This paper describes the use of random forests (ensembles of decision trees) to predict streamflow in various basins in the Pacific Northwest at one-day lead time. The authors use two methods to understand the most important predictor variables, and they also investigate the effect of other basin characteristics (not included as predictor variables) on the performance of the random forest. With some improvements this work could be a valuable contribution to the literature, especially given the analyses of predictor importance and confounding variables (basin characteristics not included as predictors). However, at this time I have chosen to reject, due to several major issues with the paper. Major comments are summarized below, and inline comments are attached in a PDF.

Major comments

1. The paper has serious grammatical issues, which make it difficult to follow. I have pointed all grammatical errors in the abstract (see pham2020_annotations_abstract.pdf). After the abstract, I have mostly abstained from pointing out every grammatical error. However, the frequency of grammatical errors is approximately the same throughout the paper. I would like to make it clear that I am not rejecting on the basis of grammar alone, but it does make the paper difficult to follow (there are many sentences that I simply do not understand).

2. The explanation of machine-learning methods (the random forest and associated predictor-importance methods) is unclear and contains several false statements. See inline comments for more detail. The explanation is probably not clear enough for readers unfamiliar with ML to follow, and it contains enough false statements that readers familiar with ML will probably be left scratching their heads.

3. No significance-testing. The authors claim that their random forest outperforms the two baseline models (persistence, which they call the “naïve” model, and linear regression), but no significance-testing is conducted to support this claim. Especially for the comparison of the random forest with linear regression, the numbers are close enough (see line 277) that I doubt the differences are statistically significant.

4. Interpretation of model performance is lacking in detail and contains several confusing statements. See inline comments on lines 274, 283, 287, 293, 295, 298, 302, 304, 308, 309, and 311.