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. One point of discussion will be highlighted below.

Wrt to the use of the wording pseudo- and real-observations in the title and throughout 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 mainstream 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

1 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, prior 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 earlier 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/](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: 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 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 companion 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.

Page 3, line 15: consider changing the word “simulations” to “forecasts”, as it is confusing otherwise.

Page 3, line 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, line 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 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 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?

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 same skill metrics?

Page 5, line 6: please consider changing the word “simulations” to “forecasts” here.

Page 5, line 7: 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 graph 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, Crochemore et al. (2016), Demir 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 where 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 former 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 I 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 cells 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! Are those numbers equal or different for other target and initialisation months?

Page 8, line 35: I would rephrase the explanation of the here.

Page 8, line 38: which result 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 ROC 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 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 2 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? Thehindcasts 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 lead 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 rephrase 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 time lines, 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 is R?

Could you 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 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 R.

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 left.

Could you please add an x-axis label for both plots to say if these are the initialisation or target months?

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 above each plot.

Because plot d does not show much 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 months?

Would it be worth adding lines for the middle tercile in this same plot?

Page 17, line 10: I think that the “minus” does not belong here.

**Technical corrections:**

- General:
  - Could you please only use of the two terms: “basins” or “catchments”?
  - Please consider changing “lead month” 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 you 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”.
  - Could 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 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° x 0.5° resolution, i.e. ~ 50km) 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
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 I3-4, 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 will 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 I31-33). 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 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? (p5 l23-24). In particular, zero flow simulations can be extremely important to depict droughts. Why excluding them?
  
o How is a skillful forecast defined? (p5 I37-38; p6 I1-2) What is the threshold used to define a re-forecast as ‘skilful’? Is this based on statistically significance test? Is it the value of 0.31 quoted in caption fig 1? This needs to be made clear 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 I34-36)

- Analysis/ interpretation
  
o Influence of catchment size on theoretical vs actual skill (p8 I4-17). 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
catchment sizes, rather than dividing the sample in 2 categories. It also needs to be linked with the model performance.

- 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?

- 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)

- 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 achieve 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’ [p3 l39-40]. 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 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 3.3 out 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
- 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 l9-15, 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 fig 2 shows.

- Figure 3 is excellent.

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

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

Wageningen University and Research

all authors:

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

correspondence to ronald.hutjes@wur.nl
Abstract

Seasonal predictions can be exploited among others to optimize hydropower energy generation, navigability of rivers and irrigation management to decrease crop yield losses. 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 Prediction system), applied here to Europe. The present paper presents the development and the skill evaluation of the system. In WUSHP hydrology is simulated by running the Variable Infiltration Capacity (VIC) hydrological model with forcing from bias-corrected output of ECMWF’s Seasonal Forecasting System 4. The system is probabilistic. For the assessment of skill, we performed hindcast simulations (1981-2010) and a reference simulation, in which VIC was forced by gridded meteorological observations, to generate initial hydrological conditions for the hindcasts and discharge output for skill assessment (pseudo-observations). Skill in hindcasting runoff and discharge is analysed with monthly temporal resolution, up to 7 months of lead time, for the entire annual cycle. Using the pseudo-observations and taking the correlation coefficient as metric, hot spots of significant skill in runoff were identified in Fennoscandia (from January to October), the southern part of the Mediterranean (from June to August), Poland, northern Germany, Romania and Bulgaria (mainly from November to January) and western France (from December to May). Generally skill decreases with increasing lead time, except in spring in regions with snow rich winters. The spatial pattern of skill is fading with increasing lead time but some skill is left at the end of the hindcasts (7 months). On average across the domain, skill in discharge is slightly higher than skill in runoff. This can be explained by the delay between runoff and discharge and the general tendency of decreasing skill with lead time. Theoretical skill as determined with the pseudo-observations was compared to actual skill as determined with real discharge observations from 747 stations. Actual skill is mostly and often substantially less than theoretical skill. This effect is stronger for small than for large basins, which is consistent with a conceptual analysis of the structural differences between the two types of verification. Qualitatively, results are hardly sensitive to the different skill metrics considered in this study (correlation coefficient, ROC area and Ranked Probability Skill Score) but ROC areas tend to be slightly larger for the below normal than for the above normal tercile.
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) 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
observations for verification. We will argue that the fact that the pseudo-observations are obtained with the same model as the hindcasts logically contributes to an overestimation of the skill when the pseudo-observations are used for verification.

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

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

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

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

Svensson et al. (2015) found skilful winter river flow forecasts across the whole of the UK due to a combination of skilful winter rainfall forecasts for the north and west, and strong persistence of initial hydrological conditions in the south and east. Strong statistical correlations between NAO and winter precipitation in Iberia lead to skilful forecasts of JFM river flow and hydropower production (Trigo et al., 2004). Céron et
al. (2010) and Singla et al. (2012) set up a high resolution river flow forecasting system (8 km) over France, for which seasonal climate forecast improved MAM skill over northern France, but worsened it over southern France (compared to a river flow model with proper initialisation of soil moisture, snow etc., but random atmospheric forcing). Demirel et al. (2015) found that both two physical models and one neural network over-predict runoff during low-flow periods using ensemble seasonal meteorological forcing for the Moselle basin, and as a result more extreme low flows are less reliable than more moderate ones.

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” (misleadingly named “true discharge” in Bierkens and Van Beek, 2009; more appropriately “synthetic truth” in Shukla and Lettenmaier, 2011; “reanalysis” in Singla et al., 2012; “a posteriori estimates” in Shukla et al., 2014). We prefer the term “pseudo-observations” over “re-analysis” since the latter has a meteorological connotation that often implies the use of some form of (variational) data assimilation. We did not attempt any form of assimilating observed hydrological variables, such as discharge, in our reference run.

Pseudo-observations have the important 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. Finally, assessment of skill based on pseudo observations excludes model errors from the analysis, which is especially useful when addressing various sources of skill (Wood et al., 2016), something we will do in the companion paper.

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 previously done by Bierkens and Van Beek (2009) and Van Dijk et al. (2013). We will analyse and discuss conceptual differences between using pseudo- and real observations for verification. We hypothesise that the fact that the pseudo-
observations are obtained with the same model as the hindcasts logically contributes to an overestimation of the skill when the pseudo-observations are used for verification.

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

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

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

2 System, models, data and methods of analysis

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

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

In the following subsections we will describe each component.
Figure 10: Diagram illustrating 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: 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.,
for the period of 1979-2010, of which the first two years were used to spin up
the states of snow, soil moisture and discharge, and not used in further analysis. The
reference simulation has the dual aim, namely to create the pseudo-observations for
verification purposes and to create a best estimate of the temporally varying model
state, which is then used for the initialisation of the hindcasts.

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

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

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

VIC was run for the period of the S4 hindcasts (1981 – 2010), for spin up periods. In for the reference simulation two extra years (1979 – 1980) were
simulated to spin up the states of snow, soil moisture and discharge. The hindcast
simulations were initialised with states of soil moisture and snow from the reference
simulation, so for these variables spin up was not needed. However, due to the set-up
of the routing module of VIC, the state of discharge could not be saved and loaded.
Hence to spin up discharge, each 7-month hindcast simulation was preceded by a-one
month simulation with WFDEI forcing. Simulations were performed on a 0.5˚ x 0.5˚
grid for all 15 members of the bias corrected S4 hindcasts. Though the forcing
consisted of daily values, the simulations were done with a three-hourly time step.
Because snow may contribute significantly to the seasonal predictability of other hydrological variables, VIC was run with the option of elevation bands. This means that for each cell calculations were carried out at up to 16 different elevations, with the aim of simulating the elevational gradient of snow. Since the hindcasts cover 30 years with 12 dates of initialisation each and consist of 15 members, a total of 5400 hindcast simulations was carried out. VIC was run in naturalised flow mode meaning that river regulation, irrigation and other anthropogenic influences were not considered.

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

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

All validation results discussed in these two paragraphs are for the VIC model forced by E-obs (v9, Haylock et al. 2008). Our forcing, WFDEI shows higher precipitation (+104 mm/yr) across most of Europe, except the Alps, Scotland and western most Norway. This leads to higher mean discharge, higher inter annual variability and higher Q95 (not Q5) of simulated discharge for almost all stations.
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 and mapped these two station data sets onto the VIC grid with gridded versions with a resolution of 0.5° x 0.5° and a time step of a month. The first contained only observations for catchments larger than 9900 km² (“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 catchment basin 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 only observations for catchmentbasins larger than 9900 km² (“large basins”). The second contained only observations for catchmentbasins smaller than the area of the grid cells, i.e. smaller than about 2530 km² in southern Europe (at 35°N) or < 1050 km² at 70°N (“small basins”). This subdivision was also used 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:
If the nearest model cell did not have an influx from any of the neighbouring cells, its simulations were matched with the observations. Otherwise, all of the eight neighbouring cells without influx were selected. Their simulations were averaged and matched with the observations.

We further discarded all observations with less than 21 years of data within the simulation period (1981-2010) for any of the months of the year. The final data sets within our European domain contained 111 cells with observations for large basins and 636 cells with observations for small basins smaller than a model gridcell. These data sets do not include any variable or parameter characterising the level of human impact. To enable analysis of the effect of anthropogenic flow modifications on predictive skill, we quantified the human impact by performing two model simulations with the Lund-Potsdam-Jena managed Land (LPJmL) model (Rost et al., Schaphoff et al., 2013). This model was operated at the same spatial resolution (0.5° x 0.5°) and with the same river network (DDM30) as VIC, but the former does include dams (GRanD database; Lehner et al., 2011) and associated reservoir management. From the discharge output of a naturalized run and a run with reservoir operation and irrigation, the human impact at cell level was quantified by computing the so-called Amended Annual Proportional Flow Deviator (AAPFD; see Marchant and Hehir, 2002). Subsequently, we selected all discharge observations for large basins with an AAPFD < 0.3, i.e., basins with a relatively small degree of human impact (about half of all 111 basins).

2.3 Methods of analysis

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

Instead of analysing skill per target season and/or for a number of consecutive lead months, we analysed skill for every combination of per 12 target and per 7 lead months. The thus achieved higher temporal resolution of the skill metrics enables a more accurate determination of the beginning and end of periods of skill. Moreover, skill at a monthly resolution provides the possibility to determine the consistency of the skill where we define consistent skill as skill that persists during at least two consecutive target or lead months. In accordance with Hagedorn et al. (2005) we designated the first month of the hindcasts as lead month zero, so target month number...
is equal to the number of the month of initialisation plus the lead month number. In discussing the results we will pay relatively little attention to lead month zero because seasonal prediction deals with forecasts beyond the first two weeks.

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

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

To a large extent, we found that our results and conclusions, in terms of spatio-temporal patterns of skill, are independent of the chosen metric. Hence, and because among the three metrics the correlation coefficient is the easiest to understand, we will
discuss results mostly in terms of the correlation coefficient, which is in line with
Doblas-Reyes et al. (2013). The sensitivity to the chosen metric and significant
differences between these metrics will be discussed in Sect. 4.4.2.

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

3 Results

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

3.1 Spatiotemporal variation of skill in runoff forecasts

Eighty-four maps of skill of the runoff hindcasts were drawn for all 12 initialisation
months of initialisation and all 7 lead months (all are presented in supplementary
material S1). Two cross-cuts through that collection are shown in Figs. 1 (for a single
initialisation month) and 2 (for a single lead month). The seven panels of Fig. 2 show
the skill of the hindcasts initialised on April 1 as a function of lead time. Cells with an
insignificant amount of skill are tinted yellow; cells where no metric could be
computed remain white. In lead month 0, significant skill is found across almost the
entire domain (99% of the cells). After the first lead month, the fraction of cells with
significant skill gradually decreases to reach 16% at the longest lead time (lead month
6). This is more than expected for the case of completely unskilful simulations (5% of
the cells), so at the end of the hindcast simulations significant skill that does not occur
due to chance is still present in some regions. The general impression is that the
pattern of skill does not move in space but that skill is fading, i.e. for individual grid
cells R is mostly decreasing with increasing lead time.
Figure 1: Skill of the runoff hindcasts initialised on April 1 for all seven lead months. Skill is measured in terms of the Pearson correlation coefficient between the median of the hindcasts and the observations (R). White, terrestrial cells correspond to cells where observations or hindcasts consist for more than one third of zeros or one sixth of ties. The threshold of significant skill lies at 0.31, so yellow cells have insignificant skill, dark red cells have (most) skill. White, terrestrial cells correspond to cells where observations or hindcasts consist for more than one third of zeros or one sixth of ties. The legend provides the fraction of cells with significant values of R (at the 5% level) and the domain-averaged value of R.
Figure 2: Annual cycle of skill (R) of runoff hindcasts of lead month 2. More explanation is given in the caption of Fig. 1.

The twelve panels of Fig. 2 show the annual cycle of skill of the hindcasts for lead month 2. Consistent skill (persistent during at least 3 consecutive target months) is found in (causes of skill are reproduced here from the companion paper, Greuell et al., 2017):

- Fennoscandia. Much skill is present during the entire year, except for November and December, and there is a dip in skill in April. On average across the entire region, skill reaches a maximum in May and June, i.e. the end of the melting season, and as shown in the companion paper—largely due to initialising snow.
Compared to the rest of the peninsula, there is generally less skill along the Scandinavian Mountain range. The companion paper shows some evidence this may be due to high variability of orographic rain, ill-represented in the re-forecasts.

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

- Western France, more or less from Paris to Brittany and roughly from December to May. Skill derives from initialisation of soil moisture.

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

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

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

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

These results can be compared to those of Bierkens and Van Beek (2009). They found maxima in predictability of winter discharge in North Sweden, Finland, the region between Moscow and the Baltic Sea, Romania and Bulgaria, and East Spain. For the winter there is crude agreement with the current study about North Sweden, Romania and Bulgaria but not about the other regions. For the summer, Bierkens and Van Beek (2009) compute maxima in skill for South Spain, Sardinia, West Turkey and South-west Finland. This pattern agrees to some extent with the locations of the summertime maxima in skill of the present study (most of Fennoscandia and southern part of the Mediterranean region).
Figure 3: Number of months in which the year with significant skill (R) in the runoff forecasts of lead month 2.

Figure 3 displays a synthesis of Fig. 2 in the form of a map with the fraction of the 12 months of the year with significant skill for lead month 2. Many of the regions with very little or no skill all over the year are coastal regions (e.g., northern coast of...
Spain), especially coastal regions on the western side of land masses (e.g. 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.
Figure 4: At left a) Fraction of cells with significant skill (in terms of R), and 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) shows 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 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 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).
Similar summary plots for the other skill metrics are presented in the supplementary material S2, and discussed in section 3.4.
3.2 Spatiotemporal variation of skill in discharge forecasts

Figure 5: Comparison of the skill of the hindcasts of discharge and runoff. The two maps display R for runoff (a) and discharge (b) for hindcasts initialised on May 1 and target lead month 2 (July), (see further explanation in Fig. 1).
Panel c depicts the annual cycle of the domain-averaged R for runoff (red) and discharge (blue) for lead months 0 to 4. The horizontal line at 0.31 is the threshold of significance for a single cell. Panel d is a box plot of the difference between R for discharge and runoff as a function of the catchment basin size. Each bin i contains the results for all catchment basins with a maximum of 2^i cells and more than 2^{i-1} cells, e.g. bin 4 is for all catchment basins with a size from 10 to 16 cells. Boxes represent the interquartile range and the median; and whiskers extend to minimum and maximum by 1.5 times the interquartile range from the box top and bottom. Values found in the bin. All values are average differences over the twelve months of the year and results are shown for three different lead times. The value above the abscissa give the number of cells in each bin.

3.2 Spatiotemporal variation of skill in discharge forecasts

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

Domain summary statistics of skill also differ slightly between runoff and discharge. Figure 5c compares the annual cycle of the skill in discharge with the skill in runoff at five different lead times. Here we show the difference in the domain-averaged R instead of the fraction of cells with a significant R because in lead month 0 that fraction is close to one for both variables. In terms of the domain-averaged R, predictability is higher for discharge than for runoff for the first lead month. On average over the 12 months of the year, the difference is 0.049. We ascribe this result to the combined effect of the delay between runoff and discharge and the general tendency of decreasing skill with lead time. The curves for the different lead times in Fig. 5c show that the difference in skill between the two variables gradually disappears
with increasing lead time (an annual average of 0.020 and 0.012 for lead months 1 and 2, respectively). This is compatible with the given explanation for the difference and the fact that the rate with which skill is lost gradually decreases with increasing lead time.

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

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

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

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

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

Figure 7: Comparison between verification of discharge with pseudo- (red) and real (blue) observations in terms of the annual cycle of the domain mean R. The horizontal line at 0.31 is the threshold of significance for a single cell.
Results are shown for cells representing large basins (left) and cells representing small basins (right). Both panels depict cycles for lead months 0 and 2 only.

Comparing skill for small basins with skill for large basins in Fig. 7, we notice two differences. Firstly, in terms of the domain mean $R$, theoretical skill is higher for large basins than for small basins (0.39 and 0.29, respectively, for the annual mean and lead month 2). However, this result holds for the cells with observations. If all cells of the domain are considered, this difference becomes insignificantly small and almost vanishes. On average, all cells with an upstream catchment larger than 10000 km$^2$ have a mean $R$ of 0.396 and all cells with an upstream catchment smaller than 2500 km$^2$ have a mean $R$ of 0.384. So, the apparent difference in theoretical skill between large and small basins can be blamed on the geographical distribution of the discharge monitoring stations, with stations on small basins being relatively more often located in regions with relatively little skill like Germany, France and the British Isles than large basin stations.

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

Finally, we investigated to what extent these results are affected by human interference, keeping in mind that the simulations are naturalized, while the observations include human impacts to a variable but unknown degree. Human interference is expected to have a negative effect on actual skill and hence on the ratio of actual to theoretical skill. We quantified the human impact by performing two model simulations with the Lund–Potsdam–Jena managed land (LPJmL) model (Rost et al., Schaphoff et al., 2013) that was operated at the same spatial resolution (0.5° x 0.5°) and with the same river network (DDM30) as VIC. From the discharge output of a naturalized run and a run with reservoir operation and irrigation, the human impact at cell level was quantified by computing the so-called Amended Annual Proportional Flow Deviator (AAPFD, see Marchant and Hehir, 2002). Subsequently, we selected all discharge observations for large basins with an AAPFD $< 0.3$, i.e. basins with a relatively small degree of human impact (about half of all 111 basins). For relatively natural basins (AAPFD $< 0.3$; see end of section 2.2), this selection the ratio of actual skill.
to theoretical skill was computed in terms of the domain mean $R$, averaged across all target months and for lead month 1. We found a ratio of 0.686, which should be compared to a ratio of 0.667 for the entire set of large basins (see above). So, as expected the ratio is larger for basins with less impact. However, since the difference between the two ratios is small we conclude that the effect of evaluating the combination of naturalised runs against observations that are obviously affected by human interference, contributes only little to the difference between actual and theoretical skill. A similar analysis was not applied to the collection of small basins with observations, since these are generally smaller than the spatial resolution of the simulations.

Comparing skill for small basins with skill for large basins in Fig. 7, we notice two differences. Firstly, in terms of the domain mean $R$ theoretical skill is higher for large basins than for small basins (0.39 and 0.29, respectively, for the annual mean and lead month 2). However, this result holds for the cells with observations. If all cells of the domain are considered, the difference almost vanishes. On average, all cells with an upstream catchment larger than 10000 km$^2$ have a mean $R$ of 0.306 and all cells with an upstream catchment smaller than 2500 km$^2$ have a mean $R$ of 0.284. So, the apparent difference in theoretical skill between large and small basins can be blamed almost entirely to the geographical distribution of the stations, with small basin stations being relatively more often located in regions with relatively little skill like Germany, France and the British Isles than large basin stations.

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

### 3.4 Results for other skill metrics

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

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

The agreement that we find between the patterns of the different metrics is in accordance with a result mentioned in a global analysis of seasonal streamflow predictions by Van Dijk et al. (2013) who found high spatial correlation between the different skill metrics they used (among which R, the RPSS and the ranked correlation coefficient).
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 below-normal tercile, c) the Ranked Probability Skill Score (RPSS) and d) the difference in ROC area between the
Below-Normal and Above-Normal terciles. In panels a, b and c skill is not significant in cells with a yellow colour. Legends provide the fraction of cells with significant values of the metric and the domain-averaged value of the metric.

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

Figure 9 compares the BN with the AN tercile in terms of the fraction of cells with a significant ROC area across all target and initialisation months. The main finding is that in all combinations of lead and target month the fraction of significant cells is larger for the BN than for the AN tercile. This is perhaps not as expected from the VIC performance in reproducing historic flows, which is better for high flows than for low flows (Greuell et al., 2015; recall that their high/low flows are defined as p95 and p5, respectively, while here they are p67 and p33; see also Section 2.1). However, the AN and BN two fractions tend to become equal (i) when these ROC areas approach 1.0, (ii) when they approach the limit of no skill (5%) and (iii) during target months from October to January.
Figure 9: Skill of the runoff hindcasts in the Below Normal (BN) minus compared to the skill of the runoff hindcasts in the Above Normal (AN) tercile. The plot depicts annual cycles of the fraction of cells with a significant ROC area for the two terciles and for four lead months.

4 Discussion

For verification of the hindcasts two options were considered in this paper. We determined the skill of the hindcasts by comparing predicted discharge with the output of the reference simulation (the “pseudo-observations” leading to “theoretical skill”) and with observations of real discharge (“real observations” leading to “actual skill”). To obtain a basis for understanding the differences in skill that we found, Fig. 10 presents a streamflow diagram of the three relevant physical systems, namely the real world and the model systems that generate the hindcasts and the pseudo-observations. In each system, confined in the diagram by a box, meteorological and initial conditions force and initialize hydrology, of which discharge is the relevant component here. There are two complications. First, the initial conditions themselves are generated by meteorological forcing during the spin up period, initial conditions at the beginning of the spin up period and hydrology. This is represented by the upper left branch in each box, omitting initial conditions at the beginning of the spin up period for simplicity. Second, due to measurement errors real observations of
discharge generally differ from real discharge (Juston et al., 2014) as illustrated in the upper right corner of the figure.

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

2. Models are imperfect, in terms of physics and in terms of spatial and temporal discretisation, so model hydrology differs from real world hydrology. Hindcasts and the pseudo-observations are produced with the same model, so imperfections in model hydrology are expected to lead to more theoretical than actual skill. One assumption implicitly made in the diagram is that the basin of the observation station and the model basin are identical. This is not the case (see Sect. 2.2), so differences between observation and model basin form an additional cause of disagreements between theoretical and actual skill. Again, this will favour theoretical skill with respect to actual skill since basins are identical in the hindcasts and the reference simulation. In particular, differences in meteorological forcing between the basin of the observation station and the model basin 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.

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

4. Initial conditions are absent in this list of differences since in WUSHP they are not independent components but entirely determined by two components of the system listed above, namely meteorology and hydrology. Alternatively, initial hydrological conditions could be taken from observations or by assimilation of observations into model calculations. In that case, initial conditions would become an independent or semi-dependent component of the system. However, again, while model initial conditions would, of course, differ from real initial conditions, the two model system had identical initial conditions.
Hence, again, this difference would again be expected to lead to more theoretical than to actual skill.

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

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.

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

4.2 Results and uncertainties

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

This prior work also invokes some warnings. Greuell et al. found that seasonal flow cycles show a too late and too broad spring peak in (northern ) Fennoscandia. This suggests that our theoretical forecast skills may also be too high at too long lead times in that region and season, (as was also already revealed by comparing Figure 6b vs 6c).
In a future extension of our work, an objective method like cluster analysis could reveal regions where skill has a similar signature. This could lead to an improved assessment of the physical and climatological factors that are responsible for the spatial variations in skill found in this and its companion paper.

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

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

Other seasonal forecasting systems, based on different couple ocean-climate models, exist that could have been used, such as CFSv2 (Saha et al., 2014), GloSea5 (MacLachlan et al., 2014), etc., as some of these have recently become more accessible or will become open access soon. Given that, at least at large scales, multi model ensembles exhibit better climate forecast skill, it is interesting to investigate if that additional skill also propagates into river flow forecasts. While this seems to be true for the Eastern United States (Luo & Wood, 2008) it is not known if similar conclusions could be drawn for Europe. A similar reasoning can also be extended to the hydrological models: using a multi climate model ensemble to force a multi hydrological model ensemble might also provide improved skill, as the latter models may be complementary in the regions and seasons of best model performance. Bohn et al. (2010) showed some advantage of using an ensemble of three hydrological models (but with a single forcing), over using only the best of the three, but only after bias correcting the hydrological output and making a linear combination of them with monthly varying weights.
4.3 Implications and recommendations

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

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

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

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

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

Overall the present analysis shows that especially in winter, spring and early summer, there is potentially good skill to forecast runoff and discharge in large parts of Europe, with considerable lead time. While this broadly confirms previously published work, the present study (while being specific to or model setup) gives much more spatial and temporal (season and lead time) details. As such it provides a good basis to support
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 part of the Mediterranean (from June to August), Poland, northern Germany, Romania and Bulgaria (mainly from November to January) and 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 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 km².

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
degree. ROC areas tend to be slightly larger for the Below Normal than for the Above Normal tercile but not during target months from October to January.
References


