Hydrol. Earth Syst. Sci. Discuss., 6, C2895–C2903, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C2895/2009/

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.

M.-A. Boucher et al.

marie-a.boucher.1@ulaval.ca

Received and published: 14 December 2009

We would like to thank the second reviewer for his very useful comments that will help us improving the manuscript. Our answers follow:

1.Construction of the ensembles and sources of uncertainties

It is true that the ensemble construction methodology does not account for all the sources of uncertainties in the hydrological process. Considering the options given by the referee, the ensemble used in the project described in the manuscript were obtained by training one network for each of the fifty bootstrapped dataset (option ÂńÂăbÂăÂż). Therefore, we partially consider both the uncertainty due to the train-

ing data and to the model.

This choice is supported by a small experiment that was done to study the outcome of different ensemble construction strategies. First, we produced five bootstrapped series and trained ten networks with each of those series. Second, we produced ten bootstrapped series and trained five networks for each series. We also repeated this process with a variant, where we allowed each observation (from the original training dataset) to appear exactly 50 times in the 50 new training series. This was also tested with the fifty bootstrapped series with one network trained with each series. Overall, we obtained five ensemble construction scenarios, in addition to the one presented in the paper (fifty bootstrapped series, each one used to train one network) and another scenario where fifty networks were trained using the original, not bootstrapped, training dataset. For each of the aforementioned scenarios, the fifty networks forming the ensemble were trained for fourty epochs, in the same fashion as in the experiment presented in the manuscript. For each epoch, the CRPS, MAE and logarithmic scores were computed, in order to assess if a gain could be made by adopting a certain construction pattern. The results, in terms of CRPS, MAE and logarithmic score values, were quite similar for each ensemble construction scenarios. We could not identify a scenario that would provide a more reliable ensemble or for which all three scores would be optimized at an earlier epoch than when using fifty bootstrapped series and one network per series.

Finally, this leads us to the interrogations regarding to the various sources of uncertainties and whether or not they are included in the ensemble. As pointed out by the referee, the underdispersion problem of the ensembles is "mainly due to the fact that all sources of uncertainty are not considered properly". However, our work is an exploration, not a final say on this issue.

2.About the main conclusion of the experiment

The main conclusion is not that one needs less than ten iterations for ensemble predic-

tions with neural networks. It is often the case that ensemble forecasts are underdispersed, whether they are issued with conceptual or physics-based hydrological models (e.g. Buizza et al 2005) or neural networks (e.g. Boucher et al 2009). In all instances, it is difficult to capture all sources of uncertainty perfectly. Also, as it is explained in the next section of this document, reliability of the ensemble must prevail over precision of the mean forecast, since one of the main interest of producing ensemble forecasts is to be able to estimate confidence intervals or to calculate probabilities of the streamflow being above (or under) a certain threshold value. On one hand, since the networks parameters are randomly initialized, it is expected that the ensemble is overdispersed at the beginning of the process. On the other hand, if each network is trained until convergence, the precision of the forecast is great but we are overconfident regarding the uncertainties (Boucher et al. 2009). We wanted to explore the learning process during which the ensemble goes from overdispersed to underdispersed and see if there is a possibility of compromising between reliability and sharpness of the distribution. That process is now documented.

3.Reliability of the confidence interval

If a confidence interval is reliable, it means that its nominal confidence level corresponds to its observed coverage. For instance, the 90% confidence interval for the forecasted streamflow value, should, on average, contain the observation 9 times out of 10. To verify whether or not the confidence intervals computed with the neural networks are reliable, we followed the process described at section 3.4.4 of the manuscript, leading to the reliability diagram, which precisely aims at assessing the reliability of the confidence interval. More precisely, the calculation of the limits for the various confidence intervals were obtained in a non-parametric fashion, using the sorted ensemble members to deduce the limits of the various confidence intervals. For instance, for the 90% interval, the lower bound was the mean of the second and third member (0.05*50 members) and the higher bound was the mean of the 47th and 48th members (0.95*50 members).

This definition of the reliability diagram is somewhat different than the usual one, as noted by the referee. However, the definition of the reliability diagram, as the “graph of the observed frequency of an even plotted against the forecast probability of an event”, is suitable only for probabilistic forecasts where the outcome can only take a limited number of specific values. Since streamflow is a continuous variable, we adapted the usual reliability diagram to better suit our needs. However, its interpretation remains the same, meaning that for a reliable forecasting system, the observed frequency must match the forecast probability, which is here represented in terms of how the nominal confidence level of various intervals is exact.

In addition, we chose to plot the mean effective probability of the intervals as a function of their length (or width, as suggested by the referee), but we could as well have plotted the mean effective probability as a function of the nominal confidence level. Instead, the nominal confidence level was explained in the text. The first dot is the 10% confidence interval, the second is 20%, and so on. The terminology “width” instead of “length” will be used in the subsequent version of the manuscript.

The reliability of the ensemble does decrease with the increase of individual network performance (as the training progresses), not “with the increase of the ensemble mean”. As is pointed out by the referee, there is clearly a trade off to be made between reliability and precision. To answer the question regarding “which one is better”, we adopt the point of view proposed by Gneiting and Raftery (2007) that we think can be extended to all types of ensemble forecasts. According to them, a good ensemble forecast “maximizes sharpness, subject to calibration”. While “sharpness” refers to the precision of the ensemble members, meaning that a in a sharp ensemble forecast all members are close to one another and centered around the observed value, “calibration” refers to reliability of the confidence interval. We subscribe to this view that the reliability (calibration) precedes precision (sharpness) in ensemble or probabilistic forecasting since the final goal is most often to express the forecast in terms of a probability, or to provide the user with a way of assessing the uncertainty on the next

outcome. Therefore, if the ensemble is not reliable, it is not useful and can even lead to wrong decisions. Of course, for many equally reliable ensembles, then the best one is the more precise (sharp).

4.Separation of the database by means of a Self-Organizing Map (SOM)

The separation of the database into statistically homogenous training and testing database is not a new process. The methodology used in our experiment is very similar to what was proposed in Abrahart and See (2000). Also, we have used this type of self organizing map successfully in previous study on the same catchments (Anctil and Lauzon 2004) and also in a water quality experiment on an agricultural watershed (Anctil et al. 2009). It is in fact very effective to ensure that the training and testing dataset are statistically equivalent, as can be seen from Table 1 and 2. Table 1 compares the statistical properties of two group of data (streamflow) obtained by splitting the chronologically ordered database in half. The values for the different statistics are similar for both halves of the database.

However, in table two, which shows the same statistics computed on the training and testing databases obtained through self-organizing maps, those values are even more similar.

With those four databases, we could have separated the database in two, keeping the chronological order intact. However, we consider that the use of a self organizing map for the purpose of data separation is a good practice that we systematically apply in our group. We could as well have had data for which there was a great statistical disparity between the first and the second half of the recording. As shown in Table 2, the use of the self organizing map is quite effective to solve this problem and ensure that the two datasets are equally representative of all range of events that can happen on the catchment. Also, the use of a self-organizing map is more critical in the context of model building, where results obtained with the training datasets can be compared to results obtained with the testing datasets, if both have similar statistical properties.

Finally, it can be useful to assess the model performance for specific classes of data (for instance the class regrouping the highest streamflows).

There is a drawback to this procedure, which is that it makes impossible the plotting of a continuous hydrograph of the training or testing dataset, since they are formed by data that is discontinuous in time. However, the performance of the model can still be assessed graphically, using scatter plots. Also, since the training and testing datasets are similar, they can be put back together in order to plot a hydrograph using all data. It doesn't mean however that the time correlation is evacuated from the problem. As this issue was raised by referee 1, please refer to the answer to referee 1 comments for more details about the time correlation is the series.

1.Minor Comments

We thank the referee for pointing out minor corrections such as spelling mistakes and bibliographical notices that will of course be corrected in the next version of the manuscript.

a)Improvement of the literature review

We will include more about other ensembling techniques with neural networks, including the two references suggested by the referee.

b)P. 6271, L23 and Figure 2

It is true that a 50-member ensemble does not provide the lowest mean CRPS value. However, this graph presents the results obtained after all networks have been trained to convergence. Since this experiment focuses more on the learning process than on the results obtained after the networks are fully trained, we believe that fifty members are sufficient to provide accurate results without being time consuming. The explanation on this page will be modified accordingly in the next version of the manuscript.

c)On the use of the testing dataset

The testing dataset was never used in the model building process. As mentioned at the end of section 3.3, “The networks’ parameters are stored for each epoch in order to be able to apply the partially trained network to the training dataset.” This means that the entire training process is done using exclusively the bootstrapped training datasets. Since all parameter for each network and each epoch are stored, it is possible, after the training is entirely completed, to apply partially trained networks to the testing dataset. This does not affect the model building process at all, since it is already complete and respects the wide spread “split sample” strategy as proposed by Klemeš (1986).

d)P.6278 L5-7

This sentence will be reformulated.

e)Plotting of the individual performances of the MLP as a function of the number of iterations

Figure 1 included with this answer shows the evolution of the errors of individual networks as a function of the number of training epochs. In order to simplify the plot, the results for only 10 neural networks are included in the figure, but all the other networks exhibit the same behavior. This figure will be added to a subsequent version of the manuscript.

f)Catchment areas

The catchments area will be added to Table 1. As for the daily streamflow and precipitation for Sanjuan, please see the answer to referee 1.

g)Daily CRPS value across time for some epochs

The CRPS for a probabilistic forecast behaves very similarly to the absolute error for a deterministic forecast. To our knowledge, one seldom plots the absolute error as a function of time, since the error made by a model on one particular time step is difficult to interpret on its own. More commonly, one looks at the mean absolute error, averaging over all time steps. The CRPS has the same interpretation. It may be higher

for some time steps than for others, but in the end, the goal is to minimize its average value over a certain period of time. For this reason, we prefer not to plot the graph of the CRPS across time.

References:

Boucher, M.-A., Perreault, L. and Anctil, F. 2009: Tools for the assessment of hydrological ensemble forecasts obtained by neural networks, Journal of Hydroinformatics, 11(3-4), 297-307.

Buizza, R, Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. 2005: A comparison of the ECMWF, MSC ans NCEP global ensemble prediction systems, Monthly Weather Review, 133, 1079-1097.

Abrahart, R.J., and See, L. 2000: Comparing neural network and autoregressive moving average technique for the prevision of continuous river flow forecasts in two contrasting catchments, Hydrological Processes, 14, 2157-2172.

Anctil F. and Lauzon, N. 2004: Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions, Hydrology and Earth System Sciences, 8(5), 940-958.

Anctil, F., Filion, M. and Tournebize, J. 2009: A neural network experiment on the simulation of daily nitrate-nitrogen and suspended sediment fluxes from a small agricultural catchment, Ecological Modeling, 220 (6), 879-887

Klemeš, V. (1986): Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31(1), 13-24.

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

Table 1: Statistical properties of the first and second half of the databases, in chronological order.

First Half						
	Mean	Standard deviation	minimum	maximum	kurtosis	skewness
Leaf	1,24	2,77	0,07	58,25	84,52	7,24
Sanjuan	7,52	11,88	0,17	143,45	25,41	3,94
Serein	0,54	0,78	0	6,86	14,74	2,91
Le Golo	2,21	2,26	0,31	19,44	12,53	2,57
Second Half						
	Mean	Standard deviation	minimum	maximum	kurtosis	skewness
Leaf	1,5	3,02	0,1	64,02	75,49	6,52
Sanjuan	6,66	11,66	0,13	128,41	30,7	4,46
Serein	0,68	0,94	0	8,95	13,17	2,72
Le Golo	2,58	2,55	0,27	19,99	10,79	2,4

Table 2: Statistical properties of the training and testing databases obtained using a self-organizing map, as in the manuscript.

Training						
	Mean	Standard deviation	minimum	maximum	kurtosis	skewness
Leaf	1,36	2,74	0,07	64,03	88,57	6,77
Sanjuan	7,16	11,79	0,13	143,45	27,98	4,2
Serein	0.61	0,9	0	8,95	17,06	3,15
Le Golo	2,41	2,4	0,27	20	10,92	2,37
Testing						
	Mean	Standard deviation	minimum	maximum	kurtosis	skewness
Leaf	1,35	2,92	0,07	58,25	91,33	7,31
Sanjuan	7,19	11,95	0,14	128,26	27,6	4,2
Serein	0,61	0,86	0	8,25	13,44	2,76
Le Golo	2,31	2,36	0,3	19,44	12,92	2,65

Fig. 1.