

Evaluation of Data Display Methods in a Flash Flood Prediction Tool

Elizabeth M. Argyle¹(✉), Chen Ling², and Jonathan J. Gourley³

¹ Cooperative Institute for Mesoscale Meteorological Studies,
Norman, OK, USA
emargyle@ou.edu

² University of Akron, Akron, OH, USA
cl199@uakron.edu

³ National Severe Storms Laboratory/NOAA,
Norman, OK, USA
jj.gourley@noaa.gov

Abstract. Flash flooding can be difficult to predict using traditional, rainfall threshold-based approaches. New initiatives like the Flooded Locations and Simulated Hydrographs (FLASH) project provide real-time information using rainfall observations to force distributed hydrologic models to predict flash flooding events. However, in order to address the goal of creating a weather-ready nation, system designers must not only possess tools that relay useful information, but such tools must also be able to communicate environmental threats to stakeholders in a clear and easy-to-use interface. Where previous research has addressed the performance of forecasting models, the present study uses a human factors approach to enhance FLASH's ability to present information to decision-makers (i.e., forecasters).

Keywords: Flash flood · Decision Making · Forecasting

1 Introduction

The mission of the National Weather Service (NWS) is to “provide weather, water, and climate data, forecasts and warnings for the protection of life and property and enhancement of the national economy” [1]. One does not have to search very long or very hard in order to find examples of when forecasts have made a difference in public response to severe weather events. Events such as the May 20, 2013 tornado in Moore, Oklahoma or the September 2013 flash floods in Colorado show that timely forecasts can inform emergency managers, broadcasters, and the general public about weather-related risks.

Within the weather forecasting system, decisions must be made on many different scales. On a spatial scale, appropriate decisions may differ based on geographic region or between state or county borders. On a temporal scale, forecasters must synthesize information within strict deadlines each day; depending upon the role of the forecaster, forecasts may be for weather events in the very near term—i.e. 0–6 h forecasts, as in the case of severe weather—or in the longer term, such as for a

weekly forecast. In observations of forecasting operations, Morss and Ralph [2] found that lead time directly affected the weights that forecasters placed upon model guidance, model output statistics, radar observations, climatology, and personal background knowledge.

In addition, forecasters may apply events that match prototypes of weather phenomena more readily due to extensive training in class recognition [3]. Forecasters are trained to recognize “textbook” examples weather events, such as the synoptic patterns associated with tornado development [4]. This pattern matching technique is indicative of the representativeness heuristic, which can lead to errors if similar data sets are applied incorrectly to weather events. In a further example of this, Heinselman, LaDue, and Lazrus [5] found that a forecaster’s prior experiences with a particular type of weather event affected the forecaster’s ability to predict a similar weather event.

Another factor that may affect the forecaster’s reasoning and judgment about the weather condition is the display method used in weather decision-making products. In this study, we explore how different display methods used for flash flood prediction affect judgments of the severity level of the event.

2 Flooded Locations and Simulated Hydrographs (Flash)

Many weather decision-making products exist to give forecasters information about potential weather-related threats. The Flooded Locations and Simulated Hydrographs (FLASH) project is one such product for flash flooding prediction. FLASH is a real-time system that uses rainfall observations to force hydrologic models to predict flash floods. Potential users would be forecasters at both the national and regional scales in the United States, including, but not limited to, National Weather Service Weather Forecast Offices, River Forecast Offices, and national centers. Both at the national and regional scale, FLASH is designed to assist national forecasters identify areas of risk and then work with local forecasters to predict specific threats.

In July 2013, the Hydrometeorological Testbed at the Weather Prediction Center (HMT-WPC) hosted the first Flash Flooding and Intense Rainfall (FFaIR) experiment [6]. The purpose of the experiment was to evaluate the utility of a set of experimental forecast models, including FLASH, on a sample of professional forecasters and weather researchers. Over the three-week period, forecasters assessed standard and experimental computational models to create probabilistic forecasts of heavy rainfall and flash flooding events in the United States. As part of the experiment, the researchers observed forecaster behavior when creating the forecasts and identified patterns of information processing. Through daily observations of three independently acting forecasters, the researchers observed that the design of the information display affected how well forecasters were able to interpret the data modeled in FLASH.

Of particular interest was a suggestion that FLASH’s visual design led forecasters to predict false alarms. This comment was attributed to FLASH’s sampling method and visualization design. Specifically, FLASH divides the continental United States into a 1 km-by-1 km grid. The model calculates a measure of flash flood risk,

the return period, for every cell within the grid. However, when depicting the entire country, the system is not able to display each individual grid cell. The system developers solved this by developing an algorithm to sample the maximum grid cell value out of the collection of grid cells contained within one pixel, and the map of all the maximum values displays at the national level. In practice, this means that while the true predicted return period values are displayed when a viewer zooms in to a local level, the national view displays an adjusted value of the data by displaying the maximum value.

Forecaster comments from FFaIR led the researchers to hypothesize that a display algorithm which takes the average of sampled grid cells (henceforth called the average-based display) would produce different task performance than the maximum-based display. The present study seeks to identify differences in terms of error rate and task completion time when comparing two different display algorithms on the national-scale maps. It is hypothesized that the average method will produce fewer false alarms, but the maximum method will produce more hits. In terms of task completion time, it is hypothesized that the average display will take more time for participants to evaluate; it is thought that due to the color scheme and the larger size of represented regions, the design of the maximum-based display draws attention to events more rapidly than the average-based display would.

3 Method

3.1 Participants

A sample of 30 participants was recruited from the student and post-doctoral population at the University of Oklahoma. Participants were required to have a degree in meteorology or atmospheric science, or to be currently pursuing one. Though participants had little experience working with FLASH, this expectation ensured that they had adequate experience with reading weather prediction visualizations.

The sample consisted of 19 males and 11 females between the ages of 21–41 years old, with an average age of 25.0 years. Participants were randomly assigned to one of the between-subjects display conditions (the maximum-based algorithm or the average-based algorithm).

3.2 Materials

A set of 40 image pairs was created by taking screen captures of FLASH; each image pair consisted of one image of FLASH at a national, full-view level, and a second image of the same date and time, but zoomed in to a local, county- or state-level. It is important to note that while participants in the two display groups viewed different representations of the weather event at the national scale, the local images that participants viewed were identical between groups.

The dates and times were selected based on flash flooding events that were reported between April and July 2013 in the National Climatic Data Center Storm Events Database. When selecting the events from the database, the researcher categorized

events into “severe” and “not severe” flash flooding. Unlike tornado events and the Fujita scale, there is not yet a standardized scale for flash flooding severity, so the research team defined severe flash flooding to be those that caused \$500,000 or more of property and crop damage ($n = 20$, $\mu = \$9.86 \text{ M}$; $\sigma = \$22.33 \text{ M}$). Events that were placed in the “not severe” category had less than \$500,000 of property and crop damage ($n = 20$, $\mu = \$38.75 \text{ K}$; $\sigma = \$84.59 \text{ K}$).

Images were randomly presented to participants using PsychoPy [7], an open-source software which allows researchers to present stimuli and collect data from participants. Each evaluation was conducted on an Asus A53U laptop with a 15-inch screen; each image was displayed at a size of 869×680 pixels.

3.3 Experimental Design

The between-subjects independent variable in the present study was the display algorithm which varied across two levels—the maximum-based display and the average-based display. Using a Signal Detection Theory framework [8], the error rates were calculated using a detection task. Though property damage was used as a measure of severity to select the image pairs, it was not an independent variable itself—within the study framework, participants had to detect which images represented severe events. Likewise, while participants viewed images at varied spatial scales, the local images were identical no matter which display algorithm each participant viewed at the national level.

3.4 Procedure

At the beginning of the study, participants were informed about the study’s purpose and tasks. After completing an informed consent form, participants received an excerpt from the FLASH training manual that explained how to read and interpret the FLASH display. During the instruction stage, participants were given the opportunity to ask questions about FLASH, how to interpret the display, and what the study would involve.

Once participants stated that they felt comfortable with the FLASH interface, they answered a series of demographic questions (age, gender, and academic classification). Following this, a series of image pairs was presented in a randomized order. In each image pair, the first image showed an event in FLASH on the national-level (see Fig. 1A). Participants were asked, “Based on the information that is modeled in this image, would you expect for this event to produce flash flooding with severe levels of property damage? ($> \$500,000$).” Participants reviewed the image, and then pressed “y” for yes or “n” for no after making their decision. The next image would be a representation of the same weather event, but the image would be zoomed in to a local scale (see Fig. 1C). The participants answered the same question about severity based on the new presentation of information. All forty image pairs were presented in a randomized order. When participants finished with the last pair, they were debriefed (Fig. 2).

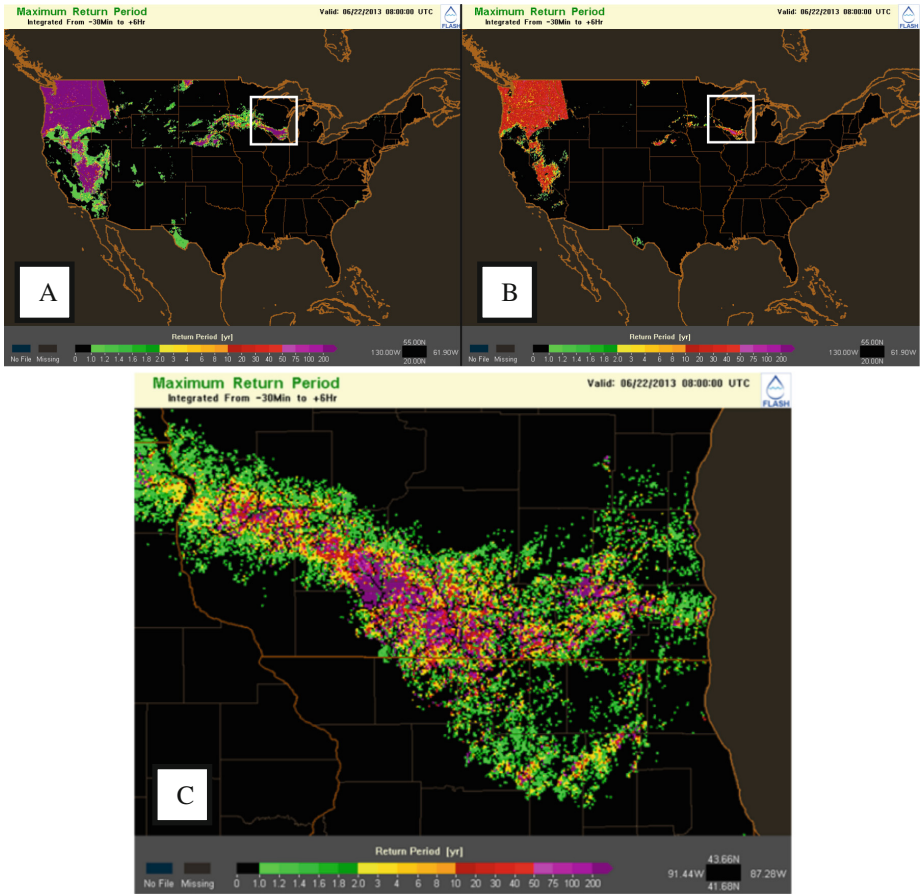


Fig. 1. Images showing the same severe hydrometeorological event, displayed at the national scale with the maximum-based display algorithm (A), the average-based display algorithm (B), and the local view associated with both (C).

4 Results

4.1 Error Rates

After collecting the participants' responses, an error rate in terms of the Signal Detection Theory framework [8] was calculated for the severity judgment associated with the average-based and maximum-based display styles and for the national and local images. The data were compared using t-tests. The results show that there is a significant difference between the display methods. A summary of the results is shown in Table 1. The maximum display produced a higher hit rate than the average display, but the average display minimizes false alarms.

Similarly, an analysis of participant judgments for the local-level images used the Signal Detection Framework. Though all subjects saw the same images in this

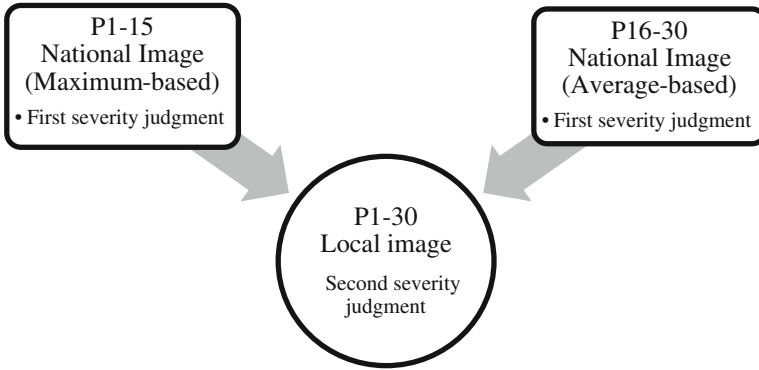


Fig. 2. Participants viewed pairs of national and local images of identical weather events. Half the participants viewed national images visualized with the maximum-based algorithm and the other half with the average-based algorithm.

Table 1. Comparison of average-based and maximum-based display types in terms error rates

	Hit rate	False alarm rate
Average-based	0.57	0.25
Maximum-based	0.85	0.50
p-value	<0.0001	<0.0001

Table 2. Error rates for viewing the local-level events (n = 30)

	Hit rate	Miss rate	Correct rejection rate	False alarm rate
Local	0.50	0.50	0.76	0.24

category, responses were compared between the maximum-based and the average-based participant groups in order to ensure parity. As expected, a t-test found no significant difference between how participants in either test group when judging the local-level images. Still, as shown in Table 2, participants did not make perfect judgments, which may in part be due to lack of participant experience with flash flood forecasting.

4.2 Task Completion Time

Task time was recorded from the time the national display was shown to the participant still they made a severity judgment on the national image. Though the original hypothesis was that the average display would produce slower response times than the maximum display, no significant difference was found between the two. A further analysis of task completion time for the time taken during hits, misses, false alarms, and correct rejections also failed to find any statistically significant differences. A summary of results is shown in Table 3.

Table 3 Average time (in seconds) taken to produce a hit, miss, correct rejection, or false alarm, analyzed with a t-test.

	Hit	Miss	Correct rejection	False alarm
Average-based	4.02	4.49	4.34	4.12
Maximum-based	4.63	5.23	5.81	5.45
p-value	0.58	0.81	0.12	0.57

5 Discussion and Conclusions

The results of this study show that there is a significant difference between display styles in terms of error rates, but not in terms of task completion time. Though the original hypothesis was that the average display would cause participants to review the image for a longer period of time, this in fact was not observed. When examining the images that participants commonly had trouble judging correctly, common causes of confusion occurred for events that had particularly different visual representations between the national and local level. For example, one event looked like a very small storm when visualized with the national-level average-based algorithm, but was actually very severe looking after zooming closer. Participants often judged the national image to be insignificant, but changed their minds after viewing the local level.

Design recommendations based on these results for future weather information displays must rely on the risk management values of the system designers. While the maximum display style maximized hits, it also produced many more false alarms than the average display. In weather forecasting, excess numbers of false alarms can consume valuable time that forecasters could be using to analyze true threats. However, while the average display style produced fewer false alarms, participants were much more likely to miss an event; this could also result in critical consequences.

In the case of a flash flooding prediction system such as FLASH, the recommendation from these results would be to use the maximum display algorithm. Flash flooding is by nature a rapidly occurring event that can have life-threatening consequences if not predicted with enough lead time. For such a system, having a design that promotes more hits, even at the expense of producing false alarms, would ensure that forecasters' attentions would be drawn to severe events in a timely manner.

In the future, more studies should examine the effects of a forecaster's first severity decision on the second severity decision; for example, understanding how the viewing the national map first affects how the forecaster interprets the local level information may shed light on other aspects of the forecaster's decision-making process. Additionally, future work could address limitations of the present study; namely, while participants all had some background in meteorology and forecasting, few had specifically studied flash flood forecasting. A similar study to the present work, but run with a sample of professional flood forecasters may produce new and valuable information.

References

1. National Weather Service Strategic Plan. http://www.nws.noaa.gov/com/stratplan/files/plan_v01.pdf
2. Morss, R.E., Ralph, F.M.: Use of information by national weather service forecasters and emergency managers during CALJET and PACJET-2001. *Weather Forecast.* **22**, 539–555 (2007)
3. Doswell III, C.A.: Weather forecasting by humans—heuristics and decision making. *Weather Forecast.* **19**, 1115–1126 (2004)
4. Johns, R.H., Doswell III, C.A.: Severe local storms forecasting. *Weather Forecast.* **7**, 588–612 (1992)
5. Heinselman, P.L., LaDue, D.S., Lazrus, H.: Exploring impacts of rapid-scan radar data on nws warning decisions. *Weather Forecast.* **27**, 1031–1044 (2012)
6. Barthold, F.E., Workoff, T.E., Cosgrove, B.A., Gourley, J.J., Novak, D.R., Mahoney, K.M.: improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Am. Meteorol. Soc.* e-View (2015). <http://dx.doi.org/10.1175/BAMS-D-14-00201.1>
7. Peirce, J.W.: PsychoPy - psychophysics software in python. *J. Neurosci. Meth.* **162**, 8–13 (2007)
8. Wickens, T.D.: *Elementary Signal Detection Theory*. Oxford University Press, Oxford (2001)