We would like to thank the reviewer for the positive and constructive remarks. We provide here below our answers to the points discussed:

It might be more reader friendly to combine these sections and take the question-answer pairs one after another. We have thought about it when writing the paper but some feedback from the first readers found that it would be better to separate methodology from results.

Specific comments:

- P4L16-19: Did the participants know from the beginning that there are two rounds? → According to P6L7 they didn't. And P6L7: So, at the beginning, the participants didn't know that there were two rounds to play? P6L7 indeed suggests that the participants were not aware that 2 rounds would be played in total. However, participants were given a worksheet at the beginning of the game which contained a table for both round 1 and round 2. Moreover, they were told that 2 rounds of 5 cases each would be played during the game introduction. This will be clarified in the revised version of the paper.
- P4L27: Total number of flood events is kept equal in order to give equal chances to all participants to win. This statement doesn't hold. Participants surely did not have equal chances to win the game. This depended on the forecast set type and the river they were given. For the aim of the study it is not required that every participant has the same chance to win, but this statement is wrong. → 2.1.1. P5L3. This is a good comment and the reviewer is right for pointing it out. An equal total number of floods is not a needed criteria for this specific experiment and does not give equal chances to participants to win the game, given the other influencing factors. This sentence will be removed from the revised paper.
- P4L29-30: the number of flood events was different for every river, but not for every round → the Green River has the same number of events in round 1 and 2, whereas it changes for Yellow River and Blue River (which again influences their chances to win). These flood frequencies were chosen in order to: 1) explore the influence of different flood frequencies in round 1 on the participants WTP for a second forecast set, 2) investigate the change (yellow and green rivers) or no change (blue river) in flood frequency between round 1 and round 2 on the participants' strategies throughout the game. This was presented in Section 3.5 (P15L6-13). We will clarify those motivations in the revised version of the paper.
- P6L26-27: How did the participants that purchased a forecast perceive the quality of the forecasts compared to round 1? Did participants that had biased forecast sets in round 1 notice the better performance of the forecasts in round 2? This is an interesting point. It was explored during the analysis of the game results but only partially stated in the paper (see Section 3.5, P15L14-15). Among the 44 participants who bought a forecast in round 2 and perceived it of "good quality", 18 had played round 1 with a biased forecast set, and 19 with an unbiased forecast set. The 3 remaining participants stated that their forecasts for the second round were 'neither good nor bad' or 'quite bad'. These 3 participants all had biased forecasts in the first round and their behaviour during round 2 suggested that they might have been influenced by the bias in their forecasts during round 1. These additional results will be added to the paper (Section 3.5).
- Figure 3: - Yellow and Green River –> the ticks do not match the x-axis labels! - It would be nice if the plot was arranged according to table 3 - Change "forecasted final purse" to "final purse if following median forecast" - Labels of the "columns" could be changed to "pos. biased", "unbiased", "neg. biased" (same could be done for fig. 4). We thank the referee for the suggestions and will change the figures accordingly.
- QA 1 (2.2.1/3.1):
    - Fig 4: please change order of the bars such that forecast type 1 is first in line. This will indeed make the figure easier to read. We will change as suggested.

- Equal chances…: - the disadvantage for participants with positive bias is smaller if the observed value is above/equals the flood threshold. Thus for the blue river, which experiences three floods in five cases (1st round), the participants with forecast set type 1 are expected to perform better than those with forecast set type 3. This is a good point, and it will therefore be added to the paper. We indeed expected an influence of the game setup on the participants' behaviours and on their performances. The costs chosen for this experiment is an example of those factors. It was mentioned on P12L7-11 that, as the damage cost was set to twice the protection cost, this might have influenced the participants' tolerance to misses and false alarms.

- QA2 (2.2.2/3.2): You state that the percentage of negative perceptions of the quality of the forecasts increases with increasing or decreasing forecast bias. This seems to be quite consistent for positively biased forecasts, but how do you explain that the distribution of the ratings from the participants with the most negative bias (0) looks almost the same as from the participants with unbiased forecasts? → Fig.5. Participants with a bias of 0 belonged to the yellow river and had negatively biased forecasts in round 1. There was only one flood for river yellow, which occurred at the end of round 1. The negatively biased forecasts for river yellow thus missed this flood. During the analysis of the results, it was observed that only about 25% of the yellow river participants given the negatively biased forecasts did not protect for this flood. This could be because the participants had time to learn about their forecasts' quality until the occurrence of the flood, during case 5 of round 1. This low number of participants who actually suffered from their negative bias and the presence of only 1 miss out of the 5 cases of round 1 could explain the good rating of their forecasts by those participants. This will be made clearer in the text of the revised version.
- QA3 (2.2.3/3.3): At a first glance Figure 6 looks completely fine. However, there were five levels of perceived performances (very bad to very good) the participants could choose from. So the graph should not show perceived performances higher than five or lover than one. You could change the graph to a simple scatterplot and choose the point size proportional to the number of participants that fall onto a specific perceived actual-performance combination. (Same for Figure 6 b). This is true and a good suggestion. The figure will be modified accordingly.
- QA4 (2.2.4/3.4):
  - P13L3-4: It would be more straightforward if you just stated the average percentage the participants were willing to spend from the tokens left in their purse. We think that both, the average percentage and the actual amount, are needed for the sake of clarity.
  - P14L4-5: … 48+32+21=101 … round to 20 % for blue river. Yes, good point! However, it would be more informative what percentage of river group members purchased a forecast for the second round → 36% of the yellow river group, 41% of green rivers, 23% of blue rivers purchased forecasts for the second round. And also for the forecast set type groups → pos. biased 30%, unbiased 42% and neg. biased 31%. The percentages within each group will be added to follow the suggestions of the reviewer. We also think it will be clearer for the reader.
- QA6 (2.2.6/3.6):
  - P15L18-19: Could you give the average final purse for the two groups "with and without forecast in the second round" separately. On average, the participants without a second forecast set ended the game with 3341 tokens, compared to 2778 tokens on average for participants with a second forecast set. These final average purses are however not only a reflection of the participants' overall performances, but also of the setup of the game. If the number of cases had been larger in round 2 (not feasible due to the restricted amount of time in our case), we would possibly

have seen a larger benefit of having a forecast set in round 2 and as a result, a larger final purse for participants with a second forecast set. This is why the analysis of the winning and losing strategies (presented in Section 3.6) is largely qualitative. The final purses will be mentioned in Section 3.6.

  - Table 5: you could do that additionally by forecast set type. It is a possibility, but we think this would make the results more difficult to interpret since the table refers to round 2 and the forecast set types played a stronger role in round 1. Also, the average avoided cost makes sense for the same river as it will only depend on the possession of a second forecast set or not (as the flood frequency is the same within a river). If we consider also the different forecast quality groups, the average avoided cost will depend on the possession of a second forecast set but also on the number of floods for the second round and on the distribution of the participants among the different rivers. This would stratify too much our sample and results would lose robustness.

- Discussion/Conclusion:
  - P18L9: gambling was considered a reason for not buying a forecast set by other few participants. → According to P14L22 this was just one participant. Thank you for pointing this out. It will be corrected in the revised version of the paper.
  - P18L12-15: "This further demonstrates that more work is needed not solely to provide guidance on the use of probabilistic information for decision-making,..." Comment: It is questionable if the setup of the game is not to some extent counterproductive and doesn't help to improve this. The winners of the games had mostly biased forecasts in the first round and no forecast in the second round … If the game should have an educational merit, shouldn't the game then be set up in a way so that people who have no bias in the forecasts of the first round and who purchase a forecast in the second round have better chances to win? By designing this experiment, we did not intend to create an educational game as such, but rather a role-play situation where some important topics in hydrological forecasting could be reflected upon and discussed within a group. The main purpose of the game was to answer the questions explored and presented in this paper, within the limits of its feasibility and results' interpretation. The game contributes to the discussion about how forecasts are used and how much one is ready to pay to have them in their decision-making situation. Much of what one can learn from the game relies on the discussions the group may have after playing it. We believe that issues on the importance of forecasts for decision-making can be raised by participants once they have been introduced to the topic with this game experiment and are more at ease to share their own perceptions and knowledge.

Minor Comments: The required corrections will be made for the following comments.

- P2L3: remove dashes
- P3L15: remove "one"
- P7L8: his purse
- P11L17: … the lowest percentage of participants not following the median forecast are for the unbiased forecast set type 2.
- P13L16: … (the river that experienced most floods in round 1 and for which players thus ended the first round with on average the lowest amount of tokens left in their purse)
- P16L5-6: … than the "avoided cost" of each river. On average participants paid 1000 tokens more …
- P18L3: … necessary. Seifert …

- General: it would be easier for the reader if you used the terms "neg. biased forecast", "unbiased forecast" and "pos. biased forecast" instead of the terms "forecast set type 1-3" in the text, tables and graphs.