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

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 preparing a revised manuscript, we will shorten long sentences and simplify their structure. We will also address 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 planning to share an anonymised version of the code we developed for other users. 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.

We will add additional information about observations

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 will add and clarify these points in the discussion of the paper.

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 tradeoff 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 will rewrite this section to make this point clearer.

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 will replace 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. 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 will further clarify this point in 2.3.1 and we will improve 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) (lines 264-266) and not to the probabilistic approaches. We will further clarify this in the manuscript (2.3.1 and 2.3.6) and provide with an example. To further clarify this, we will add as an appendix the equations of the model (without the parameter values, which are confidential) and of the optimisation problem.

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. 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 will clarify this in 2.3.2.

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 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 469-470). 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 will clarify and justify this expectation that the forecast skill generally improves due to bias correction by citing in the Discussion several studies such as the ones above.

We agree, the sentence in lines 387-388 is wrong and will be replaced by: "However, the result points at a possible intrinsic contradiction in the very idea of bias correcting based on climatology." In lines 391-392 what we aimed to communicate is that since both bias correction and ESP forecast are based on climatology, the bias corrected DSP forecast skills tend to become equivalent to ESP forecast skills. However, ensuring this skill level with bias correction (Crochemore et al. 2016) may not be enough especially under conditions significantly drier or wetter than climatology, which are likely the ones when water managers can extract more value from forecasts. We will further clarify this point in the reviewed manuscript.

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 will be included in the discussion.