The research aims to understand and improve the important challenge of communicating three-dimensional flood map uncertainty to various end-users through a series of qualitative surveys. The manuscript is well written and structured, there are several wordy tables that could be presented differently (see suggestions below). The probabilistic visualisation prototypes presented represent a significant step-forward in terms of communicating uncertainty to forecast end-users. The work would benefit from more emphasis on which uncertainties are being represented in the visualisations and how the prior knowledge of the surveyed participants is assessed and how this impacts their opinions and the conclusions drawn from the results. The research questions proposed in the introduction are reasonable, they should be re-addressed again in conclusion. Consideration of the limitations of the survey approach and applications of this approach outside of Quebec would enhance the manuscript. Once these concerns are addressed, I feel that the article would make a valuable contribution to HESS.

Thank you for taking the time to review our manuscript. Your comments and suggestions are much appreciated and will help us to prepare an improved revised version of the manuscript.

More specifically:

1. Where do the uncertainties originate from? Are they based on uncertainties in the precipitation inputs to the hydrological model? Or are they uncertainties relating to model parameters/antecedent conditions/underlying data used to determine the flood maps such as the DTM? Or are they compound and include all the above? The paper would benefit from some discussion of these aspects of uncertainty in relation to the flood forecasting system used.

We will add a short sub-section in the methodology section to provide more information about the operational forecasting chain. In short, deterministic streamflow forecasts are obtained by feeding Hydrotel (a distributed physics-based model) with deterministic meteorological forecasts (precipitation and temperature) from Environment and Climate change Canada. Then, the deterministic streamflow forecasts are dressed statistically using a method (« dressing ») that is explained in detail in Huard (2013). This method is based on an analysis of previous errors between forecasts and observations, and it varies according to season. This can be seen as post-processing, and encompasses many sources of uncertainty all at once. Also, there is significant manual data assimilation that is performed by the forecasters at the onset of the forecast. They apply perturbations to the most recent meteorological observations and re-run Hydrotel in simulation mode to obtain new state variables from which the forecast will start from. As for the hydraulics model (HEC-RAS), at the moment it is deterministic, in the sense that it has a fixed parameterisation, and the uncertainty is provided only through the probabilistic streamflow forecasts that are used as inputs to HEC-RAS.

Huard, D. (2013) Analyse et intégration d'un degré de confiance aux prévisions de débits en rivière, Tech. Rep., David Huard Solution, Quebec.

2. What determines that this is a large-scale survey? How does it compare to previous similar surveys in Canada or elsewhere?

The most commonly used criterion for estimating sample size in qualitative research is saturation. In studies involving focus group interviews like ours, the interviews are recorded, a verbatim is transcribed and then used to code all the information provided by the participants. Saturation is reached when no new information can be obtained by

conducting additional interviews (i.e., the new participants repeat information that was already provided by previous participants). In a recent literature review, Henrick et al (2022) came to the conclusion that in most studies (in medicine and various domains of social sciences), saturation was reached after a maximum of 17 focus groups. Another comparative study by Hagaman et al (2017) about the number of interviews needed to reach saturation show that larger sample size, between 20 to 40 interviews, were generally needed to reach saturation when a qualitative study aims at covering a large territory with potential cultural differences between participants. Interestingly, the study of Hagaman et al (2017) is based on a cross-cultural study on water issues. It involved 132 respondents in four different countries, but they found that saturation was reached after much less than 132 interviews. Note that the study of Hagaman et al (2017) was included in Henrick et al. (2022) literature review, and their sample size was by far the largest among the 23 qualitative studies they reviewed. The second largest had a sample size of only 60.

In our study, even though there are differences between the characteristics of the four groups of respondents, they all have a similar general background, with no prior experience with flood maps, and they also come from the same country. Therefore, it is a relatively homogenous group, and we also noted that saturation was reached early. We still maintained our initial plan of covering a large spatial territory, and therefore 139 participants is indeed considered a large number for that type of long (2-3 hours) focus groups. In the revised version of the manuscript, we will include further comparison with sample size from similar qualitative studies, to better support our affirmation.

Hennink M. and Kaiser B.N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests, Social Science and Medicine 292, 114523

Hagaman A.K. and Wutich A. (2017). How Many Interviews Are Enough to Identify Metathemes in

Multisited and Cross-cultural Research? Another Perspective on Guest, Bunce, and Johnson's (2006) Landmark Study, Field Methods, 29(1), 23-42.

3. How is the 'limiting the confusion of decision makers' (abstract L20) of end-users measured/known?

It was not measured. This sentence does not state a result of our study, but rather our intention when we designed the prototypes. We will rephrase this sentence to make it clearer.

4. Section 4.1.1 What was the prior experience of the participant groups at using and interpreting flood maps (probabilistic or otherwise). This seems to be critically linked to the users' preferences.

None of the participants had any experience with using and interpreting flood maps, because they did not exist in the study area previously. When specifically questioned about that, none of the participants mentioned using flood maps from other sources (a global model like GloFAS, for instance). We will clarify this point in section 4.1.1 in the revised version of the manuscript.

5. How were the visualisation prototypes developed, and by whom?

The visualisation prototypes were developed by the four coauthors together. At the beginning of the project, a literature review was performed by V. Jean and M-A Boucher, with the aim of identifying best practices for visualising probabilistic forecasts, in hydrology but also in other fields (e.g., hurricanes, forest fires, etc). On the one hand, this literature review confirmed that very little has been done on the specific topic of visualising probabilistic flood maps in hydrology, but it also provided some general guidelines, for instance regarding the choice of a colour scale (tones of blue and « traffic light » scales were often recommended). The discussions between the four coauthors started from those recommandations, and we designed the prototypes according to other elements we wanted to verify: the use of wording vs the use of numbers, different ways to phrase the validity time of the forecast, different ways to separate probability categories, but mostly, different ways of expressing the probabilities themselves. Once the concepts were agreed on, the prototypes were produced (in French) by the last author (D. Roussel) and his colleagues at the DEH. They were constructed using screenshots of HEC-RAS for the Jacques-Cartier River, modified in Microsoft Powerpoint. They were then translated in English for this manuscript by M-A Boucher.

6. Tables 2, 3 and 4 could be presented graphically to enable readers to visualise results and aid comparison. Tables 5, 6, 7, 8, 9 and 10 should be ordered/sectioned by participant group to improve readability.

Thank you very much for this nice suggestion. We will do our best to transform Tables 2 to 4 into figures and we will improve Tables 5 to 10 following your comments.

7. Are the survey findings applicable in other places/countries or should this type of survey be repeated elsewhere? Adding recommendations would be beneficial to readers.

Most findings are quite general and are applicable in any culturally similar context, for participants who are not familiar with flood forecast maps. We will add recommandations in the conclusion.

8. What are the limitations of this interview style survey approach? Could a quantitative survey be used to draw more specific conclusions such as linking prior experience/understanding to visualisation preferences? Also, how can the probabilistic forecasts be linked to impacts and with users' actions. The next step to this would be to link the likelihood of impact (or flow scenario from prototype 2) with appropriate actions. These points could be developed further in the discussion/ conclusions.

Yes, a quantitative survey could complement the interviews nicely. In particular, it could indeed provide quantitative data regarding users' prior experience, and could even be used to test (to some extent) their understanding of probability concepts. At the beginning of this research, we initially considered a more quantitative approach involving a survey sent and collected by mail. Such an approach would have involved a completely different methodology. In addition, because this is the first time that such a study about users preferences is performed in Quebec (and in Canada, in fact), we preferred to talk to the users in person (or through Zoom, because of the pandemic). We are considering a follow-up study, ideally once the flood maps become operational, and this follow up study could involve a more quantitative approach. We will develop those elements for future studies in the discussion/conclusion, as per your suggestion. At the end of each focus group, the participants were asked if they would be willing to take part in such a follow up study. They all answered positively.

Please see supplement for minor comments.

Thank you for taking the time to annotate the manuscript. All your suggestions will be included in the revised version of the manuscript. However, regarding Montreal Urban community, we will have to limit ourselves to a relatively brief explanation. Their decision to not participate in our study rests on political concerns more than scientific ones.