Comment on hess-2022-65

Anonymous Referee #1

Referee comment on "Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts" by Richard Laugesen et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-65-RC1, 2022

Summary

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.

Comments

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. 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.

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

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?

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?

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

L98: "first convert them"

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

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.

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

Equation 2: At this stage o is not defined.

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.

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 comment).

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?

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.

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.

L230: "the absolute value of a specific decision"

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.

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".

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

Section 4.2: Could you briefly state why you chose this station and basin? To which extent do you expect your results (sensitivities) to differ in a catchment with different hydrometeorological characteristics?

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.

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

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"

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

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%".

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?

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.

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?

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.

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

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.

L520: "making decisions with fixed critical probability thresholds leads to"

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

L576: "summarizes"/"summarises"

L680 and Table 5: In the text, you mention that the formulation of C_t depends on the value of p, but in Table 5, the formulation of C_t 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?

L701: The link to the companion dataset is missing.