Referee #2 – response from the authors

Referee comments are copied below in black. Author replies are indented in blue.

Please find attached my review of the manuscript.

> The authors would like to thank this referee for a careful and detailed review. We are grateful for your suggestions which will help us prepare an improved version of the paper.

1. Scope

The scope of the article is inside the scope of HESS.

2. General evaluation

The authors present a method to improve seasonal hydrological forecasting in Great Britain, in which they conditioned the numerical meteorological forecast with historical weather data and used this forecast as input to a hydrological model to produce seasonal forecasts.

In general, I found the paper difficult to follow. In my opinion, it lacks a consistent story telling. Also, I think in most cases the findings are not conclusive, and the language is not in line with scientific discussion (see examples below). Moreover, the correlations presented in Figure 2 are quite weak, compromising the efficiency of the method. Consequently, I believe major revisions are necessary before moving on with the review process.

> We thank the referee for these general observations. We propose (following comments by referee #1) to separate out the description of the forecast evaluation methods from the results. This should clarify the presentation. General language, and particularly the examples identified by the referee will also be improved in revision.
>
> A detailed response to the referee's observations that "the results are not conclusive…", "the correlations… are quite weak" pertaining to the quality of our rainfall forecasts is given in our response to comments on section 3 below.
>
> We consider the most important results in the paper to be those for river flows (Figures 3, 4 & 5). These are stronger than those for rainfall (Fig. 2). This shows that, despite examples of marginal and no skill in the rainfall forecasts, we can still produce skilful river flow forecasts thanks to the contributions of hydrological initial conditions. Moreover, in Fig. 6 we demonstrate that these rainfall forecasts improve the winter river flow skill in parts of the country where initial conditions are less influential. Thus we argue that these rainfall forecasts provide skill complementary to that from the hydrological initial conditions.

3. Specific comments

Section 1: Introduction

Line 33: You should define sub-seasonal to seasonal timescales (weeks, months…). Also, the temporal resolution of the forecast (daily, sub-daily, etc.)

> We will do so. In this work we are talking about both seasonal forecasts (three-month total precipitation and mean river flows) and monthly forecasts (one-month total precipitation and mean river flows).

Line 40: It would be beneficial to give examples of low- and no-regret actions.

Examples will be given in revision. For instance, when a forecast predicts water scarcity, low-regret responses may include advertising to encourage reduced water use. If a forecast predicts high flows, flood management assets may be repositioned to more vulnerable regions, staff availability may be reviewed and stored water may be released to create space in reservoirs.

Line 57: What do you mean reasonably well?

This will be rephrased to make clear that SAAR-based downscaling provides *improved* skill over uniform downscaling. This is shown by the cited work Kay et al. (2023), who demonstrate that the accuracy of simulated river flows based on SAAR-downscaled rainfall are much closer to that achieved with natively high-resolution rainfall, while uniform downscaling yields worse performance. (Median Nash-Sutcliffe efficiency across the tested catchments are 0.65 for SAAR-downscaled, 0.7 for 1km rainfall and 0.6 for uniform downscaling).

Line 64. I would suggest removing "Section 2 describes our hydrological forecasting scheme in detail." as it is stated below, where you give the structure of the paper.

Line 65-80: This is part of the methods and not the introduction. I would suggest moving this section.

Both suggested changes will be implemented.

Line 71-75: I do not understand what you are explaining. Can you elaborate further? Why 3 subsamples, where does the 63 come from? If the pseudo-observations come from historical records, how are you including physically plausible extreme rainfall events that have not previously been observed?

Line 75-80: This also requires further explanation. What are you matching? The large-scale atmospheric circulation patterns with what?

The technical questions above are addressed in the cited work Stringer et al. (2020). These paragraphs were intended to differentiate Historic Weather Analogues (HWA) from other approaches to seasonal forecasting. In the revised text we will first provide a description of the HWA method before this discussion which should put these comments into context.

In general, I think what lacks in the introduction is the importance / impact that this seasonal forecast can have. A small paragraph justifying why they are used, what type of decisions can be taken, what are the limitations.

This is done in lines 33-41 of the introduction. Following your feedback we will raise the prominence of these lines and also include some examples as requested above.

Section 2: Forecasting methodology

We thank the referee for their careful attention to this section and interest in the detailed implementation of the rainfall forecasting method (referee comments relating to lines 135-158 below). Elements of this method have been published previously in Stringer et al. (2020), Donegan et al. (2021), both cited here (line 134), and, since this preprint was submitted, by Chan et al. (2025).

Unfortunately, the Stringer et al. (in prep) work that was going to synthesise all these details and which we mention on line 143 has been delayed. We therefore propose to add a technical appendix describing the full implementation, with an overview given here to maintain the flow of the paper. In this response to the referee we will briefly answer specific questions, all of which will be included in the technical appendix.

Line 110: How do you pass from daily total rainfall to 15-minute inputs that the G2G model needs (mentioned in line 107)?

We disaggregate the rainfall uniformly across the day. The text will be updated to state this.

Line 127: Which is the original scheme?

This was described in detail on lines 48-60, referenced on line 123 and is discussed in Bell et al. (2013, 2017) which are cited in both places. We will ensure that revised version is clear that we are talking about the same scheme on this line too.

Line 135: What is MSLP (this has not been introduced before)

This is the Mean Sea-Level Pressure (MSLP). We will ensure this is spelled out in the revised version.

Line 137: What criteria do you use to select analogue years? How did you match them?

This was covered in Stringer et al. (2020) and will also be described in the new technical appendix.

Line 139: NAO was not introduced before

This is the North Atlantic Oscillation (NAO). This will be stated in the revised text.

Line 140: What is the "than" referring to here?

Here we are referring to the level of variance in the ensemble mean of forecasts that would be statistically consistent with its correlation with observations. The consequence of this is that the forecast is underconfident in predicting deviations from 'normal' conditions. This 'signal to noise paradox' is discussed in detail by both Scaife and Smith (2018) and Stringer et al. (2020). We will expand the text here to discuss this phenomenon more clearly.

Line 143: Why are you referring to a method that will be explained in a publication that has not been published yet? This is not how references work. You need to either explain it here or wait for the other paper to have at least a preprint and refer to that.

The required description will now be contained in a technical appendix.

Line 147: What is "any" referring to here? Also, by using a proxy and then using that proxy to calculate a quantity, you are not avoiding biases, you are just changing the model you are using.

This will be rephrased – not using a rainfall generation model allows us to avoid its particular biases, but we are affected instead by the biases in predicting the proxy (and the scatter in the correlation between proxy and rainfall).

Line 148: Less good = worse.

This will be rephrased.

Line 150: I do not agree with that; observational means will vary depending on the resolution. Hourly, daily and monthly precipitation means are different.

This will be rephrased – we intended to say that the use of observations *implies* what follows.

Line 153: Three sub-periods are 3 months?

Yes, it will be clarified in the text that for a seasonal forecast these are three months.

Line 158: Why are the extremes better sampled?

This will be explained better in the revised draft. When too small an ensemble is sampled from a continuous distribution the distribution of the samples will deviate substantially from the original continuous distribution. If you increase the ensemble size, the deviation will become smaller. It would be more correct for us to say in the paper that the whole distribution is better sampled when the ensemble size is larger, but here we want to specifically highlight that the extremes of the distribution (which include the low-probability, high impact events of most concern) are substantially improved.

Line 155-162: I believe this needs to be explained better, right now the writing is confusing and there are a lot of ideas not properly explained. How are you sampling not-independent samples?

We appreciate the feedback on these lines. On further consideration, the discussion is ancillary to the main aims of the paper. We will therefore summarise our argument more clearly (below) and provide a full discussion in an appendix.

- We don't know the true distribution of the predictand (rainfall), but we approximate it using an ensemble from a deterministic model. The larger the ensemble, the more accurately it will reflect the true distribution (assuming the model is unbiased).
- Although we appear to have a large ensemble in this work (e.g. 510 members for seasonal forecasts), this ensemble is derived by resampling a small number of deterministic model forecasts (51 runs of GloSea).
- Ensemble members derived from the same GloSea run will not be independent of one another. Their distribution is therefore a potentially biased sample of the true distribution we are looking for.
- In this work, resampling has increased the ensemble size, making it a better approximation to the true distribution of rainfall. However, it is not as good as if we were able to run GloSea a similar number of times (which is too computationally expensive to attempt).

Are you quantifying the covariances you talked about?

We have now calculated the correlations between ensemble members using the UK-mean rainfall forecast by each ensemble member in our hindcast sample. For DJF, we find that the average Pearson correlation between ensemble members associated with the same GloSea member is 0.74±0.10 (indicating that they are not independent) whereas for ensemble members associated with different GloSea members the average correlation

is 0.16±0.21 (both ±1σ). The former is not consistent with 1, indicating that resampling offers some increase in effective ensemble size. The latter is consistent with zero, indicating the GloSea members are nearly independent.

We haven't quantified the magnitude of this effect on the quality of our forecasts – this is outside the scope of this work and would require a large (computationally expensive) GloSea ensemble to compare the resampled ensemble against. However, Stringer et al. (2020) did explore the effect of varying the number of resamples per GloSea member.

Line 172: I believe the references for section 2.1 and 2.2 are mixed.

Thank you, this will be corrected.

Section 3: Comparison between forecast and observed rainfall

I have multiple questions about this section, especially the correlation values that you presented in Figure 2. If I understand correctly you used samples of the hindcast period to evaluate the rainfall forecast model (so comparing the model results against past observed data).

This is correct.

Moreover, on the top right of figure 2 you present the mean correlation.

We acknowledge that the 'mean' correlation (not to be confused with the correlation of spatially-averaged river flow forecasts) is not a statistically robust measure but is intended to provide a simple summary of how the skill differs between months/seasons. The far more important results here are the spatial patterns – we will update the text to emphasise this.

The values you have there are extremely low. Saying that the "correlations are not particularly strong" is a huge overstatement. A correlation of 0 means no correlation at all, and most of your cases are values around 0. Also, you have negative correlations. If you are calculating the correlations between the forecasted and the observed variable (so same variable), and the correlation is negative, it is a clear sign that the model is not working at all.

We understand your points so far. We accept that our discussion in this section lacks nuance by not discussing where rainfall forecasts are not skilful (Pearson r<significance level), and this will be revised. However:

(1) Although the regions/seasons where rainfall correlations are statistically significant still only have low correlations, this level of correlation is not unusual in seasonal forecasting in the extra-tropics. For UK seasonal precipitation forecasts, few models show statistically significant correlations with observations, and these are still low ($r<0.5$ even for the best-performing UK Met Office and Météo-France models; see e.g. Dunstone et al. 2018, Quaglia et al. 2022, Nikraftar et al. 2024). Other models (e.g. ECMWF, CMCC, DWD) show no correlation (not statistically significant, typically $|r|<0.2$; see e.g. Johnson et al. 2019, Quaglia et al. 2022), or negative correlations at some times of year (see previously referenced works). The MSLP patterns that drive UK rainfall are similarly predicted with low skill, again with only GloSea and Meteo-France achieving statistically significant correlations (and these $0.38<r<0.5$ across the UK; Baker et al. 2018); the NAO index is only skilfully predicted at best with $r\approx0.4$ (Thornton et al. 2022). The text will be improved to make these comparisons.

(2) The historic weather analogues method was introduced by Stringer et al. (2020) specifically to improve winter (DJF) forecasts. However, it is important to test whether the scheme is applicable all year round, or just in DJF. Our Fig. 2 thus demonstrates where and when rainfall forecasting method has skill *and when it doesn't*. We agree our results show that the forecasts are not skilful where we have negative correlations, such as in the summer. In these cases, better river flow forecasts could be made by instead using traditional ensemble streamflow prediction (using all historical rainfall patterns without selection). We will revise the text to say this explicitly and discuss the probable drivers of non-skilful cases in more detail.

(3) These results in Figure 2 should be compared to those in Figure 3, which shows the same skill score for the river flow forecasts. The correlations there are mostly much better, indicating that although the rainfall forecasts are only skilful in some places, and often only marginally so, the skill provided by the inclusion of hydrological initial conditions generally compensates for it. This is discussed extensively in Section 5.

Moreover, how did you get that a threshold of 0.33 to define statistical significance correlation?

We will clarify the text. This is the critical value above which we would reject the null hypothesis (that the forecast ensemble mean and observation are not positively linearly correlated). Using a confidence level of p=0.05 the one-tailed significance level for the Pearson correlation for a sample of size n=24 is 0.330 (the value for n=23 is PCorr>0.337). These values are found easily in published tables (e.g. Fisher and Yates, 1963) and can be calculated using software packages e.g. scipy.stats.

I think the evidence in this section suggest that the forecast model used to predict precipitation does not work.

We do not agree with this conclusion. Section 3 shows explicitly that the rainfall forecasting method adopted in this work is skilful in some places and times *and* not at others. The strength of using these forecasts for hydrological forecasting is that, as we argue in Section 5, the inclusion of additional information in the hydrological initial conditions can somewhat compensate for the lack of skill in the rainfall forecasts.

Is there any reference of how other models perform? How do this compare, for example, with the seasonal forecast of ECMWF?

This will be addressed in the revised text. As we have discussed above, other seasonal forecasting models do not appear to perform any better, and many are worse. In particular, the ECMWF SEAS5 seasonal precipitation forecasts for DJF and JJA have no correlation with observations over the UK (-0.2<r<0.2; Figure 20 of Johnson et al. 2019). Similarly, Weisheimer & Palmer (2014) concluded that the earlier SEAS4 were only 'marginally useful' over Northern Europe for wet winters and summers, 'not useful' in dry winters and were 'dangerous' [misleading] in dry summers. Nikraftar et al. (2024) compared seasonal forecasting models included in the Copernicus Climate Change Service and concluded that "UK-Met [GloSea6] and Météo-France consistently outperform other models".

I would suggest that if your model is not working you take another model.

As discussed above, we don't think there is another model available that performs better.

Section 4: River flow forecast performance

Line 232: I agree that taking another model is as the "ground truth" give you the advantage to validate in ungauged locations, but it might be worth to also compare against observed values where you have them and present both metrics. In GB there is the CAMELS-GB dataset that have discharge locations in over 600 stations, which you can compare against.

> We agree that this is possible; there is no shortage of available data. We are concerned that this becomes repetitious in an already long paper. We appreciate the suggestion and will consider it for future studies.

Line 235: What do you mean reasonably well? Also, if you use a dataset as CAMELS-GB there is a label that says what basin are relative unaffected by human impact, so the argument that comparing against observed flows would be unfair is not an actual argument.

> We will rephrase "reasonably well" to be more specific. The sentence that comparing against "observed flows… would be unfair" will be removed.

Line 257: What do you mean by "the performance is good overall"? In a scientific article you should indicate the metrics also in the text, otherwise is completely arbitrary.

> We will provide some headline statistics and ranges in the text (e.g. fraction of river pixels with PCorr > critical value) and refer the reader to the figure for the detail.

Line 260: How did you define the statistically significant threshold? How does it compare to other models? Do you have any reference hydrological model? Why do you have negative correlations for JJA in figure 3?

> See above comments about statistical significance testing, this will be clarified here also.
>
> We do not use a reference hydrological model. Our comparison is against the results of Bell et al. (2017), which used the same hydrological model but different precipitation forecasts in Section 4.4. We extend their work as follows.
>
> (1) We now assess river flow forecast performance at 1km resolution (rather than only when averaged over large regional scales).
> (2) Our assessment now spans three performance metrics (ensemble mean correlation, (continuous) ranked probability skill score, and relative operating characteristic), rather than just one.
> (3) We show that the regional relative operating characteristic (ROC) scores for seasonal river flow forecasts (the only metric published by Bell et al. 2017) are improved when we use the historic weather analogues approach (our Section 4.4, Table 3).
>
> Comparisons with other hydrological models are outside the scope of this paper.
>
> Negative correlations in JJA have already been discussed in the paper (lines 257-259). We will say explicitly here that this work demonstrates clearly that this forecasting methodology as currently configured should not be relied upon to provide accurate JJA forecasts.

Your use of language is quite general and not in line with a scientific article. Expressions as:

- "are statistically significant over much of the country in many months"

- “forecast ensemble means perform less well in northern England, where hydrological memory is not particularly long, nor rainfall forecasts especially skilful. ”

are subjective and do not quantify anything.

We will tighten the language in our revision.

Figure 3. The name of the figures should be descriptive by its own, not a reference to other figures.

This will be corrected in a revised version.

Section 4.2. I think here also comparing your results against observed data would be useful. It would give you an idea of how your model performs against real data.

Please see earlier comments relating to line 232 that addresses this.

References:

- Baker et al.: An Intercomparison of Skill and Overconfidence/Underconfidence of the Wintertime North Atlantic Oscillation in Multimodel Seasonal Forecasts, Geophysical Research Letters, 45(15), 7808-7817, doi:10.1029/2018GL078838, 2018.
- Bell et al.: Developing a large-scale water-balance approach to seasonal forecasting: application to the 2012 drought in Britain. Hydrol. Process., 27(20), 3003-3012, doi:10.1002/hyp.9863, 2013.
- Bell et al.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain. Hydrol. Earth Syst. Sc., 21(9), 4681-4691, doi:10.5194/hess-21-4681-2017, 2017.
- Chan et al.: UK Hydrological Outlook using Historic Weather Analogues, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-2369, 2025.
- Donegan et al.: Conditioning ensemble streamflow prediction with the North Atlantic Oscillation improves skill at longer lead times, Hydrol. Earth Sys. Sc., 25(7), 4159-4183, doi:10.5194/hess-25-4159-2021, 2021.
- Dunstone et al.: Skilful Seasonal Predictions of Summer European Rainfall, Geophysical Research Letters, 45(7), 3246-3254, doi:10.1002/2017GL076337, 2018.
- Fisher and Yates: Statistical tables: for biological, agricultural and medical research (6th ed.), Oliver & Boyd: Edinburgh, 1963.
- Johnson et al.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12(3), 1087-1117, doi:10.5194/gmd-12-1087-2019, 2019.
- Kay et al.: Spatial downscaling of precipitation for hydrological modelling: Assessing a simple method and its applications under climate change in Britain, Hydrol. Process., 37(2), 14823, doi:10/1002/hyp.14823, 2023.
- Nikraftar et al.: Impact-Based Skill Evaluation of Seasonal Precipitation Forecasts, Earth’s Future, 12(11), e2024EF004936, doi:10.1029/2024EF004936, 2014.
- Quaglia et al.: Temperature and precipitation seasonal forecasts over the Mediterranean region: added value compared to simple forecasting methods, Climate Dynamics, 58, 2167-2191, doi:10.1007/s00382-021-05895-6, 2022.
- Scaife and Smith: A signal-to-noise paradox in climate science, npj Climate and Atmospheric Science, 1, 28, doi:10.1038/s41612-018-0038-4, 2018.

- Stringer et al.: Improving Meteorological Seasonal Forecasts for Hydrological Modelling in European Winter, J. Appl. Meterolol. Clim., 59(2), 317-332, doi:10.1175/JAMC-D-19-0094.1, 2020.
- Thornton et al.: Seasonal Predictability of the East Atlantic Pattern in Late Autumn and Early Winter, Geophysical Research Letters, 50(1), e2022GL100712, doi:10.1029/2022GL100712, 2022.
- Weisheimer & Palmer: On the reliability of seasonal forecast predictions, J. R. Soc. Interface, 11, 20131162, doi:10.1098/rsif.2013.1162, 2014.