Snieder et al.
hess-2020-430
RC1 Supplement

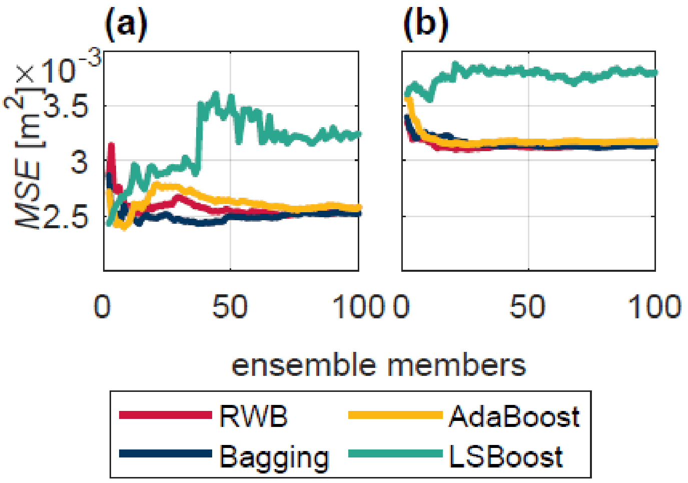

Figure 6. Test MSE across ensemble size for RWB (red), Bagging (blue), AdaBoost (yellow), and LSBoost (green) for the Don (left) and Bow River (right).

“Fig. 6 illustrates the change in test performance as the ensemble size increases from 2 to 100 for each river. This grid search is performed only for the base ensemble methods (RWB, Bagging, AdaBoost, and LSBoost) without any resampling. The Bow River plot indicates that AdaBoost and LSBoost tend to favour a small ensemble size (2-15 members), whereas the generalisation of RWB and Bagging improves with a larger size (>20 members). The performance of LSBoost rapidly deteriorates as the ensemble size grows, likely as the effects of overfitting become more pronounced. Similar results are obtained for the Don, except that RWB, Bagging, and AdaBoost all improve with larger ensemble size, while LSBoost does not offer competitive performance for any ensemble size. Again, a larger ensemble size (>20 members) produces favourable MSE.”

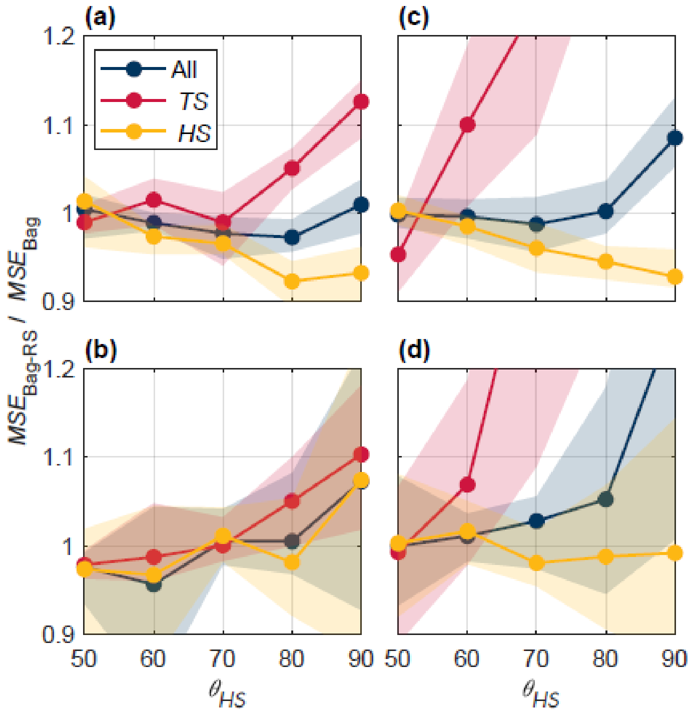

Figure 11. Calibration and test MSE ratio between Bagging and SMOTER-Bagging models for the Bow (a) and (c) and Don (b) and (d) Rivers across high stage threshold values ranging from 50% to 90%.

“As discussed in Sect. 2.1, a fixed threshold is used to distinguish between high and typical stages. Fig. 11 shows the effects of the fixed threshold increasing from the 50th to 90th percentile of the stage distribution. These plots show the relative effects of SMOTER-Bagging compared to simple Bagging. A performance ratio greater than 1 indicates that the SMOTER-Bagging model has greater error compared to the Bagging model, 1 indicates that they have the same performance, and less than 1, improved performance. The error (MSE) is presented for all stages as well as the TS and HS subsets. The calibration plots illustrate an asymmetric trade-off between HS and TS error. For a given θ_{HS} value, the error ratio of the TS subset increases more than the decline in HS error. More importantly, the improvements in HS performance obtained in calibration are considerably less pronounced in the test dataset, despite a loss in TS performance”.

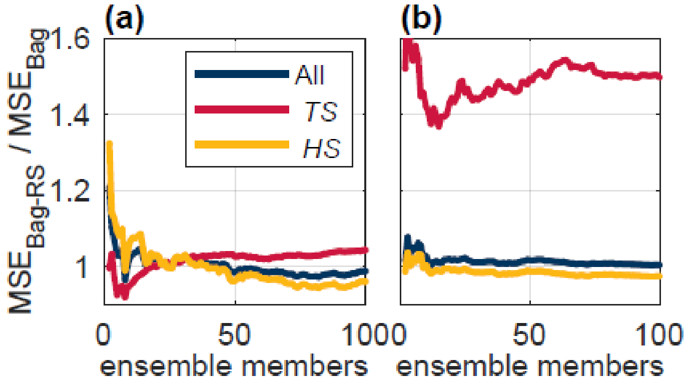

Figure 12. Test MSE ratio between Bagging and SMOTER-Bagging models for the Bow (a) and the Don (b) across ensemble size.

“Fig. 12 illustrates the effects of varying the ensemble size, thus, number of resampling repetitions, for the SMOTER-Bagging model, relative to the simple Bagging model. The plot shows the relative improvement in HS produced by the SMOTER resampling as the ensemble size increases, reaching a steady value at an ensemble size of approximately 70 for both models. This is larger than that required for the simple Bagging model to reach steady performance, indicating that SMOTER requires more resampling than simple resampling with replacement in order to reach stable performance. Consistent observations made from Fig. 11, an asymmetric trade-off between typical and high stage performance is noted.”

Appendix B: Pseudocode

Algorithm 1 Random undersampling

```
Require:
  Set S containing X input features and Y observations, $(x_1, y_1), ..., (x_m, y_m)$
  High stage threshold, $\theta$
  $S_{TS} = S$ where $Y < \phi_{TS}$
  $S_{HS} = S$ where $Y \geq \phi_{HS}$
  $S'_{TS} \leftarrow sample(S_{TS}, N_{HS})$
  $S'_{HS} \leftarrow sample(S_{HS}, N_{HS})$
  $S' = S'_{TS} \bigcup S'_{HS}$
```

Algorithm 2 Random oversampling

```
Require:
  Set S containing X input features and Y observations, $(x_1, y_1), ..., (x_m, y_m)$
  High stage threshold, $\theta$
  $S_{TS} = S$ where $Y < \phi_{TS}$
  $S_{HS} = S$ where $Y \geq \phi_{HS}$
  $S'_{TS} \leftarrow sample(S_{TS}, N_{TS})$
  $S'_{HS} \leftarrow sample(S_{HS}, N_{TS})$
  $S' = S'_{TS} \bigcup S'_{HS}$
```