Hydrol. Earth Syst. Sci. Discuss., 11, C1994–C2002, 2014
www.hydrol-earth-syst-sci-discuss.net/11/C1994/2014/

Hydrology and Earth System Sciences
Discussions

Interactive comment on "The skill of seasonal ensemble low flow forecasts for four different hydrological models" *by* M. C. Demirel et al.

Anonymous Referee #2

Received and published: 23 June 2014

The article presents the application of two conceptual hydrological models and two models based on artificial neural networks (ANN) for seasonal low-flow forecasting on the Moselle River basin. This study is a follow-up of a study published last year by the same authors in WRR, in which they had focused on the medium range forecast using the same conceptual models on the same basin.

Here the authors use seasonal ensemble meteorological forecasts to test the models and combine various forecast inputs to evaluate model sensitivity. The authors conclude that the most complex conceptual model and one ANN model perform best.

Although the article raises an interesting topic, I found it difficult to follow and I am unsure whether the conclusions are general enough to be much helpful for other mod-

ellers. I found that several aspects should be improved:

- Section 3, which presents the methodological aspects, is not detailed enough. Hence it is sometimes difficult to fully understand what was done by the authors. This section should be improved.

- The test results are either presented on two specific years (2002 and 2003) of the test period, or using criteria calculated on the full period (2002-2005). The authors draw conclusions from all these results, without discussing to which extent the results presented on the two specific years are general or not. Therefore it is difficult to evaluate the generality of the conclusions proposed here.

- The authors introduce a new objective function and a new evaluation criterion, but do not provide any justification of the added value of these criteria compared to existing ones. Since there is already plethora of criteria in the literature, the authors should demonstrate why they found necessary to introduce these new ones.

- While the tests were made on a quite large basin, results are shown only for a single gauging station. It would be useful that the authors include results on other gauging stations (e.g. nested catchments within the Moselle basin, or basins elsewhere), with catchments possibly showing different geological settings, to evaluate whether the same conclusions are drawn. This would give more general conclusions. I provide detailed comments below. I advise major revision before the article can be reconsidered for publication in HESS.

Specific comments

1. Title: The title is too general. At least it should be mentioned that models are tested on the Moselle basin only.

2. Abstract: The abstract should be modified in light of the modifications made by the authors to answer the comments above and below.

3. Section 2.1: Can this basin be considered as natural? If there are influences, this

could be mentioned as it may influence the evaluation of simulated low flows.

4. Section 2.2.1: As mentioned in my major comments, I think it would be useful to test the models on a set of gauging stations, not a single one. This would make conclusions more general and more useful for practitioners.

5. Section 2.2.1: It seems that a groundwater indicator (G) is used by ANN-I. What are the corresponding data used to compute this indicator?

6. Section 2.2.2: What is the control members compared to the other members? It is said that forecasts are available with a 184-day lead time, but then forecasts are made only up to a 90-day lead time. Is there a reason for this difference? How often the seasonal meteorological forecast is issued within the year? Every day? Every first day of each month? Other? If not every day, what is considered as seasonal forecasts for the other days when making the modelling tests?

7. Section 2.2.2: Nothing is said on the quality of the seasonal P and PET forecasts? Did the authors calculate some skills on these ensembles? This may help better discussing the results, by distinguishing possible sources of errors. It could also be said whether the P and PET forecasts are joint (i.e. member i for P correspond to member i for PET) or are independent.

8. Section 3.1: This section is very important to fully understand what the authors did, but I think it should be more detailed and clarified (see comments below).

9. Section 3.1.1: The authors detail the parameters here for GR4J but not for the other model (section 3.1.2). This makes the presentation a bit unbalanced. The authors could refer to Table 7 instead.

10. Section 3.1.1 and 3.1.2: The authors could shortly explain how models are updated to make the article more self-contained (one sentence is given later in section 3.1.4 but it refers to another article).

11. Section 3.1.3, p. 5386: I must say that I did not fully understand how the ANN-E

and ANN-I models were built. The authors should better explain how the models work, using more precise notations (for example Q(t) instead of Q) and better distinguishing between observed inputs up to the day of forecast t and inputs over the forecasting horizon (t+1 to t+90). For example, for the ANN-E model, is the Q forecast at t+j used as input to the model to make the forecast at t+j+1? For ANN-I, what is G? For ANN-I, it is mentioned that the model uses historical Q (do you mean Q observed at the day of issuing the forecast?), but how is it done in practice since Q does not appear as a model input in Fig. 1? For ANN-I, can we say that this model assumes no P and PET in the future, or alternatively, that it intends to make a forecast only based on previous conditions? In the second case, does it mean that this model assumes that the catchment has at least a 90-day memory of antecedent conditions?

12. Section 3.1.4: What is the warm-up period used in calibration and validation periods? Comments on model updates should be placed in section 3.1.1 and/or 3.1.2.

13. Section 3.1.4: The objective function aggregates mean absolute error on low flows (MAE_low) and MAE on inverse low flows. It means that almost no weight is given to intermediate or high flows. Although this focus on low flows may appear logical for a study on low flows, it neglects the fact that low flows are the results of flow recession: if high or intermediate flows are not well simulated, this may have an impact on the simulation of low flows. Besides, since most of the annual water volume generally flows during floods, this may limit the good identification of parameters responsible for the water balance. Could the authors discuss this point and better justify their choice? Besides, could the authors better explain why it was deem useful to aggregate these two MAE criteria? Are not they redundant to some extent, since they seem to similarly focus on low flows? Why is it so important to introduce this new objective function here? Was it found much better than other objective functions more classically used for studies on low flows? Please also explain how the value of epsilon was chosen. Moreover, although this may not be important numerically, I found not really correct to write an equation (Eq. 4) that deliberately neglects homogeneity in units. Last, it

is unclear whether the models were optimized for a forecasting objective (i.e. computing the errors between Qfor(t+90) produced by updated models and Qobs(t+90)) or for a simulation objective (i.e. computing the errors between Qsim(t) produced by models without update and Qobs(t)). In the second case, I would not understand how calibration is performed for ANN. What is "sim 113"?

14. Section 3.1.5: I did not fully understand how the forecasting tests were made. Over the 2002-2005 period, was the 90-day ahead forecast made for each day of the period? Why the authors did not include a reference deterministic forecast in which the models would be run with the a posteriori observed meteorological inputs as forecasts? This would help distinguishing the role of modelling uncertainty from input uncertainty.

15. Section 3.2.3: Missing end in the last sentence of the paragraph.

16. Section 3.2.4: There is a wide range of existing criteria to evaluate deterministic and/or probabilistic forecasts. Why introducing a new score is necessary here? The authors should explain why the existing scores do not answer their question and whether this new score has expected statistical properties.

17. Section 4.1: Adding neurons in the hidden layer does not only not improve the performance but even strongly degrade it. Do the authors have any explanation for this strong decrease?

18. Section 4.1: What the authors mean by "GR4J, HBV and ANN-I are also calibrated accordingly"?

19. Section 4.1, last paragraph: I do not see obvious reason why the better performance of HBV in validation should be explained by its higher complexity. Although this may be the case in calibration, since more complex models have more degrees of freedom, more complex models may also be less robust and therefore less performing in validation.

20. Section 4.2: This section is based on the analysis of two specific years, which

were "carefully selected". Although illustrations are always useful to better understand model behaviour, I think it is quite dangerous to try to draw general conclusions from such specific cases as was done here by the authors. I also found that some of the conclusions drawn from the visual analysis (note that it is not clear on which ground one forecast is said better than the other) presented in section 4.2. and those given in section 4.3 (based on criteria) appear contradictory. For example, from the analysis of Fig. 3, it seems that the behaviour of GR4J and HBV are quite similar, and that ANN-E is poor in case of extreme low flows (year 2003). However, from the analysis on criteria, it seems that HBV and ANN-E are very close, and that GR4J is comparatively poor. Why is it so? Are the other years very specific, with different model behaviours, which could explain this difference? In that case, why the years 2002 and 2003 were chosen if they are not representative of the actual model behaviour? There are also some of the comments in this section that are not so clear when looking at the illustrations (p. 5392, l. 14-15; l. 17-18; p. 5393, l. 13).

21. Section 4.2, p. 5393, l. 2-4: Do the authors have any explanation for this behaviour? This period is the closer to the preceding winter high flows. Maybe high flows are poorly simulated which may impact the forecasts for these first months of low flows (see comment above on the objective function).

22. Section 4.2: The ANN-E model seems to have an erratic behaviour in Fig. 4b (shifting between two values). How this can be explained? Is not that a bit difficult to use such a model in an operational context?

23. Section 4.2, p. 5394, l. 1: Probably this is obtained only by chance.

24. Section 4.2, p. 5394, l. 7: Why results are not shown? Please include them, e.g. in Fig. 5.

25. Section 4.2, p. 5394, l. 11-13: This conclusion is two strong based on a single result shown here.

26. Section 4.3, p. 5394, l. 19: I did not fully understand which tests evaluated the sensitivity to initial conditions here. Maybe the authors should introduce tests in which models are not updated to make forecasts, to better evaluate the importance of "recalibrating" the model on observed values at the day of forecast issue.

27. Section 4.3, p. 5394, l. 24-26: I did not understand how the shape of the curves can be explained. Could the authors detail this a bit? The limit does not seem to be 20 days for all models.

28. Section 4.3, p. 5395, first paragraph: Although the authors seem to find some skill to the ANN-I model, it is basically useless operationally since it is unable to forecast any threshold crossing and it is worse than climatology. Therefore, I don't understand why the authors find reasons in their conclusions to encourage the use of this model.

29. Section 5: The authors could also further discuss why models in forecasting mode seem to be differently impacted by uncertainty in meteorological forecasts. For example, the ANN-E model seemed much poorer than the two conceptual models in validation, but is judged to perform as well as the best conceptual model in forecasting mode. Conversely, the most simple conceptual model appears much poorer than ANN-E whereas it was better in the validation test.

30. Section 5, p. 5396, l. 14-26: Again, I found the usefulness of the ANN-I very limited in practice.

31. Section 6: As mentioned above, this section mixes conclusions apparently based on the visual inspection of two specific years and conclusions based on criteria calculated on the full period. This creates some unbalance, as the first group of conclusions may not be as general as the others. A more general evaluation should be sought to draw general conclusions. Besides, the conclusion on the ANN-I model should be more balanced. The apparently good model behaviour in low flow conditions is probably due to the fact that low flows are very slowly varying on this catchment. If other catchments with more dynamical low flows had been used (see major comment above), the

conclusions may have been a bit different.

32. Section 6, p. 5397, l. 9-10 and l. 13-14: Again, why introducing these criteria was so necessary?

33. Table 1: Maybe the ranges of annual Q, P and PET could be added. What about G?

34. Table 3: For ANN-I, why Q is not detailed in the "temporal resolution" comlumn? I did not fully understand what the lag is used for.

35. Table 4: The number of members for P and PET could be added in each case. On which period is the climate mean calculated?

36. Table 6: This table could probably be improved, by providing the examples on a separate table.

37. Table 7: CFLUX is calibrated at its maximum value (1.0), which means that the model probably would prefer a larger value. Would the optimized value (and performance) be different if the upper bound had been set at a larger value?

38. Fig. 1: I found the graphs of ANN models confusing. For ANN-E, the use of a store is a bit misleading as there is no actual internal state in the model. The authors should try to distinguish between observed (past) and forecast (future) values.

39. Figs. 3-4: The authors should explain why there are gaps in the series (observed values above threshold?). Indicate on the graph the name of the model on each line. A horizontal line could indicate the low flow threshold. The graph for ANN-I is unclear: it is very difficult to distinguish between observed and forecast (use different colours and/or symbols).

40. Fig. 5: The authors could use a presentation similar to Figs. 3-4.

41. Fig. 6: Why the hit rate first increases for the ANN-E model.