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 account for her comments.

Comment 1) 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.

Madadgar, Shahrbanou, Hamid Moradkhani, 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

Comment 2) 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. The 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 procedure.

Comment 3) 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 4) 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 period using ensemble seasonal meteorological forcing.”

Comment 5) 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.

Comment 6) 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.

Comment 7) 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 mean altitude of the catchments (h) have been obtained from the German Federal Institute of Hydrology (BfG) in Koblenz, Germany.

Comment 8) 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 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).

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

Comment 9) 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/

Comment 10) 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.

Comment 11) 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.

Comment 12) 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 the Reviewers #2 and #3. We also noticed a typo in ANN-I part as "*Q: State update*" is relevant only for ANN-E and other two conceptual models.*

It should be noted that we will remove ANN-I from the revised version of the manuscript. We will also revise the Table 3 and 7 accordingly. The number of weights is 4 for both ANN-I and ANN-E models: 3 weights connecting input layer to hidden layer, 1 weight connecting hidden layer to output layer. ANN-E and ANN-I have 3 inputs. ANN-E has P, PET and Q, whereas ANN-I has P, PET and G as inputs. In short, the illustration in Table 7 is correct.

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

Comment 13) 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 systematic biases of the model structure were not assessed as the main focus of this study was input uncertainty.

Comment 14) 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).

Comment 15) 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.

Comment 16) 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.

Comment 17) 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.

Comment 18) 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.

Comment 19) 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 references 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 ANN model with one hidden neuron structure to the seasonal low flow forecast problem.

Comment 20) 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 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 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 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 fall in this period. Since the objective function used in this study solely focuses on low flows, the high flow period is implicitly ignored. The low flows occurred 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.

Comment 21) 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.

Comment 22) 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.

Comment 23) 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.

Comment 24) 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 to the revised version of the manuscript for brevity of the paper. Since the review reports are also public and shown together with the papers, the readers may read and refer to these figures below.

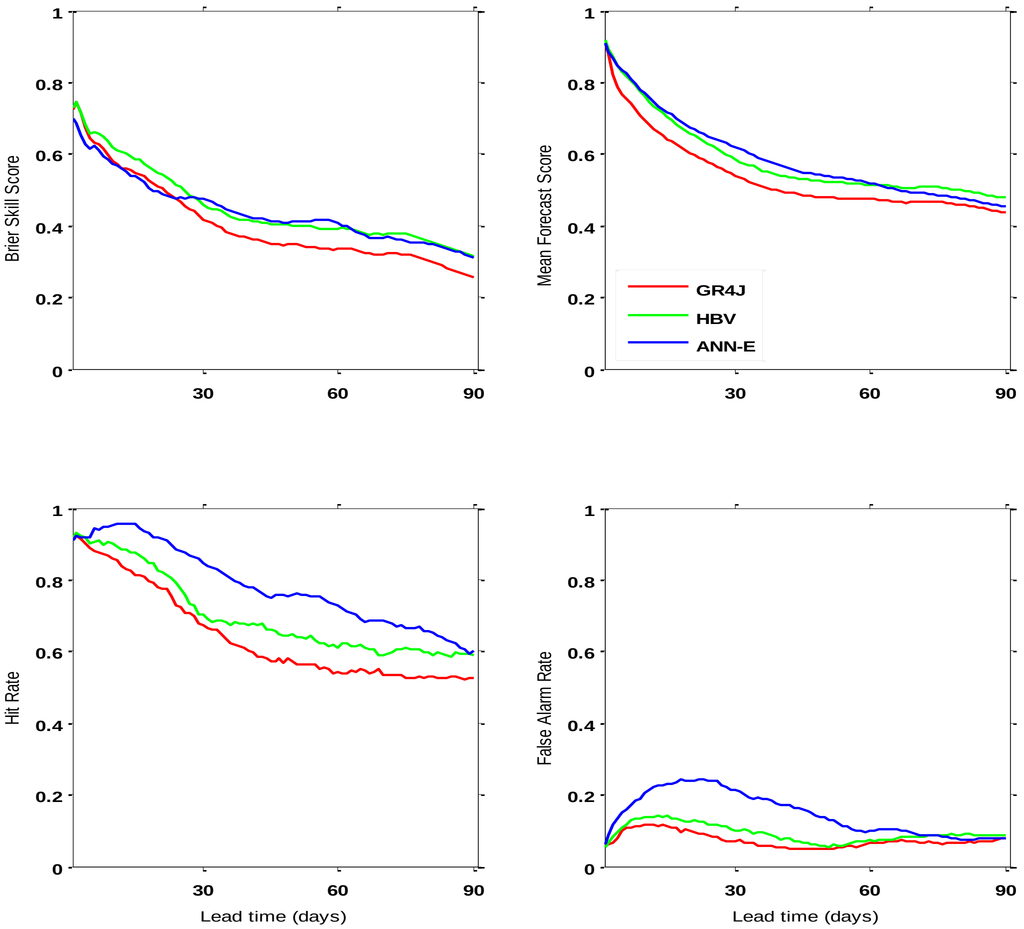

Figure: Skill scores for case-2

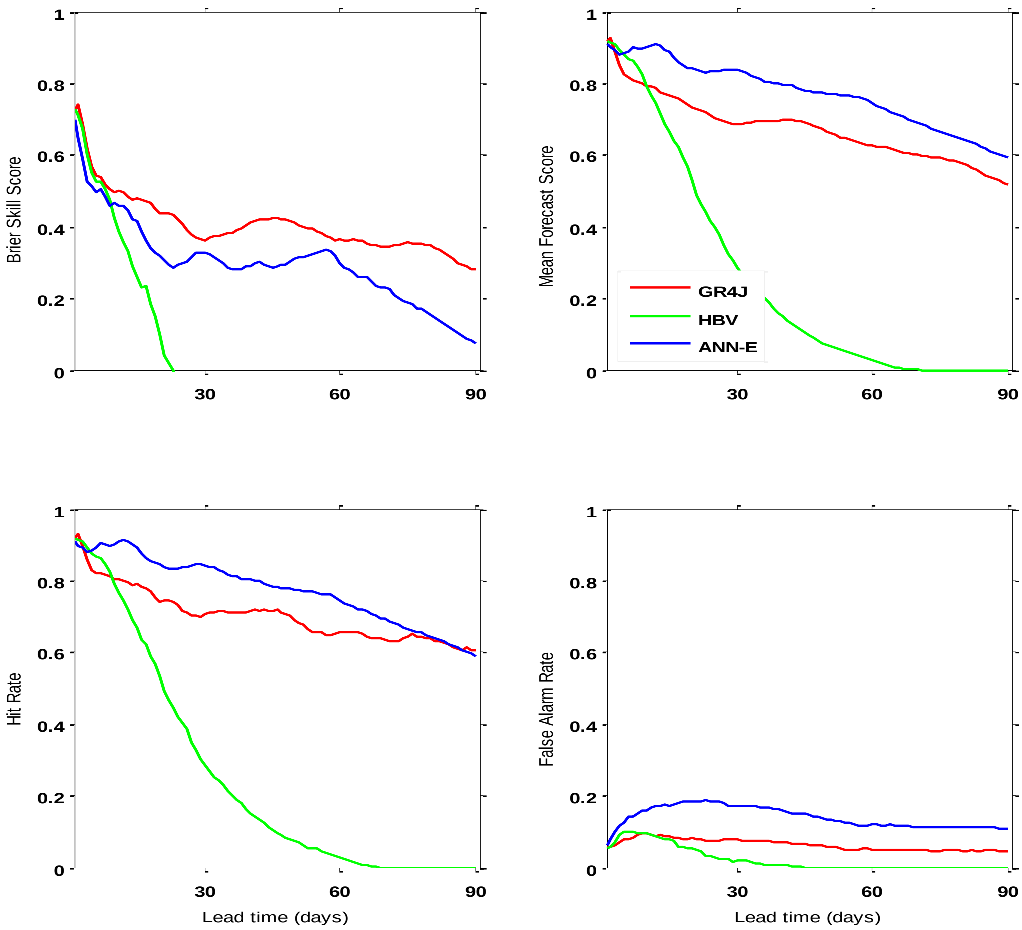

Figure: Skill scores for case-3

Comment 25) 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.

Comment 26) 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.

Comment 27) 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 28) P 5380 L12: change recipitation to precipitation

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

Comment 29) P5380 L27: Start a new paragraph with: 'The first approach...'.

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

Comment 30) 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.

Comment 31) P5386 L6: Replace NN-E with ANN-E.

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

Comment 32) P 5386 L11: Please introduce G also in the text.

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

Comment 33) 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.

Comment 34) P 5389 L3: delete ‘where’

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

Comment 35) 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.

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				
Conceptual	**Data-driven**		**Temporal resolution of input**	**Lag between forecast issue day and final day of temporal averaging (days)**	**Model time step**	**Model lead time (days)**
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	**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

APPENDIX-B: Table 6

Table 6 Low flow contingency table for the assessment of forecasts

		Observed low flows	
		Deterministic	**Probabilistic**
Forecasted low flows	**Deterministic**	*O_j=1 (Low flow observed)* *F_j=1 (Low flow forecasted if more than half of the ensemble members indicate low flows) otherwise 0*	*O_j= Observed frequency based on long term climate (e.g. 34/50 years indicates low flow for day j)* *F_j=1 or 0*
	Probabilistic	$O_j = 1$ *F_j=Forecast frequency based on 40 ensemble members (e.g. 23/40 members indicate low flows for day j)*	*O_j= Observed frequency based on long term climate* *F_j=Forecast frequency based on 40 ensemble members*

References

de Vos, N. J., and Rientjes, T. H. M.: Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation, Hydrol. Earth Syst. Sci., 9, 111-126, 10.5194/hess-9-111-2005, 2005.
Maier, H. R., and Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environ Modell Softw, 15, 101-124, 2000.
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.
Yuan, H. C., Xiong, F. L., and Huai, X. Y.: A method for estimating the number of hidden neurons in feed-forward neural networks based on information entropy, Computers and Electronics in Agriculture, 40, 57-64, 10.1016/s0168-1699(03)00011-5, 2003.