Harrigan et al. (2017) first response to reviewers - Reviewer 3: RC3 (Anonymous)

All reviewer 3 comments are labelled consecutively, for example, comment 1 is R#3-1, with our responses to reviewers given in blue text.

R#3-1. This paper investigates the performance of the ESP forecast method in the United Kingdom. The authors investigate when, where and why the ESP is skillful, based on a set of 314 catchments and 50 years of hindcasts generated with the GR6J model and data from the UK National River Flow Archive. The forecasts are evaluated with a deterministic and a probabilistic criterion, and compared to modelled streamflow climatology. The authors conclude that the skill decreases exponentially with lead time. Higher skill are observed in forecasts initialized in summer months for lead times up to one month, and in winter and autumn months for seasonal and annual lead times. Higher skill is observed in slow responding catchments with high soil moisture and groundwater reservoirs and less skillful in highly responsive catchments.

General comment

I think that this paper is very well-written and of great quality. The objectives and methods are clearly defined, and therefore easy to read and to follow the scope of the paper. The length of the article and the number of figures were appropriate and the content was always relevant. In addition, this paper fits nicely in the Subseasonal-to-seasonal special issue. This study provides a useful diagnostic of ESP over the UK. I particularly enjoyed how the authors made the link between the spatial and temporal skill patterns and catchment characteristics and seasonal features. I listed some comments and questions below, most of them dealing with methodological aspects, and none of them being major.

We that the reviewer for very supportive comments on our manuscript. The comments and questions around the methodological issues have been assessed and we have decided to take your suggestion about focusing on CRPSS on board throughout the manuscript. We discuss the impact this will have on the revised manuscript below.

Major comments and general questions

R#3-2. In both Twedt et al. (1977) and Day (1985), the abbreviation ESP actually stands for “Extended Streamflow Prediction”. It is true that “Ensemble Streamflow Prediction” is widely used, but I think that the original term better conveys the purpose of the method and should be used instead.

We acknowledge the terminology associated with ESP has changed over the years, and recognise that we did not quote appropriately Twedt et al. (1977) and Day (1985). We will edit the text on Pg2; L7-8 to “(Day, 1985; Twedt et al., 1977; originally stood for Extended Streamflow Prediction)”.

As per comment from R#1-2, it is now standard practice to describe the traditional ESP approach as ‘Ensemble Streamflow Prediction’ (e.g., wood et al. (2016), as well as papers within this special issue: e.g., Beckers et al. (2016), Crochemore et al. (2017), and Arnal et al. (2017)). As per R#2-2, we have now made it clearer that we are talking about the ‘traditional formulation of ESP’ whereby historic meteorological sequences are resampled. We would like to keep our terminology consistent with these papers but could change it if deemed necessary by the editor.

R#3-3. P5 L24-25 : “Each of the 51 generated hindcast time-series were then temporally aggregated to provide a forecast of streamflow volume with seamless lead times of 1-day to 12-months, resulting in 365 lead times LT per forecast (leap days were removed).” Do I understand correctly that the streamflow volume for 30

days is obtained by aggregating daily forecasts from day 1 to day 30, and that the streamflow volume for the year aggregates all daily forecasts from day 1 to day 365? If not, could you please clarify? If so, I was confused by the word "lead time" and the analysis involves more factors than just the lead time. Rather than an analysis on lead times, it is an analysis on both aggregation periods and lead times that can be argued to be between 0 days and the last day of the aggregation period. I don't believe this to be real issue, but maybe the authors could be more careful in the way they used the term "lead time". To be more specific, it is the occurrence of "lead times" in Figures 3, 4 and 5 and Section 3.1.1 that triggered this comment.

We thank the reviewer for pointing out that this needs more clarification in the manuscript (which was also queried by 'R#1-4'). The streamflow time-series the evaluation metrics are calculated on is equivalent to the volume of water which flowed from the first (forecast initialisation date) to the last day of the forecast. For simplification, it is expressed in the manuscript in equivalent average daily streamflow (evaluation results are identical for both). We will insert the following text after Pg 5; L25 for clarification: "Note that lead time in this paper refers to the aggregation of mean streamflow over the period from the forecast initialisation date to n days/months ahead in time. So a January ESP forecast with 1-month lead time is the mean streamflow from 1 January to the end of January and a January forecast with 2-month lead time is the mean streamflow from 1 January to end of February". We now hope this is now clearer.

R#3-4. P5 L28 : Regarding the implementation of the L3OCV method, I was wondering why the authors excluded the subsequent two years but not the preceding two. My guess would be that, operationally, the preceding two years are always available, in any case, while the succeeding two are still missing on the day of the forecast, and adding them will add missing and non-independent information to the calibration-validation procedure. Could the authors say a bit more on that?

Yes, this is correct. Operationally we have meteorological forcing data to drive ESP up until the forecast initialisation date. In the hindcast experimental design, we will never have exactly the same conditions as the operational case, because we are driving the ESP in the hindcast (e.g. 1965) with precipitation and PET sequences from 'future' periods (e.g. 1967), which clearly we would not have operationally. To make sure the hindcast experiment is as close to operational conditions as practically possible we do not use the current or two succeeding years (i.e. L3OCV), as large-scale climate phenomenon such as the NAO has shown to have multi-season/year persistence in some parts of the UK. We were motivated by an insightful HEPEX blog post by Robertson et al. (2016) which we also cite in the original manuscript: https://hepex.irstea.fr/how-good-is-my-forecasting-method-some-thoughts-on-forecast-evaluation-using-cross-validation-based-on-australian-experiences/.

We will change the section on Pg5; L26-30 to: "Although it is not possible to create a hindcast experiment under exactly the same conditions experienced in operational mode, effort was made to ensure historic climate sequences did not artificially inflate skill (Robertson et al., 2016) by using leave-3-years-out cross-validation (L3OCV) whereby the 12-month forecast window and the two succeeding years were not used as climate forcings. This was done to account for persistence from known large-scale climate-streamflow teleconnections such as the North Atlantic Oscillation with influences lasting from several seasons to years (Dunstone et al., 2016) as this climate information is not available in operational forecasting it should be included either in the hindcast experiment".

R#3-5. P6 L25-27 : "It was found in testing that ESP skill was artificially advantaged (disadvantaged) if cross-validation was not carried out in historic climate forcings (benchmark forecasts), in some cases by +/-15 %." Could you please clarify this sentence?

This sentence also relates to a point made in the Robertson et al. (2016) HEPEX blog post "Forgetting to cross-validate reference forecasts can unfairly *disadvantage* your forecast method. Remembering to cross-validate the reference forecast (e.g. climatology) is just as important as cross-validating forecasts".

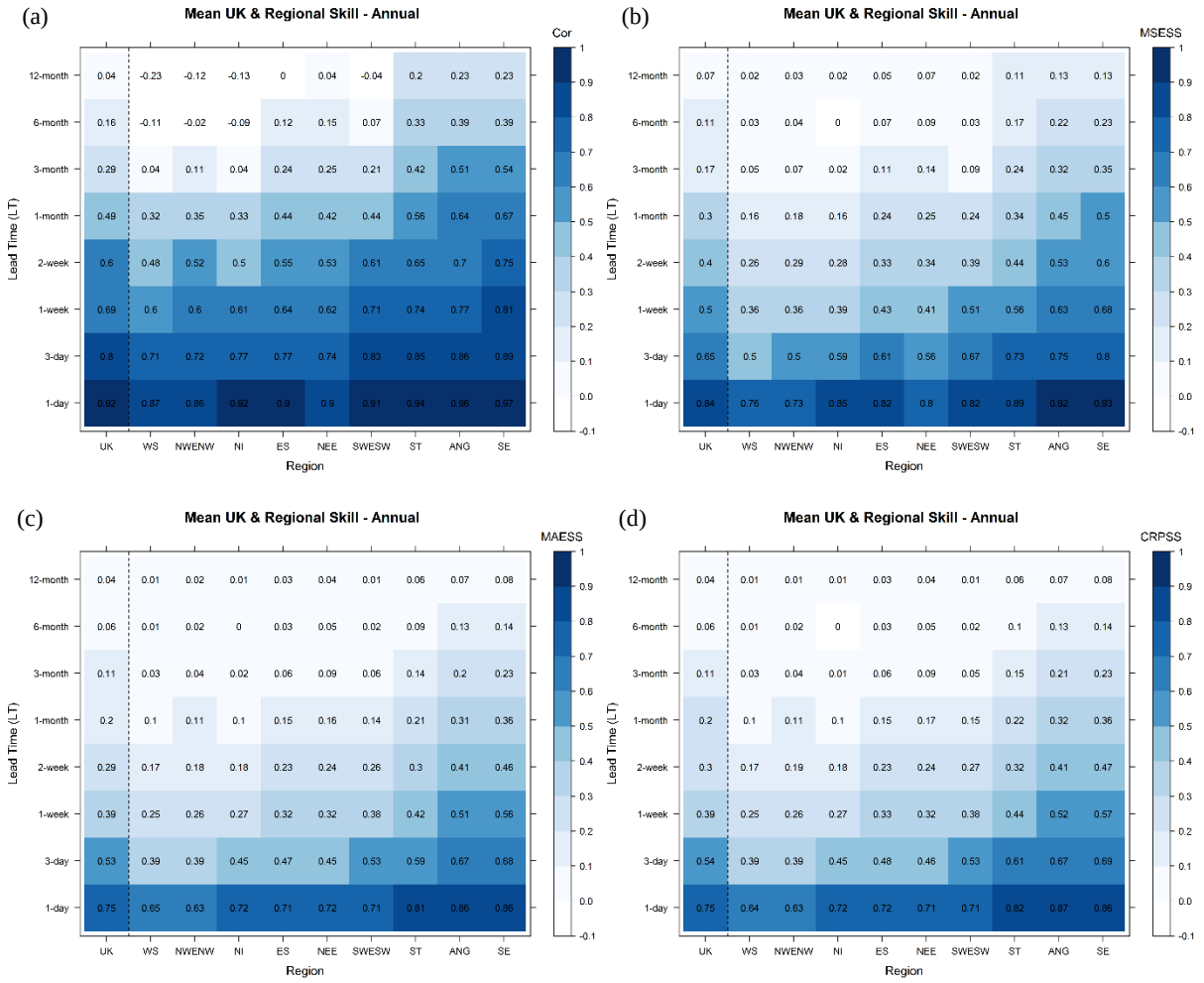
We will replace the text on P6; L25-27 with: "In testing, we performed the evaluation of ESP skill with and without cross-validation of historic climate forcing sequences and climatological benchmark forecasts. It was found that cross-validation was important as in some cases failing to cross-validate historic sequences inflated skill scores (advantaged ESP forecasts) whereas failing to cross-validate climatological benchmark forecasts deflated skill scores (i.e. the benchmark forecast were advantaged thereby disadvantaging ESP forecasts), in some cases by +/-15 %". We hope this is now clearer.

R#3-6. I was wondering about the authors' choice to use the MSE as deterministic score in this case. If the purpose of the two scores is simply to distinguish between deterministic and probabilistic performances, I would recommend using the Mean Absolute Error (the CRPS value of a deterministic forecast is MAE, Hersbach, 2000) so that, when comparing both scores (e.g. Figure 3), the difference in value is solely due to considering the probabilistic side of the forecast.

We thank the reviewer for their recommendation that we have decide to proceed with in the revised manuscript. There is not yet consensus within the hydrological forecasting community on which is the 'best' skill scores to use. We originally decided on MSESS for the deterministic evaluation purely as it has been widely applied and recommended elsewhere. It also has the advantage to being analogous the Nash-Sutcliffe Efficacy (NSE) metric used very widely in hydrological modelling. However, after consideration of your comment and in testing with the MAESS it became clear that the MSESS and CRPSS are not comparable – as you point out for any single ESP you cannot conclude that the ensemble mean (deterministic) is more skilful than the full ensemble (probabilistic) if the MSESS value is higher than the CRPSS value (i.e. Figure 3) – a point we've also responded to R#2-3.

We have taken this suggestion on board and have further tested four of the most common used metrics for assessing hydrological forecasts: Pearson's correlation coefficient (not a skill score: x = ensemble mean, y = proxy obs), MSESS (deterministic), MAESS (deterministic), and the CRPSS (probabilistic). Results from this analysis show that scores from the MAESS and CRPSS are very similar (see figure S2 below), and that the there is virtually no difference between the skill ensemble mean and full ensemble across lead times or regions (Figure S2 c and d). The results for correlation (Figure S2a) and MSESS (Figure S2b and same as Figure 6 in the original manuscript) are systematically higher than MAESS and CRPSS, not due to IHC influence etc. but simply due to the different formulation of these metrics. Their values on a 0 to 1 scale are not directly comparable. However, it must be made clear that it is only the ***magnitude*** of values that is different – the results/interpretation of ESP skill remain the same no matter which metric is used (so most/least skilful region, skill across initialisation months etc.).

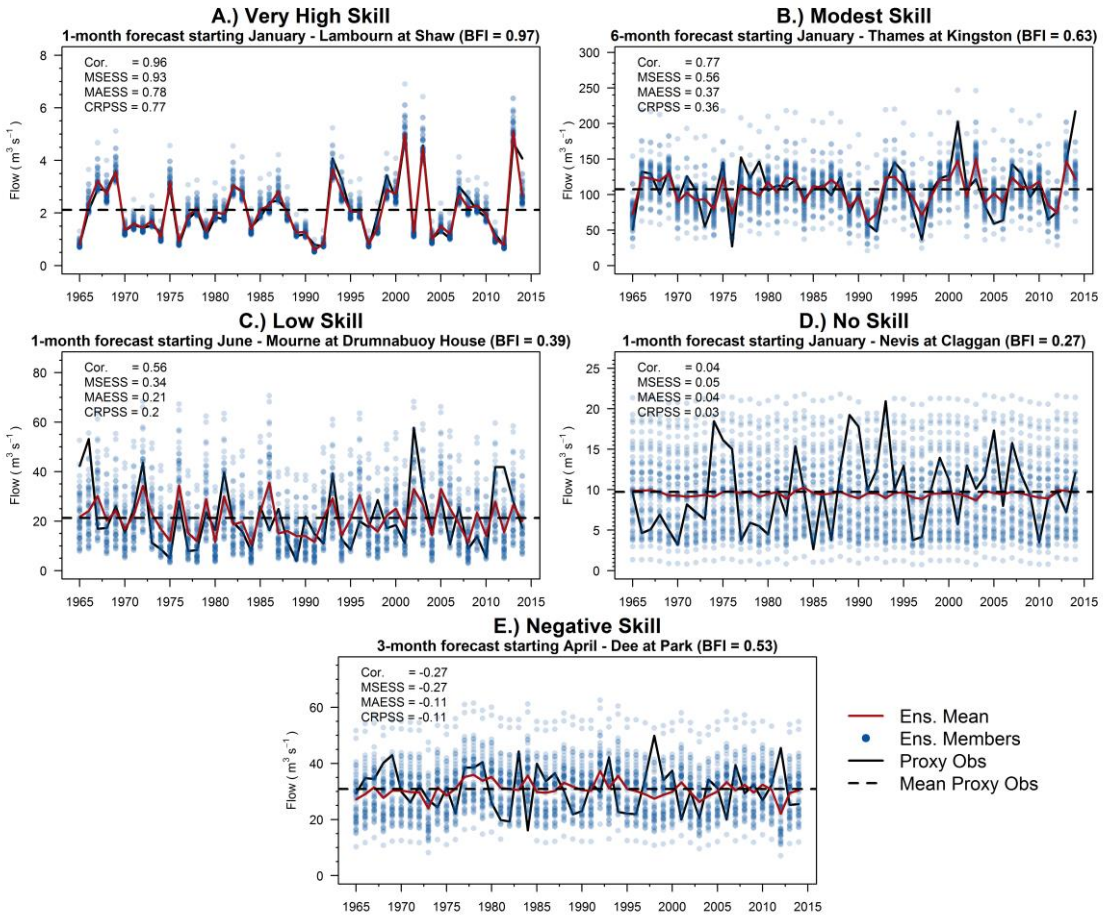
We will concentrate on CRPSS in the revised manuscript, as ESP is a probabilistic method. Given results are so similar between the full ensemble and deterministic ESP forecasts using MAESS, in the revised manuscript we will only use CRPSS (instead of MSESS) in Figures 3, 4, 5, 6, 7, and 8. Therefore, the text in Pg8; L2-4 referring to your will be modified accordingly. We think it's important to include the results of the comparison of the four scores and will include in as **supplementary Figure S2**.



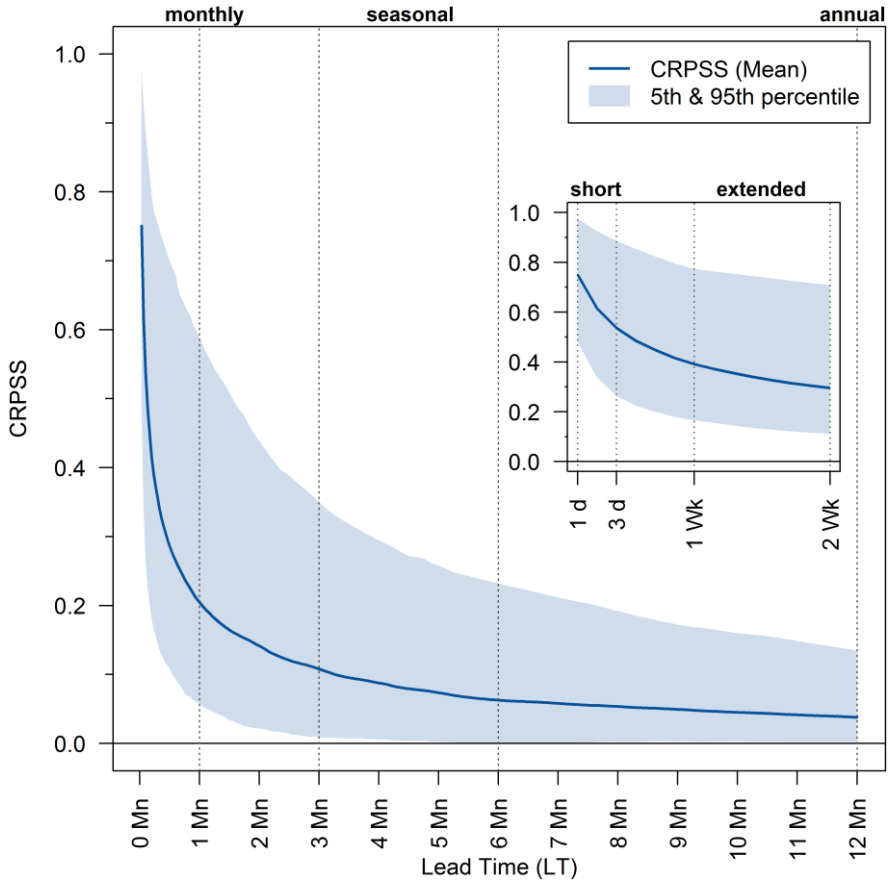
Supplementary Figure 2: Heatmap of mean ESP skill across all 12 forecast initialisation months for the UK and for each of the nine hydroclimate regions ordered from least to most skilful (horizontal axis) at eight sample lead times (vertical axis). Skill is given by the a.) Pearson correlation coefficient (Cor.), b.) Mean Squared Error Skill Score (MSESS), c.) Mean Absolute Error Skill Score (MAESS), and d.) Continuous Ranked Probability Skill Score (CRPSS). Darker (lighter) shades showing higher (lower) skill; individual mean skill values are shown within each cell.

R#3-7. Still on the evaluation criteria, given that ESP is a probabilistic ensemble that translates the uncertainty from climatology, I would have liked the authors to focus more on the CRPS than on the MSE, e.g. in Figures 6, 7 and, possibly, 8). Was there a reason to focus on MSE instead?

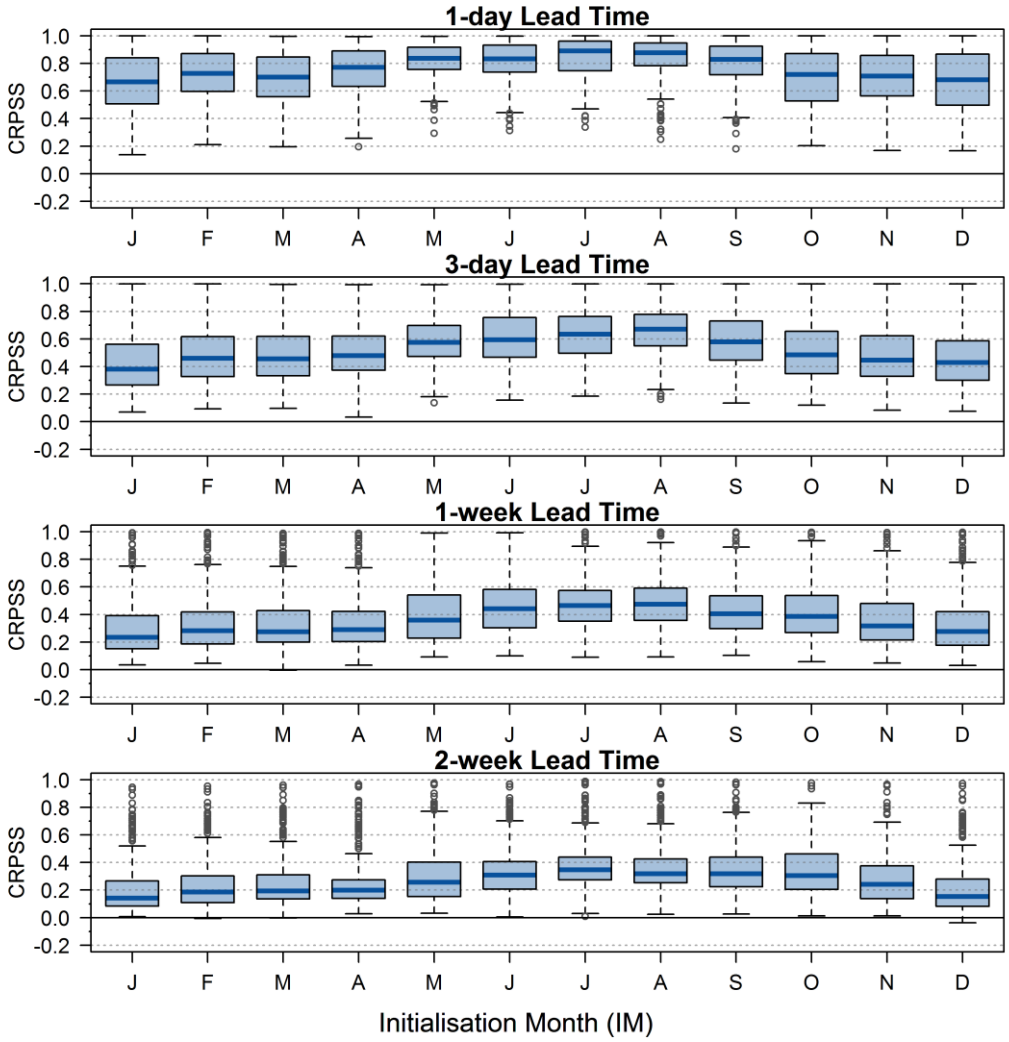
As per our response to R#3-6, we have now redrawn figures using CRPSS. Below are the redrawn figures 2-8. Basing results on CRPSS does not change the conclusions of the paper in terms of ESP skill, however the reported skills in text using MSESS will be replaced with CRPSS, and the magnitude of the skill is lower. This highlights that the qualitative threshold of what is a 'highly skilful' forecast is strongly metric dependent. For example, the CRPSS for the 6-month January ESP forecast in the Thames is 0.36 with the Pearson correlation coefficient is 0.77 (new Figure 2b below). Sect. 3.4 will be modified to reflect this change. Also, we have revisited Figure 7 and added a new threshold in grey (+/- 0.05) to show the difference between CRPSS values near zero.


A.) Very High Skill
1-month forecast starting January - Lambourn at Shaw (BFI = 0.97)
Cor. = 0.96
MSESS = 0.93
MAESS = 0.78
CRPSS = 0.77
B.) Modest Skill
6-month forecast starting January - Thames at Kingston (BFI = 0.63)
Cor. = 0.77
MSESS = 0.56
MAESS = 0.37
CRPSS = 0.36
C.) Low Skill
1-month forecast starting June - Mourne at Drumnabuoy House (BFI = 0.39)
Cor. = 0.56
MSESS = 0.34
MAESS = 0.21
CRPSS = 0.2
D.) No Skill
1-month forecast starting January - Nevis at Claggan (BFI = 0.27)
Cor. = 0.04
MSESS = 0.05
MAESS = 0.04
CRPSS = 0.03
E.) Negative Skill
3-month forecast starting April - Dee at Park (BFI = 0.53)
Cor. = -0.27
MSESS = -0.27
MAESS = -0.11
CRPSS = -0.11
Flow (m3 s-1)
Ens. Mean
Ens. Members
Proxy Obs
Mean Proxy Obs

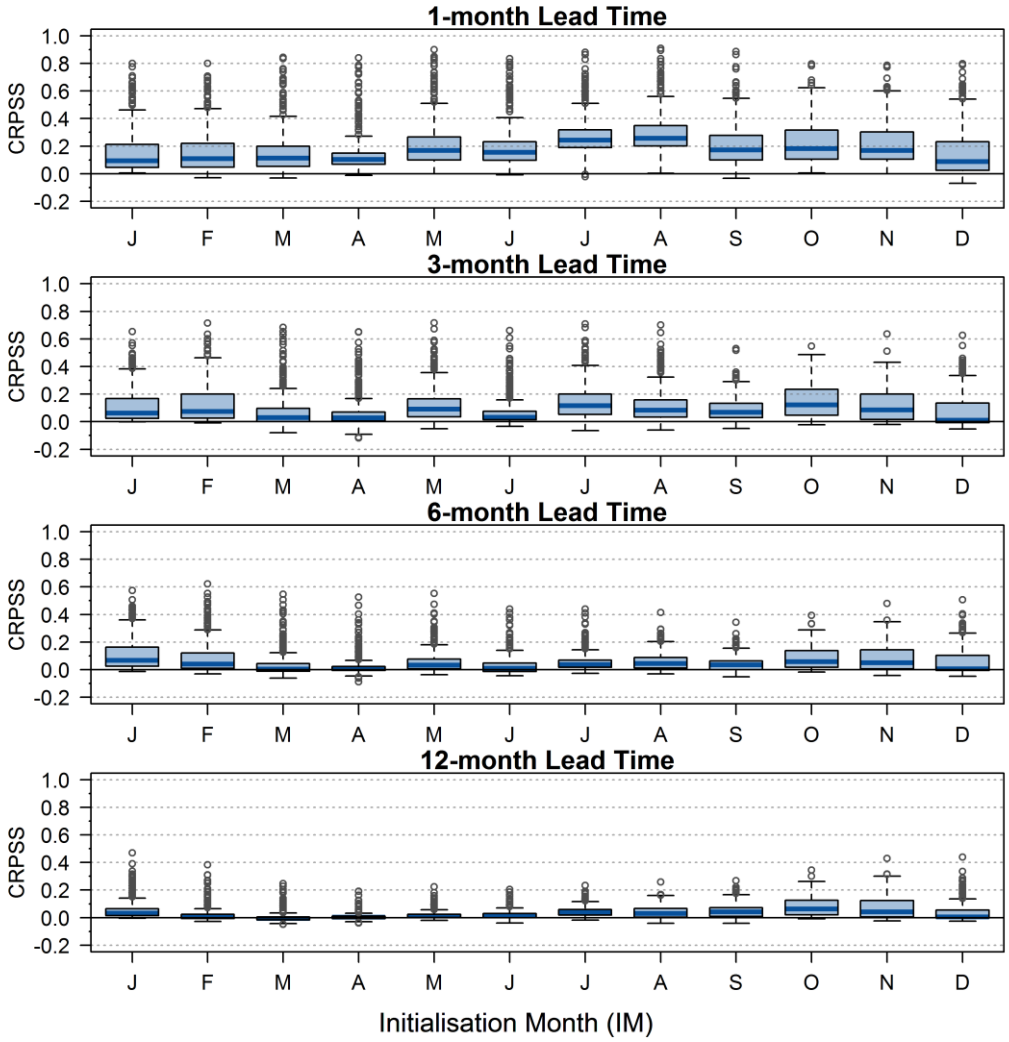
New Figure 2


monthly
seasonal
annual
CRPSS (Mean)
5th & 95th percentile
short
extended
CRPSS
1.0
0.8
0.6
0.4
0.2
0.0
1 d
3 d
1 Wk
2 Wk
0 Mn
1 Mn
2 Mn
3 Mn
4 Mn
5 Mn
6 Mn
7 Mn
8 Mn
9 Mn
10 Mn
11 Mn
12 Mn
Lead Time (LT)

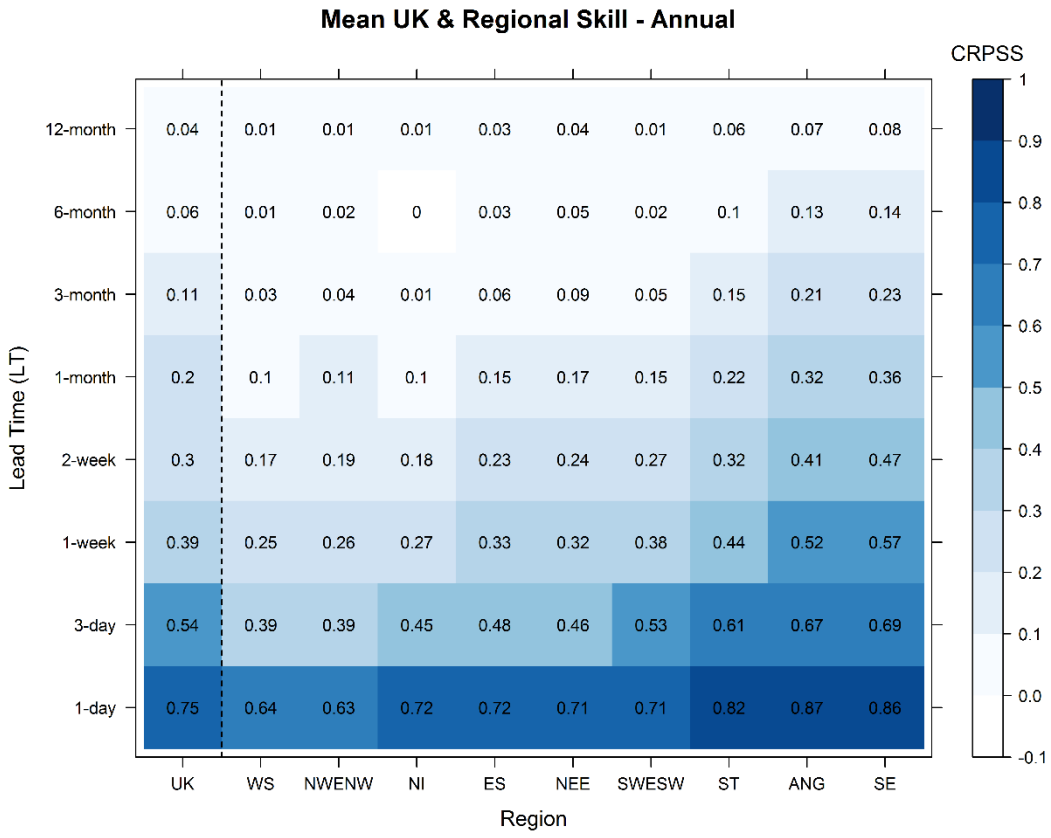
New Figure 3


1-day Lead Time
3-day Lead Time
1-week Lead Time
2-week Lead Time
CRPSS
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
J F M A M J J A S O N D
Initialisation Month (IM)

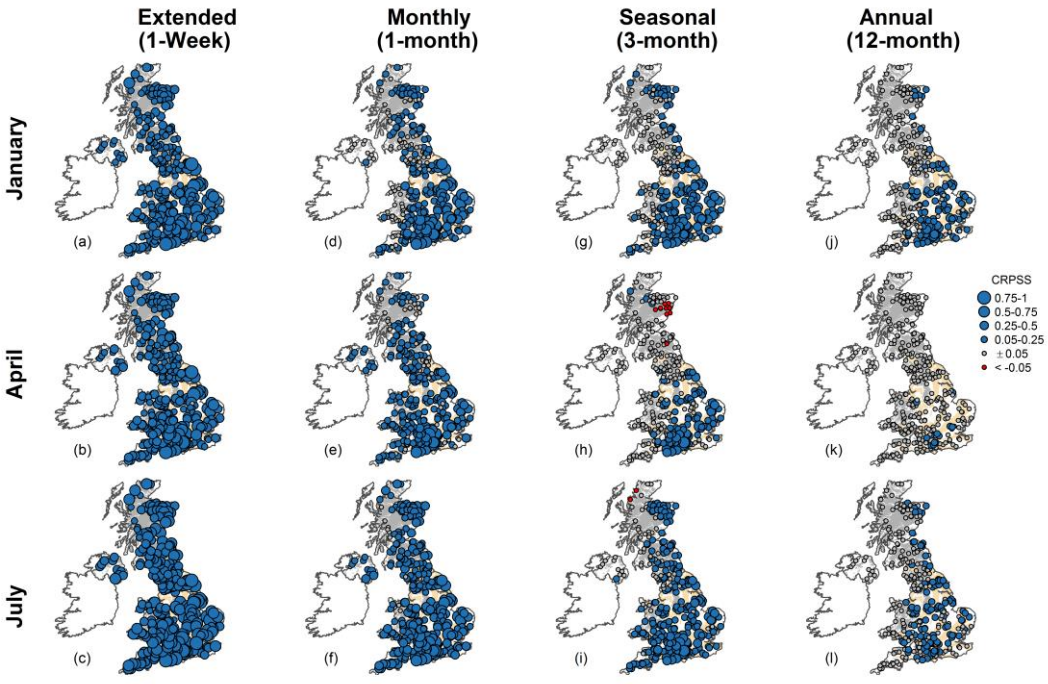
New Figure 4


1-month Lead Time
3-month Lead Time
6-month Lead Time
12-month Lead Time
CRPSS
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
J F M A M J J A S O N D
Initialisation Month (IM)

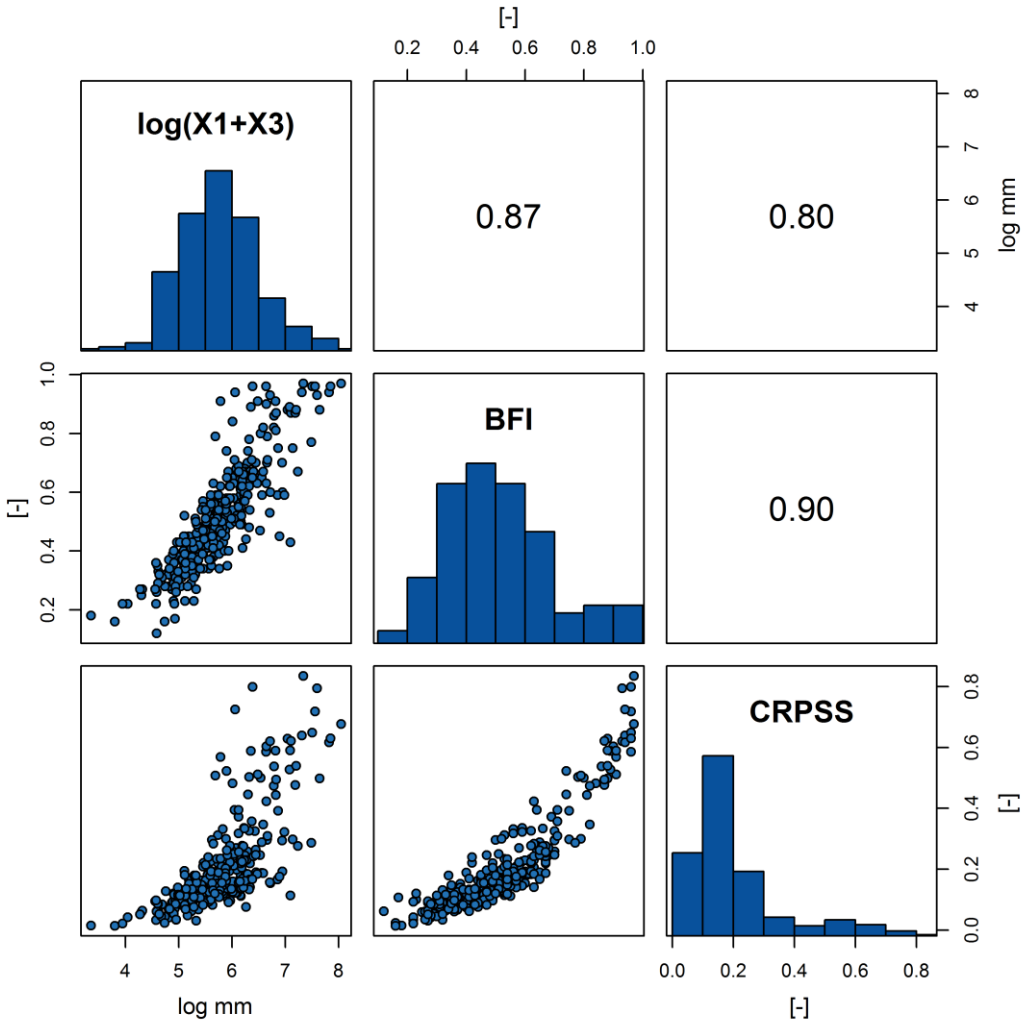
New Figure 5



New Figure 6



New Figure 7



New Figure 8

R#3-8. P7 L17-21 : Is the scale defined for MSESS values or CRPSS values? In the interpretation of Figure 6, it also seems that the threshold value for "Very Low" has shifted to (0, 0.1).

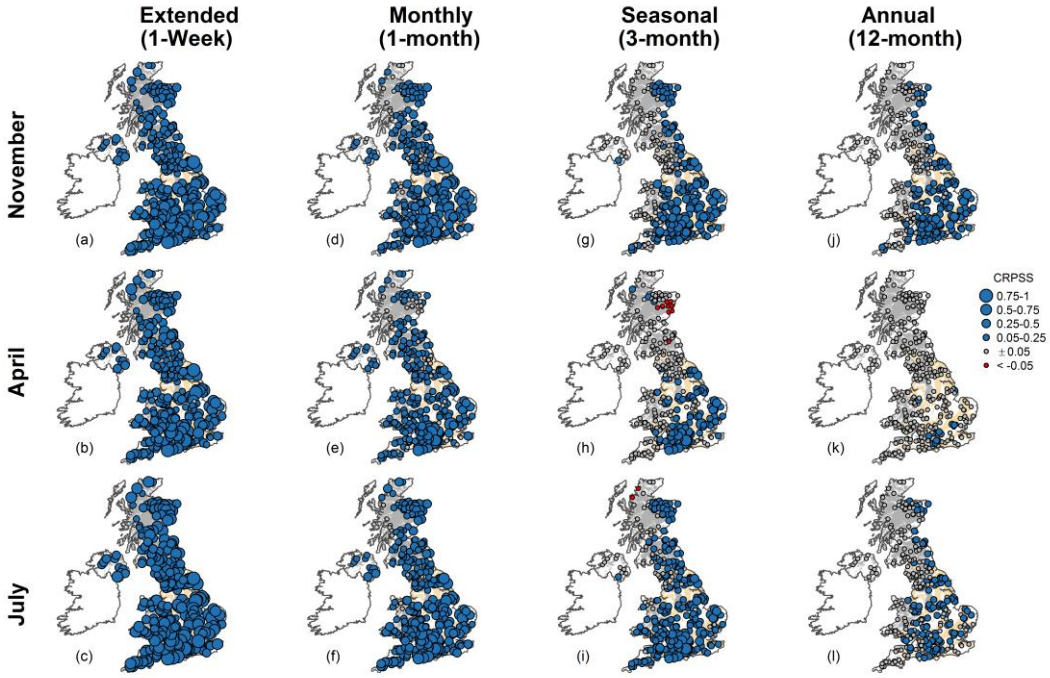
Figure 6 does not discuss these qualitative skill categories but rather shows skill per lead time and hydroclimate region having sequential increments at 0.1.

R#3-9. Figure 4 and Table 2: To which extent does the performance of GR4J for each month of the year explains the results obtained for short to medium lead times and presented in Figure 4?

This is a good point, and was also brought up by R#2-4. We will edit the text as per our response to R#2-4.

R#3-10. Figure 7: Here, I would have liked to see the maps for November which is cited earlier in the analysis.

The new version of Figure 7 is also plotted for November (instead of January in the below figure). As you can see there is very little difference between skill of November and January. The reference to November on Pg8; L13 also includes January. We would therefore prefer to keep Figure 7 with January as these are only sample initialisation months to demonstrate these points.



New Figure 7 – with November instead of January.

Minor comments

R#3-11. P2 L27 : Please change “out to at a least 7-month lead time” to “out to at least a 7-month lead time”

The study in question did not assess lead times beyond 7-months, so we cannot conclude ESP is not skilful after a 7-month lead time, hence why we used ‘out to at least’. We will however make this more clear in the revised manuscript.

R#3-12. P3 L28 : “132 catchments that are part the new version” to “132 catchments that are part of the new version”

Thank you this will be changed. We also note that the number of UK benchmark catchment is 128, not 132. This error will be corrected in the revised manuscript.

R#3-13. P6 L2: Please change “initilisation” to “initialisation”

Will be changed.

We thank the reviewer again for taking the time to give such a positive and considered review of our manuscript. The advice around the best skill scores to use have led us to revise the figures and presentation aspects of the paper and we believe the revised manuscript will be stronger, and more comparable with other papers using ensemble hydrological forecasting.

Kind regards,

Shaun.