Response letter of hess-2021-497

Dear Editor and Anonymous Referee #1,

Please find the responses to the comments.

Comments made by the reviewer were highly insightful. They allowed us to greatly improve the quality of the manuscript. We described the response to the comments.

Each comment made by the reviewer is written in *italic* font. We numbered each comment as (n.m) in which n is the reviewer number and m is the comment number. In the revised manuscript, changes are highlighted in yellow.

We trust that the revisions and responses are sufficient for our manuscript to be published in *Hydrology and Earth System Sciences*

Sincerely
Yohei Sawada, Rin Kanai, Hitomu Kotani

Responses to the comments of Referee #1

I am not entirely satisfied with the revision of the paper in reply to both reviewers' comments. Specifically:

(1.1) Comment 1.6: the authors' response here is that the their assumption about trust is explained with the sentence: "Previous studies pointed out that the recent forecast accuracy and false alarm ratio affected the performance of preparedness actions (Simmons and Sutter 2009; Trainor et al. 2015; Ripberger et al. 2015; Jauernic and van den Broeke 2017)." This talks about the performance of preparedness actions, not about an increase or decrease in trust, or even about the implementation of measures. It does not explain why it is reasonable to assume that trust in FEWS increases (decreases) when prediction succeeds (fails). This needs to be better substantiated with evidence from the literature.
→ LeClerc and Joslyn directly investigated the relationship between trust ratings and false alarms by the controlled experiment. They found that trust ratings are increased by the decreased false alarm levels. Our model of social collective trust is based on this finding. This point was indeed unclear in the original version of the paper. We have clarified this issue in the revised version of the paper.

> Lines 198-201: In the controlled experiment of LeClerc and Joslyn (2015), medium-range trust ratings are increased by decreased false alarm levels. Their experiments revealed that trust ratings are based on the pattern of forecasts and observations over the previous month.

(1.2) Comment 1.7: in response to my point that it is not realistic that preparedness increases only because of trust while there is no memory of an event, the authors' decided to remove this discussion from the text, rather than update the model so that it is more realistic. Even if this does not happen often, this behaviour is not realistic and therefore implies a flaw in the model structure. This should be addressed or at the very least discussed as a limitation.
→ We examined this point again, and currently we believe that $P_r(t) > 0$ with $E(t) = 0$ is not so unrealistic. In an individual level, even if people have never experienced damages by themselves, they may be able to take preparedness actions based on information from their trusted authorities. Forecasters strongly expect this behavior. If $P_r(t) = 0$ with $E(t) = 0$ in an individual level, investment in disaster prediction systems cannot be justified in most of highly protected urbanized areas such as Tokyo in which most of citizens have never experienced water levels above damage thresholds by themselves. Since $E(t)$ depends only on damages, $E(t) = 0$ does not necessarily mean that they have no memory about flood events nor they fully forget the existence about flood. $E(t)$ should be interpreted as collective *personal* experiences of flood damages and should not be interpreted as a simple memory and knowledge about flood. Many disasters prevention measures such

as education, evacuation drills, and weather forecasting are designed to help people avoid the risk of flooding even if they have no personal experiences of flood disasters. To evaluate the effectiveness of these measures, $P_r(t) = 0$ with $E(t) = 0$ is not an appropriate behavior of the model although we admit that the effectiveness of forecasting highly depends on $E(t)$ as Girons Lopez et al. (2017) discussed.

Therefore, we still believe that the current additive form of the equation (7) is appropriate even at the limit of $E(t) \to 0$ with $T(t) \gg 0$ although it rarely happens as we discussed in the previous round. In the revised version of the paper, we have justified the additive form of the equation (7) using the discussion written above:

> Lines 231-243: The additive form of the equation (7) implies that preparedness actions are taken even if either social collective memory $E(t)$ or social collective trust $T(t)$ goes to zero. Note that $E(t) \approx 0$ does not mean that a community does not know the existence of a flood event while it means most of citizens have never experienced water levels above damage thresholds by themselves. Many disasters prevention measures such as education, evaluation drills, and FEWS are designed to let people take preparedness actions even if they have no personal experiences of flood disasters. Forecasters expect that people take preparedness actions based on information from their trusted authorities even if they have never experienced damages by themselves. To evaluate the effectiveness of these measures, $P_r(t) = 0$ with $E(t) = 0$ is not an appropriate behavior of the model although the effectiveness of FEWS highly depends on $E(t)$ as Girons Lopez et al. (2017) found. Therefore, we chose the additive form of the equation (7) rather than the other simple alternatives such as multiplicative forms.

(1.3) Comment 2.2: I believe the authors should more thoroughly discuss what their model results add to the discussion of whether the cry wolf effect exists or not. In my opinion, simply stating "the current evidence suggests the importance to understand the effect of false alarms on behavioral responses to warning in order to design efficient flood early warning systems." is not the same as discussing your results in the context of the current debate in the scientific literature.

→ Our modeling study does not give any evidences to conclude the debate on the existence of cry wolf effects and the purpose of this study is not to reveal the existence of cry wolf effects. We successfully justified the current behavior of forecasters, which balance the number of false alarms with that of missed events, more realistically than the existing dynamic models. This point has been discussed in the original version of the paper:

> Lines 522-528: While the GL model realistically simulates the behavior of the optimal warning threshold only if unrealistically high costs of mitigation and protection actions are assumed, our

> stylized model needs no costs of mitigation and protection actions to realistically simulate the behavior of the optimal warning threshold. Our stylized model is more consistent to the previous works in which the costs of mitigation and protection actions responding warnings were found to be negligibly small (e.g., Schroter et al. 2008; Hallegatte 2012; Pappenberger et al. 2015).

Note that this result implies that forecasters believe the existence of cry wolf effects, which is one of the important implications in our modeling work, but it does not necessarily mean that cry wolf effects exist. This point has been emphasized in the revised version of the paper.

> Lines 528-531: Our results justify the optimal warning thresholds which balance false alarms with missed events and imply that forecasters believe the existence of cry wolf effects, although it does not necessarily mean that cry wolf effects exist.

It is necessary to perform and accumulate more sophisticated field surveys and econometric analyses to contribute to the debate on the existence of cry wolf effects. Although our modeling work cannot directly contribute to it, we can obtain some useful implications for the design of future field surveys. First, our results show the sensitivity of relative loss to predefined probability threshold around the optimal value is small in many cases. In many field surveys such as Simmons and Sutter (2009) and Trainor et al. (2015), pairs of false alarm ratio and damage in many regions of one country are collected and compared to show the increase of false alarm ratio increases damage. Assuming that nationwide criteria of issuing warnings are almost optimal, our study implies that the observable signal of cry wolf effects in this approach is weak. It may be the reason why several field surveys contradict with each other. Our modeling results imply that it is difficult to quantify cry wolf effects using time-mean performance of warnings and damages at least for the flood disasters. We recommend analyzing the temporal change in behaviors responding to recent forecast outcomes, although this strategy seems to be costly and time-consuming.

Second, Figure 3 of the paper implies that it is better to choose technological societies as a research field because it is more difficult to distinguish the contributions of experience and trust in flood-prone areas. These points were indeed unclear in the original version of the paper. We have included them in the revised version of the paper.

> Lines 554-571: As discussed above, systematic econometric analyses and field surveys on cry wolf effects have not been implemented for flood disasters, so that it is important to design such kinds of analyses. Our modelling work provides useful implications for the design of future field analyses. First, our results show that the sensitivity of relative loss to predefined probability threshold is small around its optimal value in many cases. In many field surveys such as Simmons and Sutter (2009) and Trainor et al. (2015), pairs of false alarm ratio and damage in many regions

of one country are collected and compared to show the increase of false alarm ratio increases damage. Assuming that nationwide criteria of issuing warnings are near-optimal, our study implies that the detectable signal of cry wolf effects in this approach is weak. Our modeling work implies that it is difficult to quantify cry wolf effects using time-mean performance of warnings and damages. It may be the reason why several field surveys contradict with each other and the negative effect of false alarm ratio cannot be found in some surveys (Lim et al. 2019). We recommend analyzing the temporal change in behaviors responding to recent forecast outcomes, although this strategy is costly and time-consuming. Second, our experiment 3 implies that it is better to choose technological societies as a research field because it is more difficult to distinguish the contributions of experience and trust in less protected areas.

(1.4) Comment 2.6: in reply to the reviewers' question about the choice of value for γ the authors reply there is no evidence from the literature on which to base this value. The justification for choosing 0.5 (i.e. not having to consider asymmetric contributions of E and T) is not good enough. The addition of trust in the model is the main contribution, if there is a parameter that determines the influence of trust on your model outcome and there is no evidence that suggests a certain value, one should at least investigate the effect of the parameter's value on the end results of the study. I would suggest to add a sensitivity analysis of the study's results to this parameter.

→ We have added how γ affects the relationship between predefined probability threshold and relative loss in the revised version of the paper.

Lines 292-293: The behavior of the models with the different γ is also discussed in the supplement material.

Lines 398-404: Figure S2 shows how γ in the equation (7) affects the relationship between relative loss and predefined probability threshold. When the contribution of social collective trust to social preparedness increases (i.e., γ gets smaller), the “implicit cost” of false alarms induced by relatively small predefined probability thresholds increases. Figure S2 also shows that moderate changes of γ from the default setting of the SKK model (i.e. 0.5) do not qualitatively change the relationship between relative loss and predefined probability threshold.

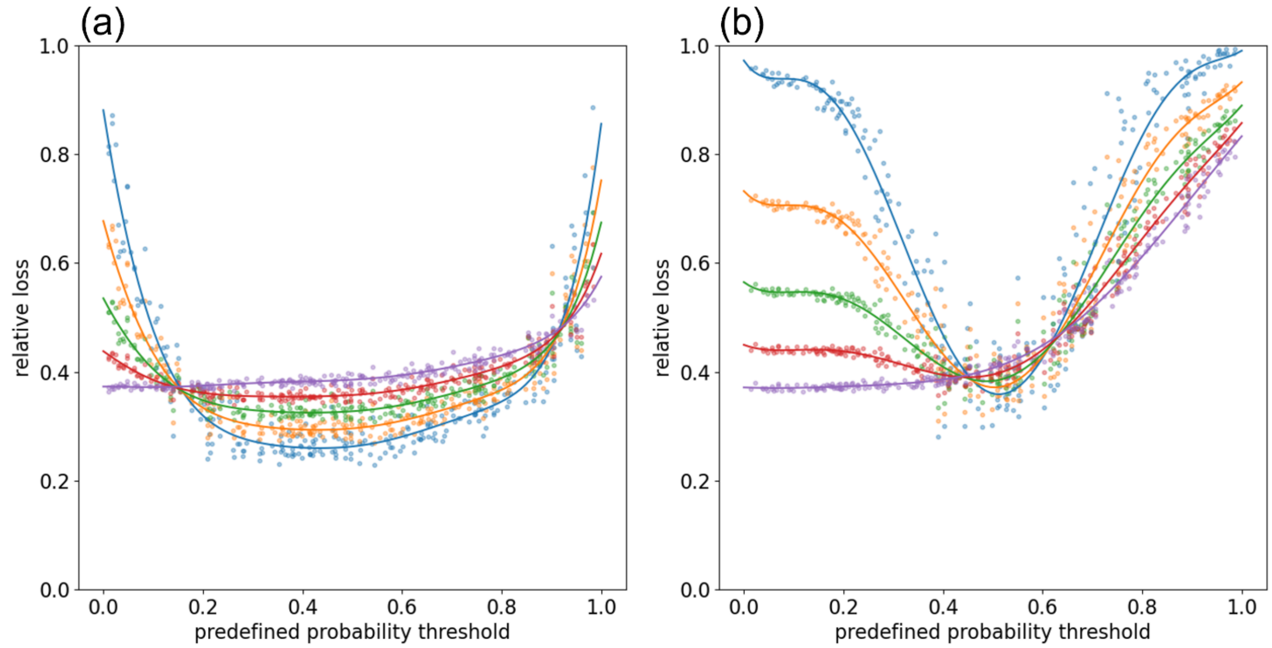

Figure S2. Sensitivity of γ in equation (7) to the relationship between relative loss and predefined probability thresholds. No cost of the mitigation and protection action is assumed. In (a), the high prediction accuracy that is same as experiments 2.1 and 2.4 is assumed. In (b) the lower prediction accuracy that is same as experiments 2.2 and 2.5 is assumed. Blue, orange, green, red, and purple dots and lines are results with $\gamma = 0, 0.25, 0.5, 0.75, and\ 1$, respectively. Each dot shows the result of the individual Monte-Carlo simulation and we smoothed them by Gaussian process regression. See also Table 4 for detailed parameter settings.

(1.5) Comment 2.9: while the chosen time range may clearly show the difference between the two models, it would also be good to include a figure that shows the behaviour of the model over the entire time range.

→ We have shown it in a supplement material.

> Lines 347-349: While Figure 1 shows the subset of the entire timeseries to clearly demonstrate the differences between two models, the entire timeseries can be found in Figure S1 of the supplement material.

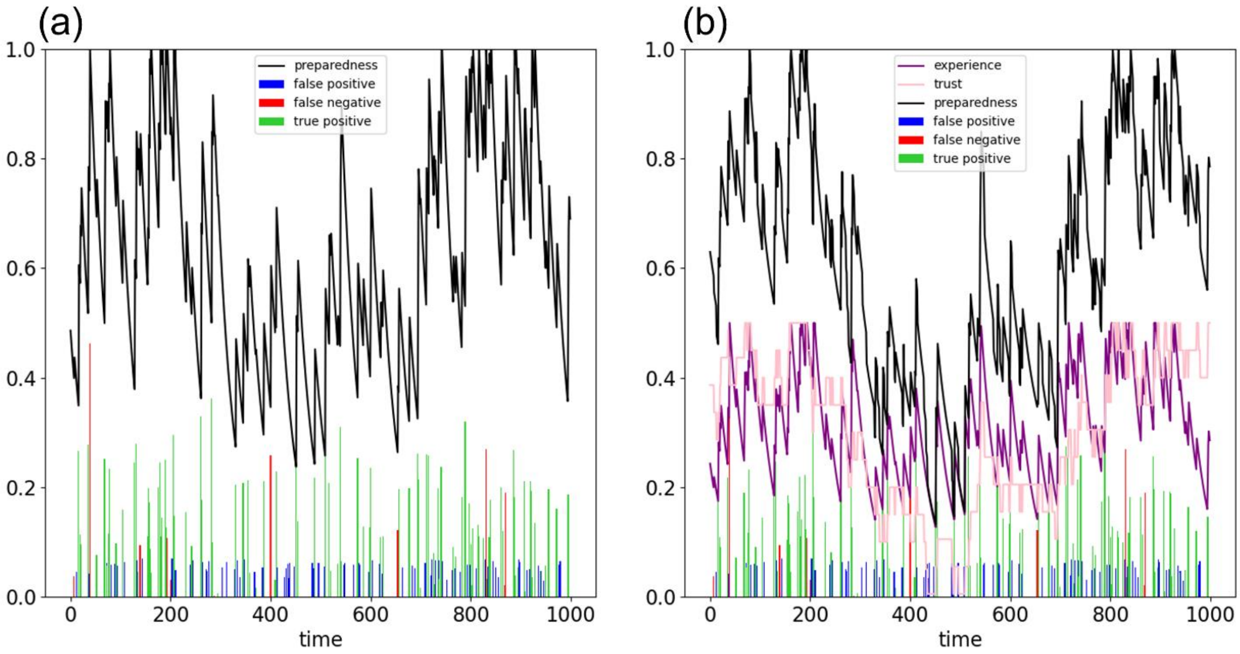

Figure S1. Same as Figure 1 but for the entire time range.

(1.6) Comment 1.1, 2.11, 2.12: The authors do now discuss their results in relation to the literature that they discuss in the introduction, but only in relation to the literature that supports the existence of the cry wolf effect. Why do the other papers find that this effect does not exist? And how do the results of this study relate to that? How do they support the evidence that the effect does exist?

→ For example, the survey design of Trainor et al. (2015) and Lim et al. (2019) is similar, but the conclusions are completely different. Especially, the contribution of actual false alarm ratio is debating in the literature. As discussed in our response to the comment (1.3), we found that the sensitivity of relative loss to predefined warning threshold is small around the optimal value of the threshold. It implies that it is difficult to obtain the signal of cry wolf effect in the real world, which may be able to explain the contradiction of previous works. This point has been included in the revised version of the paper. See also our response to the comment (1.3).

Lines 557-566: First, our results show that the sensitivity of relative loss to predefined probability threshold is small around its optimal value in many cases. In many field surveys such as Simmons and Sutter (2009) and Trainor et al. (2015), pairs of false alarm ratio and damage in many regions of one country are collected and compared to show the increase of false alarm ratio increases damage. Assuming that nationwide criteria of issuing warnings are near-optimal, our study implies that the detectable signal of cry wolf effects in this approach is weak. Our modeling work implies that it is difficult to quantify cry wolf effects using time-mean performance of warnings and damages. It may be the reason why several field surveys contradict with each other and the negative effect of false alarm ratio cannot be found in some surveys (Lim et al. 2019).

(1.7) In addition, after reading both reviewer's comments and the author's replies I tried to reproduce the results presented in the paper and was unable to do this with the information provided in the paper. Therefore, I have some further points that I believe should be addressed before the paper can be published:

→ Thank you very much for checking the details of the study. We have made our source code publicly available. See https://gitlab.com/ysawada/sociometeorology as well as our responses to the comments below.

Lines 590-592: **Code and Data Availability**

The code to perform the numerical experiments is available in a public repository (https://gitlab.com/ysawada/sociometeorology).

(1.8) Using a gaussian distribution for the variance of the forecast implies you get negative values, which is not possible, how is this solved? Is a truncated distribution used? If so, this should be mentioned.

→ A truncated distribution was used. This point was indeed unclear in the original version of the paper, and we have clarified this point in the revised version of the paper:

Lines 128-129: Negative $N(\mu_v, \sigma_v^2)$ is truncated to 1.0e-6 to prevent from obtaining negative values of variance.

(1.9) Similarly, in general (even when assuming the above mentioned variance is set to be >0), the forecast distribution yields negative forecasts for the discharge. This does not make sense, the discharge cannot be negative.

→ We fully agree with this comment. We did not deal with this issue. It does not affect the consequences of the dynamic model since the concern is if forecasted flood level is above the (positive) damage threshold. However, this is the limitation of our model due to the oversimplification of a forecasting system and should be discussed. In the revised version of the paper, we have mentioned this issue.

Lines 133-136: Although this simplification of the forecasting system unrealistically assigns non-

> zero probability to negative values of discharge, it does not affect the process dynamics since the model evolution depends only on whether forecasted discharge is above the damage threshold, as we explain in the next paragraph.

(1.10) In addition, the parameter settings used for the discharge imply a discharge of 0.2 corresponds to the 2.4 year flood and a discharge of 0.45 corresponds to a 21 year flood. This last value, 0.45, is used as the damage threshold for the technological society, but a protection level protecting against a 21 year flood is rather low for a technological society. I would expect it to protect at least against a 100 year flood (i.e. a damage threshold between 0.6 and 0.61)

→ We believe that the reviewer's expectation of technological society, which is to fully protect at least against a 100-year flood, is too severe. In Japan, most of the largest river basins do not achieve this level. If the reviewer's criterion for technological society is applied, we guess few technological societies exist in Asian monsoon regions which have severer rainfall than European countries. To check how the models work with a larger damage threshold, we performed additional experiments with a damage threshold of 0.60 and Figure R1 shows the result. It does not change the take home messages in the experiment 2. The notably difference from the original experiments is that 1000-year averaged relative loss are strongly affected by random processes (see dots in Figure R1). This is because the number of floods in the 1000-year simulation is very small and the forecast outcomes of this several events fully determined the overall performances, which increases sampling errors. Although we decided not to include Figure R1 in the paper, we have discussed this point in the revised version of the paper.

> Lines 434-437: These behaviors of the models can be found when damage threshold is further increased to 0.6, although the 1000-year averaged statistics are strongly affected by random processes due to the insufficient number of disaster events within the 1000-year computation period (not shown).

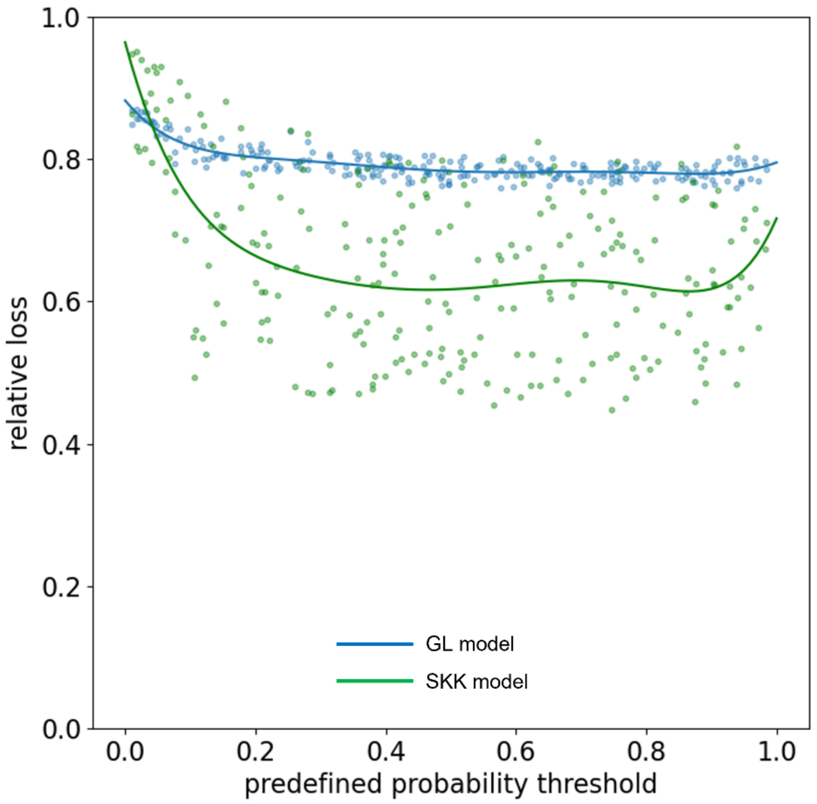

Figure R1. Same as Figure 3a or 3c, but for damage threshold of 0.6.

(1.11) The initial values for E and T are not reported. And related to that: do they remain the same across all experiments or do you vary them? The initial values may have a significant effect on the model results and therefore the conclusions.

→ The initial conditions of E and T are randomly selected from 0-1. This point was written in the section 3.2 of the original version of the paper:

> Lines 336-338: In experiments 2–4, we performed the 250-member Monte-Carlo simulation by randomly perturbing a predefined probability threshold, π, and the initial conditions of social collective memory and social collective trust in FEWS. We analyzed the sensitivity of the efficiency of FEWS to predefined probability thresholds.

Although we randomly choose 250 initial conditions, we used the same 250 combinations of initial conditions for all experiments. The initial conditions do affect the model results, which in part was shown in the variance of dots in our Figures 2-4, although this effect becomes small by integrating the model for 1000-year and averaging this long-term simulation results. However, the difference of initial conditions does not affect the differences between experiments (e.g., between experiments 2.1 and 2.2) since we used the same combination of initial conditions. This point was indeed unclear in the original version of the paper, and we have clarified that the differences between experiments do not depend on any random processes in the revised version of the paper.

> Lines 338-340: We used the same random seed to generate 250-member Monte-Carlo simulation in each experiment, so that the differences between experiments do not depend on random processes.

Lastly, the initial conditions of the experiment 1 are also randomly chosen, and we have explicitly written the values of them in the revised version of the paper.

> Lines 297-298: The initial conditions of E and T are randomly chosen and set to 0.49 and 0.77, respectively.

(1.12) Similarly to the initial values, the order of the floods will influence the model results. Running the model for several different floods time series and then averaging the outcome metric (e.g. the relative loss) over those runs yields more robust results.

→ We have performed the experiments with 10 different discharge timeseries, and Figure S3 shows the results. We found that the uncertainty induced by the order of the floods is comparable to the uncertainty quantified by 250 Monte-Carlo simulations with different initial conditions and forecast outcomes. Our take home messages are qualitatively unchanged by different discharge timeseries sampled from the same gamma distribution. Since we believe that it is straightforward to compare different experiments under a single sequence of events and the accurate estimation of the outcome metric is not the main purpose of this theoretical study, we still use a single timeseries in the revised version of the paper. However, we have clarified the discussion written here using Figure S3 in the main manuscript of the revised paper.

> Lines 404-407: In addition, the qualitative behavior of our SKK model is robust to different discharge timeseries (Figure S3). Figure S3 reveals that the uncertainty induced by different discharge timeseries is comparable to that quantified by 250 Monte-Carlo simulations with different initial conditions and forecast outcomes.

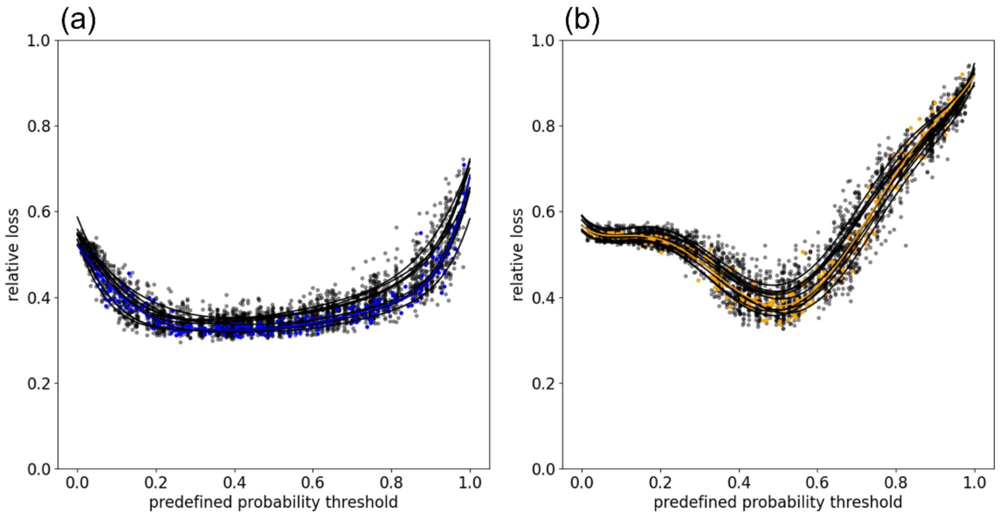

Figure S3. The relationship between relative loss and predefined probability thresholds in (a) the experiment 2.4 and (b) the experiment 2.5. In (a), the blue line shows the original result shown in Figure 2(b) and black lines show the results with 10 different river discharge timeseries which are sampled from the gamma distribution shown in the equation (1). In (b), the orange line shows the original result shown in Figure 2(b) and black lines show the results with 10 different river discharge timeseries which are sampled from the gamma distribution shown in the equation (1). Each dot shows the result of the individual Monte-Carlo simulation and we smoothed them by Gaussian process regression. See also Table 4 for detailed parameter settings.

(1.13) In the form as it is written in the paper, the trust T explodes in a negative or positive direction, rather than being constrained between 0 and 1. Similarly the experience E is not constrained to be below 1. Therefore the preparedness can become very large or negative, giving negative damage.

→ We totally agree with this comment. We constrained E and T to 0-1. This point was indeed unclear in the original version of the paper, and we have clarified this point in the revised version of the paper.

Lines 190-191: When E becomes larger than 1, it is truncated to 1.

Lines 205-206: When T becomes larger than 1, it is truncated to 1. When T becomes smaller than 0, it is truncated to 0.

(1.14) For most experiments conducted and reported in the paper, π is varied from 0 to 1. But for the first comparison between the two models, a fixed value is used. What is this value?

→ It was set to 0.40. We have clarified this point in Table 3 of the revised version of the paper.