

**Reply on reviews on Greuell et al. Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo and real observations. HESS Discuss. doi: 10.5194/hess-2016-603**

Anonymous Referee #1 (hess-2016-603-RC1)

We are pleased with the generally very positive evaluation of our paper by RC1 and thank him/her for his/her detailed textual comments, many of which will be acted upon as suggested and will improve the readability of the paper. We adopt most of the suggested textual improvements and specified our action to every remark in the annotated report *hess-2016-603-RC1-author-reply.pdf*<sup>1</sup>. One point of discussion will be highlighted below.

Wrt to the use of the wording *pseudo- and real-observations* in the title and through-out the paper: we have thought about the wording used and are aware of the meteorological 'convention' to use the word *analysis* or *reanalysis* instead of pseudo observations. However, the methods used in meteorological (re)analysis nearly always involve some sort of assimilation of data, be it simple Newtonian nudging or more complex types of 4D variational analysis to adjust models states to observed values. That is not the case here. We simply simulate the hydrological state of a region by forcing the hydrological model with the 'best possible reconstruction' of near surface meteorology present at the start of our research. Moreover, the use of the word (re)analysis is not main stream in hydrology as shown on lines 71-74 page 3 of our paper. Since only 603-RC1 and 604-RC1, which we believe to be the same person (given similarities in style), and none of the other three reviewers make the point to change the wording, we will stick to our wording, while stressing the distinction between meteorological analysis and our type of analysis even more in support of this choice (in the same paragraph, p3 line 70 etc)

Referee #2 Christel Prudhomme (hess-2016-603-RC2)

We thank Christel for her more critical but very constructive review. Some of her remarks are in line with those of RC1, some are additional. Below we will discuss the main remarks, details can be found in *hess-2016-603-RC2-author-reply.pdf*.

A number of remarks have to do with the structure of the paper:

- RC2 requests a better description of the main findings of cited literature mostly in the introduction, but also in other parts of the paper, and how these influenced the objectives of our study. We now recognize that this indeed can, should and will be improved, together with some additional references as suggested by RC1.
- With RC1, RC2 suggests to move the first part of the discussion, the part explaining the present figure 10 to the Methodology section. We will do so.
- RC2 asks repeatedly for suggestions/recommendations for further analysis. We recognize this omission, but of course have thought about that rather extensively. We will add such suggestions where appropriate in the discussion

As a result of these 3 points, and some others, both the introduction and the discussion section will be largely rewritten.

Some remarks pertain more to the science of our analysis:

- RC2 asks for a better description of the deterministic performance of the model used (VIC) prior to its use in a probabilistic seasonal forecasting context. We will do so based on and referring to previously published work, both from our own group (Greuell et al. 2015, Haddeland et al., 2012; van Vliet et al. 2012) and from others. More in particular we will try to relate good and bad forecasting skill for certain regions/basins and seasons in Europe to previously identified strengths and weaknesses in VIC performance, i.e. strengths/weaknesses to reproduce historical river flows across Europe.
- This issue partially overlaps with the RC2 request to better analyse the potential relation between basin size and model hindcast skill. Without focussing on individual basins (which is one the directions for future work we'd like to take). We will prepare, present and discuss a graph

---

<sup>1</sup> unfortunately the page line numbering in the RC reports do not agree with *hess-2016-603-manuscript-version1.pdf*, sometimes it is not entirely clear which statement exactly is addressed by the reviewer

similar to the present Fig 5d, but then relating difference between actual and theoretical discharge skill to basin size. This will be a new piece of analysis leading to a yet unknown outcome. Thus we will also increase the relative importance of section 3.3 better justifying the title of this paper.

Altogether, we believe that by following most of the recommendations by both RC1 and RC2 we will be able to significantly improve the structure and readability of the paper, as well as improving the scientific quality by some additional analysis and especially much better 'embedding' in previous work, both our own and that of others. Finally, priori to resubmission we'll have a language check done by a native speaker.

## 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 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  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 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”.

## ***Interactive comment on “Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations” by Wouter Greuell et al.***

**C. Prudhomme (Referee)**

chrp@ceh.ac.uk

Received and published: 9 February 2017

General

The paper is the first of 2 companion papers on a pan-European seasonal streamflow forecasting system. This paper focuses on the verification of the re-forecast for a 30-year period (1981-2010).

Streamflow forecasting beyond medium range is still a relatively new area of research in Europe, and has received more attention in the past few years, following the availability of seasonal climate re-forecasts. Skilful hydrological forecasts at monthly to seasonal lead time would have great potential use in Europe as it would help planning

C1

and management of water resources for a huge variety of sectors including transportation, agriculture, public and domestic water supply or energy. Whilst the skill of dynamic rainfall forecasts is relatively limited at lead times over 10 days in temperate climates such as Europe, the existence hydrological memory due to catchment storage raised the question of potential higher skill in hydrological seasonal forecast than in its climate forcing data. As such, the paper addresses a topical subject with a large readership interest. I have however some concerns about some of the analyses undertaken here, detailed below. I hence suggest a major revision.

The streamflow forecasting system developed and used in this paper relies on two major sources of information and tool: 1) climate forcing data, here based on the ECMWF System 4 re-forecasts; and 2) a gridded hydrological model that transforms the weather signal into runoff and routed discharge. Inherent to any modelling exercise, simulations and re-forecasts are likely to be associated with bias and errors.

The authors run a gridded hydrological model forced by observed climate for 1 month, as spin-up to set-up initial conditions, and run the model with re-forecast climate forcing. They then evaluate the skill of the re-forecasts by comparing the results with 1) hydrological simulations forced by observed climate (runoff and routed discharge; called ‘theoretical skill’); 2) observed discharge (called ‘actual skill’). For actual skill, they use discharge time series from the GRDC and EWA database, and match the location of the river gauges with the routed network used in the model (at a  $0.5^\circ \times 0.5^\circ$  resolution, i.e.  $\sim 50$ km) so that gauged flows can be compared with the correct modelled discharge. Three metrics are used for the theoretical skill assessment, but most discussion is based on correlation coefficients, also applied to actual skill. The seasonal variation of the spatial distribution of the theoretical skill is described and compared for runoff and discharge, mainly for a 2-month lead time. Overall pan-European theoretical and actual skill compared for 2 classes of catchment size, and some causes of degradation between theoretical and actual skill discussed, but not formally tested.

Whilst the findings of pan-European hydrological seasonal forecasting skill are really

C2

relevant, I have some reservation regarding some methodological decisions and interpretations presented in the paper, detailed below.

- Actual skill analysis. The analysis must be better justified, and the discussion strengthened. Below are some points that need to be added to the paper:

o Is simulated discharge comparable to actual discharge? There is no data assimilation at the beginning of the forecast to reduce potential bias in the simulated discharge. So the hydrological re-forecasts include both hydrological modelling errors and climate forcing errors, without any attempt to reduce the former.

o Is the catchment matching exercise working? The hydrological model has a relatively coarse resolution, and a catchment area error of up to 15% (for large catchments) is deemed acceptable [the choice of this threshold should be justified]. For small catchments, there is no attempt to scale the discharge from the hydrological model scale to the gauged catchment scale. This could introduce some discrepancies between simulated and observed discharge. In fact p8 l3-4, the authors do state that '[the] small basins (...) are generally smaller than the spatial resolution of the simulations'

o Is the hydrological model performance influencing the actual skill results? Poor hydrological model performance introduce errors for both initial states and re-forecasts. One hypothesis is for 'actual skill' to be much lower for seasons and locations where the hydrological model is known not to reproduce well the hydrological processes. Comparison of hydrological model performance and actual skill is necessary for a meaningful interpretation of the results. This is only mentioned briefly in the discussion (2.5 lines) as second point (p9 l31-33). This should be the first point of the analysis when regarding actual skill.

- Re-forecast simulations

o Is the spin-up period long enough? It is not clear what actual spin up is used, with 1-month spin-up period suggested (p3 l29), but this sounds really short compared to

C3

expected storage in some parts of Europe (e.g. snow pack in high latitude/ high elevation and/or groundwater storage in large aquifers).

- General methodology

o How are the catchments classified as small/ large? There is no surface area mentioned, and not physical justification, but size is the only physical measure used to attempt explaining the difference between theoretical and actual skill

o What is the justification for the non-calculation of skill metrics? (p5l23-24). In particular, zero flow simulations can be extremely important to depict droughts. Why excluding them?

o How is a skilful forecast defined? (p5 l37-38; p6 l1-2) What is the threshold used to define a re-forecast as 'skilful'? Is this based on statistically significant test? Is it the value of 0.31 quoted in caption fig 1? This needs to be made clearer within the text

o Human influence analysis. This is fully based on the assumption that LPJmL has identified and reproduces accurately all the human interventions, and the derived Amended Annual Proportional Flow Deviator is a realistic representation of the degree of influence. This is a strong assumption that needs to be caveated in the text. This modelling exercise needs to be described in the methods section and not so late in the paper (p7 l34-36)

- Analysis/ interpretation

o Influence of catchment size on theoretical vs actual skill (p8 l4-17). I found the analysis difficult to follow, the paragraph confusing, and the language used is inappropriate 'apparent difference in (...) skill (...) can be blamed almost entirely to the geographical distribution of stations'. What does 'this results holds for the cells with observations' mean? Is the difference between 'large basins' skills (0.396) and 'small basins' skills (0.384) significant? Is this to be linked with the scale of the hydrological modelling? The analysis would be more thorough if conducted by looking at relationships with

C4

catchment sizes, rather than dividing the sample in 2 categories. It also needs to be linked with the model performance.

o Section 3.4 (p8). Is this conducted on pseudo observations? Why is this not after section 3.2? What is the implication of the findings? Can a physical explanation be given? Can the authors recommend skill metrics following their analysis?

o Discussion (p9-10). I found it unclear and difficult to follow, and some description of methods (model calibration technique) don't fit well (this should be in methods). The authors here describe some hypotheses for the difference between theoretical and actual skill: this should come at the beginning of the paper, and being tested within the study. Moreover, the analysis between theoretical and actual skill is short and not very thorough, yet is discussed at length; this does not reflect well the study. Some points are not clear (e.g. p9 l26-30; p9 l39-42)

o Statements not justified. There is a lack of evidence of the authors' claim that 'optimisation of the model system could, and would in many case, lead to a degradation of the theoretical skill'. What is the reason for that? What is the evidence? Have the authors conducted a sensitivity analysis? I agree that perfect theoretical skill does not adequate with perfect re-forecast, when main processes are not accounted for in the models. But the whole section needs careful re-wording, and better scientific justification, references, or suggestions for further analysis for verification of the hypotheses.

Main points of suggested improvement

Science

- There is no information on the hydrological model performance, albeit it is written to be 'on average across all basins considered, more or less in the middle ranking of the five models' [p3l39-40]. This is not enough and does not provide any information of the actual performance (it could be middle ranking of an ensemble with very low skill). Reference of a paper is not enough in this case. This is critically important when

C5

the re-forecast skills are compared with what the authors call real- observations, as it would be expected that lower hydrological modelling performance would result in lower skill in reproducing the real observations.

- There is not enough discussion on the role of initial conditions, hydrological memory and catchment storage that can bring predictability: catchment storage could include groundwater, lakes, and snow pack. At the very least, reference to some of the findings of part 2 could be made.

- There is a lot of discussion about the quality of measurements and their implication on lower actual skill, and much less on modelling error. I found this out of proportion.

- Current conclusion is a summary of the research. I would expect the discussion to be opened to future research and application.

- The reference to the companion paper (page 2) is very limited, and it is difficult to see the link between both. At least the conclusions could be brought in the discussion, rather than exposed in the introduction and not referred to later onto justify the writing up of the study in 2 parts.

Structure

- The title does reflect the bulk of the paper. The analysis of 'real-discharge' is only done in section 3.1 but of 4 analysis sections.

- The structure is not logic: 3.1, 3.2 and 3.4 all analyse the results in a 'pseudo-observations' [modelled] world whilst 3.3 looks at the results in 'real- observations' world.

- Description of the model set-up/ calibration is given in the discussion (p10 l29-33), but this should be in the methods section when the model is introduced

Other points

Science

C6

- The explanation of matching gauges locations with the 0.5 grid needs to be improved  
Structure/ description
- Introduction Most of the introduction is dedicated to the methods, data and tools used in the paper, and is not a review and discussion of the state of the art, with a judgment of the conclusions obtained from previous studies, and how to move forward. A typical example is p2 19-15, with a list of papers without any discussion, and a description of some of the analysis, and even a discussion of the results, which should not be in introduction. I found this very confusing. The whole section needs to be greatly improved, with a more traditional layout of state of the art, research gaps identified, and then at the end aims of the paper, without details of the methods and tools used.
- Section 3.1: Inconsistency in figure references; first sentence of page 6 does not describe what figure shows.
- Figure 3 is excellent.

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-603, 2016.

1 **Seasonal streamflow forecasts for Europe – I. Hindcast verification**  
2 **with pseudo- and real observations**

3

4 Wouter Greuell, Wietse H. P. Franssen, Hester Biemans and Ronald W. A. Hutjes

5

6 Wageningen University and Research

7

8 all authors:

9 Water Systems and Global Change (WSG) group, Wageningen University and  
10 Research, Droevendaalsesteeg 3, NL 6708 PB Wageningen, Netherlands

11

12 correspondence to [ronald.hutjes@wur.nl](mailto:ronald.hutjes@wur.nl)

13

14



15 **Abstract**

16

17 Seasonal predictions can be exploited among others to optimize hydropower energy  
18 generation, navigability of rivers and irrigation management to decrease crop yield  
19 losses. This paper is the first of two papers dealing with a model-based system built to  
20 produce seasonal hydrological forecasts (WUSHP: Wageningen University Seamless  
21 Hydrological Prediction system), applied here to Europe. The ~~present~~-paper presents  
22 the development and the skill evaluation of the system. In WUSHP hydrology is  
23 simulated by running the Variable Infiltration Capacity (VIC) hydrological model with  
24 forcing from bias-corrected output of ECMWF's Seasonal Forecasting System 4. The  
25 system is probabilistic. For the assessment of skill, we performed hindcast simulations  
26 (1981-2010) and a reference simulation, in which VIC was forced by gridded  
27 meteorological observations, to generate initial hydrological conditions for the  
28 hindcasts and discharge output for skill assessment (pseudo-observations). Skill in  
29 hindcasting runoff and discharge is analysed with monthly temporal resolution, up to 7  
30 months of lead time, for the entire annual cycle. Using the pseudo-observations and  
31 taking the correlation coefficient as metric, hot spots of significant skill in runoff were  
32 identified in Fennoscandia (from January to October), the ~~southern~~southern-part of the  
33 Mediterranean (from June to August), Poland, ~~North~~northern Germany, Romania and  
34 Bulgaria (mainly from November to January) and ~~West~~western France (from  
35 December to May). Generally skill decreases with increasing lead time, except in  
36 spring in regions with snow rich winters. The spatial pattern of skill is fading with  
37 increasing lead time but some skill is left at the end of the hindcasts (~~7 months~~). On  
38 average across the domain, skill in discharge is slightly higher than skill in runoff.  
39 This can be explained by the delay between runoff and discharge and the general  
40 tendency of decreasing skill with lead time. Theoretical skill as determined with the  
41 pseudo-observations was compared to actual skill as determined with real discharge  
42 observations from 747 stations. Actual skill is mostly and often substantially less than  
43 theoretical skill. This effect is stronger for small than for large basins,—which is  
44 consistent with a conceptual analysis of the structural differences between the two  
45 types of verification. Qualitatively, results are hardly sensitive to the different skill  
46 metrics considered in this study (correlation coefficient, ROC area and Ranked  
47 Probability Skill Score) but ROC areas tend to be slightly larger for the ~~B~~below  
48 ~~N~~normal than for the ~~A~~above ~~N~~normal tercile.

49

50

51

Formatted: Highlight

## 1 Introduction

Society may benefit from seasonal hydrological forecasts, i.e. hydrological forecasts for future time periods from more than two weeks up to about a year (Doblas-Reyes et al., 2013). Such predictions can e.g. be exploited to optimize hydropower energy generation (Hamlet et al. 2002), navigability of rivers in low flow conditions (Li, et al., 2008) and irrigation management- (Mushtaq et al. 2012; Ghile and Schulze 2008) to decrease crop yield losses. In order to be of any value in decision making processes of such sectors, forecasts must be credible, i.e. be skilful in predicting anomalous system states, as well as being relevant and legitimate to the decision making process (e.g. Bruno Soares and Dessai, 2016). In this paper we will introduce WUSHP (Wageningen University Seamless Hydrological Prediction system), a dynamical\_ (i.e. model-based) system (see Yuan et al., 2015) that was built around the Variable Infiltration Capacity (VIC) hydrological model and ECMWF's Seasonal Forecast System 4. to produce seasonal hydrological forecasts. It will be applied to Europe. The usefulness of the system depends partially on the level of its skill and the paper will therefore describe the system and then focus on an extensive assessment the determination of WUSHP its skill. The usual method of assessing skill of predictive systems is by analysing hindcasts, a strategy that will be adopted here as well.

~~It is quite common in seasonal hydrological forecasting (e.g. Shukla and Lettenmaier, 2011, Singla et al., 2012, Mo and Lettenmaier, 2014, and Thober et al., 2015) but also in medium range forecasting (Alfieri et al., 2014) to determine prediction skill by comparing the hindcasts with the output from a reference simulation. A reference simulation is a simulation made with the same hydrological model as the hindcasts, except that the forcing is taken from meteorological observations or from a gridded version of meteorological observations. The reference simulation can best be regarded as a simulation that attempts to make a best estimate of the true conditions (in terms of e.g. discharge, soil moisture and evapotranspiration), using the modelling system. We will refer to the output of such a reference simulation as "pseudo observations" ("true discharge" in Bierkens and Van Beek, 2009; "synthetic truth" in Shukla and Lettenmaier, 2011; "reanalysis" in Singla et al., 2012; "a posteriori estimates" in Shukla et al., 2014). Pseudo observations have the advantages of being complete in the spatial and the temporal domain and to be available for all model variables. Also, they are suitable for the quantification of small sensitivities, e.g. to bias correction of the meteorological forcing, which would be hard to detect with real observations.~~

~~The downside of pseudo observations is, of course, that they are not equal to real observations. In this paper we will determine the performance of the prediction system not only with pseudo observations but also with real observations of discharge (like e.g. Koster et al., 2010, and Yuan et al., 2013) and compare the skill found with the two different approaches ("theoretical and actual skill", according to Van Dijk et al., 2013), which was earlier done by Bierkens and Van Beek (2009) and Van Dijk et al. (2013). Also, we will analyse conceptual differences between using pseudo and real~~

Formatted: Default, Line spacing: Multiple 1.15 li

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman, 12 pt

96 ~~observations for verification. We will argue that the fact that the pseudo-observations~~  
97 ~~are obtained with the same model as the hindcasts logically contributes to an~~  
98 ~~overestimation of the skill when the pseudo-observations are used for verification.~~

Formatted: Justified

100 During recent years, a number of systems for seasonal hydrological forecasts have  
101 been developed. Examples are the forecasting model suite ~~for France described by~~  
102 ~~Céron et al. (2010)~~, the University of Washington's Surface Water Monitor (SWM;  
103 Wood and Lettenmaier, 2006) and the African Drought Monitor (Sheffield et al.,  
104 2014).

106 Seasonal hydrological forecast systems for the entire continent of Europe are scarce  
107 (Bierkens and van Beek, 2009; Thober et al., 2015), but a few more concentrate on  
108 smaller domains such as the British Isles (Svensson et al., 2015), Iberia (Trigo, 2004)  
109 or France (Céron et al., 2010; Singla et al., 2012).

111 Thober et al. (2015) forced a mesoscale hydrological model (mHM) with  
112 meteorological hindcasts of the ~~North-North~~ American Multi-Model Ensemble  
113 (NMME) to investigate the predictability of soil moisture in continental Europe  
114 (excluding the British Isles and Fennoscandia. Evaluating a number of forecasting  
115 techniques that produced distinctive variations in the magnitude of skill, they found  
116 that spatial patterns in skill were remarkably similar among each other, as well as  
117 compared to the autocorrelation (persistence) of reference soil moisture. High skill  
118 was found in eastern Germany and Poland, Romania, southern Balkans and eastern  
119 Ukraine as well as north-western France. ~~Less~~ skill was found in the mountainous  
120 areas of Alps and Pyrenees, the northern Adriatic and Atlantic Iberia. Most skill was  
121 found for winter months (DJF), least for autumn (SON), this minimum shifting to  
122 summer (JJA) at long lead times (6 months).

Commented [RH1]:

Commented [RH2R1]: Fennoscandia , british isle not evaluated by Thober2015

124 Bierkens and van Beek (2009) developed an analogue events method to select annual  
125 ERA40 meteorological forcings on the basis of annual SST anomalies in the ~~North~~  
126 northern Atlantic and then made hydrological forecasts with a global-scale  
127 hydrological model applied to Europe. Evaluating only winter and summer half year  
128 aggregated skill, they found wintertime skill in large parts of Europe with maxima in  
129 eastern Spain and a zone from southern Balkans and Romania through eastern Poland  
130 and the western Russia, the Baltic states and Finland. Summertime skill was about  
131 50% lower, and even more around the Alps and Adriatic. NAO based climate forecast  
132 added significant skill only in limited areas, such as Scandinavia, the Iberian  
133 Peninsula, the Balkans, and around the Black Sea.

135 Svensson et al. (2015) found skilful winter river flow forecasts across the whole of the  
136 UK due to a combination of skilful winter rainfall forecasts for the north and west, and  
137 strong persistence of initial hydrological conditions in the south and east. Strong  
138 statistical correlations between NAO and winter precipitation in Iberia lead to skilful  
139 forecasts of JFM river flow and hydropower production (Trigo et al., 2004). Ceron et

140 al. (2010) and Singla et al. (2012) set up a high resolution river flow forecasting  
141 system (8 km) over France, for which seasonal climate forecast improved MAM skill  
142 over northern France, but worsened it over southern France (compared to a river flow  
143 model with proper initialisation of soil moisture, snow etc., but random atmospheric  
144 forcing). Demirel et al. (2015) found that both two physical models and one neural  
145 network over-predict runoff during low-flow periods using ensemble seasonal  
146 meteorological forcing for the Moselle basin, and as a result more extreme low flows  
147 are less reliable than more moderate ones.

148  
149 It is quite common in seasonal hydrological forecasting (e.g. Shukla and Lettenmaier,  
150 2011, Singla et al., 2012, Mo and Lettenmaier, 2014, and Thober et al., 2015) but also  
151 in medium range forecasting (Alfieri et al., 2014) to determine prediction skill by  
152 comparing the hindcasts with the output from a reference simulation. A reference  
153 simulation is a simulation made with the same hydrological model as the hindcasts,  
154 except that the forcing is taken from meteorological observations or from a gridded  
155 version of meteorological observations. The reference simulation can best be regarded  
156 as a simulation that attempts to make a best estimate of the true conditions (in terms of  
157 e.g. discharge, soil moisture and evapotranspiration), using the modelling system. We  
158 will refer to the output of such a reference simulation as “pseudo-observations”  
159 (misleadingly named “true discharge” in Bierkens and Van Beek, 2009; more  
160 appropriately “synthetic truth” in Shukla and Lettenmaier, 2011; “reanalysis” in Singla  
161 et al., 2012; “a posteriori estimates” in Shukla et al., 2014). We prefer the term  
162 “pseudo-observations” over “re-analysis” since the latter has a meteorological  
163 connotation that often implies the use of some form of (variational) data assimilation.  
164 We did not attempt any form of assimilating observed hydrological variables, such as  
165 discharge, in our reference run.

166  
167 Pseudo-observations have the important advantages of being complete in the spatial  
168 and the temporal domain and to be available for all model variables. Also, they are  
169 suitable for the quantification of small sensitivities, e.g. to bias correction of the  
170 meteorological forcing, which would be hard to detect with real observations. Finally,  
171 assessment of skill based on pseudo observations excludes model errors from the  
172 analysis, which is especially useful when addressing various sources of skill (Wood et  
173 al., 2016), something we will do in the companion paper.

174  
175 The downside of pseudo-observations is, of course, that they are not equal to real  
176 observations. In this paper we will determine the performance of the prediction system  
177 not only with pseudo-observations, but also with real observations of discharge (like  
178 e.g. Koster et al., 2010, and Yuan et al., 2013) and compare the skill found with the  
179 two different approaches (“theoretical and actual skill”, according to Van Dijk et al.,  
180 2013), which was previously done by Bierkens and Van Beek (2009) and Van Dijk et  
181 al. (2013). We will analyse and discuss conceptual differences between using pseudo-  
182 and real observations for verification. We hypothesise that the fact that the pseudo-

183 observations are obtained with the same model as the hindcasts logically contributes to  
184 an overestimation of the skill when the pseudo-observations are used for verification.

185  
186 ~~The hydrological hindcasts are produced by WUSHP by running the Variable~~  
187 ~~Infiltration Capacity (VIC) hydrological model using bias corrected output of~~  
188 ~~hindcasts from ECMWF's Seasonal Forecast System 4 as meteorological forcing.~~  
189 ~~The system is probabilistic. In addition, a reference simulation is carried out, in which~~  
190 ~~VIC is forced by gridded meteorological observations (WATCH Forcing Data Era-~~  
191 ~~Interim, i.e. WFDEI), with the aims of generating pseudo-observations and initial~~  
192 ~~hydrological conditions. Details about WUSHP are provided in Sect. 2.~~

Commented [RH3]: ..in this paper..

Commented [RH4]: ..for the Hindcast generation.

193  
194 This paper aims to analyse to what extent WUSHP is able to predict runoff and  
195 discharge in Europe over the full annual cycle and for lead times up to 7 months. We  
196 aim to assess skill at maximum resolution, i.e. at monthly resolution instead of  
197 seasonal or semi-annual aggregates. Where many studies use correlation coefficient as  
198 main skill metric we will assess skill also for the more probabilistic metrics ROC area  
199 and RPSS (see section 2.3). The second aim is to get a better understanding of the  
200 effects of using pseudo-observations, as opposed to using actual observations, for the  
201 verification of hindcasts. In the next section we describe the concept and details of our  
202 modelling (Sect. 2.1) and analysis approach (2.2 and 2.3). We will start the result  
203 section by assessing theoretical skill of the runoff hindcasts (Sect. 3.1) and then  
204 proceed to theoretical skill of the discharge hindcasts and a comparison between  
205 theoretical skill of discharge and runoff in (Sect. 3.2). Differences between theoretical  
206 and actual skill of discharge will be presented ~~using our data (Sect. 3.3)~~ followed  
207 by an analysis of differences in skill when comparing various metrics in Sect. 3.4.  
208 The discussion starts with a conceptual analysis of reasons for differences in actual  
209 and theoretical skill (Sect. 4.1), followed by a discussion of uncertainties (Sect. 4.2)  
210 and implications (4.3). Additional figures are published in a supplement of this paper.

211  
212 In a companion paper (Greuell et al., 2017<sup>6</sup>) we analyse the reasons for the presence  
213 or source of skill and the lack of skill discussed in the present paper, using two  
214 different methods. Firstly, skill in the forcing and other directly related hydrological  
215 variables— like evapotranspiration are analysed. Secondly, a number of Ensemble  
216 Streamflow Prediction (ESP) and reverse-ESP experiments, which isolate different  
217 causes of predictability, are discussed. In the present results and discussion sections  
218 we will occasionally look forward to the identified causes of skill. ~~The main~~  
219 ~~conclusions from the companion paper are that, in Europe, a) skill beyond the first~~  
220 ~~lead month is almost exclusively caused by initial hydrological conditions and not by~~  
221 ~~skill in the meteorological predictions and b) at most times and locations the initial~~  
222 ~~state of soil moisture contributes more to skill than the initial state of snow.~~

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

## 224 **2 System, models, data and methods of analysis**

225 To assess the forecast quality of our system, two approaches ~~For verification of the~~  
226 hindcasts two options were ~~are~~ used/considered in this paper. First, ~~We~~

227 ~~determin~~determined the skill of the hindcasts by comparing predicted discharge with  
228 the output of a ~~the~~ reference simulation (the “pseudo-observations” leading to  
229 “theoretical skill”), ~~allowing evaluation continuous in space and time.~~ Secondly, we  
230 ~~quantify skill with respect to— and with~~ observations of real discharge (“real  
231 observations” leading to “actual skill”), ~~allowing evaluation at a limited number of~~  
232 ~~locations (discharge stations) on the river network only.~~ ~~To obtain a basis for~~  
233 ~~understanding the differences in skill that we found.~~ Fig. 10 presents a streamflow  
234 diagram of the three relevant ~~physical~~ systems, namely the real world and the two  
235 model systems that generate the hindcasts and the pseudo-observations ~~respectively.~~  
236 In each system, confined in the diagram by a box, meteorological and initial  
237 conditions force and initialize hydrology, of which discharge is the relevant  
238 component here. There are ~~two~~three complications ~~when interpreting and comparing~~  
239 ~~actual and theoretical skill.~~ First, the initial conditions ~~themselves~~are generated by  
240 meteorological forcing during the spin up period, initial conditions at the beginning of  
241 the spin up period and hydrology. This is represented by the upper left branch in each  
242 box, omitting initial conditions at the beginning of the spin up period for simplicity.  
243 ~~Second, due to measurement errors real observations of discharge generally differ~~  
244 ~~from real discharge (Juston et al., 2014) due to unavoidable measurement errors as~~  
245 ~~illustrated in the upper right corner of the figure.~~ Third, obviously a difference exist  
246 between real hydrology and model hydrology, central in each box. Since the  
247 hindcasted discharge and pseudo observations share the same model hydrology ~~and~~  
248 the same initial conditions ~~and~~ both are free from any observational errors, theoretical  
249 skill will always be higher than actual skill.

250  
251 For now we simply accept, and even stress this a-priori ‘superiority’ of theoretical  
252 over actual skill. In the discussion section we will come back to this and further  
253 discuss, at least in qualitative terms, how each of the differences between the three  
254 systems affect skill assessment.

255  
256 In the following subsections we will describe each component.

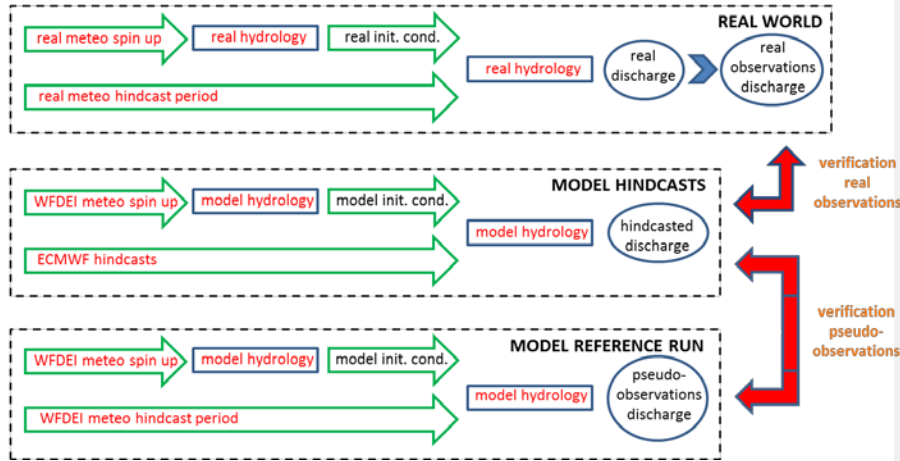


Figure 10: Diagram illustrating the conceptual setup of the present study, showing differences between verification of hindcasts (in the middle) with pseudo observations (bottom) and with observations of real discharge (top). See the text in this section and in section 4 for a detailed further explanation.

## 2.1 The hindcasts and the reference simulation

We will here describe the version of the WUSHP that has been used to generate the hindcasts for the European continent. WUSHP consists of two simulation branches; namely a single reference simulation and the hindcasts themselves. In both branches, terrestrial hydrology is simulated with the Variable Infiltration Capacity model (VIC, see Liang et al., 1994), which runs on a domain extending from 25° W to 40° E and from 35° to 72° N, including 5200 land based cells of 0.5° x 0.5° (see maps in e.g. Fig. 1). VIC is forced by a gridded data set of daily meteorological data. VIC is run in so-called 'energy balance mode' which requires resolving the diurnal cycle. Therefore, internally the model temporally disaggregates the daily input to 3-hourly data and runs at 3 hourly time step. Output of all variables is again at daily resolution. Because snow may contribute significantly to the seasonal predictability of other hydrological variables, VIC was run with the option of subgrid elevation bands. This means that for each gridcell calculations were carried out at up to 16 different elevations, with the aim of simulating the elevation gradient of snow. VIC was run in naturalised flow mode, i.e. river regulation, irrigation and other anthropogenic influences were not considered.

In the reference simulation VIC is forced by a gridded data set of meteorological observations, namely the WATCH Forcing Data Era-Interim (WFDEI; Weedon et al.,



290 2014)- for the period of 1979-2010, of which the first two years were used to spin up  
291 the states of snow, soil moisture and discharge, and not used in further analysis. The  
292 reference simulation has ~~the~~ dual aim, ~~namely~~ to create the pseudo-observations for  
293 verification purposes and to create a best estimate of the temporally varying model  
294 state, which is then used for the initialisation of the hindcasts.

295  
296 The second branch, the hindcasts, consists of three steps. Seasonal predictions of daily  
297 meteorological variables are taken from ECMWF's Seasonal Forecast System 4 (S4  
298 hereafter). These are then ~~corrected~~ bias-corrected for bias using WFDEI ~~again, here~~  
299 as the reference data set. Finally, VIC is run with the bias-corrected S4 hindcasts as  
300 forcing, taking initial states from the reference simulation. ~~The whole system is~~  
301 probabilistic.

302  
303 The S4 hindcasts used in the present study include 15 members, cover the period from  
304 1981 to 2010 and consist of 7 month simulations initialised on the first day of every  
305 month (see Molteni et al., 2011 and ECMWF Seasonal Forecast User Guide,  
306 online[http://www.ecmwf.int/en/forecasts/documentation-and-support/long-  
307 range/seasonal-forecast-documentation/user-guide/introduction](http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/introduction)). The S4 ensemble is  
308 constructed by combining a 5-member ensemble analysis of the ocean initial state with  
309 SST perturbations of that state and with activation of stochastic physics. The whole  
310 system is thus probabilistic.

311  
312 The variables taken from the S4 hindcasts are daily values of precipitation, minimum  
313 and maximum temperature, atmospheric humidity, wind speed and incoming short-  
314 and long wave radiation, since these are all needed to force VIC. All of these variables  
315 were regridded with bi-linear interpolation from the 0.75 x 0.75° lat-lon grid of the S4  
316 hindcasts to a 0.5° x 0.5° grid. Since bias correction generally improves forecasting  
317 skill.~~Next~~, the quantile mapping method of Themeßl et al. (2011) was applied to bias-  
318 correct the forcing variables, taking the WFDEI as reference. For each variable and  
319 grid cell, 84 correction functions were established and applied by separating the data  
320 according to target month (12) and lead month (7). Such empirical distribution  
321 mapping of daily values has been successful in improving especially forecast  
322 reliability (rather than sharpness and accuracy; Crochemore et al., 2016).

323  
324 VIC was run for the period of the S4 hindcasts (1981 – 2010). ~~and in a~~ Additionally,  
325 for spin-up periods. In for the reference simulation two extra years (1979 – 1980) were  
326 simulated to spin up the states of snow, soil moisture and discharge. The hindcast  
327 simulations were initialised with states of soil moisture and snow from the reference  
328 simulation, so for these variables spin up was not needed. However, due to the set-up  
329 of the routing module of VIC, the state of discharge could not be saved and loaded.  
330 Hence to spin up discharge, each 7-month hindcast simulation was preceded by ~~a~~ one  
331 month simulation with WFDEI forcing. Simulations were performed on a 0.5° x 0.5°  
332 grid for all 15 members of the bias corrected S4 hindcasts. Though the forcing  
333 consisted of daily values, the simulations were done with a three hourly time step.



334 ~~Because snow may contribute significantly to the seasonal predictability of other~~  
335 ~~hydrological variables, VIC was run with the option of elevation bands. This means~~  
336 ~~that for each cell calculations were carried out at up to 16 different elevations, with the~~  
337 ~~aim of simulating the elevational gradient of snow. Since the hindcasts cover 30 years~~  
338 ~~with 12 ~~dates of~~ initialisation ~~dates~~ each and consist of 15 members, a total of 5400~~  
339 ~~hindcast simulations was carried out. VIC was run in naturalised flow mode meaning~~  
340 ~~that river regulation, irrigation and other anthropogenic influences were not~~  
341 ~~considered.~~

342  
343  
344 Simulations of historic discharge made with VIC (and four other hydrological models)  
345 were validated with observations from large European rivers by Greuell et al. (2015)  
346 and Roudier et al. (2016). ~~For making seasonal predictions the most interesting results~~  
347 ~~of that validation study are the skills of simulating interannual variability and the~~  
348 ~~annual cycle. In both aspects VIC performed, on average across all basins considered,~~  
349 ~~more or less in the middle of the ranking of the five models. VIC exhibits a fairly~~  
350 ~~small average bias (across 46 stations) of +23 mm/yr (=7%) and overall differentiates~~  
351 ~~well between low and high runoff basins with a spatial correlation coefficient of 0.955.~~  
352 ~~However, specific discharge was overestimated in the Mediterranean and under~~  
353 ~~estimated in northern Fennoscandia. Annual cycles are fairly well reproduced across~~  
354 ~~Europe, though VIC somewhat overestimates its amplitude. In northern Fennoscandia~~  
355 ~~the spring peak is too late and too long. Annual cycles of rainfed rivers are best~~  
356 ~~reproduced (central Europe) while also those for rivers with significant snow~~  
357 ~~dynamics are good (Alps). However, the annual cycle in basins with strong soil~~  
358 ~~freezing dynamics (northern Fennoscandia) or strong damping of discharge amplitudes~~  
359 ~~by large lakes (southern Finland) is more poorly reproduced.~~

Commented [RH5]: hier of omhoog?

360  
361 Perhaps more relevant in the present context is the model's capability to reproduce  
362 inter-annual variations in discharge. The standard deviation of simulated ( $\sigma_m$ ) annual  
363 discharge was 9% higher than observed ( $\sigma_o$ ) and the correlation between the two  
364 0.935. Like most models, VIC is better in simulating high flows (95 percentile: Q95)  
365 than low flows (Q5); the first is slightly overestimated, the second more seriously  
366 underestimated. The inter-annual variation in Q5 is overestimated in central Europe  
367 and the Alps, but underestimated in Fennoscandia (overall correlation across Europe  
368 0.40). The inter-annual variation in Q95 shows no clear spatial pattern and the overall  
369 correlation is 0.7.

370  
371 All validation results discussed in these two paragraphs are for the VIC model forced  
372 by E-obs (v9, Haylock et al. 2008). Our forcing, WFDEI shows higher precipitation  
373 (+104 mm/yr) across most of Europe, except the Alps, Scotland and western most  
374 Norway. This leads to higher mean discharge, higher inter annual variability and  
375 higher Q95 (not Q5) of simulated discharge for almost all stations.

378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420

## 2.2 Discharge observations

For the assessment of skill with real discharge observations, two data sets were acquired from the Global Runoff Data Centre, 56068 Koblenz, Germany (GRDC); namely the GRDC data set proper and the European Water Archive (EWA) data set. ~~These data sets do not include any variable or parameter characterising the human impact.~~ We ~~converted~~ mapped these two station data sets ointo the VIC grid ~~two gridded versions~~ with a resolution of  $0.5^\circ \times 0.5^\circ$  and a time step of a month. ~~The first contained only observations for catchments larger than 9900 km<sup>2</sup> (“large basins”). The second contained only observations for catchments smaller than the area of the grid cells (“small basins”). The subdivision enabled to investigate the effect of catchment size on skill.~~

Initially, in many cases the location of observation stations did not match with the corresponding river in the digital river network used in the routing calculations (DDM30, see Döll and Lehner, 2002). We corrected for this issue by matching the observations with the simulations by means of catchmentbasin size. The size of ~~the~~ model catchmentbasins (“model catchmentbasin area”) was determined by the DDM30 network. The size of the catchmentbasins upstream of the observation station (“station catchmentbasin area”) was taken from the meta data of the observations. The mapping procedure varied slightly with the size of the basins grouped in two classes. The first comprised ~~contained only observations for catchmentbasins larger than 9900 km<sup>2</sup> (“large basins”).~~ ~~The second contained only observations for catchmentbasins smaller than the area of the grid cells, i.e. smaller than about 2530 km<sup>2</sup> in southern Europe (at 35°N) or < 1050 km<sup>2</sup> at 70°N (“small basins”).~~ This subdivision was also used ~~enabled to investigate the effect of catchmentbasin size on skill.~~

First the station catchmentbasin area was compared to the model catchmentbasin area of the cell that is nearest to the station (“nearest model cell catchmentbasin area”).

For large basins we then proceeded as follows:

- If the station and the nearest model cell catchmentbasin area differed by less than 15%, the observations were matched with the model calculations for the nearest model cell.
- Otherwise, the station catchmentbasin area was compared with the model catchmentbasin area of the eight cells surrounding the nearest model cell.
- The minimum of the eight differences was determined.
- If that minimum was less than 15%, the simulations for the corresponding cell were matched with the observations.
- Otherwise, the station was discarded.

For small basins we proceeded as follows:

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Formatted: Highlight

Commented [RH6]: Any rationale for this number (RC2 comment)

Formatted: Highlight

- 421 - If the nearest model cell did not have an influx from any of the neighbouring cells,
- 422 its simulations were matched with the observations.
- 423 - Otherwise, all of the eight neighbouring cells without influx were selected.
- 424 - Their simulations were averaged and matched with the observations.

425  
 426 We further discarded all observations with less than 21 years of data within the  
 427 simulation period (1981-2010) for any of the months of the year. The final data sets  
 428 within our European domain contained 111 cells with observations for large basins  
 429 and 636 cells with observations for ~~small~~ basins smaller than a model gridcell.

430  
 431 These data sets do not include any variable or parameter characterising the level of  
 432 human impact. To enable analysis of the effect of anthropogenic flow modifications  
 433 on predictive skill, we ~~We~~ quantified the human impact by performing two model  
 434 simulations with the Lund-Potsdam-Jena managed Land (LPJmL) model (Rost et al.,  
 435 Schaphoff et al., 2013). This model that was operated at the same spatial resolution  
 436 (0.5° x 0.5°) and with the same river network (DDM30) as VIC, but the former does  
 437 include dams (GRanD database; Lehner et al., 2011) and associated reservoir  
 438 management. From the discharge output of a naturalized run and a run with reservoir  
 439 operation and irrigation, the human impact at cell level was quantified by computing  
 440 the so-called Amended Annual Proportional Flow Deviator (AAPFD;- see Marchant  
 441 and Hehir, 2002). Subsequently, we selected all discharge observations for large  
 442 basins with an AAPFD < 0.3, i.e. basins with a relatively small degree of human  
 443 impact (about half of all 111 basins).

Commented [RH7]: How many in our domain?

### 444 2.3 Methods of analysis

445  
 446 From the model output, consisting of daily means, monthly mean values were  
 447 computed, which were then used for the analysis. The analysis is restricted to runoff,  
 448 defined here as the amount of water leaving the model soil either along the surface or  
 449 at the bottom, and discharge, defined here as the flow of water through the largest  
 450 river in each grid cell. Discharge accumulates all runoff from cells that are upstream in  
 451 the model river network, with delays due to transport inside cells and through the river  
 452 network. Hence, whereas runoff represents only local hydrological processes,  
 453 discharge aggregates hydrological processes occurring in the entire ~~upstream~~  
 454 catchmentbasin- upstream of a particular cell.

455  
 456 Instead of analysing skill per target season and/or for a number of consecutive lead  
 457 months, we analysed skill for every combination of per 12 target and per 7 lead  
 458 months. The thus achieved higher temporal resolution of the skill metrics enables a  
 459 more accurate determination of the beginning and end of periods of skill. Moreover,  
 460 skill at a monthly resolution provides the possibility to determine the consistency of  
 461 the skill where we define consistent skill as skill that persists during at least ~~two~~  
 462 consecutive target or lead months. In accordance with Hagedorn et al. (2005) we  
 463 designated the first month of the hindcasts as lead month zero, so target month number  
 464

465 is equal to the number of the month of initialisation plus the lead month number. ~~In~~  
466 ~~discussing the results we will pay relatively little attention to lead month zero because~~  
467 ~~seasonal prediction deals with forecasts beyond the first two weeks.~~

468  
469 Three skill metrics (see Mason and Stephensen, 2008, for a good discussion of the  
470 why and how of these) were computed; ~~namely~~ i) the correlation coefficient between  
471 the observations and the *median* values of the ~~simulations–hindcasts~~ (shortly  
472 “correlation coefficient” or R), ii) the area beneath the Relative Operating  
473 Characteristics (ROC) ~~graph–curve~~ (shortly “ROC area”) and iii) the Ranked  
474 Probability Skill Score (RPSS).— The ROC area is computed for each month  
475 separately and for three categories of the (pseudo and real) observations and hindcasts  
476 with an equal number of values, with the categories containing the one third highest,  
477 ~~lowest~~ west and the remaining values (upper, lower and middle tercile, resp.; “above”,  
478 “below” and “near-normal”, AN, BN and NN categories), ~~respectively~~. The same  
479 subdivision of observations and hindcasts in terciles was made to compute the RPSS.  
480 Since none of these metrics is sensitive to systematic biases in the forecasting system,  
481 no attempt was made to correct simulated runoff or discharge for any such errors prior  
482 to computing the skill metrics, e.g. by scaling simulated discharge with the ratio of  
483 real world basin area over model world basin area. So we focus our evaluation on the  
484 models capability to predict river flow anomalies rather than absolute rover flows.

485  
486 All three skill metrics quantify, though in different ways, how well the ranking of the  
487 ~~annual~~ hindcasts matches the ranking of the observations. The correlation coefficient  
488 is a measure of the association between (pseudo-) observation and forecast ensemble  
489 median; we used the Pearson correlation coefficient. The ROC area is a measure of  
490 resolution or discrimination and indicates whether the forecast probability of an event  
491 (i.e. value falling in the considered tercile) is higher when such an event occurs  
492 compared to when not. The RPSS is a measure of accuracy and summarizes in a single  
493 number the skill of a forecast system to make ~~correct~~ forecasts ~~of~~ with the correct  
494 percentage of ensemble members events falling in any of the defined terciles. Perfect  
495 forecasts have values of 1 for all three skill metrics. Climatological forecasts  
496 (probabilistic forecasts that ~~are identical in our case~~ each year predict a 0.33 chance of  
497 a high or low anomaly occurring) lead to values of 0 for R, 0.5 for the ROC area and 0  
498 for the RPSS. Random forecasts were used to determine the significance of the  
499 metrics. In the case of the ~~Ranked Probability Score (RPSS)~~, these random forecasts  
500 were generated by sampling randomly from the multinomial distribution with  $p = (1/3,$   
501  $1/3, 1/3)$  and  $N = 15$  (the number of ensemble members), which is the distribution of  
502 climatological ensemble forecasts. Each metric will be designated as significant for p-  
503 values less than 0.05. This implies association is significant for  $R > 0.31$ , resolution is  
504 significant for ROC area  $> 0.69$  and accuracy is significant for  $RPSS > 0$ .

505  
506 To a large extent, we found that our results and conclusions, in terms of spatio  
507 temporal patterns of skill, are independent of the chosen metric. Hence, and because  
508 among the three metrics the correlation coefficient is the easiest to understand, we will

509 discuss results mostly in terms of the correlation coefficient, which is in line with  
510 Doblus-Reyes et al. (2013). The sensitivity to the chosen metric and significant  
511 differences between these metrics will be discussed in Sect. 4.4.2.

512  
513 All metrics were computed using the low and high level R packages  
514 “SpecsVerification” (Siegert et al., 2014) and “easyVerification” (Bhend et al., 2016)  
515 respectively. Metrics ~~cannot~~will not be computed if observations or hindcasts within  
516 the entire 30 year period consist for more than one third of zeros or one sixth of ties  
517 (i.e. equal values). Such skill gaps (i.e. the white terrestrial cells in Fig 1 and 2) only  
518 occur in the far North due to rivers that are frozen for at least a month in winter.

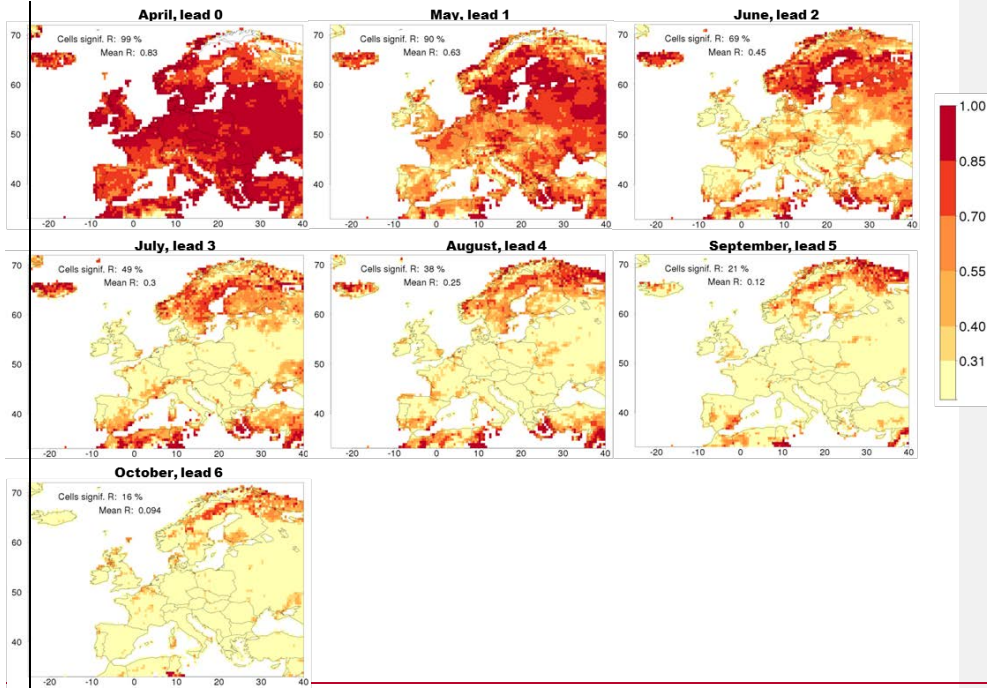
### 521 3 Results

522  
523 ~~In this section we present the skill of monthly mean values of hindcasted runoff and~~  
524 ~~discharge. First, skill as determined with the pseudo-observations is discussed, starting~~  
525 ~~with runoff (Sect. 3.1) and then continuing with a comparison between runoff and~~  
526 ~~discharge (Sect. 3.2). Next, Sect. 3.3 analysis differences in skill found by using~~  
527 ~~pseudo- and real observations for verification. In the first three sub-sections skill is~~  
528 ~~measured in terms of— the correlation coefficient between the observations and the~~  
529 ~~median values of the simulations (R). Section 3.4 deals with results for other skill~~  
530 ~~metrics.~~

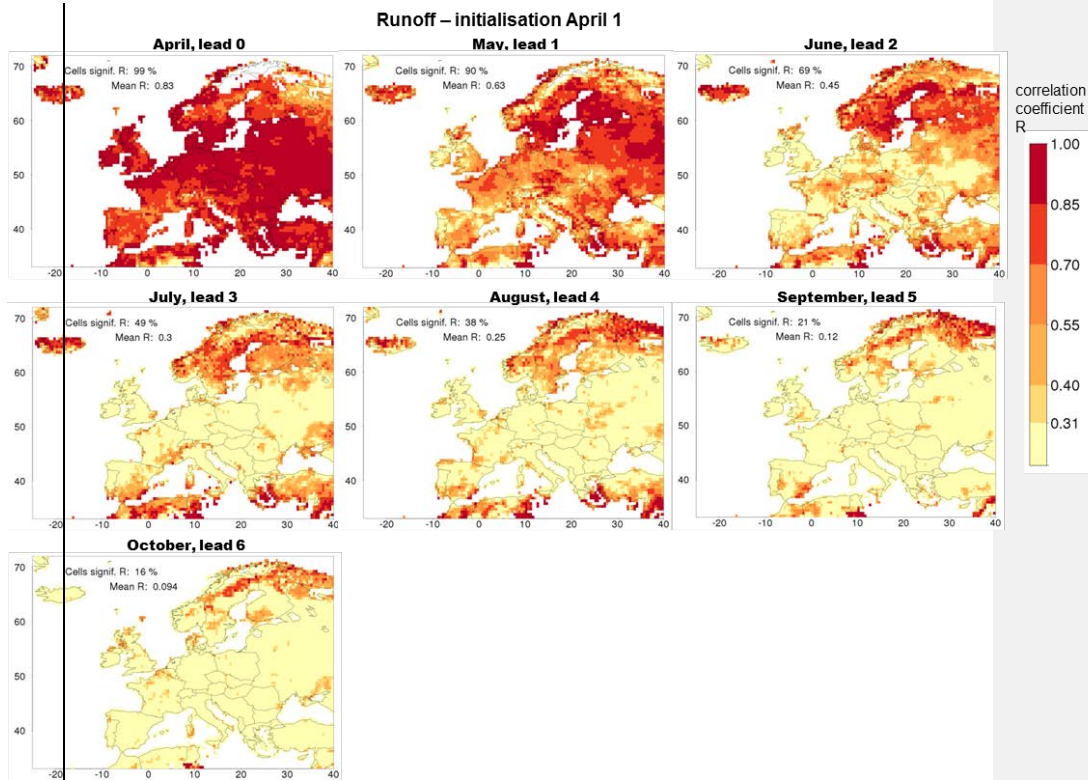
#### 533 3.1 Spatiotemporal variation of skill in runoff forecasts

534  
535 Eighty-four maps of skill of the runoff hindcasts were drawn for all 12 initialisation  
536 ~~of initialisation~~ months and all 7 lead months (all are presented in supplementary  
537 material S1). Two cross-cuts through that collection are shown in Figs. 1 (for a single  
538 initialisation month) and 2 (for a single lead month). The seven panels of Fig. 2 show  
539 the skill of the hindcasts initialised on April 1 as a function of lead time. Cells with an  
540 insignificant amount of skill are tinted yellow; cells where no metric could be  
541 computed remain white. In lead month 0, significant skill is found across almost the  
542 entire domain (99% of the cells). After the first lead month, the fraction of cells with  
543 significant skill gradually decreases to reach 16% at the longest lead time (lead month  
544 6). This is more than expected for the case of completely unskilful simulations (5% of  
545 the cells), so at the end of the hindcast simulations significant skill that does not occur  
546 due to chance is still present in some regions. The general impression is that the  
547 pattern of skill does not move in space but that skill is fading, i.e. for individual grid  
548 cells R is mostly decreasing with increasing lead time.

Commented [RH8]: check number for new SM

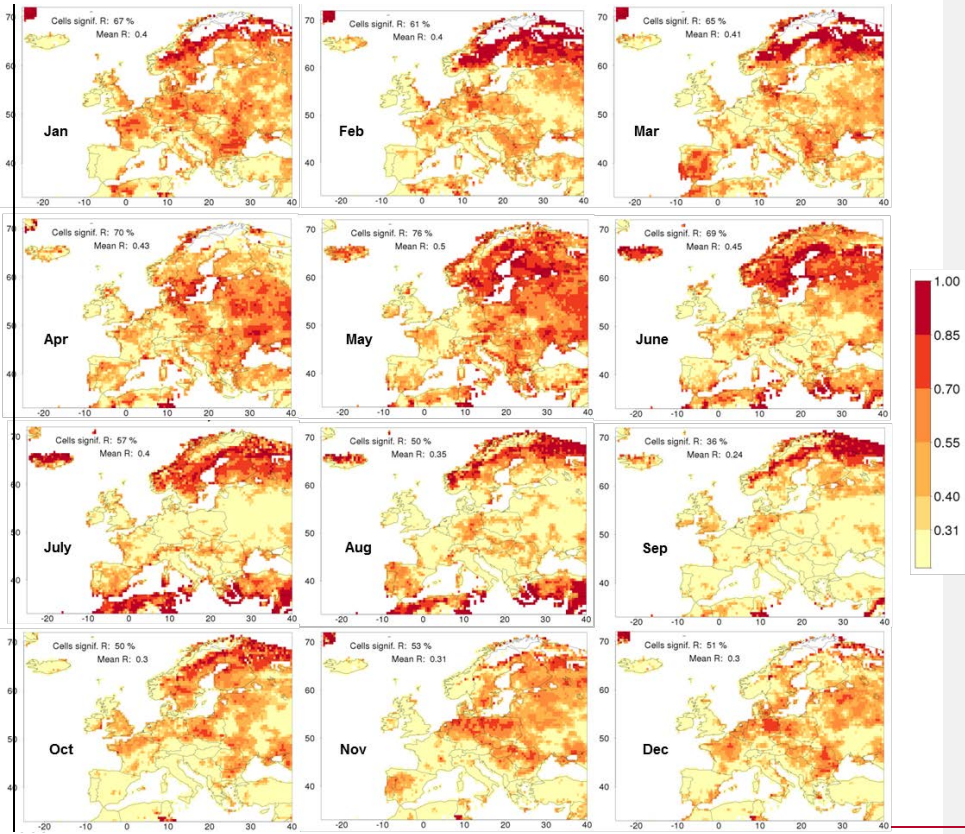




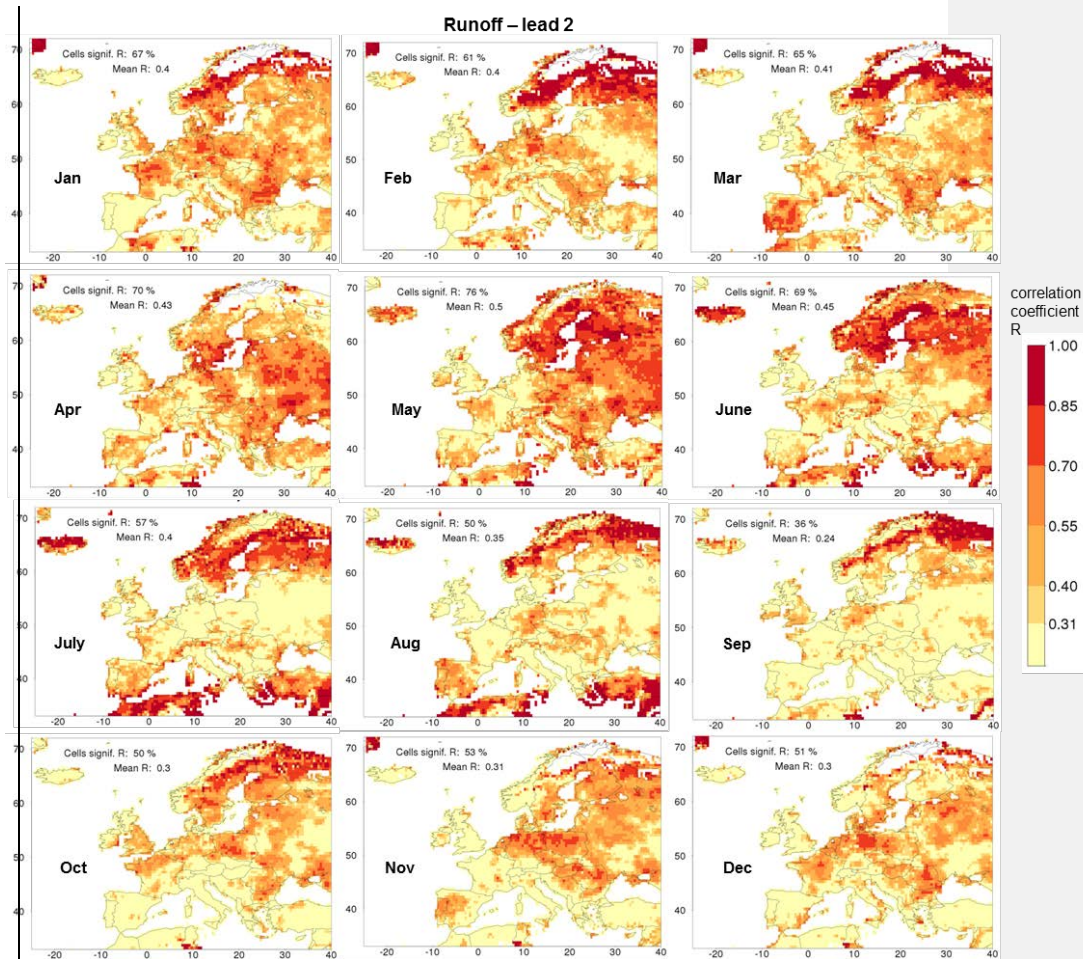


552

553 Figure 1: Skill of the runoff hindcasts initialised on April 1 for all seven lead  
 554 months. Skill is measured in terms of the Pearson correlation coefficient  
 555 between the median of the hindcasts and the observations (R). White,  
 556 terrestrial cells correspond to cells where observations or hindcasts consist  
 557 for more than one third of zeros or one sixth of ties. The threshold of  
 558 significant skill lies at 0.31, so yellow cells have insignificant skill, (dark)  
 559 red cells have (most) skill. White, terrestrial cells correspond to cells  
 560 where observations or hindcasts consist for more than one third of zeros or  
 561 one sixth of ties. The legend provides the fraction of cells with significant  
 562 values of R (at the 5% level) and the domain-averaged value of R.







565

566 Figure 2: Annual cycle of skill (R) of runoff hindcasts of lead month 2. More  
 567 explanation is given in the caption of Fig. 1.

568

569

570 The twelve panels of Fig. 2 show the annual cycle of skill of the hindcasts for lead  
 571 month 2. Consistent skill (persistent during at least ~~east-east~~ 3 consecutive target months) is  
 572 found in [\(causes of skill are reproduced here from the companion paper, Greuell et al.,  
 573 2017\)](#):

- 574 - Fennoscandia. Much skill is present during the entire year, except for November  
 575 and December, and there is a dip in skill in April. On average across the entire  
 576 region, skill reaches a maximum in May and June, i.e. the end of the melting  
 577 season, and –as shown in the companion paper- largely due to initialising snow.

578 Compared to the rest of the peninsula, there is generally less skill along the  
579 Scandinavian Mountain range. The companion paper shows some evidence this  
580 may be due to high variability of orographic rain, ~~ill-represented~~ in the re-  
581 forecasts.

Commented [RH9]: see fig 12 (org) Explanation paper

582 - Poland and ~~North-northern~~ Germany. The core period lasts from November to  
583 January, but it is extended with periods of less skill into October and the months  
584 from February to May. -Here both initialisation of soil moisture and snow -are  
585 important for skill.

586 - ~~West-western~~ France, more or less from Paris to Brittany and roughly from  
587 December to May. Skill derives from initialisation of soil moisture.

588 - The eastern side of the British Isles from December to February up to lead month  
589 2. Also here skill derives from soil moisture initialisation.

590 - Romania and Bulgaria. The core as well as the whole period are the same as that  
591 for Poland and ~~North-northern~~ Germany. In addition to causes mentioned there, in  
592 this part of Europe also summer P and ET are forecasted fairly well.

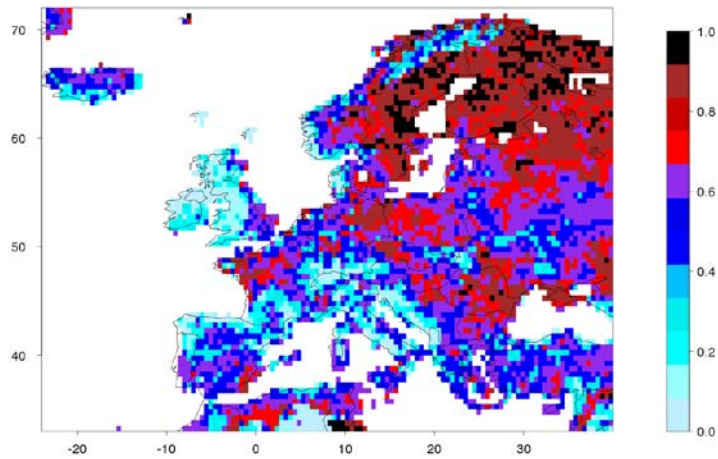
593 - The ~~southern~~southern- part of the Mediterranean region from June to August. The  
594 high amounts of skill are limited to the coastal parts of ~~North-northern~~ Africa,  
595 Sicily, ~~South-southern~~ Greece, Turkey, Syria and Lebanon.

596 - The Iberian peninsula from January to March up to lead month 2, and July and  
597 August like the other parts of the Mediterranean mentioned before. Skill derives  
598 from soil moisture in initialisation and in winter also from some skill in  
599 precipitation.

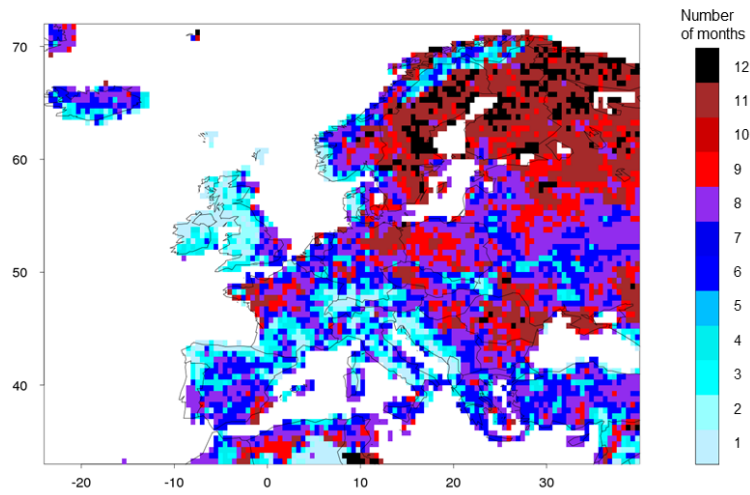
600

601 ~~These results can be compared to those of Bierkens and Van Beek (2009). They found~~  
602 ~~maxima in predictability of winter discharge in North Sweden, Finland, the region~~  
603 ~~between Moscow and the Baltic Sea, Romania and Bulgaria, and East Spain. For the~~  
604 ~~winter there is crude agreement with the current study about North Sweden, Romania~~  
605 ~~and Bulgaria but not about the other regions. For the summer, Bierkens and Van Beek~~  
606 ~~(2009) compute maxima in skill for South Spain, Sardinia, West Turkey and South-~~  
607 ~~west Finland. This pattern agrees to some extent with the locations of the summertime~~  
608 ~~maxima in skill of the present study (most of Fennoscandia and southern part of the~~  
609 ~~Mediterranean region).~~

610



611



612

613

614 Figure 3: ~~Number of months Fraction of the 12 months in a~~ of the year with  
 615 significant skill (R) in the runoff forecasts of lead month 2

616

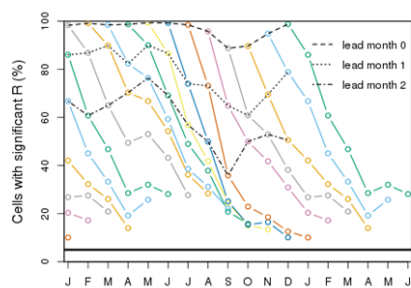
617

618 Figure 3 displays a synthesis of Fig. 2 in the form of a map with the fraction of the 12  
 619 months of the year with significant skill for lead month 2. Many of the regions with  
 620 very little or no skill all over the year are coastal regions (e.g. ~~north-northern~~ coast of

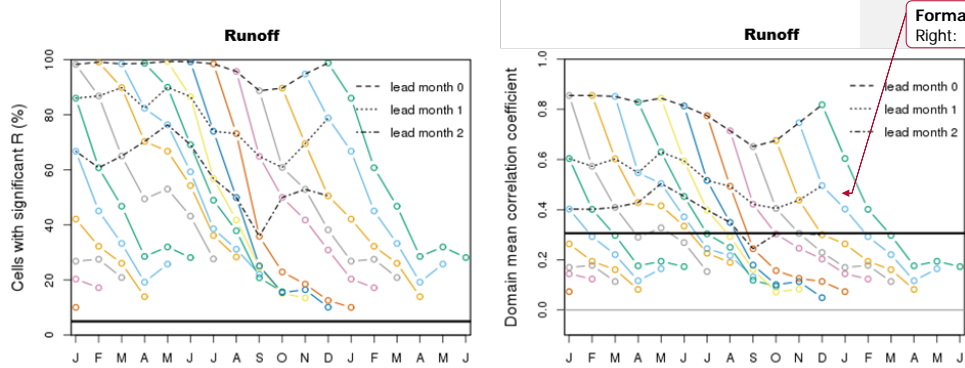
Spain), especially coastal regions on the western side of land masses (e.g. ~~west~~  
~~western~~ coasts of Denmark, ~~South-southern~~ Norway, Croatia and the British Isles), and  
mountain regions (e.g. the Alps, mountains in ~~North-northern~~ Norway and Sweden and  
on the Tatra on the border of Poland and Slovakia). The ~~entire~~ British Isles exhibit  
~~very~~ little skill, except for the ~~east-eastern~~ coast of Great Britain in late winter and  
early spring (JFMA). The companion paper shows that for regions with skill during a  
large part of the year, this skill is derived from complementary periods of skill due to  
initial conditions of snow and/or soil moisture.

These pan-European results can be compared to those of Bierkens and Van Beek  
(2009). They found maxima in predictability of winter discharge in Northern Sweden,  
Finland, the region between Moscow and the Baltic Sea, Romania and Bulgaria, and  
Eastern Spain. For the winter there is crude agreement with the current study about  
Northern Sweden, Romania and Bulgaria, but not about the other regions. For the  
summer, Bierkens and Van Beek (2009) compute maxima in skill for Southern Spain,  
Sardinia, Western Turkey and South-western Finland, a pattern that broadly agrees  
with the locations of the summertime maxima in skill (most of Fennoscandia and  
southern part of the Mediterranean region) we find.

Singla et al. (2012) found considerable skill in the Seine basin for low flows from June  
– September, a bit more eastern from the region where we found skill. Trigo et al.  
(2004) using a statistical model based on December NAO indices found skill for JFM  
discharge (and hydropower production) for the Douro, Tejo and Guadiana basins  
covering most of central and western Iberia. We confirm this skill which last till about  
May here, when initialised in January. In addition (not analysed by Trigo) we find skill  
beyond lead zero also in summer but then more concentrated around the south eastern  
coast of Iberia. Svensson et al. (2015) using a statistical model, based on NAO indices  
and river flow persistence, found good skill for winter river flows on the eastern side  
of the British Isles, consistent with our findings, and barely significant skill on its  
western coast that we do not reproduce.



Formatted: Left



Formatted: Left, Indent: Left: 0 mm, First line: 0 mm, Right: -0 mm

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

Figure 4: At left a) Fraction of cells with significant skill (in terms of R), and\_ at right b) domain average correlation in the runoff\_ hindcasts, as a function of initialisation month and lead time. Each coloured curve corresponds to the hindcasts initialised in a single month. For better visualisation, parts of the curves that end in the next year are shown twice, namely at the left hand and the right hand side of the graph. Black lines (dashed, dotted and dashed-dotted) connect the results for identical lead times. The horizontal line in a) show gives the expected fraction of cells with significant skill, in the case that the hindcasts have no skill at all (5%), in b) the minimal magnitude of the correlation for it to be statistically significant-

Figure 4a summarizes skill across the domain in terms of the fraction of cells with significant R for all initialisation and lead months. Overall there is a considerable amount of significant skill, with a minimum roughly from August to November and a maximum in May. For lead month 2 the fraction of cells with significant skill varies between 36% (September) and 76% (May). In all of the 84 combinations of initialisation and lead month, the theoretical value of no skill at all (5%) is exceeded, implying there are (small) pockets of skill even at lead month seven. Individual curves show the loss of skill with increasing lead time. The exception is formed by hindcasts starting in November, December and January which gain skill when they progress from April to May, a phenomenon caused by initial conditions of snow that takes longer or shorter to melt in (late) spring. For details, see the companion paper. A graph similar to Fig. 4b shows but for the domain-averaged R instead of the fraction of cells with a significant R (not shown here) shows identical behaviour including the mentioned exception to the overall trend of skill decaying with lead time decay and gain trends of domain averaged skill. It shows that a forecast initialised in February exhibits persistent domain average skill into June (5 lead months), while one starting in July does so only into August (2 months).

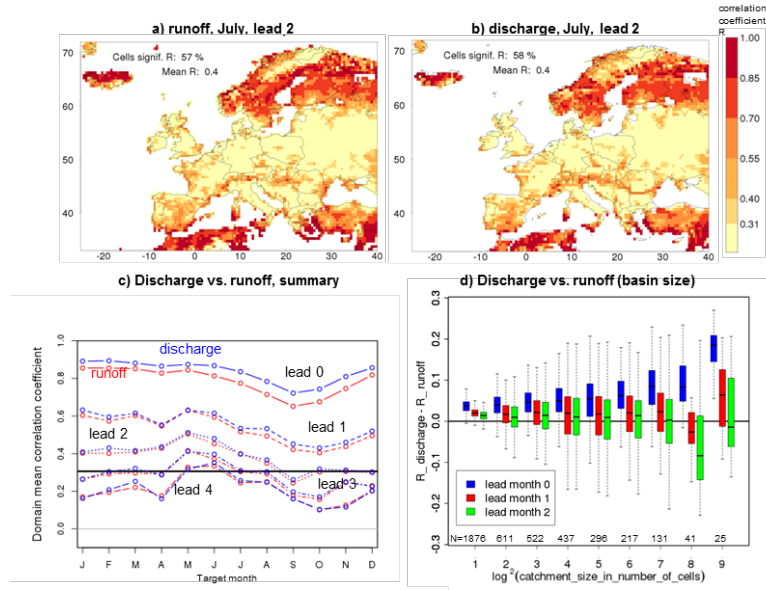
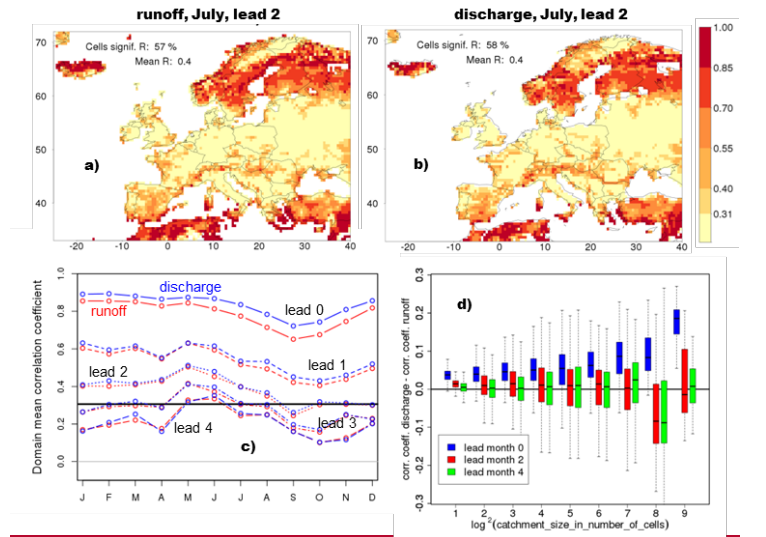
686 Similar summary plots for the other skill metrics are presented in the supplementary  
687 material S2, and discussed in section 3.4.

688

689

Commented [RH10]: add correct number for new SM

690 **3.2 Spatiotemporal variation of skill in discharge forecasts**



693  
694  
695 Figure 5: Comparison of the skill of the hindcasts of discharge and runoff. The two  
696 maps display R for runoff (a) and discharge (b) for hindcasts initialised on  
697 May 1 and target lead month 2 (July); (see further explanation in Fig. 1).

Formatted: Indent: Left: 0 mm, First line: 0 mm, Right: 0 mm, Space After: 10 pt

Formatted: Indent: Left: 0 mm, Right: 0 mm, Space After: 10 pt

Formatted: Left, Indent: Left: 0 mm, First line: 0 mm, Space After: 10 pt



698 Panel c depicts the annual cycle of the domain-averaged R for runoff (red)  
699 and discharge (blue) for lead months 0 to 4. The horizontal line at 0.31 is  
700 the threshold of significance for a single cell. Panel d is a box plot of the  
701 difference between R for discharge and runoff as a function of the  
702 catchmentbasin size. Each bin  $i$  contains the results for all catchmentbasins  
703 with a maximum of  $2^i$  cells and more than  $2^{(i-1)}$  cells, e.g. bin 4 is for all  
704 catchmentbasins with a size from 10 to 16 cells. Boxes represent the  
705 interquartile range and the median; and whiskers extend to minimum and  
706 maximum by 1.5 times the interquartile range from the box top and bottom  
707 values found in the bin. All values are average differences over the twelve  
708 months of the year and results are shown for three different lead times. The  
709 value above the abscissa give the number of cells in each bin.  
710  
711

### 712 3.2 Spatiotemporal variation of skill in discharge forecasts

713  
714 This sub-section compares skill for discharge with skill for runoff. The two maps of  
715 Fig. 5, which depict the skill in ~~the~~ runoff and ~~the~~ discharge hindcasts for July as lead  
716 month 2, show a high degree of similarity in terms of the patterns and the magnitude  
717 of the skill. The same holds for other target months and lead times (not shown). There  
718 are, however, subtle differences ~~though~~ because rivers ~~aggregate~~ average the skill, or  
719 lack of skill, from the whole upstream part of their catchmentbasin. As a result, cells  
720 containing rivers with large catchmentbasins may contrast against adjacent cells if  
721 these contain rivers with a small, local catchmentbasin. Indeed, some downstream  
722 parts of large rivers stick out in the skill map for discharge, but not in the skill map for  
723 runoff. An example in Fig. 5b are the reaches of the Danube along the Romanian-  
724 Bulgarian border, which show more skill than local small rivers in adjacent cells,  
725 because some upstream parts of the Danube have more skill than the region around the  
726 Romanian-Bulgarian border. An example that demonstrates the opposite is the  
727 downstream part of the Loire showing less skill than local small rivers, because  
728 upstream parts of the Loire have less skill than small, local rivers in the downstream  
729 part.  
730

731 Domain summary statistics of skill also differ slightly between runoff and discharge.  
732 Figure 5c compares the annual cycle of the skill in discharge with the skill in runoff at  
733 five different lead times. Here we show the difference in the domain-averaged R  
734 instead of the fraction of cells with a significant R because in lead month 0 that  
735 fraction is close to one for both variables. In terms of the domain-averaged R,  
736 predictability is higher for discharge than for runoff for the first lead month. On  
737 average over the 12 months of the year, the difference is 0.049. We ascribe this result  
738 to the combined effect of the delay between runoff and discharge and the general  
739 tendency of decreasing skill with lead time. The curves for the different lead times in  
740 Fig. 5c show that the difference in skill between the two variables gradually disappears



741 with increasing lead time (an annual average of 0.020 and 0.012 for lead months 1 and  
742 2, respectively). This is compatible with the given explanation for the difference and  
743 the fact that the rate withby which skill is lost gradually decreases with increasing lead  
744 time.

745

746 We finally analysed whether the difference in skill between discharge and runoff was  
747 a function of the size of the catchmentbasin (Fig. 5d). For the first lead month, when  
748 on average there is more skill in discharge than in runoff, the difference increases with  
749 the size of the catchmentbasin. Again, this can be explained by the combination of the  
750 skill decaying with time and the delay between runoff and discharge, with the delay  
751 increasing with the size of the catchmentbasin. For longer lead times (lead months 2  
752 and 4), when the domain-averaged difference in skill has become very small (Fig.  
753 5panel-c), panel- the figured shows no effect of the catchmentbasin size. So, rReferring  
754 to the comparison between runoff and discharge in panels Fig. 5a and 5b for lead  
755 month 2, cases like the Danube (more skill than local rivers) and the Loire (less skill  
756 than local rivers) tend to cancel when the entire domain is considered.

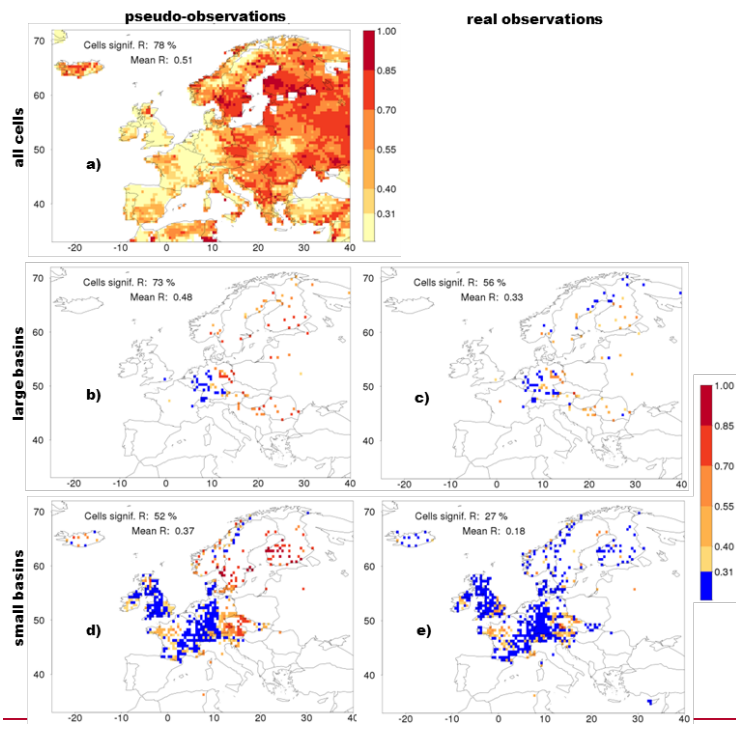
757

758

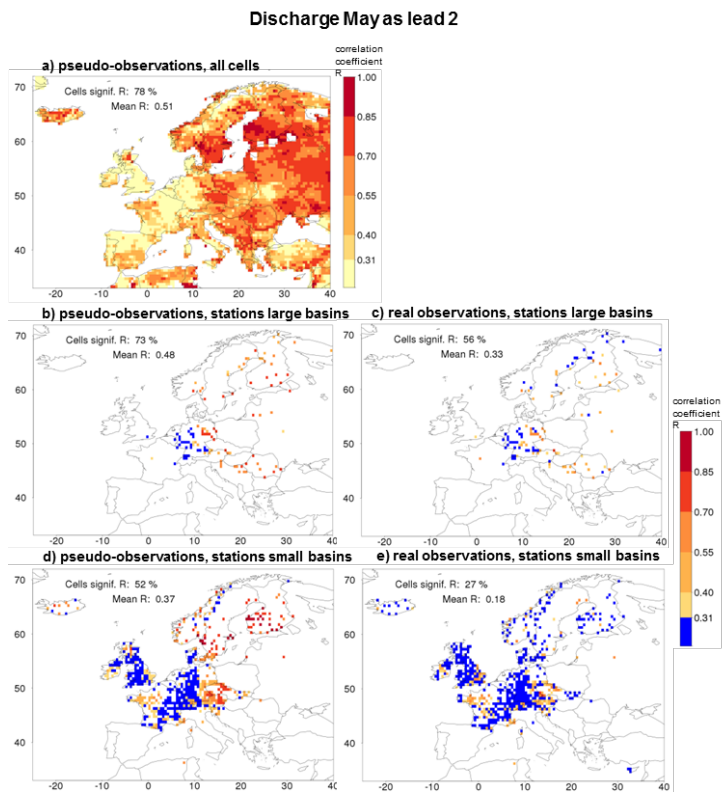
### 759 3.3 Verification of discharge with pseudo- and real observations

760

761 So far, all skill was determined by using the discharge generated with the reference  
762 simulation. i.e. with pseudo-observations. In this section, this “theoretical skill” will be  
763 compared with the skill determined with real discharge as observed at gauging stations  
764 (“actual skill”) from the GRDC and EWA data bases. Figure 6 compares the  
765 theoretical skill (Fig. 6panels-b and 6d for large and small basins, respectively) with  
766 actual skill (Fig. 6panels-c and 6e for large and small basins, respectively) for a single  
767 combination of a target month (May) with a lead month (2).



768



769

770

771

Figure 6: Skill ( $R$ ) of the discharge hindcasts for May as lead month 2 (initialisation on March 1). In sequence: a) discharge verified with pseudo-observations, b) as a but for cells representing large basins only, c) discharge verified with real observations for large basins. The two final panels (d) and e) are identical to b) and c), respectively, but for cells representing small basins. More explanation is given in the caption of Fig. 1 but in panels d) and -e) cells with insignificant skill are coloured blue instead of yellow for better contrast.

772

773

774

775

776

777

778

779

780

781

For this combination of May forecasts initialised in March target and lead month of Fig-6, a substantial degradation in skill is found when the pseudo-observations are replaced by real observations. In terms of the fraction of cells with significant skill, the reduction is from 73 to 56 % for large basins and from 52 to 27 % for small basins and the domain-averaged  $R$  decreases from 0.48 to 0.33 for large basins and from 0.37 to 0.18 for small basins. Of the larger basins especially those in northern Fennoscandia lose all skill when using actual observations, a region where VIC also performed poorly in reproducing historic flows: there specific discharge was underestimated and

782

783

784

785

786

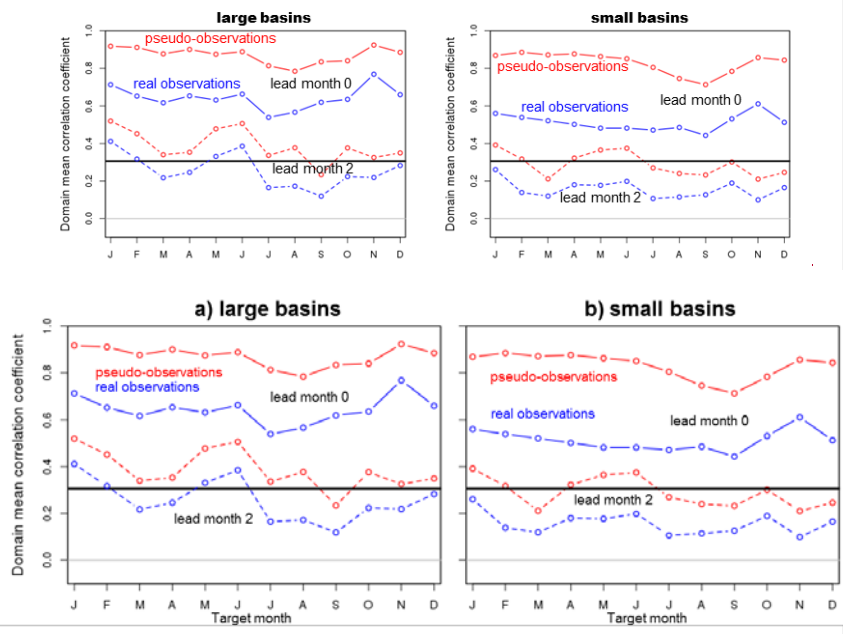
787

788

789 the annual cycle was poorly reproduced (especially the spring peak occurred too late  
 790 and too long (Greuell et al 2015). In central Europe useful skill remains when using  
 791 real observations, a region where VIC well reproduced annual cycles, though  
 792 interannual variation in low flows were overestimated in that area. With respect to  
 793 the latter it should be stressed that (n Greuell et al. 2015 consider the 5 percentile as  
 794 low flows (Q5) where here we consider the 33 percentile as below normal.

796 Figure 7 compares actual with theoretical skill for all target months and two lead times  
 797 by considering the domain-mean R. Similar figures for the other skill metrics are  
 798 presented in supplementary material S4 and discussed in the next section 3.4. The  
 799 reduction in skill occurs for all combinations of target and lead months and does not  
 800 exhibit a clear annual cycle. On average across all target months and for lead month 2,  
 801 the ratio of actual to theoretical skill is 0.667 (0.258 divided by 0.387) for large basins  
 802 and 0.538 (0.156 divided by 0.290) for small basins. This is can be comparabled  
 803 to Van Dijk et al. (2013), who found a ratio of actual to theoretical skill of 0.54 for 6192  
 804 catchmentbasins worldwide in terms of the ranked correlation coefficient.

Commented [RH11]: check number for new SM



809  
 810  
 811 Figure 7: Comparison between verification of discharge with pseudo- (red) and real  
 812 (blue) observations in terms of the annual cycle of the domain mean R.  
 813 The horizontal line at 0.31 is the threshold of significance for a single cell.

814 Results are shown for cells representing large basins (left) and cells  
815 representing small basins (right). Both panels depict cycles for lead  
816 months 0 and 2 only.

817

818

819 Comparing skill for small basins with skill for large basins in Fig. 7, we notice two  
820 differences. Firstly, in terms of the domain mean  $R$ , theoretical skill is higher for large  
821 basins than for small basins (0.39 and 0.29, respectively, for the annual mean and lead  
822 month 2). However, this result holds for the cells with observations. If all cells of the  
823 domain are considered, ~~this~~ difference ~~becomes insignificantly small, almost vanishes.~~  
824 ~~On average, all cells with an upstream catchment larger than 10000 km<sup>2</sup> have a mean~~  
825  ~~$R$  of 0.396 and all cells with an upstream catchment smaller than 2500 km<sup>2</sup> have a~~  
826 ~~mean  $R$  of 0.384.~~ So, the apparent difference in theoretical skill between large and  
827 small basins can be ~~blatantly~~ attributed almost entirely to the geographical distribution of  
828 the ~~discharge monitoring stations, with stations on small basins stations~~ being  
829 relatively more often located in regions with relatively little skill like Germany, France  
830 and the British Isles than large basin stations.

831

832 The second effect of the size of basins is that ~~skill-reduction~~ between theoretical and  
833 actual skill is larger for small basins than for large basins. This is perhaps even more  
834 clear from Fig. S3 in the supplementary material. We speculate that this is due to a  
835 combination of two effects. ~~Firstly, there is more skill in simulations of historic~~  
836 ~~streamflow in large basins than in small basins (Van Dijk and Warren, 2010,~~  
837 ~~confirmed for VIC in Europe by Greuell et al. 2015).~~ Secondly, as Van Dijk et al.  
838 (2013) demonstrated, the ratio of actual to theoretical skill is almost linear in the skill  
839 of simulating historic streamflow. Combining these two relationships confirms the  
840 relationship that we found, namely an increase in the ratio of actual to theoretical skill  
841 with basin size.

842

843

844 Finally, ~~We~~ investigated to what extent these results are affected by human  
845 interference, keeping in mind that the simulations are naturalized, ~~—~~ while the  
846 observations include human impacts to a variable but unknown degree. Human  
847 interference is expected to have a negative effect on actual skill and hence on the ratio  
848 of actual to theoretical skill. ~~—~~ We quantified the human impact by performing two  
849 model simulations with the Lund Potsdam Jena managed Land (LPJmL) model (Rost  
850 et al., Schaphoff et al., 2013) that was operated at the same spatial resolution (0.5° ×  
851 0.5°) and with the same river network (DDM30) as VIC. From the discharge output of  
852 a naturalized run and a run with reservoir operation and irrigation, the human impact at  
853 cell level was quantified by computing the so-called Amended Annual Proportional  
854 Flow Deviator (AAPFD, see Marchant and Hehir, 2002). Subsequently, we selected all  
855 discharge observations for large basins with an AAPFD < 0.3, i.e. basins with a  
856 relatively small degree of human impact (about half of all 111 basins). For relatively  
857 natural basins (AAPFD < 0.3; see end of section 2.2), ~~this selection~~ the ratio of actual

Commented [RH12]: also for VIC? see Greuell 2015

858 to theoretical skill was computed in terms of the domain mean  $R_s$  averaged across all  
859 target months and for lead month 2. We found a ratio of 0.686, which should be  
860 compared to a ratio of 0.667 for the entire set of large basins (see above). So, as  
861 expected the ratio is larger for basins with less impact. However, since the difference  
862 between the two ratios is small we conclude that the effect of ~~evaluating the~~  
863 ~~combination of~~ naturalised runs ~~against with~~ observations that ~~are obviously are~~  
864 affected by human interference, contributes only little to the difference between actual  
865 and theoretical skill. A similar analysis was not applied to the collection of small  
866 basins with observations, since these are ~~generally~~-smaller than the spatial resolution  
867 of the simulations.

868  
869 ~~Comparing skill for small basins with skill for large basins in Fig. 7, we notice two~~  
870 ~~differences. Firstly, in terms of the domain mean R theoretical skill is higher for large~~  
871 ~~basins than for small basins (0.39 and 0.29, respectively, for the annual mean and lead~~  
872 ~~month 2). However, this result holds for the cells with observations. If all cells of the~~  
873 ~~domain are considered, the difference almost vanishes. On average, all cells with an~~  
874 ~~upstream catchment larger than 10000 km<sup>2</sup> have a mean R of 0.396 and all cells with~~  
875 ~~an upstream catchment smaller than 2500 km<sup>2</sup> have a mean R of 0.384. So, the~~  
876 ~~apparent difference in theoretical skill between large and small basins can be blamed~~  
877 ~~almost entirely to the geographical distribution of the stations, with small basin~~  
878 ~~stations being relatively more often located in regions with relatively little skill like~~  
879 ~~Germany, France and the British Isles than large basin stations.~~

880  
881 ~~The second effect of the size of basins is that skill reduction is larger for small basins~~  
882 ~~than for large basins. We speculate that this is due a combination of two effects.~~  
883 ~~Firstly, there is more skill in simulations of historic streamflow in large basins than in~~  
884 ~~small basins (Van Dijk and Warren, 2010). Secondly, as Van Dijk et al. (2013)~~  
885 ~~demonstrated, the ratio of actual to theoretical skill is almost linear in the skill of~~  
886 ~~simulating historic streamflow. Combining these two relationships confirms the~~  
887 ~~relationship that we found, namely an increase in the ratio of actual to theoretical skill~~  
888 ~~with basin size.~~

### 891 3.4 Results for other skill metrics

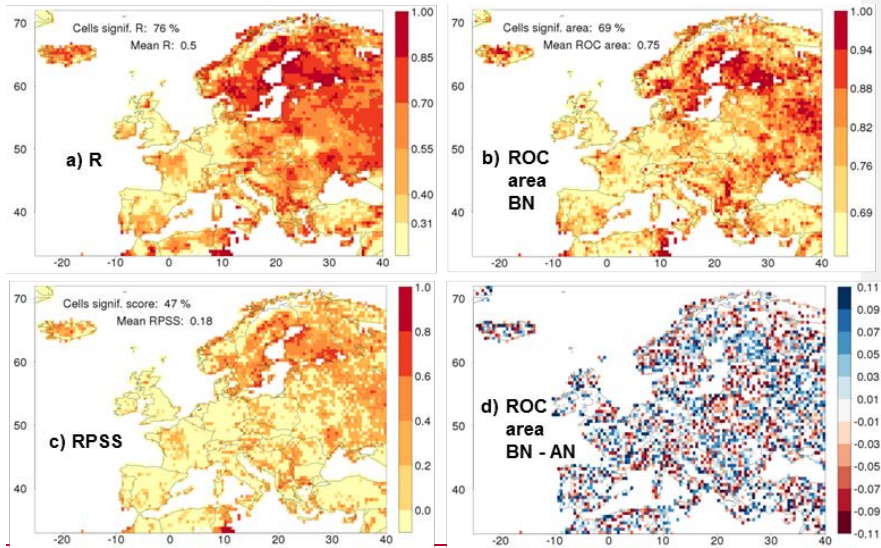
892  
893 So far, skill was measured in terms of the correlation coefficient between the median  
894 of the hindcasts and the observations (R) only. This section compares those results, ~~for~~  
895 ~~runoff-~~ with results in terms of other skill metrics. Figure 8 gives an example for one  
896 particular target month (~~May~~) and lead month, ~~i.e. target May initialised in March~~  
897 ~~(lead 2)-2. Fig. 8a Panels a, 8b and 8c show the skill patterns for R, for the ROC area~~  
898 ~~for Below Normal (BN) years and for the RPSS. The three patterns are spatially~~  
899 ~~similar to a large degree, noting that differences in colour are partly due to the~~  
900 ~~interplay between differences in the domain averaged magnitude of the skill metrics~~  
901 ~~and the choice of the colour intervals though the magnitudes and number of significant~~

902 cells do differ. The pattern of the map of the ROC area for Above Normal (AN) years  
903 (not shown here) is also similar to the patterns of the three maps shown. On average,  
904 across all lead and target months, 89% of the cells that have significant R also have  
905 significant ROC scores for the BN tercile, 84% also for the ROC scores for the AN  
906 tercile. Finally, 65% of the cells that have significant R also have significant RPSS  
907 scores.- The fraction of cells with no significant R, but with significant ROC or RPSS  
908 remains below the 5% level across all target and lead months, and thus such cases are  
909 likely due to chance.

Commented [RH13]: in suppl

910  
911 The agreement that we find between the patterns of the different metrics is in  
912 accordance with a result mentioned in a global analysis of seasonal streamflow  
913 predictions by Van Dijk et al. (2013) who found high spatial correlation between the  
914 different skill metrics they used (among which R, the RPSS and the ranked correlation  
915 coefficient).





Runoff May as lead 2

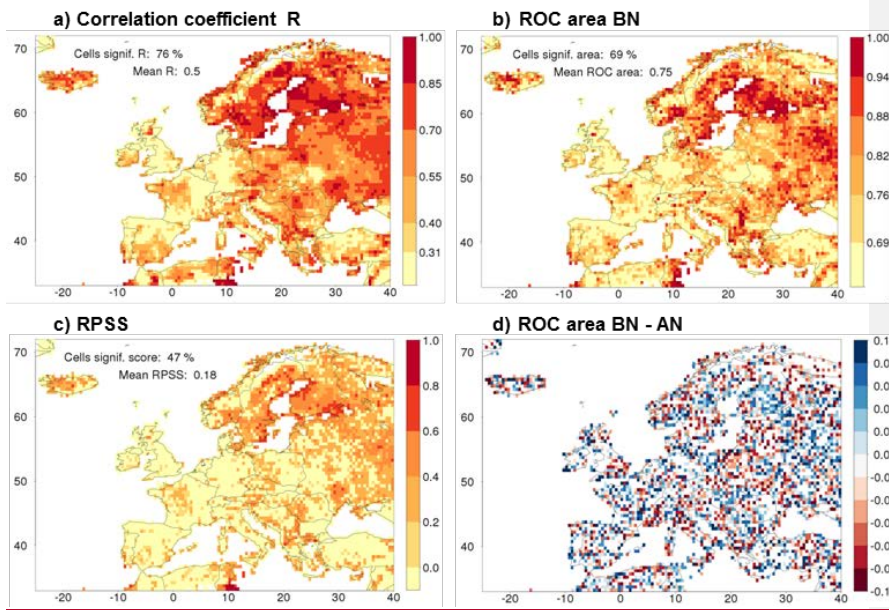


Figure 8: Maps of different skill metrics for one combination of a target month (May) and a lead month (2) of the runoff hindcasts. Panels show a) R, b) the ROC area for the ~~B~~below-~~N~~normal tercile, c) the Ranked Probability Skill Score (RPSS) and d) the difference in ROC area between the

923 ~~BN~~Below-Normal and ~~AN~~the Above-Normal terciles. In panels a, b and c  
924 skill is not significant in cells with a yellow colour. Legends provide the  
925 fraction of cells with significant values of the metric and the domain-  
926 averaged value of the metric.

927  
928  
929 ~~Though~~Although the different nature of the different metrics does not enable a  
930 quantitative comparison of the metrics, ROC areas for the different terciles can be  
931 compared among each other. For the particular combination of May target month and  
932 lead month two-shown in Fig. 8, the domain-mean ROC area is largest for the BN  
933 tercile (0.75), slightly smaller for the AN tercile (0.73) and much lower for the near-  
934 normal (NN) tercile (0.58, see Fig. S2c and S2d not shown here; 0.5 corresponds to  
935 climatological forecasts). A similar tendency is found in the fraction of cells with a  
936 significant ROC area (69%, 63% and 21%, respectively). The fraction of cells with a  
937 significant value of the RPSS is 47%, which is somewhere between the fractions for  
938 ROC areas of the three terciles because the RPSS represents "mixes" the skill to make  
939 forecasts of events falling in across all terciles. All metrics show a minimum value in  
940 the annual cycles in either September or in October, irrespective of lead time; maxima  
941 are attained in February for lead month 0 shifting to May at longer lead times (Fig.  
942 S2). Finally, panels Fig. 8d presents a map of the difference between the BN and the  
943 AN ROC area. There is no clear regional pattern in this difference, i.e. coherent larger  
944 regions with clustered positive or negative values cannot be distinguished. BN ROC  
945 values are larger than AN (blue colours) in southern Finland and central Sweden,  
946 western France, Hungary and Serbia and large parts of Russia. The reverse (ROC AN  
947 > ROC BN, red colours) is true in eastern Poland and the Baltic states, southern  
948 eastern France (Rhône basin) and eastern UK.

949 For other combinations of target and lead months the results of this analysis are  
950 similar, though numbers may vary. All of these results also hold for other combinations  
951 of target and lead month. See supplementary figures.

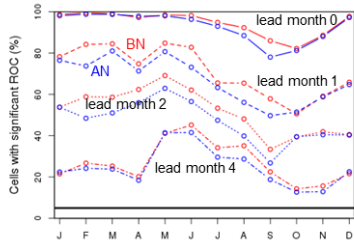
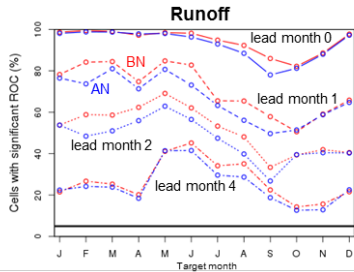
952 Figure 9 compares the BN with the AN tercile in terms of the fraction of cells with a  
953 significant ROC area across all target and initialisation months. The main finding is  
954 that in all combinations of lead and target months the fraction significant cells is  
955 larger for the BN than for the AN tercile. This is perhaps not as expected from the VIC  
956 performance in reproducing historic flows, which is better for high flows than for low  
957 flows (Greuell et al., 2015; recall that their high/low flows are defined as p95 and p5,  
958 respectively, while here they are p67 and p33; see also Section 2.1). However, the AN  
959 and BN two fractions tend to become equal (i) when these ROC areas approach 1.0,  
960 (ii) when they approach the limit of no skill (5%) and (iii) during target months from  
961 October to January.

962

963

Formatted: Justified, Right: -0.1 mm

964



965

966 Figure 9: Skill of the runoff hindcasts in the Below Normal (BN) ~~minus~~ compared to  
 967 the skill of the runoff hindcasts in the Above Normal (AN) tercile. The  
 968 plot depicts annual cycles of the fraction of cells with a significant ROC  
 969 area for the two terciles and for four lead months.

970

971

#### 972 4 Discussion

973

974 ~~For verification of the hindcasts two options were considered in this paper. We~~  
 975 ~~determined the skill of the hindcasts by comparing predicted discharge with the output~~  
 976 ~~of the reference simulation (the “pseudo observations” leading to “theoretical skill”)~~  
 977 ~~and with observations of real discharge (“real observations” leading to “actual skill”).~~  
 978 ~~To obtain a basis for understanding the differences in skill that we found, Fig. 10~~  
 979 ~~presents a streamflow diagram of the three relevant physical systems, namely the real~~  
 980 ~~world and the model systems that generate the hindcasts and the pseudo observations.~~  
 981 ~~In each system, confined in the diagram by a box, meteorological and initial~~  
 982 ~~conditions force and initialize hydrology, of which discharge is the relevant~~  
 983 ~~component here. There are two complications. First, the initial conditions themselves~~  
 984 ~~are generated by meteorological forcing during the spin up period, initial conditions at~~  
 985 ~~the beginning of the spin up period and hydrology. This is represented by the upper~~  
 986 ~~left branch in each box, omitting initial conditions at the beginning of the spin up~~  
 987 ~~period for simplicity. Second, due to measurement errors real observations of~~

Formatted: Justified, Right: -0.1 mm

discharge generally differ from real discharge (Juston et al., 2014) as illustrated in the upper right corner of the figure.

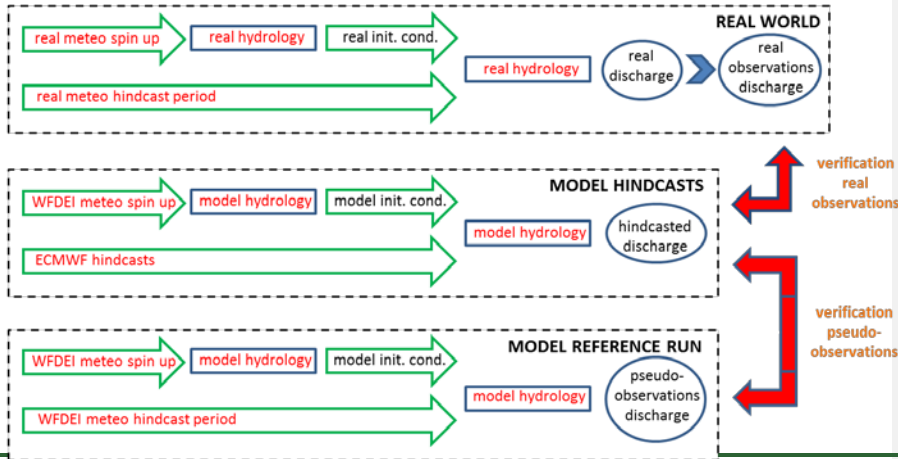


Figure 10: Diagram illustrating conceptual differences between verification of hindcasts (in the middle) with pseudo observations (bottom) and with observations of real discharge (top). See the text for a detailed explanation.

#### 4.1 Theoretical versus actual skill

The two essential questions are: 1) What are the conceptual differences between the physical systems that generate the pseudo- and the real discharge observations, i.e. between the model reference run and the real world. To answer this question, the components in the upper and the lower box of the diagram need to be compared. 2) What are the expected effects of these differences on skill, i.e. on the comparison with the hindcasts. To answer this question, the components that differ between the real world and the model reference run need to be compared with the model hindcasts. The rule then is that skill decreases with increasing disagreement between a component of the hindcast system and the corresponding component of one of the other systems. The following components (red text in diagram) differ between the real world and the model reference simulation, and their expected effect on skill are:

- 1) — Real meteorology differs from the meteorology assumed in the reference simulation (WFDEI), both during the spin up period and during the hindcast period. During spin up, model reference run and hindcasts have identical meteorological forcing (namely WFDEI), which differs from real meteorology. Therefore, this difference is expected to lead to more theoretical

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 6.3 mm + Indent at: 12.7 mm

1018 than ~~to~~ actual skill. During the hindcast period, all three systems have different  
1019 meteorological forcings. For cases with skill in the meteorological hindcasts,  
1020 one would need to have an expectation about the agreement between the skilful  
1021 hindcasts and reality, on one side, and the ~~skilful~~ skilful hindcasts and the  
1022 WFDEI data set, on the other side. Unfortunately, we do not have a well-  
1023 founded expectation about such a ~~dis~~ difference in agreement and, hence, we have  
1024 no expectation about its effect on the difference between theoretical and actual  
1025 skill. However, in Europe and beyond the first lead month almost all skill in  
1026 the seasonal forecasts is due to the initial conditions; ~~(see the companion~~  
1027 ~~paper)~~. Therefore, beyond the first lead month and in Europe differences in  
1028 forcing during the hindcast period have a negligible effect on skill.

1029 2. 2) — Models are imperfect, in terms of physics and in terms of spatial and  
1030 temporal discretisation, so model hydrology differs from real world hydrology.

1031 Hindcasts and the pseudo-observations are produced with the same model, so  
1032 imperfections in model hydrology are expected to lead to more theoretical than  
1033 actual skill. One assumption implicitly made in the diagram is that the basin of  
1034 the observation station and the model basin are identical. This is not the case  
1035 (see Sect. 2.2), so differences between observation and model basin form an  
1036 additional cause of disagreements between theoretical and actual skill. Again,  
1037 this will favour theoretical skill with respect to actual skill since basins are  
1038 identical in the hindcasts and the reference simulation. In particular,  
1039 differences in meteorological forcing between the basin of the observation  
1040 station and the model basin reduce actual skill. Van Dijk et al. (2013)  
1041 investigated this aspect by making simulations for Australia at different spatial  
1042 resolutions and verifying with networks of observations with different spatial  
1043 densities. They found that the resolution and perhaps the quality of the forcing  
1044 data contributed to at least half of the difference between theoretical and actual  
1045 skill.

1046 3. 3) — In the real world ~~a difference~~ discharge observations ~~differ~~ are subject  
1047 to from reality, i.e. a measurement error exists. Measurement errors of  
1048 discharge are not constant over time (due to varying cross sectional areas,  
1049 following erosion and sedimentation) and therefore add noise to the data; noise  
1050 always reduces skill. There is no equivalent of this error in the model  
1051 environment. Hence, as for differences 1) and 2) this difference is expected to  
1052 lead to more theoretical than to actual skill.

1054 4. Initial conditions are absent in this list of differences since in WUSHP they are  
1055 not independent components but entirely determined by two components of the  
1056 system listed above, namely meteorology and hydrology. Alternatively, initial  
1057 hydrological conditions could be taken from observations or by assimilation of  
1058 observations into model calculations. In that case, initial conditions would  
1059 become an independent or semi-dependent component of the system.  
1060 However, again, while model initial conditions would, of course, differ from  
1061 real initial conditions, the two model system had identical initial conditions.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: List Paragraph, Numbered + Level: 1 +  
Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left +  
Aligned at: 6.3 mm + Indent at: 12.7 mm

Formatted: Font: (Default) Times New Roman, 12 pt

Hence, ~~again~~ this difference would again be expected to lead to more theoretical than to actual skill.

Formatted: Font: (Default) Times New Roman, 12 pt

In summary, all of the conceptual differences between the generation of pseudo- and real observations, are expected to lead to more theoretical skill than actual skill, except for the difference in meteorology during the hindcast period, which has, in the case of Europe beyond the first lead month, a neutral effect, and otherwise an unknown effect.

~~A complication to this analysis is failure of the assumption implicitly made in the diagram that the catchment of the observation station and the model catchment are identical. This is not the case, see Sect. 2.2, so differences between observation and model catchment form an additional cause of differences between theoretical and actual skill. Again, this will favour theoretical skill with respect to actual skill since catchments are identical in the hindcasts and the reference simulation. In particular, differences in meteorological forcing between the catchment of the observation station and the model catchment reduce actual skill. Van Dijk et al. (2013) investigated this aspect by making simulations for Australia at different spatial resolutions and verifying with networks of observations with different spatial densities. They found that the resolution and perhaps the quality of the forcing data contributed to at least half of the difference between theoretical and actual skill.~~

Our data analysis, section 3.3, broadly confirms that theoretical skill exceeds actual skill.

It is interesting to discuss what would happen in the utopian case that the system of the model reference run would converge with the real world, i.e. if model meteorological forcing and hydrology would approach perfection and if measurement errors would approach zero. Equality of the two systems would, according to the analysis above, lead to equality of theoretical and actual skill. However, we like to note that at the same time optimisation of the model system ~~cannot, and would in many cases,~~ lead to a degradation of the theoretical skill if the hydrological models have unrealistic memory time scales in their storage compartments. If this memory, from stored water in either snow, soil or aquifer (or man-made reservoirs behind dams), is too strong then skill will reduce with calibrating the model towards more realistic storage accumulation. However, if this memory is too small initially then of course the reverse may happen and skill increases with optimization.

~~Hence, theoretical skill is not equal to the maximum that could be accomplished if hydrological model and meteorological forcing during the reference simulation were perfect.~~ An example proving this statement is a model that ~~is imperfect because it that~~ accumulates too much snow. The model will do so both in the initial state of the reference simulation and the initial state of the hindcasts and since more snow leads, at some stage of the melting season, to more predictive skill, theoretical skill will be overestimated. A perfect model, accumulating less but more realistic amounts of snow,



1105 would ~~show exhibit~~ less skill. Another example ~~is underlining the statement that~~  
1106 ~~theoretical skill is not the maximum that could be realized with a perfect model deals~~  
1107 ~~with~~ predictive skill caused by interannual variations in the initial amount of soil  
1108 moisture ~~and/or groundwater~~. A model that is imperfect because it overestimates the  
1109 transport speed of ~~soil moisturewater~~ through the soil and the groundwater reservoirs  
1110 will do so both in the reference simulation and the hindcasts. Predictive skill due to  
1111 soil moisture initial conditions will then occur too early. Compared to the model that  
1112 overestimates transport speed, a perfect model with smaller, realistic transport speed  
1113 would yield less theoretical skill at the early lead times.

1114 Hence, theoretical skill is not equal to the maximum that could be accomplished if  
1115 hydrological model and meteorological forcing during the reference simulation were  
1116 perfect.

1117 The version of VIC used in this study was calibrated by Nijssen et al. (2001) in a crude  
1118 way, in the sense that they assumed no spatial variation of the parameters set by  
1119 calibration within almost the entire European continent. Improving the calibration of  
1120 VIC would be an obvious candidate for trying to improve the seasonal predictions  
1121 discussed in this paper. This should lead to higher actual skill. However, the two  
1122 examples discussed in the previous paragraph show that theoretical skill may actually,  
1123 for certain locations, months of initialisation and lead months, decline due to the  
1124 recalibration.

#### 1125 **4.2 Results and uncertainties**

1126 There seems to be a broad correspondence between the probabilistic forecast  
1127 verification presented here and the model validation presented in Greuell et al. 2016,  
1128 and Roudier et al. 2016. These studies found that average discharge and inter-annual  
1129 variations therein are well reproduced, consistent with our result that all skill scores  
1130 are good for large parts of Europe in the first lead month. Their finding that high flows  
1131 are generally better reproduced than low flows seems to contradict with our fact that  
1132 BN forecasts are more reliable than AN forecasts (although by a small margin). This  
1133 discrepancy may be due to different definitions of high or low flows between these  
1134 studies and the present one. They define high and low flows by 95 and 5 percentiles,  
1135 respectively, while here we use 66 and 33 percentiles, much less extreme values. Also,  
1136 their study showed that the variability in Q5 was more overestimated than the  
1137 variability in Q95, which may be a reason for the higher skill we find in the lower  
1138 tercile (skill requires variability, see discussion of companion paper), though this  
1139 inference is hard to prove. [...]

1140 This prior work also invokes some warnings. Greuell et al. found that seasonal flow  
1141 cycles show a too late and too broad spring peak in (northern ) Fennoscandia. This  
1142 suggests that our theoretical forecast skills may also be too high at too long lead times  
1143 in that region and season, (as was also already revealed by comparing Figure 6b vs  
1144 6c).

Commented [RH14]: also need some geographically explicit remarks



1145 In a future extension of our work, an objective method like cluster analysis could  
1146 reveal regions where skill has a similar signature. This could lead to an improved  
1147 assessment of the physical and climatological factors that are responsible for the  
1148 spatial variations in skill found in this and its companion paper.

1149 There also seems to be a broad correspondence between the regions and seasons with  
1150 skill identified in the present work, with that from more spatially or temporally  
1151 confined studies based on entirely different physical or even statistical models.  
1152 Without repeating the more full description in the Introduction section and comparison  
1153 in section 3.1, Bierkens and van Beek, (2009) and Thober et al. (2015) their results  
1154 were similar at the European domain, further more confirming more regional studies  
1155 such as for the British Isles (Svensson et al., 2015), Iberia (Trigo, 2004) or France  
1156 (Céron et al., 2010; Singla et al., 2012). Though a high resolution study like the latter  
1157 may add much spatial detail, this does not change the region and season of skill

1158 Our results are based on a forcing with the 15 member, monthly initialized, 7 month  
1159 forecast version of ECMWF System 4, basically because at the start of this work their  
1160 hindcast was the only one accessible to us, but also because it allows verification at the  
1161 highest temporal resolution. Alternatively, we could have used the 51 member  
1162 seasonally initialised (4 times per year), 7 month forecast version of the same model.  
1163 That would have provided us with better constrained, more precise statistics (larger  
1164 sample size), or would have allowed assessment of more percentiles (e.g. quintiles  
1165 instead of terciles) at similar precision. But the variation of skill over a year would not  
1166 have been resolved with such detail as in the present work. Finally also a 15 member,  
1167 seasonally initialized, 12 month forecast version is available. However, as our results  
1168 show at lead month 6 only very few, small pockets of persistent skill remain,  
1169 suggesting that extending the forecast for our domain is probably not useful.

1170 Other seasonal forecasting systems, based on different couple ocean-climate models,  
1171 exist that could have been used, such as CFSv2 (Saha et al., 2014), Glosea5  
1172 (MacLachlan et al., 2014), etc., as some of these have recently become more  
1173 accessible or will become open access soon. Given that, at least at large scales, multi  
1174 model ensembles exhibit better climate forecast skill, it is interesting to investigate if  
1175 that additional skill also propagates into river flow forecasts. While this seems to be  
1176 true for the Eastern United States (Luo & Wood, 2008) it is not known if similar  
1177 conclusions could be drawn for Europe. A similar reasoning can also be extended to  
1178 the hydrological models: using a multi climate model ensemble to force a multi  
1179 hydrological model ensemble might also provide improved skill, as the latter models  
1180 may be complementary in the regions and seasons of best model performance. Bohn et  
1181 al. (2010) showed some advantage of using an ensemble of three hydrological models  
1182 (but with a single forcing), over using only the best of the three, but only after bias  
1183 correcting the hydrological output and making a linear combination of them with  
1184 monthly varying weights.

### 1186 4.3 Implications and recommendations

1187  
1188 Many conclusions drawn from this work are valid at the scale of our domain and not  
1189 necessarily at the scale of river basins. Only in some parts of our analysis, especially  
1190 where we focused on the annual cycle of the skill (Fig. 2), regional patterns at a scale  
1191 smaller than that of the domain were discussed. This was done in a qualitative way.

1192  
1193 For applications of these seasonal forecasts in decision making processes at (sub)  
1194 basin level, a more detailed skill analysis is recommended for that specific (sub)basin,  
1195 preferably after a better model calibration for that same basin. That would probably  
1196 allow not only seasonal predictions of broadly defined anomalies (terciles in our case),  
1197 but also predictions of more absolute discharge magnitudes.

1198 The facts presented in this study that anomaly correlations and ROC scores for the AN  
1199 and BN terciles are significant for large parts of the domain several lead months in  
1200 advance, supported by (fairly) positive validation results for interannual variability of  
1201 high and low flows (Greuell et al. 2016; Roudier et al. 2016), suggest these anomaly  
1202 forecasts are good enough to be used as such. However, areas of significant RPSS are  
1203 much smaller and remain significant for shorter lead times. Spatially distributed  
1204 calibration of VIC model parameters, or distribution based calibration of modelled  
1205 discharge to observed, or both, might also increase the RPSS and for a larger number  
1206 of percentiles. This might then allow forecasting of absolute discharge magnitudes and  
1207 thus inform decision making processes that involve certain absolute discharge  
1208 thresholds.

1209 In the respective Result sections we already discussed the probable reasons for skill,  
1210 which are much elaborated on in the companion paper. In general that paper shows  
1211 that for most areas skill in runoff is caused by initialising snow and /or soil moisture  
1212 properly, only in few areas and seasons skill in precipitation or skill in temperature  
1213 and ET adds to that beyond the first lead month. This has two implications: one is that  
1214 if ever the skill of seasonal climate forecasts improves for Europe of this may well  
1215 translate to improved seasonal river flow forecast too. The second is that better initial  
1216 conditions of snow water equivalent and soil moisture from observations may do the  
1217 same, but the latter only if the spatial distribution of the soil moisture storage capacity  
1218 is more realistic too (see Section 4.1).

1219 ~~In a future extension of this study, an objective method like cluster analysis could~~  
1220 ~~reveal regions where skill has similar signature. This could lead to an improved~~  
1221 ~~assessment of the physical and climatological factors that are responsible for the~~  
1222 ~~spatial variations in skill found in this study.~~

1223 Overall the present analysis shows that especially in winter, spring and early summer,  
1224 there is potentially good skill to forecast runoff and discharge in large parts of Europe,  
1225 with considerable lead time. While this broadly confirms previously published work,  
1226 the present study (while being specific to or model setup) gives much more spatial and  
1227 temporal (season and lead time) details. As such it provides a good basis to support

Commented [RH15]: more suited for companion paper

operational forecasts, and to accompany forecast certainty with forecast skill, important for proper value assessment and finally decision making.

## 5 Conclusions

This paper is the first of two papers dealing with a model-based system built to produce seasonal hydrological forecasts (WUSHP: Wageningen University Seamless Hydrological Predictions). The present paper presents the development and the skill evaluation of the system for Europe, the companion paper provides an explanation of the skill or the lack of skill.

First, “theoretical skill” of the runoff hindcasts was determined taking the output of the reference simulation as “pseudo-observations”. Using the correlation coefficient (R) as metric, hot spots of significant skill were found in Fennoscandia (from January to October), the ~~southern~~southern- part of the Mediterranean (from June to August), Poland, ~~North-northern~~ Germany, Romania and Bulgaria (mainly from November to January) and ~~West-western~~ France (from December to May). There is very little or no significant skill all over the year in some coastal and mountain regions. The entire British Isles exhibit very little skill, except for the east eastern coast of Great Britain. If the entire domain is considered, the annual cycle of skill has a minimum roughly from August to November and a maximum in May.

Runoff and discharge show a high degree of similarity in terms of the spatial patterns and the magnitude of the skill. However, when averaged over the domain and the year, predictability is slightly higher for discharge than for runoff for the first lead month (by 0.049 in terms of R). The difference then decreases with increasing lead time. These tendencies can be ascribed to the combined effect of the delay between runoff and discharge and the fact that skill decreases with lead time. We also found that the difference between discharge and runoff skill increases with the size of the ~~catchment~~basin.

Theoretical skill as determined with the pseudo-observations was compared to actual skill as determined with real discharge observations. On average across all target months and for lead month 2, the ratio of actual to theoretical skill in terms of the domain-mean R is 0.67 (0.26 divided by 0.39) for large basins and 0.54 (0.16 divided by 0.29) for small basins. So, skill reduction due to replacing pseudo- by real observations is larger for small basins than for large basins. For 10 day flow forecasts Alfieri et al. (2014) also found that, especially in mountain areas, performance drops significantly in river basins with upstream area smaller than 300 km<sup>2</sup>.

Skill patterns for the different skill metrics considered in this study (correlation coefficient, ROC area and Ranked Probability Skill Score) are similar to a large

Formatted: Right

1271 degree. ROC areas tend to be slightly larger for the ~~B~~below ~~u~~Normal than for the  
1272 ~~a~~Above ~~u~~Normal tercile but not during target months from October to January.  
1273  
1274  
1275

1276 **References**

1277

1278 Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon,  
1279 P. (2014). Evaluation of ensemble streamflow predictions in Europe. Journal of  
1280 Hydrology, 517, 913-922.

1281 Bierkens, M. F. P., & Van Beek, L. P. H. (2009). Seasonal predictability of European  
1282 discharge: NAO and hydrological response time. Journal of Hydrometeorology, 10(4),  
1283 953-968.

Formatted: English (United Kingdom)

1284 Bruno Soares, M. and S. Dessai (2016). Barriers and enablers to the use of seasonal  
1285 climate forecasts amongst organisations in Europe. Climatic Change 137(1): 89-103.

1286 Crochemore, L., Ramos, M. H., and Pappenberger, F. (2016). Bias correcting  
1287 precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol.  
1288 Earth Syst. Sci. 20(9): 3601-3618.

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

1289 Demirel, M.C., Booij, M.J. and Hoekstra, A.Y. (2015).The skill of seasonal ensemble  
1290 low-flow forecasts in the Moselle River for three different hydrological models.  
1291 Hydrol. Earth Syst. Sci., 19, 275–291, 2015

1292 Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R.  
1293 (2013). Seasonal climate predictability and forecasting: status and prospects. Wiley  
1294 Interdisciplinary Reviews: Climate Change, 4(4), 245-268.

Formatted: English (United Kingdom)

1295 Döll, P., & Lehner, B. (2002). Validation of a new global 30-min drainage direction  
1296 map. Journal of Hydrology, 258(1), 214-231.

1297 ECMWF Seasonal Forecast User Guide, retrieved from;  
1298 [http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-](http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/introduction)  
1299 [forecast-documentation/user-guide/introduction](http://www.ecmwf.int/en/forecasts/documentation-and-support/long-range/seasonal-forecast-documentation/user-guide/introduction)

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: English (United Kingdom)

1300 Ghile, Y. B., and Schulze, R. E., 2008: Development of a framework for an integrated  
1301 time-varying agrohydrological forecast system for southern Africa: Initial results

Formatted: English (United Kingdom)

1302 Greuell, W., Andersson, J. C. M., Donnelly, C., Feyen, L., Gerten, D., Ludwig, F., ...  
1303 & Schaphoff, S. (2015). Evaluation of five hydrological models across Europe and  
1304 their suitability for making projections of climate change. Hydrol Earth Syst Sci  
1305 Discuss, 12, 10289-10330.

1306 Greuell, W., W. H. P. Franssen, H. Biemans and R. W. A. Hutjes. Seasonal  
1307 streamflow forecasts for Europe – II. Explanation of the skill. Submitted to Hydrol.  
1308 Earth Syst. Sci.

1309 Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the  
1310 success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A,  
1311 57(3), 219-233.

Formatted: English (United Kingdom)

1312 Hamlet, A. F., Huppert, D., & Lettenmaier, D. P. (2002). Economic value of long-lead  
1313 streamflow forecasts for Columbia River hydropower. *Journal of Water Resources*  
1314 *Planning and Management*, 128(2), 91-101.

1315 Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New  
1316 M. (2008). A European daily high-resolution gridded data set of surface temperature  
1317 and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*  
1318 (1984–2012), 113(D20).

Formatted: Right: 0 mm, Space Before: 6 pt, Line spacing: 1.5 lines

1319 Juston, J., Jansson, P. E., & Gustafsson, D. (2014). Rating curve uncertainty and  
1320 change detection in discharge time series: case study with 44-year historic data from  
1321 the Nyangores River, Kenya. *Hydrological Processes*, 28(4), 2509-2523.

1322 Koster, R. D., Mahanama, S. P., Livneh, B., Lettenmaier, D. P., & Reichle, R. H.  
1323 (2010). Skill in streamflow forecasts derived from large-scale estimates of soil  
1324 moisture and snow. *Nature Geoscience*, 3(9), 613-616.

1325 Lehner, B., Reidy Liermann, C., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P.,  
1326 Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J., Rödel, R.,  
1327 Sindorf, N., Wissler, D. (2011): High resolution mapping of the world's reservoirs  
1328 and dams for sustainable river flow management. *Frontiers in Ecology and the*  
1329 *Environment* 9(9): 494–502.

Formatted: Right: 0 mm, Space After: 0 pt

1330  
1331 Li, H., Luo, L. and Wood, E.F. (2008). Seasonal hydrologic predictions of low-flow  
1332 conditions over eastern USA during the 2007 drought. *Atmospheric Science Letters*  
1333 9(2): 61-66.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

1334 Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple  
1335 hydrologically based model of land surface water and energy fluxes for general  
1336 circulation models. *Journal of Geophysical Research: Atmospheres* (1984–2012),  
1337 99(D7), 14415-14428.

Formatted: English (United Kingdom)

1338 Luo, L. and E.F. Wood, 2008: Use of Bayesian Merging Techniques in a Multimodel  
1339 Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. *J.*  
1340 *Hydrometeor.*, 9, 866–884, doi: 10.1175/2008JHM980.1

1341 MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M.  
1342 Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier and G. Madec,  
1343 2014. Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal  
1344 forecast system. *QJR Meteorol Soc*, doi:10.1002/qj.2396.

Formatted: English (United Kingdom)

1345 Marchant, R., & Hehir, G. (2002). The use of AUSRIVAS predictive models to assess  
1346 the response of lotic macroinvertebrates to dams in south-east Australia. *Freshwater*  
1347 *Biology*, 47(5), 1033-1050.

1348 Mason, S. J., & Stephenson, D. B. (2008). How do we know whether seasonal climate  
 1349 forecasts are any good?. In *Seasonal Climate: Forecasting and Managing Risk* (pp.  
 1350 259-289). Springer Netherlands.

1351 Mo, K. C., & Lettenmaier, D. P. (2014). Hydrologic prediction over the conterminous  
 1352 United States using the national multi-model ensemble. *Journal of Hydrometeorology*,  
 1353 15(4), 1457-1472.

1354 Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L.,  
 1355 Magnusson, L., Mogensen, K., Palmer, T., Vitart, F. (2011). The new ECMWF  
 1356 seasonal forecast system (System 4). ECMWF Technical Memorandum 656.

1357 [Mushtaq, S., Chen, C., Hafeez, M., Maroulis, J. and Gabriel, H., 2012: The economic  
 1358 value of improved agrometeorological information to irrigators amid climate  
 1359 variability. \*Int. J. Climatol.\*, 32, 567–581.](#)

1360 [Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F.  
 1361 \(2001\). Predicting the discharge of global rivers. \*Journal of Climate\*, 14\(15\), 3307-  
 1362 3323.](#)

1363 Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J., & Schaphoff, S. (2008).  
 1364 Agricultural green and blue water consumption and its influence on the global water  
 1365 system. *Water Resources Research*, 44(9), doi 10.1029/2007WR006331.

1366 [Roudier, P., Andersson, J.C.M., Donnelly, C., Feyen, L., Greuell, W. and Ludwig, F.,  
 1367 \(2016\). "Projections of future floods and hydrological droughts in Europe under a  
 1368 +2°C global warming." \*Climatic Change\* 135\(2\): 341-355.](#)

1369 [Saha, Suranjana and Coauthors, 2014: The NCEP Climate Forecast System Version 2  
 1370 \*Journal of Climate\* J. Climate, 27, 2185–2208. doi: 10.1175/JCLI-D-12-00823.1](#)

1371 [Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., & Lucht, W. \(2013\).  
 1372 Contribution of permafrost soils to the global carbon budget. \*Environmental Research  
 1373 Letters\*, 8\(1\), 014026, doi:10.1088/1748-9326/8/1/014026.](#)

1374 Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., ... & Ogallo, L.  
 1375 (2014). A drought monitoring and forecasting system for sub-Saharan African water  
 1376 resources and food security. *Bulletin of the American Meteorological Society*, 95(6),  
 1377 861-882.

1378 Shukla, S., McNally, A., Husak, G., & Funk, C. (2014). A seasonal agricultural  
 1379 drought forecast system for food-insecure regions of East Africa. *Hydrology and Earth  
 1380 System Sciences*, 18(10), 3907-3921.

1381 Singla, S., Céron, J. P., Martin, E., Regimbeau, F., Déqué, M., Habets, F., & Vidal, J.  
 1382 P. (2011). Predictability of soil moisture and river flows over France for the spring  
 1383 season. *Hydrology & Earth System Sciences*, [16-Discussions](#), 8(4): 201-216.

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: Dutch (Netherlands)



1384 [Svensson, C., Brookshaw, A., Scaife, A. A., Bell, V. A., Mackay, J. D., Jackson, C.](#)  
1385 [R., Hannaford, J., Davies, H. N., Arribas A., Stanley, S. \(2015\). "Long-range forecasts](#)  
1386 [of UK winter hydrology." Environmental Research Letters 10\(6\): 064006.](#)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

1387 [Thiemeßl, M. J., Gobiet, A., & Leuprecht, A. \(2011\). Empirical-statistical downscaling](#)  
1388 [and error correction of daily precipitation from regional climate models. International](#)  
1389 [Journal of Climatology, 31\(10\), 1530-1544.](#)

Formatted: Dutch (Netherlands)

Formatted: English (United Kingdom)

1390 [Trigo, R. M., Pozo-Vázquez, D., Osborn, T.J., Castro-Díez, Y., Gámiz-Fortis, S.,](#)  
1391 [Esteban-Parra, M.J. \(2004\). "North Atlantic oscillation influence on precipitation, river](#)  
1392 [flow and water resources in the Iberian Peninsula." International Journal of](#)  
1393 [Climatology 24\(8\): 925-944.](#)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

1394 [Van Dijk, A. I., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., & Beck, H. E.](#)  
1395 [\(2013\). Global analysis of seasonal streamflow predictability using an ensemble](#)  
1396 [prediction system and observations from 6192 small catchmentbasins worldwide.](#)  
1397 [Water Resources Research, 49\(5\), 2729-2746.](#)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

Formatted: English (United Kingdom)

1398 [Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P.](#)  
1399 [\(2014\). The WFDEI meteorological forcing data set: WATCH Forcing Data](#)  
1400 [methodology applied to ERA-Interim reanalysis data. Water Resources Research,](#)  
1401 [50\(9\), 7505-7514.](#)

Formatted: English (United Kingdom)

1402 [Wood, A. W., & Lettenmaier, D. P. \(2006\). A test bed for new seasonal hydrologic](#)  
1403 [forecasting approaches in the western United States. Bulletin of the American](#)  
1404 [Meteorological Society, 87\(12\), 1699.](#)

1405 [Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J. and Clark, M. \(2016\).](#)  
1406 [Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate](#)  
1407 [Prediction Skill. Journal of Hydrometeorology 17\(2\): 651-668.](#)

Formatted: English (United Kingdom)

1408 [Yuan, X., Wood, E. F., Luo, L., & Pan, M. \(2013\). CFSv2-based seasonal](#)  
1409 [hydroclimatic forecasts over the conterminous United States. Journal of Climate, 26,](#)  
1410 [4828-4847.](#)

Formatted: English (United Kingdom)

1411 [Yuan, X., Wood, E. F., & Ma, Z. \(2015\). A review on climate-model-based seasonal](#)  
1412 [hydrologic forecasting: physical understanding and system development. Wiley](#)  
1413 [Interdisciplinary Reviews: Water, 2\(5\), 523-536.](#)

1414

Formatted: English (United Kingdom)