Towards Ecological Validity in Evaluating Uncertainty

P. Samuel Quinan, Lace M. Padilla, Sarah H. Creem-Regehr, and Miriah Meyer

1 Introduction

Uncertainty is an inherent part of making decisions in a broad spectrum of situations, from wildfire management [14] and hurricane evacuation [3], to water use policy [4]. Studying the effects of visualizing uncertainty in these decision making contexts, however, remains a challenge due to the influence of expertise and individual differences in how experiential knowledge influences decision making [15]. This makes controlled studies of the effects of uncertainty visualization on decision making in real situations difficult, if not impossible.

While there is a growing recognition that evaluation in visualization could benefit from a greater variety of empirical methodologies [2, 7], the field remains dominated by two primary modes of evaluation: quantitatively-focused user-studies and the more qualitative case studies found in design studies. These two modes of evaluation sit on opposite ends of the experimental spectrum, with user-studies prioritizing precision over realism and case studies prioritizing the inverse. This spectrum involves tradeoffs between generalizability, precision, and realism [11], with experimental control positively correlated to the precision of measurements and results, and the amount of realism closely tied to the concept of **ecological validity**. Ecological validity specifically refers to how closely the experimental setting matches the setting in which the results might be applied [2]. Psychologists have long noted that there is an explicit trade-off between experimental control and ecological validity [8].

Existing user-studies dealing with uncertainty visualization have largely dealt with simplified low-level detailed tasks [6]. These simplifications, however, often force expert users to make judgments outside of their usual decision-making contexts, making it unclear how applicable the results of these studies are in the real-world. The qualitative case-study feedback traditionally found in design studies, on the other hand, presents its own set of challenges. Case studies can show *that* an effect exists, but it is extremely difficult to link the cause to a particular design decision. While additional feedback from activities like contextual interviews and think-aloud protocols can provide some insight into why a case study was successful, there remains a risk that the qualitative feedback conflates users' preferences or demand characteristics with performance [1, 5, 12].

In this workshop paper we will discuss why ecological validity is critical in evaluating the effects of uncertainty visualization in weather forecasting specifically, as well as describe our initial attempt at designing and running an ecologically valid user-study in this domain. Our hope is to spark a more general discussion about the need for increased realism in uncertainty visualization user-studies, the difficulties and uncertainties involved in designing such a study, and the open questions we face as a community.

2 Ecological Validity and Forecasting

Weather forecasting is a decision-making domain that is inherently uncertain and involves unique processes based on individuals' experience and knowledge [10]. A series of studies comparing domain experts' and novices' interactions with weather maps demonstrates that an understanding of atmospheric dynamics fundamentally changes the way individuals read and reason about weather maps [9]. Thus, it is unclear whether findings regarding general reasoning about geospatial maps transfer to reasoning about weather maps and vice versa. Furthermore, among forecasters there is no single model for how individuals predict weather. An individual meteorologist's process may change based on *"climate, season, experience, and a host of other factors"* [15].

The personalized nature of weather forecasting makes evaluation of uncertainty visualization effects difficult to do in a controlled setting. Previous work, however, offers no guidance on the matter of ecological validity. In their recent survey of geospatial uncertainty visualization user-studies, Kinkeldey *et al.* point out that the methods used to evaluate uncertainty visualization remain ad hoc [6]. The authors found that not only is the type and level of expertise among study participants often unclear, but the tasks are often simplified to focus on low-level details – value retrieval, aggregation, comparison, search, etc. – without any explicit justification [6]. Such simplifications are poorly matched for a study dealing with weather forecasting, where *"simply showing a complex visualization, expecting a user to extract the necessary information, and to be finished is an oversimplification of how complex visualizations are used"* [15]. Expert forecasters use visualizations to create an aggregated mental model of what is happening in the atmosphere and then use that mental model as the primary source of information for their judgements and decisions [14]. An ecologically valid study must support this complex reasoning process.

3 Designing a More Appropriate User-Study

In an effort to better understand the role of uncertainty visualization in weather forecasting, we designed and ran a pilot user-study aimed toward ecological validity by studying student "experts". The study was designed in a collaborative effort with perspectives from visualization and cognitive psychology. Furthermore, throughout the process we elicited and integrated feedback from an expert in atmospheric sciences. Weighing realism against experimental control was challenging, but we eventually settled on a design that we felt balanced our concerns satisfactorily.

The resulting 5 week longitudinal study was run with a group of 5 student forecasters during the summer of 2015. The participants were all self-selected members of the Ute Weather Center, an undergraduate club at the University of Utah which releases daily five-day forecasts for campus and several surrounding areas. The participants all had similar levels of *quasi-expertise*, sufficient for basic meteorological analysis. They were asked to forecast daily high temperatures for multiple locations, each of which had a corresponding weather station for verification of their predictions. We treated their individual forecasting process like a black box, collecting data on the information they used, but not controlling their natural forecasting process. For each of the middle three weeks in the study, the participants were asked to integrate an additional, provided uncertainty visualization product into their forecast. We compared the accuracy of their forecast judgments with and without these additional products. We also elicited qualitative feedback from the participants about the products through surveys at end of each week and an exit interview at the end of the study.

Because our aim was to keep the participants as close to their existing process as possible, we observed the workflows of several of our eventual participants when designing the study. Certain design decisions, such as what quantities the participants were asked to forecast and what data was used in the uncertainty visualization products

- *P. S. Quinan and M. Meyer are with the University of Utah School of Computing, E-mail: {psq,miriah}@cs.utah.edu.*
- *L. Padilla and S. Creem-Regehr are with the University of Utah Department of Psychology, E-mail: lace.m.k.padilla@gmail.com, sarah.creem@psych.utah.edu.*

Presented as part of the Visualization for Decision Making Under Uncertainty workshop at IEEE VIS 2015.

we provided, were directly derived from these workflow observations. Other decisions, like the number of days and locations the participants were asked to forecast each week, were decided through discussions with the participants to ensure that our study was not causing an undue amount of additional labor. We sought a straightforward task for the forecasters because we required a relatively large number of predications to offset the small number of participants.

The uncertainty visualization products that we tested were based on visualization techniques commonly used in meteorology: plume diagrams, mean and standard deviation plots, and spaghetti plots [13]. In all the products, the notion of uncertainty is derived from an **ensemble**, a collection of numerical weather prediction simulations designed to represent a set of possible forecast outcomes [14].

A number of decisions regarding our study design, however, were specifically made to strike a balance between our needs for both sufficient ecological validity and satisfactory experimental control. Specific examples include: our decision to run the study with quasi-expert student forecasters; our decision to not limit in any way what additional information a participant might choose to use in making their forecast; and our decision to use the same underlying data field for all the uncertainty products we tested.

4 Preliminary Results and Open Questions

In controlling for forecast location, individual differences in forecaster accuracy, and order effects, the study's preliminary results suggest differences between the visualization products and the baseline condition for forecast accuracy, as shown in Figure 1. While the trends in forecast accuracy generally seem to match the qualitative feedback that we received, more analysis is needed. Given that the experiment was less controlled than a typical user-study, a next step with the current data is to determine the most appropriate analysis methods.

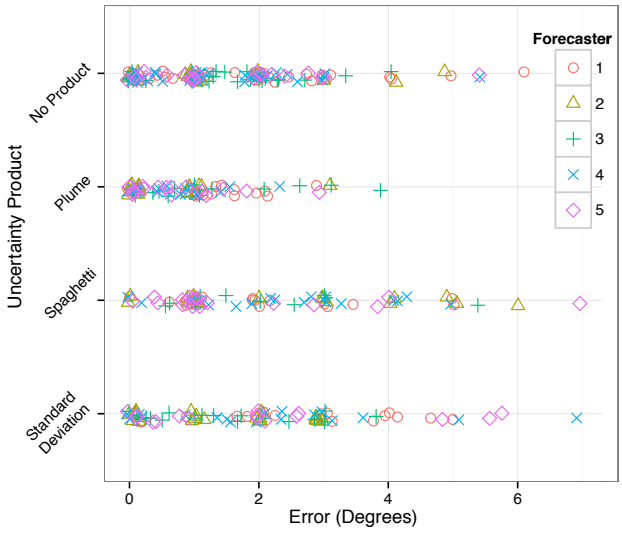

Fig. 1: The raw accuracy of the participants' forecast judgements.

We are currently discussing the appropriate modifications for expanding the pilot into a full scale study. Additionally, there remain some significant open questions, such as: to what extent will our results truly be applicable in real-world? How would one compare the results of our study to the body of existing uncertainty visualization user-studies? Is such a comparison even possible? Larger questions for the community discussion, include:

- How do we evaluate the scientific impact of studies that trade-off experimental control for ecological validity?
- Does the need for increased realism in user-studies generalize across domains that involve decision making?
- Are there other ways to increase realism and include expertise in the context of uncertainty visualization studies that still allow for good experimental control?

5 Conclusions

In this workshop paper we discuss some of the challenges facing studies of the effects of visualizing uncertainty on decision making. We argue for an increased emphasis on ecological validity in experimental design, while illustrating the problems with currently popular evaluation methodologies. Using the domain of weather forecasting as an example, we attempt to demonstrate why ecological validity is critical in studying decision-making. We describe an initial design for an ecologically valid user-study in weather forecasting and outline some core challenges and open questions. It is our sincere hope that this first step toward increased realism in controlled studies of uncertainty visualization can act as a springboard to a larger community discussion.

Acknowledgments

The authors wish thank Jim Steenburgh for his insightful feedback, and the Ute Weather Center for their participation. This work was funded by NSF grant IIS–1212806.

References

[1] B. Brown, S. Reeves, and S. Sherwood. Into the wild: Challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1657–1666, New York, NY, USA, 2011. ACM.

[2] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, pages 19–45. Springer Berlin Heidelberg, 2008.

[3] J. Cox, D. House, and M. Lindell. Visualizing uncertainty in predicted hurricane tracks. *International Journal of Uncertainty Quantification*, 3(2):143–156, 2013.

[4] S. Deitrick and R. Edsall. The influence of uncertainty visualization on decision making: An empirical evaluation. In A. Riedl, W. Kainz, and G. A. Elmes, editors, *Progress in Spatial Data Handling*. Springer Berlin Heidelberg, 2006.

[5] M. J. Intons-Peterson. Imagery paradigms: How vulnerable are they to experimenters' expectations? *Journal of Experimental Psychology: Human Perception and Performance*, 9(3):394, 1983.

[6] C. Kinkeldey, A. M. MacEachren, and J. Schiewe. How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51(4):372–386, 2014.

[7] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1520–1536, Sept 2012.

[8] J. M. Loomis, J. J. Blascovich, and A. C. Beall. Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, 31(4):557–564, 1999.

[9] R. K. Lowe. Components of expertise in the perception and interpretation of meteorological charts. In R. R. Hoffman and A. B. Markman, editors, *Interpreting Remote Sensing Imagery: Human Factors*. CRC Press, 2001.

[10] P. J. McCarthy, D. Ball, and W. Purcell. Project phoenix - optimizing the machine-person mix in high-impact weather forecasting. In *22nd Conference on Weather Analysis and Forecasting / 18th Conference on Numerical Weather Prediction*, Park City, UT, June 2007. Preprint.

[11] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, editors, *Human-Computer Interaction*, pages 152–169. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995.

[12] J. Nielsen and J. Levy. Measuring usability: Preference vs. performance. *Commun. ACM*, 37(4):66–75, Apr. 1994.

[13] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-Vis: A framework for the statistical visualization of ensemble data. In *IEEE International Conference on Data Mining Workshops*, pages 233–240, 2009.

[14] P. S. Quinan and M. Meyer. Visually comparing weather features in forecasts. *IEEE Transactions on Visualization and Computer Graphics*, 2015. Preprint.

[15] J. G. Trafton and R. R. Hoffman. Computer-aided visualization in meteorology. In R. R. Hoffman, editor, *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, Expertise: Research and Applications Series. Psychology Press, New York, NY, 2007.