Comment on hess-2022-65

Anonymous Referee #2

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-RC2, 2022

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.

General comments:

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.

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.

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). 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? 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? (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).

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?

Specific comments:

- 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/.

- Line 49: too many "and" words. Please, check.

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

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

- 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.

- 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)?

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

- 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.

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

- 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.

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

- 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).

- 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.

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

- Line 369: what do you mean by "suitable"? How? Based on data?

- Table 3, experiment 4: check typo

- Fig. 4: I am not sure it is needed to show that we come up to the same results. I would suggest putting Experiment 1 and Experiment 2 together.

- Line 437: what do you mean by "ensemble sampling error"? Please, explain.

- 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.

- 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.

- 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".

- 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?

- 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.

- Lines 513 and 514: I suggest using "developed" and "can be applied".

- Line 520: Please check deleting "is".

- 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?

- 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.

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

- Overall: please check the use (or the absence) of a comma before the word "which".

- 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.)