Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-75-RC2, 2017

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Interactive comment on "Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting" *by* Omar Wani et al.

L. Raso (Referee)

l.raso@tudelft.nl

Received and published: 21 April 2017

General comments

The manuscript explores and discusses the application of k-Nearest Neighbors (kNN) method, a non-parametric machine learning technique, to estimate the predictive uncertainty in heteroschedastic streamflow forecasting.

The paper is clearly written. It comes completed of a internet website where a user-friendly interface makes application of kNN straightforward. The innovation is well framed in the recent literature on predictive uncertainty of heteroschedastic processes in hydrology, giving particular attention to comparable methods that estimates predic-

Printer-friendly version

Discussion paper

tive uncertainty a posteriori. The authors clearly present advantages and limits of kNN with respect to other methods. Nonetheless, as already mentioned by the other referee, the limits of kNN in extrapolation, presently mentioned only in the conclusion, should deserve more emphasis.

The manuscript brings a valid and innovative contribution to its field, and I suggest its acceptance. There are two issues, however, that could contribute to make the case for this methodology in a more convincing way, and some minor issues that deserve at least to be mentioned.

The first main issue regards the selection of the k value, i.e. the number of data points considered similar to the instance to be estimated. Fixing k is a problem of kNN method. In general, when kNN is used for prediction, k is selected in order to maximize the predictive capacity, tested by a cross-validation on data. In the manuscript the criteria for selecting k is the stabilization of residuals probability distribution. Change in residuals distribution is quantified by the cumulative difference, defined at Equation (17). The reason why the stabilization of residuals distribution is a good criteria for fixing k is not clear. Moreover, this value is monotonic, hence it does not offer a clear-cut rule. The authors propose that k is to be selected when shape changes, but this rule, differently from what stated ad page 6 line 8, is not fitted to be used in an optimisation procedure.

The second issue regards the estimation of quantiles. kNN use the closest k values to build up an empirical distribution made of situations (i.e. data-points) similar to the "true" distribution that one intends to estimate. When kNN is applied for regression, the value to be predicted is the expected value, then the algorithm takes the average of k nearest data-points. In the proposed application instead, the empirical distribution is used to estimate some quantiles. Quantile estimation, however, has a different convergence rule than the expected value, particularly critical in estimating tails. Convergence rules of empirical distribution at quantiles of interest is well described in [1], chapter 21. Error in quantile estimation decreases with root square of k, and it is larger for quan-

tiles close to 1 and 0. Using the 99th value from a set of 100 points as estimator of the 99th quantile may not be sufficient in guaranteeing sufficient convergency. Quantile estimation from empirical distribution introduce an error that must be be considered, or at least discussed.

I report here other **smaller issues**, worth to be mentioned in the manuscript.

In the discussion on verification index, the authors show that they are aware of the limits in using few indicators. The authors state that "PICP and MPI [...] give a reasonable assessment of performance". But this is not further explained. There are likely good reasons to select these indicators, but this should be better explained in the text, considering also that the application is about flood forecasting.

In Equation 7, variables are standardized one at a time, losing information about covariance. Why not considering variables as a multidimensional distribution, then using the covariance matrix to standardise? This would make use of the mutual information about variables in a more efficient way.

Other comments
Page 3, line 24: add "than" after simpler
Page 5, line 12: "uncertainty in observational data is not considered", why can not it be included?
Page 8 line 23: remove extra dot.
Page 9 line 17: the adverb "just" looks like non necessary.
Page 10 line 11: The result description would be easier to follow if the reference to the figure was placed at the beginning of this paragraph (from line 19 to line 11).

[1] Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press,

2000.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-75, 2017.

Printer-friendly version

Discussion paper