Reply to reviews of second submission

Referee 1, Major comments

1) Naturally, various styles with respect to the use of citations in various parts of a manuscript exist. It is definitely not uncommon to make the introduction a concise review of previous work and to identify therein omissions/open questions that will be addressed in the new paper. Moreover in previous reviews we were actually asked to elaborate more on this. In the introduction we attempted to describe how previous studies generally focussed on smaller (sub-) regions in Europe, on particular seasons (so multiple target months combined) or on selected lead times. So we identify such limitations and from there formulate our own objectives for this paper. In the methods section we describe our system which naturally is a continuation/expansion of previous developments that need a brief description and associated citations. Given the length of the results section we prefer to incorporate descriptive comparisons to findings by others (and thus citations) directly there and leave discussion of possible reasons for consistency or inconsistency with results by others to the discussion section.

2) There is no particular reason to choose April initialisation in Fig.2. May initialisation was chosen in Fig. 6 because in the map some parts of the Danube and the Loire nicely contrast with the surrounding cells, as discussed in the paper.

   We often chose to show and discuss results for lead month 2 and added the following text where Figure 3 is discussed: *because at that lead time approximately 50% of the cells have significant skill.*

   In Figure 5 the number of dashed lines (for specific lead times) was limited to three as more lines make the figures too chaotic. Dashed lines for lead times > 2 months were omitted since these are very close together.

   Regarding the choice of the lead times in Figure 6c we have added in the paper that *the difference in skill between the two variables gradually disappears with increasing lead time*. For lead month 4 the difference is already very small, so it does not make sense to also show results for even longer lead times. More curves would again make the figure chaotic.

3) Regarding Fig. 2 we added: *The same holds for initialisation in other months (see Fig. S1 in the supplementary material), with important exceptions better identified with Fig. 5 and discussed there.*

   Regarding Fig. 3 we added: *Inspection of Fig. S1 leads to the conclusion that to a first approximation regions with skill at other lead times are equal to those listed above for lead month 2 but that the magnitude of skill decreases with increasing lead time as demonstrated in Fig. 2. (keep in mind that a change in lead time corresponds to a change in target time by the same amount). To give an example: for lead month 3 patterns in the skill maps look similar to*

those provided in Fig. 3 but colours are fainter and target months shift by one month ahead. There are many exceptions to this general rule, e.g. skill due to snow melt that suddenly appears at the end of the melt season at longer lead times while it was not present during the lead months before (see Fig. 5 and the companion paper). A more detailed regional analysis of some of these features is left for future case studies.

As to Fig. 4, we have replaced the graph for lead month 2 only by a six panel graph showing results for lead months from 1 to 6 and slightly adapted the text describing the conclusions drawn from the figure.

Indeed, Sect. 3.4 compares different metrics for runoff only. We have checked similar figures for discharge and found negligible differences with runoff. We stated this in the text: *We would finally like to note that, while in this sub-section we discussed runoff, we made similar figures and calculations for discharge. Results for these two variables are almost identical.*

4) In our opinion this is a matter of taste and both our strategy and the strategy proposed by the reviewer are good. Each map in Figs. 2 and 3 is summarised in a single point in Fig. 5. The advantage of our approach is that Fig. 5 is easier to understand after first looking at the maps (Fig. 2 and 3). A similar argument holds for Fig. 6.

We have used almost all of the specific comments to improve the paper. For details, see our annotated reply to the word document written by referee #?.

Referee #2

- The sentence that explains why skill in discharge exceeds skill in runoff during the first lead month was omitted from the abstract since the abstract was shortened by only mentioning major results. In order to specify what is meant by delay we changed a similar sentence in Sect. 3.2 as follows: *We ascribe this result to the combined effect of the delay between runoff and discharge, with variations in discharge being later in time than the corresponding variations in runoff, and the general tendency of decreasing skill with lead time.*
- The abstract was shortened.
- We have modified the structure of Section 2.1. The first paragraph provides an overview of the system (two types of simulations), the next describes the reference simulation. The topic of paragraphs 3 (overview), 4 (S4) and 5 (bias correction) are the hindcasts. The remaining paragraphs deal with VIC, namely an overview of the simulations (paragraph 6), some settings used in VIC (7) and validation of VIC (8 to 10).
- Fig. 5b. The referee was correct. August should be September. We changed this.
- Fig. 3. Done.
- Sect. 3.3. Reminder was added.

We have used all of the specific comments to improve the paper.

Editor

- Better linking the two parts

  In this manuscript results from the "companion paper" are now mentioned 12 times, spread through the manuscript. We particularly like to mention the links that we made in the list of regions with skill in Section 3.1, where we added information on the sources of that skill, a result from the companion paper.

- Comparing performance and forecasting skill
  The first three paragraphs of Section 4.2 deal with this topic and have been rewritten to a large extent:

  There seems to be a broad correspondence between the probabilistic forecast verification presented here and the model validation presented in Greuell et al. (2015) and Roudier et al. (2016). These studies found that average discharge and inter-annual variations therein are well reproduced against observations, consistent with our result that in the first lead month all skill scores, also against real observations (see Fig. S4 for the lead 0 results), are good for large parts of Europe.

  However, the relation between a model's ability to simulate historic streamflow and its ability to generate skill in seasonal forecasts is complex. There is, for instance, no reason to expect that regions with more theoretical skill than other regions would generally correspond to regions with better historic streamflow simulations. Large model stores of soil moisture and snow tend to lead to more theoretical skill, whether these stores are realistic or not. If they are not realistic, simulations of historic streamflow will be poor, despite the forecast skill. Another example of the problematic relation between validation and verification is that, even in perfect models, regions with small model stores of soil moisture and snow and regions with large interannual variation in precipitation will exhibit small amounts of theoretical and actual skill. So, regions with high quality historic streamflow simulations may for good reasons have little skill in the forecasts.

  However, what we would expect is that regions of poor model performance have little actual skill (not necessarily little theoretical skill) in the forecasts. In our work, this statement is broadly confirmed by the basins in northern Fennoscandia, which lose much of their skill when using actual instead of pseudo-observations (Fig. 7). In this region VIC indeed performed poorly in reproducing historic flows. Good model performance probably is a necessary (but not sufficient) condition for the generation of actual skill in seasonal forecasts. This is exemplified by some regions with considerable amounts of actual skill in

central Europe (e.g. northern part of the Balkans and the Elbe basin in Fig. 7), where VIC's simulations of historic streamflow are much better than in northern Fennoscandia.