more information can be utilized from the available descriptors' database to alleviate the problem. In general, ensemble learning is one of the few state-of-the-art solutions for improved short- to mid-term streamflow prediction, as individual nonlinear models are inherently instable and conventional statistical approaches sacrifices accuracy for probabilistic interpretability. As the diversity-in-learning mechanism, promoted differently by each ensemble architecture, is assumed to be the major reason to ensemble's generalization ability, resampling techniques are a major interest here. Consequently, it makes sense to study variations of ensemble learning frameworks with respect to the utilized resampling approach. This topic is increasingly gaining attention in the recent, and highly evolving, applied ML community in the field.

The paper is well-written (the chronologic format and section types are not similar to that I am used to, though). The presented results are critical, valid, and to-the-point. The resampling methods are described in organized wording and supported with references. The discussion covers most of the important aspects of this research. To this extent, I believe this paper meets HESS standards and scope, and is worthy of publication, after few important additions and modifications are implemented. Please refer below to the major and minor comments for consideration by the authors.

1. The summarization of the individual model calibration is good enough, as the focus is on the resampling-ensemble models. The use of the PC operator to select features from the predetermined bag of lags also makes sense. However, the authors should track the isolated features and report them. Are they used uniformly among all ensembles? Or does the PC based selection changes per ensemble? Could you also elaborate of the significance of the selected features as well (important for semi-physical

validation of the used features).

2. Please include pseudo-algorithm table for each resampling plan. This is very important for recreation proposes (the two utilized ensemble models, on the other hand, are well-studied in the broad literature and do not require detailed description; though I would prefer to see a mathematical description of the models to further acquaint the readers with them as ensemble learning is still not common across all fields of HESS).

3. Also, please elaborate more on the distinction between RUS and ROS (for a first glance, the wording makes RUS looks like a special case of ROS, but they are very different in reality and have distinct effect on the model performance). Please elaborate more on your choice of the ROS configuration, and why not present an array of results related to the OSR (ratio vs. performance for example).

4. Please modify the ensemble learning section to have a more concise summary of ensemble learning, diversity-in-learning concept and the effect of resampling as part of the latter. Please provide references from the pure literature.

5. It is important to note that the RWB model, contrast to what has been suggested in the applied literature, is NOT an ensemble model. Ensemble learning has three stages one of which is the resampling of the available intelligence. RWB violates this and should not be considered an ensemble for the sake of clarity. However, the randomization of weights within the individual learners are a major source of diversity, as shown in the literature, which promotes similar behaviour improvement in this model as other ensembles. RWB can be considered as middle ground between individual and ensemble learning. Please fix this issue.

6. I do not think that the authors replace the ensemble's native resampling technique with the suggested approaches, but rather add them as a preprocessing phase of the available data, as indicated in the manuscript (also, it would be impossible to replace the resampling approach in boosting!). Hence, it seems like there is "double resampling" which occurs in the modified ensemble models. this is interesting (as shown in

the obtained results). Please elaborate on the underlying effect of the ensemble's native resampling technique on the preprocessing one. For example, Bagging's uniform resampling plan has little effect, while that of Boosting has a very strong effect (I am thinking it nullifies the preprocessing resampler!).

7. Linking the previous comment with RWB, having a preprocessing resampler makes RWB a true ensemble model here. More importantly, the double-resampling effect is absent here, making it a great opportunity to elaborate on the difference between modified RWB and other ensembles! Please do so.

8. Figures 6 to 9 and Tables 5 and 6 are very important and provide most of the critical information to show how added resampling promotes improved high-flow prediction (and overall prediction also). In the same time, I sincerely think that a major result is missing here, which deals with quantifying the effect of the resampling approaches on the ensembles. I think that the paper requires a figure showing the change of ensemble performance versus ensemble size. This figure should at least depict the Bagging model and the modified Bagging model. I recommend adding the RWB and Modified RWB. This is very important to cross-examine the change of performance, per ensemble size, between the normal and modified ensembles and provide more insight on the effect of the preprocessing phase.

9. The fact that tables 5 and 6 show Boosting to perform the worst validates the results obtained, as Boosting performance deteriorates in the presence od "hard instances" in general and is more applicable to classification applications (at a larger ensemble size, the obtained combiners' weights seem to dilute in efficacy when performing regression. But in classification, they are powerful especially in binary classification due to the sign significance rather than magnitude of the weight). The authors attempt to explain the reasons to the deficiency of Boosting ensemble in the paper but I think they can elaborate more.

10. More importantly, the information in comments 3, 6 and 7 should be considered

here (when discussing tables 5 and 6). For example in table 5, when considering RWB and Bagging models, why was ROS based models the worst (I think because there are a lot of data), but in the same time SMOTER variations did provide good performance in few of the metrics. Why did SMOTED-RWB have an unusually low PITF but good performance with respect to other metrics? Please provide more details similar to this.

11. There are few typos and referencing issues. Please revise the manuscript for this. A few examples are provided below:

o Line 4: “compare three resampling;” I think it is missing a word! o Table 3: second column. Do you mean “variable” or “feature”? o Line 113: please include the term “individual learners”. o Line 163: comma is missing after the reference. o Line 164: the reference seems to be in the middle of the sentence. I noticed you do this often. Please try to minimize this to enable smother flow of the idea. o Line 168: “for handling imbalance data…” do you mean “imbalanced”. o Line 170: “that featuring each…”. Please fix. o Please italicize all symbols or feature names, such as N, theta, PI, CE, etc., across the manuscript. o Line 205: the reference format needs a fix. o Line 255 to 260: please re-write the ensemble learning summary as discussed in the major comment. o Line 262: “Ensembles members are..” please remove the “s”. o Line 310: do you mean “regional flood quantiles”?.

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