Supplement of Nat. Hazards Earth Syst. Sci., 22, 1499–1518, 2022
https://doi.org/10.5194/nhess-22-1499-2022-supplement

Natural Hazards and Earth System Sciences
Open Access
EGU

Supplement of

Effective uncertainty visualization for aftershock forecast maps

Max Schneider et al.

Correspondence to: Max Schneider (maxs15@uw.edu)

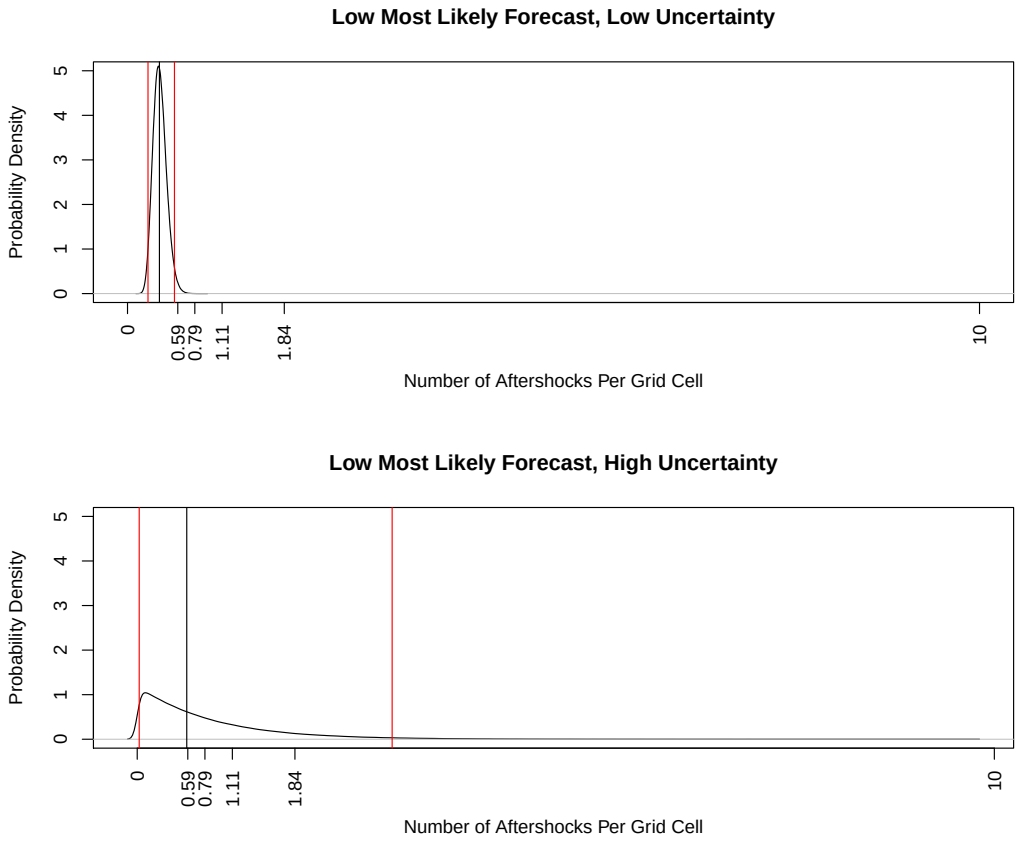

Figure S1. Examples of aftershock forecast distributions that both have low most likely forecasted aftershock rates (median; marked by black lines) but with low uncertainty (top) and high uncertainty (bottom). Red lines mark the 2.5% and 97.5% percentiles of the forecast distributions, which span a larger and more skewed interval for the forecast with high uncertainty.

S1 Summary of interviews with emergency managers

We interviewed five emergency management officials from Washington State, USA, four of whom were directors of either state or county-level emergency management offices (the fifth participant was a volunteer at a county-level emergency management office). The interviews focused on how scientific information is used when responding to natural disasters, and how the visual features of previous disaster scientific communications affected how the emergency managers used them for decision-making needs. We also asked about how uncertainty in scientific communications affects their use for disaster response. We began by posing general questions and then narrowed in on previous experiences with non-earthquake hazards, as most participants had not yet worked in earthquake response. Interviews were recorded and then summarized across participants, finding similarities and differences across the individual responses. We identified several important use cases and communication goals, around which we structured the experimental tasks to evaluate uncertainty visualizations.

S2 Selection of colors for maps

The yellow-orange-red (Yl-Or-Rd) palette is often recommended for natural hazards (Doore et al., 1993) and has appropriate color connotations across Western cultures, especially for the lowest rate level (yellow, commonly connoting caution in the natural hazards context) and highest rate level (red, commonly connoting danger) (Sherman-Morris et al., 2015; Thompson et al., 2015). We began with the Yl-Or-Rd RColorBrewer palette, known to be colorblind-friendly (Brewer and Harrower, 2009). We fixed the start and end colors as suggested by this palette: respectively, yellow (made darker to maintain distinguishability for different transparency levels) and a dark red. We then created a five-color palette that was perceptually uniform using the Hue-Saturation-Lightness color model. We uniformly decreased colors' lightness and increased hue, while keeping maximal saturation. The hue and lightness of all colors were then adjusted to be optimally discriminable in both the legends and maps. The sixth color (highest rate level) was made dark brown, a color off the Yl-Or-Rd color palette to further distinguish it from the previous colors, as it was not in any visualization aside from the intervals-based design. This was to avoid the false matching of the darkest color in the map with the darkest color on the legend, as occurred when piloting.

We used a white color with black shadow for the area marker in the Read Off and Comparative Judgment tasks to avoid simulatanous contrast affecting the perception of the color within the marker (Krauskopf et al., 1986).

S3 Developing location distance measures

We investigated whether the characteristics of trial locations in the comparative judgment task influenced how participants judged between them. We focused on primary map features (map center, zones of high-rate and -uncertainty) and whether either location was substantially closer to them.

For each location, we calculated its Manhattan distance (number of grid cells) to the map center and the nearest zone of high rate or uncertainty. We then calculated the difference between these distances for the two locations in each trial. These distance differences are only meaningful when neither location itself has high rate or uncertainty and were only computed in those cases. Furthermore, we are interested in assessing whether locations were meaningfully closer to a given map feature, as we did not expect participants to be influenced by small differences in distances. Thus, we created a new categorical variable indicating which location was at least three grid cells (the median distance across trial locations) closer to each map feature. If both locations were essentially equidistant from that map feature (difference of two grid cells or fewer) or if either location was itself in that zone, the trial's value of this variable was "neither", which was set as the reference category. These categorical variables entered our model selection procedure as potential fixed effects (see Text S5).

S4 Screenshots of conditions from experiment platform

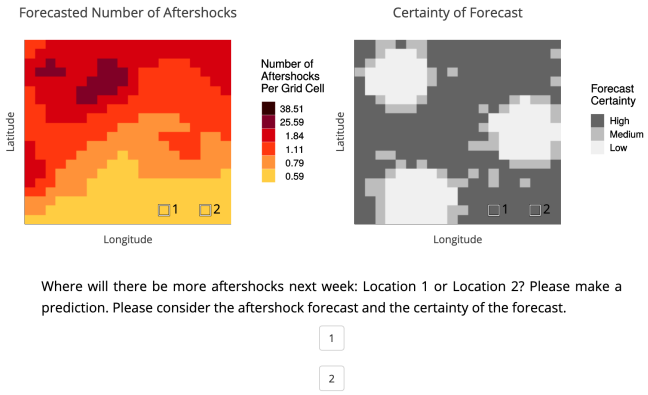

Figure S2. Screenshot of baseline trial for Comparative Judgment task with the Adjacent UV.

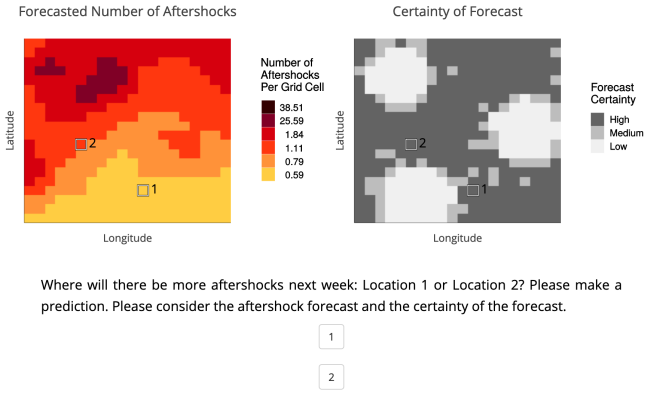

Figure S3. Screenshot of surprise trial for Comparative Judgment task with the Adjacent UV.

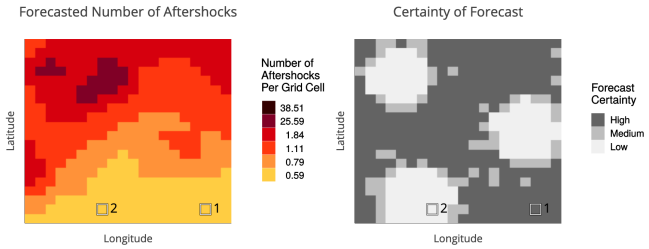

Figure S4. Screenshot of sure bet trial for Comparative Judgment task with the Adjacent UV.

S5 Multilevel model and model selection

We expected participant heterogeneity in comparative judgment, irrespective of visualization condition. We did not experimentally control for the myriad sources of this heterogeneity, nor did we know *a priori* how these may influence judgment, so we omitted participant-level effects from our main confirmatory analysis. We instead built a multilevel model of participant judgments to explore this heterogeneity by participant. We then performed a model selection analysis to identify drivers of task response when accounting for individual differences. We investigated which participant- or trial-level variables led to best model fit (across all participants and trials).

We first built a baseline multilevel regression model on Y_{ij} (participant i's judgment in trial j), with a fixed effect for condition, rate level, and their interaction as well as a random participant-level intercept. We used a treatment contrast for the visualization condition (reference category: Rate Only) and rate level (reference category: Medium rate) variables. The baseline multilevel model (built using the glmer R package, version 1.1-27) is:

$$
\begin{aligned}
logit(P(Y_{ij}=0)) = \beta_0 + \zeta_{0i} + \\
&\beta_b I(UV_i = Bounds) + \beta_t I(UV_i = Transparency) + \beta_a I(UV_i = Adjacent) + \\
&\beta_L I(Rate_j = Low) + \beta_H I(Rate_j = High) + \\
&\beta_{bL} I(UV_i = Bounds) * I(Rate_j = Low) + \beta_{bH} I(UV_i = Bounds) * I(Rate_j = High) + \\
&\beta_{tL} I(UV_i = Transparency) * I(Rate_j = Low) + \beta_{tH} I(UV_i = Transparency) * I(Rate_j = High) + \\
&\beta_{aL} I(UV_i = Adjacent) * I(Rate_j = Low) + \beta_{aH} I(UV_i = Adjacent) * I(Rate_j = High), \\
\zeta_{0i} \sim N(0, \sigma_\zeta)
\end{aligned}
$$

We considered the estimated fixed effect coefficients and compared them with the confirmatory analysis. We then built models with additional fixed effects (trial-level variables, participant-level covariates) and compared them to the baseline model, using model performance criteria. We only considered multilevel models with random intercepts (as a random "slope" on condition would not be identifiable with the data). Thus, we simply added fixed effects stepwise to the baseline model and assessed if this produced a better model fit, using multiple metrics.

A common model performance metric is the modified Bayesian Information Criterion (mBIC), which balances model goodness of fit with parsimony (Müller et al., 2013). Since we build classification (binary response) models, we also used classic metrics for classification (defined in Table S1): correct classification rate (CCR); Brier score; and area under the Receiver Operating Characteristic (ROC) curve (AUC). For all of these scores, we calculated the predicted probability of selecting the location of higher uncertainty, p_{ij}, for a given participant and trial, based on the given model. We then binarized these predicted probabilities at 0.5, obtaining binary predicted judgments b_{ij} (is location of higher uncertainty predicted to be selected?). These binary predicted judgments were then compared to the observed judgments o_{ij} with the goodness-of-fit metrics. This model selection analysis solely aimed to identify other variables beyond those in the baseline model that explain the experimental data, based on in-sample goodness of fit. Significant effects in the best-fitting model could be considered as key determinants of task response and built into future experimental designs.

S6 Supplementary experimental results

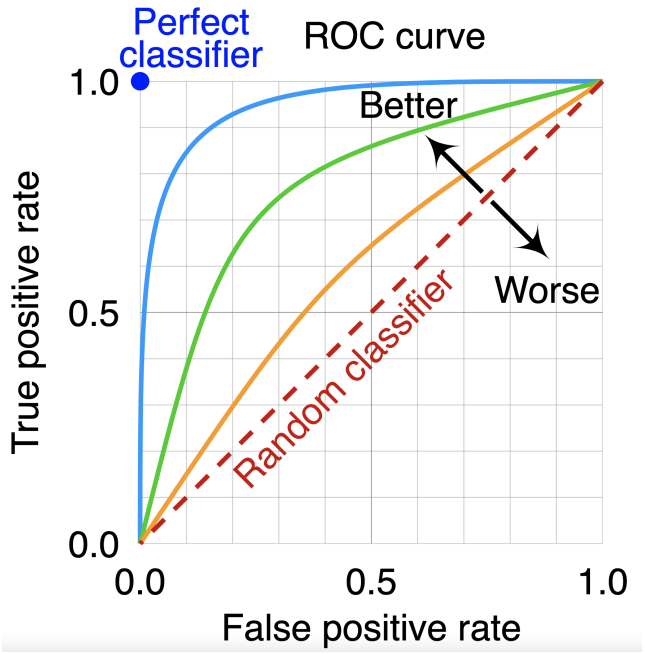

Figure S5. ROC curves as described in Thoma (2018). The area under the ROC curve (AUC) for the best possible classifier should be as close to 1 as possible, as this would maximize the true positive rate and minimize the false positive rate.

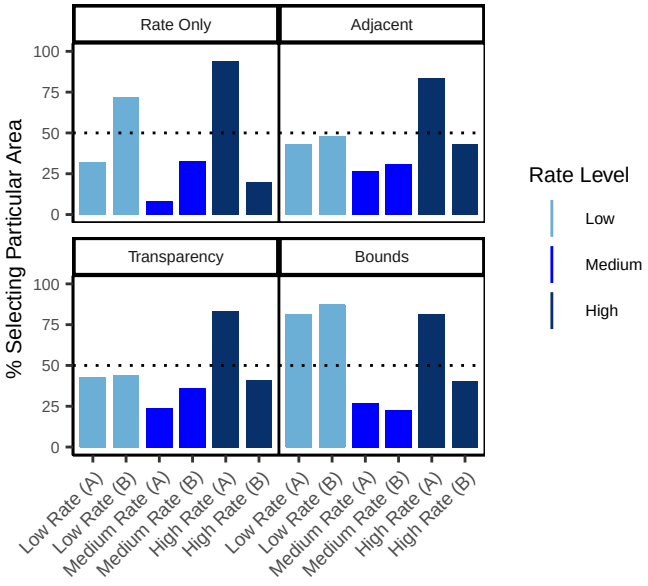

Figure S6. Percentage of participants selecting a particular location in the six baseline trials, where both locations have the same rate and same low uncertainty. This is Location 1 in the baseline trials described in Table 2 in the main article. Trials were repeated for forecast regions A and B.

Table S1. Model performance metrics. CCR is the correct classification rate, AUC is the area under the ROC curve and use n_t, the number of trials and n_p, the number of participants. $mBIC$ is the modified Bayesian information criterion. For the definition of $mBIC$, LL is the model log-likelihood, f is the number of fixed effects and r is the number of random effects estimated.

Metric	Definition	Direction
$CCR(b_{ij}, o_{ij})$	$CCR(b_{ij}, o_{ij}) = \frac{1}{n_t n_p} \sum_{i=1}^{n_p} \sum_{j=1}^{n_t} I(b_{ij} == o_{ij})$	Higher is better
$AUC(p_{ij}, o_{ij})$	Explained graphically in Fig. S5	Higher is better
$Brier(p_{ij}, o_{ij})$	$Brier(p_{ij}, o_{ij}) = \frac{1}{n_t n_p} \sum_{i=1}^{n_p} \sum_{j=1}^{n_t} (p_{ij} - o_{ij})^2$	Lower is better
$mBIC(LL)$	$mBIC = -2LL + \log(n_t n_p) \cdot (f + r)$	Lower is better

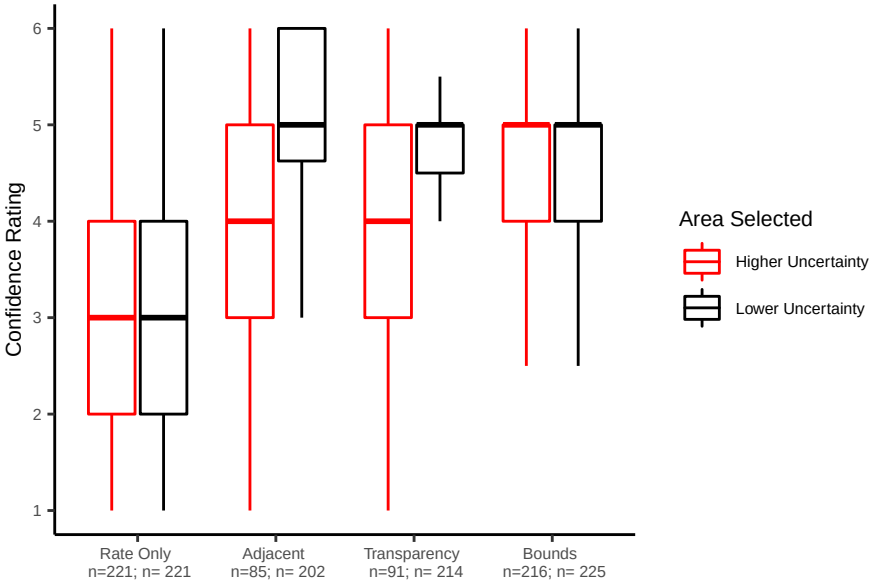

Figure S7. Boxplots for surprise judgment confidence ratings (median confidence by participant) by condition, split by which location was selected. Sample sizes are given for each condition, the number of participants in that condition that selected the given location at least once.

References

Brewer, C. A. and Harrower, M.: ColorBrewer 2.0: Color advice for cartography, http://colorbrewer2.org/, last access: 2021-08-03, 2009.

Doore, G. et al.: Guidelines for using color to depict meteorological information, Bulletin of the American Meteorological Society, 74, 1709–1713, 1993.

Krauskopf, J., Zaidi, Q., and Mandlert, M. B.: Mechanisms of simultaneous color induction, Journal of the Optical Society of America A, 3, 1752–1757, 1986.

Müller, S., Scealy, J. L., Welsh, A. H., et al.: Model selection in linear mixed models, Statistical Science, 28, 135–167, 2013.

Sherman-Morris, K., Antonelli, K. B., and Williams, C. C.: Measuring the effectiveness of the graphical communication of hurricane storm surge threat, Weather, Climate, and Society, 7, 69–82, 2015.

Thoma, M.: Receiver Operating Characteristic (ROC) curve with False Positive Rate and True Positive Rate., https://upload.wikimedia.org/wikipedia/commons/1/13/Roc_curve.svg, last access: 2022-01-17, 2018.

Thompson, M. A., Lindsay, J. M., and Gaillard, J.-C.: The influence of probabilistic volcanic hazard map properties on hazard communication, Journal of Applied Volcanology, 4, 1–24, 2015.

Table S2. Most likely estimates and 95% confidence intervals (using Wald standard errors) for fixed effects in baseline model. The intercept is the logistic of the probability of selecting the higher-uncertainty location for the Rate Only condition for medium-rate trials (reference levels for visualization and rate level). Each fixed effect give the change in the logistic of the probability corresponding to a change in visualization, rate level or their interaction (all else being held equal).

Fixed Effect	Estimated Coefficient	95% CI	
Intercept	-1.07	[-1.37,	-0.76]
Visualization			
Adjacent	-1.54	[-2.04,	-1.04]
Transparency	-1.90	[-2.40,	-1.40]
Bounds	0.65	[0.22,	1.07]
Rate Level			
Low	3.15	[2.88,	3.41]
High	0.64	[0.43,	0.86]
*Visualization*Rate Level*			
Adjacent*Low	-2.49	[-2.92,	-2.06]
Adjacent*High	-0.86	[-1.27,	-0.44]
Transparency*Low	-3.31	[-3.75,	-2.87]
Transparency*High	-1.43	[-1.87,	-0.99]
Bounds*Low	0.01	[-0.40,	0.42]
Bounds*High	-0.61	[-0.92,	-0.31]

Table S3. Model performance for baseline model and other models with trial-level fixed effects. CCR is the correct classification rate, AUC is the area under the ROC curve and $mBIC$ is the modified Bayesian information criterion (see Table S1 for definitions).

Model	Fixed Effects	CCR	AUC	$Brier$	$mBIC$
Model 1	Visualization * Rate Level	0.851	0.931	0.102	17072.6
Model 2	Visualization * Rate Level + Location Closer to Center	0.858	0.941	0.094	16700.7
Model 3	Visualization * Rate Level + Location Closer to Center + Location Closer to High Rate Zone	0.871	0.942	0.093	16670.9

Table S4. Model performance for baseline model and other models with participant-level fixed effects. CCR is the correct classification rate, AUC is the area under the ROC curve and $mBIC$ is the modified Bayesian information criterion (see Table S1 for definitions).

Model	Fixed Effects	CCR	AUC	$Brier$	$mBIC$
Model 1	Visualization * Rate Level	0.851	0.931	0.102	17072.6
Model 2	Visualization * Rate Level + Age	0.850	0.931	0.102	17078.6
Model 3	Visualization * Rate Level + log(numberEq + 0.01)	0.851	0.931	0.102	17080.3
Model 4	Visualization * Rate Level + Education	0.850	0.931	0.102	17112.6
Model 5	Visualization * Rate Level + Gender	0.851	0.931	0.102	17089.9
Model 6	Visualization * Rate Level + State	0.850	0.931	0.103	16948.4