Reply to reviewer #2

General Comments

The paper has generally improved by re-structuring the presentation of the results around the distinction between "supply objective" and "level objective" and by highlighting how the buffering role of the reservoir storage contributes to the forecast value. Still, the paper needs a revision before the publication, in my opinion. In particular, the discussion of the results is, in some cases, biased and little effort has been made to generalize the results so to be exportable to other case studies (as detailed in the specific comments below). Also, it is important to remark in the text that the results are obtained using a deterministic optimization and to highlight the potential drawbacks of this approach. Finally, I suggest revising the section "Discussion and conclusions" of the paper to highlight the limitations of the approach (as detailed below).

We thank the reviewer again for the thoughtful comments and suggestions.

Specific Comments

Page 1, Lines 17-20. The fact that "good forecast accuracy does not necessarily translate into performance improvement" is not "surprisingly" (see among others: Goddard, L., Aitchellouche, Y., Baethgen, W., Dettinger, M., Graham, R., Hayman, P., Kadi, M., Martìnez, R., and Meinke, H. (2010). Providing seasonal-to-interannual climate information for risk management and decision-making. Procedia Environmental Sciences, 1, 81-101.). The authors should revise the abstract accordingly and (at least briefly) mention in the introduction that the relationship between forecast skill and value has already been studied (including some relevant references).

We have removed the term "surprising" and cited Goddard et al. (2010) in our introduction.

Page 1, Lines 20-21. I would remove the sentence because it is not clear.

Our main conclusion from this work is that forecast skill consistency through time is essential to guarantee performance improvements in supply targeted systems (but not in level targeted systems). We wish to retain this in the abstract, but have tweaked the wording to make it clearer.

Page 2, Lines 19-20. I would remove the statement. Instead, I would enlarge the description by including more technical details on the approach (e.g., what "design parameters" are "adjusted", what optimization scheme is used, how the forecast value is computed, which forecasts are used, etc). I would remove the last line, because the paper does not deal with the risk associated to forecast-informed decisions.

We've amended the statement in question.

The suggestion to add more detail on the approach is stylistic preference. We would rather keep technical method detail inside the method section.

We disagree that the paper does not deal with the risk associated for forecast-informed decisions. We show that there is greater risk in adopting a forecast system with a supply objective (high chance of negative outcome) than with a level objective (low chance of negative outcome).

Page 4, Equation 2. The mass balance should include the spill.

Ok – thanks.

Page 4, Line 15-16. Specify what you mean for "hydrological conditions".

Ok.

Page 4, Line 20. Include the description of the release strategy when the demand can not be met in full.

Ok.

Page 6, Line 25. From this paragraph, it seems that a deterministic approach to reservoir optimization has only advantages, while it is well known that optimization approaches that explicitly account for the forecast uncertainty (included in a forecast ensembles, for instance) allow for better operating performances (e.g., Boucher, M., D. Tremblay, L. Delorme, L. Perreault, and F. Anctil (2012), Hydro-economic assessment of hydrological forecasting systems, Journal of Hydrology).

Ok—note added to inform reader that ensemble approach should glean more value + citation added.

The authors should properly comment the drawbacks of the deterministic optimization approach.

The drawback is implicit from the comment added in response to the above. Note that whilst this is a drawback for improving reservoir operations, it is not a drawback for the present study, which relies on comparison between two similar approaches.

Page 8, Lines 11-25. The whole paragraph is not fully convincing, because the authors comment only part of the results reported in Figure 5. For instance, in contrast to what it is currently commented in the text, there are also some cases in which the forecast skill is weak, but the decisions outperform the benchmark (as in Figure 5 e-f). Why does this happen? The authors should comment all the results they obtain and should try to infer the causes, so that the entire description would result more convincing and exportable to other cases.

It's possible to get very good performance when forecast skill is weak for exactly the same reason that it's possible to get poor performance when forecast skill is generally strong. You could have a terrible forecast, but if it happens, by chance, to make a good prediction at a critical moment, then the overall performance will seem good. This is why in our recommendations we suggest that forecast value estimations cannot be made on the basis of a single simulation.

In figures 5(e-f) there is also some uncertainty in the result, which can be explained by the fact that even the level targeted systems are to some extent buffered (e.g., if you're just above to the target level and if the maximum release gives sufficient control to adjust favorably irrespective of the incoming flow).

Page 8, Line 20. Which reservoirs are the "very large reservoirs" mentioned in the text? The results in Figures 5-7 should be commented also in relation to the storage ratio, which might clarify some of the results.

This is a general comment, rather than based on the result. Clarified in text.

The storage ratio should correspond to the capacity to inflow ratio. I would include also this definition in the caption of Table 2, because it is more commonly used in the literature.

Ok.

Page 8, Line 26. This comment is quite strange, because the Upper Yarra Reservoir is the reservoir with the lowest storage ratio, among the ones considered in the experiment 1 (i.e., 0.2 against 0.76, 0.35, and 2.84), which suggests that this reservoir has the lower potential for buffering inflows. Can the authors comment on this? If not, this comment should be removed.

Yes, it has lower potential for buffering inflows. That's the point. We are arguing that the buffering of inflows is what causes the uncertainty in the performance. It's the buffering that renders the forecast superfluous for long periods of operation, meaning performance becomes subject to isolated moments in the simulation rather than overall forecast skill.

Page 9, Line 13. I suggest using "difference" instead of "variance".

Why? It's the variance of forecast skill in time that will be linked to the performance uncertainty.

Page 9, Lines 17-21. Can you cite relevant literature describing this behavior? Otherwise, it should be clear that you infer this comment from your results and not that your results confirm what other works have already observed (as it seems from the current text).

What we describe here is both a well-established phenomenon of reservoir storage dynamics and an outcome observed in our results.

Page 9, Lines 22-23. The Eppalock reservoir is very big in almost all the configurations that you consider, as demonstrated by the storage ratio in Table 2. Can you comment on the reason why it is interesting or even realistic take into account such configurations?

I'm afraid we don't quite understand the point raised here. The Eppalock reservoir size varies widely across the configurations, as demonstrated by the storage capacity in Table 2 (ranging from 58 Mm^3 to 1061 Mm^3). The storage ratio is generally large because the mean inflow is low. There is no reason why these

configurations are unrealistic—in fact the actual design capacity of Eppalock (300GL) lies well inside the range we consider.

Page 9, Lines 30-31. The Eppalock and Upper Yarra reservoirs are characterized by pretty different configurations in this experiment (see for example the storage ratio). Why do they show similar results? I believe that the authors should put more effort in interpreting the results, so to give insights that allow extending the results from the specific configurations to more general settings. If this is not possible, the discussion and conclusions should acknowledge that the results described in the paper are relevant to the specific case studies.

Explanation is given in later paragraphs (FoGSS forecasts for Eppalock and Upper Yarra are consistently skillful).

Page 9, Lines 32-33. A (similar?) drawdown is visible also for the draft ratios equal to 50% and 80% reported in Figure 7. Why is the 90% draft ratio not included?

Reworded to clarify.

Page 9, Lines 33-36. Can this be a valuable justification given that the performances of the Serpentine catchment (reported in Figure 3) are similar to the performances for the Burrinjuck reservoir and better than the ones for the Eppalock reservoir?

Yes this is justified. Results reported in Figure 3 are not based on FoGSS forecasts.

Page 9, Lines 37-39. I agree with the authors that the forecasts should be assessed on the longest period possible (see the author response to the previous review). Still, I believe it would be interesting to quantify the forecast value on informing reservoir operation in case of less extreme droughts that the Millennium drought. It might result that the value is not as pronounced or that errors in the forecasts are not as decisive for properly hedging.

Point taken. Note that by varying the reservoir designs we do in fact make some of the reservoirs sensitive to dry spells unrelated to the Millennium drought.

Page 10, Line 25. What do the author mean exactly with "high sensitivity of the Serpentine reservoir"? If I am not mistaken, this is not commented in Section 3.2.

Added comment to Section 3.2 to clarify.

Page 10, Lines 33-43. What is the rationale behind computing the consistency of the forecast quality over blocks of more that 12 or 24 months? This piece of information would be relevant only for reservoirs that are able to buffer inflows on time scales that are comparable to those blocks (which does apply only to the configurations of the Eppalock reservoir, among the ones considered in this work).

Some of these reservoirs have critical drawdown periods of more than ten years. We're mainly using this analysis to show that FoGSS forecast consistency varies

across the models, and that for Serpentine and Burrinjuck we can find lots of periods < 5 years duration for which overall skill is negative. This could explain why these two reservoirs present negative instances of performance gain.

Page 11, Section 5. I have the impression that the section "Discussion and conclusions" summarizes only part of the results (for instance the sentence reported in Lines 12-14 is not true for all the configurations tested, e.g., it does not hold for Figure 5d). It should be revised in order to be more balanced.

We think it does hold for 5d. Here we see that the Upper Yarra (which has the lowest capacity to buffer) is least sensitive to forecast injected error—which is exactly what our conclusions would suggest. See reply to final comment.

In addition, the authors should mention explicitly that the results are obtained using a deterministic approach and they should comment about the drawbacks of adopting such an approach.

We state explicitly that our method makes use of the median of the ensemble forecast and that we use a deterministic approach. We've added a note and cited other studies to show that we might have used the full ensemble. It is not clear to us, however, that using the full ensemble necessarily improves performance—we did some preliminary tests applying our approach across the full spread using a scenario tree method (not reported, for brevity) and found performance to be no better, and in some cases worse. We opted to describe the simplest approach in the paper and we believe more research is required to establish the efficacy of optimization methods that use the full ensemble.

Finally, Lines 26-40 include a discussion that is not supported by the results of the paper and therefore should be removed (or moved to the introduction). For the same reason, I would remove the last sentence on Page 12, Lines 13-14.

We do not understand this comment. Our results show very clearly that negative performance gains can appear in supply targeted systems and not in level-targeted systems when skillful forecasts are applied. This must mean that operators of supply targeted systems are exposed to greater risk when adopting these services. We feel that these sentences convey the import of our findings and so we wish to retain them.

Page 12, Lines 4-5. Not always mediocre forecasts would imply "a false confidence in the forecast-informed decision scheme". For instance, forecasts with medium skill may drive good reservoir operation performances when the system is insensitive to some forecast errors (this can happen, for example, because of the buffer capacity of the reservoirs).

True, it can work both ways. But the operator needs to be aware of this uncertainty, and he or she may overlook it if looking at a case study with only one set of forecasts and one simulation.

Page 21, Figure 5. Why does the forecast value show such a high variance when the supply objective is considered (subplot a-d)? I would expect that the storage

buffer effect would make the reservoir operation insensitive to some forecast errors.

This is true, and is the major conclusion of this study. Most of the forecasts are superfluous. The performance relative to benchmark is determined by forecast skill at a few important moments (like when the system is drawn down to critical levels). So the variance in storage operation performance is inherited directly from the variance in forecast skill (even though the forecasts are of equal skill on net over the 29 year simulation). However, it's not possible to know whether a coming forecast will occur in such a critical period—all we can do is assess the consistency of forecast performance. Hence one of our major recommendations is: measuring consistency of forecast skill as a useful indicator of forecast utility in storage operations. We have added a sentence to the abstract restating this.