Hydrol. Earth Syst. Sci. Discuss., 6, C2707–C2710, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C2707/2009/


Hydrology and Earth System Sciences Discussions



# *Interactive comment on* "An experiment on the evolution of an ensemble of neural networks for streamflow forecasting" *by* M.-A. Boucher et al.

**Anonymous Referee #2**

Received and published: 24 November 2009

In this paper the authors use an ensemble of fifty neural networks for one day ahead stream flow forecasting. Each of the fifty networks initialized with different random weights is trained individually using a different bootstrap data set. The networks are evaluated for each epoch on test data set to form the ensemble of the predictions. Different probabilistic/ensemble forecasting verification statistics are also monitored during the training process on test data set. The experiments are done for four catchments. The work is basically sound from a technical perspective; however there are some issues that need to be addressed.

The construction of the ensemble members is not clear. There are 50 randomly initialized networks and 50 bootstrap data set. There are following three possibilities for



Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

ensemble: (a) training of the 50 networks (with different initial weights) on same data set (i.e. many to one) (b) training of the 50 networks on 50 bootstrap samples (i.e. one to one) (c) training of one networks on 50 bootstrap samples (i.e. one to many) The case (a) considers the uncertainty due to the initial network weights, while case (c) considers the uncertainty due to the training data. The case (b) considers partially both. The authors are advised to address it and at least mentioned what sources of uncertainty they are dealing with.

What is the main conclusion from the paper? Is it the main message that we need less than 10 iterations for ensemble predictions? Generally, as the performance of the networks increases, the uncertainty of the model prediction decreases. However, it seems that this is not the case in the paper as the authors mentioned that reliability of the confidence intervals decreases with the increase in ensemble mean. What is the reliability of the confidence interval? How do you measure it?

Reliability diagram is simply graph of the observed frequency of an event plotted against the forecast probability of an event. However in the text it is described differently. Regarding the terminology, it is better to use "width" of the intervals. In the figure 8, it is shown that the observed coverage probability for 90% confidence level at epoch 40 is around 40% which is much lower than the target 90%. However at 5 epochs the coverage probability is very close to desired 90% confidence level but with greater uncertainty (wider uncertainty bound). The question may arise then - which one is better? In epoch 5, I guess that individual networks is stopped early, thus producing wide spread outputs which is far from the observed one. However because of wider uncertainty bounds, the observed values are captured within these bounds. In epoch 40, individual member is better and produces narrow uncertainty bounds; however it fails to capture the observed one. This problem is mainly due to the fact that all sources of uncertainty are not considered properly. This is indeed multi-objective problem and one tries to minimize the width of the uncertainty bounds while maintaining the coverage probability close to the desired degree of confidence.



HESSD

6, C2707–C2710, 2009

Interactive Comment



Separation of data sets into training and test sets by SOM is not convincing; it may increase the complexity of the whole neural network model building process. It is also interesting to check if the both data sets are statistically similar. Authors are suggested to provide at least some statistical properties of both data sets. What are the input variables used in SOM? Furthermore because of random sampling of the data, these data are not contiguous in time. Then one would question the practical usefulness of such data (e.g., it is not possible to plot hydrograph of test data because of discontinuity).

Minor comments:

P 6267, L14 - Other popular ensemble techniques like boosting (e.g. Freund and Schapire, 1996, Shrestha and Solomatine, 2006) deserve to be highlighted in the present context.

P 6267, L30- "Kingston and Lambert, 2005" by Kingston et al., 2005

P 6271, L 23 and P 6285, Figure 2- 50 ensembles do not give the lowest value of the mean CRPS.

P 6276, L 5 - ".....scores calculated on the testing set". Use of test data set in evaluating the model during the training is unfair. Test data should not be used in any stage of model building process. However in the paper it is used to find the number of epochs, which could be done using cross validation data set.

P 6276, L 14-15- "After only five to ten iterations, all fifty MLPs mimic better the target data". This statement need to be justified by plotting error function of each MLPs against the epoch both in training and test data set.

P 6278, L 5-7- "When training continues, MLP forecasts gets even more similar. This leads to unreliable probabilistic distributions,.....". This is wrong statement. MLP forecasts are similar because MLPs are converging towards the best solution. In the experiments 50 MLPs are initialized with different random weights and hence consider only parameter (weight) uncertainty and partial input uncertainty (because of



HESSD

6, C2707–C2710, 2009

bootstrap samples). As commented before this is one of the reasons why most of the observed data falls outside the confidence intervals.

P 6283, Table 1 - It would be interesting to know catchment area. Leaf and Sanjuan catchments received almost equal daily precipitation, but the daily stream flow is much different. I think the daily stream flow data could be wrong in the latter.

P 6286- It is also interesting to see the daily CRPS value across the time for some epochs.

References

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In: L. Saitta (ed.), Proc. of 13th International Conference on Machine Learning, Bari, Italy, Morgan Kaufmann, San Francisco, CA, USA, pp. 148-156.

Shrestha, D. L. & Solomatine, D. P. (2006) Experiments with AdaBoost.RT, an improved boosting scheme for regression. Neural Computation, 18(7), 1678-1710.

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 6265, 2009.