Response to comments from Reviewer 2

Summary

2.1. The paper presents a generalisation of the Relative Economic Value (REV) approach, providing a flexible metric, the "Relative Utility Value" (RUV), which can be useful for decision makers on the value of probabilistic subseasonal forecasts. The results show its application and sensitivity to several factors in a case study in Australia.

The paper is well written and demonstrated. I believe it brings novel aspects in the topic of hydrometeorological forecasting, and is an excellent demonstration of how forecast producers and users should work together to enhance the usefulness of skilful forecasts.

I have just some minor general and specific comments, presented below.

Thank you for this thorough, well considered, and detailed review. We appreciate your words of encouragement and suggestions which will improve the quality of this work.

General comments:

2.2. I think some sentences need to be more carefully revised because they might convey a message that goes beyond the experimentations of this paper. For instance, concerning the first sentence of the Conclusions section, I do not believe that, overall, the value of probabilistic forecasts to making (good) decisions has not been established, as the authors say. Many public and private companies are convinced of the value of quantifying uncertainties in real-time forecasting and that is why this type of forecasts has been increasingly produced and used for many operations, from nowcasting to short-term flood forecasting and long-term inflows to reservoirs. Value has not been established (or explicitly calculated) at all lead times and users cases, I agree, but, overall, the forecasting (producers and users) community acknowledges that there is value for decision making in not being certain (or deterministic) about the unknown future. The added value of the paper, in my opinion, does not lie on bringing the "value" into discussion in forecast verification/evaluation, as this has been done in several papers previously, but in making the framework for assessing it more accessible and flexible, as the title says.

Thank you for bringing this important point to our attention. We now recognise that some statements in our paper are too general, and go beyond the experimentations in our paper. In particular, we will rewrite parts of the Conclusions to address this problem, including:

- Line 626-678: Replace "but to date their value for decision making has not been established. Forecast value methods attempt to quantify this potential" with "and forecast value methods attempt to quantify this potential.".
- Line 633: Replace "can be incorporated by the user" with "can be incorporated into RUV by the user".
- Line 638: Replace "decision-maker characteristics" with "user risk aversion and exposure to damages".

Our proposed changes in response to the following comment (point 2.3) will also align the message with the experimental results.

We also agree that the main contribution of this study is to introduce a more flexible framework for quantifying forecast value, rather than to establish the value of probabilistic forecasts in all cases, or their value over deterministic forecasts. We note that this is stated in the research aims on lines 104-108 and supported in the Introduction at lines 92-97.

2.3. I was also puzzled by the authors when they say that a decision maker who is highly exposed to damages should use the reference climatology rather than a forecast based on meteorological numerical models for binary decisions (Conclusions, lines 639-640). This might be the case for the experiment showed (and the case described in the paper), but I doubt flood forecasters (forecasting a threshold exceedance for the next 12-24 hours, for instance) would be able to say to the population they are serving that they will abandon a city located close to a river and leave than with only a climatology-based information instead of rather investing into a (good) model-based forecasting and alert system because they are highly exposed to damages. I fully understand that if the potential costs of a flood event are high, and will be incurred if the flood occurs, whatever forecast we might deliver, then no forecasting system can save us, and it is better to work on protection (decreasing costs) at first. But even in this case, using climatology might not be beneficial either (the problem is elsewhere, not in the type of forecast being used). What I mean is that out of a more explicitly presented context, some sentences might rather diverge a reader from the purposes of the paper. Therefore, I would recommend to revise some general affirmative sentences, or at least introduce more context to them to avoid misunderstandings.

Thank you for this considered comment. Reviewer 1 (point 1.3) also asks that we revise the conclusions to focus more on the outcomes of the sensitivity analysis experiments, rather than outcomes of the case study. When we re-write the conclusions, we will avoid general affirmative sentences and state that the outcomes are context dependent and provide clear information on that context.

For example, we will adjust lines 636-641 to state the case study context and outcomes more explicitly, from:

> "A case study demonstrates that subseasonal streamflow forecasts should be preferred over a reference climatology forecast for all lead-times studied (max 30 days) and almost all decision-makers regardless of their risk aversion. This positive forecast value is robust to changes in decision-maker characteristics, decision types (binary, multi-categorical, and continuous-flow), and decision-making approaches. However, beyond the second week, RUV indicates that decision-makers who are highly exposed to damages should use the reference climatology rather than the forecasts for the binary decision. This is not the case for the multi-categorical and continuous-flow decision however, where forecasts should be preferred."

to:

"A case study using a cost-loss economic model at Biggara in the Southern Murray-Darling Basin of Australia assessed the relative value of subseasonal streamflow forecasts over a fixed historical average reference climatology. This case study demonstrates that the forecasts should be preferred over the reference climatology forecast for all lead-times studied (max 30 days) and almost all users regardless of their risk aversion. This positive forecast value is robust to changes in user risk aversion, decision types (binary, multi-categorical, and continuous-flow), and decision-making approaches. However, the results indicate that users who are highly exposed to damages would gain more value using the reference climatology rather than forecasts for the binary decision in lead-time weeks 2-4. This was not the case for the multi-categorical and continuous-flow decision however, where the forecasts should be preferred. As REV is limited to binary decisions, a user making a multi-categorical or continuous-flow decision, could be misled by the REV outcomes and consider not using the forecasts when they actually have significant value as demonstrated by RUV."

2.4. a) Another general comment is about the fact that we set the context of the paper on probabilistic subseasonal forecasts (up to 30 days), but much of the demonstrations and experiments refer to 1-7 day lead-time forecasts (and many concluding sentences seem to forget this context and generalize to any type of forecast and lead time).

Thank you for highlighting this. During preliminary investigations we did generate results for other lead-times and groupings. The specific case-study features used (as described in section 4) were selected to best present the salient features of RUV and the case study results. This is described on lines 311-313. On reflection, we agree with the reviewer that some concluding statements have been generalized beyond the experimental results shown in Section 5. We will rectify this when re-writing the conclusions. Please see related responses to comments 2.2 and 2.3.

b) In many situations (but I am not sure about the case of the particular catchment of the study), a meteorological (model-based) forecast may show quality a couple of days ahead (1 to 5 days, for instance) and then be as skilful as climatology afterwards. How this difference in the quality of the forecasts might affect the results here? Is it justified to group together these lead times here?

This is an interesting point. We agree that rainfall forecasts are only skilful for short lead times (e.g. 1-5 days). However, streamflow forecasts are typically skilful at longer lead-times than rainfall forecasts due to storage effects. In particular, the subseasonal streamflow forecasts used in this case study have previously been shown to be sharper and had higher CRPS skill scores than climatology for lead times up to 30 days (see McInerney et al. (2020) for a detailed explanation and evaluation of this. Additionally,

Figure 6 demonstrates that these forecasts have higher value than climatology for longer lead-times.

We found that grouping lead times together in our analysis was beneficial in addressing the aims of this paper, namely to introduce RUV and contrast it with REV. We also found that the particular groups (1 week, 2 weeks, 3-4 weeks) demonstrated key differences between lead times.

However, grouping lead-times would not be recommended in a practical context. If the purpose of this case study was to quantify value of using forecasts to inform river operations, then we would need to analyse forecast value of individual lead-times. We will acknowledge that grouping lead-times is not recommended for practical applications in section 4.7 when we introduce the groupings. Please see our response to a related comment (1.31) from reviewer 1.

c) Would a (potential) difference in quality explain negative RUV (lines 412-414), where the authors say that climatology (as a forecast) is more useful than a (meteorological model based) forecast?

We do not believe that the differences in forecast performance across the 1st week are the root cause of negative value regions discussed on line 412-413. The negative regions are a consequence of using a fixed critical probability threshold, as explained in section 5.1 and the REV literature, see cited references Richardson (2000) and Murphy (1977). This result is seen for any forecast regardless of whether lead-times are grouped or not. We will add an additional sentence at line 417 stating the generality of this result with reference to the REV literature.

d) (note: at the end, the decision maker is always using a forecast, either from a record of historic observations – climatology – or from a coupled atmospheric-hydrologic model).

This is of course true. However typically the term "forecast" refers to a procedure more complex than "just" the marginal distribution of historical data.

Thank you for bringing this need for more clarification to our attention. We agree that the term "forecast" typically refers to a reasonably complex procedure. As this study involves a comparison of the new method RUV to the existing method REV we are limited to using the baseline reference used by the REV. This critical limitation of REV is stated on lines 65-66 of the Introduction and is a key motivation for the development of RUV. We will include the need to apply RUV with "practically relevant reference forecasts" in Section 6.3. Please see our response to a related comment (1.32) from reviewer 1.

2.5. Finally, a last overall comment I have is: why a systematic comparison with REV is so important in the development of a novel approach or metric in this topic? Is it because

REV is widely used (or supposedly widely used)? How crucial is it as motivation for the study?

The reviewer is correct that the comparison with REV is important because REV is commonly used. We feel this is an important motivation for the study, and make the following points about this in our paper:

- REV is widely used in the literature for quantifying the value of forecasts (see lines 54-61),
- The primary way to present this information is using a value diagram (see lines 89-90).
- As the value diagram is a compelling way to communicate value across a range of decision-makers, and the community is familiar with it, we felt it was important to be able to leverage this with any new approach (see lines 265-266).

Specific comments:

2.6. Introduction: I think the authors could introduce some literature on works done on forecast value and links between forecast quality and value with respect to inflows to hydropower reservoirs. These cover a large range of cases and lead times, and also use optimisation-based economic models to link forecast production (quality) to usefulness (economic value). It would be interesting to give here this broader view to the topic, I think, and then replace better the context of the paper (to which the conclusions drawn will specifically apply). Besides the paper mentioned in the discussion (Penuela et al.), some others that might be interesting are: https://doi.org/10.1002/2015WR017864; https://hess.copernicus.org/articles/23/2735/2019/; https://doi.org/10.1029/2019WR025280; https://hess.copernicus.org/articles/25/1033/2021/.

Thank you for this suggestion and for providing these references. Originally, we did not want to add context to the introduction beyond forecast value and the case study, and instead left broader context to the discussion. Your reasoning has convinced us to include some text on the links between forecast quality and value, and the optimisation-based modelling in hydropower. Thanks

2.7. Line 49: too many “and” words. Please, check.

We do not see any “and” words on line 49 and feel the sentence reads fine.

2.8. Line 50: “better verification implies more value”: I think you refer to “quality” and not “verification”. Please, check.

Correct. Thanks for catching this.

2.9. Line 88-89: not clear to me. Please, check.

Thanks. We will rewrite this sentence for clarity.

2.10. Line 90, 102: when you refer to "the authors" I am sometimes a bit confused if you mean "you" or the authors in Matte et al. Please, check.

Thanks for noting this. We will replace "the authors" with "we" in these two sentences.

2.11. Line 192-193: maybe it is not reported in scientific papers, but are you sure it is not commonly used by water managers in practice? Have you conducted a survey or any other study not reported here to assess it (i.e., real-world practices)?

Good point. This statement was based on our knowledge of the scientific literature and 15 years professional experience providing forecasts to water managers in Australia. However, we recognise that this statement was too broad, and will soften the language accordingly. In particular, we will mention that, to the best of our knowledge, making real-work decisions with $p_t = \alpha$ has not been reported in the published literature.

2.12. Line 227-230: again too many "and" words. I found the sentence unclear. Please, check (maybe also correct to "a specific decision").

Thanks for spotting this. We will remove the unnecessary "and" from line 227. We assume that you are referring to the sentence on lines 228-230 as being unclear. We agree and will split this into two sentences, as well as adding a little more context.

2.13. Line 280: I am not fully convinced that information on amount spent, damage etc. at each time step is something valuable to a user. Is that so? Can you provide examples or a justification for that? I believe that users might be more interested in the long-term performance of a forecast system (in particular when it comes to reservoir operations), while a flood alert user would be interested in the whole flood event duration performance (and less on each time step). Maybe I misunderstood something here.

Thank you for highlighting this.

We will provide an example of users who may find this additional information useful.

One example is a user applying alternative economic models or tuning damage functions to match real-world data. This user would require the amount spent and damages incurred at individual time steps to determine that the model is behaving as expected.

A second example is a decision-maker who has finite funds to spend on the mitigation of damages. Understanding when the maximum spend amount exceeds the funds available would require investigation of spend and damage amounts at individual time-steps.

2.14. Line 309: I do not think "Methodology" is a good title for the section. I would suggest "Application" or "Experiment".

We agree with this suggestion, and will change this heading to better describe the content.

2.15. Line 310-311: I guess that by "different decision-makers" you mean "different levels of aversion of decision-makers". I think it is not the person themselves you are talking about but the theoretical level of aversion that you are modifying in the experiments.

Correct. We are referring to the level of risk aversion of an individual decision-maker, and their exposure to damages (α). We will reword "decision-makers" to "decision-makers with different exposure to damages and different levels of risk aversion".

Thank you for bringing this to our attention.

2.16. Section 4.1: I think part of it could go to the Introduction.

Thank you for this suggestion. We agree that some of this material could be included in the Introduction. The intent for Section 4.1 is to provide background and motivation which is specific to the case study application introduced in Section 4. Therefore, we will move lines 315-320 to the Introduction and adjust the remaining sentences for clarity.

2.17. Line 339: maybe place the references in the right place would help the reader (ex. Perrin et al., after GR4J, and not after RRP-S).

Thank you for this suggestion. We will move the references from the end of the sentence and place after each modelling component.

2.18. Line 343: “seamless” has usually another meaning in the literature. It usually refers to a system that forecasts in a coherent and homogeneous way from minutes to hours and months. It is not usually related to performance across scales. Please, check.

In this context, “seamless” refers to forecasts which are coherent and homogenous and have similar forecast quality across time scales. We will rephrase the sentence on line 343-344 sentence to clarify how we are using the term.

2.19. Section 4.4: I think part of it could go to the Introduction (lines 346-354).

Thank you for this suggestion. We agree that the Introduction does not provide enough context or motivation for the different decision types. While lines 53-54 briefly introduce the decision-types we feel that having more context up front would help the reader and will add additional sentences to the Introduction based Section 4.4 (lines 346-354).

2.20. Line 369: what do you mean by “suitable”? How? Based on data?

Good point. We did not define what we meant by “suitable”. We will clarify that it reproduced the real-world assumptions described in the previous sentence.

2.21. Table 3, experiment 4: check typo

Well spotted. We will remove the “of” typo

2.22. Fig. 4: I am not sure it is needed to show that we come up to the same results.

Although the value diagrams in panel (a) and (b) are identical, they are not the same thing. They are the same outcome from two different methods. We feel it is important to show this equivalence in an explicit way as it supports the research aims of this study. However, on reflection we can see how it may have caused confusion and will change the caption of panel (b) to "RUV with restrictive assumptions equivalent to REV" to make the difference clearer.

2.23. I would suggest putting Experiment 1 and Experiment 2 together.

Thank you for this suggestion.

Although the results shown in Figures 4 and 5 are similar, they are different experiments with different reasoning and explanations. We prefer to keep the two experiments distinct for clarity and introduce the content in a staged way. We agree that combining the results

together as a single figure would be more efficient in terms of space; however, it would be more difficult to explain and describe the differences. By keeping the results separate it is easier for the reader to understand the reasons for the similarity and differences between REV and RUV, which is a research aim of the paper.

2.24. Line 437: what do you mean by “ensemble sampling error”? Please, explain.

Thank you for noting this. We were referring to errors due to the small ensemble size. We will replace "ensemble sampling error" with "sampling errors due to small ensemble size".

2.25. Line 458: please, clarify the sentence (see my general comments above) in terms of saying that a “decision-maker should avoid using forecasts” in certain conditions.

Thanks for pointing this out. This point is discussed in our response to reviewer comment 2.3.

2.26. Line 464-466: Does this correspond to reality? Have you discussed the results with the Murray-Darling Basin managers, for instance? It would be interesting to link mathematical calculations to reality in the field, providing supporting to some sentences on the results and overall conclusions drawn in the paper.

We agree that it would be interesting to make a strong link between forecast value methods such as RUV and applications "in the field".

As pointed out by reviewer 1, this paper is not aimed at providing a detailed and conclusive evaluation of forecast value for Murray Darling Basin managers. It is aimed at introducing RUV and contrasting it to REV by demonstrating that RUV can handle factors which are important to real-world decision making. This is demonstrated through experiments showing the conditions under which RUV and REV are equivalent and illustrating the sensitivity of forecast value to different decision types (binary, multi-categorical, and continuous flow) and levels of risk aversion. For the case study in this paper, we have made reasonable assumptions about the damage functions and decision thresholds, and necessarily used the cost-loss economic model for comparison with REV. We feel this is sufficient for the purposes of this paper, which is to demonstrate that forecast value is sensitive to decision-type and levels of risk aversion.

Now that we have established forecast value is sensitive to these choices, in the future we can undertake a realistic evaluation of forecast value for Murray Darling Basin managers using RUV. We will add this topic to the future work section, mentioning the need to “calibrate” the damage function, decision thresholds, and economic model based on the real-world experience of decision-makers.

2.27. Fig. 7: I think it should be more commented. The differences we see in the column on the right do not seem to be “moderate”.

We will expand the commentary on lines 471-473 and lines 486-487 to ensure it is clear that the impact is “moderate”, except for the case of highly risk aversion decision-makers with continuous flow. Thanks for pointing this out.

2.28. Experiment 5: could you justify the choice of adopting a binary decision and alpha = 0.2 here? Also, why are you showing week 1 if the focus of the paper is on longer-term forecasts?

Good point. The choices to use a binary decision and α =0.2 are illustrative to simplify the explanation of a complicated idea. Similar results were found when we conducted this experiment with different values of α and forecast lead-time, and additionally with multi-categorical decisions with different numbers of flow classes. This is noted on line 500-503 however we inadvertently forgot to mention forecast lead-times which will add to this sentence. We will also add an additional sentence on line 500 to clearly state that the chosen values (i.e. binary decision, α =0.2, 1st week of lead-time) were selected to simplify the explanation.

2.29. Line 510-511: is this a general conclusion? Over any lead time and situation? Not all probabilistic streamflow forecasts are skilful and reliable. Do you mean for the case study of the paper? Please, clarify.

Thanks for spotting this. We will make this sentence specifically about the forecasts used in this case study and add an additional sentence on the availability of streamflow post-processing methods to improve skill and reliability of raw forecasts.

2.30. Lines 513 and 514: I suggest using “developed” and “can be applied”.

Well spotted. We will correct this.

2.31. Line 520: Please check deleting “is”.

We will fix this typo.

2.32. Line 553: I do not understand what you mean by “a single forecast user” (single forecast or single user)? Please, clarify. Also “they” here refers to whom? The users?

Thank you for noting this. We are referring to a single decision-maker and will therefore replace the "forecast user” with "decision-maker". We will also replace “they” with “the decision-maker” so the sentence is clear.

2.33. Line 569: by “mitigation” do you mean “real time mitigation of damages”? Sometimes mitigation is more related to “prevention” (out of real time) for some users. Please, clarify.

Thanks for raising this important point. We will add an additional sentence explaining these two different types of “mitigation”. In the context of a cost-loss economic model, mitigation refers to “preventive” action taken ahead of time. This is therefore what is meant by mitigation in REV and in our application of RUV. However, as RUV is general purpose and any economic model can be used, so it could in principle consider either of these types of mitigation. Exploration of this dynamic decision-making process over lead-times and forecast updates is left for future work and will be noted in section 6.3.

2.34. Section 6.3: I suggest using “could” instead of “will” when talking about possible future pathways for further research/future works.

Thank you. We will change this.

2.35. Overall: please check the use (or the absence) of a comma before the word “which”.

Sure, we will check this.

2.36. Figures/tables: overall, please check the use of colours in black and white printing (maybe use italics in Table 3 instead of red, for instance; use dotted lines instead of colours in other figures, etc.)

Thank you for this suggestion. We will use italic and red in table 3 and will add additional line styles to improve ease of reading the figure. In addition, we will ensure the colour choices are colour blind friendly.

References McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. *Water Resources Research*, *56*(11). https://doi.org/10.1029/2019wr026979

Murphy, A. H. (1977). The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation. *Monthly Weather Review*, *105*(7), 803-816. https://doi.org/10.1175/1520-0493(1977)105<0803:tvocca>2.0.co;2

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649-667. https://doi.org/10.1002/qj.49712656313