Reply to reviews explanation paper

On the basis of the suggestions made by the referees, we substantially improved this paper. Some important changes are:

- Skill of forcing is determined after bias correction
- New names for specific hindcasts
- Better explanation of the set-up of the specific hindcasts and better motivation of their set-up
- Thorough discussion of the differences between our specific hindcasts and ESP
- We produced an ESP which we compared with the InitSH (previously called ESP)
- Better motivation for the choice of metrics, e.g. why they are insensitive to biases
- Addition of versions of Figure 4 and 5 for other metrics to the supplementary material
- Addition of an analysis of the reliability of the system (Appendix B)
- Addition of a description of the detrending procedure
- Addition of some discussion about possible future post-processing

We address most of the minor points of referee 1 in a reply to the document written by the referee. The remaining, more important points, are addressed in this document.

We put the analysis of reliability in an appendix to this paper. However, we feel that this analysis would fit better into the first paper. Would it be possible to publish it as an addendum to the first paper?

Referee 1

Major point 1) Type of forcing that is verified

In the evaluation of the skill of the meteorological forcings in Section 3.1, we have replaced the non-bias-corrected data by bias-corrected data, as suggested by the referee. Since we believe that the evaluation of the forcing has considerable intrinsic value (meteorologists might find it interesting), we added an Appendix (A) about the skill in the raw S4 output. Differences in skill between bias-corrected and raw forcing data are small or negligible.

Major point 2) About the ESP experiments.

First of all, we have given our specific hindcasts new names that do not contain any reference to ESP, instead of the ESP names that we used before (exception is the newly produced conventional ESP, which we just call "ESP"). Despite this renaming, our specific hindcasts resemble ESP experiments to a considerable degree. We briefly point this out in the introduction and then discuss differences between our specific hindcasts and ESP extensively in the discussion section (4.3 Relation of the specific hindcasts with ESP).

We have also rewritten in the introduction the paragraphs that introduce our experiments and refer to the ESP. Section 4.3 demonstrates that our InitSH (formerly called ESP) and the usual ESP yield almost identical skill, since both have meteorological forcing that does not vary from year to year. On the other hand we describe that the skill from our MeteoSH (formerly called reverse-ESP) and the usual reverse-ESP differ enormously and serve completely different purposes. A usual reverse-ESP does not fit into the present study.

The referee suggests sampling the meteorological forcing randomly from the available 30 years of forcing for the InitSH (formerly ESP), with the argument that taking identical forcing for each year could result in some artificial skill/signal. This approach would be novel as far as we know. In addition, we do not agree that identical forcing could lead to an artificial signal. To our understanding, there will be no signal (and no noise) due to the forcing if the forcing is identical each year. In addition we believe that taking

random samples of the forcing for each year, as compared to taking identical forcing for each year, would add noise, which dampens the signal from the initial conditions more than in a simulation with each year the same forcing, see our explanation for the difference between the FullSH and the InitSH. We thus speculate that in terms of skill at longer lead times a specific hindcast with random forcing would be very close to the FullSH. We consider doing another experiment with random sampling for each year of the hindcasts as outside the scope of the present study (experiments are very expensive).

Page 4, lines 19-23

Mainly see main point 1 of this referee.

For the evaluation of skill we used the same metrics throughout the paper. All these metrics are insensitive to biases, as motivated in the introduction;

The version of VIC that we used was only crudely calibrated (by Nijssen et al., 2001). Hence, streamflow computed by the present version of the system may be expected to deviate substantially from observations, both in terms of the mean and in terms of the spread of the ensemble of forecasts. Also, within WUSHP no post-processing of discharge is carried out to correct for such deficiencies. This makes the system unsuitable to issue forecasts of absolute amounts of discharge but the system can be used to provide information on how likely it is that in a coming month or season discharge will be above or below normal. Consequently, the criteria for the selection of skill metrics (see Sect. 2.2) are their ability of discrimination, and their insensitivity to biases and to the spread of the forecasts.

Metrics that are sensitive to bias are certainly not useful for the comparison of the skill in the raw S4 forecasts with the skill in their bias-corrected version. More skill would be a trivial result. Thus, our bias-insensitive metrics are appropriate.

We write in Appendix A:

The fact that differences are small is not surprising because the bias corrections do not change the ranking of the values while the value of the correlation coefficient largely depends on the ranking of the hindcasts relative to the ranking of the observations. Differences would have been much larger if metrics sensitive to biases had been used for verification.

Page 4, line 36 – page 5, line 3

First bullet. One member is selected from all of the odd years. We reformulated:

FullSH ..... consists of an ensemble of fifteen S4 hindcasts. More specifically, we selected member 1 from the 1981 hindcasts, member 2 from the 1983 hindcasts, etc.

Second bullet:

The referee suggests taking random forcing instead of each year the same forcing: see our reply under major point 2

Third bullet:

We now refer in Section 4.3 to the advantage of ESP as follows:

An advantage of ESP is that its production is relatively cheap because no climate model forecasts are needed.

To explain why we use S4 forcing in InitSH, we write in Sect. 4.3:

However, while in ESP the forcing is selected from historic observations, it is selected from the S4 hindcasts in InitSH in order to retain an inter-member variability and other statistical characteristics of the time series similar to that in the FullSH.

And some sentences below:

So, both in ESP and in the InitSH the meteorological forcing is identical for all years of the hindcasts with the aim of eliminating the skill due to the forcing. If indeed all skill due to the forcing is removed, the remaining skill, due to the annually varying initial conditions, which are identical in ESP and InitSH, should logically be the same in both types of hindcasts.

We then show that ESP and InitSH have virtually identical skill. Because characteristics of the forcing in InitSH (from S4) match better with the characteristics of the forcing in the FullSH (from S4) than the characteristics of the forcing in ESP (observations) do, we prefer the InitSH above the ESP in the main line of the paper.

Referee 2

Point 1) The choice of R as metrics.

Misschien kun je hier de volgende zin uit Barnston et al. (2011) gebruiken: Case-to-case discrimination (as indicated by correlation skill; Murphy 1988) is often considered the most important component of final skill, since many calibration (bias-related) problems are correctable, while discrimination reflects a more fundamental ability of the prediction model.

The correlation between model predictions and observations reflects purely the discrimination ability of the models, since biases of various types do not affect this metric.

We have much better illustrated, described and explained in the main text of **paper 1** the spatio temporal differences found, between the various skill metrics, adding to the supplementary material also maps for the other skill scores. Though the spatio temporal *patterns* of the dynamics of skill are similar between the three metrics (R, RPSS and ROC), we agree with the referee that their statistical significance levels are not similar (generally the fraction of with significant skill decrease from R > ROC-AN/BN > RPSS >> ROC-NN, see figure S2 in paper 1. We believe that in paper 2 the choice of the metrics is less relevant than in paper 1 since paper 2 is about (dynamics of) the sources of skill. So, in the main text, we only present R. In the supplementary material we added several figures identical to Figs. 4 and 5, but for the other metrics (Figs. S1-S4). Conclusions are almost the same:

Figures similar to Fig. 4 (Figs. S2-S5) illustrate that the skill reversal is found for all of the metrics considered in this study and also for the domain-mean of R.

Figures similar to Fig. 5 but for all metrics of the present study are included in the supplementary material (Fig. S4). The graphs for the ROC areas for the Above Normal (AN) and Below Normal (BN) terciles are qualitatively similar to the graph for R. This also holds for the RPSS though fractions of the domain with significant RPSS are almost always lower than for the other metrics. An exception is the relatively large amount of significant skill in the SnInitSH when RPSS is used as metric.

In the first paragraph the referee writes that "skill scores that describe performance in relation to climatology (like RPSS) make it much easier to understand the value of the forecasting system (even against the 'pseudo observations' used in this paper) than correlations". Reply: the skill score of R is equal to R itself because R (climatology) = 0 and R (perfect forecast) = 1. So, R is a skill score.

Second paragraph of referee point 1:

What we claim is that the "patterns of skill are similar for the different metrics", not that "R and RPSS themselves are similar". For instance in the abstract of the companion paper we write:

Qualitatively, the use of different skill metrics (correlation coefficient, ROC area and Ranked Probability Skill Score) leads to broadly similar spatio-temporal patterns of skill, but the level of skill decreases, and the area of skill shrinks, in the following order: correlation coefficient, ROC area below normal tercile, ROC area above normal tercile, Ranked Probability Skill Score and finally, ROC near normal tercile.

We have read the paper by Murphy (1988). He shows convincingly that there are differences between the skill score of the mean-square-error (SSMSE) and R. The differences are due to conditional and unconditional biases, which do affect the SSMSE but not R. However, we have for motivated reasons (see reply to the Page 4, lines 19-23-point of referee 1) chosen metrics that are insensitive to conditional and unconditional biases. So, it was correct to take R and not the SSMSE as metric. As far as we understand this, the RPSS is also insensitive to biases, so this metric would encounter the same theoretical objections that Murphy (1988) formulated against R.

In the new version of paper 1 we have better explained how the statistical significance of R is computed.

We use the word "skill" like Mason and Stephenson (2008), namely in a general and not precisely defined sense. Wikipedia writes: " The term 'forecast skill' can be used both quantitatively and qualitatively. In the former case, skill could be equal to a statistic describing forecast performance, such as the correlation of the forecast with observations. In the latter case, it could either refer to forecast performance according to a single metric or to the overall forecast performance based on multiple metrics."

To avoid misunderstanding we added at the beginning of Section 2.2:

Discrimination skill (briefly skill from now on) is measured in terms of the correlation coefficient between the median of the hindcasts and the observations (R).

2) Reliability

We analysed the reliability of the system. Results are shown and discussed in Appendix B. However, we believe that this analysis would better fit in the companion paper.

3) Cross validation

VIC was calibrated by Nijssen et al. (2001) and was validated by Greuell et al. (2015). This can be considered as a cross-validation since the observations used by Greuell et al. (2015) are almost completely independent of the observations used by Nijssen et al. (2001).

The ways of selecting meteorological forcings in the InitSH (formerly ESP) is now better described (Section 2.3):

In the InitSH ... while for each year the meteorological forcing is identical and consists of an ensemble of fifteen S4 hindcasts. More specifically, we selected member 1 from the 1981 hindcasts, member 2 from the 1983 hindcasts, etc.

Yes, forcings are from different years.

In making the bias corrections of the forcing we established the equations on the basis of all 30 years of data. We agree with the referee that the strategy of leave-the-year-itself-out is superior. However, we believe that in practice differences between the two approaches are small. Also and unfortunately, production of each set of (experimental) hindcasts is computationally very expensive. Hence, we have

not reproduced the hindcasts with the leave-the-year-itself-out.technique We will seriously consider that technique when we make bias corrections for a future version of the system.

4) Detrending

We added some sentences about the detrending procedure in Sect. 2.2 (Methods of analysis and observations):

To investigate the possible contribution of trends to skill, skill in the meteorological forcing and in runoff was determined before and after removing the trend from the observations and the hindcasts. Data were detrended by first constructing time series of 30 years (1981-2010) where each time series is for a single variable, month of the year and grid cell. We then removed the trend from each time series of by first fitting a line with the method of least squares to each time series and then subtracting the time series corresponding to the line from the undetrended data. In the case of the hindcasts, time series were constructed for the mean of the ensembles and the resulting best fit was subtracted from all 15 individual members of the ensembles.

The referee asks to explain how trends could impact predictability. We have reformulated the explanation as follows:

It should be noted here that trends can only cause correlation between hindcasts and observations, and hence skill in the hindcasts, if they are present in both time series. A random time series of hindcasts is not correlated with a time series of observations with a trend and vice versa.

To further explain this (but we think this is too detailed for the paper): if observations and hindcasts tend to be low in the beginning of the time series and high towards the end of the time series, low values of the observations will tend to be associated with low values of the hindcasts and high values of the observations will tend to be associated with high values of the hindcasts. So, there is a positive correlation between observations and hindcasts, and hence skill in the hindcasts. However, if observations show no tendency with time while hindcasts tend to be low in the beginning of the time series and high towards the end of the time series, low and high values of the hindcasts are associated (on average) with similar values of the observations. So, there is no correlation between observations and hindcasts, and hence no skill in the hindcasts.

5) Other set-up of MeteoSH (formerly revESP)

One of the principles on which we built our experimental design is that each hindcast should consist of the same number of ensemble members (15) and that the initial conditions should be deterministic. We formulate that as follows:

Thus, like the FullSH, all specific hindcasts for a single starting date consist of 15 members, which is important since ensemble size affects skill metrics (Richardson, 2001). Also, in all of hindcasts the probabilistic character is exclusively due to the meteorological forcing (15 members) while initial conditions are deterministic. This consistency is important since the main aim of the various specific hindcasts is to compare them with each other.

The experimental MeteoSH (previously rev-ESP) hindcasts suggested by the referee would have m (numbers of in initial conditions) times more members than 15. We will not produce the suggested hindcasts since:

- These do not consist of the same number of members as the other hindcasts
- The enormous computational cost
- One condition for the MeteoSH (our previously rev-ESP) is that the initial conditions are the same for each year of the simulations. This condition is based on the assumption that, if initial conditions are identical each year, they will not affect the skill of the hindcasts. If that assumption is justified,

then indeed it does not matter what those time-invariant initial conditions are. We speculate that perhaps with exotic initial conditions the assumption would not be justified but that with the 30-year mean the assumption holds.

6) The ESP experiment

See major point 2, referee 1. We have rewritten the sentences about our specific hindcasts in the introduction section, as well as the text about the forcing of the InitSH (formerly ESP):

In the InitSH ... while for each year the meteorological forcing is identical and consists of an ensemble of fifteen S4 hindcasts. More specifically, we selected member 1 from the 1981 hindcasts, member 2 from the 1983 hindcasts, etc.

We also renamed the specific hindcasts.

7) Context, reliability

We extensively put our study in the context of other systems in the companion paper (paragraphs 2-5 of its Introduction and in its Sects. 3.1 and 4.2).

We analyse reliability in Appendix B.

Indeed Yuan (2016) applies a post-processing step but he does not demonstrate that this increases the reliability of his system, nor does he mention that this step is meant or known to increase reliability. So, we do not cite Yuan (2016) in the context of reliability.

In the introduction we added a sentence about the general aims of the present version of WUSHP:

In the present and in the companion paper (Greuell et al. 2016), WUSHP is used as a tool for purposes of academic interest.

Also in the introduction, we write about the crude calibration of VIC, the lack of post-processing and the selection of metrics:

The version of VIC that we used was only crudely calibrated (by Nijssen et al., 2001). Hence, streamflow computed by the present version of the system may be expected to deviate substantially from observations, both in terms of the mean and in terms of the spread of the ensemble of forecasts. Also, within WUSHP no post-processing of discharge is carried out to correct for such deficiencies. This makes the system unsuitable to issue forecasts of absolute amounts of discharge but the system can be used to provide information on how likely it is that in a coming month or season discharge will be above or below normal. Consequently, the criteria for the selection of skill metrics (see Sect. 2.2) are their ability of discrimination, and their insensitivity to biases and to the spread of the forecasts.

In Section 4.4 we added a paragraph about an operational version of the system, post-processing and reliability:

We plan to launch an operational version of WUSHP. That version might include a post-processing procedure with the aims of removing biases in discharge and making the system more reliable. This could perhaps be done with statistical calibration (e.g. Gneiting et al., 2005, and Schepen et al., 2014), a technique that, contrary to quantile mapping, considers information that is available from correlations between hindcasts and observations (see Wood and Schaake, 2008, and Madadgar et al., 2014). Reliability of the present system is evaluated in Appendix B.

8) Post-processing

See the reply to the previous point.

9) RPSS

We improved the description of the determination of significance of the RPSS in the companion paper:

In the computation of significance of the RPSS, sampling errors, i.e. the limited number of ensemble members, constitute a problem. They cause a bias in the RPSS when climatology is used as reference (Mason and Stephenson, 2008). Therefore, the reference for the calculation of the RPSS was generated by sampling randomly from the multinomial distribution with p = (1/3, 1/3, 1/3) and N = 15 (the number of ensemble members).

Specific comments

Page 2, line 18	We write in Sect. 4.3: *ESP is not only used as an experimental tool in science but is also widely used to produce forecasts in operational mode (Day, 1985)*
Page 2, line 20	We need to stick to "reference simulation" in order to be consistent with the companion paper.
Page 2, line 21	Could you, please, provide a reference to an ESP with annually varying forcing?
Page 4, line 21	This statement was moved to Appendix A. For this study we have chosen to use only metrics that are not sensitive to biases, see our reply to major point 7 of this referee.
Page 5, lines 1-2	See our reply to major point 5 of this referee.
Page 5, lines 14-15	This is a good point. We acknowledge this at the end of Section 2: *A disadvantage of the small ensemble size of the forcing is the sampling uncertainty, see Sect. 4.2 of the companion paper.*
Page 7, line 5	done
Page 7, line 5	see our reply to major point 4 of this referee
Page 6, line 6	We added this point in the following sentence in Section 2.2: *Data were detrended by first constructing time series of 30 years (1981-2010) for a single variable, target and lead month, and grid cell.*
Page 7, line 35	We do not think so, see our reply to major point 5 of this referee.
Page 8, lines 1-3	This is an interesting mechanism. It is worthwhile to describe how it really works. We re-described it as follows: *We explain the enhanced skill in runoff mainly by an indirect effect. Skill in the precipitation forcing of the first lead month leads to skill in the states of soil moisture and snow at the end of that month. These model states then serve as the source of skill during the next lead months, when the precipitation forcing has no skill at all.*
Page 8, line 31-32	We reformulated this sentence as follows: *A good example is Scandinavia, where the earliest skill (in April; lead month 1) occurs at low elevations near the coasts of Southern Norway and Sweden, at the end of the local snow season. The latest skill (in July; lead month 4) occurs in the Norwegian mountains, again at the end of the local snow season*

Page 9, line 20 There are three reasons for incorporating a section about skill in ET. In the introduction we write:

Since evapotranspiration has a large effect on runoff, the analysis is complemented with an analysis of the skill in this variable. Predictions of evapotranspiration also have independent value because they are useful for planning of water level control in polders and for planning of water use for irrigation and fertiliser application

And then at the beginning of Section 3.3:

Because hindcasts of evapotranspiration are useful in themselves, because evapotranspiration affects runoff (see Sect. 1), and in order to demonstrate the rich possibilities of the pseudo-observations and the specific hindcasts, this section analyses skill in the hindcasts of evapotranspiration.

Page 9, line 21 This was reformulated, see the previous point. In this paper we hardly describe VIC. Hence, we are also brief about the computation of ET in VIC (beginning of Sect. 3.3):

In VIC evapotranspiration is computed with the Penman-Monteith method (see Shuttleworth, 1993).

More information about VIC is provided in the companion paper.

Page 10, line 24-26 That could be an interesting question: Is there generally agreement between the VIC and the S4 initial snow/soil moisture states? We would be very surprised if snow (water equivalent) was assimilated in S4. We found no literature to answer this question.

Page 11, line 18 We changed the first sentence of this paragraph: *Comparing our results with those of Bierkens and van Beek (2009), both studies agree that initial conditions form the dominant source of skill.*

And then continue about the difference in the contribution to skill by forcing, found in the two studies.

Page 11 line 28 This issue is about the sentence *The question is to what extent hotspots of skill (see Table 1) linked to soil moisture initialisation are due to the cause of the skill and to what extent they are due to a lack of interannual variability in the processes that eliminate the skill?*

We are giving here an explanation of the skill, or lack of skill. Grid cells with large inter-annual variation in soil moisture are more likely to have skill due to the initial amount of soil moisture. This is depicted in Fig. 12a as a standard deviation. Grid cells with large inter-annual variation in rain fall are more likely to loose skill due to the initial amount of soil moisture. This is depicted in Fig. 12b as a standard deviation. So, we plotted standard deviations in these two panels and do not see how skill scores could be useful here.

Page 11 line 32 See the previous issue.

Page 12, line 25 We will consider these suggestions in casef a post-processing step is added to WUSHP

Page 12, line 36 We have added some sentences about this topic, see our reply to major point 8 of this referee.

Referee 3

1) We have changed the title following a suggestions by referee 1 and we prefer a short title.

   The definition of seasonal forecasts is "forecasts for future time periods from more than two weeks up to about a year" (from Doblas-Reyes et al., 2013). So, seasonal refers to the forecast horizon and not to the aggregation time. So, a monthly aggregation time (as we use) is not in conflict with the title.

   Skill in temperature and evaporation are analysed because skill in run-off may be due to skill in temperature and evaporation. This is formulated in the introduction:

   The current paper deals with the sources of the skill in WUSHP and is structured in two main parts. First, an analysis of the skill in the most important meteorological forcing variables (precipitation, two-meter temperature and incoming short-wave radiation from S4) is carried out.

   Since evapotranspiration has a large effect on runoff, the analysis is complemented with an analysis of the skill in this variable. Predictions of evapotranspiration also have independent value because they are useful for planning of water level control in polders and for planning of water use for irrigation and fertiliser application.

2) About half of the figures provide lumped results; the other half (the maps of the various figures) provides regional and/or temporal detail. The power of the lumped results is large.They give the average levels of skill for the different variables and experiments, allowing comparison between variables and experiments. Just a few examples
   - FullSH versus InitSH (Fig. 5)
   - Detrended versus undetrended temperature (Fig. 2c)
   - Compare precipitation with temperature, radiation and run-off

3) Our modelling system is a dynamic system, so auto-regressive effects are implicitly represented, e.g. streamflow persistence is accounted for by computing conduction and other processes in the soil layers of the model (VIC), by the inclusion of the Unit Hydrograph in VIC and by computing discharge with the Saint Venant's equations. We provided more detail about VIC in the companion paper.

4) Human impact on discharge has been dealt with in the companion paper. In this paper real discharge observations were only used in Figs. 4c and 4d. In addition, we produced an extra figure for basins with smaller human impact, which we added to the supplementary material and we extended the text about the reversal in skill as follows:

   As discussed in the companion paper, domain-average actual skill is less than domain-average theoretical skill but the reversal of skill after lead month 1 found with the pseudo-observations is confirmed with real observations, both for large and for small basins. While Fig. 4c is based on the data for all 111 large catchments, a similar graph was produced for a selection of the large catchments with relatively little human impact (about half of the 111 basins; Fig. S1 in the supplementary material; see the companion paper for a description of the selection procedure). Again the reversal occurs after lead month 1, so this phenomenon is also confirmed by real observations from relatively pristine basins.

5) We reformulated the following sentences to clarify theoretical and actual skill. In Section 2.1 we state

   Secondly, the output of the reference simulation, e.g. discharge, is used for verification of the hindcasts. This output will be named "pseudo-observations" here.

And in Section 2.2 :

Unless mentioned otherwise, prediction skill of the hydrological variables is determined against the pseudo-observations (see Sect. 2.1). These have the advantages of being complete in the spatial and the temporal domain and to be available for all model variables. We will refer to this type of skill as "theoretical skill". In the companion paper theoretical skill for discharge was compared to "actual skill", which is the skill assessed with real observations.

6) Validation of VIC has been quite extensively discussed in the companion paper.

7 and 8) We found no important differences for different metrics (R, ROC and RPSS) and also tested our main conclusions for domain-averaged values of the metrics. We added graphs and texts about verification with other (than R) metrics, see the extensive reply to point 1 of referee 2. We also added a figure showing results for the domain mean of R to the supplementary material (Fig. S5) and described the result as follows:

Figures similar to Fig. 4 (Figs. S2-S5) illustrate that the skill reversal is found for all of the metrics considered in this study and also for the domain-mean of R.

9) The statement is that the relative contribution to skill by the meteorological forcing is smaller in the present study than in Bierkens and van Beek. We do not state that we found lower overall performance!

We reformulated the implications of the smaller contributions as follows (Section 4.1):

One might speculate that in Europe their semi-statistical forcing is more skilful than the S4 forcing used in WUSHP. This suggests that there is room for improvement of climate model seasonal forecasts, so the relative contribution of the meteorological forcing would grow in future. In any case, that contribution depends and will depend on the climate model used (e.g. S4 or GloSea).

The proposed methodology is extensively compared with ESP in Sect. 4.3, which has almost completely been revised, see our reply to major point 2 of referee 1.

Using different (not S4) forcings, e.g. GloSea, and comparing the results would be an interesting experiment, which is outside the scope of this paper.

10) Assimilation of soil moisture and snow water equivalent data is an interesting topic for future research but it is outside the scope of the present study.

11) We have followed almost all of the many suggestions for shortening and improvement of the language by referee 1.

Review of "Seasonal streamflow forecasts for Europe – II. Explanation of the skill" by W. Greuell et al.

Reviewed: December 2016

Recommendation: The manuscript is acceptable with major revisions.

In this paper, the authors present the sources of skill of a model-based seasonal hydrological forecasting system, which produces hydrological forecasts for up to seven months of lead time over Europe. Seasonal hydrological forecast systems over Europe are scarce, as well as the analysis of the sources of the skill over this region, which makes this work relevant to HESS and to the wider hydro-meteorological community.

The authors analyse the sources of skill in the seasonal meteorological, discharge, runoff and evapotranspiration forecasts, using a variety of skill metrics and experiments. This complete and very interesting analysis enables to disentangle the relative contributions of initial hydrological conditions of soil moisture and snow and of meteorological forcing on the skill of seasonal hydrological forecasts up to several months of lead time. The results would however largely benefit from being more concise and structured, in order to guide the readers throughout the paper. To this end and for reproducibility, some results would also need more explaining in the methods section of the paper (i.e., climate change).

I would like to raise two major comments about the methods of this paper, which should be addressed by the authors. These are:

1. Presenting the skill of the meteorological forecasts is a great idea as it highlights the importance of meteorological forecasts for seasonal hydrological forecasting. Furthermore, it shows the differences between the skill in seasonal meteorological and hydrological forecasting, which is a prime motivation for producing hydrological forecasts at seasonal time scales. Nevertheless, I do not fully understand the choice of presenting the skill of the raw S4 forecasts rather than the bias-corrected S4 forecasts, ultimately used to produce the hydrological forecasts in WUSHP. Therefore, I believe that it would make more sense if the skill of the bias-corrected S4 forecasts was presented in the results of the paper, with appropriate skill metrics. Otherwise, readers might question your choice to bias-correct the S4 forecasts in WUSHP.
2. The ESP-experiments carried out for this paper are very original! However, I am not sure if I agree with the way that you designed your ESP-experiments (ESPall, ESPsoilm, ESPsnow). The use of an identical meteorological forcing resampled from the S4 hindcasts for each year might in the end produce an artificial signal, which might lead to a biased analysis of the skill in this paper and is not the aim of the experiment design. Please read and address the detailed comments I have made about this in the rest of the revision. Additionally, the reasons you give for designing all of the experiments differently from the widely used standard ESP and reverse-ESP are not sufficient to argue your choice. In the methods, you state that it is in order for these experiments to be closer to the Full Hindcasts, which would not be the case when using the standard ESP and reverse-ESP formulations. However, I do not see the need for the experiments to be close to the Full Hindcasts, at least in the context of this paper. After reading the detailed comments I have made in this revision referring to the ESP-experiments, please consider either arguing more thoroughly why you have decided

to take a different approach from the standard ESP and reverse-ESP or changing the experiment designs.

The paper is overall written in a generally fluent and precise language. As a whole, I thought that this paper provided valuable results and I would therefore be pleased to see it published in HESS, after major revisions. Below are comments which will hopefully help the authors to improve the paper and highlight the value of the results it contains.

Title: The title is pertinent with regards to the contents of the paper. However, the formulation "Explanation of the skill" could be rephrased in order to sound more scientific. You could consider rephrasing it to for example "Sources of the skill".

Abstract: Overall, the abstract provides a complete summary of the paper. It could however benefit from being more concise; here are a few suggestions in that direction:

- Page 1, line 12: could you please consider rephrasing "hindcast simulations […] were carried out" to "hindcasts […] were generated"? The term simulations could be confusing here as you are referring to forecasts.
- Page 1, line 17: please change "simulations" to "hindcasts" for the two instances.
- Page 1, line 20-21: this sentence could be removed from the abstract, not hindering the content of the abstract and making it more concise overall. Please consider doing so.
- Page 1, line 23-24: the sentence could be shortened and thus made more concise by removing "to all potential sources of skill but".

Introduction: The introduction is interesting and introducing this paper's content with a summary of the previous paper's main findings is a great idea. The introduction however contains a lot of overlap in what is being said. Here are a few suggestions that could maybe help to make the introduction more concise and structured:

- Page 1, line 26: 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 28: it would be good to add references for other applications of the seasonal predictions, as done for the energy generation sector.
- Page 1, lines 30-31: please consider rephrasing the beginning of the sentence to "WUSHP produces hydrological simulations and forecasts from the Variable Infiltration Capacity […]".
- Page 2, line 1: could you please rewrite "[…] in runoff is fading […]" to "[…] in runoff was found to be fading […]"?
- Page 2, lines 1-2: please consider rephrasing this to "[…], but some significant skill remained up to 7 months of lead time".
- Page 2, line 3: could you please change the word "causes" to "sources"? which is widely used in this context.
- Page 2, line 3: please consider changing "[…] along two lines" to "[…] and is structured in two main parts". Then in the following paragraphs introduce the two parts by saying something similar to: "First, an analysis of the skill […] is carried out" and "In a second part, sources of predictability are analysed […]" (page 2, line 12).
- Page 2, line 4: could you please specify here what variables of the S4 meteorological forcing are analysed in this paper?
- Page 2, line 7: "starting date" or "initialization month" is better here than "start period".

- Page 2, lines 8-11: these very lines sound too much like results, it should sound more like a literature review. Please consider rephrasing these sentences to sound more like an introduction material. Referring to specific maps of the paper is for instance not adequate here.
- Page 2, lines 14-17: could you please consider combining this part of the introduction with page 2, lines 23-24? which is essentially a repetition of the former.
- Page 2, line 17: please add the word "contributions" after "soil moisture and snow initial conditions".
- Page 2, line 18: the ESP refers to a forecasting technique rather than modelling.
- Page 2, line 19: instead of "as realistic as possible and vary from year to year", a clearer formulation would be for example "our best estimates of the current initial conditions for this specific forecast starting date".
- Page 2, lines 26, 30, 38 and 39: please change "simulations" to "forecasts" or "hindcasts".
- Page 2, line 26: consider rephrasing "is as realistic as possible and is" to "it is our best estimate of the current meteorological conditions,".
- Page 2, lines 32-33: I would rather use the term "climatological information" instead of "no information at all", which is not accurate, and then ", which is the case when they have a climatological distribution" could be removed.
- Page 2, lines 35-36: the sentence "All of these studies basically looked at uncertainty in seasonal forecasts." could be removed as it is not necessary and does not sound so good.
- Page 2, lines 37-38: rephrase this to "(2010), we will first look at the skill of the ESP hindcasts which we will then compare to the standard […]".
- Page 3, line 1: could you please specify here the total skill of what hydrological variables will be quantified by removing one or more sources of skill?
- Page 3, lines 1-3: the sentence starting with "It should be noted" is a repetition of what was already said earlier in the introduction. Could you please consider removing it here?
- Page 3, line 4: could you please rephrase this to "the sources of skill for seasonal hydrological forecasting over Europe"?
- Page 3, line 4: it would be good to specify what is dominated by initial conditions.
- Page 3, lines 13-14: it would be nice if you could add references for this use of evapotranspiration prediction.
- Page 3, line 15: please change "of evapotranspiration" to "in evapotranspiration forecasts".
- Page 3, lines 18-22: could you please specify the sections for each of these different analysis parts? As was done on page 7, lines 10-14.
- Page 3, lines 19 and 20: it is not clear from the introduction what is meant here by "the various ESP experiments". It becomes clearer after reading the methods section though. Could you thus rephrase this here to "the ESP and reverse-ESP experiments"?
- Page 3, line 22: please change "evaporation" to "evapotranspiration" here.
- Page 3, line 22: the sentence about additional figures is not appropriate here. Please consider moving it to the methods or results section of this paper.

Section 2.1:

- Page 3, lines 25-29: in this description section, it would be nice if the time step of the simulations, as well as downscaling of the meteorological forcing for the hydrological simulations was mentioned.
- Page 3, line 26: please specify that the bias correction is for the meteorological forcing.

- Page 3, line 27: could you specify hindcasts of which variables are used for this paper? Namely runoff, discharge and evapotranspiration.
- Page 3, line 33: "each of" could be removed here.

Section 2.2:

- Page 4, line 6: are those terciles of the observations or of the forecasts? It would be good to specify here.
- Page 4, lines 6-9: I would explain here that for that reason this paper presents results only in terms of the correlation coefficient.
- Page 4, line 16: please specify what is the maximum area that a basin can reach in order to be called a small basin in this paper.
- Page 4, lines 16-18: I would move these results earlier, after the sentence about the comparison between theoretical and actual skill on page 4, lines 12-13, where it fits better.
- Page 4, lines 19-23: I do not understand why you decided to analyse the skill of the non-bias-corrected (raw) S4 forecasts here as you are using the bias-corrected S4 forecasts to produce your hydrological forecasts for this paper. It would thus make more sense to present the skill analysis of the bias-corrected forecasts here. Also, the fact that there are only negligible differences between the bias-corrected and the raw S4 forecasts is, as you mention it, due to your choice of the metrics to compare them. I would thus suggest to use different skill metrics for this specific comparison analysis. If there are still only negligible differences with appropriate skill metrics, it would be ideal to mention that, as previously shown by Wood et al. (2016), negligible meteorological forcing skill improvements can lead to large seasonal streamflow skill improvements (as you mention it in your discussion section), which is why you decide to bias-correct the S4 forecasts here.
- Page 4, line 25: please add "in the scores overview" after "high temporal resolution".
- Page 4, lines 27-28: "lead month zero" is present in many results in this paper, I would thus remove this sentence which is not accurate.
- Page 4, line 29: please specify what will be analysed at the level of the entire domain, the skill?
- Page 4, lines 30-32: please remove the example, sit does not fit here.

Section 2.3: for this section, it would be good to make a figure of the various "ESP experiments", this would help the readers understand exactly what was done here.

- Page 4, line 34: you could also refer to the ESP experiments with "ESP" and the reverse-ESP experiment with "reverse-ESP" or "revESP", which would be much clearer. Also, please make sure that you use either the term "ESP experiments" or "ESP-experiments" if you decide to keep this terminology.
- Page 4, line 36-page 5, line 3: ESPall
    - It is not clear to me how the ESPall can have 15 members, since there are 28 years of hindcasts from which the members can be selected. Are some years not used? This should be clarified here.
    - It is mentioned in the results that the same meteorological forcing is used each year (this should be made clearer in the methods). However, I am not sure if this is a good resampling strategy. It could indeed be that the members selected lead to an artificial and persistent skill/signal in certain regions and for some initialisation dates. I would suggest to resample the members for the ESP in a random way for each year individually, the forcing would thus vary for each year of the forecasts.

Alternatively, you could show here that using the same meteorological forcing each year, or using a different meteorological forcing by randomly resampling the members for each year, leads to the same results and that you thus decided to use the former and simpler resampling method.

- It is in theory a nice idea to resample from the S4 hindcasts instead of the observed meteorological conditions. However, the wider reason for using the standard ESP as a reference to analyse the skill of a seasonal hydrological forecasting system is because it is a computationally cheap method (ideal for operational forecasting) invented when seasonal meteorological forecasts were not skilful enough and based on the assumption that previous years' meteorological conditions are a good indication of future meteorological conditions for the same time of the year. The standard ESP is furthermore known to be a skilful reference and having a more skilful seasonal hydrological forecasting system (here called Full Hindcasts) would guarantee that it is skilful. Here, resampling from the S4 hindcasts is not computationally cheap since you first have to produce those meteorological hindcasts. I also do not entirely understand why you would want this ESP experiment to be as close to the Full Hindcasts as possible. Also, avoiding to reproduce the reference simulation is not a good argument here as this can also be avoided in the standard ESP by simply not selecting the current year. You thus have to argue the choice for this alternative ESP method better in order to use it for your paper. Otherwise, please consider redesigning the experiments. This will impact all other "ESP experiments" of this paper (including the revESP).

- Page 5, lines 4-9: the ESPsoilm and ESPsnow are really clever!

Section 3.1:

- Page 5, lines 26-27: please rephrase to "significant skill approaches 5%, the no skill line. Hence […]".
- Page 5, line 28: rephrase to "there is more skill in January February […] than during the other months".
- Page 5, line 32: please add "(see Fig. 1a)" after "coastal regions".
- Page 5, lines 34-36: this climate change analysis comes as a surprise here, it should be explained in the methods part of the paper to guide the readers throughout the paper. If it is an analysis done in another paper, this paper should refer to it.
- Page 5, line 38: please rephrase to "the theoretical no skill limit".
- Page 6, line 10: specify here that the three summer months are JJA.
- Page 6, line 12: please add "for the summer months" after "function of lead time".
- Page 6, lines 12-13: this is however a completely different area, compared to Europe. Can you really compare the two
- Page 6, line 30: "mix" is not appropriate here, maybe using "are a combination of" would be better?
- Page 6, lines 36-38: since there is not much to show in the figure and this paper already contains many figures, I would suggest to move the figure to the supplementary material and say that it is not shown here in the text. You could in the text then say what fraction of the domain has skill for "lead month 0".

Section 3.2:

- Page 7, line 2: could you please specify here what other sources is referring to? Initial conditions?
- Page 7, lines 3-6: this climate change analysis was not mentioned in the methods, please mention it there.
- Page 7, lines 10-14: I find that this whole paragraph describing the content of the following results breaks the results section. I would remove it or remind the readers of the results structure in the methods rather.

Section 3.2.1:

- Page 7, line 18: this is however not entirely true, in the companion paper, some key differences were highlighted between runoff and discharge. Please consider rephrasing this to say that they show a high degree of similarity in terms of magnitude and spatial patterns of skill, or remove the sentence as a whole.
- Page 7, line 20: please specify that the reverse occurs beyond "lead month 1" for most target months, because it does not occur for all.
- Page 7, lines 21-23: there are quite a few differences between the large and the small basins plots (Fig. 4c and Fig. 4d respectively). The differences are however not highlighted here are it is not the main focus of this paragraph. This questions the existence of the two figures. I would either merge small and large basins in one figure, since there is not distinction in the text, or raise the differences (even briefly) in the text.
- Page 7, lines 26-34: this is quite a nice explanation for the reversal of the skill! However, this is hence due to the ESPall experiment design: the use of the same S4 forcing for each year of forecast produced. As discussed in the methods, this should probably be changed to using different random forcing for each year. Because it could be that this specific selection of S4 forcing made here leads to non-random weird skill patterns, in other words to some random signal. This is however not the goal of the method, which is as you said to assess the importance of initial conditions for seasonal discharge and runoff forecasting and the impact of losing the knowledge about the future meteorological forcing, and using random meteorological forcing from previous years of hindcasts as proxy for future meteorological forcing, on the seasonal runoff/discharge forecasting skill.
- Page 7, lines 35-36: it would be interesting to know whether for specific regions in Europe the revESP is more or as skilful as the ESPall for certain target months-lead times combinations. Could this be discussed and added here?
- Page 7, lines 38-39: the parentheses content is not needed here, the readers can go back to this part of the results if they want to read the specific numbers and it breaks this part of the results. These numbers were however not stated in section 3.1 and should be moved there.
- Page 8, lines 1-4: this is a very interesting observation!

Section 3.2.2:

- Page 8, lines 6-7: this is true compared to the ESPsnow experiment and should be specified.
- Page 8, lines 8-9: is this however true for all lead times?
- Page 8, line 14: please, first say what the figure shows in general.
- Page 8, lines 14-18: could you please specify which ESP experiment (ESPsoilm, ESPsnow or ESPall) you are referring to when you write those results, it will help the readers to understand them faster.
- Page 8, line 17-18: I am not sure what is meant by "combined initialisation map". Please rephrase.

- Page 8, lines 19-22: I would move this section earlier, when you are talking about Figure 5, to make it more structured. You can then refer back to this feature when looking at the maps of Figure 6
- Page 8, lines 25-26: it is hard to understand this point, please consider rephrasing
- Page 8, lines 27-30: this is a very interesting observation!
- Page 9, lines 2-9: this is a very interesting observation, it would be interesting to show the ESPsnow for soil moisture for May with lead 0, for a comparison. So it is also proving that spin-up is important for hydrological modelling
- Page 9, lines 6-9: this is a repetition of page 9, lines 2-6. Please consider removing this repetition or combining both explanations.
- Page 9, lines 10-12: this paragraph should be moved earlier, when Figure 5 is described, which would make it more structured and hence clearer to read
- Page 9, line 12: this order rather depends on the month, not the season, because you are talking in this paper in terms of months. Please change.
- Page 9, lines 13-15: could you please state here what was done exactly with those maps? Or maybe say this in the methods section.
- Page 9, lines 13-19: so none of these hindcasts have skill thanks to the meteorological forcing?

Section 3.3:

- Page 9, line 21: I would repeat here the intrinsic value of evapotranspiration hindcasts.
- Page 9, lines 21-22: "the power […] ESP experiments" is an odd phrase, rephrase or remove.
- Page 9, lines 23-24: I would remove the piece of the sentence about the April and July decomposition. It is not needed and distracts the readers from the first analysis.
- Page 9, line 25: please explain the overall Fig. 9a before entering into details.
- Page 9, line 25: specify that the levels of predictability in Fig. 9a are for the Full Hindcasts.
- Page 9, line 27: could you please remind the readers in between parentheses what the three ESP experiments are.
- Page 9, line 29: specify that you are talking about the evapotranspiration hindcasts.
- Page 9, lines 33-34: in this sentence, add in between parentheses the ESP experiment you are talking about (revESP, ESPsoilm, ESPsnow), to guide the readers through the results nicely.
- Page 10, lines 1-2: this is during the snowmelt season, please mention.
- Page 10, lines 3-10: this part of the results will benefit greatly from explaining the climate change analysis in the methods section of the paper.
- Page 10, line 11: remind the readers the skill of what variable they are currently looking at.
- Page 10, line 20-21: specify that this is for evapotranspiration hindcasts in April.
- Page 10, lines 21-26: the fact that both temperature and evapotranspiration hindcasts may have the same predictability source is a hypothesis here. Please rephrase the sentences to sound like one.
- Page 10, lines 29: please remind the readers once again what three ESP experiments you are referring to.
- Page 10, lines 30-31: please consider saying that this result can be drawn from the fact that the ESPsoilm shows a higher skill than the ESPsnow and revESP for the Mediterranean.
- Page 10, lines 34-35: please specify which ESP experiments those are
- Page 11, line 4: is Figure 11f really needed? There are already a lot of figures so I would consider removing it.

Discussion:

- Page 11, line 8: please use the terms runoff or discharge rather than streamflow here, to be consistent with the rest of the paper.
- Page 11, line 13: it is not really clear what you are referring to when you say the "uncertainty strategy", could you please rephrase?
- Page 11, line 23: remind the readers that these hotspots regions and periods of skill were identified in the companion paper.
- Page 11, lines 24 and 27: the term "sources of skill" is preferred over "causes of skill".
- Page 11, line 28-page 12, line 8: this analysis is very interesting!
- Page 12, line 10: instead of "replaces" I would say "becomes less skilful than […]".
- Page 12, lines 11-13: you can however not exactly compare your results to results from other papers here as their ESP experiment was different from yours I suppose. You could still cite their results to compare to yours but make sure to highlight this difference.

Conclusions: the conclusion is overall too long. I would shorten it to keep only the main results of the paper. Here are some suggestions:

- Overall, please don't refer to figures here.
- Page 13, line 11: describe quickly the different ESP experiments.
- Page 13, lines 14-15: when you say "other ESP-experiments" mention that these were performed in this paper. Otherwise it sounds like you are talking about ESP experiments from other papers
- Page 13, lines 18-19: I would remove the piece of sentence "Similar […] domain".
- Page 13, lines 22-29: I would summarise this whole paragraph in just a few sentences. Just to convey the main conclusion from these results. The detailed results that you are currently describing can be found in the results section of the paper.
- Page 13, lines 30-3[illegible]ame as above.
- Page 13, line 37-page 14, line [illegible] would move this paragraph in the discussion section of the paper.

Figure 1:

- Please consider swapping Figure 1a and 1b as you are first describing result from Figure 1b in the results.
- Figure 1a: could you please add a label for the colour bar saying that this shows R?
- Figure 1b:
    - Please consider making the y-axis a log scale so that we can see what is happening around the 5% line?
    - Would making a colour scale for the initialisation months be possible? This would maybe make the plot more understandable.
- Caption:
    - Please specify that the legend which provides the percentage of cells with significant R values is in the top left corner of Figure 1a.
    - Could you also state that darker red colours signify a better skill?

Figure 2:

- Figures 2a and 2b:

    - These figures look quite messy for lead times 1 and 2. Would using a log scale for the y-axis help with that?
    - Also consider making a colour bar for the different starting months, as suggested for Figure 1a.
- Figure 2c: the labels of this figure are quite messy. Could you please put a legend outside the figure instead?
- Please remove the general title, it is already said in the caption.
- Specify that the colour bar is for R by adding a label next to it.
- Caption:
    - Instead of writing “As Fig. 1” I would mention here again what this figure is. Because it is easier to read directly under the figure than having to jump from a figure caption to the other figure.
    - Consider removing the exclamation mark after “not the trend itself”.

Figure 3: I would suggest to remove this figure and put it in the supplementary material. In any case, all comments made to Figure 1b apply here, as well as the caption explanation instead of writing “As Fig. 1b”.

Figure 4:

- These figures are too messy, please make a common legend to explain what the different lines are.
- Add an x-axis label specifying that these are target months.
- You could remove the x-axis tick labels for Figure 4b as it is shown in the Figure 4d below.
- You could remove the y-axis tick labels for Figures 4b and 4d as they are the same as in the Figures 4a and 4c.
- Caption:
    - Please explain what the figures show in the caption instead of writing “As Fig. 2c”.
    - Instead of “first two panels” say “top two panels” and instead of “other two panels” say “bottom two panels”.

Figure 5:

- Please remove the title as it is specified in the caption already.
- Could you please make a legend for the different lines, it is currently quite messy?
- Add an x-axis label specifying that these are target months.
- Explain what the figures show in the caption instead of writing “As Fig. 4”.

Figure 6:

- Consider removing the main title.
- Please add a label for the colour bar.
- Caption:
    - Explain what the three ESP experiments are.
    - Could you also explain what the figures show in the caption instead of writing “For more explanation, see Fig. 1a”?

Figure 7: same comments as for Figure 6, except the comment regarding the different ESP experiments.

Figure 8: same comments as for Figure 7, except the comment about the main title. Additionally, the caption should not describe the results.

Figure 9:

- Consider removing the main title.
- Figure 9a: same comments as for Figures 1b, 2a and b.
- Figure 9b: same comments as for Figure 5.
- Figure 9c: same comments as for Figures 2c and 4a, b, c and d.
- Caption:
    - Explain what the figures show in the caption instead of writing “for more explanation, see Fig. 1b”.
    - Please specify what the different ESP-experiments are.

Figure 10:

- Please add a label for the colour bar.
- Specify what the figures show instead of saying “for more explanation, see Fig. 1a”.

Figure 11:

- Consider removing the main title.
- Add a label for the colour bar.
- Caption:
    - Could you specify what the figures show instead of saying “for more explanation, see Fig. 1a”?
    - “The final is panel f and depicts the skill […]”.
- Consider removing Figure 11f.

Figure 12:

- You do not need a colour bar for each sub figures: Figures 12a and b can share one, and Figures 12c and d as well.
- Add the labels for the two different colour bars.
- Caption:
    - The caption describes the results, it should not.
    - Please specify what the figures show instead of saying “for more explanation, see Fig.s 1”.

Technical corrections:

- General:
    - Could you please add the word “meteorological” in front of “forcing” when you refer to meteorological forcing. It will make it clearer to the readers what you are talking about.
    - Please consider changing “lead month” to “month of lead time” or “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: could you please add a comma after "In WUSHP"?
- Page 1, line 12: could you please change "To explain skill" to "To explain the skill"?
- Page 1, line 13: please consider using the term "analysed", or something more scientific sounding instead of "looked at".
- Page 1, line 13: please change "of the first […]" to "for the first […]".
- Page 1, line 14: instead of "later", consider using the word "subsequent".
- Page 1, line 14: "Seasonal forecasts of temperature".
- Page 1, line 30: consider removing "that was" and adding a comma before "built […]".
- Page 1, line 33; page 2, line 3: please change the "[…] and lack of skill […]" to "[…] or lack thereof […]".
- Page 2, line 4: please add a comma in "For S4, this was done […]".
- Page 2, line 5: rephrase "with initialisation at the" to "initialisation on the".
- Page 2, line 16: rephrase "and separated" to "to separate".
- Page 2, line 17: remove the second dot.
- Page 2, line 30: remove the comma after "(2008)".
- Page 2, line 35: please change the sentence to "changes in the information of the meteorological forcing and the initial conditions.".
- Page 2, line 37: the word "However," can be removed.
- Page 3, line 4: add "the" in front of "sources".
- Page 3, lines 10-11: move "also" to before "play a role".
- Page 3, line 12: add "the" in front of "skill".
- Page 3, line 16: "Thus" is not needed here.
- Page 3, line 29: change to "so a total of 5400 simulations (30 years * 12 months * 15 members) was carried out.".
- Page 3, lines 31-32: consider changing one "namely" to a synonym.
- Page 4, line 5: please add "the" in front of "Relative".
- Page 4, line 7: change "are similar" to "were similar".
- Page 4, line 9: if you put capital letters for Below Normal and Above Normal please also add the abbreviation in parentheses after the term is introduced.
- Page 4, line 15: is the hyphen needed in between "large" and "basins"?
- Page 4, line 16: the quotation marks are not needed around the word "observations" here.
- Page 4, line 24: add "a" in front of "relatively".
- Page 4, line 19: please add a comma after "Here".
- Page 4, line 26: please add a comma after "(2005)".
- Page 4, line 29: please add a comma after "result sections".
- Page 4, line 29: please consider changing the term "remarkable" to "outstanding" or "noteworthy".
- Page 4, line 30: please add the word "intend to" in front of "provide".
- Page 4, line 34: please add a comma after "total".
- Page 5, line 23: please write "are used here as a reference".
- Page 5, line 24: "A summary of the skill".
- Page 5, lines 24-25: "in Fig. 1a with statistically".
- Page 5, line 29: rephrase to "target months (not shown here) hot spots […]".
- Page 7, lines 3-4: "the question of how much […]".

- Page 7, line 33: remove the “at” in front of “some time”.
- Page 9, line 22: please add a comma after “First”.
- Page 9, line 27: add “the” in front of “three ESP”.
- Page 10, line 14: add a comma after “in revESP”.
- Page 10, line 14: add “the” before “revESP”, for both instances.
- Page 10, line 30: “From the ESP experiments, it can be concluded […]”.
- Page 11, line 4: there is a space missing between “July 1” and “(panel f)”.
- Page 11, line 22: “The same is probably true for the S4 hindcasts”.
- Page 12, line 8: replace “less than” with “lower than”.
- Page 12, line 19: add “the” in front of “ESP”.
- Page 12, line 33: please add “for” in front of “practical”.
- Page 12, line 36: add “also” in front of “demonstrates”.
- Page 13, line 17: please add a comma after “melt season”.