Reviewer #2

The authors thank Reviewer#2 for her/his constructive and elaborated comments on the manuscript. We agree with most of the points of view she /he expressed and we explain how we will modify the text to account for her/his comments.

Comment 1) 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.

Reply from authors: We agree with the comment. The Reviewer #3 highlighted similar points for this section. Based on the comments from both reviewers, we will include more details about the models and their state update procedure.

Comment 2) 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.

Reply from authors: We referred to the figures when presenting the results at each paragraph. The figures include the period (year) information. However, we will review the results again and be clear about which result apply to which period or year in the revised of the manuscript.
An example from the conclusion part is shown below. The green part will be added in the revised version of the manuscript.

"Based on the results of the comparison of different model inputs for two years i.e. 2002 and 2003, the largest range for 90 day low flow forecasts is found for the GR4J model when using ensemble seasonal meteorological forecasts as input."

Comment 3) 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.

Reply from authors: We partly agree with the comment. There are well-known objective functions (e.g. Nash-Sutcliffe, Kling-Gupta, RMSE) and skill scores (e.g. Brier Skill Score) in

the hydrology literature. Most of the objective functions target the mean discharge values and can be sensitive to the high discharge values. The proposed objective function in this study is a combination of one very strict and one less strict low flow oriented objective functions i.e. MAE_low and MAE_inverse respectively. The inverse or log transforms are commonly used approaches to suppress the high flows (See Table 3 in Ref-1 below). However, calculating the performance in only low flow period is new and the added value of this study.
Regarding the new skill score i.e. MFS, it is solely focusing on the low flow forecast performance and ignoring the correct negatives. The main advantage of this skill score is the simplicity of the calculation compared to the Brier Skill Score. This will avoid any mistakes in the calculations. Moreover, the hydrology literature can definitely benefit from this simple and effective skill score to evaluate the probabilistic ensemble forecasts and simulations.

Ref -1: Pushpalatha, R., Perrin, C., Le Moine, N., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171–182, doi:10.1016/j.jhydrol.2011.11.055, 2012.

Ref -2: Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol., 411, 66–76, 2011.

Comment 4) Title: The title is too general. At least it should be mentioned that models are tested on the Moselle basin only.

Reply from authors: The methods described in this paper can be applied to any (European) river. We think that it is not necessary to change the title as it may affect the scientific visibility of the article.

Comment 5) Abstract: The abstract should be modified in light of the modifications made by the authors to answer the comments above and below.

Reply from authors: We agree with the comment. We will revise the abstract based on the revisions in the text.

Comment 6) 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.

Reply from authors: The River Rhine, in general, has been heavily canalised for river navigation and flood prevention. There are many dams in the upstream of the River in Switzerland. However, the human influence in Moselle River is assumed to be negligible, therefore, not mentioned in the text.

Comment 7) 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.

Reply from authors: We agree with the comment. However, it is outside the scope of this study.

Comment 8) 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?

Reply from authors: Groundwater levels from numerous stations in the Rhine basin were included in this study. The individual groundwater stations' measurements, shown in figure 3 at Ref-1 below, were aggregated to the scale of seven sub-basins using standardised data.

For details it should be referred to:
Ref-1 Demirel, M. C., Booij, M. J. and Hoekstra, A. Y. (2013), Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times. Hydrol. Process., 27: 2742–2758. doi: 10.1002/hyp.9402

Comment 9) 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?

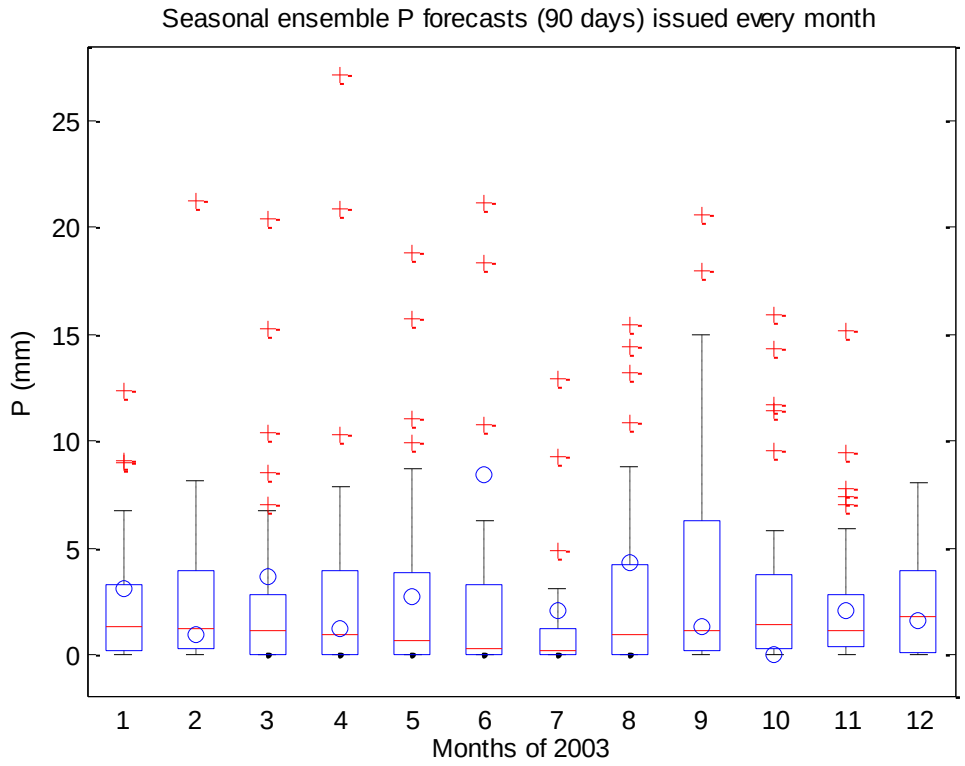
Reply from authors: The control member is the unperturbed forecast. The forecast lead time of 90 days is assumed to be appropriate for seasonal scale as the utility of the forecasts for more than three months lead time is highly questionable. Moreover, the major river users i.e. river navigation and energy sector can benefit from 90 days low flow forecasts (see Ref-1 below). We used daily meteorological forecast data issued every first day of each month for a lead time of 184 days (Ref - 2).

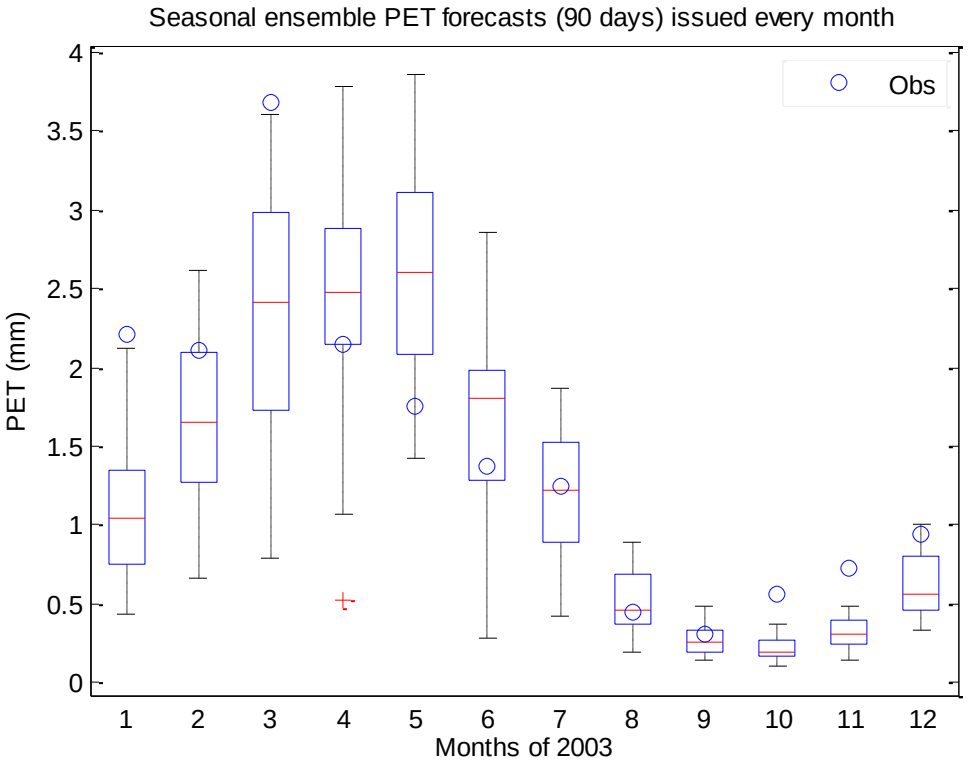
Ref-1: http://hepex.irstea.fr/colloquium-seasonal-forecasting-current-challenges-and-potential-benefits-for-decision-making-in-the-water-sector/

Ref -2: http://old.ecmwf.int/publications/manuals/mars/

Comment 10) 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.

Reply from authors: We agree with the comment. The skills of 90 day a head P and PET forecasts issued in each month are summarized in two box plots below. The range of ensemble members and outliers are shown in these figures. The basin averaged observed P and PET are also presented (blue circles). These figures are definitely useful to distinguish possible sources of errors as they indicate the large range and the significant number of members with high precipitation amount. For brevity we plan not to include these figures to the manuscript. However, they will be used in elaboration of the discussion part of the revised manuscript. Regarding the last part of the reviewer comment, P and PET forecasts are joint in our modelling practice. For example, if the first ensemble member is called from P then the first member from PET is also called to force the hydrological model.


Seasonal ensemble P forecasts (90 days) issued every month
P (mm)
0
5
10
15
20
25
1
2
3
4
5
6
7
8
9
10
11
12
Months of 2003


Seasonal ensemble PET forecasts (90 days) issued every month
Obs
PET (mm)
0
0.5
1
1.5
2
2.5
3
3.5
4
1
2
3
4
5
6
7
8
9
10
11
12
Months of 2003

Comment 11) 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).

Reply from authors: We will improve this section especially to clarify the forecast scheme and model state update procedure.

Comment 12) 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.

Reply from authors: The model parameters are indeed detailed in Table 7. We will revise the GR4J model part as indicated by the reviewer.

Comment 13) 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).

Reply from authors: As we mentioned in Reply #11, we will include a section where we explain the state update procedure used in this study.

Comment 14) 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?

Reply from authors: We agree with the comment. We will use the precise time notation (t) to describe the modelling scheme. The ANN-E model works as conceptual models. Therefore, P(t), PET(t) and Q(t-1) are used to simulate discharge at the same day i.e. Q(t). The Q(t-1) represents the model state from previous day.

Question: is the Q forecast at t+j used as input to the model to make the forecast at t+j+1?

Answer: *Yes*

Question: For ANN-I, it is mentioned that the model uses historical Q (do you mean Q observed at the day of issuing the forecast?),

Answer: *No, Q observed is not used in the model. There has been a significant typo in Table 3, second column (see below for the corrected second column for ANN-I). Probably this caused the major confusion. The ANN-I model uses different historical inputs (low flow indicators P, PET and G). The G is the groundwater levels at appropriate lags and temporal resolution shown in Table 3, columns 3-5.*

ANN-I	P: Observed PET: Observed G: Observed	110-day mean P 180-day mean PET 90-day mean G	P: 0 PET: 210 G: 210	Daily	90

Question: 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?
Answer: *Yes*

Question: In the second case, does it mean that this model assumes that the catchment has at least a 90-day memory of antecedent conditions?
Answer: *Yes, the correlation analysis in our previous study (Ref-1) revealed significant correlations between low flow indicators and Q for a lead time of 90 days.*

Ref-1: Demirel, M. C., Booij, M. J. and Hoekstra, A. Y. (2013), Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times. Hydrol. Process., 27: 2742–2758. doi: 10.1002/hyp.9402

Comment 15) 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.

Reply from authors: A period of three years was used as warm-up period in calibration and validation periods. We will include the model state update procedure in a separate sub-section.

Comment 16) 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"?

Reply from authors: We elaborated a similar question in Comment #3. We will include a paragraph in the revised version of the manuscript where we justify the selection of the hybrid objective function more clearly. It should be noted that we didn't fully neglect the high and intermediate flows using MAE_inverse metric, whereas only low flow period is considered in MAE_low. This is one of the advantages of using MAE_hybrid metric which also avoids the redundancy. The explanation for introduction of a new objective function was also given in Comment #3.

Following Pushpalatha et al (2012), the value of epsilon was set at one hundredth of the mean flow (see Ref-1 below).

Ideally it is always better (and correct) to keep the unit homogeneity/consistency in an equation. However, the two components of MAE_hybrid were not normalised as the different units had no effect on the calibration results. The ANN-E model was calibrated for simulation objective and the ANN-I model was calibrated for forecasting objective. As the reviewer indicated, it wouldn't make much sense to calibrate the ANN-I model for the simulation objective.

sim 113: ~ 113 (sim is used in Latex typesetting to indicate "approximately"). This typo will be corrected in the revised version of the manuscript.

Ref -1: Pushpalatha, R., Perrin, C., Le Moine, N., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171–182, doi:10.1016/j.jhydrol.2011.11.055, 2012.

Comment 17) 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.

Reply from authors: The daily meteorological forecast data issued every month for a lead time of 184 days. The two figures presented in Comment #10 will be used to distinguish between possible sources of error. However, assessing the model uncertainty is out of scope of this study.

Comment 18) Section 3.2.3: Missing end in the last sentence of the paragraph.

Reply from authors: There should be reference to the Table 5 at the end of the sentence. We will include it in the revised version of the manuscript.

Comment 19) 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.

Reply from authors: Please refer to the reply for comment 3.

Comment 20) 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?

Reply from authors: We partly agree with the comment as the degradation of MAE_hybrid was around ~0.5mm after adding the second hidden neuron. Over-fitting can be the main reason for the degradation (Ref - 3).

Ref-3: Shamseldin, A. Y.: Application of a neural network technique to rainfall-runoff modelling, J. Hydrol., 199, 272-294, 10.1016/s0022-1694(96)03330-6, 1997.

Comment 21) Section 4.1: What the authors mean by “GR4J, HBV and ANN-I are also calibrated accordingly”?

Reply from authors: We will remove the confusing word "accordingly" form the sentence. Only calibration of the ANN-I model was based on the selected number of the hidden neurons for the ANN-E model.

Comment 22) 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.

Reply from authors: We partly agree with the comment. The calibrated and sophisticated hydrological model is assumed to learn the basin behaviour at different weather conditions better than the simple models. Moreover, the sophisticated model is expected to repeat this successful behaviour for a different period if the input quality remains the same.

Comment 23) 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).

Reply from authors: We partly agree with the comment. The Figure 3b shows the results for one year, whereas Figure 6 shows the results for the 2002-2005 period. As mentioned in the manuscript, the two years were selected based on their characteristics i.e. wet and dry.

Comment 24) 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).

Reply from authors: We agree with the comment. The poor performance of the models during the spring period can be explained by the high precipitation amount fall in this period. The poor simulation of high flows can have effect on the forecasts too.

Comment 25) 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?

Reply from authors: The two hydrological models used in this study have well defined surface and ground water components. Therefore, they react to the weather inputs in a physically meaningful way. However, in black box models, the step functions (transfer functions or activation functions) may affect the model behaviour. The ANN model will then fire (i.e. react) to a certain range of inputs based on the objective function. This feature of ANN is the main reason for the erratic behaviour in Figure 4b and small (and uniform) uncertainty range in the figures (e.g. Figure 3).

Comment 26) Section 4.2, p. 5394, l. 1: Probably this is obtained only by chance.

Reply from authors: We present the results for 2002 below. The GR4J model still performs better than other two models.

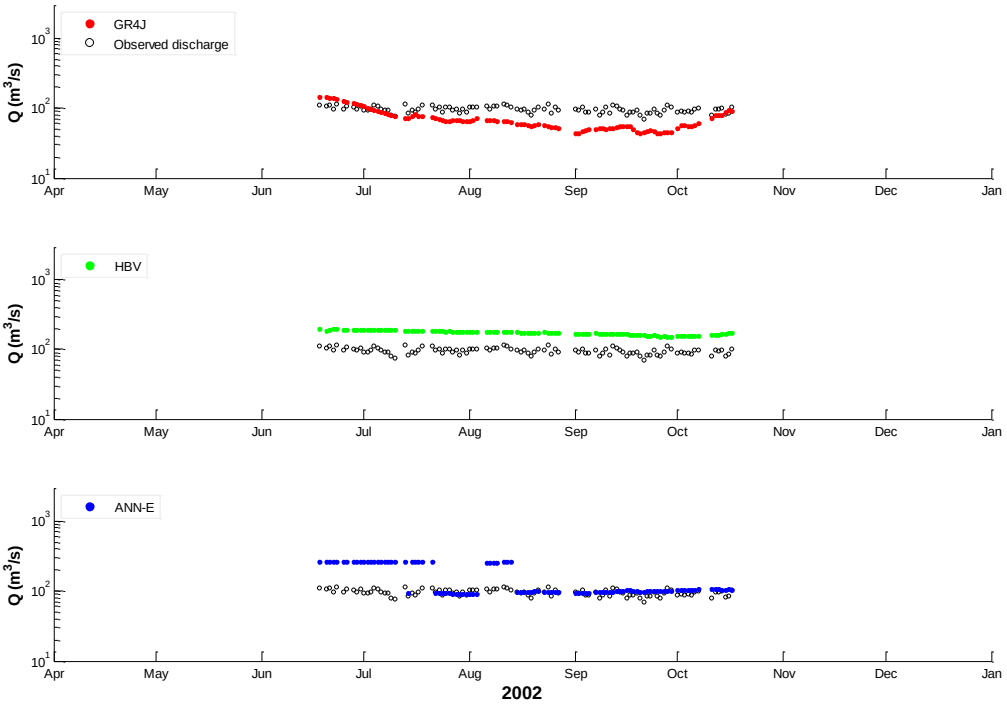


Figure: Low flow forecasts in 2002 for a lead time of 90 days using both climate mean P and PET as input for GR4J, HBV and ANN-E models (case 4).

Comment 27) Section 4.2, p. 5394, l. 7: Why results are not shown? Please include them, e.g. in Fig. 5.

Reply from authors: We have removed the figure for brevity of the results after the suggestion of the associated editor. We will include the figure in the revised version of the manuscript.

Comment 28) Section 4.2, p. 5394, l. 11-13: This conclusion is two strong based on a single result shown here.

Reply from authors: We agree with the comment. We will remove the text in the revised version of the manuscript.

~~"Interestingly, the results of ANN-E are relatively better than the other two conceptual models showing the ability of partly data-driven models for seasonal low flow forecasts."~~

Comment 29) 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.

Reply from authors: We agree with the comment but the main focus of the study is assessing the effect of the model inputs.

Comment 30) 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.

Reply from authors: We agree with the comment. The limit is around 20 days for ANN-E and shorter for other models. When the forecast is issued as day (t), the model states are updated using the observed discharge on that day (t) using the deterministic state update procedure proposed in our previous paper (Ref-1 below). However, the models probably spin-up after some days and improve the results for false alarm rate.

Ref-1: Demirel, M. C., M. J. Booij, and A. Y. Hoekstra (2013), Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49,4035–4053, doi:10.1002/wrcr.20294.

Comment 31) 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.

Reply from authors: We agree with the comment. We have already mentioned the utility of ANN-I in Conclusions as shown below. We will remove ANN-I from the text in the revised version of the manuscript.

"The skill score results of ANN-I may seem contradictory, but they show that ANN-I is useless to predict whether a low flow (as defined, below a threshold) will occur or not. For that purpose, one of the other three models will be required."

Comment 32) 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.

Reply from authors: We partly agree with the comment. The calibration, validation and test periods are all different periods. The results are based on the test period runs using ensemble forecasts from ECMWF, whereas the observed meteorological inputs are used for the calibration and validation.

Comment 33) Section 5, p. 5396, l. 14-26: Again, I found the usefulness of the ANN-I very limited in practice.

Reply from authors: We agree with the comment. We will remove ANN-I in the revised version of the manuscript.

Comment 34) 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.

Reply from authors: The entire test period 2002-2005 is used for the Figures 6 and 7. We will review our conclusions and clarify the difference between the two groups of figures (i.e. Figures 3-4-5 and Figures 6-7) in the revised version of the manuscript.

Comment 35) Section 6, p. 5397, l. 9-10 and l. 13-14: Again, why introducing these criteria was so necessary?

Reply from authors: Please refer to comment #3.

Comment 36) Table 1: Maybe the ranges of annual Q, P and PET could be added. What about G?

Reply from authors: We agree with the comment. We will include the ranges to the Table 1.

Comment 37) 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.

Reply from authors: The Q should not be in Table 3. It is a significant typo as explained in Comment #14. Moreover, the ANN-I model will be removed from the revised manuscript.

Comment 38) Table 4: The number of members for P and PET could be added in each case. On which period is the climate mean calculated?

Reply from authors: All available historical data (1951-2006) were used to estimate the climate mean. For example the climate mean for January 1st is estimated by the average of 55 January 1st values in the available period (1951-2006).

Comment 39) Table 6: This table could probably be improved, by providing the examples on a separate table.

Reply from authors: We will separate the example and present in a different table.

Comment 40) 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?

Reply from authors: The upper and lower limits are selected based on previous works (Booij, 2005; Eberle, 2005; Perrin et al, 2003; Pushpalatha, 2011; Tian et al, 2013).

Booij, M. J. (2005), Impact of climate change on river flooding assessed with different spatial model resolutions, *J. Hydrol.*, **303**(1–4),176–198.

Pushpalatha, R., C. Perrin, N. L. Moine, T. Mathevet, and V. Andréassian (2011), A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, *J. Hydrol.*, **411**(1–2), 66–76.

Tian, Y., M. J. Booij, and Y. P. Xu (2012), Uncertainty in high and low flows due to model structure and parameter errors, *Stochastic Environmental Research and Risk Assessment*, doi: 10.1007/s00477-013-0751-9.

Pushpalatha, R., C. Perrin, N. L. Moine, and V. Andréassian (2012), A review of efficiency criteria suitable for evaluating low-flow simulations, *J. Hydrol.*, **420–421**, 171–182, doi:10.1016/j.jhydrol.2011.11.055.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, **279**(1–4), 275–289.

Comment 41) 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.

Reply from authors: We agree with the comment. We will improve the figure 1.

Comment 42) 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).

Reply from authors: We agree with comment. We will revise the figures as the reviewer indicated. The figures 3, 4 and 5 show only low flows and the high flow periods are censored i.e. shown as gap.

Comment 43) Fig. 5: The authors could use a presentation similar to Figs. 3-4.

Reply from authors: We will revise the Figure 5 to make it similar to Figure 3 and 4.

Comment 44) Fig. 6: Why the hit rate first increases for the ANN-E model.

Reply from authors: Please refer to Comment #30 for explanation.

APPENDIX-A: Table 3

Table 3 Model descriptions. PET is potential evapotranspiration, P is precipitation, G is groundwater, Q is discharge and *t* is the time (day).

Model Type		Input	Temporal resolution of input	Lag between forecast issue day and final day of temporal averaging (days)	Model time step	Model lead time (days)
Conceptual	**Data-driven**					
GR4J		P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET	P: 0 PET: 0 Q: 1	Daily	1 to 90
HBV		P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET	P: 0 PET: 0 Q: 1	Daily	1 to 90
ANN-E		P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET Daily Q	P: 0 PET: 0 Q: 1	Daily	1 to 90
	ANN-I	P: Observed PET: Observed G: Observed	110-day mean P 180-day mean PET 90-day mean G	P: 0 PET: 210 G: 210	Daily	90