Added value of seasonal hindcasts for UK hydrological drought outlook
Response to reviewers

Reviewer #3

This study assesses ECMWF SYS5 for use in drought outlooks over the UK. The paper is beautifully written, the figures are nicely presented and I could clearly follow at every point what the authors did. I really regret to say, then, that I thought the paper was ultimately misguided and that I do not recommend the paper to be published, for the following reasons:

1) The paper makes use of methods from climate studies (notably assessments of climatological distributions from the UNSEEN project) that are inadequate for ensemble forecast verification. As I note in the specific comments below, the accuracy and skill of forecasts can be assessed directly using well-established forecast verification methods, as can the appropriateness of ensemble spread, which consider the correlation of forecasts with observations. I recommend the authors familiarise themselves with the fundamentals of forecast verification (see, e.g., Joliffe & Stephenson 2011) before reconfiguring their paper.

2) To me, the use of story lines is fundamentally at odds with the aims of ensemble forecasting. The conception of story lines is appropriate in climate projections, where ensembles of different GCMs are not formally statistically exchangeable, and thus should not be used to express, for example, quantitative confidence intervals. Story lines in climate change projections can be thought of as hypotheses. In ensemble forecasting, however, ensemble members should be formally exchangeable (i.e., each member is equally probable), and we can directly test the appropriateness of the ensemble following (1). This means ensembles can be used to assign probabilities to events. Developing story lines based on climate drivers essentially does away with this probabilistic information in preference to a narrative-drive prediction method. The purpose of ensemble weather and climate forecasting can be thought of as an attempt to get away from narrative-style predictions: basically, the uncertainty in weather is irreducible and cannot be distilled to one or two 'story lines'. Because weather is chaotic, when ensembles are constructed correctly, they should be on average more accurate than any single-value forecast.

3) Following on from (2), it's not appropriate to assess probabilistic forecasts - especially of extremes - on a single event as is done in this paper. Probabilistic forecasts must be assessed on a population of events, and the population must be unbiased. Selecting such a population on the basis of when an extreme event (e.g., a drought) is observed isn't correct: it produces a biased population. The reasons for this are both intuitive and highly technical - e.g., it's not possible to assess false alarms when an extreme event is always observed in your population; for more technical reasons see Lerch et al. (2017) and Bellier et al. (2017)

4) Given (1) and (2) , it's not appropriate to assess the usefulness of ECMWF SYS5 for drought prediction solely on its ability to replicate a climatology; nor to use it to characterise climate drivers of rainfall. If the authors wanted to reconfigure their paper, reusing some of their techniques to identify climate drivers, the authors might consider:

i) Comparing whether observed teleconnections (as seen in, e.g., reanalyses) are reproduced in SYS5, and whether this changes with lead time

ii) Whether drivers of teleconnections are well forecast by SYS5 (following (3), when the forecasting system predicts such an event, rather than if one is observed)

iii) Using (i) and (ii) to address questions on the conditional skill of SYS5 forecasts. For example, it could be used to answer questions such as: a) how well does SYS5 predict key drivers of drought? (b) how well does it do at reproducing observed teleconnections and (c) how do (a) and (b) relate to forecast skill?

These are merely suggestions of course; I leave it to the authors as to what they may wish to pursue. In any case, if the authors did follow these recommendations, I think it would be a fundamentally different paper to what is presented, and hence my recommendation to reject (rather than revise the paper).

It's always difficult to reject a paper like this, in which the authors show a good command of statistical and climate analyses, not to mention clear scientific writing and presentation, but which is in other ways seriously flawed (at least in my view). I really hope the authors don't find my review too discouraging - I wish them well reconfiguring their work and analyses to better align with the precepts of forecast verification.

Response: We would like to thank reviewer #3 for their detailed review of our paper. We are disappointed that the reviewer recommended a rejection of our paper. We note that similar concerns were not expressed by reviewers #1 and #2, and both reviews were positive with minor changes requested. However, reviewer #3's comments have brought to our attention that the aims and objectives of our paper have not been explained clearly enough and we believe this has led to a fundamental misunderstanding of the aims of our study. We apologise for the confusion generated. In a revised version, we will add early on in the introduction to clearly explain the objectives of the work and throughout the manuscript to emphasize our objectives and, more importantly, what our paper is not aiming to do.

1) This paper is not about forecast verification. We are not attempting to predict the 2022 drought event or assess the skill of the SEAS5 hindcasts in predicting the drivers of a particular drought year. Instead, the aim of this paper is to

explore “what-if” situations of the 2022 drought should winter 2022/23 resemble specific atmospheric circulation patterns. We do this by pooling a large sample of hindcast data and clustering them using various circulation indices known to drive winter rainfall in the UK to create circulation storylines. We believe our use of the word “outlook”, particularly in the paper title, could have contributed to the confusion that this paper is about forecast verification. We propose to amend the title to “Added value of seasonal hindcasts for UK hydrological drought storylines”. The drought storylines created allow users to explore in real-time what the plausible worst case could be during an ongoing drought event, hence can be considered a worst-case outlook of the drought event. We will make it clear throughout the manuscript that our study is not attempting to predict the outcome of the 2022 drought and amend any potentially confusing wording to reflect this.

2) In relation to point 2 raised by the reviewer, we disagree that the use of storylines in this way is “at odds with the aims of ensemble forecasting”. We did not advocate for the replacement of ensemble weather forecasting with storylines. The reviewer is correct that storylines can be thought of as hypotheses. Appending the circulation storylines in place of winter 2022/23 therefore enables the exploration of plausible worst cases – e.g. how bad could the 2022 drought have been if followed by a dry winter characterized by a specific combination of atmospheric circulation indices. As outlined in our response to reviewer #1, we propose to expand our discussion to explore the utility of this approach. From a user’s (e.g. water resources manager) perspective, this approach is valuable as the skill of available forecasts, though continuously improving, is currently not perfect. Having the information on what a plausible worst-case might look like is therefore useful from a water management perspective for planning purposes. For example, during a period of prolonged dry weather, conditional storylines can be created several months/seasons ahead to explore the potential range of outcomes should upcoming seasons resemble specific atmospheric circulation patterns with a particular focus on potential impacts of worst cases, in order to plan accordingly. While there have been advances in probabilistic forecasts, plausible worst cases will by definition lie in the tail of the distribution and their likelihood will not be well represented by finite-sized ensembles. Given the irreducibility of uncertainties in weather as the reviewer correctly identifies, we believe that understanding plausible worst cases during an event can be valuable as a “perfect” probabilistic forecast may not be attainable.

   This paper argues that the use of conditional storylines could be a complementary tool (hence the “Added value” in our title) to ensemble forecasting. The well-established ensemble streamflow prediction (ESP) technique, where historical years are repeated, can also be thought of as

following a storyline approach (i.e. what if the rainfall sequence from a historical year is repeated). This paper extends this methodology by adding information on atmospheric circulation drivers of rainfall and by exploring a wider range of outcomes. Additionally, the UK Environment Agency, the public body tasked with overseeing water abstraction licences, managing water transfers and preserving environmental flows, already uses storylines in their operational water resources management, especially during on-going droughts, where hydrological models are driven to produce 'possible worst-case' with hypothetical synthetic rainfall time series, which are simply % of long-term average, but with no consideration of how physically plausible that rainfall occurrence is. The approach proposed in our study is a clear improvement to this, as we sample physically plausible rainfall occurrence modelled by SEAS5. As noted in our response to reviewer #1, we propose to expand on our discussion of how storylines can complement traditional ensemble forecasting approaches, and to include an additional figure comparing results from a traditional ESP framework and the circulation storylines.

3) In response to point 3 raised by the reviewer, we did not "assess probabilistic forecasts … on a single event". As noted previously, we pooled SEAS5 hindcasts and clustered them according to known drivers of winter rainfall in our study region. We thus end up with a large sample of plausible winters separated by rainfall response from different combinations of circulation patterns. Contrary to the reviewer's concern, an extreme event is not always observed in the SEAS5 hindcast population – the outlooks clearly show that there are winters that are wet enough to have abruptly terminated the 2022 drought across the Anglian region.

4) We thank the reviewer for providing suggestions for research ideas outlined in point 4 - they are all interesting topics worthy of future research. However, the aims of these research ideas are fundamentally different to the aims of our paper. We believe our paper outlines a novel approach in relation to the use of meteorological information to aid decision-making by water resources managers and enhance risk awareness during a drought.

Specific comments

P3 L78-79 "Each ensemble member is perturbed with different initial atmospheric and ocean initial conditions" On first reading this I though this wording implied that the authors are (re) perturbing ensemble members, which I would think is unlikely given the computational demands of SYS5 and the authors' earlier declaration that

they are using the retrospective forecast dataset. I assume the authors mean something like: "SYS5 ensemble members are generated by perturbing initial conditions", so if I'm right I suggest the authors go with something like this. In addition, it's my understanding ECMWF perturbs model physics in SYS5, which the authors might also want to mention.

Response: We will make this clearer.

P4 Section 2.1. I found this method of forecast verification basically inappropriate, as follows:

1) Using the UNSEEN framework isn't really appropriate here: that paper tried to put a flood event in climatological context, therefore assembling a large ensemble to assess describe the climatology makes sense. But we are dealing with forecasts here, which are expected to be correlated with observations. Assessing a simple model climatology is not enough to demonstrate the value of forecasts.

2) Even given (1), the idea that the model performs well at simulating a climatology isn't well demonstrated by Figure 1. For example, the figure shows that variance of winter rainfall is understated by SYS5. Further, variance can change with lead time, as information from initial conditions wanes and the model reverts to its internal climatology. Finally, it is quite possible to demonstrate bias etc. across multiple sites and lead times, rather than restricting it to a single site and pooling lead times, which occludes valuable information about the utility of the forecasts.

3) As noted in (1) Forecasts are expected to be correlated with observations. This means that the utility of forecasts is usually measured by:

1) Forecast skill (i.e., forecast errors with respect to a climatological forecast, computed with appropriate error scores such as the continuous ranked probability score), conditioned on both location and lead time.

2) The reliability of ensemble spread, using appropriate measures such as probability integral transforms, attributes diagrams or spread-error diagrams.

In addition, useful information about forecasts is also:

a) The sharpness of the forecasts

b) ability to replicate observed climatology - e.g., with bias/variance/etc. It is only this last criterion that is assessed in the paper.

Response:

1) In response to high-level points #1 and #2 in the specific comments, we reiterate that this paper is not about forecast verification; its aim is not to assess whether the SEAS5 forecasts accurately specific drought events. We will improve the clarity of the explanation regarding the purpose of the UNSEEN model fidelity tests, as it appears that this aspect has not been entirely clear and seems to have led to some misunderstanding. The UNSEEN fidelity test introduced in Thompson et al. (2017) aims to assess whether the model data can be considered as alternative realizations of the real world. Observed winter rainfall from the historical period is just one realisation out of many possible alternative realisations that could have happened. As there is only one observed value per year, the purpose of the UNSEEN framework is to consider the spread of simulated winter rainfall in the SEAS5 hindcasts compared to the observations. If the observed sample statistic falls within 95% of the modelled distribution, the modelled rainfall is considered to be statistically indistinguishable from the observations, as seen in Figure 1a in the original manuscript. This follows the now well-established use of initialised ensembles where most of the initial-condition skill has been lost, either by pooling climate model hindcasts as in Thompson et al. (2017) or by pooling seasonal hindcasts as in Kelder et al. (2020) and Brunner and Slater (2022) among others. Individual forecasts are not discernibly correlated (as is mentioned in the next paragraph), and the different lead times are not discernibly dependent, so there is little skill in these forecasts. That is why we believe our use of the SEAS5 hindcasts, considered over all years rather than just for single years, as plausible realisations of winter weather is appropriate.

2) In relation to point #3 in the specific comments, the reviewer raised the concern that forecasts are expected to be correlated with the observations. This does not detract from the results of this study. As our study is not about forecast verification, the wide range of plausible outcomes from pooling the seasonal hindcasts is important given our aims of exploring "what-if" situations and plausible worst cases. The reviewer's point on model climatology is valid, which is why we had tested for the independence and stability of the SEAS5 hindcasts across the three lead times. As detailed in Kelder et al. (2022), techniques have been developed to assess the model fidelity of seasonal hindcasts. First, the stability test recognizes that forecasts may drift towards the model climatology over time. This is tested by a comparison of the distribution of simulated winter rainfall between the three lead times. As seen from Figure 1b of the manuscript, the distribution of simulated winter rainfall is similar across the three lead times, indicating that there is no drift within this time scale. Second, the independence test assesses whether each ensemble member for each lead time is correlated with each other. This is done by

considering the correlation between each pair of ensemble members over the 1982-2015 period for each lead time. As seen from Figure 1c in the manuscript, the median correlation for all three lead times is close to zero, thus showing that the ensemble members can be considered independent from each other.

References

Brunner, M. I. and Slater, L. J.: Extreme floods in Europe: going beyond observations using reforecast ensemble pooling, Hydrology and Earth System Sciences, 26, 469–482, https://doi.org/10.5194/hess-26-469-2022, 2022.

Kelder, T., Müller, M., Slater, L. J., Marjoribanks, T. I., Wilby, R. L., Prudhomme, C., Bohlinger, P., Ferranti, L., and Nipen, T.: Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes, npj Clim Atmos Sci, 3, 1–13, https://doi.org/10.1038/s41612-020-00149-4, 2020.

Kelder, T., Marjoribanks, T. I., Slater, L. J., Prudhomme, C., Wilby, R. L., Wagemann, J., and Dunstone, N.: An open workflow to gain insights about low-likelihood high-impact weather events from initialized predictions, Meteorological Applications, 29, https://doi.org/10.1002/met.2065, 2022.

Thompson, V., Dunstone, N. J., Scaife, A. A., Smith, D. M., Slingo, J. M., Brown, S., and Belcher, S. E.: High risk of unprecedented UK rainfall in the current climate, Nature Communications, 8, 107, https://doi.org/10.1038/s41467-017-00275-3, 2017.