Throughout this response, the reviewer's text is presented in black, our response in blue

Dear authors,
Thank you for this interesting research, written up in a well-organised and clear paper. Overall, I have not hesitation to recommend your paper for publication. I do agree with you, that this kind of research, with continuous simulation of operational water management to test new information sources, methods, or strategies, is valuable for science and in particular for bringing findings forward to practice in an informed and iterative way.
I appreciate in particular your Conclusions section, and clear description of data used, methodology, and presentation of results.

We thank the reviewer for their kind words.

General comments
I have the following main comments:

- The authors assumed the water demand to be known in advance and to be equal to the observed reservoir releases (line 220). This may be an important assumption. If more than needed was pumped-in for storage, this would perhaps also lead to releasing more than the actual demand. Could the authors reflect on this? The actual releases are the result of the current water management priorities, which focus on water resources availability. Could this have led to the forecast value also being maximum for the rap scenario's? Could the authors address this with a limited sensitivity analysis, varying water demand? (if time, sensitivity analysis of other aspects would be interesting as well, e.g. towards the set-up and settings of the NSGA optimisation experiment.)

The release data that we use are only the controlled releases (from the outlet tower) and do not include spills, so we believe that our assumption that they reflect the demand is quite ok. Moreover, in the system, we have only modelled the refill period during winter where demands on the system are fairly stable/predictable. The forecast value is quantified with respect to the benchmark and both benchmark and DSP under all the scenarios assume the same demand. So very unlikely that changing the demand is going to have an influence on how better DSP does with respect to the benchmark. The point made about not knowing demand in advance is more relevant if this case were expanded to understand what happens with demand in the following summer, and whether we would really need to refill the reservoir or not. We will clarify these points in the revised manuscript. A sensitivity analysis is a good idea that we would like to address in future publications, but this is out of the scope of this study and besides, reviewers complained about this paper being too long. The message that we want to convey with this study is that yes, we can improve the performance of a realistic reservoir system using seasonal forecast using common and readily available methods and forecast products. Nevertheless, in the revised manuscript we will further discuss the possible reasons that have led to the forecast value being maximum for the rap scenario, such as the one pointed by the reviewer.

- To my view, the results show that the bias correction applied, did not work in this particular case study, Figure 3 (only changed sign of MAE from under- to over-prediction, as authors also indicate in line 364). Could the authors reflect on this in their section on limitations of the research? What may be the reason? Could other bias correction methods work better or is this not to be expected (e.g. perhaps higher forecast skill is needed to begin with, for post-processing methods to be effective)? The poor performance of the bias correction also connects to the following comment.

The main reason for the bias correction to fail is the DSP forecast was already doing relatively good in terms of skills in 3 exceptionally dry years (Figure 3) and worse in the rest, which are less dry and hence closer to the average climate conditions. After bias correction we worsened the forecast skills these 3 exceptionally dry years, but we improved the skills in the rest. Rather than having higher skills to begin with, in this case the bias correction would have performed better if these 3 dry years would have not been considered, i.e. under less exceptional climate conditions the bias correction would have been more effective.
We think that to find the best bias correction method as well as the best skill score is out of the scope of this study but would be very interesting to look at this in future publications together with the sensitivity analysis mentioned above. We will include this in the section 4.1 Limitations and perspective for future research and implementation.

- The Discussion section contains notes and even recommendations on the use of bias correction. In my view, the poor performance of the bias correction in this case study, and the fact that, as the authors point out, indeed there is only one particular case study analysed here, do not warrant such discussion on the merits of bias correction. Could the authors reflect on this, and depending on whether they agree or disagree, adjust the Discussion accordingly.

In the manuscript we neither recommend nor reject the use of bias correction. We conclude that more studies are needed to investigate the benefits of bias correction when seasonal hydrological forecasts are specifically used to inform water resource management (lines 392-393). Firstly, because this is a case study and secondly because what a skill score reflects may not be representative of the benefits or costs in terms of forecast value of applying bias correction. We will revise the manuscript to make sure that our conclusions do not sound like a recommendations.

- The Discussion section also recommends use of ensemble (probabilistic) forecasts in operational management, which is supported by the research findings, but then connects this to UK policy recommendations on long-term water resources planning. This I think is a bridge too far, and not needed. I would favour the Discussion to be less broad, and stay focused on the research findings presented (see my detailed comments for specific suggestions on where and how to make the Discussion section more specific). This leads to my next comment.

As a case study we do not aim to make general recommendations but rather bring into attention for future studies and practical applications the importance of some aspects or factors such as the uncertainty consideration. We believe that this reference to planning helps the reader understand that while rarely considered currently in short term management, risk-based approaches attempting to deal with the range of potential future conditions expected are already starting to become standard methods in the industry for long term planning. These are two fields that are strongly linked, where seasonal and long term planning are often the responsibility of the same practitioners/teams within companies, or at least teams that strongly interact, and that (could) apply fairly similar methodologies at different time scales.

- I miss a more in-depth discussion on the forecast skill of the DSP used, and the influence of forecast lead time throughout the analysis chain. The CRPSS results nicely show that only for the first two months the uncorrected forecasts have skill (the bias-forecasts do not have skill). Is this positive skill utilised by the operational water management strategies simulated. Could the authors suggest ways on how to capitalise more on this positive skill, e.g. by using DSP for the first 2-month lead time, and using ESP for months 3 to 6?

The Reviewer suggestion is interesting and potentially worth exploring. But we are not sure we will include it because of need to keep the paper concise and because it is not said that what

brings more skill also brings more value. We believe that a more in-depth analysis of the forecast skills-lead time relationship would need a sensitivity analysis what would fit better in a potential future publication already mentioned above. It's difficult to say a priori how a mixed DSP-ESP will perform. Our results overall suggest that inferring the forecast value from its skill may be misleading, given the weak correlation between the two (at least as long as we use skill scores that are not specifically tailored to water resources management). Running simulation experiments of the system operation, as done in this study, can shed more light on the value of different forecast products (lines 402-406).

- Lastly, to come back to the motivation of the authors to bridge science to practice, I would like to see observed and simulated releases for sample priority scenarios and years. These actual releases throughout a season is what operators will recognise and this will enable a discussion on how and to what extend the use of ensemble seasonal forecasts would lead to changes in operation.

We thank the reviewer for this suggestion. While observed releases may not be representative of the system in study, which is a simplification of the real system, we will consider this interesting suggestion, compatibly with the need to keep the manuscript not too long and with confidentiality issues (the reservoir system data used are property of the water company).