# Author's Response

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

## Reviewer 1

This an interesting and timely study into the value of forecasts for improving the performance of a simple water supply reservoir system with an operational trade off between augmentation of stored water through pumping and associated energy cost. The selected case study is appropriately simple and also informative for this type of analysis. Results are quite difficult to follow and key details are omitted from the method. The set of forecasts selected for use in the simulation are also poorly justified. Finally, I feel that the paper attempts to answer too many questions and would benefit significantly from more focus. For example, the analysis of the dynamical forecast product and its failure to provide skill over ESP is an interesting study in its own right, demanding much more in-depth analysis and interpretation than is offered in the paper. The operational section then addresses ESP vs dynamical and the additional question relating to importance of incorporating ensemble uncertainty. The paper would be much stronger if you were to focus on just one of these areas and deliver a more compelling conclusion backed up with in-depth analysis of a specific question. I recommend that the paper would be publishable if significant changes are made to simplify the overall story and provide further method detail as outlined in the comments below.

We thank the reviewer for their overall positive evaluation of our manuscript and the suggestions for improvement. We think the analysis of different scenarios is interesting (and other reviewers also seem to agree) and so we intend to keep it. Nevertheless, we appreciate that the manuscript writing is sometimes overly complex and that some analyses (for instance Fig. 6 in the original manuscript) raise more questions than they answer, so in the revised manuscript we have simplified some aspects of the Results section, in particular we have deleted Fig.6 and integrated some of its content into Fig.7 now Fig.6 in the revised manuscript). We have also modified Figure 4 to accommodate the Reviewers' comments (more below) and edited the manuscript throughout for more clarity.

- It's not clear what optimization framework is used to deal with the forecast ensemble. The deterministic approach using rolling horizon (e.g., ANGHILERI et al., 2016) is quite common and there are very few successful examples in the literature where the full ensemble is used to inform the decision. Please outline exactly how the ensemble is used in your optimization and then justify the approach. If this is a new approach it perhaps needs to be described in its own, separate publication.

In our optimisation framework we minimise the expected values of the two objective functions based on the 25 ensemble members (whereas Anghileri et al. 2016 optimised the objective functions evaluated at the ensemble mean). In the revised manuscript, we have given a clearer explanation of this in Sec. 2.3.3 and in the Formulation of the optimization problem (Supplementary material). Regarding the justification of this approach, as we mention on Lines 431-436 of the revised manuscript, the use of the full ensemble has been proved to improve the forecast value in several studies with shorter lead time (i.e. days instead of months), while deterministic approaches such as the one applied by Anghileri et al (2016) did not show significant value in seasonal forecasts. So in our work we used the full ensemble at seasonal scale and also analysed the impact of the ensemble size, which confirm the value of using the full ensemble (see Section 3.2.2 and Figure 5): as discussed in lines 430-431, RTOS outperforms the current operation when using the ensemble forecasts, but it does not if uncertainty is

removed and the ensemble mean is used. We have clarified this also in the Supplementary Materials, where the equations of the model and of the optimisation problem are shown.

- Given the skill scores achieved for the dynamical forecasts, it's not clear why these were pursued in the operational part of the study. What is the justification for using a forecast product that is demonstrated to be unskillful relative to ESP?

One of the objectives of our work was to explore the skill-value relationship and whether one can extract value from forecasts in support of water resource management even if their skill is still relatively low. This is what led us to pursue the evaluation of DSP forecasts as well as ESP. As we mention in the Discussion (lines 407-408) our results suggest that inferring the forecast value from its skill may be misleading. This study indeed contribute to show that seasonal forecasts can deliver benefits to inform operational decisions even if their skill is low (as often the case in extra-tropical areas, such as the UK), and that under certain scenarios DSP can provide higher value than ESP despite its relatively similar skill.

- I found the results quite difficult to follow, partly because it's hard to keep track of the various operational settings. Why not simplify by showing the Pareto front for each forecast set (as opposed to five schemes with different symbols/colors). This would be both more comprehensive and easier to understand. Also, the emojis in the key figures are not appropriate.

We agree with the reviewer that our Figures are quite dense; on the other hand, showing each Pareto front in a different plot may make comparison across sets more difficult. However, in the revised manuscript, we have modified Figure 4 by adding coloured circles to group points under the same operational priority scenario and dashed lines to link points using the same forecast product.

## Reviewer 2

The main merit of this paper is the proposal of the methodology. The paper forms a valuable contribution to the methodology of quantifying the value of forecasts, here in terms of water availability and energy cost. Probably the methodology is more widely applicable. Generally, the paper is well written, although sentences tend to be too long and their structure could sometimes be made clearer by repeating some short words. Unfortunately, the conclusions from this paper are not really valuable. The problem with the first two conclusions, namely 1) seasonal forecasts can increase value and 2) ESP is hard to beat, is that they are case specific, as acknowledged by the authors (line 470). The third conclusion (the relationship between forecast skill and value is complex) is a trivial one. Below, there is a quite long list of main points, which the authors have to address in my opinion: information about the observations should be given (p1), any procedure based a scenario or forecasts with more inflow than in the worst-case scenario seems beneficial, e.g. taking the median of the historical years (p2), the methodology should be better explained (ps 3, 5 and 6), Mliters are not a valid unit (p4), there is an issue with the bias correction (p7), different processing for the benchmark and other forecasts is questionable (p8), it is strange that the value of the driest years with DSP and ESP processing is not almost equal to the benchmark (p10) and the first part of the discussion section should perhaps be removed (p9). In my opinion should be published after making the suggested major revisions.

We thank the reviewer for their overall positive evaluation and suggestions for improvement. In the revised manuscript we have tried to shorten long sentences and simplify their structure. We have also addressed the specific points raised by the Reviewer, as detailed below.
As for the generalisability of our work, we agree that the methodology here employed is widely applicable, and we are sharing an anonymised version of the code we developed for other users. We have added the link to this toolkit in Section 4.1. As for the generalisability of the results and conclusions, we do not fully agree with the Reviewer. We believe that case studies are necessary to advance our understanding and they allow in-depth, multi-faceted explorations of complex issues. We think that while the results are case specific the conclusions have more general practical implications. First, the study demonstrates that higher forecast skills do not necessarily translate into higher forecast value in reservoir operation and that seasonal forecasts can deliver benefits to inform operational decisions even if their skill is low. Second, we show that the hydrological conditions and the decision maker priorities can have as much or even higher influence on the forecast value than the forecast skill. Third, the study demonstrates the importance of accounting for the forecast uncertainty and highlight the potential benefits with respect to deterministic approaches.

A section about observations (discharge and meteorological forcing) should be added.

In the supplementary material we have added additional information about observational hydroclimatic data

One of the results of this paper is that by basing the operational procedure on the forecasts, less energy for pumping is used while ensuring similar water availability (statistically over the years), compared to basing operational procedures on the worst-case scenario (driest historical year). It is my impression that any operational procedure based on forecasts or scenarios with more inflow into the reservoirs than in the worst-case scenario leads to less pumping and similar water storage, provided the increases are realistic. The authors confirm this in lines 395-397 for the case of applying a bias correction, which increases the inflows to the reservoirs and hence increases the value of the forecasts. So, the worst-case scenario is possibly easy-to-beat by any scenario with more inflow. Somewhere in the paper (in the discussion section?) the following

points need to be discussed. Can this effect on value of increasing the inflow be generalized? What is the value of the forecasts if the operators base their procedure on the scenario of the year with the median value of the historical inflows? I suggest making a calculation with such a scenario. By the way: are the calculations in the worst-case scenario deterministic?

We choose the worst-case scenario as a forecast value benchmark (instead of the median) as this is representative of the current operation of the system, and thus it enables us to show the potential benefits of using seasonal forecast with respect to the current approach. This scenario is actually not so 'easy-to-beat': our results (Figure 5) already demonstrate that deterministic optimisation under a scenario with higher inflows ("DSP-corr deterministic" in Figure 5) does not beat the worst-case scenario (which is also deterministic). In fact, while "DSP-corr deterministic" improves energy savings, it decreases the resource availability for any decision maker priority. We have clarified these points in the discussion (Lines 429-430).

I did not understand Section 2.1 – 1b and c. These paragraphs need to be rewritten. At this stage this paragraph is too abstract. Perhaps providing an example of each concept (operation objective function, optimizer, set of operational decisions) would help. Perhaps merging Sections 2.1 and 2.3 helps. Moreover, after 1b the "set of operational decisions is determined", so why is the operator again "selecting a set of optimal decisions" in 1c? Perhaps lines 124-126 helped me to understand a little bit of what you try to explain, namely that you use hindcasts to evaluate the performance of RTOS. If this is correct, just write that you use hindcasts to evaluate RTOS and discuss a possible operational application in the discussion section.

In this section we try to represent the process that a reservoir operator would follow. In 1.b the operator obtains a set of possible optimal decisions as a result of the optimization of the reservoir system in response to the forecasted inflows. Given that the optimisation problem has multiple objectives, it does not provide one optimal solution but set of Pareto-optimal solutions, each realising a different trade-off between the conflicting objectives. This is why in 1.c. the operator needs to select, according to their priorities, one of the optimal decisions among the ones obtained in 1.b. We have revised our description in section 2.1 to make the point clearer and we have also added examples (Lines 120-121, 146-148).

Replace all appearances of Ml and Ml per some time unit by $m_3$/s (per day is also ok), the common unit in the hydrological literature, e.g. in Figure 2, 3, 6 and 7.

We have replaced Ml by $m^3$ as suggested.

I did not completely understand 2.3.1. Was river R fed by the outflow of reservoir S1 before the dam of S1 was built? Also, it sounds ridiculous to pump water, that was released by gravity from S1 to R, back to S1. So, is the water in R at the location where it is pumped out of the river partly fed by rivers that are not connected to S1? Is S1 located at lower elevation than D, so the water flow needs to be pumped?

As mentioned in the manuscript, the gravity releases from S1 are used to support downstream abstraction during low river (R) flows/season (essentially in Summer). In contrast, during the high flow season (Nov to Mar), pumped inflows from R to S1 may be operated to supplement the natural inflows to S1. The water pumped out in R is fed by rivers that are not connected to S1. We have further clarified this point in Section 2.3.1 and we have improved the system schematic (Figure 2) to make clear that R is fed by both a natural catchment and the gravity release from S1.

I did not really understand those “rule curves” (lines 173-179). Add a figure with a rule curve. It is not clear to me how the refilling ($U_{R,S1}$) is done. Is the “missing water” immediately refilled or is the refilling spread over time until April 1, using the optimizer? In the latter case, how does the optimizer work? Can you give an example of an operational decision? What level is targeted on April 1? How does the operational procedure work for probabilistic forecasts? Since there is variation in resource availability by April 1 (e.g. in Figure 4), storage is not equal to the target on April 1. Can storage be larger than the target or is all water above the target spilled? Can storage be less than the target? Perhaps only in S2 and not in S1 because water can be pumped into the latter basin?

The rule curve applied in the current operation procedures defines the storage level at which pumps are triggered. By the 1 April the objective is to be at full storage. Water is only spilled when the storage is higher than the reservoir capacity. The rule curve is only applied in the current operation approach (benchmark) and not in the RTOS approach. We have further clarified this in section 2.3.6 (Lines 270-272) and in the Supplementary material.

The method of bias correction is not correct (199-203). The number of years used to compute the multiplication factors differ per target year. I suggest using the common leave-one-year-out-method, i.e. the factor for each target year is computed from the data of all other years, including years later than the target year. Your method suggests that it is not allowed to use data from future years but there is no problem in doing so if different years are independent of each other.

Using the leave-one-year-out method works statistically but it does not represent what the operator could have achieved historically if using seasonal forecasts, because at each simulated decision time-step the operator would have only been able to use data up to that moment, and not future years. Given that our methodology aims to simulate the behaviour of the operator and the operational decision-maker process, we must assume that the operator can only have access to past data and hindcasts for the bias correction. We have clarified this point in Section 2.3.2 (Lines 207-208).

Line 264 “but with three main variations”: Why do you treat the benchmark differently? This implies that if the forecast is equal to the benchmark, the forecast value differs, which seems undesired.

We treat the benchmark differently because it represents the current operation (Lines 263-266) procedures and we aim to assess the potential of using a real-time optimization system informed by seasonal forecasts in place of current procedures (Lines 102-103). It is virtually impossible that the forecast is equal to the benchmark because it is not possible that the ensemble members are all equal to the worst-case inflow sequence.

The first general lesson in the discussion is “First, we found that the use of bias correction to improve the skill and value of DSP forecast is less straightforward than possibly expected” (lines 381-382). I do not agree. Such an expectation, namely that the forecast skill generally improves due to bias correction (is that the expectation?), just does not exist. Your study indeed confirms that this is a naive expectation. So, remove lines 378-393 or reformulate them. By the way: your bias corrections are based on observations of precipitation and temperature and not on the output (hydrological variables!) of ESP forecasts. So, I did not understand the sentences related to ESP in lines 387-388 and 391-392.

Reading the literature, we have the impression that studies tend to show the benefits of bias correction and it is often recommended or even required for impact assessments. Here some examples:

- From Crochemore et al, 2016: "ECMWF forecast **skill is generally improved** when applying bias correction"
- From Ratri et al (2019): "**Uncorrected meteorological forecasts are not suitable** as direct input for quantitative models, such as those used in agriculture and water management (Schepen et al. 2016). The bias **should be corrected** because it can lead to significant errors in impact assessments (Murphy 1999)." https://doi.org/10.1175/JAMC-D-18-0210.1
- From Schepen et al. 2016: "GCM forecasts suffer from systematic biases, and forecast probabilities derived from ensemble members are **often statistically unreliable**. Hence, it is necessary to postprocess GCM forecasts **to improve skill and statistical reliability**." https://doi.org/10.1175/MWR-D-13-00248.1
- From Zalachori et al 2012: "To **improve the quality** of probabilistic forecasts and provide **reliable estimates** of uncertainty, statistical processing of forecasts is **recommended** (Schaake et al., 2010)" https://doi.org/10.5194/asr-8-135-2012
- From Jabbari and Bae 2020: "Numerical weather prediction (NWP) models produce a quantitative precipitation forecast (QPF), which is vital for a wide range of applications, especially for accurate flash flood forecasting. Since NWP models are subject to many uncertainties, the QPFs **need to be post-processed**. The NWP biases should be corrected **prior to their use as a reliable data source** in hydrological models." https://doi.org/10.3390/atmos11030300

We have clarified this expectation that the forecast skill generally improves due to bias correction by citing in the Introduction several studies such as the ones above (Lines 71-73).

We agree that the sentence on lines 387-388 (original manuscript) was badly formulated and we have now replaced it by: "However, the result points at a possible intrinsic contradiction in the very idea of bias correcting based on climatology." What we aimed to communicate in this paragraph is that since both bias correction and ESP forecast are based on climatology, the bias corrected DSP forecast skills tend to become closer to ESP skills. However, ensuring this skill level with bias correction (Crochemore et al. 2016) may not be the best approach especially under conditions significantly drier or wetter than climatology, which are likely the ones when water managers can extract more value from forecasts. We have further clarified this point in the reviewed manuscript (Lines 386-397)

It is strange that the increase in the value of the system with DSP or ESP forecasts relative to the value of the system based on the worst-case scenario is highest in the driest years (e.g. lines 408-409), while those driest years resemble the worst-case scenario more than the other years. You need to explain this.

The benchmark tends to pump more water during the driest years because the lower storage level is more likely to cross the rule curve and trigger the pumped inflows. This explanation has been now included in the discussion (Lines 414-416).

**General text points**

1) The authors often use the term "bias correction" without mentioning what is corrected. As far as I understood, the forcing of the hydrological model is corrected but if you do not repeat mentioning this now and then, it is confusing because the

output of the hydrological model, i.e. inflow to the reservoirs, can also be bias-corrected. So, replace at numerous places “bias correction” by “bias correction of the meteorological forcing” or “bias correction of the forcing”.
This has been corrected accordingly

2) In general sentences are too long, making the manuscript difficult to read. So, shorten sentences where there is an opportunity and make the structure of long sentences clearer, especially by adding some words in sentences with “and”. I made some suggestions below (e.g. 15-16 in the abstract).
In the revised manuscript we have tried to shorten long sentences and simplify their structure.

3) “Uncertainty (considerations)” is used to discuss effects of ensemble size and the probabilistic nature of the forecasts. Replace throughout the paper the vague term “uncertainty (considerations)” by the more explicit terms “ensemble size” and “probabilistic/deterministic nature of the forecasts”.
We have clarified wherever appropriate that uncertainty in forecast is represented through an ensemble.

**Minor points**
16 Insert “to” before “other factors”. Changed accordingly

17 “Some of these factors” is too vague. Write which factors have a significant correlation with forecast value (see point Figure 6 below). Corrected accordingly

24 Add reference to endorse the statement that climate variability is increasing. A new reference has been added

44-45 Replace “it provides” by “they provide” and add “that they” before “reflect”. Changed accordingly

54 I miss the logic behind “i.e. ESP …”. Replace this part of the sentence or clarify the logic. We have removed it

55-56 Reorder sentence to “The possible improvements of supply-hydropower systems operation due to the use of ESP were assessed by Alemu …” Changed accordingly

69-71 Remove this sentence: this distracts too much. We have kept this sentence because it raises an important point that is later discussed, is bias correction necessary?

78 weaker compared to ….? We have added “than in hydropower production or flood management systems”

83-92 Remove these sentences about some of the many existing metrics. It is not efficient to read about metrics not used in this paper and the metrics used in this evaluation are introduced 2.3.5. We have kept these sentences because they raise an important point that is discussed in this study, i.e. inferring the forecast value from its skill may be misleading and the need for skill scores better tailored to the purpose of the studied system, e.g. such as water resources management

94 Replace “this” by “the”. Changed accordingly

99 Replace “simulate and compare” by “assess”. Simulate performance sounds strange and it is not clear from the rest of the sentence what is compared with what. Changed accordingly

151 Insert “diagram” after “Pareto front”. Changed accordingly

163 Consider removing all text about two companies. It is irrelevant for your story while it is making your story more complex. This text has been removed, and water company has been replaced by system operator.

168 Insert “(R)” after “river” Added accordingly

188 Remove period. Removed

198 Did you also use a multiplicative factor for temperature? No, we use an additive factor, we have corrected this in the text

211 $U_{S1,D}$ is also a pumped water flow according to Fig. 2. Yes, that’s right, as described in section 2.3.1

214 Replace “The first objective function” by “Pumping savings” and “The second function” by “Resource availability”. Changed accordingly

219 Replace “the 15%” by “only 15%”. Changed accordingly

227 Rephrase sentence as follows: “They represent five different trade-offs of operational priorities, according to their relative importance” The sentence has been rephrased as follows: “They represent five different trade-offs of operational priorities, according to the relative importance given to each performance objective”

237 Remove sentence. Removed

247 Remove “and for a given lead time”. The role of lead time comes some sentences below. Removed

251 Replace “lead time” by “range of the lead time (we use monthly ranges)” and “CRPS values” by “individual CRPS values”. Changed accordingly

252 Replace “CRPS” by “individual CRPS values”. Changed accordingly

253 I suggest to replace “mean error” by “discharge bias” since bias in the common word for mean error and the addition of “discharge” helps to distinguish this bias from that in the forcing. I also find the equation redundant. Just write that the bias is the difference between the means of the forecasts and the observation over all …….” We have kept the original terminology, i.e. “mean error”, because we believe that using the term “discharge bias” may infer that it can be corrected with bias correction.

262 Add “(1975-76)” after “drought on records”. Added

Figure 3 Is this the sum of the inflows to both reservoirs? Are these results for the whole year or a specific part of the year? In the legend of the lower panel “2006” should be replaced by

"2016". This has been clarified in the Fig 3 caption. 2006 is correct, as mentioned in the caption the 3 particularly dry winters are represented and one of them corresponds to 2005-2006.

334 and 338 Replace "Figure 6" by "Figure 5". Changed accordingly

Figure 6 If I just look at these graphs, I get the impression that there is no significant relationship in any of these graphs. However, according to your p-values relationships are significant at the 90% confidence level in panels b and e. Is the calculation of the p-values correct? Or are those low p-values due to using the Spearman coefficient instead of the Pearson coefficient? I think you should use Pearson unless you have good reasons to use Spearman. We have removed this figure and any reference to correlation or p-values.

344 Replace "skill" now and then by "forecast skill", to remember the reader what type of skill this is. This part has changed in the new version of the manuscript

358 Replace "year" by "years". Changed accordingly

395 Replace "reduce" by "reduces". Changed accordingly

399 "improvement of forecast accuracy in some direction". What do you mean by "accuracy"? For me this is something like the root-mean-square-error, which means that there is only one desired direction, namely towards 0. Do you mean something like "a change towards either higher or lower values can be more valuable than a change in the other direction"? We agree the sentence was confusing and we have now removed it and revised the paragraph.

411 Is "Initial storage (total storage value)"equal to "Initial storage" in Figure 6? For clarity, be consequent in the use of specific terms. Moreover, panels c and h in Figure 6 do not show a significant correlation (p-values of 0.21 and 0.80). As mentioned above this figure has been removed

428 Remove ")" Removed

447 Replace, for clarity, "seasonal forecasts" by "seasonal meteorological forecasts". Changed accordingly

465 Insert "of" before "the institutional" Changed accordingly

466 Insert "of" before "the most" Changed accordingly

478-479 Replace "but also the methodology in the first place" by "but in the first place by the methodology". Changed accordingly

## Reviewer 3

This paper addresses comprehensively a topic that increasingly requires investigations: the value of seasonal forecasts for real-life applications. Until recently, many studies have worked on assessing or improving forecast skill, but still few manage to link this skill to value. In addition to being innovative, this study investigates the issue through different uncertainty lenses which makes it a very strong contribution to the field and results in valuable findings both for the scientific community but also for water management stakeholders. Overall, the paper, its ideas, structure and methodology are of high quality. However, additional information, for example about the data used and the forecast skill evaluation, would be necessary to support the analysis. In addition, I have some concerns about the validity of the results on linking skill and value due to the chosen methodology. These are detailed hereafter.

We thank the reviewer for their kind words.

The paper would benefit from a Data section, presenting for instance the data used as reference in the bias correction, the inflows used as reference in the forecast evaluation, or the demand model/observations.

We have added a section titled 'Observational hydrological data' in the Supplementary material (as other Reviewers commented that the manuscript is too long already).

Some additional information would be needed in Section 2.2 on the forecast skill evaluation. More specifically, (1) In Figure 2, two inflows feed the two reservoirs (Is1, Is2), which inflow is being considered when evaluating forecasts? (2) Which time period is used to evaluate the forecasts? In Figure 8, November to April is shown, but in the forecast methodology (Sections 2.2 and 2.3.2) or in Figure 3, no specific time period is mentioned. I would recommend mentioning these two points in 2.1/2a.

The inflow to S1 is used to evaluate the forecast skills from Nov to Apr. We have clarified and mentioned these two points in the case study section (2.3) and in particular in Sec. 2.3.5 (we would rather keep sections 2.1 and 2.2 as general as possible because this methodology is meant to be generalizable and applied by others).

Since the goal of the paper is to assess the added value of dynamical seasonal forecasts for water management, and since authors evaluate the skill and performance of these forecasts against observations, it would be important: (1) to add information about how HBV was calibrated and setup for the area, or at least, to mention its performance (mean error) in simulating past inflows to the reservoirs (giving the possibility to make a parallel with the results in Figure 3); (2) to mention even briefly the hydrological regime upstream the reservoir system, as well as the interannual variability, which will define the added value of a method like DSP over ESP.

We have added this information in the 'Observational hydrological data' section of the Supplementary material.

I have concerns about the methodology chosen to link value and skill, and therefore about some of the subsequent results (first two paragraphs of Section 3.2.3). The authors are trying to link a skill obtained from comparing dynamic forecasts with forecasts based on past climatology (1981-20XX), with a value obtained from comparing dynamic forecasts with a worst-case scenario (1975/1976). On the one hand, the skill is based on the comparison of DSP-corr with ESP, meaning that its value solely indicates the gain in performance from using ECMWF SEAS5 instead of historical precipitation and temperature. On the other hand, the value is obtained by

comparing DSP-corr with a benchmark based on a worst case scenario (1975-1976). These choices result in skills and values that cannot be compared. Instead, the skill computed in the paper could be related to the gains/losses in value between DSP-corr and ESP (difference between the blue and green bars of Figure 7). Reversely, the value computed in the paper could be compared to a skill whose benchmark would be the performance obtained by using the worst case scenario (1975-76) as forecast in all years. This is a major point that should be addressed prior to publication to ensure the validity of the conclusions.

We understand the reviewer concerns and we agree that there is somehow an inconsistency in the definition of the skill and the value, given their different benchmarks. However, we think it is not easy to resolve such inconsistency. In fact, we use ESP as a benchmark for the forecast skill because this is the standard practice in the literature (Pappenberger et al., 2015, Harrigan et al., 2018) as it is more likely to demonstrate the "real skill" of the hydrological forecasting system (lines 253-255), whereas the worst case scenario is never applied as a forecast skill benchmark and it is not likely to demonstrate the "real skill". As for the forecast value, we use the worst-case-scenario as benchmark because it is the scenario applied in the current operation approach and hence it is a benchmark that can demonstrate the "real value" of moving away from that approach towards using seasonal forecasts (lines 265-266). We would thus be reluctant to change any of these benchmarks, as they are appropriate for their different purposes.
However, we agree with the Reviewer that this inconsistency should prevent one from directly comparing the numerical values of the forecast skill and value. So, we have now removed Figure 6 of the original manuscript, which incorrectly suggested that such value-by-value comparison can be drawn. On the other hand, we think one can still compare the ranking of forecast products induced by the forecast skill with the ranking based on the value, which is the main point of our work. This is what we refer to when discussing the "forecast skill-value relationship" (line 98). With this clarification, we think the year-by-year analysis of the different optimisation results (Figure 7 in the original manuscript, now Figure 6 in the revised version) still provides interesting insights, and supports the main conclusion that "the relationship between the forecast skill and its value for decision-making is strongly affected by the decision maker priorities and the hydrological conditions in each specific year" (lines 491-492).

**Specific comments**
L45: "and reflects the risk-averse attitude"

L68: "to continue increasing in coming years"

L75-76: More literature review would be needed on past works on seasonal (climate or hydrological) forecasts value. Here are a few examples to consider:

Bruno Soares, Marta, Meaghan Daly, and Suraje Dessai. 'Assessing the Value of Seasonal Climate Forecasts for Decision-Making'. *WIREs Climate Change* 9, no. 4 (2018): e523. doi:10.1002/wcc.523.

Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti. 'From Skill to Value: Isolating the Influence of End-User Behaviour on Seasonal Forecast Assessment'. Hydrology and Earth System Sciences Discussions 2020 (2020): 1–20. doi:10.5194/hess-2019-659.

Parton, Kevin A., Jason Crean, and Peter Hayman. 'The Value of Seasonal Climate Forecasts for Australian Agriculture'. *Agricultural Systems* 174 (2019): 1–10. doi:10.1016/j.agsy.2019.04.005.

Thank you for the interesting references. In this part of the literature review we refer to the lack of pilot studies demonstrating the value of seasonal forecast products in UK and Europe but it was not very clear in the manuscript so we have further clarified this in the new version of the

manuscript. We make reference to past general studies on seasonal forecast value in reservoir operation in the 2nd paragraph of the Introduction.

We thank the reviewer for the references. The Bruno Soares et al (2018) study has been included it in the Introduction as well as Giuliani et al. (2020). The latter is the first pilot application in Europe demonstrating the value of seasonal forecast products, in this case ECMWF, that we have found in the literature. It must be noted that this study is still under review.

The study by Parton et al (2019) is a review of the use of seasonal forecast in agriculture in Australia. However, in the context of our study, we could not find any reference using seasonal forecast for reservoir operation.

L107: In this sentence, the authors seem to assume that their results could be generalized to extra-tropical regions and that the potential for use is region-dependent. However, as the authors state, results (forecast value) are in fact highly dependent on the investigated system, and therefore this sentence may sound a bit too ambitious compared to the paper objectives.

We have deleted the reference to extra-tropical areas.

Section 2.1: I found this section very clear and helpful.

We thank the reviewer for this positive comment.

L118: Here, I was wondering what type of demand model was used. It is later in the discussion that the authors mentioned that the demand is based on observations. This would be worth mentioning earlier in the text, and would fit in a Data section.

This is mentioned in section 2.3.3. As we would like to keep this section as generalizable as possible we mention this later, in the Case study, in particular in section 2.3.3.

L142 (2.a): More information would be needed at this stage about the forecast skill evaluation, for example by mentioning the reference and the benchmark used.

Since we would like this section to be used by others and hence to be as generalizable as possible, we have deleted this reference to the specifics of our case study (explained in section 2.3) in 2.a as well as 2.b

L160-161: The notation for units is not consistent between the text (Ml) and Figure 2 (ML).

This has been corrected. In the new version of the manuscript we use m3 instead of ML

L176: Did you mean a “reasonable chance”?

Yes, thank you

L182: Please consider adding the following reference for ECMWF SEAS5:

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H. and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12(3), 1087–1117, doi:10.5194/gmd-12-1087-2019, 2019.

Thanks, this reference has been added.

L182-183: I suggest writing “The ECMWF SEAS5 hindcast dataset” because it is the hindcast dataset that includes 25 members, while the real-time operational ECMWF SEAS5 has more members.

Changed accordingly.

L187: There are two points on this line.

Corrected

L195: This statement needs to be refined. Linear scaling and distribution mapping may lead to similar results in terms of bias removal, but they will have very different results on the forecast themselves, e.g. linear scaling will just shift the ensemble, while distribution mapping will have a different correction for each member, resulting in larger impacts on the spread.

We have added this clarification

L196-198: Usually a difference approach is applied to calculate the correction factor for temperatures (see e.g. Lucatero et al. 2018, Teutschbein and Seibert 2012). Similarly the correction factor is added/subtracted to correct the raw forecasts. If a ratio approach was applied for temperatures, I would suggest adding a short justification of this choice. For instance, how are negative temperatures handled?

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J. and Jensen, K. H.: On the skill of raw and post-processed ensemble seasonal meteorological forecasts in Denmark, Hydrology and Earth System Sciences, 22(12), 6591–6609, doi:10.5194/hess-22-6591-2018, 2018.
Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456–457, 12–29, doi:10.1016/j.jhydrol.2012.05.052, 2012.

We used an additive factor in the case of temperature but we did not mention this in the manuscript. We have clarified this in the new version of the manuscript

L206: “with what was done”

Corrected

L208: The ensemble size has an impact on the CRPS (Ferro et al. 2008) and therefore, on the skill when the benchmark has a different ensemble size than the evaluated system. This could impact the results presented in this paper, and I would suggest adding comment about the impact of the varying ESP ensemble size on your skill results.

Ferro, C. A. T., Richardson, D. S. and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Met. Apps, 15(1), 19–24, doi:10.1002/met.45, 2008.

We thank the reviewer for the comment and reference suggested. However, the ensemble size is very unlikely to have a considerable effect on the results of this study because a) the ESP forecast quality in this study is too limited and the time range too short to have an important effect on the forecast skills and b) the results already show that the forecast value is fairly insensitive to changes in the ensemble size (Figure 5) unless a very low number of members (5 or less) are considered. Nevertheless, it would be an interesting question to explore by means of a sensitivity analysis in a future publication.

Section 2.3.5: I would suggest presenting the CRPS (L245-252) before presenting the CRPSS (L239-244) since it would define the CRPS notation before using it in the skill equation. There are also inconsistencies between the first two equations (CRPS and CRPSS) and the third one (mean error), such as the notation for the observations (*y* vs $I_{Obs}$), the notation for the system (*Sys* vs *Syst*), the mention of the lead time only in the third equation but not in the previous ones.

We have corrected the inconsistencies and we have also changed the order and now CRPS is presented before CRPSS as suggested by the reviewer

L237-238: This sentence is repeated twice.

We have removed the duplicated sentence

L243: By choosing ESP as a benchmark, you also make the decision of only analysing the added skill from using dynamic weather forecasts over meteorological history, it would be worth mentioning.

*We have mentioned this as suggested by the reviewer.*

L256: Do I understand correctly that the mean error is computed over a time window that varies with the lead time you consider, for example to evaluate lead time 6 weeks, you average the mean error from week 0, 1, … to 6? Is the same calculation method applied when computing the CRPS?

The average CRPS or mean error for a given lead time is equal to the average of the individual CRPS or mean error values obtained for the forecasts published across the time frame. We have clarified this in the revised manuscript.

L279: I suggest replacing “gets lower” with “gets larger in absolute value”.

Here we refer to DSP-corr which gets lower and not larger for longer lead times

L334: The reference should be to Figure 5 instead of Figure 6.

Corrected

L358: “two specific years”

Corrected

L381-393: In this paragraph, conclusions are likely only valid for linear scaling, and not for any bias correction. I would suggest being more specific throughout the paragraph.

We have mentioned that in particular we refer to linear scaling

L446: “that should be kept in mind”

Corrected

Figure 6: I wonder why, for each of the five operation scenarios, the authors do not display both DSP-corr and ESP in the same graphs.

This figure has been removed in the revised version of the manuscript after comments from other reviewers

Figure 6b and 6g: Shouldn't the absolute mean error be used to assess the correlation between the increase of resource availability and the performance, since both high positive and low negatives reflect poor performance?

This figure has been removed in the revised version of the manuscript after comments from other reviewers but we wanted to keep the sign of the error because we have observed that underestimation has different impacts on the forecast value as compared to overestimation.

Figure 7: It could be interesting to see the mean error and the CRPSS in this figure, even if they are not correlated.

We have added both.

Supplementary material – Figure 8: In the legend, I could not remember that 1975-1976 was the worst-case scenario. It could be worth mentioning it again at the end of the caption.

We have specified this in the end of the caption too

Supplementary material – Figures 9 to 13: The CRPSS is used to try and explain the gains in value from using ESP. Here, it is not clear which benchmark the performance of ESP was compared to in order to compute the skill, if it is indeed the CRPSS of ESP being shown. If not, then please refer to the fourth general comment.

These figures have been removed in the revised version of the manuscript after comments from other reviewers

## Reviewer 4

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 reasonable, and they should not exceed the actual demands, as suggested by the Reviewer. Moreover, in our experiment we have only simulated the refill period during winter when demands are fairly stable and predictable, hence also the assumption that demand are known in advance should not be too limiting. Finally, the forecast value is quantified with respect to the benchmark and both benchmark and DSP under all the scenarios assume the same demand, so it is unlikely that changing the demand is going to have an influence on how better DSP does with respect to the benchmark. We have clarified these points in section 2.3.3 and in the Details of the optimization (Supplementary material). This said, we agree that one could test the influence of the demand assumption through a sensitivity analysis, but we would leave this experiment to future studies, especially as the manuscript is already quite long (and other Reviewers asked to shorten it already). We have thus only mentioned the point in the Discussion of future research (Lines 462-467).

- 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 others, which are closer to the average climate conditions. After bias correction, the forecast skills are worsened

in these 3 exceptionally dry years, but improved in the others. We have included this explanation in the Discussion. 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 interesting to look at this in future works together with the sensitivity analysis mentioned above. We have included this in the section 4.1 Limitations and perspective for future research and implementation (Lines 462-467).

- 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.

We do not intend to 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. We have now revised the manuscript to make sure that our conclusions do not sound like recommendations (Lines 395-397)

- 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 in future studies. We have not included it because of the 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. It is difficult to say a priori how a mixed DSP-ESP would perform. Besides, 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). We have included this in Sec. 4.1. as a suggestion for future studies (Lines 462-467)

- 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 the interesting suggestion. We cannot, however, publish the company data on reservoir releases. The value of the data may also not be representative also of the simplified system in study, which is part of a broader conjunctive use system, where releases may be driven by broader resource considerations.

**Detailed comments:**

- line 275: Indeed the bias corrected DSP "improves" skill for longer lead times, but only from negative skill to less negative skill. I would suggest to point that out.

We have pointed this out.

- line 281: Yes, with bias correction there is "some improvements for some lead times", but still with negative skill. Rather than pointing out "some improvements for some lead times", I think it is more relevant to point out here that the bias correction as applied here, in this case study, is not working and even has adverse effect on forecast performance.

We have modified this paragraph to reflect the in summary the bias correction does not produce an improvement in the forecast skills

- line 311: The question is why? Again (See my first general comment, and note that resource availability in the results varies only with 1-2%) this may indicate a constraint set-up, favouring a focus on rao. Please reflect on this.

Rather than a constraint set-up, this can simply be explained by the reduction of the pumping costs after bias correction, which is a consequence of the overestimation of the inflows. We have now included this explanation to rap and *bal* Pareto dominating the benchmark which can be also applied to *rao* (lines 400-401).

- line 382: "Our results show that on average bias correction slightly improves the DSP forecast skill (as measured by CRPSS and mean error)". I do not agree here. When looking at the results, also on average, bias correction reduced the CRPSS for the first 2 months lead time where the DSP had skill (bias correction made skill less negative for further lead times, but still negative), and it changed the sign of the average error but did not reduce it, so bias correction was not working well.

We agree on this with the reviewer and we have modified this paragraph accordingly.

- line 392: I agree with the authors. Based on this particular study, not much can be discussed on the merits of bias correction. Instead I would recommend to focus discussion in the paper more on the skill of the ensemble DSP (slightly positive for the first 2 months), how and why this is or is not being used in the water system operations simulated (see my last two General comments above).

We thank the reviewer for this interesting suggestion. As mentioned above, we have included the possibility of combining DSP with ESP as a way to improve seasonal forecast as a suggestion for future studies.

- lines 399-404: While I agree with this statement in general, I do not think the research presented provides sufficient supporting evidence. Nor is it a surprise or a new finding.

It is true that this cannot be generalized but we believe that the results of this particular study support the need to further investigate the adequacy of different skill scores to evaluate the forecast value for water management purposes. To our knowledge, the literature on forecasts evaluation often overlook the issue of defining scores that are specifically tailored to a particular purpose.

- lines 413-416: This is quite a strong recommendation not substantiated with any analysis/numbers on what these 'costs' are. The authors could consider leaving this out.

This was not meant to be a recommendation but a hypothetical and explanatory scenario where one product could be preferred over the other. As mentioned above, 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. We have revised the manuscript to make sure that they do not sound like recommendations.

- lines 433-436: I am not sure if the link with Long-term water resources planning is appropriate, and also I think it is not needed to make the case of risk-based operation on the basis of ensemble (probabilistic) forecasts. The results of your study do show this nice enough. The authors could consider leaving this out.

As already mentioned above, we believe that this reference to planning may be useful for some readers (particularly UK practitioners) who are maybe familiar with risk-based approaches in long-term planning, and point them to the fact that similar concepts may be usefully applied in short-term management.

- lines 458-467: Yes, developing such toolkit and making available to the water management organisation is very valuable. Indeed question here would be to what extend the toolkit is customised to the specific case study, and how much time/effort customisation to a new case study would require. Could the authors reflect on this in the text?

This toolkit was tailored to this study case. It is an interesting and relevant question that we are still not able to answer but we are now testing the use of the toolkit as a mechanism to support knowledge transfer to practitioners and to evaluate how easily they can customise it, and this will be the focus of a future publication. We have clarified this in section 4.1 (Lines 475-478).

- line 475: DSP is now more readily available from international weather forecast centres and more easily processed, such that this by-sentence on "ESP being more easily derived", in my view is perhaps becoming less relevant. Also because, as the authors describe, they have provided a Toolkit for ease-of-use.

From our own experience and through our collaboration with practitioners at water companies, downloading and post-processing the seasonal forecasts still needs a considerable level of expertise and while weather forecast centres, such as ECMWF, are gradually facilitating the access to the data, the process of bias correction is still quite difficult. Not only because they still do not provide with such tools but because first we need to decide whether applying bias correction or not and also select (and understand) the most adequate bias-correction method. And as we have seen in this study this is still not clear. This has been added to the Discussion (Lines 418-422)

Please see for suggested technical changes (editorials) the annotated pdf.

Apart from the pdf with the reviewer comments we couldn’t find an annotated pdf.

Thank you and with best regards.

# List of all relevant changes

Dear editor,

Thank you for the opportunity to respond to the reviewers' comments and submit a revised manuscript.

The reviews were positive overall and very helpful to improve the manuscript. However, sometimes the reviewers asked for changes or additions to the paper in different directions so that we found it impossible to accommodate all of them simultaneously. Some reviewers appreciated the applied nature of our work and suggested to give more details about the case study, while another reviewer asked for more details about the methodology; all reviewers suggested further analyses of various aspects of our modelling chain, while also asking for a more concise, shorter paper. Some of these suggestions for further analysis are very interesting and are worthwhile exploring in future publications, but in our opinion are beyond the scope of this paper.

So, we have tried to accommodate as many as possible of the reviewers' comments while also maintaining the focus on the key question of our work (what is the value of seasonal forecast for water resource management?) and mentioned in the Discussion session the further analyses that we think are beyond the scope of this paper.

List of relevant changes. We have:

- edited the manuscript throughout for more clarity
- added more references
- improved the following aspects of the Methodology:
    - clarified the formulation of the optimization problem
    - further justified the use of ensemble modelling instead of deterministic, the use of past data for bias correction method and the use of ESP as forecast value benchmark
- improved the following aspects of the Case study description:
    - the system schematic (Fig. 2)
    - the explanation of how the rule curve operates
    - added information about the observed hydrological data in the Supplementary material
- simplified the Results section:
    - removed Fig. 6 (in the original manuscript), which we realised was making the storyline unnecessarily complex and was raising more questions than it was answering. Some of the content of that Fig. 6 has been integrated into what was Fig.7 (and is now Fig.6 in the revised manuscript).
- improved the Discussion section:
    - further clarified the reasons of why bias correction deteriorates the forecast skills but improves the forecast value and further discussed whether ESP or DSP should be applied
    - added further possible future research.
- shared a link to an anonymised version of the code for application of our methodology.

Marked-up manuscript version

# Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK

Andres Peñuela[1], Christopher Hutton[2], Francesca Pianosi[1, 3]

[1]Civil Engineering, University of Bristol, Bristol, BS8 1TR, UK
[2]Wessex Water Services Ltd, Bath, BA2 7WW, UK
[3]Cabot Institute, University of Bristol, BS8 1UH, UK

*Correspondence to*: Andres Peñuela (andres.penuela-fernandez@bristol.ac.uk)

**Abstract.** Improved skill of long-range weather forecasts has motivated an increasing effort towards developing seasonal hydrological forecasting systems across Europe. Among other purposes, such forecasting systems are expected to support better water management decisions. In this paper we evaluate the potential use of a real-time optimisation system (RTOS) informed by seasonal forecasts in a water supply system in the UK. For this purpose, we simulate the performances of the RTOS fed by ECMWF seasonal forecasting systems (SEAS5) over the past ten years, and we compare them to a benchmark operation that mimics the common practices for reservoir operation in the UK. We also attempt to link the improvement of system performances, i.e. the forecast value, to the forecast skill (measured by the mean error and the Continuous Ranked Probability Skill Score) as well as ~~other factors such as~~ to the bias correction of the meteorological forcing, the decision maker priorities, hydrological conditions and ~~level of uncertainty consideration.~~the forecast ensemble size. We find that ~~some of these factors control the forecast value much more strongly than~~in particular the decision maker priorities and the hydrological conditions exert a strong influence on the forecast skill~~.~~ value relationship. For the (realistic) scenario where the decision-maker prioritises the water resource availability over energy cost reductions, we identify clear operational benefits from using seasonal forecasts, provided that forecast uncertainty is explicitly considered~~.~~ by optimising against an ensemble of 25 equiprobable forecasts. These operational benefits are also observed when the ensemble size is reduced up to a certain limit. However, when comparing the use of ECMWF-SEAS5 products to ensemble streamflow predictions (ESP), which are more easily derived from historical weather data, we find that ESP remains a hard-to-beat reference not only in terms of skill but also in terms of value.

## 1. Introduction

In a water-stressed world, where water demand and climate variability ~~are increasing, it is essential to improve the efficiency and lifespan of existing water infrastructure along with, or possibly in place of, developing new one~~(Stocker et al., 2014) are increasing, it is essential to improve the efficiency of existing water infrastructure along with, or possibly in place of, developing new assets (Gleick, 2003). In the current information age~~,~~ there is a great opportunity to do this by improving the ways in which we use hydrological data and simulation models (the 'information infrastructure') to inform operational decisions (Gleick et al., 2013, Boucher et al., 2012).

Hydro-meteorological forecasting systems are a prominent example of information infrastructure that ~~has a huge potential for improving~~could be used to improve the efficiency of water infrastructure operation ~~efficiency~~. The usefulness of hydrological forecasts has been demonstrated in several applications, particularly to enhance reservoir operations for flood management (Voisin et al., 2011, Wang et al., 2012, Ficchì et al., 2016) and hydropower production (Faber and Stedinger, 2001, Maurer and Lettenmaier, 2004, Alemu et al., 2010, Fan et al., 2016). In these types of systems, we usually find a strong relationship between the forecast skill (i.e. the forecast ability to anticipate future hydrological conditions) and the forecast value (i.e. the improvement in system performance obtained by using forecasts to inform operational decisions). However, this relationship becomes weaker for water supply systems, in which the storage buffering effect of surface and groundwater reservoirs may reduce the importance of the forecast skill (Anghileri et al., 2016, Turner et al., 2017)~~,~~ particularly when the reservoir capacity is large (Maurer and Lettenmaier, 2004, Turner et al., 2017). Moreover, in water supply systems, decisions are made ~~taking into consideration~~by considering the hydrological conditions over lead time of several weeks or even months. Forecast products with such lead times, i.e. 'seasonal' forecasts, are typically less skilful compared to the short ~~or medium -~~range forecasts used for flood control or hydropower production applications.

When using seasonal hydrological scenarios or forecasts to assist water system operations, three main approaches are available: worst case scenario, ensemble streamflow prediction (ESP) and dynamical streamflow prediction (DSP). In the worst-case scenario approach, operational decisions are made by simulating their effects against a repeat of the worst hydrological droughts on records. Worst-case forecasts clearly have no particular skill, but their use has the advantage ~~that it provides~~of providing a lower bound of system performance and reflect the risk-adverse attitude of most water ~~resource~~ management practice. This approach is commonly applied by water companies in the UK ~~for reservoir operation~~ and it is ~~recommended by~~reflected in the water resource management planning guidelines of the UK Environment Agency (EA, 2017).

In the ensemble streamflow prediction (ESP) approach, a hydrological forecasts ensemble is produced by forcing a hydrological model using the current initial hydrological conditions and historical weather data over the period of interest (Day, 1985). Operational decisions are then evaluated against ~~such~~the ensemble. The skill of the ESP ensemble is mainly due to the updating of the initial conditions. ~~However, since~~Since ESP ~~is limited to~~forecasts are based on the range of past observations, ~~ESP forecasts~~they can have limited skill under non-stationary climate and where initial conditions do not dominate the seasonal hydrological response (Arnal et al., 2018). Nevertheless, the ESP approach is popular among operational agencies thanks to its simplicity, low cost, efficiency and its intuitively appealing nature (Bazile et al., 2017)~~, i.e.~~. Some previous studies assessed the potential of seasonal ESP ~~is coherent~~ to ~~the human tendency to examine a situation according to past experiences. Seasonal ESP was used to assess possible improvements~~improve the operation of supply-hydropower systems ~~operation, e.g. by~~. For example, Alemu et al. (2010) ~~who~~ reported achieving an average economic benefit of 7% with respect to the benchmark operation policy, ~~and by~~whereas Anghileri et al. (2016) ~~who however did not observe~~reported no significant improvements (possibly because they only used the ESP mean, instead of the full ensemble).

Last, the dynamical streamflow prediction (DSP) approach uses ~~seasonal~~numerical weather forecasts produced by a dynamic climate model to feed the hydrological model (instead of

historical weather data). The output is also an ensemble of hydrological ~~forecast~~forecasts, whose skill comes from both the updated initial condition ~~as well as~~and the predictive ability of the numerical weather forecasts~~,~~. The latter is due to global climate teleconnections such as the El Niño Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO). Therefore, ~~these~~DSP forecasts are generally more skilful in areas where climate teleconnections exert a strong influence, such as tropical areas, and particularly in the first month ahead (Block and Rajagopalan, 2007). In areas where climate teleconnections have a ~~weak~~weaker influence~~, instead~~, DSP can have lower skill than ESP, particularly beyond the first lead month (Arnal et al., 2018, Greuell et al., 2019). Nevertheless, recent advances in the prediction of climate teleconnections in Europe, such as the NAO (Wang et al., 2017, Scaife et al., 2014, Svensson et al., 2015), means that seasonal forecasts skill is likely to continue increasing in ~~next~~the coming years. Post-processing techniques such as bias correction can also potentially improve seasonal streamflow forecast skill (Crochemore et al., 2016). ~~However, studies~~Studies assessing the benefits of bias correction for seasonal hydrological forecasting are still rare in the literature~~, while~~. While bias correction is often recommended or even required for impact assessments to improve forecast skills (Zalachori et al., 2012, Schepen et al., 2014, Ratri et al., 2019, Jabbari and Bae, 2020) studies on long-term hydrological projections (Ehret et al., 2012, Hagemann et al., 2011) highlighted a lack of clarity on whether bias correction should be applied or not. In recent years, meteorological centres such as the European Centre for Medium-Range Weather Forecast (ECMWF) and the UK Met Office, have made important efforts to provide skilful seasonal forecasts, both meteorological (Hemri et al., 2014, MacLachlan et al., 2015) and hydrological (Bell et al., 2017, Arnal et al., 2018) in the UK and Europe, and encouraged their application for water resource management. ~~To our knowledge, however, pilot applications demonstrating the value of such seasonal forecast products to improve operational decisions are still lacking.~~To our knowledge, however, pilot applications demonstrating the value of such seasonal forecast products to improve operational decisions are mainly lacking and have only very recently started to appear (Giuliani et al., 2020).

While the skill of DSP is likely to keep increasing in the next years, ~~this~~it may still ~~not produce considerable improvement in water system operations soon, especially in~~remain low at lead times relevant for the operation of water supply systems ~~where the forecast skill-value relationship is weaker~~. Nevertheless, a number of studies have demonstrated that other factors, which are not necessarily captured by forecast skill scores, may also be important to improve the forecast value ~~of short-term and seasonal forecasts.~~. These include accounting explicitly for the forecast uncertainty in the system operation optimization (Yao and Georgakakos, 2001, Boucher et al., 2012, Fan et al., 2016), using less rigid operation approaches (Yao and Georgakakos, 2001, Brown et al., 2015, Georgakakos and Graham, 2008) and making optimal operational decisions during severe droughts ~~(Turner et al., 2017).~~(Turner et al., 2017, Giuliani et al., 2020). Additionally, the forecast skill itself can be defined in different ways, and it is likely that different characteristics of forecast errors (sign, amount, timing, etc.) affect the forecast value in different ways. Widely used skill scores for hydrological forecast ensembles are the rank histogram (Anderson, 1996), the relative operating characteristic (Mason, 1982) and the ranked probability score (Epstein, 1969). The ranked probability score is widely used by meteorological agencies ~~and it is the recommended score for evaluation of overall performance~~ since it provides a measure of both the bias and the spread of the ensemble into a single factor, while it can also be decomposed into different sub-factors in order to look at the different attributes of the ensemble forecast (Pappenberger et al., 2015, Arnal et al., 2018). However, whether these skill score definitions are relevant for

the specific purpose of water resources management, or whether other definitions would be better proxy of the forecast value, remains an open question.

In this paper, we aim at contributing to the ongoing discussion on the value of seasonal weather forecasts in decision making (Bruno Soares et al., 2018) and at assessing the value of DSP for improving water system operation by application to a real-world reservoir system, and in doing so we build on ~~this~~the growing effort to improve seasonal hydrometeorological forecasting systems and make them suitable for operational use in the UK (Bell et al., 2017, Prudhomme et al., 2017). Through this application we aim to answer the three following questions: 1) can the efficiency of a UK real-world reservoir supply system be improved by using DSP forecasts?, 2) does accounting explicitly for forecast uncertainty improve forecast value (for the same skill)? and 3) what other factors influence the forecast skill-value relationship?

For this purpose, we will simulate ~~and compare the performance of~~ a real-time optimization system informed by seasonal weather forecasts over a historical period for which both observational and forecast datasets are available, and we will ~~benchmark~~compare it to a worst-case scenario approach~~, which is commonly used to inform water supply management in the UK~~ that mimics current system operation. As for the seasonal forecast products, we will assess both ESP and DSP derived from the ECMWF seasonal forecast products (Tim et al., 2018). We will also compare the forecast skill and value before and after applying bias correction ~~to the ECMWF forecast products~~, and for different ~~degrees of forecast uncertainty (i.e. different~~ ensemble sizes~~). To account for decision-making uncertainty,~~. System performances will be measured in terms of water availability and energy costs, and we will ~~also simulate the performance of the system under~~investigate five ~~operating~~different scenarios ~~representing different operational priorities~~for prioritising these two objectives depending on the decision-maker preferences. Finally, we will discuss opportunities and barriers to bring such approach into practice.

Our results are meant to provide water managers with an evaluation of the potential of using seasonal forecasts in ~~extra-tropical areas, such as~~ the UK~~,~~ and to give forecasts providers indications on directions for future developments that may make their products more valuable for water management.

# 2. Methodology

## 2.1. Real-time optimization system

An overview of the real-time optimization system (RTOS) informed by seasonal weather forecasts is given in Figure 1 (left part). It consists of three main stages that are repeated each time an operational decision must be made. These three stages are:

1.a Forecast generation. We use a hydrological model forced by seasonal weather forecasts to generate the seasonal hydrological forecasts. The initial conditions are determined by forcing the same model by (recent) historical weather data for a warm-up period. Another model determines the future water demand during the forecast horizon. Although not tested in this study, in principle such a demand model could also be forced by seasonal weather ~~forecast~~forecasts.

1.b Optimization. This stage uses (i) a reservoir system model to simulate the reservoir storages in response to given inflows and operational decisions, (ii) a set of operation objective functions to evaluate the performance of the system, for instance, to maximize the resource

availability or to minimize the operation costs, and (iii) ~~an~~ a multi-objective optimizer to determine the ~~set of~~ optimal operational decisions ~~that realise~~. When a problem has multiple objectives, optimisation does not provide a single optimal solution (i.e. a single sequence of operational decisions over the forecast horizon) but rather it provides a set of (Pareto) optimal solutions, each realising a different trade-~~off~~off between the ~~objective functions.~~conflicting objectives (for a definition of Pareto optimality see e.g. (Deb et al., 2002)).

1.c Selection of one trade-off solution. In this stage, we represent the performance of the optimal trade-off ~~decisions~~solutions in what we call a "pre-evaluation Pareto front". The terms "pre-evaluation" highlights that these are the anticipated performances according to our ~~models and~~ hydrometeorological forecasts, not the actual performances achieved when the decisions are implemented (which are unknown at this stage). ~~Among this set of optimal decisions~~By inspecting the pre-evaluation Pareto front, the operator will select one Pareto-optimal solution according to their priorities, i.e. the relative importance ~~given~~they give to each operation ~~objectives~~objective. In a simulation experiment, we can mimic the operator choice by setting some rule to choose one point on the Pareto front (and apply it consistently at each decision timestep of the simulation period).

## 2.2. Evaluation

When the RTOS is implemented in practice, the selected operational decision is applied to the real system and the RTOS used again, with updated system conditions, when a new decision needs to be made or new weather forecasts become available. If however we want to evaluate the performance of RTOS in a simulation experiment (for instance to demonstrate the value of using RTOS to reservoir operators) we need to combine it with the evaluation system depicted in the right part of Figure 1. Here, the selected operational decision coming out of the RTOS is applied to the reservoir system model, instead of the real system. The reservoir model is now forced by hydrological inputs observed in the (historical) simulation period, instead of the seasonal forecasts, which enables us to estimate the actual flows and next-step storage that would have occurred if the RTOS was used at the time. This simulated next-step storage can then be used as the initial storage volume for running the RTOS at the following timestep. Once the process has been repeated for the entire period of study, we can provide an overall evaluation of the hydrological forecast skill and the performance of the RTOS, i.e. the forecast value. This evaluation (Figure 1) consists of two stages:

2.a Forecast skill evaluation. ~~In order to evaluate~~The forecast skill is evaluated based on the ~~capacity of the~~ differences between hydrological ~~forecast to predict~~forecasts and observed reservoir inflows over the simulation period. For this purpose, we can calculate the ~~observed inflows we apply forecast skill scores and~~ absolute ~~error indicators. In this paper, we will use~~differences between the observed and the forecasted inflows or we can use forecast skill scores such as the continuous ranked probability skill score (CRPSS~~) and the absolute difference between the observed and forecasted inflows.~~).

2.b Forecast value evaluation. The forecast value is presented as the improvement of the system performance obtained by using the RTOS over the simulation period, with respect to the performance under a simulated benchmark operation. Notice that, because the RTOS deals with ~~a~~ multi~~-objective problem~~ objectives and ~~we have to implement~~hence provides a ~~rule to select one solution out~~set of ~~the pre-evaluation~~ Pareto ~~front~~optimal solutions, in principle we could run a different simulation experiment for each ~~possible definition~~point of the ~~selection rule~~pre-evaluation Pareto front, i.e. for each possible definition of the operational priorities. However, for the sake of simplicity, we ~~only~~will simulate ~~five different~~a smaller number of

relevant and well differentiated operational priorities~~, and thus obtain a~~. The simulated performances of these solutions are visualised in a "post-evaluation" Pareto front ~~with five points.~~ In this Pareto front diagram, the origin of the coordinates represents the performance of the benchmark operation, and the performances of any other solution are rescaled with respect to the benchmark performance. Therefore, a positive value along one axis represents an improvement in that operation objective with respect to the benchmark, whereas a negative value represents a deterioration. When values are positive on both axes, the simulated RTOS solution dominates (in a Pareto sense) the benchmark; the further away from the origin, the more the forecast has proven valuable for decision-making. If instead one value is positive and the other is negative then we would conclude that the forecast value is neither positive ~~nor~~or negative, because the improvement of one objective ~~by the RTOS~~ was achieved at the expenses of the other.

## 2.3. Case study

### 2.3.1. Description of the reservoir system

The reservoir system used in this case study is a two-reservoir system in the South West of the UK (schematised in ~~Figure 2).~~Figure 2). The two reservoirs are moderately sized with storage capacities in the order of 20,000 ~~megalitres (Ml)~~,000 $m^3$ (S1) and 5,000 ~~Ml~~,000 $m^3$ (S2) (the average of UK reservoirs is 1,377 ~~Ml~~,000 $m^3$ (EA, 2017)). ~~The system is partially shared between two different water companies, reservoir S1 being the system element used by both companies.~~ The gravity releases from ~~this~~ reservoir S1 ($u_{S1,R}$) ~~are used by the owner company~~feed into river R, and thus contribute to support downstream abstraction during low ~~river (R)~~ flows~~. The other company can also use pumped~~ periods. Pumped releases from S1 ($u_{S1,D}$) ~~to complement~~and gravity releases from reservoir S2 ($u_{S2,D}$) ~~from their own reservoir (S2) in supplying D in a wider conjunctive use system. Both reservoirs are required to make environmental compensation releases.~~

are used to supply the demand node D. A key operational aspect of the system is the possibility of pumping water back from river R into ~~the shared~~ reservoir S1. Pumped inflows ($u_{R,S1}$) may be operated in the winter months ~~to~~ (from 1st November till 1st April) to supplement natural inflows, provided sufficient ~~water~~discharge is available in the river~~.~~ (R). This facility provides additional drought resilience~~, as it allows~~ by allowing the ~~companies~~operator to increase reservoir storage ~~if natural inflows are insufficient during the~~ in winter ~~months (from 1st November till 1st April) to~~ to help ensure ~~meeting~~that the demand in the following summer ~~demand. The two companies that operate the system liaise regularly, particularly regarding the pumped storage operation, which is constrained by rule curves, and has operated in eleven years since 1995~~can be met. As the pump energy consumption is costly, there is an important trade-off between the operating cost of pump storage and ~~achieving~~ drought resilience.

The pumped storage operation is constrained by a rule curve ~~applied~~, and has operated in ~~the current operation procedures~~eleven years since 1995. The rule curve defines the storage level at which pumps are triggered. Each point on the curve is derived based on the amount of pumping that would be required to ~~refill~~fill the reservoir ~~under the worst historical observed inflows between that point in time and~~ by the end of the pump storage period (1st April~~).~~), under the worst historical inflows scenario. The pumping trigger is therefore risk-averse, which means there is a reasonable ~~change~~chance of pumping too early on during the refill period~~.~~ ~~This increases~~ and increasing the likelihood of reservoir spills if spring rainfall is abundant~~,~~ ~~which means~~ . This may result in unnecessary expenditure on pumping. Informing ~~the~~ pump ~~operations~~operation by using seasonal forecasts of future natural inflows ($I_{S1}$ and $I_{S2}$) may thus

help to reduce the volume of water pumped whilst achieving the same reservoir storage at the end of the refilling period.

### 2.3.2. Forecast generation

In this study we generated dynamical streamflow predictions (DSP) by forcing a lumped hydrological model, the HBV model (Bergström and Singh, 1995), with the seasonal ECMWF SEAS5 weather hindcasts ~~(Tim et al., 2018). The ECMWF SEAS5~~(Tim et al., 2018, Johnson et al., 2019). The ECMWF SEAS5 hindcast dataset consists of an ensemble of 25 members starting on the 1st day of every month and providing daily temperature and precipitation with a lead time of 7 months. The spatial resolution is 36 km which compared to the catchment sizes (28.8 $km^2$ for S1 and 18.2 $km^2$ for S2) makes it necessary to ~~bias correct and~~ downscale the ECMWF hindcasts. Given the lack of clarity in the potential benefits of bias correction (Ehret et al., 2012), we will provide results of using both non-corrected and bias corrected forecasts. ~~.~~The dataset of weather hindcast is available from 1981, whereas reservoir data are available for the period 2005-2016. Hence, we used the period 2005-2016 for the RTOS evaluation and the earlier data from 1981 for bias correction~~.~~ of the meteorological forcing. While limited, this period captures a variety of hydrological conditions, including dry ~~ones~~winters in 2005-06, 2010-2011 and 2011-12, ~~relatively~~which are close to the driest period on records (1975-1976) (see more details in ~~(~~Figure 7~~)~~ of the Supplementary Material). This is important because, under drier conditions, the system performance is more likely to depend on the forecast skill and the benefits of RTOS may become more apparent (Turner et al., 2017). Daily inflows were converted to weekly inflows for consistency with the weekly time step applied in the reservoir system model.

A linear scaling approach (or "monthly mean correction") was applied for bias correction~~.~~ of precipitation and temperature forecasts. This approach is simple and often provides similar results in terms of bias removal as more sophisticated approaches such as the quantile or distribution mapping (Crochemore et al., 2016). A correction factor is calculated as the ratio (for precipitation) or the difference (for temperature) between the average daily observed value and the forecasted value (ensemble mean~~) values of the variable of interest (precipitation or temperature) ~~), for a given month and year. The correction factor is then applied as a multiplicative factor (precipitation) or as an additive factor (temperature) to correct the raw daily ~~forecast values~~forecasts. A different factor is calculated and applied for each month and each year of the evaluation period (2005-2016). For example, for November 2005 we obtain the precipitation correction factor as the ratio between the mean observed rainfall in November from 1981 ~~until~~to 2004 (i.e. the average of 24 values) and the mean forecasted rainfall for ~~the same~~those months (i.e. the average of 24x25 values, as we have 25 ensemble members). For November 2006, we re-calculate the correction factor by also including the observations and forecasts of November 2005, hence taking averages over 25 values; and so forth. The rationale of this approach is to best mimic what would happen in real-time, when the operator would likely access all the available past data and hindcasts for the bias correction.

As anticipated in the Introduction, the ESP is an ensemble of equiprobable weekly streamflow forecasts generated by the hydrological model (HBV in our case) forced by meteorological inputs (precipitation and temperature) observed in the past. ~~In our case and for~~ For consistency with ~~what done for~~ the bias correction ~~of~~ approach used for the ECMWF SEAS5 ~~forecasts~~hindcasts, we ~~use~~produce the ESP using meteorological observations ~~(precipitation and temperature)~~ from 1981 until the year before the simulated decision timestep ~~to produce~~

~~the ESP.~~. This ~~also produces~~leads to producing an ensemble of ~~similar~~increasing size (from 24 to 35 members) ~~with respect~~ but roughly similar to the ECMWF ensemble size (25 members).

### 2.3.3. Optimization: Reservoir system model, ~~operation~~ objective functions and optimiser

The reservoir system dynamics is simulated by a mass balance model implemented in Python. The simulation model is linked to an optimiser to determine the optimal scheduling of pumping ($u_{R,S1}$) and release ($u_{S1,D}$ and $u_{S2,D}$) decisions. ~~As~~For the optimiser we ~~used~~use the NSGA-II multi-objective evolutionary algorithm (Deb et al., 2002) implemented in the open-source ~~python~~Python package Platypus (Hadka, 2018). We set two operation objectives for the optimiser: to minimize the overall pumping energy ~~costs~~cost and to maximize the water resource availability at the end of the pump storage period. The ~~first objective function~~pumping cost is calculated as the sum of the weekly energy costs associated to pumped inflows and pumped releases ($u_{R,S1}$ and $u_{S1,D}$) over the optimisation period. The ~~second function~~ resource availability is the mean storage volume in S1 and in S2 at the end of the optimisation period (1st April). ~~The release of S1 $u_{s1,}$~~When optimisation is run against a forecast ensemble, the two objective functions are evaluated against each ensemble member and the average is taken as final objective function value. The gravity releases from S1 ($u_{S1,R}$ ~~is~~) are not considered ~~a~~as decision ~~variable~~variables and ~~is defined by~~they are set to the observed values during the period of study. This choice is ~~however not likely~~unlikely to have important implications on the optimization results because $u_{S1,R}$ on average ~~only~~ represents ~~the~~only 15% of the total ~~S1~~ releases from S1 ($u_{S1,D} + u_{S1,R}$). Also, we ~~made the simplifying assumption~~assume that ~~the~~ future water ~~demand is~~demands are perfectly known ~~at each time step~~in advance, and ~~thus defined D by~~set them to the sum of the observed releases from S1 ($u_{S1,D}$) and S2 ($u_{S2,D}$) for the period of study~~, instead of using a demand model. The~~. This simplification is reasonable for our case study as the water demand is fairly stable and predictable in winter, and it enables us to focus on the relationship between skill and value of the seasonal hydrological ~~forecast skill and the forecast value while avoiding the influence of non-perfect water demand~~ forecasts~~.~~ while assuming no error in demand forecasts. More details about the reservoir simulation model and the optimisation problem are given in the Supplementary Material.

### 2.3.4. Selection of the trade-off solution

~~In order to take into account the uncertainty in~~We use five different rules for the selection of the trade-off solution~~, we estimate the forecast value under five operating scenarios out of the 20 available in~~ from the pre-evaluation Pareto front (see Figure 1~~). They represent five selection rules based on different operational priorities, according to the relative importance given to each performance objectives: 1) resource availability only (*rao*), 2) resource availability prioritised (*rap*), 3) balanced (*bal*), 4) pumping savings prioritised (*psp*) and 5) pumping savings only (*pso*). The same selection rule is~~), and apply them consistently ~~applied~~ at each decision timestep of the simulation period. ~~The relative importance of the objectives is quantified as the percentile of the performance improvement along the axes of the pre-evaluation Pareto front. For instance, *rao* is~~The five rules correspond to five different scenarios of operational priorities. They are: 1) resource availability only (*rao*), which assumes that the operator consistently selects the extreme solution ~~in the pre-evaluation Pareto front~~ that delivers the largest improvement in resource availability; ~~*rap* is~~2) resource availability prioritised (*rap*) selects the solution delivering the 75% percentile in resource availability increase ~~among the 20 operation scenarios available; *bal* delivers~~; 3) balanced (*bal*) selects the

solution delivering the median improvement~~; etc~~ in resource availability; 4) pumping savings prioritised (*psp*) selects the solution delivering the 75% percentile in energy cost reductions; and 5) pumping savings only (*pso*), which selects the best solution for energy saving.

### 2.3.5. Forecast skill evaluation

We ~~used~~use two metrics ~~to evaluate the forecast skill:~~, a skill score and the mean error~~.~~, to evaluate the quality of the hydrological forecasts over our simulation period (from November to April).

A skill score evaluates the performance of a given forecasting system with respect to the performance of a reference forecasting system. As a measure of performance, we use the continuous ranked probability score (CRPS) (Brown, 1974) (Hersbach, 2000). The CRPS is defined as the distance between the cumulative distribution function of the probabilistic forecast and the empirical distribution of the corresponding observation. At each forecasting step, the CRPS is thus calculated as:

$$CRPS(p(x), I^{Obs}) = \int (p(x) - H(x < I^{Obs}))^2 dx$$

where $p(x)$ represents the distribution of the forecast; $I^{Obs}$ is the observed inflow [$m^3$]; and $H$ is the empirical distribution of the observation, i.e. the step function which equals 0 when $x < I^{Obs}$ and 1 when $x > I^{Obs}$. The lower the CRPS, the better the performance of the forecast. ~~As a measure of performance, we use the continuous ranked probability score (CRPS) (Brown, 1974).~~In this study weekly forecast and observation data were used to compute individual CRPS values. The skill score is then defined as:

$$CRPSS = 1 - \frac{CRPS^{Sys}}{CRPS^{Ref}}$$

When the skill score is higher (lower) than zero, the forecasting system is more (less) skilful than the reference. When it is equal to zero, the system and the reference have equivalent skill. Following the recommendation by Harrigan et al. (2018) we used ensemble streamflow predictions (ESP) as a "tough to beat" reference, which is more likely to demonstrate the "real skill" of the hydrological forecasting system (Pappenberger et al., 2015)~~.~~ based on dynamic weather forecasts.

~~The continuous ranked probability score appearing in the above equation is defined as the distance between the cumulative distribution function of the probabilistic forecast and the empirical distribution of the corresponding observation. At each forecasting step, and for a given lead time, CRPS is thus calculated as:~~

$$~~CRPS(p(x), y) = \int (p(x) - H(x < y))^2 dx~~$$

~~Where $p(x)$ represents the distribution of the forecast; y is the observation; and $H$ is the empirical distribution of the observation, i.e. the step function which equals 0 when $x < y$ and 1 when $x > y$. The lower the CRPS, the better the performance of the forecast. The average CRPS for a given lead time is equal to the mean of the CRPS values across the time frame. In this study weekly forecast and observation data were used to compute CRPS.~~

The mean error measures the difference between the forecasted and the observed inflows~~.~~ (at monthly scale). The mean error is negative when the forecasts tend to underestimate the

observations and positive when the forecasts overestimate the observations. The mean error for a given forecasting step and lead time *T* [~~weeks~~months] is:

$$mean\ error = \frac{1}{M}\sum_{\sout{m=0}}^{\sout{M}}\left(\frac{\sout{1}}{\sout{T}}\sum_{\sout{t=0}}^{\sout{T}}\left(I_{\sout{t,m}}^{\sout{Syst}} - I_{\sout{t}}^{\sout{Obs}}\right)\right)\sum_{m=0}^{M}\left(\frac{1}{T}\sum_{t=0}^{T}\left(I_{t,m}^{Sys} - I_{t}^{Obs}\right)\right)$$

where *I* is the inflow [~~ML~~$m^3$], *t* is the timestep [~~week~~month] and *M* the total number of members (*m*) of the ensemble.

### 2.3.6. Forecast value evaluation and definition of the benchmark operation

To evaluate the ~~forecast~~ value of ~~DSP (before and after bias correction) and ESP~~the hydrological forecasts, we compared the simulated performance of the RTOS ~~(Figure 1)~~ informed by these ~~seasonal weather forecast products~~forecasts with the simulated performance of a benchmark operation. The benchmark mimics common practices in reservoir operation in the UK, whereby operational decisions are made against a worst-case scenario – a repeat of the worst hydrological drought on records~~. We can~~ (1975-76). This comparison enables us to show the potential benefits of using seasonal forecast with respect to the current approach. We simulate the benchmark operation using similar steps as in the RTOS represented in Figure 1, but with three main variations. First, instead of seasonal weather forecasts, we use the historical weather data recorded in Nov 1975-Apr 1976~~.~~ (the worst drought on records). Second, the optimiser ~~only~~ determines the optimal scheduling of reservoir releases ($u_{S1,D}$ and $u_{S2,D}$~~), whereas~~) but not that of pumped inflows ($u_{R,S1}$~~)~~). Instead, these are determined by the rule curve applied in the current operation procedures. Specifically, if at the start of the week the storage level in S1 is below the storage volume defined by the rule curve for that calendar day, the operation triggers the pumping system during that week (we assume that the triggered pumped inflow is equal to the maximum pipe capacity). Third, the optimiser only aims at minimising pumping costs, whereas the resource availability objective is turned into a constraint, i.e. the mean storage volume of the two reservoirs must be maximum by the end of the pump storage period (1st April) and no trading-off with pumping costs reduction is allowed.

# 3. Results

## 3.1 Forecast skills

First, we analyse the skill of DSP hydrological forecasts. Figure 3a shows the average CRPSS at different lead times before (red) and after (blue) bias correction~~.~~ of the meteorological forecasts. We compute the average CRPSS for a given lead time as the average of the CRPSS obtained for each forecast used for the simulation of the reservoir system in that time frame. For instance, the forecast for a 3-month lead time, since the simulation time frame is in this case 1 Jan to 1 Apr, we average the CRPSS values obtained for the 1 Jan-1 Apr, 1 Feb-1 Apr and 1 Mar-1 Apr forecasts.

Before bias correction, the average ~~forecast~~ skill ~~is highest~~score is positive, i.e. the forecast is more skilful than the benchmark (ESP), only at 1~~-~~-month or 2-month lead ~~time and decreases with larger lead time~~times (solid red line). ~~Furthermore, the skill~~CRPSS is higher than average in the three driest winters, i.e. 2005-2006, 2010-2011, 2011-2012 (dashed lines). If we compare DSP to DSP-corr (red and blue solid lines), we see that bias correction deteriorates the average ~~skills~~skill scores for shorter lead times (1 and 2 months) while it improves ~~it~~them

for longer ones (3,4 and 5 months~~).~~) but the value is still negative, i.e. the forecast is less skilful than the benchmark (ESP). In the driest years (dashed lines) bias correction deteriorates the skill score for most lead times.

~~The~~ We compute the average mean error for a given lead time as for CRPSS. The computed mean error values (Figure 3b) ~~indicates~~indicate that DSP systematically underestimates the inflow observations but less so in the three driest winters. After bias correction (DSP-corr), this systematic underestimation turns into a systematic overestimation. Also, the average mean error gets lower for longer lead times, though not as much in the driest years.

In summary, we can conclude that bias correction does not seem to produce ~~a systematic~~an improvement in the forecast skill for our observation period~~, but only some improvement at some lead times~~. On the other hand, what we find in our case study is a clear signal of bias correction turning negative mean errors (inflow underestimation) into positive errors (overestimation). So, while the magnitude of errors stays relatively similar, the sign of those errors changes. We will go back to this point later on, when analysing the skill-value relationship.

## 3.2 Forecast value

The forecast value is presented here as the simulated system performance improvement, i.e. increase in resource availability and in pumping cost savings, with respect to the benchmark operation.

### 3.2.1 Effect of operational priority scenario and forecast product on the forecast value

We start by analysing the average forecast value over the simulation period 2005-2016 (Figure 4) for the three seasonal ~~weather~~ forecast products (DSP, DSP-corr and ESP) and the perfect forecast, under five operational policy scenarios (rao: resource availability only; rap: resource availability prioritised; bal: balanced; psp: pumping savings prioritised; and pso: pumping savings only).

Firstly, we notice in Figure 4 that the monthly pumping energy cost savings vary widely with the operational priority. The range of variation depends on the forecast type, going from £20,000 to £48,000 for the perfect forecast and from -£77,000 to £48,000 for ~~the three seasonal weather forecasts.~~DSP, DSP-corr and ESP. For all forecast products, the improvement in resource availability shows lower variability, with an improvement of less than +2% (of the mean storage volume in S1 and in S2 at the end of the optimisation period) for *rao*, and a deterioration of -2% for *pso*. While this seems to suggest a lower sensitivity of the resource availability objective, variations of few percent points in storage volume may still be important in critically dry years.

As for the forecast value, we find that the perfect forecast brings value (i.e. a simultaneous improvement of both objectives) in the two scenarios that prioritize the increase in resource availability (*rao* and *rap*), DSP brings no value in any scenarios, DSP-corr has positive value in the *rap* and *bal* scenario, and ESP in the *bal* only. In other words, real-time optimisation based on seasonal forecasts can outperform the benchmark operation, but whether this happens depends on both the forecast product being used and the operational priority.

An interesting observation in Figure 4 is that the distance in performance between using perfect forecasts and real forecasts (DSP, DSP-corr, ESP) is very small under scenarios that prioritise energy savings (bottom-right quadrant) and much larger under scenarios prioritising resource availability (top quadrants). This indicates a stronger skill-value relationship under the latter scenarios, i.e. improvements in the forecast skill are more likely to produce improvements in the forecast value if resource availability is the priority.

Last, if we compare DSP with DSP-corr we see that the effect of bias ~~correction~~correcting the meteorological forcing is mainly a systematic shift to the right along the horizontal axis, i.e. an improvement in energy cost savings at almost equivalent resource availability. Thanks to this shift, in the scenario that prioritises resource availability (*rap*), DSP-corr outperforms ESP. In fact, using DSP-corr is win-win with respect to the benchmark (i.e. the *rap* performance falls in the top-right quadrant in Figure 4) while using ESP is not, as it improves the resource availability at the expenses of pumping energy savings (i.e. producing negative savings).

#### 3.2.2 Effect of ~~uncertainty consideration on~~the forecast ensemble size on the forecast value

We now analyse the effect that different characterisations of the forecast uncertainty have on the DPS-corr forecast value. We start by the extreme case when uncertainty is not considered at all in the real-time optimisation, i.e. when we take the mean value of the DSP-corr forecast ensemble and use it to drive a deterministic optimisation. The results are reported in ~~Figure 5~~Figure 5, which shows that the solution space shrinks to the bottom-right quadrant and, no matter the decision maker priority, the deterministic forecast has no value because energy savings are only achieved at the expenses of reducing the resource availability.

We also consider intermediate cases where optimisation explicitly considers the forecast uncertainty~~,~~ (i.e. it is based on the average value of the objective functions across a forecast ensemble) but the size of the ~~forecast~~ ensemble varies between 5 and 25 members (the original ensemble size). For clarity of illustration, we focus on the resource availability prioritised (*rap*) scenario only. We ~~chose~~choose this scenario because it seems to best reflect the current preferences of the system managers, whose priority is to maintain the resource availability while reducing pumping costs as a secondary objective. Moreover, the previous analysis (Figure 4) has shown that the optimised *rap* has a larger window of opportunity for improving performance with respect to the benchmark and could potentially improve both operation objectives if the forecast skill was perfect.

For each chosen ensemble size, we randomly choose 10 replicates of that same size from the original ensemble, then we run a simulation experiment using each of these replicates, and finally average their performance. Results are again shown in ~~Figure 6.~~Figure 5. For a range of 10 to 20 ensemble members, the forecast value remains relatively close to the value obtained by considering the whole ensemble (25 members). However, if only 5 members are considered, the resource availability is definitely lower and cost savings higher, so that the trade-off that is actually achieved is different from the one that was pursued (i.e. to prioritise resource availability). Notice that the extreme case of using 1 member, i.e. the deterministic forecast case (green cross in ~~Figure 6),~~Figure 5), further exacerbates this effect of 'achieving the wrong trade-off' as resource availability is even lower than in the benchmark.

### 3.2.3 Year-by-year analysis of the forecast value

Last, ~~We now study more in detail~~we investigate the ~~year-by-year relationship between~~temporal distribution of the forecast skill and value~~, and between~~ (i.e. increased resource availability and energy cost savings) along the simulation period and compare it to the hydrological conditions ~~and~~ observed in each year (Figure 6). The "hydrological conditions" is the sum of the initial storage value~~.~~ and the total inflows during the optimisation period, hence enabling us to distinguish dry and wet years. Again, for the sake of simplicity we focus on the simulation results in the most relevant priority scenario of resource availability priority (rap). ~~For this scenario, Figure 6 plots the improvement in system performance achieved in every year against different indicators of skill and hydrological conditions (the plots for the other scenarios are reported in the Supplementary Material).~~

~~The two top and bottom panels on the left (a,b, f and g) show that the forecast skill, measured by either the CRPSS or the mean error, is in general weakly correlated to the system performances (Spearman coefficient < 0.5 and p-value > 0.05). Similarly, weak correlation was found in the other priority scenarios (see Supplementary material). The other panels (c-e,h-j) show that the Initial storage (on November, 1st), the Total inflows (from November to the end of April), and their sum (called 'Hydrological conditions') are more strongly correlated to the performance. In particular, the correlation is strongest and with highest confidence (Spearman correlation = -0.60, p-value = 0.05) between the Hydrological conditions and the Increase in resource availability (Figure 6e). The correlation between the Initial storage and the Increase of resource availability (Figure 6c) is lower (Spearman correlation = -0.41, p-value = 0.21), although visually we can observe a threshold effect with a sharp increase of the value in the two years with the lowest initial storage (2011-2012 and 2010-2011). This result may have interesting operational implications, as further discussed in the next Section.~~

~~Last,~~ ~~in Figure 7 we investigate the distribution of benefits (i.e. increased resource availability, top, and energy cost savings, bottom) along the simulation period. We compare three different forecast products, DSP, DSP-corr and ESP, in the *rap* scenario.~~ First, we observe that two specific ~~year~~years play the most important role in improving the system performance with respect to the benchmark: 2010-11 for pumping cost savings ~~(bottom panel~~(Figure 6e) and 2011-12 for resource availability ~~(top).~~(Figure 6d). These years correspond to the driest conditions in the period of study (see ~~inflow and initial condition data in the top panel~~Figure 6a , and the Supplementary Material for further analysis of the inflow data~~).~~) but not to the highest forecast skills either quantified with CRPSS or mean error (Figure 6b and c). In general, the temporal distribution of the average yearly forecast skill does not show any correspondence with the yearly forecast value. When comparing DSP-corr with DSP (blue and grey bars), we observe that they perform similarly in terms of resource availability but DSP-corr performs better for energy savings. This difference was observed already when looking at average performances over the simulation period (Figure 4) and can be related to the change in sign of forecasting errors induced by the bias correction of the meteorological forcing (Figure 3b). In fact, without bias correction, reservoir inflows tend to be underestimated, which leads the RTOS to pump more frequently and often unnecessarily (e.g. in 2005-06, 2006-07, 2007-08, etc.). With bias correction, instead, inflows tend to be overestimated, and the RTOS uses pumping less frequently. Interestingly, the reduction in pumping still does not prevent to improve the resource availability with respect to the benchmark. This is achieved by the RTOS through a better allocation of pump and release volumes over the optimisation period. When comparing DSP-corr with ESP, we find that the largest improvements ~~with~~

~~respect to the benchmark~~ are gained in the same years ~~for~~by both products, i.e. in the driest ~~years~~ones. As already emerged from the analysis of average performances (Figure 4), we see that ESP achieves slightly better resource availability than DSP-corr but with less pumping cost savings. ESP in particular seems to produce 'unnecessary' pumping costs in 2006-07, 2011-12 and 2013-14, where DSP-corr achieves a similar resource availability ~~(top panel~~(Figure 6d) at almost no cost ~~(bottom).~~(Figure 6e). It must be noted that for the ESP approach, these three specific years, 2006-07, 2011-12 and 2013-14. play the most important role in decreasing the pumping energy cost savings with respect to the benchmark~~, 2006-07, 2011-12 and 2013-14 (Figure 7b), which together with 2010-11 have the lowest initial storage (Figure 7a).~~.

# 4. Discussion

Our study provides some insights on the complex relationship between forecast skill and its value for decision-making. Although these findings may be dependent on the case study and time period that was available for the analysis, they still enable us to draw some more general lessons that could be useful also beyond the specific case investigated here.

First, we found that the use of bias correction, and in particular linear scaling of the meteorological forcing, to improve the skill and value of DSP forecast is less straightforward than possibly expected. Our results show that on average bias correction ~~slightly improves~~does not improve the DSP forecast skill (as measured by CRPSS and mean error) ~~but it~~and can ~~reduce~~even deteriorate it in dry years (Figure 3). This is because in our system DSP forecasts systematically underestimate inflows (before bias correction), which means their skill is relatively higher in exceptionally dry years and is deteriorated by bias correction. To our knowledge, no previous study reported such difference in skill for the ECMWF SEAS5 forecasts in dry years in the UK, hence we are not able to say whether our result applies to other systems in the region. However, the result points at a possible intrinsic contradiction in the very idea of bias correcting based on climatology~~-based forecast (e.g. ESP). In fact~~. In this study, the main reason for the bias correction to fail in improving the forecast skills is that the DSP forecast before bias correction was already performing relatively well in terms of skills in the three particularly dry winters (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 of these three exceptionally dry winters, but we improved the skills in the rest. In this case the bias correction would have performed better if these three dry years were not considered, i.e. under less exceptional climate conditions the bias correction would have been more effective. More generally, by pushing forecasts to be more alike climatology, one may reduce the 'good signal' that may be present in the original forecast in years that will indeed be significantly drier (or wetter) than climatology. As exceptional conditions are likely the ones when water managers can extract more value from forecasts, the argument that bias correction ensures average performance at least equivalent to climatology or ESP (e.g. Crochemore et al. (2016)) may not be very relevant here. We would 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.

While we could not find an obvious and significant improvement of forecast skill after bias correction, we found a clear increase in forecast value (Figure 4). ~~In fact,~~ RTOS based on bias-corrected ~~DPS~~DSP considerably ~~reduce~~reduces pumping costs with respect to the original ~~DPS~~DSP, while ensuring similar resource availability. A consequence of this is that decision maker priorities *rap* (resource availability prioritised) and *bal* (balanced) dominate (in a Pareto sense) the benchmark. We explained this ~~finding~~reduction in pumping costs by the change in

the sign of forecasting errors induced by bias correction – from a systematic underestimation of inflows to a systematic overestimation. While this change is again case specific, a general implication is that not all forecast errors have the same impact on the forecast value~~. From a water resource management perspective, the improvement of forecast accuracy in some directions can be more 'valuable' than others. This also implies that~~, and thus not all skill scores may be equally useful and relevant for water resource managers. For example, in our case a score that is able to differentiate between overestimation and underestimation ~~error~~errors, such as the mean error, seems more adequate than a score such as CRPSS, which is insensitive to the error sign. This said, our results overall suggest that inferring the forecast value from its skill may be misleading, given the weak ~~correlation~~relationship 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.

While we found a weak ~~correlation~~relationship between forecast skill and value, we found that forecast value is more strongly ~~liked~~linked to hydrological conditions (~~Figure 6).~~Figure 6). As expected, a forecast-based RTOS system is particularly useful in dry years, where we find most of the gains with respect to the benchmark operation (~~Figure 7).~~Figure 6). This is consistent with previous studies for water supply system, e.g. Turner et al., 2017. ~~An interesting finding in~~ In our ~~system is that the value of forecast-based~~case study, RTOS ~~seems correlated to the Initial conditions (total storage value) of the system. Given that this initial condition is known at the beginning of the pumped-storage season, in practice this indicator could be used to decide whether to use the forecast-based RTOS approach in the coming months or~~ not~~. In fact, using the RTOS has a cost in that downloading seasonal weather forecasts, transforming them into hydrological forecasts and bias correcting, running optimisation, etc. takes time. So, water managers may choose to use RTOS~~ only ~~in those~~improve resource availability but also reduce pumping costs because, in the dryer years~~ where they expect it will lead to considerable improvements of system performance~~, storage levels are more likely to cross the rule curve and trigger pumping in the benchmark operation.

~~Similarly, in~~In light of the pre-processing costs of seasonal weather forecasts, it is interesting to discuss whether their use is justified with respect to a possibly simpler-to-use product such as ESP. While weather forecast centres are increasingly reducing the pre-processing costs by facilitating access to their seasonal weather forecast datasets, bias correction still needs a considerable level of expertise. This is not only because the necessary tools are currently not provided but also because we first need the knowledge to decide whether applying bias correction is appropriate for the specific case study. Further, once deciding that it is appropriate, we then need to select and understand the most adequate bias-correction method. In this study, we found ESP to be a 'hard-to-beat' reference not only in terms of ~~skills~~skill (as previously found by others, e.g. (Harrigan et al., 2018)) but also in terms of forecast value (Figure 4). In fact, the use of DSP-corr delivers higher energy savings with respect to ESP (without compromising the resource availability) at least in the most relevant operating priority scenario (the *rap* scenario, see ~~Figure 7).~~Figure 6). However, whether these cost-savings are large enough to justify the use of DSP-corr, or whether water managers may fall back on using simpler ESP, it is difficult to argue and remain an open question with the simulations results available so far.

One point where our results instead point to a univocal and clear conclusion is in the importance of explicitly considering forecast uncertainty (~~Figure 5).~~Figure 5). In fact, RTOS

outperforms the current operation when using ensemble forecasts, but it does not if uncertainty is removed and the ~~ensemble mean is used within a deterministic optimisation approach.~~system is optimised against the ensemble mean. In this case, in fact, DSP-corr improves energy savings but it decreases the resource availability under all operational priority scenario. This is in line with previous results obtained using short-term forecasts for flood control (Ficchì et al., 2016), who found that consideration of forecast uncertainty could largely compensate the loss in value caused by forecast errors~~),,~~ hydropower generation (Boucher et al., 2012) and multi-purpose systems (Yao and Georgakakos, 2001). It is also consistent with previous results by Anghileri et al. (2016), who did not find significant value in seasonal forecasts while using a deterministic optimisation approach (they did not explore the use of ensemble though).

Finally, we tried to investigate whether we could evaluate the effect of the ensemble size on the value of the uncertain forecasts. We found that in our case study we could reduce the number of forecast members down to about 10 (from the original size of 25) with limited impact on the forecast value (Figure 5). This is important for practice because by reducing the number of forecast members one can reduce the computation time of the RTOS. While we cannot say if such 'optimal' ensemble size would apply to other systems too, we would suggest that future studies could look at how the quality of the uncertainty characterisation impacts on the forecast value, and whether a 'minimum representation of uncertainty' exists that ensures the most effective use of forecasts for water resource management.

From the UK water industry perspective, we hope our results will motivate a move away from the deterministic (worst-case scenario) approach that often prevails when using models to support short-term decisions, and a shift towards more explicit consideration of model uncertainties. Such a move would also align with the advocated use of "risk-based" approaches for long-term planning (Hall et al., 2012, Turner et al., 2016, UKWIR, 2016a, UKWIR, 2016b), which have indeed been adopted by water companies in the preparation of their Water Resource Management Plans (SouthernWater, 2018, UnitedUtilities, 2019). The results presented here, and in the above cited studies, suggest that greater consideration of uncertainty and trade-offs would also be beneficial in short-term production planning. ~~Last, we tried to investigate whether we could evaluate the effect of the ensemble size on the value of the uncertain forecasts. We found that in our case study we could reduce the number of forecast members down to about 10 (from the original size of 25) with limited impact on the forecast value (Figure 5). This is important for practice because by reducing the number of forecast members one can reduce the computation time of the RTOS. While we cannot say if such 'optimal' ensemble size would apply to other systems too, we would suggest that future studies could look at how the quality of the uncertainty characterisation impacts on the forecast value, and whether a 'minimum representation of uncertainty' exists that ensures the most effective use of forecasts for water resource management.~~

### 4.1 Limitations and perspective for future research and implementation

Our study is subject to a range of limitations that should be kept in mind when evaluating our results. First, the current (and future) skill of seasonal meteorological forecasts varies spatially across the UK depending on the influence of climate teleconnections and particularly the NAO. Given that our case study is located in the ~~West~~South-west of the UK, where the NAO influence has been found to be stronger than in the East (Svensson et al., 2015), our simulated benefits of using DSP seasonal forecasts may be particularly optimistic. Second, the general validity of the results is limited by the relatively short period (2005-2016) that was available for

historical simulations, and which may be insufficient to fully characterise the variability of hydrological conditions and hence accurately estimate the system's performances (see for example discussion in Dobson et al. (2019)). Hence, we aim at continuing the evaluation of the RTOS over time as new seasonal forecasts and observations become available. Another limitation ~~of evaluation of the RTOS~~ is that we used the observed water demand, hence implicitly assuming that operators know in advance the demand values for the entire season with full certainty.

Future studies should extend the testing of the RTOS over a longer time horizon and evaluate the influence of errors in forecasting water demand. To improve our understanding of the forecast skill-value relationship and the benefits of bias correction it would also be interesting to test the sensitivity of our results to the use of different skill scores and bias correction methods. The results of this study and in particular the higher DSP forecast skills than ESP for 1 or 2-month lead times, suggest that combining DSP for the first two months and ESP for the rest of the forecast horizon may be way worthwhile to explore in the future studies.

The Python code developed to: generate the seasonal inflow forecasts~~,~~ from weather forecasts; to optimise the system operation and; to visualise the pre-evaluation Pareto front (with its uncertainty), has been implemented in a set of interactive Jupyter Notebooks, which we have now transferred to the water company in charge of the pumped-storage decisions. The general code and Jupyter Notebooks for application of our methodology to other reservoir systems are available as part of the open-source toolkit iRONS (https://github.com/AndresPenuela/iRONS). This toolkit aims at addressing some of the problems identified in the literature for the implementation of forecast informed reservoir operation systems, by providing better "packaging" (Goulter, 1992) of model results and their uncertainties, enabling the interactive involvement of decision makers (Goulter, 1992) and creating a standard and formal methodology (Labadie, 2004) to support model-informed decisions. Besides supporting the specific decision-making problem faced by the water company involved in this study, through this collaboration we aim at evaluating more broadly ~~the effectiveness of~~how effective our toolkit is to promote knowledge transfer from the research to the professional community and how easily the toolkit can be adapted for different purposes. Through the use of the toolkit, we also hope to gain a better understanding of how decision-makers view forecast uncertainty, of the institutional constraints limiting the use and implementation of this information (Rayner et al., 2005) and of the most effective ways in which forecast uncertainty and simulated system robustness can be represented.

# 5. Conclusions

This work assessed the potential of using a real-time optimization system informed by seasonal forecasts to improve reservoir operation in a UK water supply system. While the specific results are only valid for the studied system, they enable us to draw some more general conclusions. First, we found that the use of seasonal forecasts can improve the efficiency of reservoir operation, but only if the forecast uncertainty is explicitly considered ~~(e.g. via ensemble forecast).~~ Uncertainty is characterised here by a forecast ensemble, and we found that the performance improvement is maintained also when the forecast ensemble size is reduced up to a certain limit. Second, while dynamical streamflow predictions (DSP) generated by numerical weather predictions provided the highest value in our case study (under a scenario that prioritise water availability over pumping costs), still ensemble streamflow predictions (ESP), which are more easily derived from observed meteorological conditions in previous years, remain a hard-to-beat reference in terms of both skill and value.

Third, the relationship between the forecast skill and its value for decision-making is complex and strongly affected by the decision maker priorities and the hydrological conditions in each specific year. It must be noted that in practice the decision-making priorities are not solely related to the selection of a specific Pareto-optimal solution, but ~~also the methodology~~ in the first place by the methodology, i.e. the "risk" taken in using something other than the worst-case scenario approach and in applying bias correction of the meteorological forcing or not. We also hope that ~~this~~the study will ~~contribute to show that seasonal forecasts can deliver benefits to inform operational decisions even if their skill is low; and~~ stimulate further research towards better understanding the skill-value relationship, and in finding ways to extract value from forecasts in support of water ~~resource~~resources management.

*Data availability*. The reservoir system data used are property of Wessex Water and as such cannot be shared by the authors. ECMWF data are available under a range of licences~~. For more information please visit http://www.ecmwf.int.~~, for more information please visit http://www.ecmwf.int. A generic version of the code used for implementing the RTOS methodology is available at https://github.com/AndresPenuela/iRONS.

*Author contributions*. AP developed the model code and performed the simulations under the supervision of FP. CH helped to frame the case study and in the interpretation of the results. All the authors contributed to the writing of the manuscript.

*Competing interests.* We declare that there are no competing interests.

*Acknowledgments*. This work is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant EP/R007330/1. The authors are also very grateful to Wessex Water for the data provided. The authors wish to thank the Copernicus Climate Change and Atmosphere Monitoring Services for providing the seasonal forecasts generated by the ECMWF seasonal forecasting systems (SEAS5). Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains

# References

ALEMU, E. T., PALMER, R. N., POLEBITSKI, A. & MEAKER, B. 2010. Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *Journal of Water Resources Planning and Management,* 137**,** 72-82.

ANDERSON, J. L. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of climate,* 9**,** 1518-1530.

ANGHILERI, D., VOISIN, N., CASTELLETTI, A., PIANOSI, F., NIJSSEN, B. & LETTENMAIER, D. P. 2016. Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research,* 52**,** 4209-4225.

ARNAL, L., CLOKE, H. L., STEPHENS, E., WETTERHALL, F., PRUDHOMME, C., NEUMANN, J., KRZEMINSKI, B. & PAPPENBERGER, F. 2018. Skilful seasonal forecasts of streamflow over Europe? *Hydrology and Earth System Sciences,* 22**,** 2057.

BAZILE, R., BOUCHER, M. A., PERREAULT, L. & LECONTE, R. 2017. Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate. *Hydrol. Earth Syst. Sci.,* 21**,** 5747-5762.

BELL, V. A., DAVIES, H. N., KAY, A. L., BROOKSHAW, A. & SCAIFE, A. A. 2017. A national-scale seasonal hydrological forecast system: development and evaluation over Britain. *Hydrology and Earth System Sciences,* 21**,** 4681.

BERGSTRÖM, S. & SINGH, V. 1995. The HBV model. *Computer models of watershed hydrology.***,** 443-476.

BLOCK, P. & RAJAGOPALAN, B. 2007. Interannual Variability and Ensemble Forecast of Upper Blue Nile Basin Kiremt Season Precipitation. *Journal of Hydrometeorology,* 8**,** 327-343.

BOUCHER, M. A., TREMBLAY, D., DELORME, L., PERREAULT, L. & ANCTIL, F. 2012. Hydro-economic assessment of hydrological forecasting systems. *Journal of Hydrology,* 416-417**,** 133-144.

BROWN, C. M., LUND, J. R., CAI, X., REED, P. M., ZAGONA, E. A., OSTFELD, A., HALL, J., CHARACKLIS, G. W., YU, W. & BREKKE, L. 2015. The future of water resources systems analysis: Toward a scientific framework for sustainable water management. *Water Resources Research,* 51**,** 6110-6124.

BROWN, T. A. 1974. Admissible scoring systems for continuous distributions (Report P-5235).

BRUNO SOARES, M., DALY, M. & DESSAI, S. 2018. Assessing the value of seasonal climate forecasts for decision-making. 9**,** e523.

CROCHEMORE, L., RAMOS, M. H. & PAPPENBERGER, F. 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.,* 20**,** 3601-3618.

DAY, G. N. 1985. Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management,* 111**,** 157-170.

DEB, K., PRATAP, A., AGARWAL, S. & MEYARIVAN, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation,* 6**,** 182-197.

DOBSON, B., WAGENER, T. & PIANOSI, F. 2019. How Important Are Model Structural and Contextual Uncertainties when Estimating the Optimized Performance of Water Resource Systems? 55**,** 2170-2193.

EA 2017. Water resources management planning guideline: Interim update.

EHRET, U., ZEHE, E., WULFMEYER, V., WARRACH-SAGI, K. & LIEBERT, J. 2012. HESS Opinions "Should we apply bias correction to global and regional climate model data?". *Hydrol. Earth Syst. Sci.,* 16**,** 3391-3404.

EPSTEIN, E. S. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology,* 8**,** 985-987.

FABER, B. A. & STEDINGER, J. R. 2001. Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. *Journal of Hydrology,* 249**,** 113-133.

FAN, F. M., SCHWANENBERG, D., ALVARADO, R., ASSIS DOS REIS, A., COLLISCHONN, W. & NAUMMAN, S. 2016. Performance of Deterministic and Probabilistic Hydrological Forecasts for the Short-Term Optimization of a Tropical Hydropower Reservoir. *Water Resources Management,* 30**,** 3609-3625.

FICCHÌ, A., RASO, L., DORCHIES, D., PIANOSI, F., MALATERRE, P.-O., OVERLOOP, P.-J. V. & JAY-ALLEMAND, M. 2016. Optimal Operation of the Multireservoir System in the Seine River Basin Using Deterministic and Ensemble Forecasts. *Journal of Water Resources Planning and Management,* 142**,** 05015005.

GEORGAKAKOS, K. P. & GRAHAM, N. E. 2008. Potential Benefits of Seasonal Inflow Prediction Uncertainty for Reservoir Release Decisions. *Journal of Applied Meteorology and Climatology,* 47**,** 1297-1321.

GIULIANI, M., CROCHEMORE, L., PECHLIVANIDIS, I. & CASTELLETTI, A. 2020. From skill to value: isolating the influence of end-user behaviour on seasonal forecast assessment. *Hydrol. Earth Syst. Sci. Discuss.,* 2020**,** 1-20.

GLEICK, P. H. 2003. Global Freshwater Resources: Soft-Path Solutions for the 21st Century. *Science,* 302**,** 1524-1528.

GLEICK, P. H., COOLEY, H., FAMIGLIETTI, J. S., LETTENMAIER, D. P., OKI, T., VÖRÖSMARTY, C. J. & WOOD, E. F. 2013. Improving understanding of the global hydrologic cycle. *Climate science for serving society.* Springer.

GOULTER, I. C. 1992. Systems Analysis in Water‐Distribution Network Design: From Theory to Practice. *Journal of Water Resources Planning and Management,* 118**,** 238-248.

GREUELL, W., FRANSSEN, W. H. P. & HUTJES, R. W. A. 2019. Seasonal streamflow forecasts for Europe – Part 2: Sources of skill. *Hydrol. Earth Syst. Sci.,* 23**,** 371-391.

HADKA, D. 2018. *A Free and Open Source Python Library for Multiobjective Optimization* [Online]. Available: https://github.com/Project-Platypus/Platypus [Accessed 06/11/2019].

HAGEMANN, S., CHEN, C., HAERTER, J. O., HEINKE, J., GERTEN, D. & PIANI, C. 2011. Impact of a Statistical Bias Correction on the Projected Hydrological Changes Obtained from Three GCMs and Two Hydrology Models. 12**,** 556-578.

HALL, J. W., WATTS, G., KEIL, M., DE VIAL, L., STREET, R., CONLAN, K., O'CONNELL, P. E., BEVEN, K. J. & KILSBY, C. G. 2012. Towards risk-based water resources planning in England and Wales under a changing climate. 26**,** 118-129.

HARRIGAN, S., PRUDHOMME, C., PARRY, S., SMITH, K. & TANGUY, M. 2018. Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences,* 22**,** 2023.

HEMRI, S., SCHEUERER, M., PAPPENBERGER, F., BOGNER, K. & HAIDEN, T. 2014. Trends in the predictive performance of raw ensemble weather forecasts. 41**,** 9197-9205.

HERSBACH, H. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting,* 15**,** 559-570.

HOUGH, M. N. & JONES, R. J. A. 1997. The United Kingdom Meteorological Office rainfall and evaporation calculation system: MORECS version 2.0-an overview. *Hydrol. Earth Syst. Sci.,* 1**,** 227-239.

JABBARI, A. & BAE, D.-H. 2020. Improving Ensemble Forecasting Using Total Least Squares and Lead-Time Dependent Bias Correction. 11**,** 300.

JOHNSON, S. J., STOCKDALE, T. N., FERRANTI, L., BALMASEDA, M. A., MOLTENI, F., MAGNUSSON, L., TIETSCHE, S., DECREMER, D., WEISHEIMER, A., BALSAMO, G., KEELEY, S. P. E., MOGENSEN, K., ZUO, H. & MONGE-SANZ, B. M. 2019. SEAS5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.,* 12**,** 1087-1117.

LABADIE, J. W. 2004. Optimal Operation of Multireservoir Systems: State-of-the-Art Review. *Journal of Water Resources Planning and Management,* 130**,** 93-111.

MACLACHLAN, C., ARRIBAS, A., PETERSON, K., MAIDENS, A., FEREDAY, D., SCAIFE, A., GORDON, M., VELLINGA, M., WILLIAMS, A. & COMER, R. 2015. Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society,* 141**,** 1072-1084.

MASON, I. 1982. A model for assessment of weather forecasts. *Aust. Meteor. Mag,* 30**,** 291-303.

MAURER, E. P. & LETTENMAIER, D. P. 2004. Potential Effects of Long-Lead Hydrologic Predictability on Missouri River Main-Stem Reservoirs. *Journal of Climate,* 17**,** 174-186.

PAPPENBERGER, F., RAMOS, M. H., CLOKE, H. L., WETTERHALL, F., ALFIERI, L., BOGNER, K., MUELLER, A. & SALAMON, P. 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology,* 522**,** 697-713.

PRUDHOMME, C., HANNAFORD, J., HARRIGAN, S., BOORMAN, D., KNIGHT, J., BELL, V., JACKSON, C., SVENSSON, C., PARRY, S., BACHILLER-JARENO, N., DAVIES, H., DAVIS, R., MACKAY, J., MCKENZIE, A., RUDD, A., SMITH, K., BLOOMFIELD, J., WARD, R. & JENKINS, A. 2017. Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales. *Hydrological Sciences Journal,* 62**,** 2753-2768.

RATRI, D. N., WHAN, K. & SCHMEITS, M. 2019. A Comparative Verification of Raw and Bias-Corrected ECMWF Seasonal Ensemble Precipitation Reforecasts in Java (Indonesia). *Journal of Applied Meteorology and Climatology,* 58**,** 1709-1723.

RAYNER, S., LACH, D. & INGRAM, H. 2005. Weather forecasts are for wimps: why water resource managers do not use climate forecasts. *Climatic Change,* 69**,** 197-227.

ROBINSON, E., BLYTH, E., CLARK, D., COMYN-PLATT, E., FINCH, J. & RUDD, A. 2016. Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2015)[CHESS-PE].

ROBINSON, E., BLYTH, E., CLARK, D., COMYN-PLATT, E., FINCH, J. & RUDD, A. 2017. Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2015)[CHESS-met] v1. 2.

SCAIFE, A. A., ARRIBAS, A., BLOCKLEY, E., BROOKSHAW, A., CLARK, R. T., DUNSTONE, N., EADE, R., FEREDAY, D., FOLLAND, C. K., GORDON, M., HERMANSON, L., KNIGHT, J. R., LEA, D. J., MACLACHLAN, C., MAIDENS, A., MARTIN, M., PETERSON, A. K., SMITH, D., VELLINGA, M., WALLACE, E., WATERS, J. & WILLIAMS, A. 2014. Skillful long-range prediction of European and North American winters. *Geophysical Research Letters,* 41**,** 2514-2519.

SCHEPEN, A., WANG, Q. J. & ROBERTSON, D. E. 2014. Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs. *Monthly Weather Review,* 142**,** 1758-1770.

SOUTHERNWATER. 2018. *Revised draft Water Resources Management Plan 2019 Statement of Response* [Online]. Available: https://www.southernwater.co.uk/media/1884/statement-of-response-report.pdf [Accessed 6/11/19 2019].

STOCKER, T. F., QIN, D., PLATTNER, G.-K., TIGNOR, M. M., ALLEN, S. K., BOSCHUNG, J., NAUELS, A., XIA, Y., BEX, V. & MIDGLEY, P. M. 2014. Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of IPCC the intergovernmental panel on climate change. Cambridge University Press.

SVENSSON, C., BROOKSHAW, A., SCAIFE, A. A., BELL, V. A., MACKAY, J. D., JACKSON, C. R., HANNAFORD, J., DAVIES, H. N., ARRIBAS, A. & STANLEY, S. 2015. Long-range forecasts of UK winter hydrology. *Environmental Research Letters,* 10**,** 064006.

TANGUY, M., DIXON, H., PROSDOCIMI, I., MORRIS, D. G. & KELLER, V. D. J. 2014. Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2012) [CEH-GEAR]. NERC Environmental Information Data Centre.

TIM, S., MAGDALENA, A.-B., STEPHANIE, J., LAURA, F., FRANCO, M., MAGNUSSON, L., STEFFEN, T., FRÉDÉRIC, V., DAMIEN, D., ANTJE, W., CHRISTOPHER, D. R., GIANPAOLO, B., SARAH, K., KRISTIAN, M., HAO, Z., MICHAEL, M. & MONGE-SANZ, B. M. 2018. SEAS5 and the future evolution of the long-range forecast system. ECMWF.

TURNER, S. W. D., BENNETT, J. C., ROBERTSON, D. E. & GALELLI, S. 2017. Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrol. Earth Syst. Sci.,* 21**,** 4841-4859.

TURNER, S. W. D., BLACKWELL, R. J., SMITH, M. A. & JEFFREY, P. J. 2016. Risk-based water resources planning in England and Wales: challenges in execution and implementation. *Urban Water Journal,* 13**,** 182-197.

UKWIR 2016a. WRMP 2019 Methods - Decision making process: Guidance.

UKWIR 2016b. WRP19 Methods – Risk-Based Planning.

UNITEDUTILITIES. 2019. *Final water resources management plan 2019* [Online]. Available: https://www.unitedutilities.com/globalassets/z_corporate-site/about-us-pdfs/wrmp-2019---2045/final-water-resources-management-plan-2019.pdf [Accessed 6/11/19 2019].

VOISIN, N., PAPPENBERGER, F., LETTENMAIER, D. P., BUIZZA, R. & SCHAAKE, J. C. 2011. Application of a Medium-Range Global Hydrologic Probabilistic Forecast Scheme to the Ohio River Basin. *Weather and Forecasting,* 26**,** 425-446.

WANG, F., WANG, L., ZHOU, H., SAAVEDRA VALERIANO, O. C., KOIKE, T. & LI, W. 2012. Ensemble hydrological prediction-based real-time optimization of a multiobjective reservoir during flood season in a semiarid basin with global numerical weather predictions. *Water Resources Research,* 48.

WANG, L., TING, M. & KUSHNER, P. J. 2017. A robust empirical seasonal prediction of winter NAO and surface climate. *Scientific Reports,* 7**,** 279.

YAO, H. & GEORGAKAKOS, A. 2001. Assessment of Folsom Lake response to historical and potential future climate scenarios: 2. Reservoir management. *Journal of Hydrology,* 249**,** 176-196.

ZALACHORI, I., RAMOS, M. H., GARÇON, R., MATHEVET, T. & GAILHARD, J. 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Adv. Sci. Res.,* 8**,** 135-141.

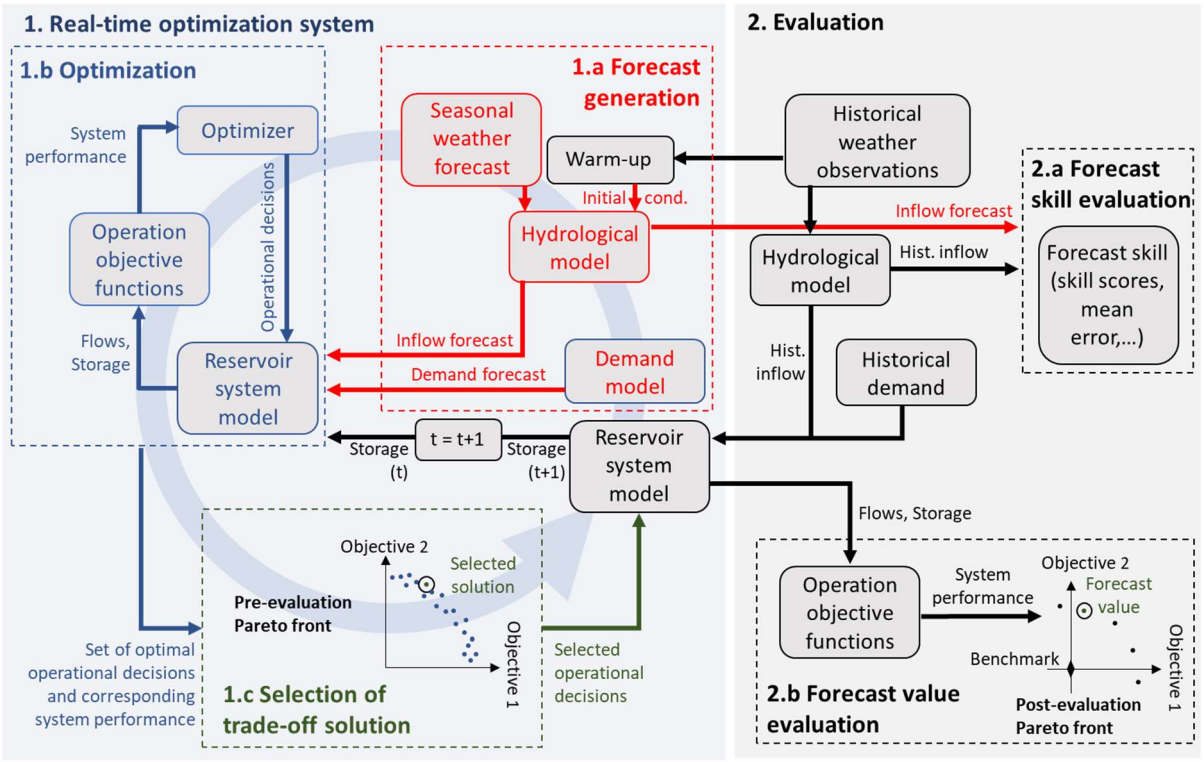



**Figure 1 Diagram of the methodology used in this study to generate operational decisions using a Real-time optimisation system (RTOS) (left) and to evaluate its performances (right). In the evaluation step, the RTOS is nested into a closed loop simulation where at every time step historical data (weather, inflows and demand), along with the operational decisions suggested by the RTOS, are used to move to the next step by updating the initial hydrological conditions and reservoir storage.**

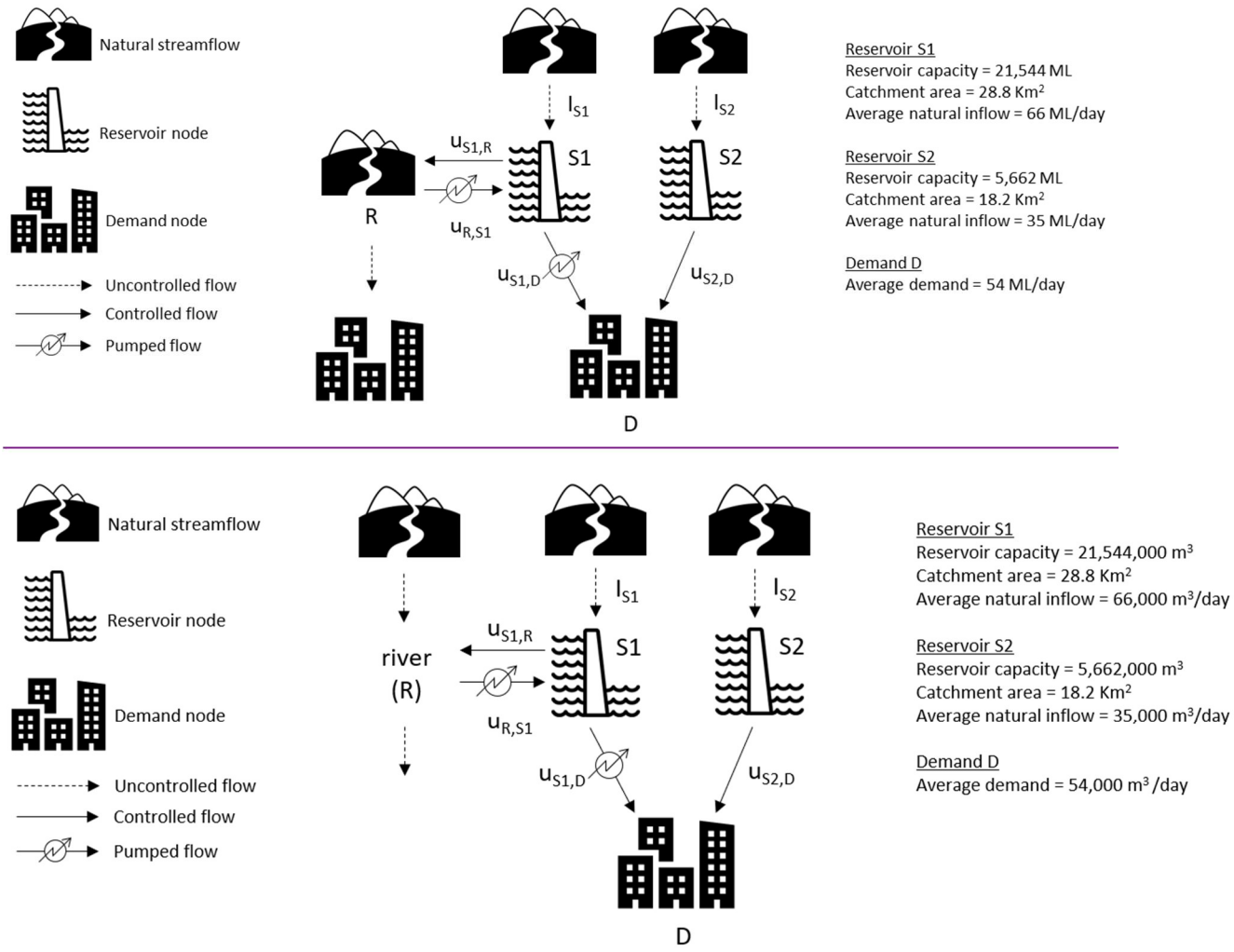
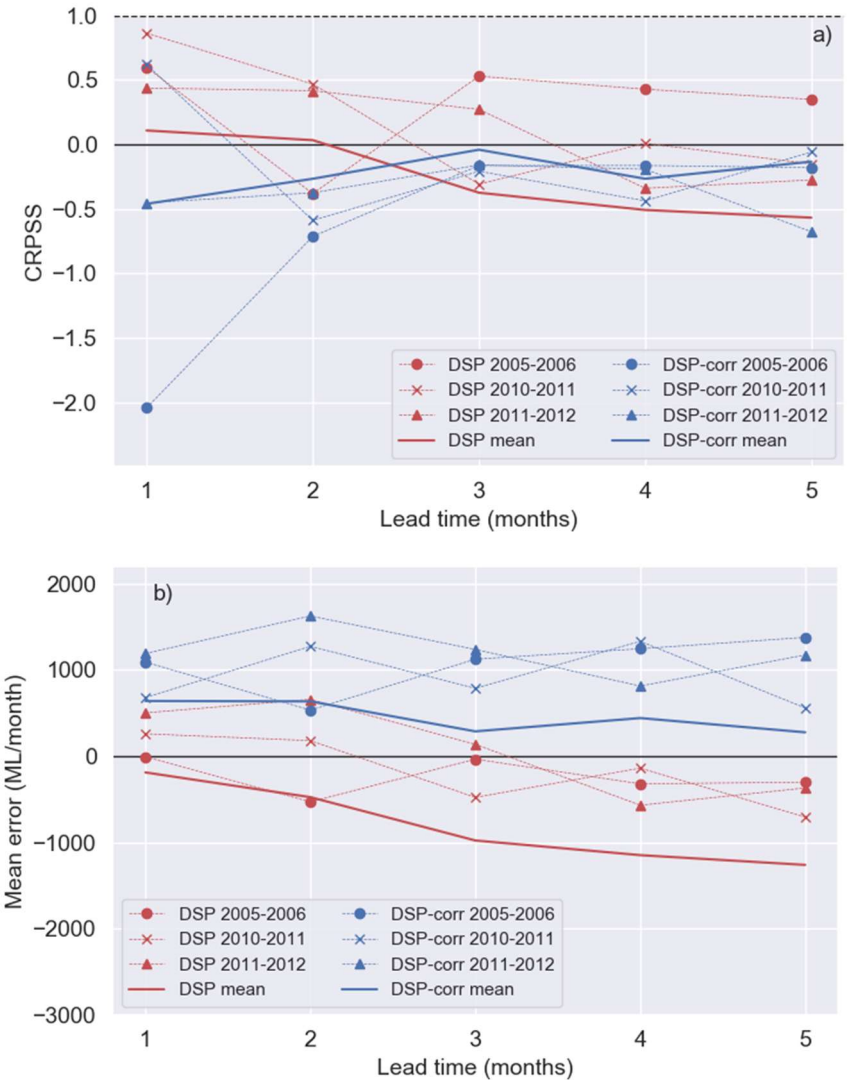
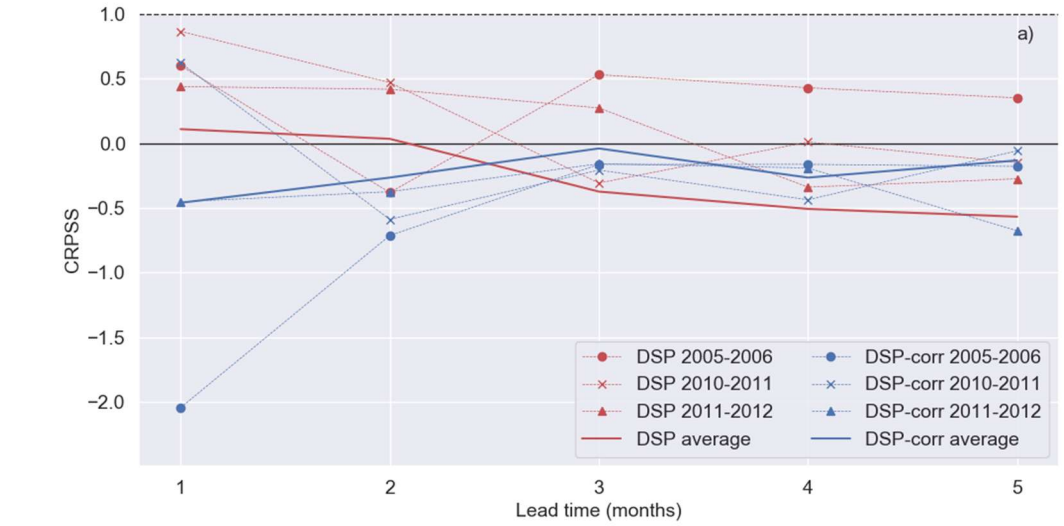



**Figure 2 A schematic of the reservoir system investigated in this study to test the Real-time optimization systems.** ~~**I is**~~ **Reservoir inflows from natural catchments are denoted by I, S1 and S2 are the two reservoir** ~~**inflow, S reservoir node**~~**nodes, u denote controlled inflows/releases, R** ~~**river**~~ **is the river from/to which reservoir S1 can abstract and release, and D** ~~**human**~~**is a demand node.** ~~**The system is a two-reservoir system where S1 both supports downstream abstraction during low river (R) flows and use pumped releases to complement gravity releases from S2 in supplying D. The system has the possibility of pumping water into S1 from Nov to Apr.**~~


a)
CRPSS
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
DSP 2005-2006
DSP 2010-2011
DSP 2011-2012
DSP mean
DSP-corr 2005-2006
DSP-corr 2010-2011
DSP-corr 2011-2012
DSP-corr mean
1
2
3
4
5
Lead time (months)
b)
Mean error (ML/month)
2000
1000
0
−1000
−2000
−3000
DSP 2005-2006
DSP 2010-2011
DSP 2011-2012
DSP mean
DSP-corr 2005-2006
DSP-corr 2010-2011
DSP-corr 2011-2012
DSP-corr mean
1
2
3
4
5
Lead time (months)



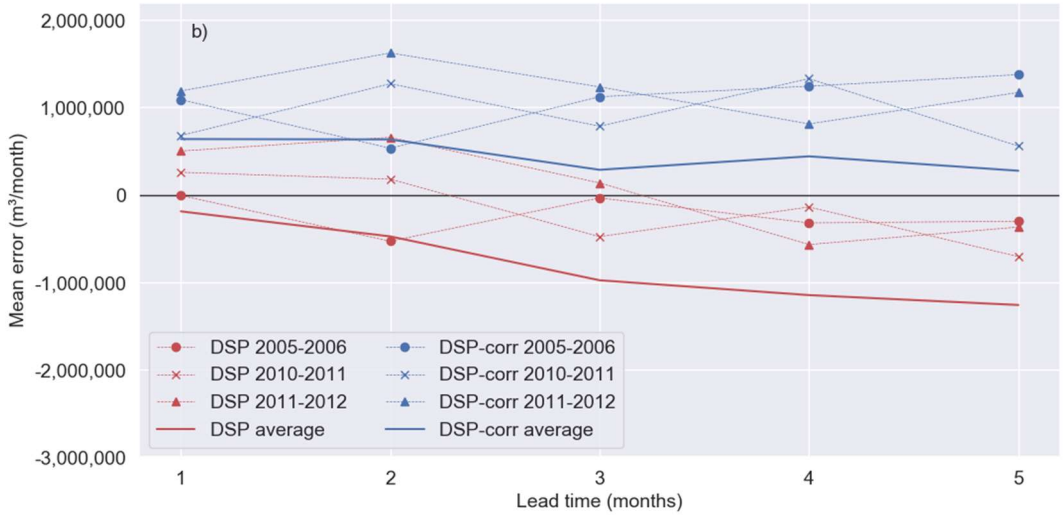




**Figure 3 Skill of the hydrological forecast ensemble (inflow to reservoir S1) during the pumping licence window (1 Nov - 1 Apr) measured by the CRPSS (a) and the mean error (b) for different lead times. from 1 Nov. Red lines represent the skill without bias correction of the ~~non-bias-corrected~~ meteorological forcing (ECMWF seasonal ~~forecast,~~forecasts), blue lines represent the skill after bias correction. ~~The solid line~~Solid lines represents the ~~mean~~average skill over the period 2005-2016, while circles, crosses and triangles represent the skill in 3 particularly dry winters. (Nov-Apr). CRPSS = 1 represents the perfect forecast and CRPSS = 0 the no skill threshold with respect to the benchmark (ESP).**

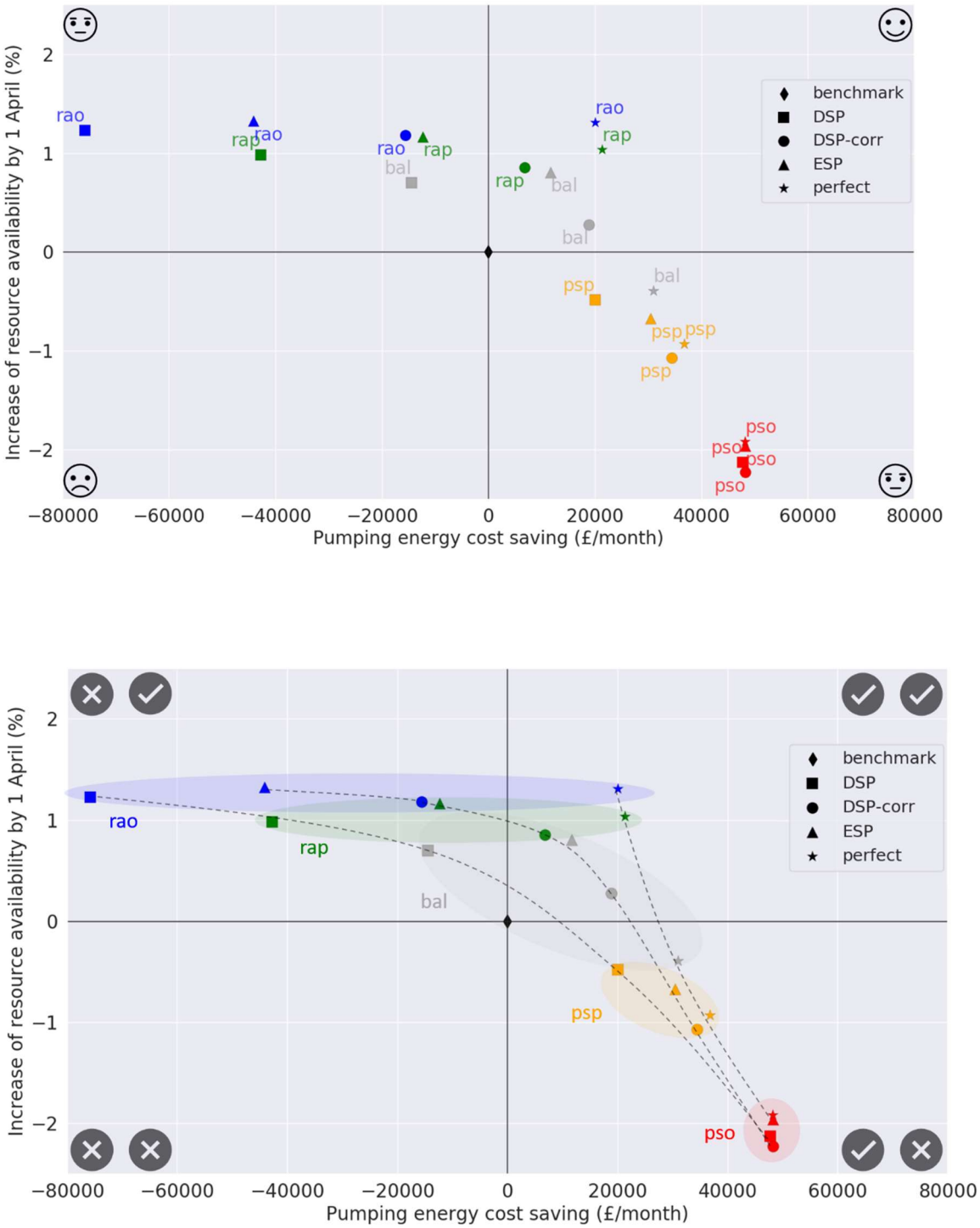



**Figure 4 Post-evaluation Pareto fronts representing the average system performance improvement (over period 2005-2016) of the real-time optimization system during the pumping licence window (1 Nov - 1 Apr) with respect to the benchmark (black diamond), using four forecast products: non-corrected forecast ensemble (DSP), bias corrected forecast ensemble (DSP-corr), ensemble streamflow prediction (ESP) and perfect forecast. For each of the four forecast products, five ~~decision making~~scenarios of operational priorities are represented ~~depending on the dominant priority from 100% priority to maximize resource availability (top left) to 100% priority to maximize cost savings (bottom right):~~: resource availability only (rao; in blue), resource availability prioritised (rap; in green), balanced (bal; in grey), pumping savings prioritised (psp; in green) and pumping savings only**

**(pso; in red). For visualization purposes, the coloured circles group points under the same operational priority scenario and the dashed lines link points using the same forecast product. The pumping energy cost is calculated as the sum of the energy costs associated to pumped inflows and pumped releases and the resource availability as the mean storage volume in both reservoirs (S1 and S2) at the end of the optimisation period. ~~The annotation is the corresponding operational priority scenario for each point~~Both objective values are rescaled with respect to the performances of the benchmark operation.**

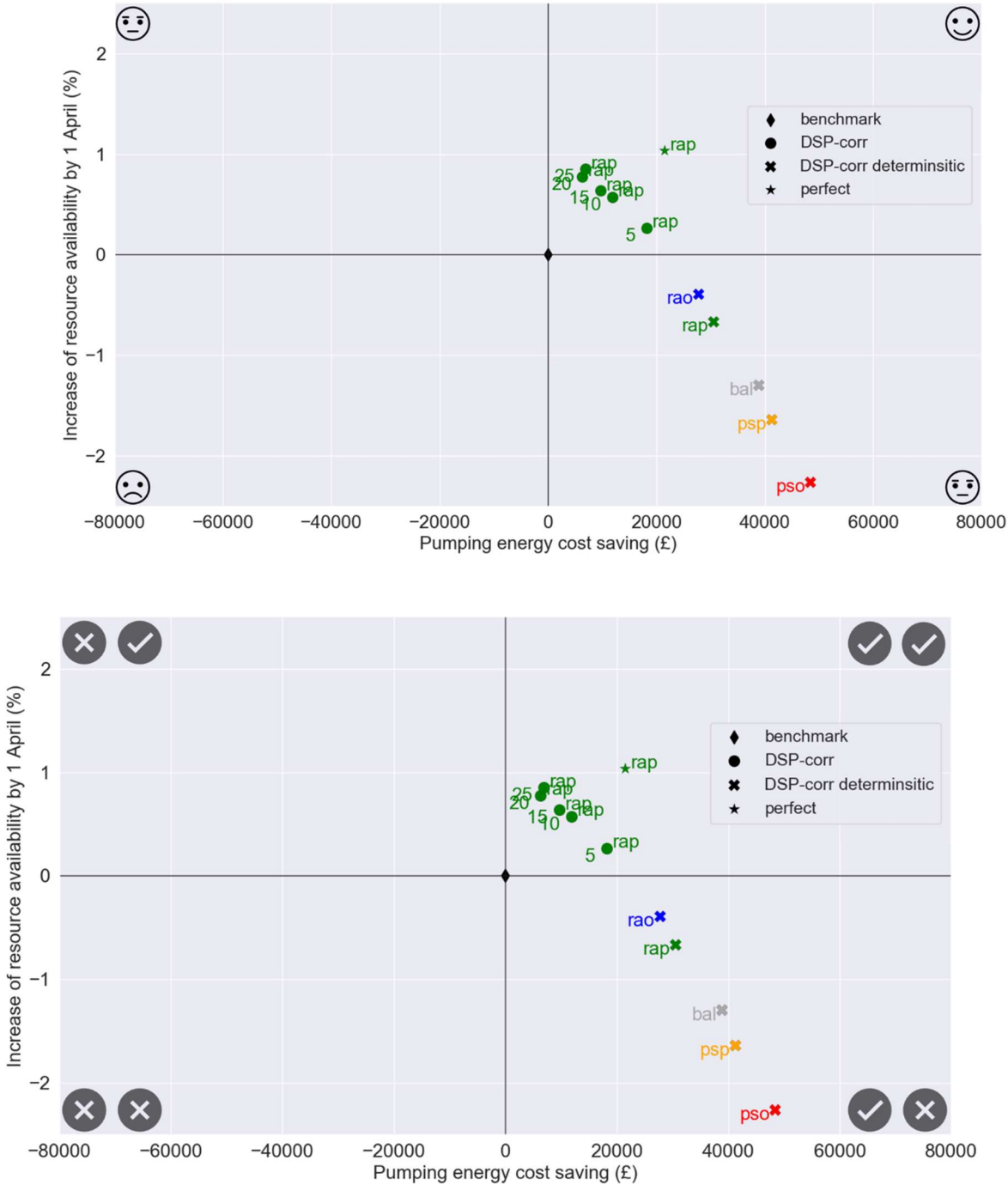
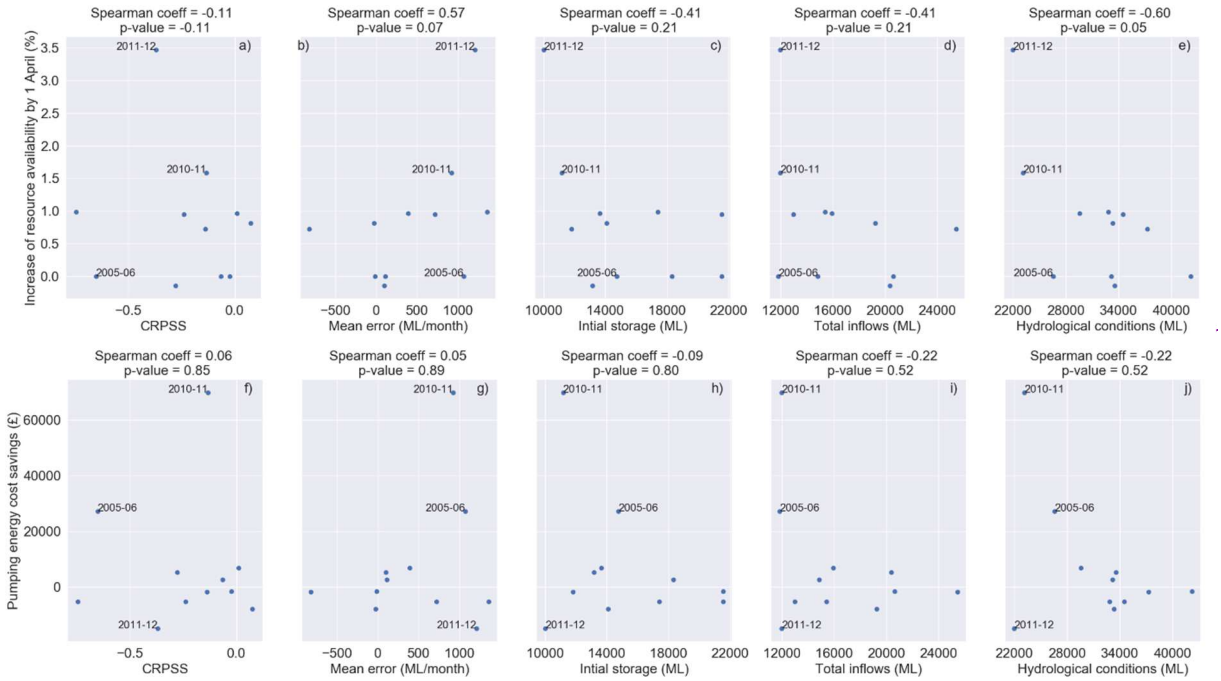



**Figure 5 Post-evaluation Pareto fronts representing the average system performance (over period 2005-2016) of the real-time optimization system during the pumping licence window (1 Nov - 1 Apr) with respect to the benchmark (black diamond), using bias corrected forecast ~~deterministic~~ensemble (DSP-corr ~~deterministic~~))**

**with different ensemble size, and ~~bias correctedthe~~ mean of the forecast ensemble (DSP-corr~~) with different ensemble size.~~ deterministic). For practical purposes, only the "resource availability prioritised" scenario (rap) is represented for the DSP-corr. ~~The annotation is the corresponding operational priority for each point.~~ The annotation numbers ~~are~~refer to the ~~number of~~ ensemble ~~members considered. The perfect forecast for rap is also represented for reference purposes~~size.**


Spearman coeff = -0.11
p-value = -0.11
Spearman coeff = 0.57
p-value = 0.07
Spearman coeff = -0.41
p-value = 0.21
Spearman coeff = -0.41
p-value = 0.21
Spearman coeff = -0.60
p-value = 0.05
Increase of resource availability by 1 April (%)
CRPSS
Mean error (ML/month)
Intial storage (ML)
Total inflows (ML)
Hydrological conditions (ML)
Spearman coeff = 0.06
p-value = 0.85
Spearman coeff = 0.05
p-value = 0.89
Spearman coeff = -0.09
p-value = 0.80
Spearman coeff = -0.22
p-value = 0.52
Spearman coeff = -0.22
p-value = 0.52
Pumping energy cost savings (£)
CRPSS
Mean error (ML/month)
Intial storage (ML)
Total inflows (ML)
Hydrological conditions (ML)

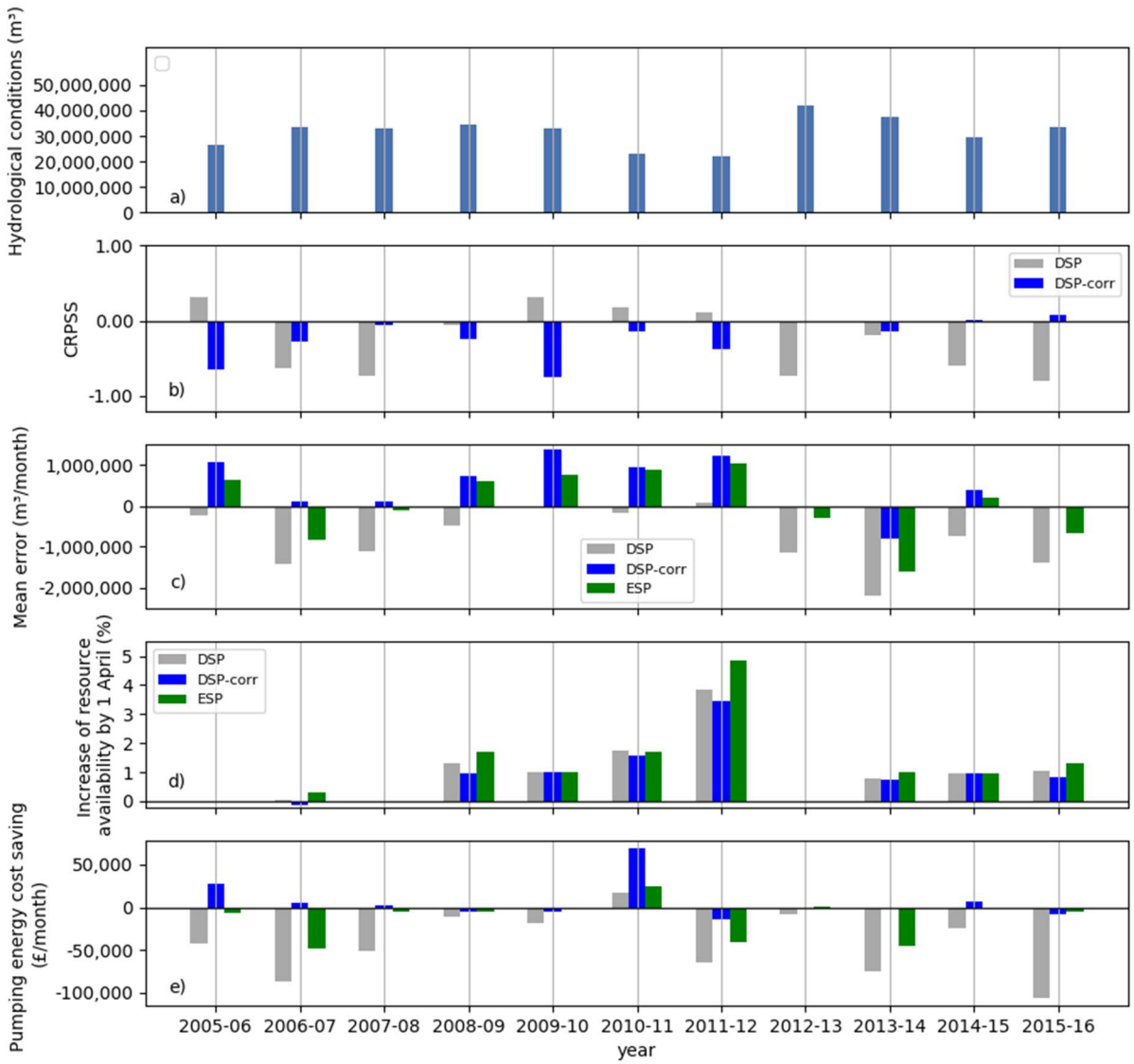
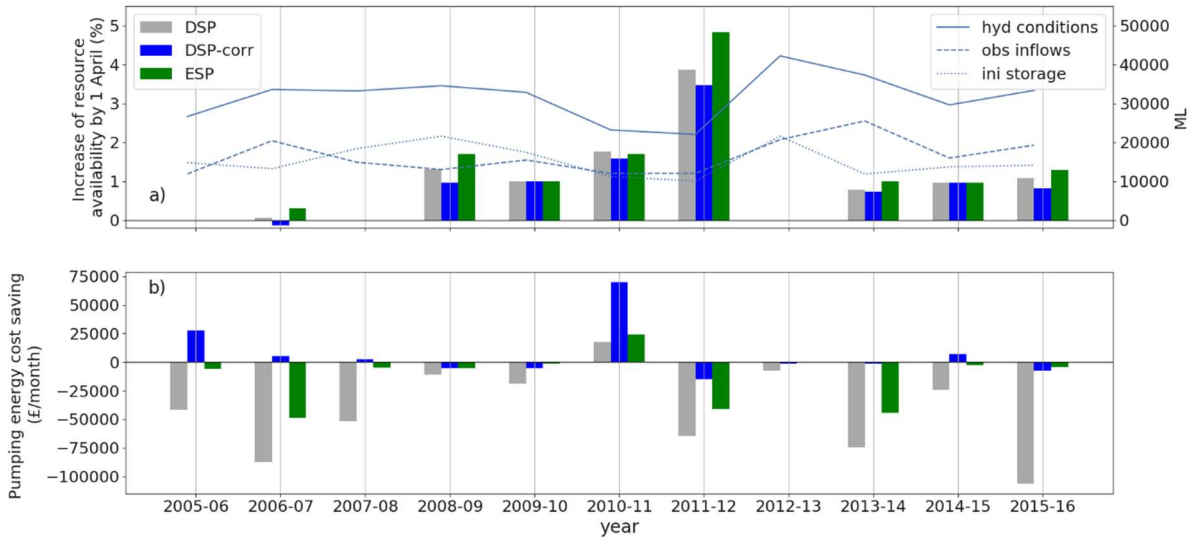



**Figure 6** ~~Bias corrected forecast ensemble (DSP-corr) for the "resource availability prioritised" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~




**~~Figure 7~~ Year-by-year a) ~~Total observed inflows (1 Nov – 1 Apr), Initial reservoir storage (1 Nov) and~~ Hydrological conditions (Total observed inflows + Initial storage~~) (right hand y-axis))~~; forecast skills of the meteorological forcing: b) CRPSS and c) Mean error; d) Increase of resource availability ~~(left hand y-axis)~~ and ~~b~~e) Pumping energy cost savings of the real operation system informed by: the dynamical streamflow prediction (DSP), the bias corrected dynamical streamflow prediction (DSP-corr) and the ensemble streamflow prediction (ESP) for the "resource availability prioritised" (rap) scenario. Please note that ESP is not shown in b) as it is the CRPSS benchmark.**

# Supplementary material

## ~~Details of the reservoir~~Reservoir system model

We use weekly resolution to simulate the system and its operation for both the benchmark and the real-time optimization system (RTOS) approaches. For each reservoir (S1 and S2), the volume of stored water (*s(t+1)*) is equal to the previous week's storage (*s(t)*) plus natural and controlled inflows minus releases, evaporation and spills. The mass balance equations are:

S1: $s_{t+1} = s_t + \left(I_{S1,t} + u_{R,S1,t}\right) - \left(u_{S1,D,t} + u_{S1,R,t} + evap_t + spill_t + env_t\right)$

S2: $s_{t+1} = s_t + \left(I_{S2,t}\right) - \left(u_{S2,D,t} + evap_t + spill_t + env_t\right)$

Spills are calculated by imposing the hard constraint that the storage at next time-step should never exceed the reservoir capacity, hence they are either equal to zero or to the excess volume generated by the storage plus inflows minus outflows:

S1: $spill_t = max\left(s_t + \left(I_{S1,t} + u_{R,S1,t}\right) - \left(u_{S1,D,t} + u_{S1,R,t} + evap_t + env_t\right) - s_{max}, 0\right)$

S2: $spill_t = max\left(s_t + \left(I_{S1,t}\right) - \left(u_{S2,D,t} + evap_t + env_t\right) - s_{max}, 0\right)$

where $s_{max}$ the reservoir storage capacity in ~~ML~~$m^3$. Controlled inflows and outflows (u) are limited by the real-world system capacity. Besides, pumped inflows are limited such that flow downstream of R ~~will~~does not drop below a legal constraining value, unless using water released from S1. Evaporation fluxes (evap) are computed as the product of the reservoir surface area by the potential evaporation rate. Environmental compensation flows (env) are given by prescribed values that are kept constant over the year.

## ~~Details~~Formulation of the optimization problem

Both the release scheduling of the benchmark approach and the release and pumped inflow scheduling of the real-time optimization system (RTOS) approach are optimized using the NSGA2 genetic optimization algorithm included in the Platypus Python package (https://platypus.readthedocs.io/). ~~The~~ In the RTOS, the optimization decision variables are both the weekly reservoir releases ($u_{S1,D}$ and $u_{S2,D}$) ~~for both reservoir operation approaches~~ and the weekly pumped inflows ($u_{S1,R}$~~) for the RTOS approach~~); in the benchmark operation, the decision variables are the reservoir releases only~~.~~, while the pumped inflows are calculated according to the control curve. We assume that the future water demand is perfectly known in advance, and equal to the sum of the observed releases from S1 ($u_{S1,D}$) and S2 ($u_{S2,D}$) for the period of study. Unless physically unfeasible, the sum of reservoir releases ($u_{S1,D}$ + $u_{S2,D}$) is always forced to meet such demand.

~~For~~When simulating the benchmark ~~approach the reservoir S1~~ operation ~~rule curve, and not the optimizer, defines, according to the storage level and date, when pumped inflows ($u_{R,S1}$) are triggered. The~~ , the optimization ~~decision variables are the weekly reservoir releases ($u_{S1,D}$ and $u_{S2,D}$). As an optimization constraint, the storage volume~~ is constrained to achieve maximum storage for both reservoirs (S1 and S2) ~~is set to be maximum~~ by the end of the pumping license period window (1 April~~) and the~~). The (single) optimisation objective is to minimize the sum of the energy costs for pumped release ($u_{S1,D}$) ~~energy costs:~~and pumped storage ($u_{R,S1}$):

$$\sum_{t=0}^{T} c_{R,S1}\, u_{R,S1,t} + \sum_{t=0}^{T} c_{S1,D}\, u_{S1,D,t}$$

where *c* is the pumping energy cost ~~per ML~~ and T is the lead time in weeks.

For the RTOS approach, the optimization decision variables are the weekly reservoir releases ($u_{S1,D}$ and $u_{S2,D}$) and the weekly pumped inflows ($u_{S1,R}$) and the optimization ~~objective is to minimize the following objective functions~~objectives to be minimized are two:

1) Sum of the pumping energy costs (same equation as above)

~~1)~~2) Average ~~of the~~ difference between the ~~storage~~reservoir capacity and storage volume by 1 April ~~for~~in the two reservoirs (S1 and S2):

$$\frac{(s_{S1,max} - s_{S1,T}) + (s_{S2,max} - s_{S2,T})}{2}$$

~~where $s$ is the reservoir storage volume in ML and $s_{max}$ the reservoir storage capacity in ML.~~

~~2) Sum of the pumping energy costs (only applied on the multi-objective optimization of the RTOS approach):~~

$$\sum_{t=0}^{T} c_{R,S1}\, u_{R,S1,t} + \sum_{t=0}^{T} c_{S1,D}\, u_{S1,D,t}$$

~~where $c$ is the pumping energy cost per ML and T the lead time in weeks.~~

$$\frac{1}{M}\sum_{m=0}^{M} \frac{(s_{S1,max} - s_{S1,T,m}) + (s_{S2,max} - s_{S2,T,m})}{2}$$

where $s_{max}$ is the reservoir storage capacity in $m^3$, $s$ is the reservoir storage volume in $m^3$, $T$ is the final week of the optimisation period, and $M$ the total number of ensemble members. Notice that, as denoted by the subscript $m$, the final storage of S1 and S2 will differ depending on the inflow forecast ensemble member that is used to force the simulation, even if the set of pumping and release decisions remain the same. Hence, at each iteration of the optimisation procedure, the same set of decisions is evaluated against each ensemble member and then the objective value is obtained by averaging across all the simulations (with the exception of the "deterministic" case presented in Sec. 3.2.2, where the ensemble forecast is replaced by the mean forecast and therefore averaging is not needed as only one simulation is run against any set of decisions).

For both operation approaches ~~the optimization consisted of 100,000 runs per iteration and~~, benchmark and RTOS, the population size for the multi-objective optimization of the RTOS approach was 20.

## Observational hydrological data

Daily rainfall in the study area from 1981 to 2016 was derived from the UK Centre for Ecology and Hydrology (CEH) Gridded Estimates of Areal Rainfall (CEH-GEAR) dataset (Tanguy et al., 2014) and daily temperature and PET data for the period was derived from the CEH Gridded CEH-CHESS dataset (Robinson et al., 2016, Robinson et al., 2017). CEH-GEAR is a gridded 1km product derived from the interpolation of observed rainfall across all daily and monthly rain gauges in the UK. CEH-CHESS is a gridded 1km product derived from the Met Office 40km gridded MORECS dataset (Hough and Jones, 1997). We used the HBV model forced by observed weather data to simulate a proxy of the daily observed inflows (Figure 7). The HBV model was previously calibrated against observed hydrographs from 1972 to 2003 in the Wimbleball catchment. For the Wimbleball catchment the average observed yearly inflow is 24,462,227 $m^3$/year with an interannual standard deviation equal to 4,340,594 $m^3$/year. Given the lack of good calibration data for the Clatworthy catchment, we applied the Wimbleball

parameter set to the Clatworthy catchment, given that they are adjacent to each-other. The averaged mean error from 1972 to 2004 of the Wimbleball calibrated inflow is 33,283 $m^3$/month (from 1 Nov to 1 Apr).

## Supplementary figures

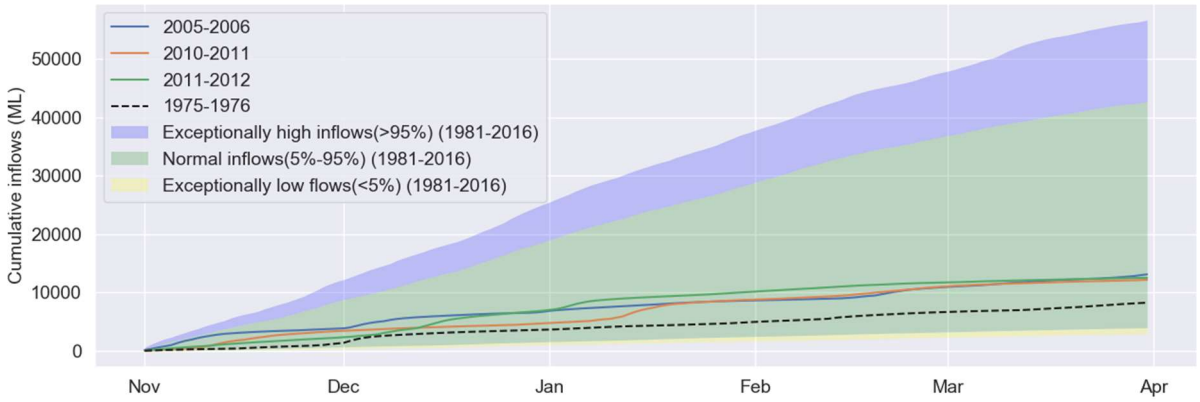
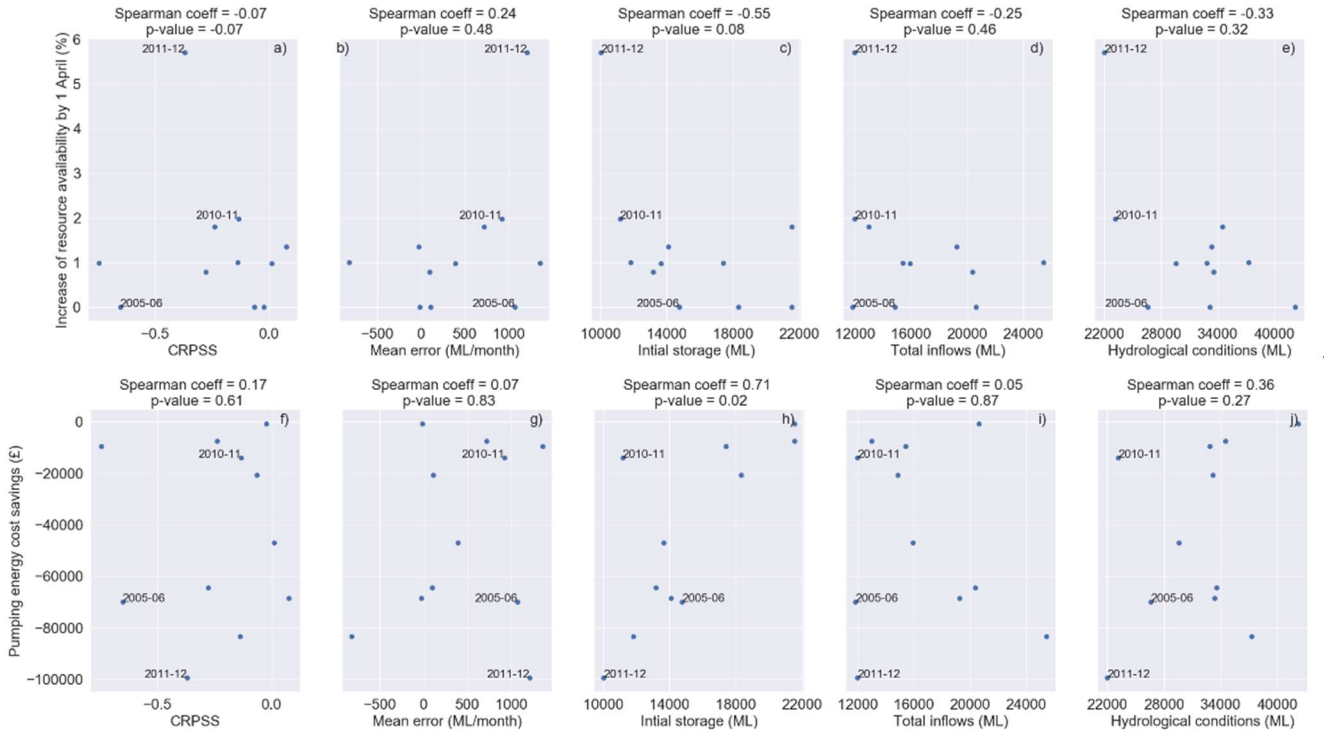
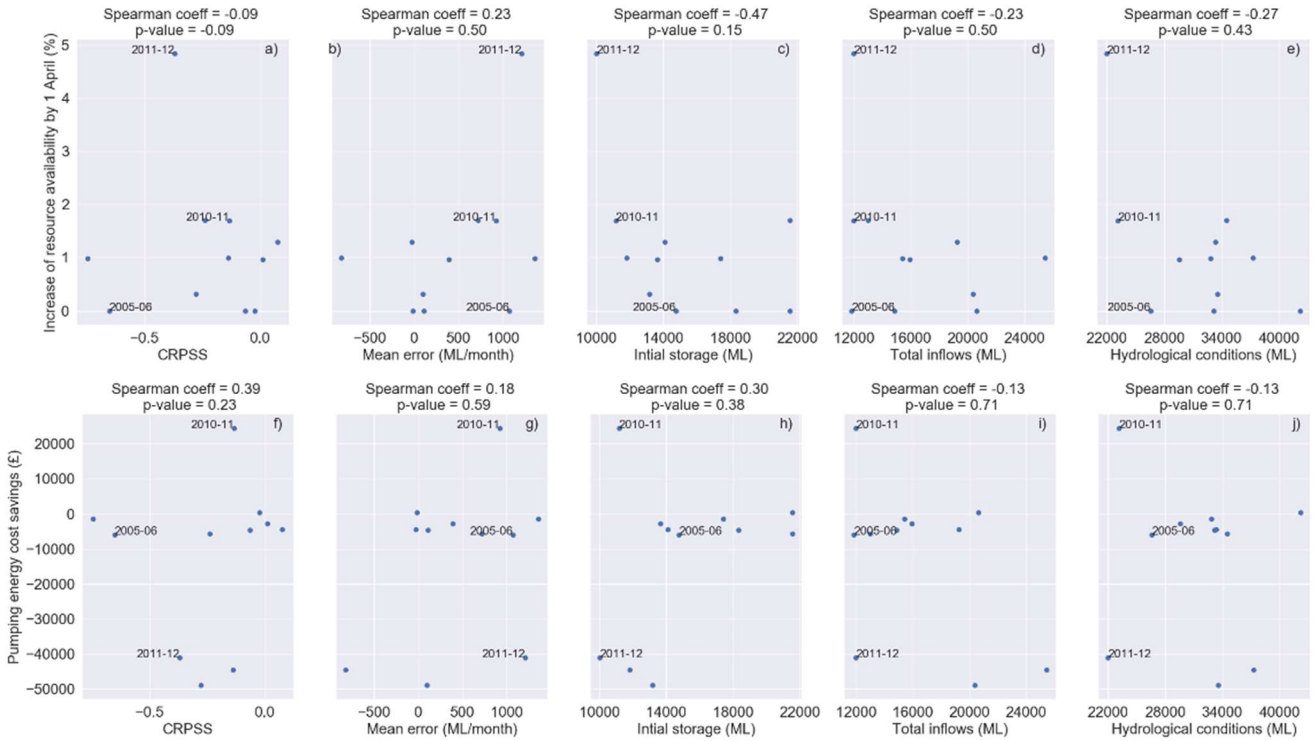



**Figure 7 Cumulative inflows to the S1 reservoir in the worst-case scenario (1975-1976) and in the three driest years (2005-2006, 2010-2011 and 2011-2012) of the period used for the simulation of the RTOS (2005-2016). Only data relative to the pumping licence window (Nov to Apr) are shown. Shaded areas show the weekly inflow distribution calculated on the period used for the forecast bias correction of the meteorological forcing and ESP generation (1981-2016). Notice that the three driest years are relatively close to the worst-case scenario. (1975-1976).**

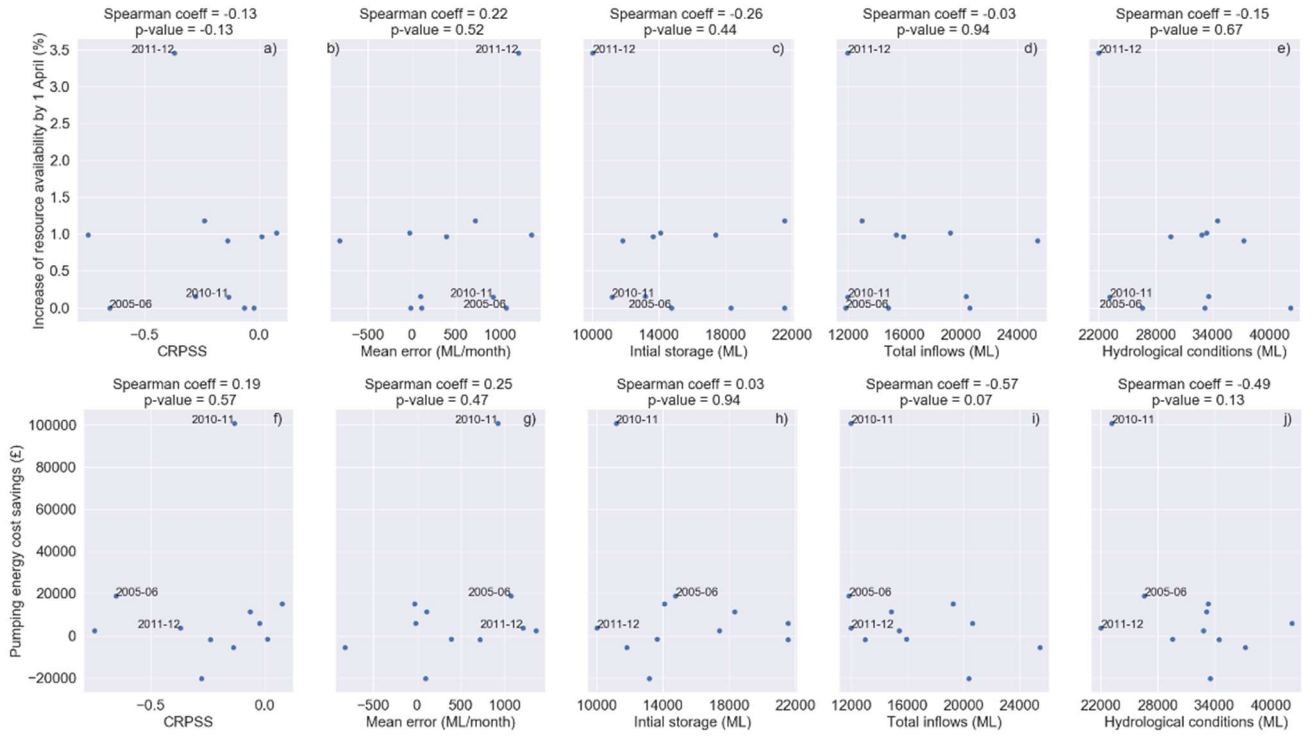



~~**Figure 9 Ensemble streamflow prediction (ESP) for the "resource availability only" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological**~~

~~conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~

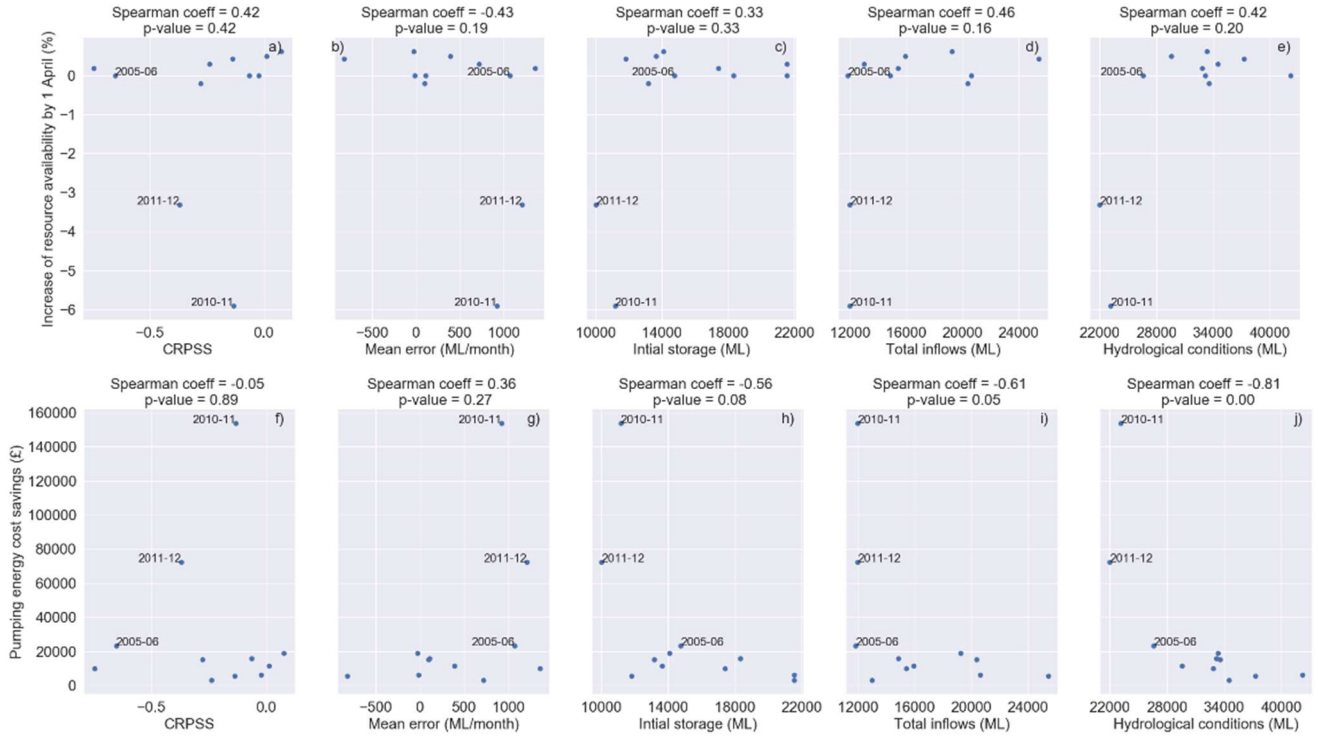



~~Figure 10 Ensemble streamflow prediction (ESP) for the "resource availability prioritised" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~

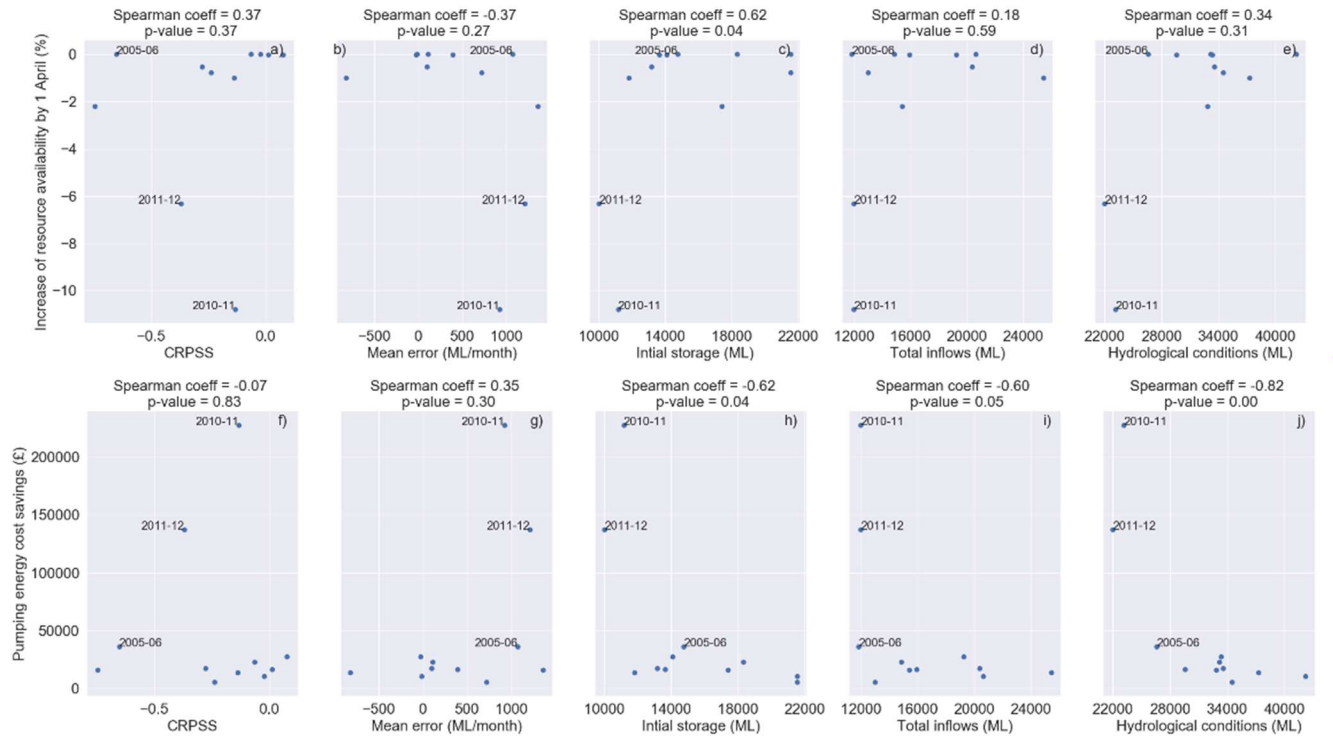



~~Figure 11 Ensemble streamflow prediction (ESP) for the "balanced" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr)~~

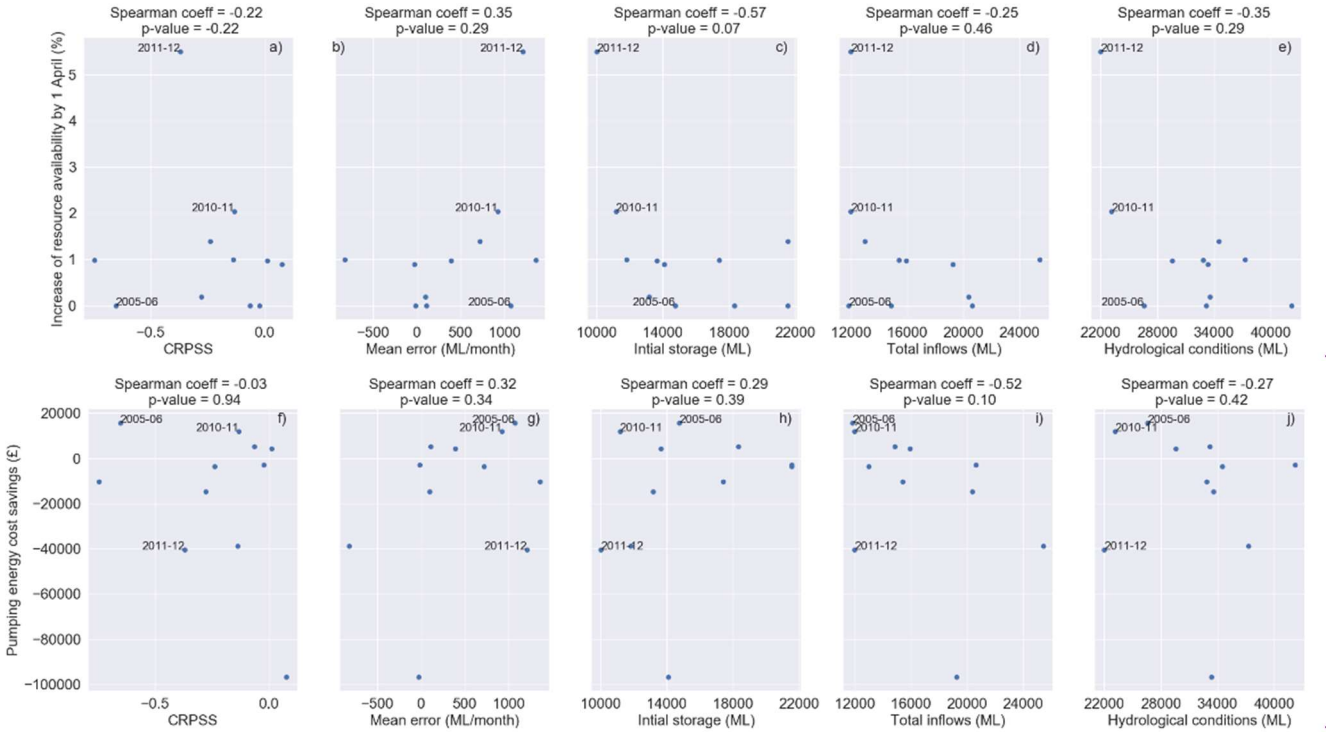
~~and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~




~~Figure 12 Ensemble streamflow prediction (ESP) for the "pumping savings prioritised" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~

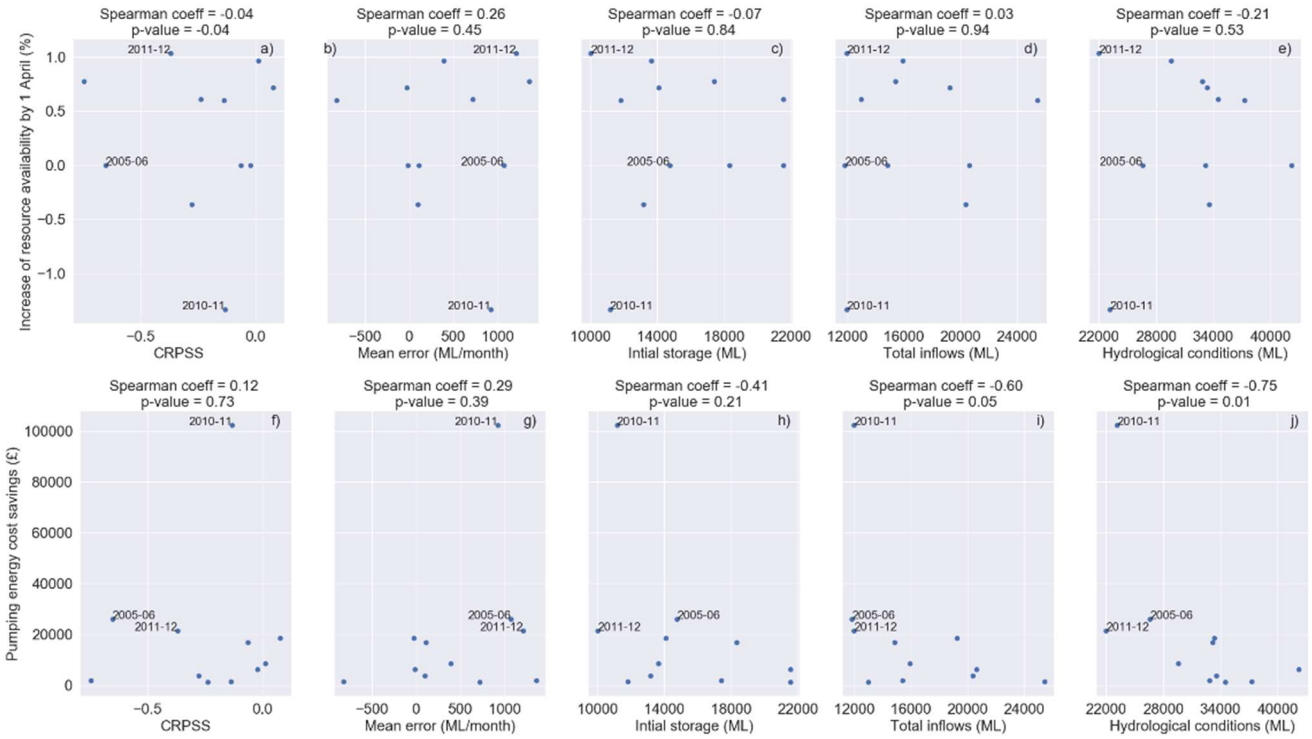



~~**Figure 13 Ensemble streamflow prediction (ESP) for the "pumping savings only" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.**~~

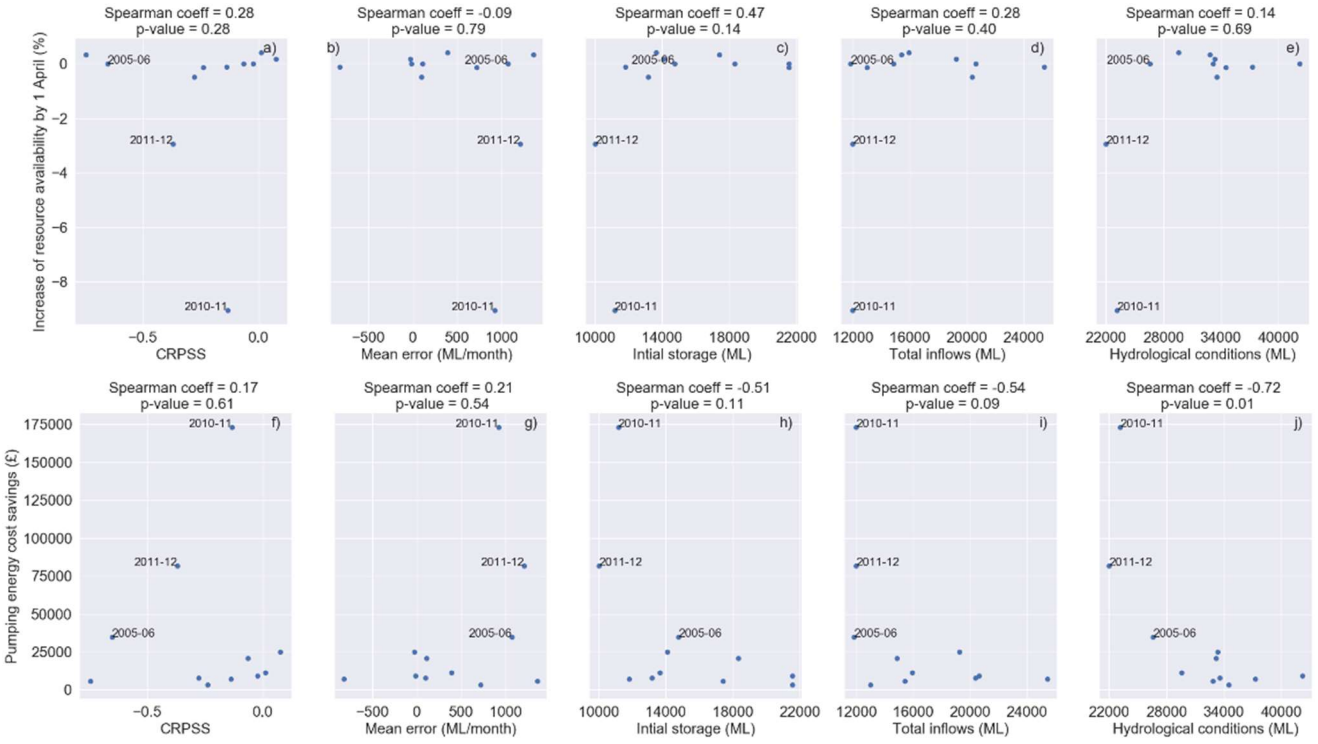



~~**Figure 14 Bias corrected forecast ensemble (DSP-corr) for the "resource availability only" scenario - correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological**~~

~~conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~

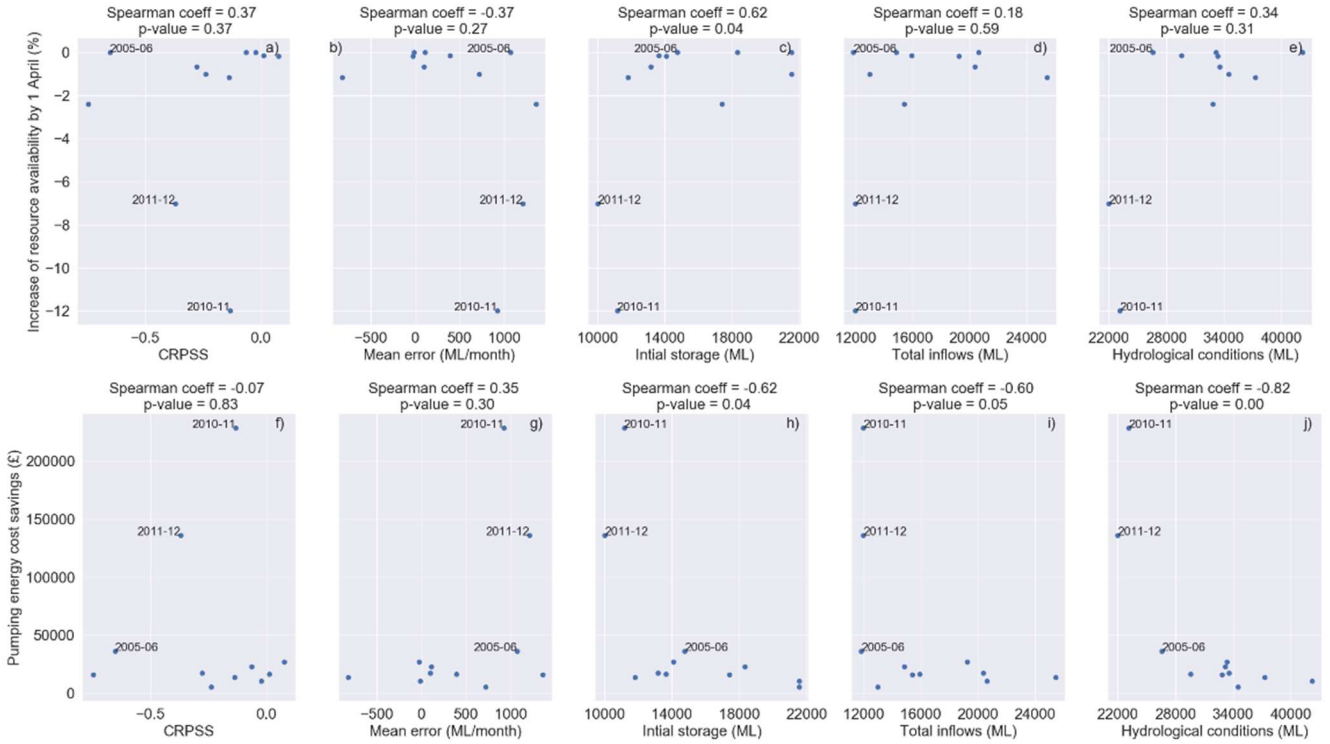



~~Figure 15 Bias corrected forecast ensemble (DSP-corr) for the "balanced" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~




~~Figure 16 Bias corrected forecast ensemble (DSP-corr) for the "pumping savings prioritised" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d)~~

~~total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~




~~Figure 17 Bias corrected forecast ensemble (DSP corr) for the "pumping savings only" scenario – correlation between Increase of resource availability and a) CRPSS, b) mean error, c) initial storage (1 Nov), d) total inflows (1 Nov – 1 Apr) and e) hydrological conditions (initial storage + total inflows) and between Pumping energy cost savings and f) CRPSS, g) mean error, h) initial storage (1 Nov), i) total inflows (1 Nov – 1 Apr) and j) hydrological conditions (initial storage + total inflows). Each point represents a year. Correlation and its significance are quantified by the Spearman coefficient and the p-value, respectively.~~