We would like to thank the reviewer for the positive and constructive remarks. We provide here below a response to the main points discussed:

- Experiment Setup:
  - How conclusions could be impacted by the choice of the parameters for the experiment? e.g. protection cost or damage costs? This is a good point. We have mentioned a possible impact of these choices in the results section (P12L9-11). When designing the game, we have tried to balance initial money in purse, costs and number of rounds in a way that players could focus more on using the information provided than on money left in hand. However, we had expected that the willingness to pay (WTP) for a service would depend on how much money you have available. That is why we have examined the bids as a function of the money left in hand after round 1. The results are presented in Section 3.4 (page 12). They show that, in our configuration, the money left in hand did not influence significantly the WTP (P13L23-24). We believe that if the damage costs were higher, for instance, participants would probably want to protect all the time, which would make forecasts rather useless (i.e., whatever is forecast to happen, participants would pay for protection anyway). We did not want to have this situation, as we would not know how they were using (or willing to use) their forecasts. It would be interesting to test the influence of the many parameters of the setup, but then we would need to make constant some other parameters and play the game differently. For instance, without different qualities of the forecasts or no differences in flood events among rivers. This could be interesting for further developments and we will mention it in the conclusion section, where we have also mentioned some other limitations of the setup (P16L19-28).
  - How river levels and increments were sampled??? What about the uncertainty of forecast increments? How such choice could impact conclusions? River levels and increments were set in a way that the number of floods was kept equal for each river, and river level values (of initial river level and river level increment) were randomly generated. This is mentioned in the experiment setup section (P4L27-30). As for the forecast increments, it was not the uncertainty (here expressed as interquantile ranges of the boxplots representing the forecasts) that differed among the different forecast sets, but the position of the observations inside the forecast distribution (which made the forecast underpredict or overpredict the observation). This was explained in the round 1 setup section (P5L16-19), but we will make it more clear in the revised version of the paper. The impact on conclusions comes therefore from the different quality (or perceived quality) of the forecasts on the decisions made. This is an issue explored in the results presented in Section 3.1 (P10). We could expect that using forecasts with different levels of uncertainty would make participants spend more money on protection when forecast increments are too uncertain (large boxplots) and take more risks (by not protecting) when forecasts are sharper. This could be an option for another setup of the game (i.e., instead of using forecasts of different accuracy).

- Page 5.Line 10 Include a figure (or a new panel on fig. 1) showing a diagram with the sequences of rounds, auction, etc., to summarize section 2 (experiment setup). This is a good idea, we believe that it will make the game structure clearer to the readers. Therefore, a diagram summarising the setup of the experiment will be added as a panel on Figure 1.

- Page 10 Line 15. Figure 2. How this distribution compare to the distribution of people actually making the decisions? This is a difficult question to assess. We have not included a question in the worksheet about it and have not discussed this point after the game. It would be certainly interesting to know the percentage of each category that had actual decision-making duties in their work. This is an issue that we can add in the discussion and conclusions section for further improvements of the game.

- Page 11, Lines 5-22. Figure 3. It is really not clear from this figure 3 if participants changed strategy during the 5 cases. It would be easier to ask this question to the participants in the form. This is a good point, but one that raises several challenges. In fact, participants were not aware that there could have been a bias in their forecasts. It was important that they did not know that in advance. If we had a direct question on the worksheet (which they have in hand from the beginning of the game), they would have discovered this feature (or be suspicious) before starting to play. Also, if we had put a question asking if they had looked at the median, or the upper or the lower quantiles, we would already be suggesting in advance that they should be looking at one of these. We did not want to influence their strategy. It is a difficult issue when designing a game experiment: how much do we present before they start playing and how much do we leave for the participants to decide (i.e., to play freely)? We have opted to leave it free to participants and then analyse the responses under some hypotheses and in terms of "could the participants have discovered the bias in their forecasts?" This was motivated by the fact that several participants came to talk to us after the application of the game saying that they had seen that their forecasts were biased. Therefore, this was our main driver in the analysis of "possible influencing factors" in Section 3.1.

- Page 11, Lines 20-25. Figure 3. It seems that participants also used information on the other percentiles different from median. For example, in case 4 for type 2, 5th and 95th percentiles indicate flood, so all participants chose the same and correct action. On the other hand, in cases where 5th and 95th percentiles fall above and below the flood threshold, more people did not follow the median (case 1 type 1, case 2 type 3, case 3 type 2, case 5 type 3). This is a very good point and it will be added to the results (Section 3.1).

- Page 12, Line 5, Figure 6b. This is very interesting. Participants attribute good decision making performance with good forecast quality, but they forget that their personal strategy adopted for decision making plays a major role, as there are several ways to interpret and use the probabilistic forecasts. What is the implication in the real world? Decision makers will tend to blame (thank) forecast providers for their wrong (good) decisions? This is a very good point which will be raised in the Discussion Section of the paper. It was mentioned in a HEPEX post "On the economic value of hydrological ensemble forecasts". (http://hepex.irstea.fr/economic-value-of-hydrological-ensemble-forecasts/). "*We once asked a decision-maker responsible for deciding on whether or not to open a control gate of a dam, and of how many meters it should be opened in case the decision was to open it, how he knew afterwards that his decision was the 'best decision'. The answer we got was (not* ipsis verbis*): 'It is the best decision: we take the best decision given the forecasts we receive and other complementary information we have on the situation. If the result is not good, the problem is not in the decision, but in the forecasts, which were not good'."*

- How the flood frequency observed in Round 1 impacted the WTP? Why participants from Green river are more WTP? The effect of the flood frequency of round 1 on participants WTP for another forecast set was briefly mentioned on P14L18-19 and in the conclusions (P17L31-33). However, nothing was said about the higher percentage of green river participants buying a second forecast set. We believe that it is a combination of the flood frequency (not as low as for the yellow river, which made it more relevant for green river participants to buy a second forecast set) and of money left in purse (on average, not as low as the blue river's participants). This will be added to the analysis section of the paper.

- Section 3.5: The analyses show that participants using forecasts had better performance in Round 2, however, participants with more money and better performance in Round 1 were willing to pay more for the forecasts. Consequently, participants with better performance in Round 1 ended buying forecasts and having better performance in Round 2. How the skill of the participants could impact the conclusion that "Decisions are better when they are made with the help of unbiased forecasts, comparatively to having no forecasts at all". We understand that the general question can be unfolded into the following questions: does the conclusions pertain only to the "good performing" participants? Is it true also for the "bad performing" participants? Do you need already to be a good decision maker to benefit from having forecasts in hand? Or do bad decision makers also improve their decisions when having forecasts? In order to investigate this issue, we first looked at the number of participants with a bad performance in round 1 and who had a forecast in round 2: all of these participants had a good performance in round 2. This is an indication that even when participants had a bad performance in round 1, when they had a forecast set in round 2, they all had a good performance in the second round. We then looked at the number of participants with a good performance in round 1 and who had no forecast in round 2. 57 out of 59 had a bad performance in round 2. This is also an indication that even if participants had a good performance in round 1, if they had no forecast set in round 2, they mostly had a bad performance. These observations will be added to the analysis in Section 3.5 in the revised version of the paper.

- Section 3.6. : Show % values, table or figure for these results. It may help the reader. We will add a table to back up the text about the winning and losing strategies.

- Is a biased forecast better than no forecast? Can you access that from this experiment? This is an interesting question: are forecasts, even if biased, useful for decision-making, comparatively to having no forecasts at all? Our experiment suggests that some participants adjusted their biased forecasts to make their decision. This can be an indication that they were useful somehow. However, our experiment setup does not allow drawing general conclusions. We cannot directly compare the performance of participants with biased forecasts in round 1 with the performance of participants without any forecasts in round 2, since the situations were not the same in both rounds (i.e., initial river levels, river level increments, money in purse, etc). In order to fully investigate this particular issue we would need to have an experiment where we play the same cases with two groups: one having no forecasts and another having biased forecasts. It is, in fact, interesting to note that when we setup a game experiment and analyse the results, several other possibilities open up for new variants of the experiment and further investigations. For us, this shows that there are still several open opportunities to enhance our understanding on how forecasts can be better used to inform decision-making.