

Review of “Seasonal streamflow forecasts for Europe - I. Hindcast verification with pseudo- and real observations”

By Wouter Greuell et al.

Summary:

This is a detailed analysis of the quality of the pan-European WUSHP seasonal runoff and discharge forecasts, measured against pseudo- and real observations. The analysis is complete and elaborate and supports the authors' conclusions. The paper could however benefit from being more logically structured, as well as more concise in places. I therefore recommend this paper to be published after minor revisions from the authors.

Major comments:

1) References to other studies should be concise in the introduction, methods and results sections of the paper. Results from the relevant literature should be described and discussed more fully in the discussion, in order to mention how these results compare to the results presented in this paper. For example: on P3L80-P4L116, P8L287-P9L315 and P16L530-550.

2) This paper contains plenty of interesting and significant results which are overall clearly described. However, the varying choice of target month and lead month used as examples in the results appears unjustified (e.g. April initialisation for Fig.2 vs. May initialisation for Fig. 6a and 6b, lead month 2 for Fig. 3, 4, 6a and 6b vs lead months 0 to 2 for Fig. 5 and lead months 0 to 4 for Fig. 6c, etc). Why aren't the same months display all the time? Could you please justify your choice here and/or in the paper?

3) The (dis)similarity of the results for other initialisation months/lead months/variables is not always discussed, and the relevant supplementary figures are not always referred to in the text. For example on P12L457-467: how do the results from Fig. 2 compare to other initialisation months? No reference is made to the corresponding supplementary figures. On P14L484-512: how do the results from Fig. 3 compare to other lead months? No reference is made to the corresponding supplementary figures. On P15L519-P16L528: how do the results from Fig. 4 compare to other lead months? I don't see any supplementary figures for this figure, why is that? In Section 3.4, the results for other skill metrics are shown and discussed for runoff only. How do these compare to the results for discharge? Can similar conclusions be made?

4) I find the presentation of some of the results not very straightforward. Indeed, in some instances, an example is used to describe the results in details, before giving a general overview of the results. For example, I would have found it more logical to discuss and show the general overview of skill through time with Fig. 5 before presenting the spatiotemporal variation in skill with selected examples (specific initialisation and lead months) presented in Fig. 2, 3 and 4. On P18L600-P19L614: I would have found it here once again more logical to discuss and show the general overview of the difference between discharge and runoff skill (Fig. 6c and d) before discussing the specific examples contained in Fig. 6a and b.

Specific (minor) comments:

- The abstract is currently too detailed and would greatly benefit from being shortened.

- It is confusing to read the word simulation referring to the hindcast runs. E.g. P2L22 (“hydrology is simulated”) or P2L25 (“hindcast simulations”), and throughout the paper. Please rephrase those instances.

- P2L27: you mention discharge but not runoff, which should probably also be mentioned here.
- P2L43-44: the sentence “which is consistent [...] two types of verification.” does not make sense to me. Consider removing it as not crucial here or rephrasing it.
- P3L62: I am not sure what is meant by “legitimate” here, consider choosing another word.
- P3L63-70: I would move the few lines “In this paper [...] adopted here as well” to the end of the introduction, before you present the aims of the paper, and after you have introduced the subject area.
- P3L76: consider adding the paper by Arnal et al. (in review), under review in this special issue, to your list of references for European studies.
- P3L80-81: are the meteorological hindcasts used by Thober et al. (2015) seasonal? Please specify.
- P3L93-P4L102: the skill of which variable(s) is this referring to?
- P4L100: you haven’t introduced what NAO stands for.
- P4L120: please specify what the timescale of “medium range forecasting” is.
- P4L128-130: using the words “misleadingly” and “more appropriately” is not appropriate here. Please remove them and simply mention the different names given to pseudo-observations by using more objective words, such as “also called ...”
- P4L140: the assessment of skill from pseudo-observations does not exclude model errors per se. Indeed, model errors are variable depending on the magnitude of the flow and can hence be different between the pseudo-observations and the hindcasts. Please consider rephrasing this sentence to reflect this.
- P5L142: please add the reference to the companion paper here.
- P5L145-154: I would suggest to move the sentences that relate to this paper’s methods from this paragraph to the following paragraph. The hypothesis could even be moved to the methods section of this paper. The reference to literature where both theoretical and actual skill are assessed can however be kept here.
- P5L158: replace “maximum resolution, i.e. at monthly resolution” with “a monthly resolution” simply. It is confusing what maximum means and the resolution could be higher than monthly aggregations.
- P5L160: replace “for the more” with “using”. Using the term “more” does not make sense here, as the correlation coefficient is not a probabilistic metric in the first place.
- P5L173-179: consider moving this paragraph to a more suitable place, e.g. Section 2.3.
- P6L211-212: the point of this sentence is not clear. I would replace it with the hypothesis from P5L151-154, which makes a clear point on the impact that this may have on the results.
- P7L245: please specify which parameters of the WFDEI are used.
- P8L269: it would be good to cite one or multiple papers that support the statement that bias correction generally improves forecasting skill.
- P8L277-279: these lines can be removed here as already previously said in this section.

- P9L324: I suppose “size” refers to the basin size?
- P10L372: I would be interested to know more about the AAPFD. Could you please provide one or two lines of explanation as to what this is?
- P11L392: it might not be clear to some readers what target and lead months refer to. Adding a short explanation of those terms here would be good.
- P11L398-399: please note that this explanation does not work >12, consider rephrasing.
- P11L401: reiterate what the benchmark for the measurement of skill is here.
- P11L405-409: it is not clear to me whether the observation terciles or the hindcast terciles were chosen as thresholds to calculate the forecast probabilities used in the ROC area computation.
- P11L422: I believe the RPSS is not only a measure of forecast accuracy. It is also for example sensitive to the forecast sharpness.
- P12L443-445: why are hindcasts or observations with more than 1/6th of ties discarded?
- P12L465-467: although there is indeed a general impression that skill is fading, it does not appear entirely true for parts of Fennoscandia. What could it be due to?
- P14L487-512: these points would be easier to read if summarised in a table.
- P15L500 and P15L509: “up to lead month 2” does not make entire sense when talking about results from Fig. 3, as it only displays results for lead month 2. Please rephrase or refer to supplementary figures.
- P19L612: mention where the Loire is, as not every reader will be familiar with this river.
- P20L666-667: the reduction in actual skill for Fennoscandia also appears true for small river basins.
- P20L670-P21L671: in Central Europe, some useful skill remains, both for small and large basins, when using real observations.
- From Fig. 7d and e, it seems that there is more skill when using real observations than pseudo-observations for small basins in NW France and SE England. What could that be due to?
- P21L675-P22L712: does the difference between theoretical and actual skill keep on decreasing for lead months > 2?
- P21L677: please specify what the reduction in skill refers to: i.e. between theoretical and actual skill.
- P21L680-681: the exact numbers which produce the ratios are not necessary to mention.
- P22L704-705: this appears evident for lead month 0 but not for lead month 2.
- P22L709-710: this sentence is not completely clear. Consider rephrasing to for example: “there is a linear relationship between the ratio of actual to theoretical skill and the hindcast skill.
- P22L745: I would remove this part of the sentence (“and thus such cases are likely due to chance”). The skill metrics all capture different attributes of the forecast and are thus not likely to be 100% similar. This is probably what causes the <5% dissimilarities you mention.
- P23L767: “0.5 corresponds to climatological forecasts” should be moved to the caption of Fig. 9.

- P24L774-778: it is very difficult to see the blue and red patterns you are mentioning here. Fig. 9d could be improved by, for example, colouring all positive [negative] values with the same shade of blue [red].
- P21L781-784: you already mention this on P23L767-769. The information could be combined.
- The discussion is overall interesting and philosophical in parts. It could however benefit from being shortened in some places, especially Section 4.1, to have more impact.
- P25L807: does “diagram” refer to Fig. 1?
- P25L821-826: these sentences could be removed or shortened as there is no clear point to them. What follows is a clear point.
- P25L842: “might reduce actual skill”, as it is not sure.
- P28L950-951: what is the 12 month forecast you are referring to?
- P30L1033-1035: I would suggest removing this sentence as it is a discussion rather than a conclusion and lengthens the conclusion.
- P30L1035-1036: this is especially true at short lead times and could be mentioned here.
- P30L1039-1042: these numbers are specific to lead month 2 and therefore not general enough for the conclusion. It could be removed from the conclusions.
- P30L1043-1045: this information should be moved to the discussion section of this paper.

Figure 1:

- Red and green should never be used jointly on a figure (similar comment for Fig. 6 and S3).
- Please specify what the ECMWF hindcasts and the hydrological model are on the figure.
- Change “real observations discharge” to “real discharge observations”.
- Add runoff to the model hindcasts and reference run boxes.
- “and Sect. 4.1 for further discussion” can be removed from the caption.

Figure 5: “a)” and “b)” are missing from the figure titles.

Figure 7: is figure a) really necessary here?

Supplementary figures: the text is quite small on Fig. S1.

Technical corrections:

- P2L24: “the bias-corrected output”.
- P2L25: change “performed” to “produced”.
- P2L29: “hindcasting” is not generally used as a verb like it is here. Consider rephrasing to “The skill in the runoff and discharge hindcasts”. Same holds in all other instances where this is done in this paper.
- P2L37: “the skill” (2 instances).
- P2L40-41: “Actual skill”.

- P2L45: "leads".
- P2L47: "in the following order:".
- P3L73: remove "forecasting model suite he".
- P3L75: "Seasonal hydrological forecasting systems".
- P3L81: change "of" to "from".
- P3L87: "the southern".
- P3L88: "the Alps".
- P3L89: "the Pyreness".
- P4L107: "the NAO".
- P4L108: "Céron".
- P4L110: "the seasonal climate forecast" and "the MAM".
- P4L140: "pseudo-observations".
- P7L233: add ° between the numbers and the letters that refer to the domain extent.
- P7L234: the ")" is missing.
- P7L237: "a three hourly time step" and "The output".
- P7L238: "at a daily resolution".
- P9L312: "except for the Alps".
- P9L322: remove "proper".
- P9L326: "the second for basins".
- P11L403-404: "shortly" does not make sense here, use for example "referred to as".
- P11L414: "river flows".
- P12L454: change "drawn" to "produced" for example.
- P12L457: "3" instead of "2" and "Fig. 2" instead of "Fig. 3".
- P14L489, P15L497, P15L499, P15L501, P15L510: "the skill".
- P15L496, P15L499: "the initialisation".
- P15L511: "soil moisture initialisation".
- P15L512: "the precipitation".
- P16L544: "which lasts until about".
- P17L554: "the left-hand side figure" instead of "At left". Same for "at right".
- P17L558-559: "namely on the left-hand side and the right-hand side".
- P17L571: "implying that there".

- P18L600: "the skill" (2 instances).
- P19L639: remove the extra ".".
- P19L647: "databases".
- P20L655: "b) as a)".
- P20L670: "and for too long".
- P21L671-672: "where VIC reproduced well".
- P21L672: "low flows were".
- P24L783: "the fraction of".
- P26L879: remove extra space before the comma.
- P27L914: "to contradict with our finding that".
- P27L919: add space after "Q33".
- P27L924: "Greuell et al. (2015)".
- P27L926: remove comma after "season".
- P29L973: "Fig. 3".
- P30L1017: "the skill or lack thereof".
- P30L1019: "was determined using the output".
- P30L1025: "mountainous regions".

References:

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci. Discuss.*, <https://doi.org/10.5194/hess-2017-610>, in review, 2017.