

## Review of “Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations” by W. Greuell et al.

**Reviewed:** December 2016

**Recommendation:** The manuscript is acceptable with minor revisions.

In this paper, the authors present a model-based seasonal hydrological forecasting system, which produces hydrological forecasts for up to seven months of lead time over Europe. As the authors state it, seasonal hydrological forecast systems over Europe are scarce, which makes this work relevant to HESS and to the wider hydro-meteorological community. Furthermore, we are currently at a turning point where model-based dynamical systems are becoming more widely used for seasonal hydrological forecasting. This is because it is only recently that dynamical modelling systems have started becoming at least as skilful as statistical modelling systems or dynamical-statistical hybrid systems. This makes the system presented in this paper state-of-the-art.

The authors analyse the skill of the seasonal runoff and discharge hindcasts against pseudo- and real observations, using a variety of metrics. This complete analysis allows to tackle many aspects of seasonal hydrological forecasting and the results are presented in a pleasant to read and concise way. This paper first demonstrates the levels of predictability reached by this system and the spatiotemporal patterns of skill. From this analysis, the authors have successfully identified regions and periods of high runoff skill. The evaluation also highlights the effect of delay between runoff and discharge on the higher discharge forecasting skill. Furthermore, by doing a comparison between hindcasts verification against pseudo- and real observations (theoretical and actual skill, respectively), the authors have shown that there is a higher theoretical skill than an actual skill in seasonal hydrological forecasting, pointing out the need for actual skill calculations. The last part of the analysis is dedicated to the overview of the metric choice on the results of the analysis, stressing the differences and similarities between the metrics.

The paper is overall clear, written in a generally fluent and precise language, and presents a large quantity of results in a structured and concise way, in a paper of appropriate length for the content. The methods are interesting and give enough details for reproducibility of this work. The paper would nevertheless benefit largely from an improvement of the introduction and the discussion sections, with the aim to set the wider context of this work to the readers.

As a whole, I enjoyed reading this paper and I will therefore be pleased to see it published in HESS. Below are minor comments which will hopefully help the authors to improve the paper.

**Title:** The title is pertinent with regards to the contents of the paper. However, I don't like the terms “pseudo-observations” and “real observations”. I would name them differently, such as “analysis” (as done in meteorology) or “simulations”, for the pseudo-observations, and simply “observations” for the “real observations”.

**Abstract:** Overall, the abstract provides a concise and complete summary of the paper. Here are however a few suggestions that could help clarify certain aspects of the abstract:

- It would be good to say that the hindcasts have 7 months of lead time earlier than on page 1, line 19. This could be mentioned for example in the sentence on page 1, line 15: “Skill is

analysed with a monthly temporal resolution, up to 7 months of lead time, for the entire annual cycle”

- Page 1, line 23: it was not clear to me what the sentence “a conceptual analysis of the two types of verification” meant. Could you please rephrase this to clarify it to the readers here? It could be rephrased to, for example, “attributed to the structural differences between the runs used for the two verification methods.”
- Before reading page 1 line 20, it wasn’t clear to me that both discharge and runoff were analysed in this paper. It would be good if you could specify it each on in the abstract.

**Introduction:** The introduction is interesting, but it could overall contain more literature review on seasonal hydrological forecasting in general: e.g., statistical versus dynamical methods and the state of seasonal hydrological forecasting over Europe, stating the current predictability in Europe (referring to work previously done on the same topic). Here are a few other suggestions that could maybe help to make the introduction more concise.

- Page 1, line 28: the word “may” sounds like society may also not benefit from such forecasts. It would therefore be interesting to refer to papers tackling this topic, such as: Viel et al. (2016), Soares and Dessai (2016), Crochemore et al. (2016), among others.
- Page 1, line 30: it would be good to add references for other applications of the seasonal predictions, as done for the energy generation sector.
- Page 1, line 33: the word “usefulness”, just like the word value, is a complex one. Indeed, the usefulness of a system does not only depend on the skill of the forecasts that it produces, but also on the way this skill is transformed into a decision within one of the sectors of interest. This is an interesting post on this topic: <https://hepex.irstea.fr/economic-value-of-hydrological-ensemble-forecasts/>. I would therefore suggest to change this sentence slightly to acknowledge this complexity in the value of probabilistic forecasts for decision-making, by saying for example: “The usefulness of the system depends partially on [...]”.
- Page 2, lines 3-4: see my comment for the title of the paper.
- Page 2, lines 6-8: another example of the use of “pseudo-observations” rather than “real observations” is in cases when the aim is to exclude the model error from the analysis in order for example to perform a sensitivity analysis to other components of the forecasting system. For example, the VESPA method introduced in Wood et al. (2016), to look at the contribution of initial hydrological conditions and seasonal climate forecast errors to seasonal streamflow forecast uncertainties. It would be worth mentioning this here.
- Page 2, line 9: you mention that the fact that “pseudo-observations” are not equal to “real observations” is a downside, which is a very good point. This however needs clarification on how it could influence an analysis of the skill of the forecasts here. The sentence on page 2, lines 14-15, could for example be rephrased to sound like a hypothesis and moved earlier.
- Page 2, lines 13-15: this description is already done in the last paragraph of the introduction (page 2, lines 33-34). It is also too methodological for this part of the introduction, which should be more focused on literature review. I would thus suggest to remove it here.
- Page 2, lines 19-23: references to these papers are very interesting. It would be even more interesting if you could also mention results of these analyses briefly, such as answers to the following questions: what is the current predictability in Europe? Where are the high skill areas?
- Page 2, line 24: could you please add “presented in this paper” after “The hydrological hindcasts”? This would then make it clear what you are talking about.

- Page 2, lines 26-28: could you please state here that the initial hydrological conditions are used for the hindcasts generation?
- Page 2, lines 30-31: could you please specify that this aim is to look at the effects of using “pseudo-observations” for the verification of the hindcasts, as opposed to using “real observations”?
- Page 2, line 34: the sentence about the supplementary figures seems out of place here. I would rather mention in the introduction paragraph of the results section of this paper.
- Page 2, lines 34-40: the results of the comparison paper are very interesting but seem out of place here as well. They should either be moved to the discussion section of this paper or mentioned earlier in the introduction, and well linked to the rest of the introduction.

### Section 2.1:

- Page 3, lines 14-15: what is the time step of these hindcasts? Daily? It would be good to mention it here.
- Page 3, line 15: consider changing the word “simulations” to “hindcasts”, as it is confusing otherwise.
- Page 3, lines 17-18: could you please specify that these are the System 4 ensembles?
- Page 3, lines 19-24: it would make the lecture of this technical description more structured if this paragraph was combined with the paragraph on page 3, lines 11-13.
- Page 3, line 25: the sentence “and in addition for spin-up periods” could be removed and the following sentence could be linked to the previous to make it clearer. This would then give: “VIC was run for the period of the S4 hindcasts (1981-2010). Additionally, for the reference simulation, two extra years (1979-1980) were run to spin up [...]”.
- Page 3, lines 29-31: why were the simulations done with a three-hourly time step? It would be good to clarify this here.
- Page 3, lines 37-38: I don’t understand what these four other hydrological models are and why they are mentioned here. If they are not used in this paper, I would suggest to remove this piece of the sentence as it might confuse the readers.
- Page 3, lines 39-40: It is interesting to note those aspects as key for seasonal predictions! However, could you please specify what is meant exactly by “more or less in the middle of the ranking of the five models”, by for example using scores to support this sentence?

### Section 2.2:

- Page 4, lines 4-5: how were the data sets converted to gridded versions? It would be useful to mention this here.
- Page 4, line 7: it would be good to mention the area of the grid cells that the catchments cannot pass in order to be considered as “small basins” here.
- Page 4, lines 23-27: what if there are 2 neighbouring cells without an influx from any of the neighbouring cells, corresponding to two small basins? How can we be sure that that nearest cell is in fact that small basin and not the other cell?
- Page 4, line 26: this sentence is not entirely clear to me. Do you mean all of the cells with no influx from the eight neighbouring cells?
- Page 4, line 27: is this method appropriate?
- Page 4, lines 29-30: could you please specify that this is over Europe, to remind the reader?

### Section 2.3:

- Page 4, lines 32-33: it would be good to repeat here again that the analysis was carried out on the 7 months of lead time.
- Page 4, lines 38-39: this explanation is slightly confusing. Could you please rephrase it to make it clearer to the readers?
- Page 5, lines 1-3: from reading the results, forecasts with zero lead time are actually still mentioned a fair amount of times.
- Page 5, line 4: it would be good to specify why you refer the readers to Mason and Stephensen (2008). Is it because they selected the skill metrics?
- Page 5, line 5: please consider changing the word “simulations” to “forecasts” here.
- Page 5, line 6: what is called the “ROC graph” here is usually called the ROC curve.
- Page 5, lines 7-9: further details are needed for the computation of the ROC score. Please consider providing more details on the following questions: Are the terciles for the ROC computed on the “pseudo-observations”? Are the terciles calculated for each month individually or for the whole period? And from monthly averages? How many bins are used for the ROC?
- Page 5, line 8: the “one third highest, lowest and the remaining values” could simply be called “upper, lower and middle terciles”.
- Page 5, lines 9-11: this is vague, it would be nice to talk about attributes of the forecasts and to mention the attributes covered by each metric.
- Page 5, line 11: by “value falling in the considered tercile” do you mean “percentage of ensemble members falling in the considered tercile”?
- Page 5, line 12: it would be good to describe the RPS first, then the RPSS. Also, what is the reference forecast used for the RPSS calculation?
- Page 5, line 13: could you please specify what is meant by “correct forecasts” here? Reliable? Sharp? Accurate?
- Page 5, line 14: is the climatology used as a reference forecast for the measure of skill then?
- Page 5, line 14: by “climatological forecasts (forecasts that are identical each year)”, do you mean an ensemble of past historical observations? This is not so clear here.
- Page 5, lines 14-15: could you also please specify what are the best values for each metric. So what value would a perfect forecast have?
- Page 5, lines 19-22: this paragraph should rather be included in the introduction of the results section I think.
- Page 5, line 20: the fact that the correlation coefficient is the easiest to understand is a valid argument. However, it doesn't sound very good to state it here as the primary reason for choosing this metric against others. I would just remove this part of the sentence.
- Page 5, lines 23-24: is it one third of zeros or one sixth of ties over the entire hindcast period? Could you also please justify that?

### Section 3:

- For the results section of this paper, more credit should be given to other papers on seasonal hydrological forecasting in Europe, where appropriate. For example, (Chemere et al. (2016), Demirel et al. (2014), Svensson (2015), Trigo et al. (2004), among others; even if these papers do not contain an analysis for the integrity of Europe.
- Page 5, lines 26-30: this description was already made in the introduction. I would not repeat it here, especially since the results section titles are quite descriptive.

### Section 3.1:

- Page 5, lines 39-40: this is a very interesting remark!
- page 5, lines 32-40: how are those results different or similar to results for the other initialisation months?
- Page 6, lines 18-19: this figure does however not look at the persistence in skill, as a single cell could have skill for 3 months in a row for example, and another for 3 months but spaced, having the same colour on figure 3. It would be worth mentioning this in the figure caption.
- Page 6, lines 27-28: that is a very interesting results. Would it be possible to say why this is? Are cells in a specific region gaining skill or is it random noise?
- Page 6, lines 28-30: a result worth mentioning however, would be the lead time at which, on average, the domain-averaged  $R \leq 0$ .

### Section 3.2:

- Page 6, line 34: could you please add "(not shown)" at the end of the sentence finishing with "target months and lead times."?
- Page 7, line 8: could you please add the word "difference" after "average"?
- Page 7, lines 15-16: this is a good point!

### Section 3.3:

- Page 7, lines 23-24: was the same observed for other initialisation and target months? It would be good to mention this here.
- Page 7, lines 23-26: with this sample of stations, is it possible to say there are regions where the difference between theoretical and actual skill is highest?
- Page 7, lines 32-40: this paragraph describes methods and should therefore be moved to the methods of analysis section of this paper.
- Page 8, lines 1-3: what about basins with an AAPFD  $> 0.3$ ? they would probably show a higher difference between the two ratios.
- Page 8, line 10: could you please add the word "observation" before "stations"?
- Page 8, line 13: is the skill reduction between theoretical and actual skill or between lead time 0 and 2? The following sentence suggests that it is the latter but it is not clear from the sentence so it would be good to specify.
- Page 8, lines 13-17: this is very interesting!

### Section 3.4:

- Page 8, line 20: it would be good to specify that we are looking at runoff again here.
- Page 8, lines 20-21: did the other initialisation and target months show similar results? It would be good to mention this here.
- Page 8, line 23: I am not sure to understand the sentence "domain-averaged magnitude of the skill metrics". Could you please clarify what it meant?
- Page 8, lines 22-24: the patterns of skill are indeed similar. However, the magnitudes appear fairly different, even given the fact that they cannot be compared exactly due to the different colour bars used for plotting. The RPSS for example shows a lower skill on average than the other scores, while R shows a higher skill on average. This is also shown by the cell signal for each score. It would be worth noting this, and also in terms of the forecast attributes.
- Page 8, line 29: this can be done with the cell signal indicated on the top left corners of the plots.

- Page 8, lines 30-32: this is very interesting. So it indeed suggests that seasonal forecasts are anomaly forecasts, which is useful for decision-making! How are those numbers equal or different for other target and initialisation months?
- Page 8, line 35: I would rephrase the explanation of the PS here.
- Page 8, line 38: which results is this referring to? All the results presented in this section so far? Could you please specify it here or mention it little by little after each result?
- Page 8, lines 39-40: this is not true for all cases, but it is on average.
- Page 8, line 41: is the 1.0 here in terms of the area?

**Discussion:** the differences between the theoretical and the actual skill stated here are very interesting. However, the discussion would benefit greatly from further examples on how to improve the actual skill of seasonal hydrological forecasts (such as the recalibration idea given on page 10, lines 29-33).

- Page 9, lines 4-12: this should be moved to the methods, together with Figure 10. Then in the discussion you could refer back to these structural differences between the systems and state the questions that these differences raise.
- Page 9, line 34: could you please rephrase the sentence “In the real world a difference discharge observations differ from reality”? It is not clear to what is meant.
- Page 9: lines 34-36: this is an interesting point. However, I do not see how it will lead to more theoretical than actual skill. Indeed, the bias in the discharge measurements could potentially mean a closer simulated discharge from the model reference run to the biased discharge observations. In other words, we do not know how this measurement bias impacts the actual skill with regards to the theoretical skill.
- Page 9, lines 37-42: it would be clearer if you made this point number 4, even if this component on Figure 10 is not in red.
- Page 10, lines 4-12: this is a very good point! I would put it in the model hydrology box, so within point 1 on page 9, or as a sub-point of point 2.
- Page 10, line 13: could you please add (see Sect. 3.3) in parentheses at the end of this sentence?
- Page 10, line 17-18: I don't understand why this would be the case? The hindcasts would also benefit from the model optimisation as they are run with the same model as the reference run. The only difference between those two systems being the meteorological forcing data used to produce the hindcasts or pseudo-observations of discharge.
- Page 10, lines 14-28: I am not sure to understand the point that you are making here. The model is the same for the reference run and the hindcasts generation, hence, even if the model is optimised to reach closer discharge simulations to the actual discharge observations, both systems would benefit from this. In the examples that you give, the predictive skill gained from wrongly forecasting this too large amount of snow or soil moisture runoff or from rightfully forecasting lower snow or soil moisture runoff should be the same, unless the metric used to calculate skill is biased towards large values, such as the MAE, for example. So the problem here is rather the choice of the metric. In case I am missing something, could you please clarify this paragraph?

### Conclusions:

- Page 11, lines 9-10: please consider adding a “for example” here to show that the British Isles are an example amongst many results of the paper.
- Page 11, lines 10-11: is this true for all times?

- Page 11, line 19: I wouldn't mention the numbers in between parentheses in the conclusion. They are already in the results of the paper, where the readers can find them if they want to.
- Page 11, lines 21-22: I would write the Ranked Probability Skill Score as RPSS since ROC is also written as an abbreviation.
- Page 11, lines 22-23: could you please replace this sentence to "The skill in terms of the ROC area tends to be slightly larger for [...]?"

**Figure 1:** these are great plots!

- Could you please put a label on the side of the colour bar to indicate that this is R?
- Please state in the figure caption that red is better.
- Could you please specify that the legend is situated in the top left corner of each plot? This is a really good idea by the way!

**Figure 2:**

- Could you please put a label on the side of the colour bar to indicate that this is R?
- Even though the caption is given in Figure 1, I would repeat it here. Because it is easier to read directly under the figure than having to jump from a figure caption to the other figure.

**Figure 3:**

- Could you please put a label on the side of the colour bar to indicate that this is R?

**Figure 4:** this figure is great, I especially like the lead times, clever!



- Could this figure be made bigger?
- In order to make it easier to read for the readers, please consider adding a colour bar for the different initialisation months.

**Figure 5:**






- I would put the a, b, c and d above each plot.
- Wouldn't it be better and easier to see the differences between plots a and b if a plot of the difference between both maps was made instead?
- In the y-axis labels of plots c and d the word correlation coefficient can be replaced with R.
- Could you please add an x-axis label for plot c to say if these are the initialisation or target months?
- Plot d is not colour blind friendly as there is both red and green. Please consider changing one of the two colours.
- Here again I would repeat the necessary information of the caption of Figure 1 for the interpretation of this figure.
- It would be good to specify the amount of catchments in each bin for plot d. This could maybe explain the negative difference for bin 8 for lead times 2 and 4.
- Could you please put a label on the side of the colour bar for plots a and b, to indicate that this is R?

**Figure 6:**



- I would put the a, b, c, d and e above each plot.
- Why isn't there a plot for the "real observations" and all stations for May and lead time 2? It would be interesting to see I think

- Could you please put a label on the side of the colour bar to indicate that this 
- Could you remind the readers what the s of small and large basins are in the caption, as well as the number of stations for both categories?




#### Figure 7:

- Could you please put letters for  plots here: a and b?
- In the y-axis labels of both plots the word correlation coefficient can be replaced with 
- I would remove the y-axis label and the tick labels of the second figure as it is already stated in the figure on the t.
- Could you please add an x-axis label for both plots to say if these are the initialisation or target mo s?
- Could you please  remind the readers what the sizes of small and large basins are in the caption, as well as the number of stations for both categories?





#### Figure 8:


- I would put the a, b, c and d a  each plot.
- Because plot d does not show n  and this paper already contains many figures, would it be maybe better to remove plot d and mention it in the text only? Figure 9 could then replace plot d for example.

#### Figure 9:

- Could you please add an x-axis label for both plots to say if these are the initialisation or target mo s?
- Would it be worth  adding lines for the middle tercile in this same plot?
- Page 17, line 10: I think that the “mi  does not belong here.

#### Technical corrections:

- General:
  - Could you please only use of the two terms: “basins” or “catchments”?
  - Please consider changing “lead th” to “lead time”, which is more widely used, and will hence be clearer for the readers even without having read the methods section.
  - Could you please replace panel” with Fig. figure# subfigure#? E.g., for Figure 5, panel c would be replaced by Fig. 5c.
  - Could  please consider renaming the terms “pseudo-observations” and “real observations”? I would for example use “analysis” (as done in meteorology) or “simulations”, for the pseudo-observations, and simply “observations” for the “real observations”.
  - Cou  you please change “North” to “Northern”, “South” to “Southern”, “West” to “Western” and “East” to “Eastern” when in front of a country’s name?

 Page 1, line 10; page 11, line 3: please consider rephrasing the sentence “The present paper presents [...]” by removing one of the words “present”.

- Page 1, line 26: the terms “below normal” and “above normal” should not be written with capital letters, unless the abbreviations “BN” and “AN” are given in between parentheses just after.
- Page 1, line 29; page 2, lines 2 and 7; page 3, line 7: “e.g.” should be replaced with “, for example,”.



- Page 2, line 11: either “like” or “e.g.” should be used here.
- Page 2, line 12: please consider changing the word “earlier” by for example “previously”.
- Page 3, lines 8 and 9: please consider changing one of the two words “namely” to a synonym of this word.
- Page 3, line 10: “which is then used for” the “initialisation of the hindcasts”.
- Page 3, line 12: please remove the word “again”.
- Page 3, line 12: does “here” mean “hereafter as”?
- Page 3, lines 16-17: this should be moved to the references section of this paper and cited here.
- Page 3, line 30: please change “Though” to “Although”.
- Page 4, line 6: should the hyphen be removed between the words “large” and “basins”?
- Page 6, lines 34-35: please rephrase this sentence to “There are however subtle differences because rivers [...]”.
- Page 7, line 9: “the rate with which” instead of “the rate by which”.
- Page 7, line 38: a “;” should be added between “AAPFD” and “see Marchant and Hehir, 2002”.
- Page 7, line 39: this should be “AAPFD”. The D is missing.
- Page 7, line 42: the “So” is not needed here.
- Page 8, line 6: there should be a “;” between “R” and “theoretical” to clarify the sentence.
- Page 8, line 10: “can be blamed on” rather than “to”.
- Page 8, line 14: there is a “to” missing between “due” and “a combination”.
- Page 9, lines 24, 36 and 42: please remove the “to” between the words “than” and “actual”.
- Page 9, line 29: please put the “see the companion paper” between parentheses.
- Page 10, line 5: please put the “see Sect 2.2” between parentheses.
- Page 10, lines 5-6: please consider changing the second “differences” in the sentence to for example “disagreement”.
- Page 11, lines 3-4: please consider adding the word “while” between just after the comma, to link the two parts of the sentence.
- Page 11, line 5: would replacing “taking” with “against” make more sense here?
- Page 11, line 5: please consider replacing the “as” with “called”.