Colour legend

□ Remarks for the editor

□ Inserted/revised text in the manuscript

Reviewer #1 (Prof Renata Romanowicz)

The authors thank Dr Renata Romanowicz for her constructive comments on the manuscript. We agree with most of the points of view she expressed and we explain how we will modify the text to account for her comments.

Comment 1) From the practitioner's point of view it would be useful if the authors presented a table with minimum and maximum prediction errors from the ensemble for each model for both examined time-periods (years 2002 and 2003).

Reply from authors: We will include a table in the revised version of the manuscript that will show the minimum and maximum prediction errors of the models.

*(**Table 8** has been added to section 4.2)*

Comment 2) The authors have set themselves a difficult task in comparing the models with different input variables and different basic assumptions. Unfortunately, the comparison does not come out sufficiently clearly. It would help if the authors did not include the ANN- I model in their comparison. It uses different inputs and obscures the message the authors want to put across.

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

*(**ANN-I** has been removed from the manuscript.)*

Comment 3) Lines 225-229: The authors distinguish between daily P and PET data and historical Q as an input. "The first model, i.e. ANN-E, requires daily P, PET and historical Q as input. Historical Q from the previous day is used to update the model states (Table 3). This is a one day memory which also exists in the conceptual models, i.e. GR4J and HBV (Figure 1). The ANN-E is assumed to be comparable with the conceptual models with similar model structures. The second model, ANN-I, uses historical Q to update initial model conditions and three low flow indicators, i.e. P, PET and G, as model input."
Does this mean: observed flow Q up to the date when the forecast is issued?

Reply from authors: We mean the observed discharge value on the forecast issue day i.e. Q(t) to forecast Q(t+1) using the two conceptual models and ANN-E, and Q(t+90) using the ANN-I model.

The sentence below will be revised for clarity.

"Historical Q from the previous day is used to update the model states (Table 3)"

Revised version:

"Observed discharge on the forecast issue day is used to update the model states (Table 3)"

(The green shaded text has been added to section 3.1.3.)

Comment 4) Line 459 Case 5: zero P and ensemble PET forecasts as input for the other three models (GR4J, HBV and ANN-E). – the figure should be shown for completeness of the discussion.

Reply from authors: We will include the figure for case 5 in the revised version of the manuscript.

(Figure 5 has been updated by including case 5.)

Comment 5) Line 477: "The decrease in false alarm rates after a lead time of 20 days shows the importance of initial condition uncertainty for short lead time forecasts. For longer lead times the error is better handled by the models." It is not clear to me how the initial conditions can affect the false alarm rate. It rather seems that the "correct negatives" are increasing in number and may be this particular indicator is not working properly for forecasts longer than 20 days? Please explain that statement in more detail.

Reply from authors: The initial conditions (i.e. model states) are important components of a reliable low flow forecast. Therefore, an estimation error in the model states can affect particularly the short term forecasts as the model improves after the spin up period. We will elaborate the discussion about Figure 6 in the revised version of the manuscript.

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 6) 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. 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 procedures.

*(A new section **(3.1.5)** has been added to describe the model storage update procedure in more detail.)*

Comment 7) 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 in each paragraph. The figure captions include the period (year) information. However, we will review the results again and will be clear about which result applies to which period or year in the revised version 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."

(The green shaded text has been added to the Conclusion section)

Comment 8) 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 hydrological literature. Most of the objective functions target mean discharge values and can be sensitive to 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 function, 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 Pushpalatha et al., (2012) below). However, calculating the performance only in the low flow period is new and the added value of this study.

The new skill score, i.e. MFS, solely focuses on the low flow forecast performance and ignores 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 hydrological literature can definitely benefit from this simple and effective skill score to evaluate probabilistic ensemble forecasts and simulations.

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 9) 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 10) 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.

(The abstract has been revised based on the changes in the manuscript content)

Comment 11) 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 part of the River Rhine in Switzerland. However, the human influence on the Moselle River is assumed to be negligible, and therefore, not mentioned in the text.

Comment 12) 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, this is outside the scope of this study.

Comment 13) 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 of Demirel et al. (2013) below, were aggregated to the scale of seven sub-basins using standardised data.

For more details please see:
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 14) 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 the seasonal scale as the utility of forecasts for

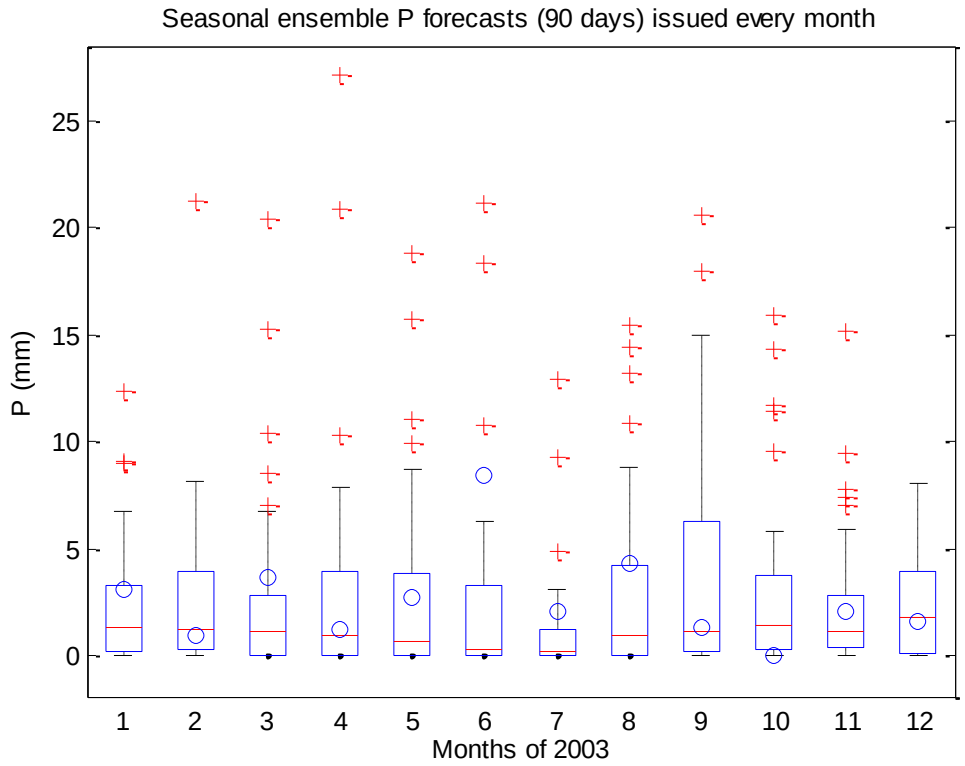
more than three months lead time is highly questionable. Moreover, the major river users, i.e. river navigation and the energy sector, can benefit from 90 days low flow forecasts (see HEPEX 2014 below). We used daily meteorological forecast data issued every first day of each month (ECMWF 2014).

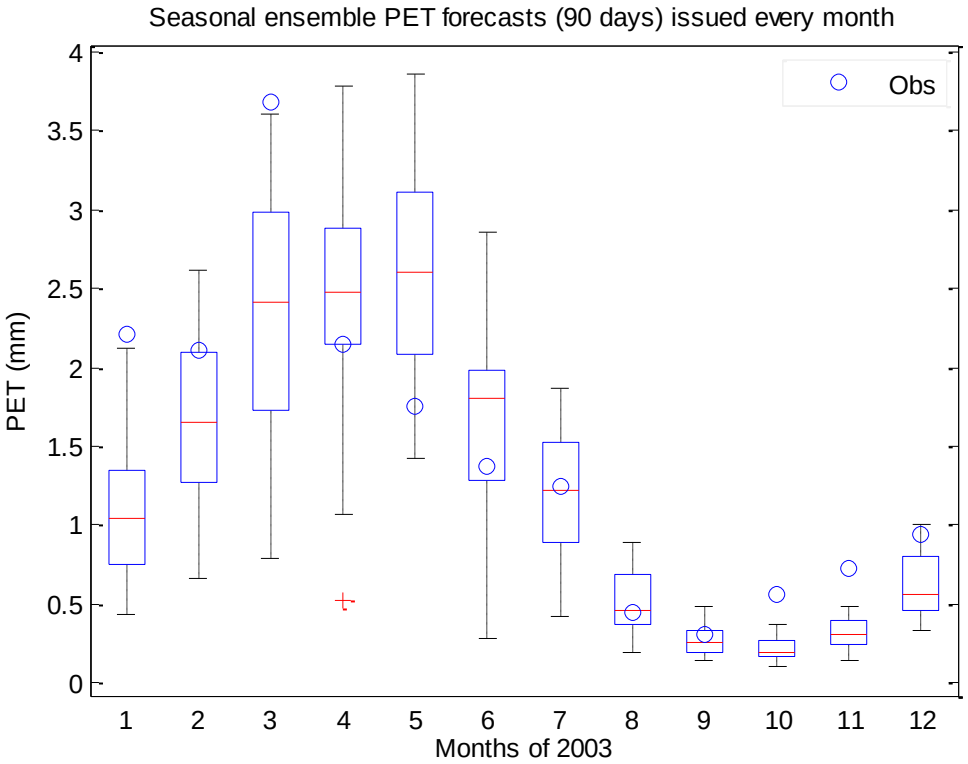
HEPEX, 2014: http://hepex.irstea.fr/colloquium-seasonal-forecasting-current-challenges-and-potential-benefits-for-decision-making-in-the-water-sector/

ECMWF, 2014: http://old.ecmwf.int/publications/manuals/mars/

Comment 15) 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 ahead 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 amounts. For brevity we plan not to include these figures into the manuscript. However, they will be used in the elaboration of the discussion part of the revised manuscript. Regarding the last part of the reviewer comment, P and PET forecasts are joint forecasts 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 16) 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.

*(A new section (**3.1.5**) has been added to describe the model storage update procedure in more detail.)*

Comment 17) 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.

(The GR4J model section has been revised and Table 7 is mentioned for parameter details.)

Comment 18) 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 #6 and #16, we will include a section where we explain the state update procedure used in this study.

*(A new section (**3.1.5**) has been added to describe the model storage update procedure in more detail.)*

Comment 19) 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) in the description of the modelling scheme. The ANN-E model works similarly as the conceptual

models. Therefore, P(t), PET(t) and Q(t-1) are used to simulate the discharge at the same day i.e. Q(t). The Q(t-1) represents the model state on the previous day.
(Figure 1 is revised accordingly)

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). G is the groundwater level at an appropriate lag 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

(Table3 has been revised as indicated above)

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 (Demirel et al., 2013) revealed significant correlations between low flow indicators and Q for a lead time of 90 days.*

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 20) 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 the calibration and validation periods. We will include the model state update procedure in a separate sub-section.

"The first three years are used as warm-up period for the hydrological model."

(The green shaded text above has been added to section 3.1.4)

Comment 21) 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 responded to a similar question in Comment #8. 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 periods are considered in MAE_low. This is one of the advantages of using the MAE_hybrid metric and also avoids redundancy. The explanation for the introduction of a new objective function was also given in Comment #8.

(The green shaded text above has been added to section 3.1.4)

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

Ideally it is always better (and correct) to use homogeneous units 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 the simulation objective and

the ANN-I model was calibrated for the forecasting objective. As the reviewer indicated, it wouldn't make much sense to calibrate the ANN-I model for the simulation objective.

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

(The typo has been corrected in the revised manuscript.)

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 22) 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 are issued every month for a lead time of 184 days. The two figures presented in Comment #15 will be used to distinguish between possible sources of error. However, assessing the model uncertainty is outside the scope of this study.

Comment 23) Section 3.2.3: Missing end in the last sentence of the paragraph.

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

(Corrected)

Comment 24) 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 #8.

Comment 25) 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 (–see also Shamseldin, 1997).

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 26) 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" from the sentence. Only calibration of the ANN-I model was based on the selected number of the hidden neurons for the ANN-E model.

(The referred sentence has been corrected and the section has been revised after removing ANN-I)

Comment 27) 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 more sophisticated hydrological model is assumed to better represent the basin behaviour for different weather conditions than the simpler models. Moreover, the sophisticated model is expected to repeat this successful behaviour for a different period in a better way if the input data quality remains the same.

Comment 28) 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. 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 29) 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 in this period. The poor simulation of high flows in the preceding winter months can have an effect on the forecasts too.

(The green shaded text has been added to the section 4.2)

Comment 30) 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 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 the small (and uniform) uncertainty range in the figures (e.g. Figure 3).

(The green shaded text has been added to section 4.2)

Comment 31) Section 4.2, p. 5394, l. 1: Probably this is obtained only by chance.

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

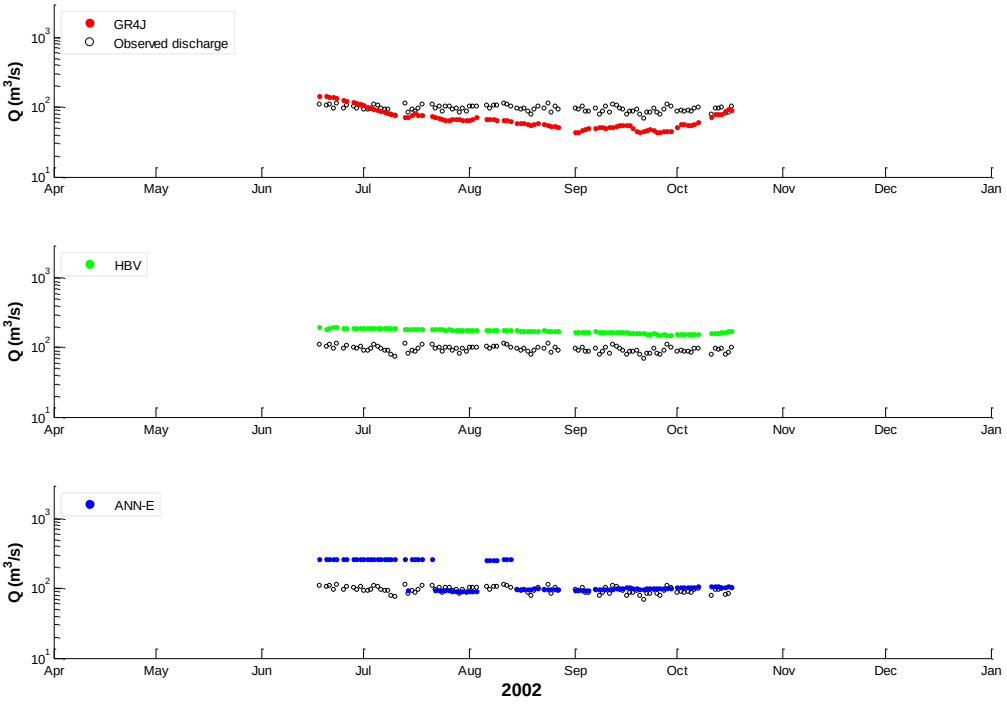


Figure: Low flow forecasts in 2002 for a lead time of 90 days using both climate mean P and PET as input for GR4J, HBV and ANN-E models (case 4).

Comment 32) 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 reasons of brevity after the suggestion of the associate editor. We will include the figure in the revised version of the manuscript.

(We included the figure for zero P, case 5, as Figure 5b)

Comment 33) 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."~~

(We removed the text above)

Comment 34) 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 on assessing the effect of the model inputs.

Comment 35) 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 the other two models. When the forecast is issued on day (t), the model states are updated using the observed discharge on that day (t) and using the deterministic state update procedure described in section 3.1.5. However, the models probably spin-up after some days and improve the results for false alarm rate are improved.

(The green shaded text has been added to section 4.3)

Comment 36) 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 the 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."

(The ANN-I model has been removed)

Comment 37) 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 38) 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.

(The ANN-I model has been removed)

Comment 39) 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 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.

(Section 6 has been partly revised)

Comment 40) 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 #8.

Comment 41) 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 in Table 1.

(We added the Q, P and PET ranges in Table 1)

Comment 42) 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 #19. Moreover, the ANN-I model will be removed from the revised manuscript.

(Table 3 has been revised since the ANN-I model is removed)

Comment 43) 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).

(The green shaded text has been added to the section 3.1.6 just above Table 4)

Comment 44) 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 45) 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 have been selected based on studies 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 (2014), Uncertainty in high and low flows due to model structure and parameter errors, *Stochastic Environmental Research and Risk Assessment*, 28, 319-332, 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 46) 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 Figure 1.

(Figure 1 has been revised based on the comment above)

Comment 47) 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.

(The caption of Figure 3 caption has been revised as below to explain the gaps in the figures)

Figure 3 Range (shown as grey shade) of low flow forecasts in **a)** 2002 (the wettest year of the test period with 101 low flow days) **b)** 2003 (the driest year of the test period with 192 low flow days) for a lead time of 90 days using ensemble P and PET as input for GR4J, HBV and ANN-E models (case 1 – 2002 and 2003). The gaps in the figures indicate non-low flow days (i.e. censored).

Comment 48) Fig. 5: The authors could use a presentation similar to Figs. 3-4.

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

(The figure was revised accordingly. Now it also shows the zero P case (i.e. case 5)

Comment 49) Fig. 6: Why the hit rate first increases for the ANN-E model.

Reply from authors: Please refer to Comment #35 for explanation.

Reviewer #3 (Stefanie Jörg-Hess)

The article addresses an interesting topic an highlights the sensitivity of different model structures to the representation of the meteorological input. The purpose of the work and the conclusions are well elaborated. The structure of the article is clear and technically sound. The Figures and Tables are well selected with room for improvement in helping the reader follow the analysis. The article would benefit from clarifying and deepening some parts.

The authors thank Stefanie Jörg-Hess for her constructive comments on the manuscript. We agree with most of the points of view she expressed and we explain how we will modify the text to take her comments into account.

Comment 50) Introduction: You spend a lot of time in introducing runoff forecasts with climate indicators and forecasted meteorological variables and present studies on different rivers. For me this is not really relevant for the following article. I would rather prefer to read more about ensemble predictions and the effect of ensembles and historic data on the runoff predictions. You could already introduce here the difference of the conceptual and the artificial neural network (ANN) models.

Reply from authors: We will revise the introduction and include recent ensemble low flow prediction literature like below. Moreover, the difference between different model types will be emphasized.

(The two references below have been added to the text)

1) *Madadgar, S., Moradkhani, H., 2013: A Bayesian Framework for Probabilistic Seasonal Drought Forecasting. J. Hydrometeor, 14, 1685–1705. doi: http://dx.doi.org/10.1175/JHM-D-13-010.1*

2) *Dutra, E., et al. (2014). "Global meteorological drought – Part 2: Seasonal forecasts." Hydrology and Earth System Sciences 18(7): 2669-2678.*

Comment 51) Methodology: Especially section 3.1.3 and 3.1.4 are difficult to follow and would benefit from some more details.

Reply from authors: We agree with the comment. Reviewer #2 highlighted similar points for these sections. Based on the comments from both reviewers, we will include more details about the models and their state update procedures.

*(A new section (**3.1.5**) has been added to describe the model storage update procedure in more detail.)*

Comment 52) Results: It would be interesting to see the effect of the ensembles also on the skill scores. I would suggest to validate the skill scores for 2 or 3 cases and add the skill scores to the figures in section 4.2.

Reply from authors: The skill scores are probabilistic scores based on the forecast ensemble. This is the main difference between deterministic and probabilistic evaluations of the forecasts. In short, the effect of ensemble members on the skill scores is implicitly shown in the figures.

Comment 53) Abstract: P 5378 L21: For avoid confusion I suggest to change 'over-predict low flows' to 'over-predict runoff during low-flow events'.

Reply from authors: We will rephrase the sentence as below.

" From the results, it appears that all models are prone to over-predict runoff during low flow periods using ensemble seasonal meteorological forcing."

(The referred sentence has been revised.)

Comment 54) P 5380 L16: On the previous page the work by Wang et al. (2011) is cited in the context of statistical approaches and here the work is cited in the context of dynamic approaches. Please clarify this.

Reply from authors: We agree with the comment. It is a typo and the year of the reference at page 5379 should be 2006, i.e. Wang et al. 2006. We will revise the reference.

(Corrected.)

Comment 55) Section 2.1: Some more information of the catchment would be helpful. What are the characteristics of the catchment and the dominant runoff processes? Further a Figure of the catchment with the distribution of the stations would be interesting. It was not clear to

me how many stations are used to estimate P and PET in the sub-basins and what is the size of the sub-basins.

Reply from authors: We agree with the comment. We will include more details about the Moselle basin and estimation of the basin averaged P and PET in the revised version of the manuscript.

(We improved the study area section and presented annual ranges for P, PET and Q in Table 1)

Comment 56) Section 2.2.1: Please mention 'h' from the Table in the Text.

Reply from authors: We agree with the comment. We will revise the text and indicate that the mean altitude of the catchments (h) have been obtained from the German Federal Institute of Hydrology (BfG) in Koblenz, Germany.

(Corrected)

Comment 57) Section 2.2.2: The ensemble forecast is available for 184-days. For your evaluation you are using the first 90 days. Is there a reason why you stop the evaluation after 90 days?

Reply from authors: The forecast lead time of 90 days is assumed to be appropriate for the 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 the energy sector, can benefit from 90 days low flow forecasts (see HEPEX 2014 below).

HEPEX, 2014. http://hepex.irstea.fr/colloquium-seasonal-forecasting-current-challenges-and-potential-benefits-for-decision-making-in-the-water-sector/, accessed on 8/10/2014

Comment 58) P 5383 L13: The Link of the reference ECMWF (2012) has been changed to http://old.ecmwf.int/publications/newsletters/pdf/133.pdf. In this newsletter I could not find any information about the MARS system 3. Please state in section 2.2.2 whether you are using daily or weekly meteorological forecast data.

Reply from authors: We agree with the comment. We will update the link and provide other references for the MARS 3 system (see Ref-2 below) in the revised version of the manuscript. We used daily meteorological forecast data issued every month for a lead time of 184 days.

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

(Link updated)

Comment 59) P 5386 L15: Here you could describe in one sentence what is the characteristic of the global approach.

Reply from authors: We agree with the comment. We will mention about the aim of the global optimisation algorithms in the revised version of the manuscript.

(*The green shaded text is added to the manuscript.*)

We used a global approach (i.e. Genetic Algorithm) to avoid local minima (De Vos and Rientjes, 2008) and tested the performance of the networks with one, two and three hidden neurons corresponding to a number of parameters (i.e. number of weights and biases) of 6, 11 and 16 respectively.

Comment 60) Section 3.1.2.: You could refer to Table 3, when you describe the model structure.

Reply from authors: We agree with the comment. We will include a reference to Table 3 in the revised version of the manuscript.

(*Revised as indicated above)*

Comment 61) Section 3.1.3: I do not fully understand the concept of ANN models. For me some more background information would be helpful. For example it is not clear to me what is the main difference between ANN-E and ANN-I. Or what are the n inputs? For me the inputs are P and PET and Q. But from equation (1) and Table 7 it seems that you use four inputs.

*Reply from authors: Apparently Table 3 is not clear to Reviewers #2 and #3. We also noticed a typo in the ANN-I part as "*Q: State update*" is relevant only for ANN-E and the two conceptual models.*

It should be noted that we will remove ANN-I from the revised version of the manuscript. We will also revise Table 3 and 7 accordingly. The number of weights is 4 for both ANN-I and ANN-E models: 3 weights connecting the input layer to the hidden layer and 1 weight

connecting the hidden layer to the output layer (see Table 7, revised). ANN-E and ANN-I have 3 inputs. ANN-E has P, PET and Q as inputs, whereas ANN-I has P, PET and G as inputs.

The revised version of the manuscript will have a clear presentation based on the reviewer comments.

(ANN-I has been removed from the manuscript and Table 3 and 7 are revised accordingly)

Comment 62) Section 3.1.4: This section needs some clarification. Please explain the meaning of the numbers (population size, reproduction elite count size, etc) and how you selected these numbers. For the calculation of observed low-flow days the Q75 of the simulation is used. How do you account for systematic biases of the model by using this threshold for observations?

Reply from authors: We agree with the comment. We will give more details about the genetic algorithm (GA) and selection of the GA parameters. Moreover, we selected the GA parameters based on the hydrological literature. The Q75 is based on the observed discharge instead of the simulated discharge as mentioned by the reviewer. However, the systematic biases of the model structure were not assessed as the main focus of this study was the input uncertainty.

(The green shaded text below has been added to section 3.1.4)

The evolution starts from the population of 100 randomly generated individuals. The population in each iteration is called a generation and the fitness of every individual in the population is evaluated using the objective function. The best 70 percent of the population (indicated as cross over fraction) survives in the process of 2000 iterations.

Comment 63) Section 3.1.5: How is the climate mean of the ensembles defined? For which period? How many members are used? Is it calculated with a moving window?

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).

(The green shaded text above has been added to the section 3.1.6)

Comment 64) P 5389 L4: Does ‘N’ (equation 6) and ‘n’ (equation 3) both refer to the total number of days? If yes, please be consistent.

Reply from authors: We agree with the comment. We will correct the notation for consistency in the revised version of the manuscript.

(Corrected)

Comment 65) P 5389 L17: You describe non-exceedance probabilities for medium to high flows. Please change this accordingly.

Reply from authors: We agree with the comment. We will replace the word "non-exceedence" with "exceedence" in the revised version of the manuscript.

(Corrected.)

Comment 66) P 5390 L17: You begin the sentence with 'These probabilities...'. For me it is not clear which probabilities. Please specify this.

Reply from authors: We agree with the comment. We will replace the confusing word "these" with "the". Table 6 is used to explain the details of the calculation. The probability of a deterministic forecast can be 0 or 1, whereas it varies from 0 to 1 for ensemble members. For example, if 22 members from an ensemble of 39 members are successful forecasts then the probability becomes 22/39.

(The green shaded text has been added to section 3.2.4)

Comment 67) Section 3.2.4: Please add some information what is the meaning of this score.

Reply from authors: We will include more information in the referenced section in the revised version of the manuscript.

(The green shaded text above has been added to section 3.2.4)

Comment 68) Section 4.1: The Figure shows that MAE is lowest for 1 hidden neuron. Did other studies find similar results concerning the optimal number of hidden neurons?

Reply from authors: There are many studies using one hidden layer for hydrological predictions (de Vos and Rientjes, 2005;Shamseldin, 1997;Yuan et al., 2003;Maier and Dandy, 2000). However, to our knowledge, this study is the only study that applies an ANN model with one hidden neuron to the seasonal low flow forecast problem.

Comment 69) Section 4.2: There is a large uncertainty of the predicted runoff with the first three models. For most low-flow events the most ensembles overestimate the runoff. Can you

explain why the spread in the conceptual models is larger than with the ANN model? Do you have an explanation why the runoff is over-predicted? I do not see your statement that the GR4J and HBV over-predict low flow after August. For me all models over-predict low-flows during the entire period of the two years. From the two years chosen my expression is that the conceptual models perform best during fall and the performance is lowest during spring. Do you have any explanation for this? In this context it would be interesting to see some scores for the forecasts (e.g. Brier skill score). Do you have an idea why the low flow in spring 2003 are not captured in the models? May be the simulation of the snow cover during winter can explain this behaviour.

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 limit model sensitivity after the training. The ANN model will then react to a certain range of inputs based on the objective function. This feature of an ANN is the main reason for the small (and uniform) uncertainty range in the figures (e.g. Figure 3). The over prediction of the models is closely related to the over prediction of the P by the ensembles. We agree with the reviewer that low flows are usually over predicted by the models for the entire period. However, there are under-predictions of low flows for some days in November-December as well. Before June, none of the low flows are captured by the ensemble members. As the reviewer indicated, the best performing period is the fall and the worst performing period is the spring period for the models. We will include skills scores in the figures. The poor performance of the models during the spring period can be explained by the high precipitation amount in this period. Since the objective function used in this study solely focuses on low flows, the high flow period is less important in the calibration. The low flows occurring in the spring period are, therefore, missed in the forecasts. The simulation of snow cover during winter and snow melt during the spring can both have effects on the forecasts too.

(The green shaded text below has been added to Discussion section)

Comment 70) P 5393 L12: State that the uncertainty range is larger in Figure 4a than in Figure 3b for the conceptual models.

Reply from authors: We will state this result in the revised version of the manuscript.

(The green shaded text below has been added to section 4.2 above the new Table 8.)

It is also obvious that the uncertainty range is larger in case 1 than in case 2 for the conceptual models. This is also what we see in Figure 3 and Figure 4 above.

Comment 71) P 5393 L 19-24: Do you have any explanation why the low-flows are not captured in Figure 4b. Please explain why the spread of the runoff forecast is narrow in this case.

Reply from authors: The precipitation information is crucial for the conceptual models to forecast low flows for a lead time of 90 days. The narrow uncertainty band indicates that the effect of PET ensemble on the forecasts is less pronounced as compared to the effect of P ensemble.

(The green shaded text has been added to section 4.2.)

Comment 72) Figure 3: Please enlarge the points of the observation, specify the points in the caption and label the plots with the according model. I would appreciate it if you could apart from the visual validation, add minimum one of the scores to Figures 3-5. Is there a reason to put different grey-scales for the ensemble forecast with the different models?

Reply from authors: We will increase the visibility of the observations by using bold and filled circles. We will include a skill score for the probabilistic forecasts. The different shades of grey were arbitrarily selected to indicate different models.

(A new Table 8 has been added to show the error ranges for each case in addition to the Figures)

Comment 73) Figure 6: It would be interesting to see a validation of these scores with ongoing lead timealso for other cases (e.g. case 2).

Reply from authors: We plotted the skill scores for the cases 2 and 3 below. The figures show the clear importance of the ensemble P input for the conceptual models, the HBV model in particular.

It should be noted that we plan not to include these figures in the revised version of the manuscript for reasons of brevity. Since the review reports are also public and shown together with the papers, the readers may read and refer to these figures below.

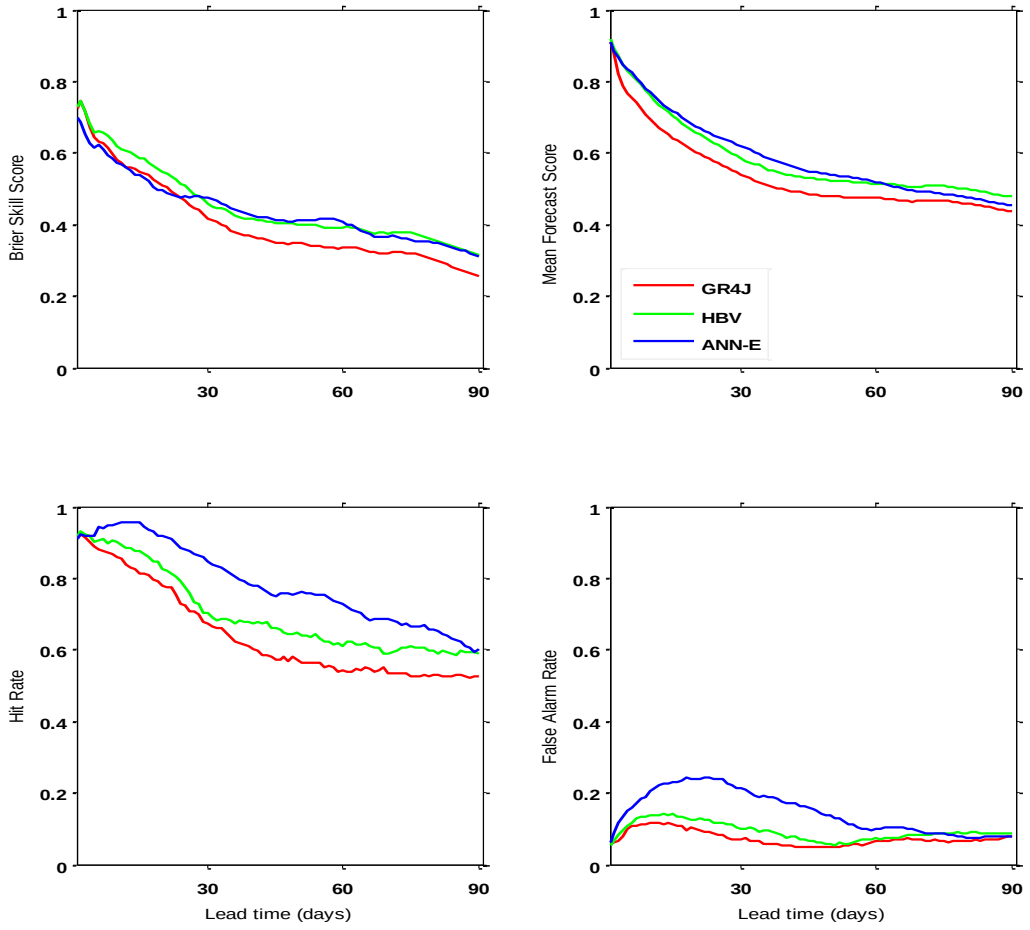


Figure: Skill scores for case-2

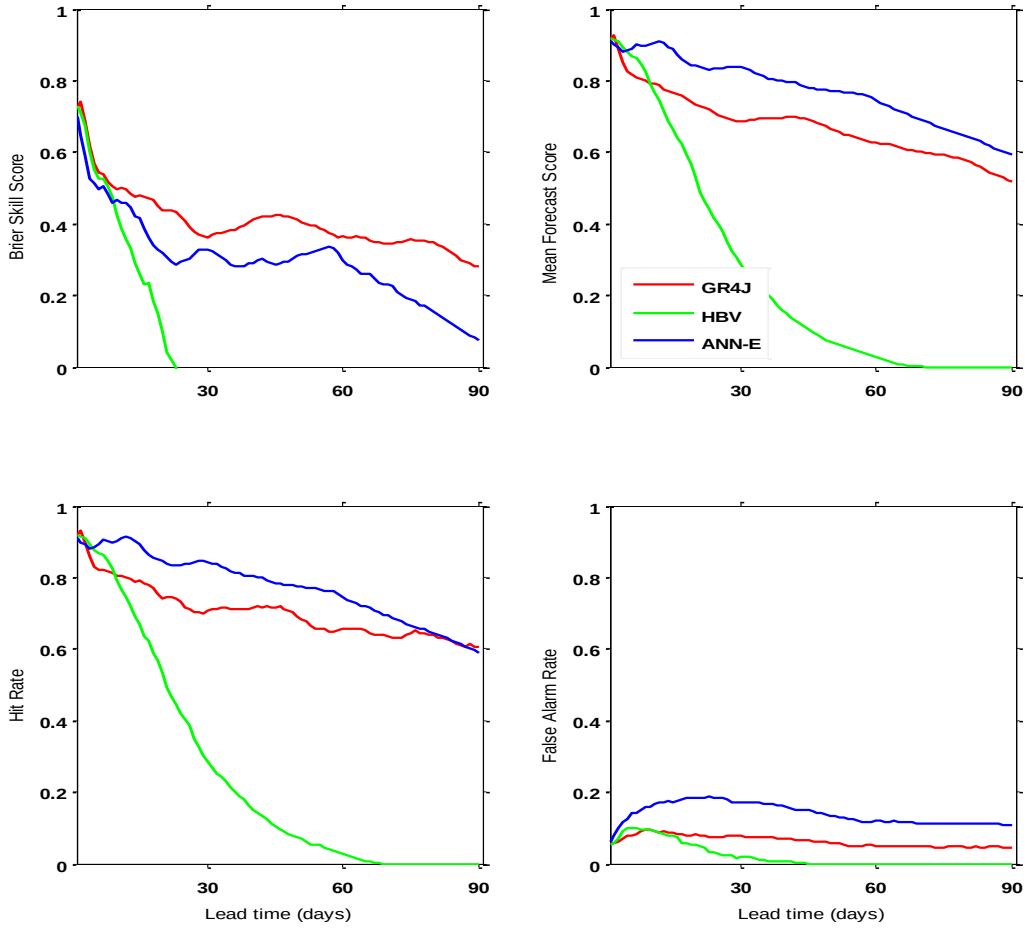


Figure: Skill scores for case-3

Comment 74) Figure 7: Please add the number of low flow events per Figure.

Reply from authors: We will include the number of low flow events in the figures in the revised version of the manuscript.

(The caption of Figure 3 has been revised and the number of low flow events are given as below).

Figure 3 Range (shown as grey shade) of low flow forecasts in **a)** 2002 (the wettest year of the test period with 101 low flow days) **b)** 2003 (the driest year of the test period with 192 low flow days) for a lead time of 90 days using ensemble P and PET as input for GR4J, HBV and ANN-E models (case 1 – 2002 and 2003). The gaps in the figures indicate non-low flow days (i.e. censored).

Comment 75) Table 5: change caption to: '... of low-flow events based on the Q75.'

Reply from authors: We will revise the caption in the revised version of the manuscript.

(Table 5 caption has been revised as below)

Table 5 Contingency table for the assessment of low-flow events based on the Q75

Comment 76) Table 6: This table could be simplified by only showing the cases that are relevant for this article.

Reply from authors: This table is used to introduce a new skill score (MFS). A simplification in this table can confuse the potential user of this skill score.

Comment 77) P 5380 L12: change recipitation to precipitation

Reply from authors: We will correct the typo in the revised version of the manuscript.
(Corrected)

Comment 78) P5380 L27: Start a new paragraph with: 'The first approach...'.

Reply from authors: We will start a new paragraph as indicated in the comment.
(Corrected)

Comment 79) P 5385 L3: Please rephrase the sentence as PET is not observed.

Reply from authors: We will revise the sentence in the revised version of the manuscript.
(Corrected)

Comment 80) P5386 L6: Replace NN-E with ANN-E.

Reply from authors: We will correct the typo in the revised version of the manuscript.
(Corrected)

Comment 81) P 5386 L11: Please introduce G also in the text.

Reply from authors: We will introduce G in the revised version of the manuscript.

(ANN-I is removed from the manuscript)

Comment 82) P 5387 L17: The formula needs to be embedded in a sentence.

Reply from authors: We include a sentence after the formula in the revised version of the manuscript.

(Embedded as below)

The hybrid Mean Absolute Error is defined as

Comment 83) P 5389 L3: delete 'where'

Reply from authors: This word has been used in every equation in the manuscript.

Comment 84) Table 3: There is a shift in the first column of the table.

Reply from authors: There are two sub-columns under the first column. The models are aligned based on their type i.e. conceptual, data-driven and hybrid. HESS uses Latex typesetting and it may have limitations for inserting textbox in a table as we originally provided the table as shown in Appendix – A.