Reviewer #1 (Prof Renata Romanowicz)

The authors thank Dr Renata Romanowicz for her constructive comments on the manuscript. We agree with the comment about the false alarm rates and further analysed the input data. The results are given below. Moreover, we explain how we will modify the text to account for her comments.

Comment 1) The authors answered all my comments and the paper reads well. I am still not convinced by the authors' explanation of the reason for the improvement of False Alarm Rates when lead time is longer than 20 days. Lines 513-514 - apart from not convincing explanation, style is wrong.

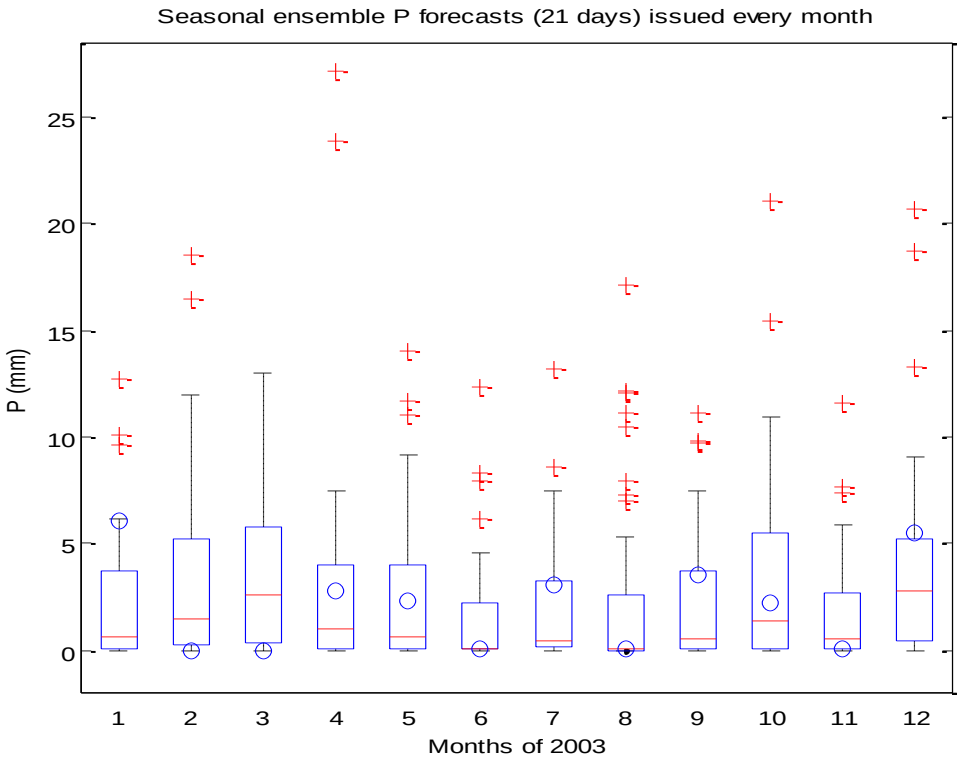
Reply from authors: We agree with the comment. We further analysed the forecasted meteorological forcing data (P and PET) to see if there is any difference between the short lead time (~20 days) and long lead time (e.g. 90 days).
This is done for three different lead times for each model when the false alarm rate was highest (i.e. 12, 15 and 21 days based on the false alarm rates of GR4J, HBV and ANN-E respectively.). We compared the boxplots from these problematic lead times with the 90 day lead time.
It is interesting to note that the ranges for P and PET are larger at 90 day lead time showing the larger uncertainty as compared to shorter lead times. However, the observed P and PET values (i.e. perfect forecasts) are covered by the large ranges resulting in higher hit rates (lower false alarm rates). In short lead times, 12, 15 and 21 days in particular, the ranges for P and PET are smaller than 90 day lead time but the observed P and PET values are usually missed caused higher false alarm rates in the results.

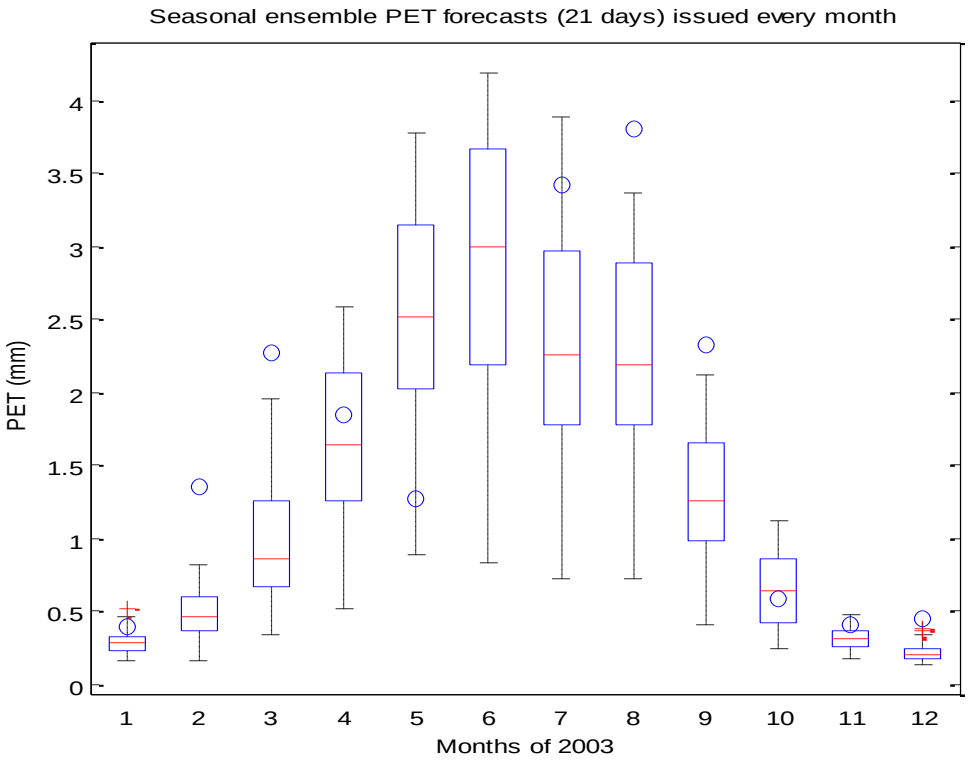
For brevity we will not include the new figures in the revised version of the manuscript. However, we will improve the text based on the findings of this analysis.

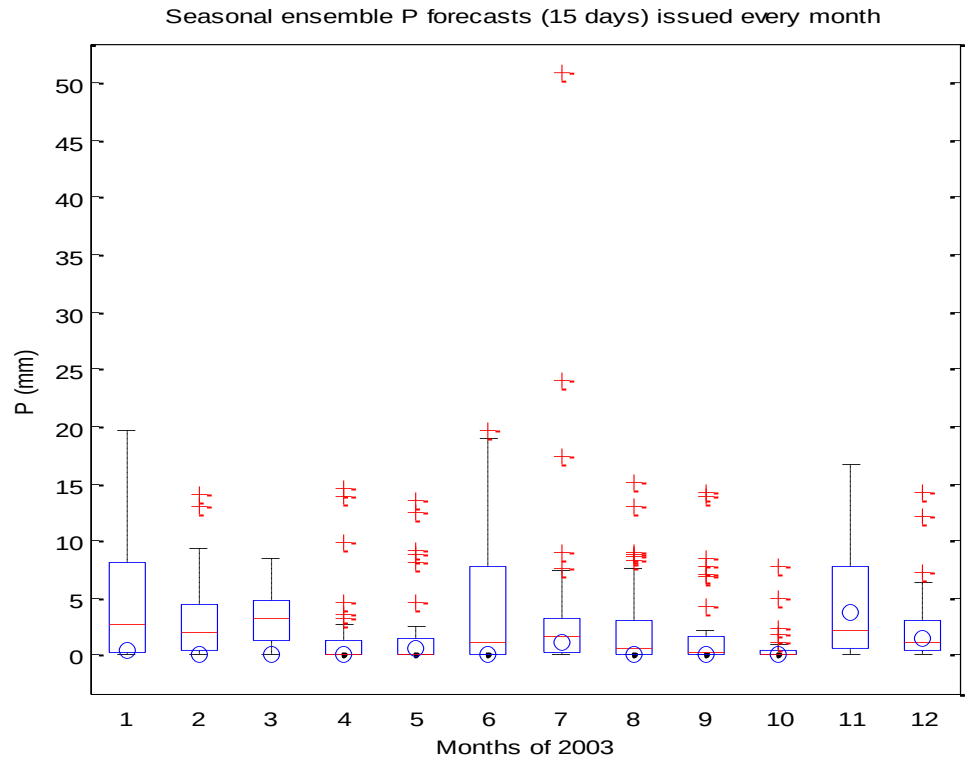
We included the green shaded text below in the revised manuscript;

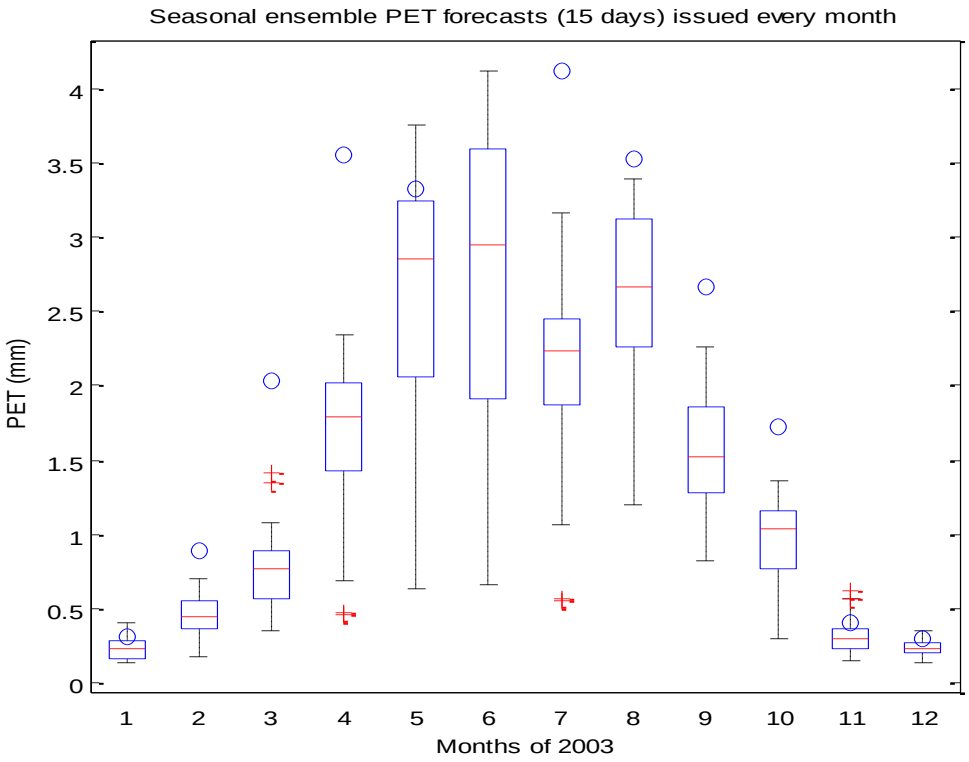
We further analysed the forecasted meteorological forcing data (P and PET) to see if there is any difference between the short lead time (~20 days) and long lead time (e.g. 90 days). This is done for three different lead times for each model when the false alarm rate was highest

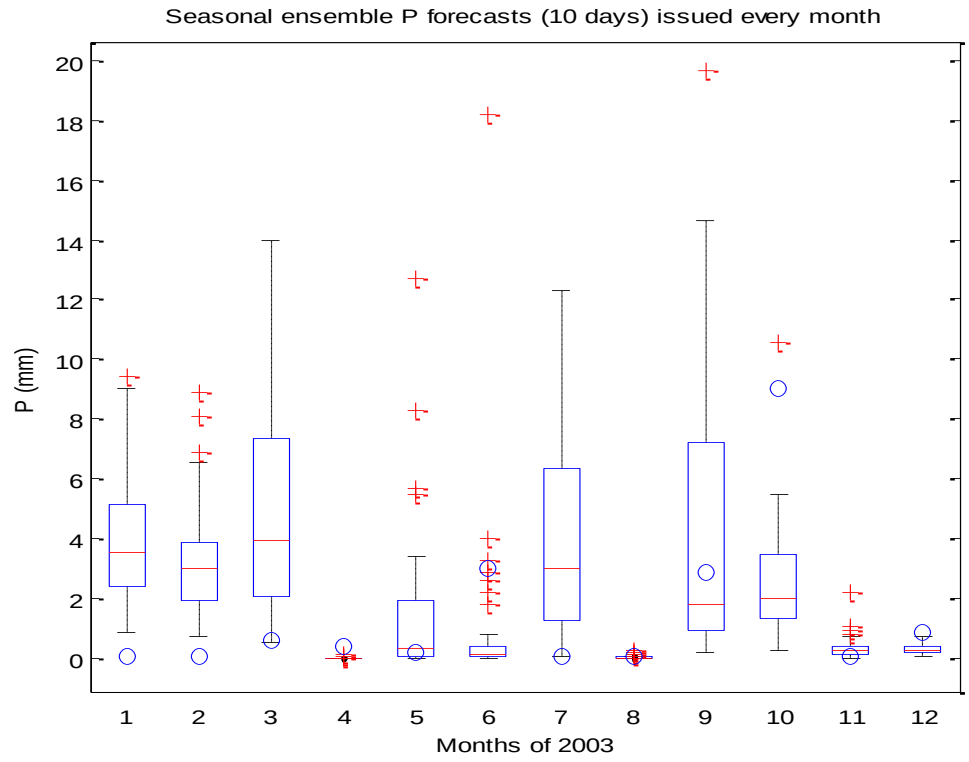
(i.e. 12, 15 and 21 days based on the false alarm rates of GR4J, HBV and ANN-E respectively.). We compared the boxplots from these problematic lead times with the 90 day lead time (not shown here but available in the review reports). It is interesting to note that the ranges for P and PET are larger at 90 day lead time as compared to shorter lead times. However, the observed P and PET values (i.e. perfect forecasts) are covered by the large ranges resulting in higher hit rates (i.e. lower false alarm rates). In other words, for short lead times, 12, 15 and 21 days in particular, the ranges for P and PET are smaller than those for the 90 day lead time but the observed P and PET values are usually missed causing higher false alarm rates in the results.

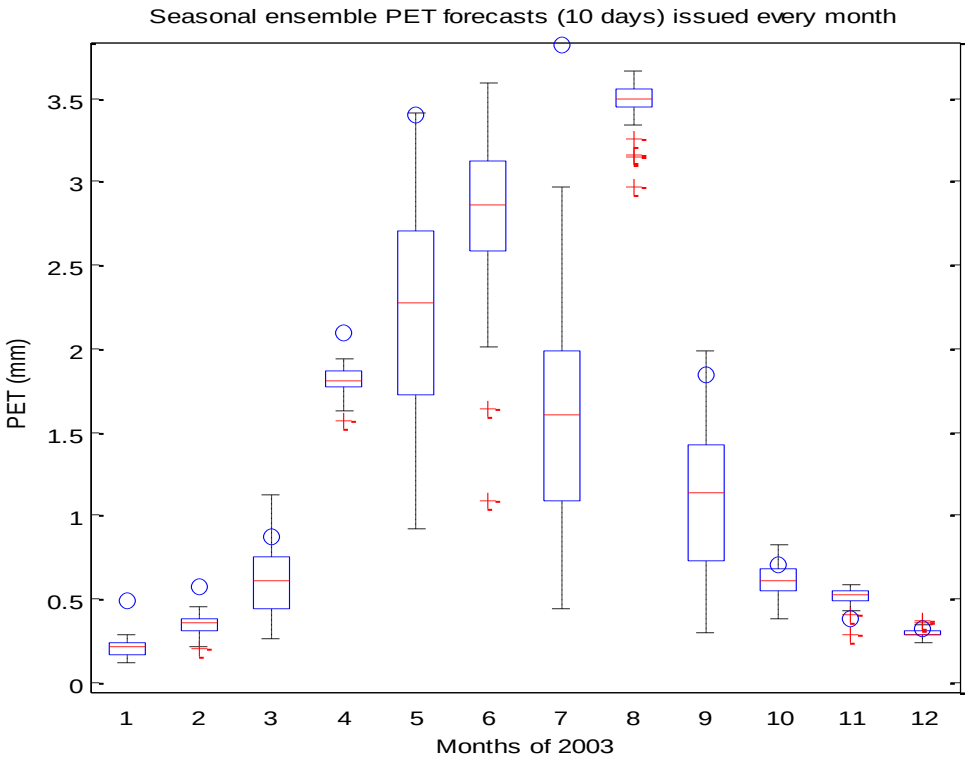


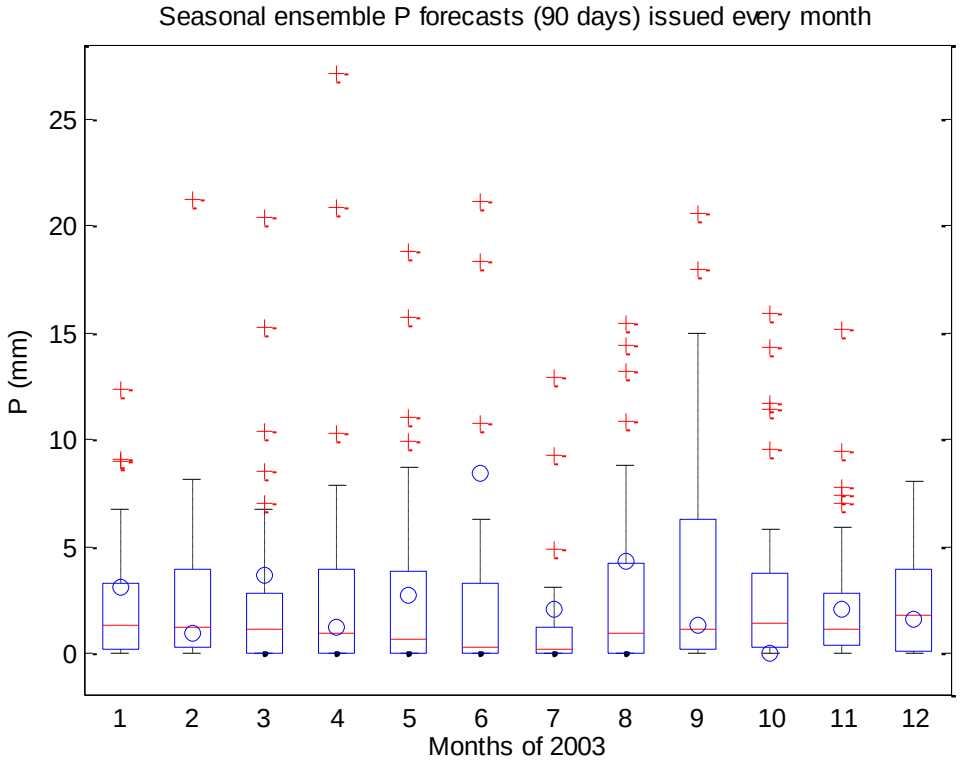

Seasonal ensemble PET forecasts (21 days) issued every month
PET (mm)
0
0.5
1
1.5
2
2.5
3
3.5
4
1
2
3
4
5
6
7
8
9
10
11
12
Months of 2003

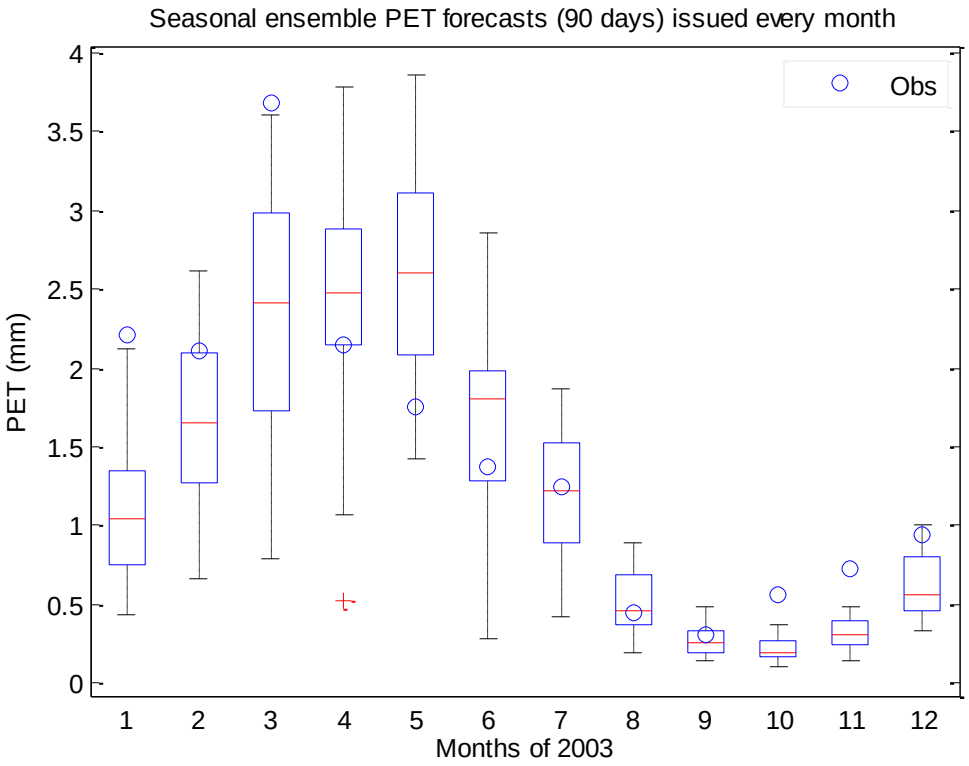

Seasonal ensemble P forecasts (15 days) issued every month
P (mm)
0
5
10
15
20
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12
Months of 2003


Seasonal ensemble PET forecasts (15 days) issued every month
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


Seasonal ensemble P forecasts (10 days) issued every month
P (mm)
0
2
4
6
8
10
12
14
16
18
20
1
2
3
4
5
6
7
8
9
10
11
12
Months of 2003


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


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



Comment 2) There are few style corrections required: lines 359-360 ...calculated for the days when low flow occurred.

Reply from authors: Corrected.

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.

First, I thank the authors for their detailed answer to the review comments and the revised version of their manuscript, which I found clearer. The removal of the ANN-I model makes the presentation more straightforward and results easier to interpret.
However, I still have some comments and requests for clarification, as detailed below. Minor revision is advised.

Specific comments
Comment 3) Title: I still think that the title is too general and should mention that the study is made on the Moselle basin, but it seems to be part of some disagreement with the authors.

Reply from authors: We included the river name in the title.
New title will be:
The skill of seasonal ensemble low flow forecasts in the Moselle River for three different hydrological models

Comment 4) L. 13: "(Precipitation P and potential evapotranspiration PET)"

Reply from authors: Corrected.

Comment 5) L. 29-30: This result should be counterbalanced by the fact that ANN-E shows the poorest false alarm rate.

Reply from authors: We removed the sentence.

~~Furthermore, the hit rate of ANN-E is higher than the two conceptual models for most lead times.~~

Comment 6) Section 2.1: As mentioned in the authors' reply, the authors make the assumption that human influences are negligible. Although I can accept this assumption to keep the study simple, I found this is a strong assumptions and nothing shows that it is actually true. So it should be clearly mentioned in the text, since this may have some impact on the evaluation of hydrological models which simulate natural conditions.

Reply from authors: We included the below text in Section 2.1.

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 in this study.

Comment 7) Section 2.2.2: I found interesting the complementary analysis on the quality of meteorological ensembles shown in the authors' reply. It could help interpreting results and the relative sensitivity to P and PET. Why did the authors choose not to include this information in their manuscript? There is space for that. Note that in the figures, the authors may prefer to show in the X axis the variables for the month of forecast instead of the month when the forecast was issued.

Reply from authors: We further analysed the quality of the ensembles as part of the reviewer-1 comments. We referred to the figures in the text but we prefer to keep them in the referee reports as the number of additional figures is eight. However, we will include the figure for the idealistic run as mentioned in the referee's 13th comment.

Comment 8) L. 224-225: This should be further explained here. If I understood well from the authors' reply, the ANN-E model receives Qobs(t) as input on the time step t when the forecast is issued, and then receives the streamflow forecast of the previous time step as input for lead times larger than 1 day (Q forecast for time step t+j is used as input to forecast Q at t+j+1). This should be more clearly stated in the text, since this is not so obvious.

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

In other words, the ANN-E model receives *Qobs(t)* as input on the time step *t* when the forecast is issued, and then receives the streamflow forecast of the previous time step as input for lead times larger than 1 day. Further, forecasted Q for time step *t+j* is used as input to forecast Q at *t+j+1*.

Comment 9) L. 241: Remove "G is groundwater" since this is not used.

Reply from authors: Corrected.

Comment 10) L. 260-274: I still did not find any strong justification to support the introduction of this hybrid objective function. If this hybrid objective function is introduced, it is certainly because the authors found some added value on model performance, compared to the use of single objective functions. Nothing demonstrates this in the results shown, so I still have the same question on the usefulness of this hybrid objective function. Besides I disagree with the authors that it is not important that this hybrid function aggregates two measures which do not have the same dimension. This is a physical non-sense since it is not possible to give any dimension to MAEhybrid. This should not appear in a scientific paper. Moreover the

variation ranges of MAElow and MAEinverse may be very different, which will tend to emphasize only one part of the hybrid objective function. This should be further analyzed.

Reply from authors: The biggest added value of our objective function is that it can evaluate solely low flows. The unit consistency is an important point. However, the unit does not change the optimum calibration point which is zero. Since both MAElow and MAEinverse will approach to zero, the calibration process will not be affected by different units. The reviewer's concern is especially important when the performance of different models is compared with this metric.

Comment 11) Section 3.1.4: The authors should state in the article that all models were calibrated in simulation mode, i.e. with no update (i.e. no use of Qobs) and no forecast scenario involved. They say in their reply that this is also the case for ANN-E. However this may induce a model that gives much weight on streamflow input. Is it the case? Besides the internal weights may be different if the ANN was more classically trained in forecasting mode. Did the authors check this? The way they use the ANN may push the model towards keeping the streamflow constant (i.e. not much changing the input information), which may explain this behavior noticed for ANN to generate almost constant predictions.

Reply from authors: The newly added Figure 6 below shows that ANN-E behaviour is comparable to the conceptual models.

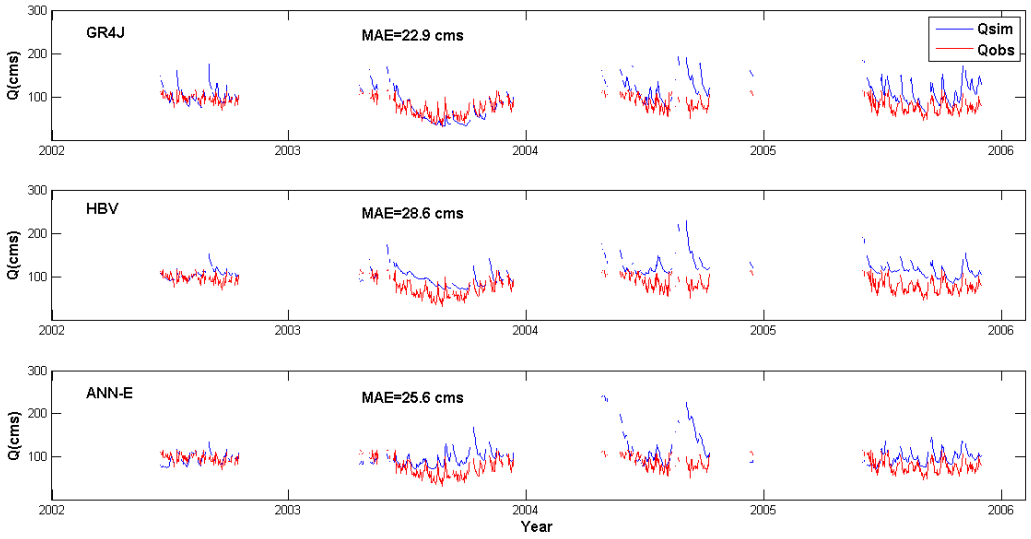


Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

Comment 12) Section 3.1.5: The authors mention later in the text (L. 511-514) that after the model updates, there may be some spin-up period. I guess this is true given their results, but it means to some extent that the way the update is done is not optimal, since one would not expect this behavior. The role of model update on model results and the possible margin of improvement in the update approaches should be further commented since updating probably plays a role in the results obtained in forecasting mode. Indeed, there is strong contrast

between the poor performance of ANN-E model in validation (Figure 2b) and its apparently good/better performance shown in Fig. 6. The difference may lie in the way models are run in forecasting mode.

Reply from authors: We explained this in detail in comment #1 of the reviewer #1. Forecasted meteorological forcing quality is further analysed and results are shown in our reply.

Comment 13) Table 4: After reading the revised version of the manuscript, I found that one benchmark test case is missing here: the case where models are run using observed P and PET as forecasts. Although this is an idealistic case, I think this would help interpreting results by giving reference performance without uncertainty on future P and PET conditions, and hence better assess the intrinsic value of the models. Could the authors add this case and shortly comment the results?

Reply from authors: We agree with the comment. We carried out the recommended run and reported the results in the revised version of the manuscript as shown below.

6 shows the performance of the three models in the test period using perfect P and PET forecasts as input. This is an idealistic case showing that GR4J model performs better than the other two models. It is interesting to note that ANN-E model does not produce constant predictions as in the previous figures showing the ability of this black box model to perform comparable to the conceptual models when configured and trained properly.

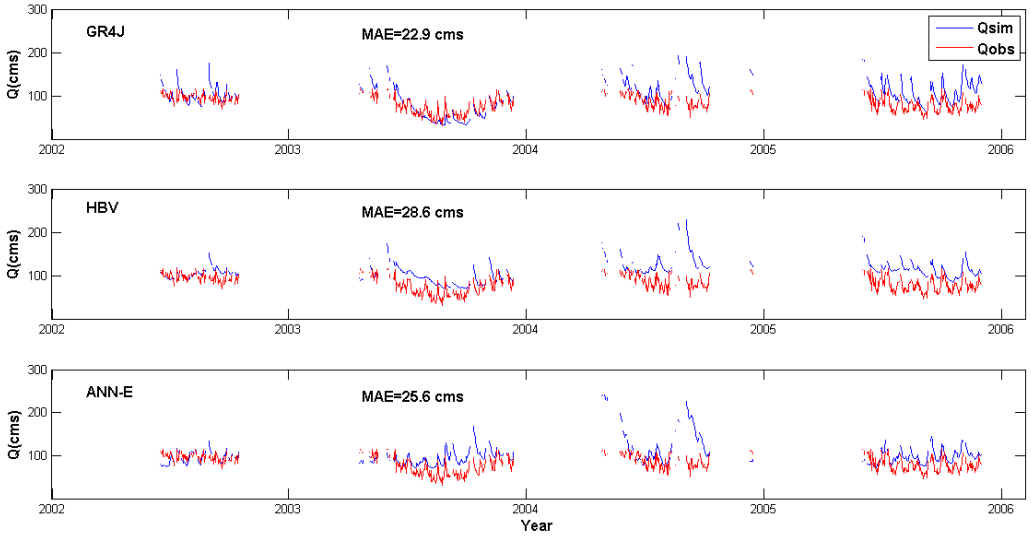


Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

Comment 14) Section 3.2.4: Like in the case of the objective function, I did not find clear justification in the text for introducing this new evaluation metric. Which forecast quality does it reflect? Are the existing criteria not sufficient to evaluate this quality?

Reply from authors: The main advantage of this skill score is the simplicity of the calculation compared to the Brier Skill Score. It shows the probabilistic forecast skill in a very straightforward way. The existing criteria evaluates the forecasts altogether without any classification or event selection. Our new evaluation criterion and new objective function are both selective evaluators for time series.

Comment 15) L. 368: So why is there a column for observed probabilistic flows in Table 6?

Reply from authors: Table 6 is a conceptual table for general cases. The user can decide to use probabilistic value (e.g. 34/50 years indicates low flow for day j) or s/he can decide to use binary cases based on the majority of the years. If 34 from 50 years are low flows then low flows (value 1). Therefore there is a column for observed probabilistic flows.

Comment 16) Table 7: The authors' reply on CFLUX (value optimized at the upper limit) is a bit disappointing. Although I understand other studies used this variation range, I do not see strong physical reasons why CFLUX could not take values above 1 mm/d in some specific cases. What can be learnt on the model behavior from the fact that the model tries to have a high CFLUX value? Is it an issue of parameter sensitivity? Does it reveal a specific behavior of the Moselle basin?

Reply from authors: This is a very good question for further analysis. However, it is beyond the scope of this study.

Comment 17) Table 7: For W1 to W3, please indicate to which input they refer respectively (P, PET, Q). Since the values of weights may depend on units, were all inputs and outputs used with the same units (mm/d) to build the ANN? Given these weights are much different, what does it mean on the respective roles of these inputs on model output? Does the ANN tend to emphasize the information of one of them?

Reply from authors: We revised the table based on the reviewer comment.

W1	[-]	-10 to +10	-2.3	Weight of connection between 1st input node (P) and hidden neuron
W2	[-]	-10 to +10	0.03	Weight of connection between 2nd input node (PET) and hidden neuron
W3	[-]	-10 to +10	-0.02	Weight of connection between 3rd input node (*Q(t-1)*) and hidden neuron

Comment 18) Figure 4b: I still did not fully understand the reason for the erratic behavior of the ANN, whose output oscillates between two values. Does not this discredit this model to some extent?

Reply from authors: The new benchmark figure below shows the skill of the ANN-E model as compared to other two conceptual models. The input quality (variation) is the main reason for the behaviour of the ANN-E in other cases.

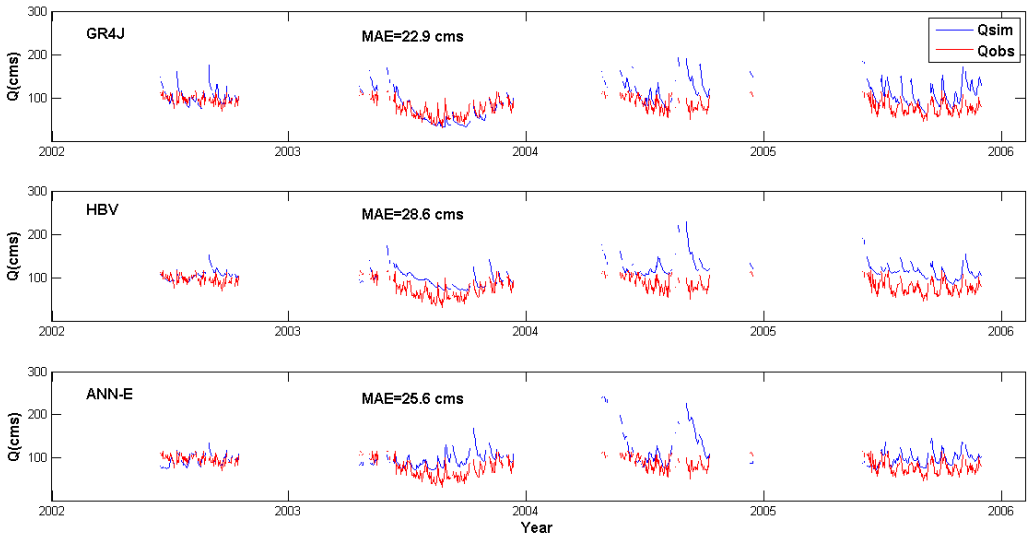


Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

Comment 19) Table 8: Indicate in the caption for which period and lead time the performances were calculated. What bold values indicate? Please add median values also between min and max.

Reply from authors: Median values are added in the table and the Table caption is revised based on the reviewer comment.

Table 8 Minimum and maximum prediction errors for low flow forecasts for a lead time of 90 days during the test period 2002-2005

Model	Minimum , **Median** and Maximum MAE (m^3/s)				
	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**
HBV	[23 101 **785**]	[23 72 **600**]	[108 119 135]	[105 105 105]	[57 57 57]
GR4J	[33 122 **906**]	[36 75 **646**]	[46 61 111]	[44 44 44]	[55 58 59]
ANN-E	[17 94 227]	[18 72 221]	[65 73 80]	[65 65 65]	[16 16 17]

Comment 20) L. 506: This is not true for HR and FAR criteria, as explained a few lines later in the manuscript.

Reply from authors: We removed the sentence to avoid ambiguity.

~~As anticipated, all scores decrease with increasing lead time.~~

Comment 21) L. 511-514: Although this explanation seems plausible in the case of the two conceptual models, does it also apply to the ANN, which does not have internal states to update per say?

Reply from authors: We agree with the comment. We revised the text as below to clearly state that we used the deterministic updating procedure only for GR4J and HBV models.

When the forecast is issued on day (t), the model states are updated using the observed discharge on that day (t). For GR4J and HBV we used the deterministic state update procedure described in section **Error! Reference source not found.**.

Comment 22) Fig. 6: Remind in the caption on which period the criteria are calculated (all the time steps or only low flow time steps) for each criterion.

Reply from authors: We agree with the comment. We included the extra information in the caption.

Figure 7 Skill scores for forecasting low flows at different lead times for three different hydrological models for the test period 2002-2005. Note that all forecasts (including high and low flow time steps) are used to estimate these skill scores.

Comment 23) L. 525, 528, 534: Is the Q95 or Q99 threshold used? Check consistency.

Reply from authors: The Q95 and Q99 are used only to classify the forecast results in different quantiles. Our study focuses mainly on Q75 low flows.

Comment 24) L. 554-567: I found this part not so clear. L.555-556: The results are somehow contrasted on the four evaluation criteria (results are different for the FAR). L559-560: It is not so clear that ANN-E is better than HBV given results shown in the previous sections of the manuscript. Besides, the argument given L. 565-567 tends to limit the practical usefulness of ANN-E.

Reply from authors: The new benchmark figure below clearly shows the MAE for ANN-E is lower than that for HBV. We revised the text as below.

Further, our study showed that data-driven models can be good alternatives to conceptual models for issuing seasonal low flow forecasts (e.g. Figure).

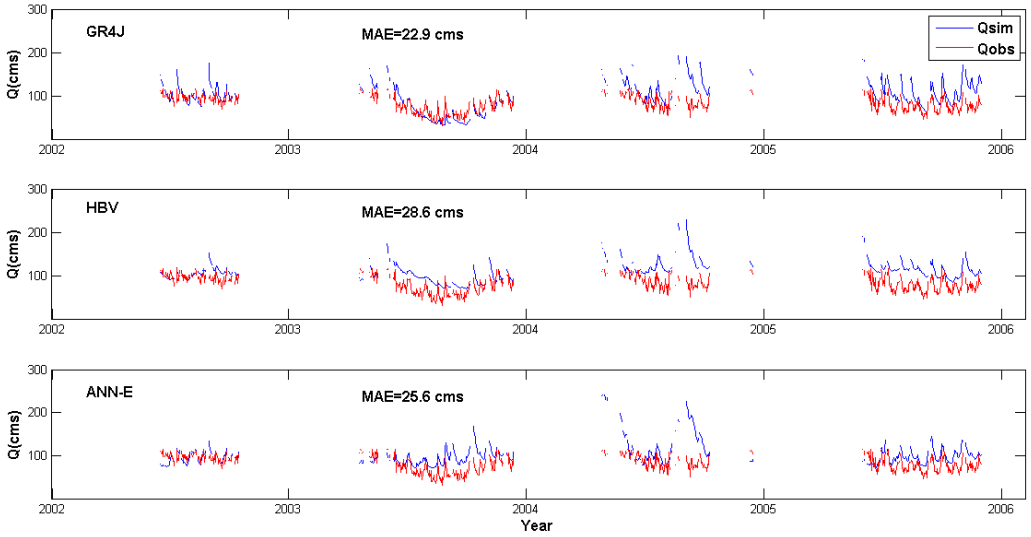


Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

Comment 25) L. 574: This somehow contradicts what is said on the objective function L. 272-273.

Reply from authors: We agree with the comment. We included the green shaded part in the revised version of the manuscript.

Since the first part of the objective function used in this study solely focuses on low flows, the high flow period is less important in the calibration.

Comment 26) L. 587-592: The authors emphasize the fact that they proposed a new objective function and a new evaluation criterion. However, as noticed above, there is no demonstration of the usefulness of these new criteria compared to existing ones.

Reply from authors: Please refer to the answer for the comment #14.

Comment 27) L. 608: "less skillful than the other two models"

Reply from authors: The sentence is removed to avoid ambiguity.

Based on the results of the comparison of forecast skills with varying lead times, the false alarm rate of GR4J is the lowest indicating the ability of the model of forecasting non-occurrence of low flow days.

Comment 28) L. 617-621: Given the strange behavior of the ANN-E model shown in the article (constant range in Figs. 3 and 4, erratic behavior in Figs. 4 and 5), I personally doubt this model would be really useful in practice and would convince practitioners. Therefore I wonder why the overall evaluation comes to an opposite conclusion. Is this model apparently working well for the wrong reasons?

Reply from authors: Based on the results shown at Figure 6 below, we believe that ANN-E type models can be applied to low flow forecast problems.

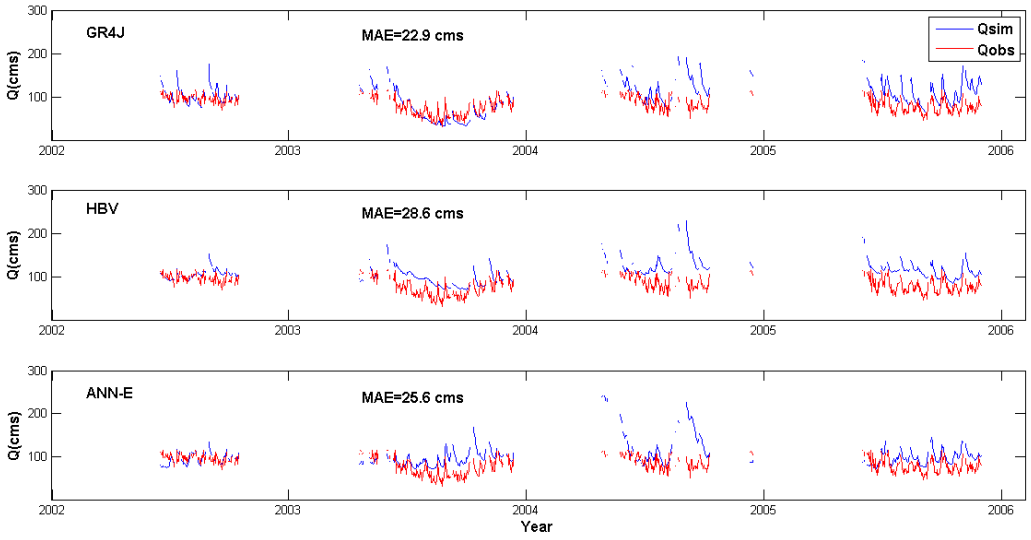


Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

Reviewer #3 (Stefanie Jörg-Hess)

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.

It is nice to see that the paper has gained in clarity and has been improved from the last version. Most of the issues listed in the first review have been addressed satisfactorily in the revised version of the work: "The skill of seasonal ensemble low flow forecasts for three different hydrological models."

I have only a few comments:

Comment 29) • P7: Thank you for the additional information on the annual ranges for P, PET and Q. It is still not clear to me how many stations are used to estimate P and PET. Is there 1 station in each sub-basin? Please include the number of meteorological stations used to estimate P and PET in the study.

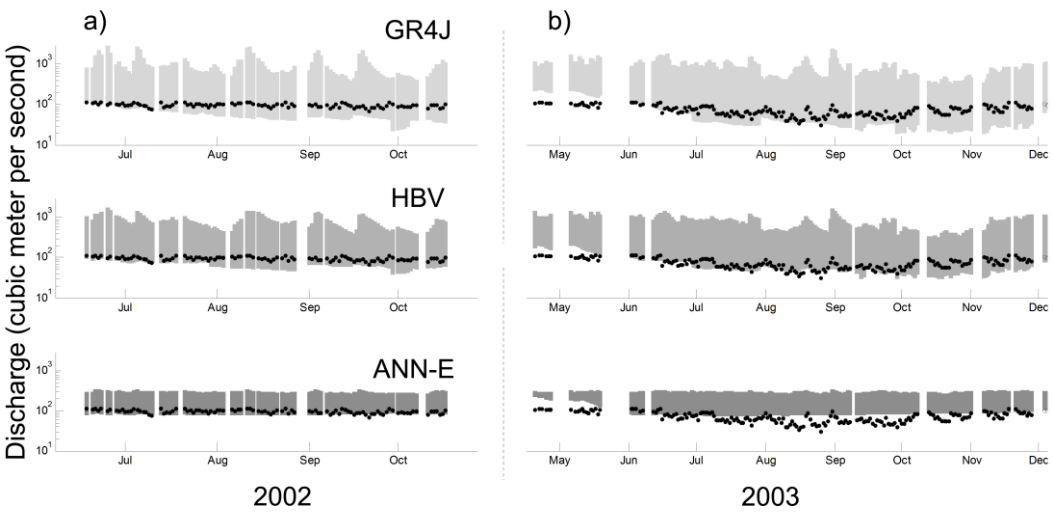
Reply from authors: We included the number of meteorological stations in the revised version of the manuscript (green shaded text).

Observed data from 12 meteorological stations in the Moselle basin (as part of 49 stations over the Rhine basin), mainly provided by the CHR, the DWD, Metéo France, are used to estimate the basin averaged input data (Görgen et al., 2010).

Comment 30) • Figures 3,4 and 5: You state in your answer that you increased the visibility of the points in these figures by using bold and filled circles. I do not see any difference in the Figures of the revised manuscript and again recommend to increase the visibility of the points.

Reply from authors: We agree with the comment. The visibility in the three figures is increased as shown below.

Example:



Comment 31) • It is interesting to see the number of low-flow events for the years 2002 and 2003. It would be nice to see the number of events also in Figure 7 for the different quantiles.

Reply from authors: We included the number of events in different quantiles in the figure caption (as shown below).

Figure 7 Reliability diagram for different low flow forecasts **a)** Low flows below Q75 threshold (584 observed events in the test period 2002-2005) **b)** Low flows below Q90 threshold (250 observed events) **c)** Low flows below Q99 threshold (20 observed events). The forecasts are issued for a lead time of 90 days for the test period 2002-2005 using ensemble P and PET as input for GR4J, HBV and ANN-E models.

References:

Görgen, K., Beersma, J., Brahmer, G., Buiteveld, H., Carambia, M., de Keizer, O., Krahe, P., Nilson, E., Lammersen, R., Perrin, C., and Volken, D.: Assessment of Climate Change Impacts on Discharge in the Rhine River Basin: Results of the RheinBlick 2050 Project, Lelystad, CHR, ISBN 978-90-70980-35-1, 211p. Available from: http://www.news.admin.ch/NSBSubscriber/message/attachments/20770.pdf (last access: 30/10/2014), 2010.