Dear authors,
One of the reviewers raised serous concerns such as the innovation of your study. Please take the comments from both reviewers seriously and decide if you will make revisions based on those comments. Thanks.
Regards,
yueping

Reviewer 1

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. We have prepared a revised version of the manuscript (and we apologise for the time it took!). We hope that we were able to address all your comments.

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 added a short sub-section in the methodology section to provide more information about the operational forecasting chain (lines 150-159):

« 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 based on an analysis of previous errors between forecasts and observations (Huard, 2013). This can be seen as post-processing, and encompasses many sources of uncertainty all at once. In addition, forecasters perform manual data assimilation 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 hydraulic component, it is based on the HEC-RAS model with a fixed parameterization (e.g. Manning coefficients). Consequently, the uncertainty that is accounted for by the current forecasting chain is strictly a result of the probabilistic streamflow forecasts used to feed HEC-RAS, and this uncertainty is estimated via the statistical dressing method of Huard (2013). »

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).

The following explanations were added in the methodology section (lines 179-194):

« In qualitative research, sample size can be determined by saturation. 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 their recent multidisciplinary literature review, Hennink and Kaiser (2022) concluded that saturation was generally reached after a maximum of 17 focus groups. Another comparative study by Hagaman and Wutich (2017) concluded that 20 to 40 interviews are generally needed to reach saturation, but this was for the case of qualitative studies covering large territories with potential cultural differences between participants. Interestingly, the study of Hagaman and Wutich (2017) is based on a cross-cultural research project on water-related 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 and Wutich (2017) is one of the 23 qualitative studies reviewed by Hennink and Kaiser (2022), and their sample size is by far the largest among the 23. The second largest study had a sample size of only 60. Similarly, the sample of Demeritt et al .(2010) includes only 50 respondents, spread across 17 European countries.

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. According to Hennink and Kaiser (2022), 140 participants is considered a very large number for that type of long (2-3 hours) interviews. Figure 2 summarizes the overall methodology. »

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.

Demeritt D., Nobert S., Cloke H., and Pappenberger F. (2010). Challenges in communicating and using ensembles in operational flood forecasting, Meteorological Applications, 17, 209-222

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 decided to simplify the sentence, which now reads (line 22-27):

«We propose several suggestions for visualizing probabilistic flood maps and also describe several potential adaptations for different categories of end users »

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. We added the following sentences in section 4.1.1 (lines 272-275):

« Operational forecasted flood map did not exist in the study territory at the time of conducting the interviews. When participants were specifically questioned about their use of such maps, they all declared no previous experience, including potential flood maps from other sources (a global model like GloFAS, for instance). However, most participants had previous experience with streamflow forecasts. »

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

The visualisation prototypes were developed by the four coauthors together. We added more details about the development process in section 4.2 (Lines 390-397):

« The four prototypes were developed using guidance from a literature review conducted jointly by the first and second authors (V. Jean and M-A Boucher). This literature review allowed to identify best practices for visualising probabilistic forecasts, in hydrology but also in other fields with more abundant literature on communication issues (e.g., hurricanes, forest fires, etc). For instance, regarding the choice of a colour map, tones of blue and « traffic light » scales were often recommended. The discussions between the four coauthors started from those recommendations, and we designed the prototypes according to other elements we wanted to verify: the choice of words, the use of numbers, different ways to separate probability categories, different ways of expressing the probabilities themselves, etc. 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 the second author (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 very good suggestion. Tables 2, 3 and 4 were transformed into pie charts (now figures 3,4 and 5). We have used the cividis colormap (Nuñez et al. 2018) available in the Matplotlib Python toolbox to ensure readability by a colorblind audience.

Tables 5 to 10 were modified following your comments. Unfortunately we have struggled quite a bit with LaTeX and we were not able to remove the enormous whitespaces that now appear in the tables. We would happily take any further suggestions.

Nuñez J.R., Anderson C.R. and Renslow R.S. (2018). Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data, PLoS ONE 13(7), https://doi.org/10.1371/journal.pone.0199239

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 applicable in any culturally similar context, for participants who are not familiar with flood forecast maps. We added recommandations in the conclusion (lines 627-634):

« Operational probabilistic forecasts of water depth and extent are only starting to be implemented and published worldwide. This study is the first one targeted at proposing and assessing visualization tools for that type of forecast. Even if our study took place in a specific geographical context (the province of Quebec, Canada), the questions that were asked to participants were general enough so that our findings are relevant and applicable in any culturally similar context, for citizens and decision makers that have never used flood forecast maps. Questions were exclusively about the visualization and communication of forecasts as well as their usefulness for decision making. Participants were not asked questions that would have been closely linked to their geographical location, such as questions about flood generating mechanisms, for instance.»

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. 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). Following your suggestion, the following was added in the conclusion (lines 654-658):

« This qualitative study could be nicely complemented in the future by a quantitative survey, especially after the new flood forecast maps have been available for some time. In fact, at the en of each interview, participants were asked if they would be willing to take part in a follow up survey or study and they all agreed. A quantitative study could be helpful to further explore the understanding of probabilities by different groups of users, but also to collect quantitative data regarding their experience with using hydrometeorological forecasts and how they use those forecasts in a variety of decision-making situations. »

9. Please see supplement for minor comments.

Thank you for taking the time to annotate the manuscript. All your suggestions have been reviewed and included in the revised version, except for the request for added details about why the Montreal Urban community wanted to be excluded from the project, for confidentiality issues. Their decision to not participate in our study rests on political concerns more than scientific ones.

Note that we have modified Figure 1 (the map) according to your suggestions, adding a North arrow, modifying the legend and adding latitudes and longitudes.

Regarding your comment about using a colour map that is readable by colour blind people, we have changed everything for the « cividis » colour map available in the Matplotlib toolbox in Python. This colour map was especially designed to be readable by colour blind people (see Nuñez et al. 2018)

Nuñez J.R., Anderson C.R. and Renslow R.S. (2018). Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data, PLoS ONE 13(7), https://doi.org/10.1371/journal.pone.0199239

Reviewer 2

This study investigates the communication of probabilistic hydrological forecasts with different types of users based on phone survey and qualitative elaboration. They show some interesting findings, for example, users' responses to uncertainty of forecasting results, similarities and differences in visualization preferences of different users, their curiosity in hydrological forecasting methods and so on. This study also shows us a blueprint of forecasting visualization schemes from a holistic view of water depth, inundation area, discharge and the uncertainty according to wide suggestions from the users' end. The paper is generally well-organized and the structure is clear. Such study can improve hydrological early warning systems, thus, benefit flood risk management.

Thank you very much for reviewing our manuscript and for your valuable comments and suggestions. We would however like to emphasise that the interviews were not conducted on the phone. They were conducted via the online platform Zoom, and this is only because of the restrictions due to the pandemic in 2020 and 2021. The initial plan was to conduct all interviews in person. We think it makes a difference to conduct the interviews on an online platform with video rather than on the phone, because the phone would have removed more of the non-verbal language of the respondents. In addition, the vast majority of interviews were group interviews, which would have been very difficult on the phone but easier with an online video platform.

However, I have several major concerns that expect to authors to address:

1. The innovation of this study needs to be further addressed (i.e., things that has not been done by previous study). In the introduction, the authors fully reviewed previous investigates on the communication of flood risks and highlight the importance of survey on probabilistic forecasts. However, the difference from or increment to previous studies is not clearly pointed out. For example, previous studies may only investigate communication of deterministic forecasts or 1-D/2-D hydrological forecasts instead of inundation map, etc. Besides, this study only survey people living in south Québec, where floods are mainly caused by snow melt. However, the situation may be different for other regions and countries. It remains known to what degree the conclusion drawn from this study can be transferred to and referenced by other places of Canada and the world.

Thank you for pointing this out. First, as explained in our response to comment #7 from Reviewer 1, the questions asked during the interview focussed on the visualisation and communication of information, and not at all on the flood-generating mechanisms. While it is true that most floods in Quebec are generated by snowmelt, it is not relevant for our study because here we are interested exclusively about the communication and visualisation of forecasts. Some participants mentioned their specific concerns about snow, but it was never directly asked to them. Our list of questions does not contain a single question about hydrological processes (snowmelt or other). If the study had taken place in Central America or elsewhere, we could have used the exact same list of questions. Please see answer to comment # 7 from Rev 1.

Regarding the innovation, at the beginning of the project, an extensive literature review was performed in order to obtain guidance to design the visualisation prototypes. It was clear from this literature review that our research is the first one focussed on the communication and visualisation of flood forecast maps, which is a novel and original contribution, with very important practical outcomes. To emphasise this, we added the following lines (lines 110-119) In the introduction:

« The above-cited research highlights the importance of conducting focused surveys of forecast users to target the optimal methods and choices for communicating and visualizing probabilistic information, but none of those studies specifically focus on flood maps, because producing flood forecast maps is an emerging practice. In fact, appart from Carr et al. (2016) and Carr et al. (2018, none of the studies cited above focused specifically of hydrological forecasts. A more abundant literature exist regarding the communication of probabilistic *weather* information, especially for extreme events like hurricanes. Even in the case of Carr et al. (2016) and Carr et al. (2018), the attention given specifically to flood forecast maps is marginal. It occupies a small portion of Carr et al. (2016), and is not studied in Carr et al. (2018), which rather focuses on the communication of more widespread ensemble streamflow forecasts. Our study is the first to concentrate exclusively on the communication of probabilistic flood forecast maps, which is an emerging product. It is a novel contribution which provides practical recommendations for the communication of this emerging type of forecasts to different groups of users. »

2. Survey should strictly take sample representativeness into account. The education background, gender and age of the participants and their living/working places may affect the results and the representativeness of samples. Thus, it will be essential to include statistics of these kind of information. For instance, a geographic distribution of the participants with flood risk map, proportion of people with/without hydrology or atmospheric education background, etc.

None of the participants had a hydrology or atmospheric science background. Participants were not asked to provide detailed information about themselves.

Appart from the citizens, participants were all nominated by their respective organization, with the mandate to represent this organization. They were also asked specifically to answer to reflect their organization's point of view, and not their's. Considering this, we did not ask any question regarding participants age, gender, education. Note that this is similar to Demeritt et al. (2010), amongst others. Retrospectively, we agree that it would have been relevant to ask this information from the citizens, but it was not done so we do not have this information.

3. I also notice that the authors design different contents of phone survey for farmers and citizens from non-farmers or citizens (i.e., drop "the themes related to the nature of the information" for farmers and citizens) but did not explain the reason for doing this too much. I think the different treatment may cause the readers wondering whether the forecast maps should **originally** be designed differently for these two kinds of users (i.e., farmers and citizens & non-farmer or citizens). Since satisfying all kind of users with a single forecast map seems to be impossible. Therefore, why did not the authors design different kind of forecast maps for them at first and then do the survey?

The content of the Zoom interview was almost the same between the different groups. Some questions were simply removed from the list for citizens and farmers because they were not applicable to them. Therefore, it is the same initial content, but slightly reduced for farmers and citizens. This was clarified in the revised version of the manuscript, which now reads (lines 258-259):

« In the case of the citizens, only the presentation of the prototypes was done, because the first part (results not presented here) was geared towards identifying which information is needed by organizations for their decision-making process. »

One of the elements we wanted to verify was if a single type of forecast map or visualisation tool would be sufficient to satisfy all kinds of users. It was part of the original mandate from our governmental partners. Therefore, we originally wanted to present all groups with the same prototypes. We also consider this approach to be more objective, in the sense that we initially provided everybody with the same information instead of taking decisions based on our own a priori for certain groups. The decision to not present Prototype 3 to the farmers and citizens came later, after the interviews with the ministries and municipalities, during which it was strongly recommended to omit it to avoid confusion.

4. The presentation is overall a bit too qualitative. Some quantitative descriptions and statistic plots are needed. For example, in L341-349, the authors can show the voting proportion of color scheme preferences with real numbers or a table or histogram. Table 7 offers too much unsorted information and words. Table 8-11 is the same without statistics and graph visualization.

Thank you for this comment, which is in agreement with comments # 6 and 8 from Reviewer 1, who also provided suggestions to transform Tables 2 to 4 into figures.

We have reorganised most tables to make them more orderly. As explained to Reviewer 1, we could not find how to remove the large vertical white spaces that are added by LaTeX when using the « wrap text » option to define the width of the columns. We are open to suggestions.

Note that our interviews followed a qualitative framework, similar to those of Demeritt et al. (2010) and Demeritt et al. (2012), for instance. Consequently, many quantitative informations were not measured. As mentioned in our answer to comment #8 from Reviewer 1, a quantitative study will complement this one nicely. In fact, after the interviews, some of the respondents were sent a quantitative survey on a follow-up topic. The results of that survey are outside of the scope of this manuscript and will be the object of another one.

Demeritt D., Nobert S., Cloke H.L. and Pappenberger F. (2012) The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management, Hydrological Processes, 27(1), 147-157.

Minor comments:

1. The structure of the abstract need to improve. The background occupies almost half of the abstract, leaving little space for results and main conclusions. The conclusion is the only one sentence with “several” statement (L19-20). And the significance of the study needs to be further stressed.

We have modified the abstract according to your comments and suggestions. We have tried to keep the context to a minimum, while emphasising the novelty and significance of our study.

2. Figure 1: The legend of the blue polygons and lines is needed. Also, please add coordinates for the map.

We have modified the map (Figure 1) according to your suggestions and those of Reviewer 1.

3. As mentioned in L112, the investigation of color scale is one of the objectives of this study, however, there is no echo in the discussion or conclusion section.

Thank you for pointing this out. The blue colour scale was preferred by a majority of participants. Therefore, we added the following recommandation at the end of the discussion (lines 619-620):

« Regarding the colour scale, the majority of participants preferred the blue scale, so we recommend its adoption. »

4. In the abstract and Figure 2, the number of the citizens and farmers are 37 in total, however, in Section 3.1.4, the author said 33 citizens plus 5 farmers. The numbers contradict. Please check. Besides, in L201, the number 11 is confusing.

The numbers have been corrected. As for the number (11) of focus groups for citizens, we have rephrased this sentence, which now reads (line 246):

«Focus groups were formed by dividing those 33 persons into 11 groups. »

5. In Section 3.2, the authors said “except for citizens and farmers, one-to-one interviews are taken for the participants”. However, in Figure 2, the interview number and respondents differ, which is confusing.

This is a mistake, thank you for pointing this out. Although the groups were small for ministries, municipalities and organizations, almost none of the interviews were one-to-one. This has been corrected.

6. The author should double check the upper and lower case of titles in the references. For example, the fifth and last reference in Page 29 use upper-case for the title, while others did not. The same problems can be found in Page 30.

This has been corrected and the format is now uniform.