Response to comments from Reviewer 1

Summary

1.1. The manuscript by Laugesen et al. introduces a new metric to assess forecast value adapting the formulation of a previously existing metric, namely Relative Economic Value, within a flexible value assessment framework based on utility. The method is then exemplified with subseasonal forecasts in the case of the Murray River, Australia, where decisions tend to target high flow values. A sensitivity analysis is carried in this case study.

The paper, which proposes a new methodology and results of high significance for the forecasting community, is detailed, very didactic and of high quality, and will undeniably be valuable to researchers who wish to carry out advanced and flexible forecast value analyses, involving decision-makers' levels of risk aversion.

I strongly recommend this paper for publication and list hereafter recommendations for clarification, as well as some minor points and typos.

We thank Reviewer 1 for their encouraging feedback and detailed review of our paper. In particular, we appreciate their thoughtful suggestions for improving the clarification of certain sections, which will make this material easier to follow.

Comments

1.2. L18-21: These two sentences seem a bit contradicting because you first announce value for all lead times, decision types and most levels of risk aversion, but then you nuance your statement beyond the second week, for binary decisions. I suggest nuancing the first statement.

We agree that these two sentences appear contradictory, and will adjust our wording to rectify this problem.

1.3. In addition, the case of the Murray-Darling basin being an example of application for sensitivity analysis rather than a stand-alone evaluation, I would consider these results as secondary compared to the advantages of the proposed RUV metric and the results of the sensitivity analysis well described in Section 6.2, which in themselves deserve to be highlighted in the abstract.

Good point. We agree that the Murray Darling experiments are an example of a sensitivity analysis, more so than an evaluation of forecasts, and will revise our text to reflect this. In particular, we will modify the abstract to highlight the outcomes of the sensitivity analysis in section 6.2, choose a new term for the "case study", and make the "case study" results secondary.

1.4. L20: "Beyond the second week" please mention that you are referring the lead time.

Good suggestion. We will implement this clarification to avoid potential confusion.

1.5. L26 (and throughout the paper): Here authors refer to the lens of "consumer" impact. The terms "user" and "decision-maker" are also used throughout the paper. Given that there are differences between these terms, I wonder whether the authors could clarify whether they use these three terms as interchangeable, or do they make a distinction. In the former case, are they actually interchangeable? Forecast datasets are increasingly open, and I am not sure whether users are indeed consumers in these cases. In the latter case, could you explicit the distinction made in an evaluation context?

This is a good point. We agree that these terms were not used consistently and will correct this by using the term "user" through the paper. We feel it is important to be clear that the "user" in this context is an individual making a decision and therefore we will replace "decision-makers" on line 11 with "decision-makers (users)".

1.6. L73-78: Based on these two examples, and purely intuitively, I would tend to consider both types of decision-makers to be risk averse (conservative approach to avoid spending in example 1 and flooding in example 2) but with a different sensitivity to forecast uncertainty. Could the authors elaborate on why they make a direct link between forecast uncertainty and risk aversion?

Thank you for raising this important point of clarification. There is potential for confusion, as the formal definition of "risk aversion" does not always align with the colloquial interpretation used in daily life. In our study we use the term "risk aversion" from the economic literature, as defined on lines 73-74. However, we see that this definition is not entirely clear, and we will improve this with specific mention to the relationship to uncertainty and cite (Mas-Colell, 1995). We will also remove example 2 from the paper, as we now appreciate that it is unnecessary and complicates the introduction of the "risk aversion" concept.

1.7. L95: Maybe reformulate "lead to improved forecast verification". For instance: "lead to improved forecast verification indicators" or "improved forecast performance".

Good suggestion. We will change this wording.

1.8. L98: "first convert them"

Thanks. We will correct this.

1.9. L125: Isn't it a 2x2 contingency matrix?

Yes it is. We will fix this typo.

1.10. L133: The term "outcome" was unclear to me here. I was unsure whether it referred to each combination of possible Action/Event in Table 1. In my understanding, E depends on each information source (reference, forecast, or perfect) but uses all possible outcomes in its weighted mean. The term "outcome" was a bit confusing, while Equation 1 and L138 were perfectly clear. Since the Supplement helped in that matter, I would suggest referring it here already.

Thanks. We will add a sentence defining "outcome", and also refer to the supplement.

1.11. Equation 1: Could you please add the range within which V should fall (-∞ to 1)?

Sure, that is a great suggestion.

1.12. Equation 2: At this stage o is not defined.

Thanks for catching this oversight. We will define $\bar{o}$ as the frequency of the binary decision event (as on line 687) in the sentence following equation 2.

1.13. Figure 1: The location of the phrase "Use reference to decide" is, I think, misleading. Based on the explanations (L162-164), it seems that for a cost-loss ratio of 0.5, for instance, the forecast outperforms climatology and should thus be used to decide, with a potential REV reaching about 0.8. Therefore, using the reference for a cost-loss ratio of 0.5 would not allow reaching a REV greater than that of the forecast. However, based on the figure, it seems that using the reference for a cost-loss ratio of 0.5 would allow reaching a REV greater than that of the forecast. Maybe the arrows pointing at the extreme intervals when the reference is indeed performing better, but this is currently not clear.

We see how the position of the "use reference to decide" text could introduce confusion. Our intent was as follows, the sentence "use reference to decide" is part of the "always act" and "never act" arrows, as in "use reference to decide and always act" and "use reference to decide and never act".

We will replace the annotations on the figure to clearly label the different regions of the value diagram using (a), (b), and (c) and add concise text on how to interpret each:

(a) $0<=\alpha<0.05$ – use reference to decide, and always take mitigating action

(b) $0.05<=\alpha<0.95$ – use forecast to decide whether to take mitigating action

(c) 0.95<=α<1 – use reference to decide, and never take mitigating action

We believe that suggested revisions to the figure will make this clear.

1.14. Additionally, it is not clear whether the arrows linked to "Always act" and "Never act" point at the interval when climatology < forecast or at the specific points (0;0) and (0;1) (see also the following response to comment 1.15).

Thanks for highlighting this. Our intention was for the arrows to referring to the intervals rather than (0,0) and (0,1). We will adjust the figure to make this clearer.

1.15. Figure 1 (and all value diagrams): If I understand correctly the meaning of α=1 (never worth acting) and α=0 (always worth acting), the decision can be taken regardless of whether the forecast or climatological information is considered. This would mean that the relative economic value should be exactly equal to 0 in both cases (α=1 and α=0). If that is correct, and that no other parameter comes into the decision of acting or not, is there a reason why the two points (0;0) and (0;1) are not represented in the value diagram?

Thanks for this comment. The above interpretation of the cases α=1 and α=0 is not quite correct. There is no conceptual reason why the relative value should be zero at these end points as this would imply that the forecast and reference climatology are both equally valuable, which is unlikely. In our illustrative example the reference climatology is more valuable.

REV uses a fixed average value for the reference climatology and there are only two possible actions:

1. Always act (when $\alpha < \bar{o}$) or
2. Never act (when $\alpha >= \bar{o}$)

(where $\bar{o}$ is the observed event frequency).

This is detailed in the derivation on lines 30-34 of the supplement.

A decision-maker *should* use climatology to make decisions when climatology is more valuable than forecasts, and therefore they will use one of these two options if their α value lies in intervals the arrows are pointing at in our illustrative example.

We will add a reference to the derivation in the supplement, and cite (Richardson, 2000) which introduces REV, the value diagram, and its interpretation.

1.16. Equation 4 (and Equation 9): Probabilities being sometimes used with powers, I would suggest to place the index m as a subscript rather than superscript.

Thanks for pointing this out. We will use subscripts for the *m* index on probabilities rather than superscripts.

1.17. L207-217: I suggest adding an example graph of μ to illustrate your explanation. For instance, I find it hard to picture the concavity of μ, especially in the case of binary decisions.

Thank you for this suggestion. We will consider the benefit of adding a second panel to figure 3 with a graph of μ for the 4 values of A used in the study. We will also add an explanatory sentence to the risk aversion section 4.6.

1.18. L230: "the absolute value of a specific decision"

Thank you for picking this up. We will fix this typo.

1.19. Equation 6: Here it is not clear to me why damage does not vary with time (in Appendix A it seems it does). It is also not clear why m, whom E, b and d depend on, appear in parenthesis in the case of b and d, and as a subscript in the case of E.

We agree that there is inconsistency in our notation, especially with the way we are indexing t and m.

The reviewer is correct that observed damages vary with time, as the realised state of the world associated with each observation changes with time. We will make this clear in our revised notation for equation 6, and across the whole paper, appendix, and supplement.

1.20. L237 "The damage function relates the streamflow magnitude to the economic damages": At this stage, you have not mentioned streamflow yet, I would suggest sticking to the term "states of the world".

We agree that would be clearer. Thanks!

1.21. L309: The previous section also comprised elements of methodology. Consider changing the name of this section.

Good suggestion. We will change this section title to a more appropriate term that describes the content.

1.22. Section 4.2: Could you briefly state why you chose this station and basin?

Good idea. We will explain that this site is significant for water resource management because it is upstream of a major water storage.

1.23. To which extent do you expect your results (sensitivities) to differ in a catchment with different hydrometeorological characteristics?

We will mention that the results are likely to be sensitive to the flow characteristics and forecast uncertainty and that other sites will be analysed in future work.

1.24. L340-341: Given that you mention a rainfall post-processing step, I would recommend stating "raw streamflow forecasts" (L340) and "the streamflow observations" (L341) to avoid any misunderstanding.

Thanks. This change will avoid potential confusion.

1.25. Section 4.3: GR4J also uses temperature or potential evapotranspiration as input. Could you say something about what you used?

Good catch. We will add that PET from the AWAP model has been used.

1.26. L345-346 "flow exceeding the height of a levee": it would be more intuitive to talk about the "water level exceeding the height of a levee"

We agree, and will make that change. Thanks.

1.27. L374: "all decision-makers share the same level of risk."

Thanks. We will fix this typo.

1.28. Table 3: (1) "Experiment 4: Impact of risk aversion on forecast value"; (2) In experiment 5, the decision thresholds says "All flow" but the decision type is "Binary", which is counter-intuitive. "All possible thresholds" might be easier to understand, or "Thresholds from bottom 2% to top 0.04%".

Good suggestion. This will make it easier to understand.

1.29. Figure 4: Here you consider two rather extreme yet probably realistic thresholds for converting the probabilistic forecast into a deterministic one. When reading the results, I was wondering whether moderate thresholds could alleviate the lack of forecast value for

high and low cost-loss ratios and provide reasonable value for all cost-loss ratios. Could you answer this by displaying intermediate probability thresholds in this experiment?

Thank you for suggesting this. We will add an additional intermediate probability threshold to further illustrate this issue.

1.30. Figure 6: To ease the reading of this figure whose lines are plain and with colors of similar intensity, I suggest adding dashes and dots to distinguish the three curves.

Good suggestion. We will add additional line styles to make it easier to read the figure and ensure the colour choices are colour blind friendly.

1.31. Figure 6: Could the authors explain the interesting difference in RUV pattern for the multi-categorical decision (also seen in other decision types) between lead week 2 and lead weeks 3 and 4? Why does value decrease with lead time for low cost-loss ratios (as expected) but increases with lead time (maybe less obvious) for high cost-loss ratios?

Thank you for bringing this pattern to our attention. While the differences between weeks 2 and 3/4 are minor they interestingly appear robust to decision-type. We will provide some possible reasons for these differences, namely lead-time dependent differences in the post-processor correction of forecast errors in low/high flow regimes, and decreasing sharpness of the forecast ensemble at longer lead-times. A definitive explanation would require a dedicated experiment and will be sought in future work that focuses on a decision-maker application and/or additional forecast locations. We will mention this future research direction explicitly in Section 6.3 on future work.

1.32. Experiment 3: In this experiment, authors look at the variation of value with the lead week. It is also common to look at the influence of the initialization month or season to appreciate the influence of different hydrological conditions on the value. Even though it would mean dividing the total forecast sample into subgroups and reducing significance, I think it could be a valuable addition to Figure 6 to show forecasts initialized in dry and wet conditions separately.

Yes, this is a great point and we agree that assessing the impact of seasonality and antecedent conditions on forecast value is an important research question. While important, we believe this assessment does not fit is not within the research aims of the current paper, which focus on the introduction of RUV and a comparison to REV.

The Biggara case study is used as a vehicle to illustrate the general application of the RUV method. The current set of case study analyses is already quite extensive and includes results from 5 experiments illustrated through 8 figures. To sufficiently assess the impact of different hydrological conditions on forecast value we would need to add an additional

experiment with dedicated figures and text to explain the impact of the seasonality and antecedent conditions with respect to the decision-types and lead-times. Further, for completeness the impact should also be assessed on the risk aversion outcomes through additional evaluation in experiments 4 and 5. We feel that this additional content would distract from the main aims of the paper and result in an unduly lengthy manuscript.

After some reflection, we feel this interesting and important research question requires a dedicated separate study and is best left for future work. Thank you for raising this important issue for practical application. We will add text to Section 6.3 that addresses this and outlines further work is needed.

1.33. Figure 7: To ease reading, consider adding a horizontal line at y=0 in graphs displaying the overspend.

This is a good suggestion. We will do this.

1.34. Experiment 4: It is currently unclear why the third line of Figure 7 is shown as it is little to not exploited in the interpretation. Please consider removing or spending some sentences to exploit this line of the figure.

Thank you for this suggestion. We assume you are referring to the third row of panels rather than third line. One intent of including the utility-difference (and overspend) results was to enable a more direct comparison of our findings with those in Matte et al. (2017), line 469-471. We agree that adding an interpretation of the utility-difference results would add value and thank the reviewer for pointing out this oversight. We will add additional text interpreting the utility-difference results at line 485.

1.35. L520: “making decisions with fixed critical probability thresholds leads to”

Good catch. We will fix this.

1.36. Sections 6.1, 6.2 and 6.3: Numbering the paragraphs is unnecessary.

Thank you for highlighting this. We will remove the paragraph numbering.

1.37. L576: “summarizes”/“summarises”

We will correct this Australian/US language issue, and check the rest of the document.

1.38. L680 and Table 5: In the text, you mention that the formulation of Ct depends on the value of p, but in Table 5, the formulation of Ct depends on whether the action is taken or not rather than on p. I could not figure out why. Are the p values you are referring to in both instances different? Could you please clarify this point?

Thank you for pointing this out. We agree that it is not clear from the text in the appendix. Section 2 in the supplement includes a more complete derivation but on reflection that is also lacking clarity.

The probability on line 680 is referring to the forecast probability of flow above the threshold, which then determines ex ante (i.e. before event has taken place) how much to spend on the action C_t^{π}. The probability in Table 5 is referring to the ex post probability that the observed flow is above the threshold. The amounts spent on action or no action in the row titles of Table 5 are the optimal costs C_t^{π} found ex ante (i.e. after event has taken place).

We will make the following changes to the main text and the supplement to improve the clarity of this derivation:

- On line 679 replace "probability is always 1 or 0" with "forecast probability is always 1 or 0"
- Change the column titles of Table 4 and Table S2 in the supplement to "Event forecast to occur" and "Event forecast to not occur".
- Change the column titles of Table 5 and Table S3 in the supplement to "Event occurred" and "Event did not occur".
- On line 673 replace "forecast probability is 1 or 0" with "event is forecast to occur (p=1) or not occur (p=0).
- On line 118 of the supplement replace "letting the probability be conditioned on observed flow above the threshold" with "letting the probability be conditioned on observed flow above the threshold, rather than the forecast flow used for the ex ante utility".

1.39. L701: The link to the companion dataset is missing.

Thanks for pointing this out. We will include a permanent link to the companion dataset in the revised manuscript.

References

Mas-Colell, A. (1995). *Microeconomic theory*. Oxford University Press.

Matte, S., Boucher, M.-A., Boucher, V., & Fortier Filion, T.-C. (2017). Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker. *Hydrology and Earth System Sciences*, *21*(6), 2967-2986. https://doi.org/10.5194/hess-21-2967-2017

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