Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE **Uncertainty of River Forecasts: Quantification and Implications for Decision- Making in Emergency Management**

PRESENTED BY **Frauke Hoss**

ACCEPTED BY THE DEPARTMENT OF

Engineering and Public Policy

Paul Fischbeck	September 2, 2014
ADVISOR, MAJOR PROFESSOR	DATE
Douglas Sicker	September 2, 2014
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Vijayakumar Bhagavatula	September 2, 2014
DEAN	DATE

Uncertainty in River Forecasts:
Quantification and Implications for Decision-Making in Emergency Management

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering & Public Policy

Frauke Hoss

MSc, Civil Engineering, Delft University of Technology
BSc, Political Science, Leiden University
BSc, Civil Engineering, Delft University of Technology

Carnegie Mellon University
Pittsburgh, PA

December 2014

Acknowledgements

People

I would like to thank my thesis committee for the support each of them gave: Paul Fischbeck, my daily adviser, is not easily satisfied with any data analysis. He kept pushing me to look deeper into the data, which resulted in much better papers. I am deeply grateful to Granger Morgan for his compassion and unreserved support while remaining critical. Kelly Klima has helped me navigate graduate student life and kept up my morale in rocky times. Finally, I appreciate working with Costa Samaras for his sharp criticism and pragmatism.

Organizing the SUCCEED summer program really made me feel part of EPP. It was a great experience and a lot of fun. Thanks, Paul Welle, for running it with me and becoming a good friend. Without the support of the other PhD students, Kelly Klima, and Inês Azevedo we could not have done it.

A very big Thank You to the EPP staff – Vicki, Patti, Adam, Steve, Barb, Elizabeth, Jaime, and Deb – for creating the welcoming environment at the office in Baker Hall. If the department had not become home so quickly, I wonder if Pittsburgh ever had.

The other EPP graduate students probably taught me more than any of the courses. Special thanks to Parth Vaishnav for making me read and think outside my bubble, and Mili-Ann Tamayao for making our house a home.

Moving to a city in the far-away Pennsylvanian forest wasn't easy. Who would have thought that I would be sad to leave this green, comfortable place? Most of all, I will miss Al Rose's Sailfish Masters swim team, the most fun and hard-working swim team I have been on in years.

Briana and Traci, I hope to see you in Boston!

Finally, I am incredibly lucky to have a rock-solid home front. When I moved to Pittsburgh, my parents surprisingly quickly found their way into the cyber world. Now they always immediately grab their iPad to skype, when I need a listening ear. While living here I talked to them more than I did, when I still lived closer by. They also visited me every single year.

Finally, I want take the opportunity to thank Pieter for all the shrink sessions we have had on Skype. Pieter and I flew many thousands of miles and skyped many hours to make this life work. Dank je wel, skattie!

Funding

I was supported by an ERP fellowship of the German National Academic Foundation and by the Center of Climate and Energy Decision Making (SES-0949710), through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (CMU).

Abstract

This dissertation focuses on (river) forecasting, but also includes a study on stormwater treatment. Using forecasts for decision-making is complicated by their inherent uncertainty. An interview-based study qualitatively and a survey empirically investigate forecast use in emergency management. Emergency managers perceive uncertainty as a given rather than as a problem. To cope with the uncertainty, decision-makers gather as much information as possible; forecasts are only one piece of information among many. For decision-making, emergency managers say that they rely more on radar than on river forecasting. However, forecasts play an important role in communication with the public, because they are the official interpretation of the situation. Emergency managers can add a lot of value to those forecasts by combining them with local knowledge, but might not do so because of accountability concerns. Forecasts must have value to emergency managers, because those with more work experience rely more on them than those without.

Another study further develops the application of quantile regression to generate probabilistic river forecasts. Compared to existing research, this study includes a larger number of river gages; includes more independent variables; and studies longer lead times. Additionally, it is the first to apply this method to the U.S. American context. It was found that the model has to be customized for each river gage for extremely high event thresholds. For other thresholds and across lead times, a one-size-fits-all model suffices. The model performance is robust to the size of the training dataset, but depends on the year, the river gage, lead time and event threshold that are being forecast.

An additional study considers the robustness of stormwater management to the amount of runoff. Impervious surfaces, such as roads and parking lots, can increase the amount of runoff and lead to more pollution reaching streams, rivers, and lakes. Best Management Practices (BMPs)

reduce the peak discharge into the storm sewer system and remove pollutants such as sediments, phosphorus and nitrogen from the stormwater runoff. Empirically, it is found that BMP effectiveness decreases sooner, steeper and deeper with increasing sizes of storm events than assumed in current computer models.

Table of Contents

List of Tables

STUDY 2

STUDY 3

STUDY 4

List of Figures

STUDY 4

Abbreviations

ABRFC – Arkansas-Red Basin River Forecast Center
ADEM – Arkansas Department of Emergency Management
ANOVA – Analysis of Variance
BAMS – Bulletin of the American Meteorological Society
BMP – Best Management Practice
BS – Brier Score
BSS – Brier Skill Score
CBP – Chesapeake Bay Program
EM – Emergency Manager
EMA – Emergency Management Association
EMI – Emergency Management Institute
EMS – Emergency Management Services
EOC – Emergency Operations Center
EPA – Environmental Protection Agency
FEMA – Federal Emergency Management Association
FTCA – Federal Tort Claim Act
HSPF – Hydrologic Simulation Program – Fortran
IAEM – International Association of Emergency Managers
LID – Low-Impact Development
MANOVA – Multivariate Analysis of Variance
M-M – Michaelis-Menten
NCRFC – North Central River Forecast Center
NIMS – National Incident Management System
NOAA – National Oceanic and Atmospheric Administration
NWS – National Weather Service
OCS – Oklahoma Climatological Survey
PBC – Perceived Behavioral Control
PCA – Principal Component Analysis
PEMA – Pennsylvania Emergency Management Association
POP – Probability of Precipitation
QPF – Quantitative Precipitation Forecast
QR – Quantile Regression
RFC – River Forecast Center
ROC – Relative Operating Characteristic
RPSS – Ranked Probability Skill Score
RSME – Root Squared Mean Error
SNS – Subjective Numeracy Scale
SUSTAIN – System for Urban Stormwater Treatment and Analysis Integration Model
TMDL – Total Mean Daily Load
TPB – Theory of Planned Behavior
USACE – U.S. Army Corps of Engineers
USGS – U.S. Geological Survey
WFO – Weather Forecast Office
WIP – Watershed Implementation Plan

INTRODUCTION

Weather forecasts are intended to provide their users lead time to prepare for the weather events to come. In the case of emergency managers, weather forecasts can give the opportunity to evacuate hazardous areas, prepare equipment to be able to respond quickly or to protect or to move valuable property before a severe weather event hits.

In practice, a number of factors impact to what extent this lead time can actually be utilized. First, there are the characteristics of the forecast itself. Forecast errors are inevitable and increase with lead time and infrequency of the expected event. Additionally, the forecasts have to be released in time and cover the areas that presently are in danger. Frequent changes in the predicted event magnitude and timing can inhibit efficient preparation activities. Second, the forecast user has to be able of making use of the lead time that the forecast potentially could provide. (S)he needs to be capable of understanding the forecast and applying it to the situation and tasks at hand. Additionally, the emergency manager has to have access to resources to implement appropriate preparation activities. Third, there are social considerations to take into account. Emergency managers possess an abundance of local knowledge that forecasts, e.g., from the National Weather Service, are not able to capture. While the emergency managers make his/her own assessment of the situation, the forecast serves as a standard. This standard informs the emergency managers' assessment, but gains renewed importance in terms of accountability if there is controversy after the event, e.g., the media looking for a scapegoat for things that have gone wrong. This dissertation sheds light on these three factors.

1.1. River Forecasts

Large forecast providers such as the National Weather Service (NWS) have an extensive infrastructure in place to collect data, compute forecasts and to distribute them. For example, in the case of river-stage forecasts, a relatively small branch of the NWS, thirteen river forecast centers (RFCs) across the nation compute the forecasts and forward them to one of the ~120 Weather Offices which then publish them through a variety of channels.

The obstacles to utilize the lead time that forecasts provide also apply to river-stage forecasting. Most importantly however, the NWS does not routinely publish any information about the expected forecast error for their short-term river-stage forecasts (for the next few days and hours). While no such case for short-term weather forecasts has been documented in the literature, other types of deterministic forecasts have certainly led to dangerous and expensive misunderstandings. The most prominent example is Grand Forks, ND. In April 1997, the people of that town along the Red River prepared their city for the forecasted river crest without understanding the uncertainty inherent in the NWS Outlook (seasonal forecast) on which they relied (Pielke, 1999; Morss, 2010). Even though the relative error turned out to be less than the average relative error, because of the lack of preparedness, the flood damage amounted to between $1 and 2 billion with many people blaming the NWS. According to Pielke (1999), the officials as well as the public "misused" the forecast by assuming that the forecast would be correct; often claiming that they would have made different decisions, if they had been aware of the potential forecast error.

In recent years, the research community and NWS have been engaged in a number of projects to improve their river-stage forecast products. The following developments are the most relevant to this dissertation: First, there has been an effort to verify the forecasts (e.g., Demargne et al., 2009; NOAA, 2006). While the hydrological and hydraulic models have been thoroughly calibrated and verified for decades, verification of the forecasts themselves was not possible because they were not being archived.[1] As Welles et al. (2007) demonstrated, this forecast verification is

[1] If they were archived at all, they were often saved as maps or graphics which are much more difficult to analyze than the underlying text files.

highly desirable. They found that forecast in Oklahoma (1993-2002) and along the Missouri River (1983-2002) did better than the persistence forecast[2] for only one day of lead time for above-flood stage water levels and up to three days of lead time for below-flood stage water levels. Additionally, they reported that the forecast skill had not improved in the past ten to twenty years, despite various updates in the forecasting process (Welles et al., 2007). While verification metrics are a measure of uncertainty that would be valuable for the forecast user, i.e., the public, it seems as if the NWS' River Forecast Verification Plan is mainly an internal effort to produce more accurate forecasts (NOAA, 2006).

Second, ensemble forecasting is slowly being implemented, so that the forecast uncertainty resulting from the major sources of uncertainty can be communicated to the user. This is done by running the forecast models several times with different input values. Together the output of these runs can be used as an uncertainty estimate (National Weather Service, 2012). So far, ensemble forecasting is only being used for seasonal forecasts in some RFCs with rather poor visualization (see figures in Study 3 for illustration). Ensemble forecasting has not been implemented for short-term weather forecasts yet, but that is planned as part of the implementation of the Advanced Hydrologic Prediction Service (NOAA, 2001).

Third, there have been a limited number of studies to make the forecasts more user-friendly. Such efforts include studying which pieces of information should be part of the forecast to facilitate decision-making. For example, Verkade and Werner (2011) find based on an hydro-economic expected annual damage model that probability forecasts lead to a lower residual flood risk than deterministic forecasts. Another effort is formatting forecasts in a way that the inevitable forecast

[2] A persistence forecast assumes that the water level is going to stay the same as the water level on the day when the forecast is published.

uncertainty is better understood. For example, Leedal et al. (2010) proposes methods to visualize the uncertainty associated with simulations of river levels in the United Kingdom.

1.2. Forecasts in Emergency Management

Studying the use of forecasts in emergency management is difficult, because of a number of reasons. No town is like another; different regions are subject to different weather hazards; each state has different regulations. Emergencies themselves are difficult to study, because they are usually unique in nature and there is often no control group. Additionally, the decision-making process during an emergency is usually not recorded, because it takes valuable time away from preventing damage and bringing people to safety. Consequently, it is often difficult to determine afterwards what consideration led to which decision, and whether that was a good decision. Unsurprisingly, this field is still very much subject to research.

As outlined above, there is a trend in the industry towards probabilistic forecasts. From an engineer's or scientist's perspective, those should be better than the deterministic ones, because they better describe the "true" weather conditions. But do such technically more sophisticated forecasts lead to better decisions?

Using forecasts for decision-making is a true case of decision-making under uncertainty, regardless of whether forecasts are deterministic or probabilistic. People consciously and unconsciously cope with this uncertainty in various ways. For example, Morss (2010) finds that experience drives expectations and therefore preparations for extreme weather events such as floods. If the event deviates from previous ones, people are often caught off-guard (e.g., representative, availability, and anchoring heuristics; Tversky, Kahnemann, 1974). The devastating flood at Grand Forks, ND, mentioned earlier, is such an example (Pielke, 1999; Morss, 2010).

Including the forecast uncertainty in the published forecast could be a solution to prevent misunderstandings that have occurred with deterministic forecasts. Morss et al. (2010) find in a survey of the general U.S. public, that most people are aware of the uncertainty in weather forecasts and can adequately interpret probabilistic forecasts. Nonetheless, people take action at different probability thresholds (Morss et al., 2010). Emergency managers develop methods to cope with such uncertainty. For example, they gather many types of information to triangulate their expectations or choose products that are less uncertain (Baumgart et al., 2007). But only about 30% of all courses that emergency managers take cover weather-related topics, even though just under 80% of all emergencies are caused by the weather (Weaver et al., 2014).

1.3. Overview Dissertation

This dissertation adds three studies to the efforts to extract greater value from (river) forecasts in emergency management. Two of the studies focus on river forecasting; one study is also applicable to other severe weather events. A fourth study on removing pollutants from urban stormwater was included to add to the portfolio of analytical skills.

The purpose of the **first** study is to better understand the role of (river) forecasts in emergency management. It consists of seventeen in-person interviews with emergency managers in Pennsylvania, Oklahoma, and Arkansas including questions on training, daily routine, emergency operations, forecasts, and forecast uncertainty. Rather than being end-users, emergency managers disseminate the forecasts to the fire, police and city departments, industry and the residents who then take action. In summary, all of the fifteen emergency managers who used river forecasts were well aware of the often substantial forecast uncertainty. For this reason, they rely mostly on radar that informs them of precipitation upstream and on real-time gage data to judge the situation. Even though they might not fully trust the published NWS forecasts themselves, they tend to disseminate them to the public without adding their interpretations because of a number of concerns. Although they hold back their own interpretation of uncertainty, the emergency managers translate the forecast to local circumstances, suggest actions to take, and use their local network and reputation to make people act. A number of measures are identified without which replacing deterministic with probabilistic forecasts is unlikely to improve emergency responses.

The **second** study is an empirical extension of the first study. In an online survey, ca. 200 emergency managers from across the U.S. answered questions about the past and intended forecast use, perceived limitations, their attitude towards forecasts and their jobs, social norms and subjective numeracy. This study included river floods, flash floods, tornadoes, hurricanes, ice and snow storms and heat waves. Work experience turns out to be the best predictor of the extent to which an emergency manager relies on forecasts and recorded weather data. Regarding the use of forecasts only, the attitude towards the weather information and having received instructions on using weather information were significant predictors. Perceived unavailability for the area, insufficient information and irregular release times and a preventive mindset tend to decrease the reliance on recorded weather data.

The **third** study is of a more technical nature. Currently, the NWS publishes their short-term river-stage forecasts (for the next few days and hours) without any uncertainty information. With the intention to devise a computationally cheap remedy, this study applies quantile regression (QR) to river-stage forecasts in order to generate probabilistic forecasts for events, defined by water levels exceeding certain thresholds. Earlier implementations of QR used the forecast itself as the only independent variable, focused on few stations in the United Kingdom, and hours of lead time (Weerts et al., 2011; López et al., 2014). This study extends the method to a larger number of river gages in the U.S. American context, i.e., to 82 river gages of North Central River Forecast Center;

includes more independent variables and a larger dataset; and studies up to four days of lead time. It was found that the model has to be customized for each river gage for extremely high event thresholds. For other thresholds and across lead times, a one-size-fits-all model suffices. Additionally, the approach was tested for robustness. Forecast quality does not depend on the size of the training dataset, but on the year, the river gage, lead time and event threshold that are being forecast.

A **fourth** study deals with hydrology and uncertainty as well, but it is unrelated to forecasting. Instead, the study focuses on methods to remove pollutants from urban stormwater. Impervious surfaces, such as roads and parking lots, can increase the amount of runoff and lead to more pollution reaching streams, rivers, and lakes. Best Management Practices (BMPs) reduce the peak discharge into the storm sewer system and remove pollutants such as sediments, phosphorus and nitrogen from the stormwater runoff. Bioretentions, dry and wet ponds, porous pavement, and many others collect the runoff and reduce the concentrations of sediments and nutrients with varying effectiveness. Empirically, it is found that BMP effectiveness decreases sooner, steeper and deeper with increasing runoff volume than assumed in current computer models.

References

Baumgart, L.A., Bass, E.J., Philips, B., Kloesel, K. (2007): "Emergency Management Decision Making during Severe Weather." *Weather and Forecasting* Vol. 23: pp.1268-1279.

Demargne, J., Mullusky, M., Werner, K., Adams, T., Lindsey, S., Schwein, N., ... & Welles, E. (2009). Application of forecast verification science to operational river forecasting in the US National Weather Service. *Bulletin of the American Meteorological Society*, *90*(6).

Leedal, D., Neal, J., Beven, K., Young, P., & Bates, P. (2010). Visualization approaches for communicating real-time flood forecasting level and inundation information. *Journal of Flood Risk Management*, *3*(2), 140-150.

López, P., Verkade, J. S., Weerts, A. H., & Solomatine, D. P. (2014). Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. *Hydrology and Earth System Sciences Discussions*, *11*(4), 3811-3855.

Morss, R.E. (2010): "Interaction among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases." *Natural Hazard Review,* August 2010: 83-96.

Morss, R.E., Lazo, J.K., Demuth, J.L. (2010): "Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication." *Meteorological Applications* Vol. 17: pp. 149-162.

National Oceanic and Atmospheric Administration (2006): *National Weather Service River Forecast Verification Plan.* Report of the Hydrologic Verification System Requirements Team, October 2006, Silver Spring, Maryland. URL [accessed 4/25/2014]: http://www.nws.noaa.gov/oh/rfcdev/docs/Final_Verification_Report.pdf

National Oceanic and Atmospheric Administration (2001): *Implementation of Advanced Hydrologic Prediction Service.* National Weather Service, August 2001. URL [accessed 4/25/2014]: http://www.nws.noaa.gov/oh/ahps/Implementation_Plan.pdf

National Weather Service (2012): *Overview of the Hydrologic Ensemble Forecast Service (HEFS)*, Release 0.1.2, June 2012, Office for Hydrologic Development. URL [accessed 4/25/2014]: http://www.nws.noaa.gov/os/water/RFC_support/HEFS_doc/HEFS_Overview_0.1.2.pdf

Pielke, R.A. (1999): "Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood", *Applied Behavioral Science Review*, 7(2), pp.83–101.

Verkade, J. S., & Werner, M. G. F. (2011). Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology & Earth System Sciences Discussions*, *8*(4).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Weaver, John, Lindsey C. Harkabus, Jeffry Braun, Steven Miller, Rob Cox, John Griffith, Rebecca J. Mazur, 2014: An Overview of a Demographic Study of United States Emergency Managers. *Bull. Amer. Meteor. Soc.*, **95**, 199–203.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S. (2011): "Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)", Hydrol. Earth Syst. Sci., 15, 255-265, doi:10.5194/hess-15-255-2011, 2011.

Welles, E., Sorooshian, S., Carter, G. and B. Olsen, 2007. Hydrologic Verification: A Call for Action and Collaboration. *Bulletin of the American Meteorological Society*, **88(4),** 503–511.

STUDY 1

How can probabilistic short-term river forecasts be designed to be useful for emergency management?[3]

Abstract

The National Weather Service (NWS) is developing probabilistic short-term river forecasts (up to five days ahead). Little has been done to understand how the reported uncertainties will be used by decision makers such as disaster responders. If this is to happen, NWS needs to understand its clients. For this study, seventeen emergency managers were interviewed and asked to describe their use of river forecasts. Rather than being end-users, emergency managers disseminate the forecasts to the fire, police and city departments, industry and the residents who then need to take action.

All of the fifteen emergency managers who used river forecasts were well aware of the often substantial forecast uncertainty. For this reason, they rely mostly on radar that informs them of precipitation upstream and on real-time gage data to judge the situation. Even though they might not trust NWS forecasts themselves, they disseminate them to the public without adding their interpretations of them because of a number of concerns. Although they hold back their own interpretation of uncertainty, the emergency managers translate the forecast to local circumstances, suggest actions to take, and use their local network and reputation to make people act. A number of measures are identified without which publishing forecast uncertainty alone is unlikely to improve emergency responses.

Keywords: National Weather Service, forecasts, river, uncertainty, emergency management

---

[3] A version of this chapter has been accepted by the Bulletin of the American Meteorological Society (BAMS) in August 2014.

1. Introduction

The twelve River Forecast Centers (RFCs) of the National Weather Service (NWS) perform a large-scale, technically-sophisticated effort to publish river forecasts for ~4,000 river gauges in the U.S. The National Hydrologic Warning Council (2002) has conservatively estimated that short-term weather forecasts (up to five days) reduce the average annual flood loss by $433 million (2000 cost levels, excluding saved lives), which is 10% of the actual flood damage. However, four characteristics of the system undermine the practical value of these *short-term weather forecast*s.

First, because of the data sparseness inherent to flood prediction, above-flood stage errors (e.g., root mean squared error) of several feet are not uncommon for short-term weather forecasts (Welles et al. 2007, Pielke 1999). This has been the case for decades (Welles et al. 2007). Second, this uncertainty is not being communicated by the NWS. The published short-term weather forecasts do not include an uncertainty range, an indication of historical average error, or any similar measure of uncertainty (Figure 1). Third, it is extremely difficult for the forecast user to estimate the expected forecast error themselves. Because of the infrequency of floods, most people cannot draw on experience to quantify the errors. Additionally, the uncertainty in river forecasts is complex. The expected absolute error (e.g., in feet) increases with forecast length and with water level. Longer forecasts are more uncertain. People do not seem to be aware that forecasts predicting high river stages – when forecasts are needed most – are subject to more uncertainty than forecasts predicting low river stage (Morss & Wahl 2007). Fourth, the RFCs have little contact with the end-users. The RFCs rarely receive feedback on how well their products serve their clients' purposes, because the ~120 NWS Weather Forecast Offices (WFOs) rather than the RFCs officially publish the forecast.[4] In short, what Parker & Handmer observed in 1998 still seems true today: "Prediction agencies have

[4] The RFCs' "guidance" product (rather than the official forecasts by the WFOs) is also published on the RFCs' websites.

often simply 'assumed' that their forecasts are conveyed to those at risk, that local needs are met and that appropriate adaptive behavior ensues."

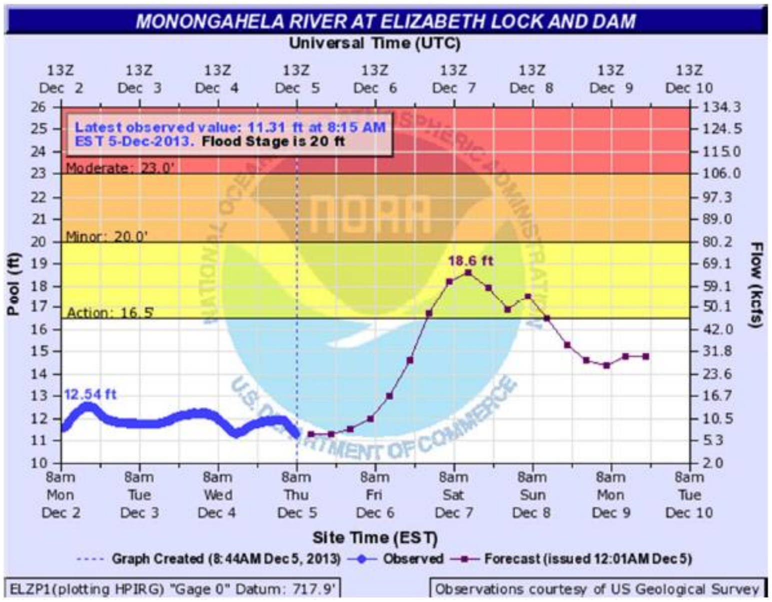


Figure 1: Example of a National Weather Service river stage forecast. Forecast for the Monogahela River at Elizabeth Lock and Dam, PA on 12/05/2013. The blue, thick line stands for the observed river stage height in the past few days. The purple dotted line represents the river forecast in six-hour intervals for the next few days. Action, minor flood and moderate flood stage indicate the different degrees of calamity. The forecast is also available in table and other formats and over the radio. In all versions, any measure of uncertainty or forecasting errors is omitted. RFCs have started or are planning to include such information. In this case, the river crested on 12/7 at 22.30pm at 17.98 feet. Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=pbz&gage=elzp1

It is likely that the above issues lead to a reduced usefulness and consequently reduced use of river forecasts. Deterministic forecasts have certainly led to dangerous and expensive misunderstandings. In April 1997, people in Grand Forks ND prepared their city for the forecasted river crest without understanding the uncertainty inherent in the NWS Outlook (seasonal forecast) on which they had relied. Even though the relative error turned out to be less than the average relative error, because of the lack of preparedness, the flood damage amounted to between $1 and 2 billion with many people blaming the NWS. According to Pielke (1999), the officials as well as the public "misused" the forecast by assuming that the forecast would be correct; often claiming that they would have made different decisions if they had been aware of the potential forecast error.

1.1. Objectives

Over the past 15 years, the NWS recognized this problem and is developing probabilistic short-term river forecasts (e.g., NWS 2011). Probabilistic seasonal river forecasts, called "outlooks," are already being communicated, while short-term river forecasts are still in development. The substantial body of literature on the communication of uncertainty suggests that this will not be an easy endeavor. This article adds to this effort by examining what role river forecasts and their inherent uncertainty play in the work of one group of NWS clients, emergency managers[5](EMs), who use short-term river forecasts to prepare their jurisdiction for an approaching flood. Based on in-depth interviews with EMs, this article investigates how short-term river forecasts that include uncertainty information (such as probabilistic forecasts) would serve EMs and what NWS can additionally do to make this innovation a success. The first part of the article describes the emergency managers and their profession. The second part investigates whether EMs use river forecasts, whether they understand the forecast uncertainty and how they cope with that uncertainty. In the third part, the role that EMs play in the dissemination of forecast information to the actual decision-makers, e.g., house owners and companies, is outlined.

1.2. Sample group

After contacting 45 emergency managers (EMs), seventeen in-person interviews with EMs in towns along rivers in Pennsylvania (7), Oklahoma (7) and Arkansas (3) were conducted and their recordings transcribed. The interviews were conversations averaging 50 minutes that were semi-

---

[5] Sometimes also called "emergency management coordinator", "emergency management director" or similar.

structured using prepared questions on training, daily routine, emergency operations, forecasts and forecast uncertainty. The interviews give an impression of the reality on the ground, but are not representative of the entire U.S. The experiences of EMs in jurisdictions along the Mississippi, where floods can be anticipated weeks in advance, are likely to be very different from those presented here. Quotes in italics from the interviews will illustrate the observations and arguments made in this article. Each section ends with a recommendation printed in bold.

Conservatively estimated, the interviewed EMs had an average work experience of twelve years. Four of them were not paid for their efforts as EM, seven were full-time emergency managers. For the remaining six, emergency management was an extra hat that they wear as part of their full-time profession. Three of the seventeen EMs were female; fifteen were responsible for a town or borough and two for a county.

Sixteen out of seventeen interviewed EMs knew the NWS website and how to access the river forecast through various channels. Fifteen interviewees made use of the forecast, only two had trouble describing it.

2. What do emergency managers do? Who are they?

> *"One thing to understand about emergency management, it is more an idea than it is a fact. If you tell me you are a fire fighter out in California, I generally know what you do. If you say, I am a cop in Arizona, I generally know what you do. If you tell me, I'm emergency management from Iowa, I know you probably have something with communications; you definitely have something with plans. But what else you do…"*

Each state makes their own requirements for emergency managers. Each political subdivision of the Commonwealth of Pennsylvania and in the state of Arkansas (i.e., counties and municipal governments) and each incorporated jurisdiction in Oklahoma is required to have an emergency management program (Pennsylvania's Emergency Management Services Code, 35 Pa. C.

S. Section 7501; Ark. Code 12-75-118; Oklahoma Statues – Title 63 Public Health and Safety Section 63-683.11A). To get a feel for numbers, there are 67 counties and 2,563 municipalities in Pennsylvania.[6] Oklahoma's Department of Emergency Management lists 285 emergency managers for 76 counties.[7] Arkansas consists of 75 counties with ~500 municipalities.[8] In practice, an EM's job description does not only vary widely from state to state but also from municipality to municipality.

A Pennsylvanian EM succinctly summed the three core features of an emergency management program:

"There has to be a plan, there has to be a person, and there has to be a place."

2.1. Person

2.1.1. Qualifications

Every state draws up its own qualification system, so that it is hard to judge if Pennsylvania, Oklahoma and Arkansas are representative. The only overview of which I am aware dates from 1990 (International Association of Emergency Management, 1990), long before recent agenda forming events like 9-11 and Hurricane Katrina had happened.[9] However, there are certainly states that require EMs to have fewer qualifications than the states described below. Alabama, for example,

[6] E.g., Pennsylvania Department of Community & Economic Development. *Local Governments Online.* URL [accessed 02/11/2014]: http://www.newpa.com/get-local-gov-support/local-governments-online

[7] Oklahoma Department of Emergency Management. *Oklahoma Emergency Management Contacts*. URL [accessed 02/11/2014]: http://www.ok.gov/OEM/Publications_&_Forms/Oklahoma_Emergency_Management_Contacts/index.html

[8] County and Municipal Information & Services. URL [accessed 02/11/2014]: http://local.arkansas.gov/index.php

[9] In 1990, the International Association of Emergency Managers (IAEM) concluded among other things that 39 U.S. states had "no state mandated minimum qualification requirements for local emergency management coordinators" (IAEM 1990). While this is not representative anymore, it does shed a light on the circumstances under which some of today's traditionally long-serving EMs have been appointed to their jobs.

does not require EMs to attend any classes or training, but puts a financial incentive on following a very thorough certification program.

In the studied three states, aspirant emergency managers do not need any prior experience in the field. To qualify for the job in Pennsylvania, the "coordinator [of emergency management] shall be professionally competent and capable of planning, effecting coordination among operating agencies of government and controlling coordinated operations by local emergency preparedness forces" (35 Pa. C.S. §7502d). In Oklahoma, an "emergency management director" has to hold U.S. citizenship, a high school diploma or equivalent, a valid Oklahoma driver license, and a social security number. Additionally, he/she cannot have been convicted of a felony in Oklahoma (63 O.S. §63-683.11B). Arkansas asks for no prerequisites but requires like all three states that a public safety officer has not engaged in subversive acts against the U.S., did not advocate forceful change of the constitutional form of the federal or state government, and did not try to overthrow the government (Ark. Code 12-75-126). In all three states, EMs have to complete between 13 (Oklahoma) and 25 (Pennsylvania) FEMA courses within one to three years of their appointment. Additionally, they have to regularly attend conferences or workshops.[10]

[10] In Pennsylvania, EMs are required to obtain "Basic Certification" within one year and "Advanced Certification" within three years of their appointment (35 Pa. C.S. §7502e). The Basic Certification can be obtained by completing a prescribed set of classes with the Federal Emergency Management Association (FEMA) and the Pennsylvania Emergency Management Association (PEMA), and attending three quarterly training sessions annually. The Advanced Certification mainly consists of following an extra set courses that brings the total number of courses to 21 (local EM) and 25 (county EM) respectively (PEMA Directive D2011-02). In Oklahoma, the EM has to complete basic emergency management training within a year of appointment. It consists of four courses regarding the National Incident Management System (NIMS) established by U.S. Department of Homeland Security and seven FEMA Independent-Study courses (Oklahoma Department of Emergency Management 2009). In Arkansas, local "emergency managers" need to complete three FEMA courses within three months of their appointment, another three within six months, an additional eleven within twelve months, four NIMS courses within 24 months and an additional five elective FEMA courses within 36 months. Annually, they have to either attend the Arkansas Emergency Management Conference or Mid-Year Workshop (Arkansas Department of Emergency Management 2014).

Like most states, Pennsylvania, Arkansas and Oklahoma do not specify whether the EM has to be paid or full-time. There is a trend to professional full-time emergency management because more colleges are now offering emergency management degrees. This trend is partly motivated by FEMA's Emergency Management Higher Education Program established in 1994. It seems to be the perception of the interviewed EMs that an increasing number of full-time emergency managers is also caused by the influx of federal money since the terrorist attacks in 2001 and the advent of the Department of Homeland Security (e.g., Federal Preparedness (Non-Disaster) Grants). Traditionally, a (retiring) police or fire chief has been named the local emergency coordinator. Subsequently, these professions have had a considerable influence on emergency management. In our sample, seven EMs had a background as fire fighters, policemen or in EMS. The others came from wide variety of professions: librarian, coal miner, oil worker, veterinarian, military, industrial safety, banker. Three EMs had a college degree in emergency management.

2.1.2. Mindset

The EMs' mindsets are as similar as their backgrounds are different. Especially in small towns, EMs are local people, deeply rooted in their community who stay on the job for decades. The following quotes illustrate both points:

"Because my city is in jeopardy, my citizens are in jeopardy. And yes, I look at them as mine. … These are the citizens I go to church with … and I see in WalMart."

"The adrenaline thing is a big deal, there is a lot of influence from police and fire. … It's people that care. [EMs] are a caring group of people that are involved in this… They want to help people and they want to help you. They want me to succeed."

2.2. Plan and other tasks

The only task that all EMs have in common is that they write and maintain emergency operation plans (35 Pa. C.S. §7503; 63 O.S. §63-683.11B; Ark. Code 12-75-118). Additionally, they take care of the communication between the various emergency services (fire, police, EMS etc.), city officials and whoever else is involved in emergency management in their jurisdiction. Beyond these shared characteristics, the EMs' task is to do everything that *"does not fit nicely into the fire department or EMS."* During disasters, EMs mainly coordinate and improvise:

> *"You know, if the police comes to the scene they bring guns, firemen bring fire trucks, the emergency manager brings a phonebook."*
>
> *"You need to get over there in a boat … I don't have a boat, but I will find a boat."*

2.3. Place and other resources

EMs in Pennsylvania are required (35 Pa. C.S. §7503) and EMs in Arkansas are authorized (Ark. Code 12-75-118) to have a physical space, the Emergency Operations Center or Public Safety Communications Center, where they can gather anybody who is involved in an emergency operation. Other than that, their resources vary wildly. On the one end, one interviewed EM just had his phone as his only resource and annually put in $2,000 of his own money. On the other end, an EM a few miles away had his own fortified office, staff, jeeps, trailers and boats.

3. How does the inherent uncertainty in river forecasts affect their use by emergency managers?

3.1. Uncertainty in Forecasts

The interviewed EMs were very well aware of the uncertainty in river forecasts. Through the daily use of rain and temperature forecasts, they have often experienced the limits of weather forecasts in general (comparable to findings for the general public: Morss et al. 2010a, Morss et al. 2008). EMs with flood experience also encountered different forms and degrees of uncertainty in river forecasts first-hand as the following quotes show:

"The river forecasts are unpredictable. … Sometimes they are close, sometimes they are way off."

- Underprediction

"During the 2007 flood, their first forecast was 19 feet. We went 10 feet above that…"

"There have been a couple of times, when I called [NWS] and said, 'My river is on flood stage and I don't see nothing on the website about it.'"

- Overprediction

"They usually overpredict, they usually predict it to be higher than it ever reaches."

As the next few quotes indicate, the uncommunicated uncertainty in river forecasts severely limits their value, i.e., providing lead time and informing what crest level to prepare for. All of the interviewed EMs stated that they normally start taking substantial measures only when they (or their monitoring crewg) see the water rising with their own eyes.

"…until the water actually comes and you know which way it's going to go and what floods, you cannot take specific measures."

"… with the topography of the area, with the hills and valleys, lead time really… you cannot rely on [the lead time]."

In practice, the interviewed EMs make very little use of the lead time that short-term weather forecasts provide. This is not surprising given that interviewed EMs only had a very superficial understanding of the weather-related products and forecasts. Participants of the OKFirst program[11] in Oklahoma were a notable exception. Through this program, EMs receive training and

[11] Brief Overview from the OK-First website: "OK-First is an outreach project of the Oklahoma Climatological Survey (OCS) and Oklahoma Mesonet. It provides training and real-time weather data to public safety officials for use in weather-impacted situations. OK-First training and data are provided at no cost to qualified applicants in Oklahoma. As of Spring 2012, more than 500 trained

real-time weather data for weather-related decisions from the Oklahoma Climatological Survey and Oklahoma Mesonet. EMs in other states mostly rely on the several hundred courses offered by FEMA`s training facility, the Emergency Management Institute (EMI). Since the National Weather Service Community Preparedness Program became part of FEMA at its creation in 1979, the NWS seems to be involved with the courses at EMI only by exception. The 2014 course catalog lists only one course that is co-taught by NWS staff (FEMA 2014).[12]

Of the EMI courses offered in 2014, ten have the words "forecast," "National Weather Service" or "weather" in their course description[13] (FEMA 2014). Another three focus on flood response (rather than floodplain management and flood insurance). Of those thirteen, only Pennsylvania requires all EMs to have completed course "IS-0271 Anticipating Hazardous Weather and Community Risk" within three years of their appointment. County EMs additionally have to take "G-271 Hazardous Weather and Flooding Preparedness" (Oklahoma Department of Emergency Management 2009). Arkansas Department of Emergency Management (ADEM) does not require any of those thirteen courses, but names three as electives for the Advanced Professional

---

public safety officials in and around Oklahoma participate in the program. OK-First operates with substantial funding support from the Oklahoma Department of Public Safety." URL [accessed January 22nd, 2014]: http://okfirst.mesonet.org/about.php

[12] Only EMI course in the 2014 course catalog that is jointly taught by Emergency Management and NWS staff (FEMA 2014): G0365 WEM: Partnerships for Creating and Maintaining Spotter Groups

[13] EMI courses in the 2014 course catalog that have the words "National Weather Service", "forecast" or "weather" in their course description or that focus on flood response (*cursive font*) (FEMA 2014):

- E0102 Science of Disaster
- IS-0247.a Integrated Public Alert and Warning System
- IS-0271.a Anticipating Hazardous Weather and Community Risk, 2nd Edition
- IS-323 Earthquake Mitigation Basics
- IS-0324.a Community Hurricane Preparedness
- G0270.3 Expedient Flood Training
- *G0271 Hazardous Weather and Flood Preparedness*
- G0272/L0098 Warning Coordination
- *G0361 Flood Fight Operations*
- G0363/L3011 Hurricane Readiness for Coastal Communities
- G0365 WEM: Partnerships for Creating and Maintaining Spotter Groups
- L0320/L0324 Hurricane Preparedness for Decision-Makers

V0007 Virtual Tabletop Exercise: Flood

Series and the Emergency Manager Certification Program that it recommends EMs to follow (ADEM 2014). None of the thirteen courses appears in Oklahoma's basic training plan, but many EMs there voluntarily participate in OKFirst. Asked about that program, they very much appreciated being taught not only how to read but also how to interpret weather radar so that they could decide whether it was justified to evacuate a ballgame; something on which they had not received instructions before the advent of OKFirst.

In the interviews, especially fire and policemen had trouble coming up with ideas how to make use of lead time, likely because they are used to dealing with disasters that occur with little or no warning:

"I'm not sure if [more lead time] would change much to be truthful."

"People that did not know what to do died [in that flood] …. So the biggest part of these things is [by the time] the fire guys get to your house, you may be dead."

While state emergency management associations (EMAs) need to ensure that every EM receives basic instruction in the use of forecasts in emergency response, the NWS needs to ascertain that the uncertainty in deterministic and probabilistic forecasts are covered in an applied "boots-on-the-ground" manner in the relevant courses, e.g., by developing course material or co-teaching courses.

3.2. Coping with Uncertainty in Forecasts

Even though they are not explicitly aware of it, the interviewed EMs use a number of strategies to cope with the uncertainty in river forecasts, which are described in the following.

3.2.1. Extensive information collection

The interviewed EMs appreciated any kind of information that they can get about approaching weather events such as floods. Especially in small towns, it seems as if many of the EMs soak in all of the information they can possibly get – even if it does not directly relate to the decisions at hand – to get an idea what is going on in their community. In their mind, the weather situation is an integral part of the state that their community is in and cannot be separated from any other on-going activities. For example, one EM appreciated being notified that airplanes on a nearby airport were being de-iced, even though the decision at hand was whether the local Santa Claus-party should take place given the winter weather. This reflects the attitude that EMs consider their job to be caring for a community as a whole rather than just ensuring safety. While not all gathered information is relevant, the value of gathering extra information should not be underestimated, Morss (2010b) observes that the most important decisions in successful flood fights were decisions to gather more information, e.g., using monitoring crews.

EMs retrieve information from a wide variety of sources: NWS publications, personnel, local media, and the personal network. One EM mentioned the importance of his personal network to find out what is happening:

> *"… I have an uncle who lives 30 miles west. He will call or somebody else's brother-in-law lives up north 60 miles and he will call "We have 6 inches of rain." … So normally, through the channels of families and business … the information trickles in from one way or the other. "*

Exploiting personal networks for information is not exclusive to EMs; other NWS clients employ the same strategy. For example, Rayner et al. (2005) tells of a water supply operations manager who relies on a brother-in-law 60 miles away for information about weather conditions.

The NWS and state EMAs need to (jointly) offer hotlines or mentoring networks where EMs can discuss the correct interpretation of forecast uncertainty. If EMs cannot reach an expert when they face uncertainty (e.g., because of jammed phone lines), they are likely to satisfy their need for information through more *ad hoc* sources. In the interviews, nine out of seventeen

EMs explicitly mentioned how much they appreciated having a close relationship with NWS personnel:

"Because we have that relationship I am able to pick up the phone and call them. And they won't talk down to me. … I make them look good, they make me look good; it is a partnership. And it has been wonderful when we can work together."

3.2. Making their own estimates

Using their local knowledge and the gathered information, EMs make their own interpretation of the information they have collected. One EM describes it as follows:

"I would have … to see what is going on on the radar and see where the rain is falling to say… I give [NWS] 75% probability that [the forecasted water level]is going to happen."

Although the NWS forecast is only one piece in the EMs' information puzzle, it is a very important piece because it is the official interpretation of the situation at hand. It serves the EMs as a standard to build on and with which to compare their own interpretations. Nonetheless, especially the less experienced EMs refrain from making uncertainty estimates, effectively ignoring its existence by considering the forecast to be perfectly correct:

"But try to analyze [the forecast]… that is not what our job is."

When floods occur, EMs gain experience to cope with uncertainty by trial-and-error, incurring potentially unnecessary damages, as the following quote illustrates:

"What I wasn't doing, and that is kind of my own fault, thinking, yeah, it is pouring rain here, but what is it doing up above?"

Even though the EMs are aware of the uncertainty in forecasts and can explain its causes, none of the interviewed EMs – not even those with flood experience – was able or willing to quantify a representative forecast error:

"I couldn't tell you. … It does vary a little bit. But 1 to 5 feet difference, I couldn't tell you."

To prevent EMs from ignoring uncertainty, the deterministic forecasts need some uncertainty information such as the historical forecast error or at least a prominent note that uncertainties exist. To facilitate EMs in making their own estimate, both deterministic and probabilistic forecasts in any format need a description of the main sources of uncertainty (e.g., temperature development with regard to snow melt, dam gate operations or expected rainfall upstream). This way, it will be easier for EMs to combine the forecast information with the other information that they have collected.

3.2.2 Using observed conditions rather than forecasts

As another coping mechanism, EMs try to use the weather information that has the least uncertainty, i.e., measured data. Rather than the river forecast, EMs use a combination of observed stage heights and weather radar. If the river water levels are high or rising and the radar indicates that more rain is to be expected upstream, they know that they are likely to experience problems with the river. Further, the river forecast is updated too irregularly, and the forecasted crests levels vary too much to base decisions on. Additionally, its uncertainty is less intuitive than that of weather radar. The precipitation observed upstream is only published a day later and therefore those data are much less useful than the weather radar. The quotes below illustrate the discussion above.

"When there is a flood forecast, certainly we are on alert. But I spend as much time driving down to the creek as I do listen to the forecast."

"The forecast can change normally every hour. Sometimes during an event, it may change every half hour."

"The NSW puts out their temperature forecast at 9.45am and 4.45pm. … The river forecast people put up forecasts whenever they want to. There is no set time. So I tell my [town's] people … that we are expecting flooding. 'Stay tuned to the NOAA weather radio. … I don't have the latest information but you can stay

updated, because I might be gone for an hour or two on a delivery and they put out a new forecast and I am not aware of it.'"

In short, because of their inherent uncertainty, river forecasts are in a poor competitive position when compared with other NWS products. This finding is consistent with Rayner *et al*'s (2005) observation that water resource managers tend to rely heavily on observed information from monitoring groups while ignoring model results.

Before starting to publish probabilistic short-term river forecasts, NWS needs to consider which purpose those forecasts serve today and how that is likely to change once they introduce uncertainty information. It seems possible that EMs would still prefer using weather radar and observed river stage heights to plan their emergency operations, because communicating uncertainty does not reduce it. For example, offering a description of the situation (e.g., how much more rain would have to fall upstream to cause serious flooding) and the sources of uncertainty would make the forecast more salient for EMs who make their own estimates of the situation.

3.2.3 Broader Context

Zooming out, the uncertainty in weather and river forecast is not the uncertainty that EMs are most concerned about. Cascading events such as clogged stormwater drainages (e.g., by a skateboard), medical needs of evacuees, hazmat spills, river traffic such as recreational boating and loaded barges, drinking water and utility outages, and snakes cause EMs the worst surprises:

"Usually, what is unexpected is when the infrastructure fails or collapses."

"It sounds stupid, but [lack of air conditioning] is a thing that can make a small emergency a big emergency."

Morss (2010b) describes a particularly turbulent situation: "In Fort Collins, in the trailer park area alone, emergency responders had to handle a water rise of 1.7 m (5 ft) in 3 min, three trailers on fire, 162 people in imminent danger requiring rescue, a building explosion, and derailment of a train that included a car carrying chlorine gas, with no advance notice."

4. What do the emergency managers do with the forecasts?

It is a mistake to think that EMs are the decision-makers and, therefore, end-users of the weather and river forecasts. Depending on the jurisdiction, roads are closed by the Public Works department; fire or police chiefs coordinate rescuing activities; the mayor signs off evacuations; etc. Citizens are responsible for their own property; even evacuation is usually optional for the individual. Ultimately, every person, including operators and owners of industries and businesses, are responsible for their own safety.

> *"I call them [at the power plant], we let them listen to the briefing, and they make their own decision based on the information they are getting."*
>
> *"Would I tell everybody to evacuate? No, I would not. I would say, here is the information, and it would probably come across very not nice. … Make your own [#!°&%] decision."*

Rather than making the decisions, the EMs act as an invaluable link in the dissemination of the NWS forecasts. Their three main tasks are to alert decision-makers to the situation, to interpret the forecast for them and to motivate them to take action. Each will be discussed in detail in the following.

4.1. Alerting People

First, EMs constantly keep an eye on the weather situation at all times and alert decision-makers (i.e., department heads, police and fire chiefs, city officials, and the citizens) if they think a critical situation is approaching. The following quotes describe how much EMs add to the dissemination of forecast information:

"We would be the receptor of … information that comes in … We bring in all the decision makers, city manager, public works director, fire, police and have them coordinate and get them working as a team."

"… I will make sure it is in the newspaper. I will make sure it is on the radio. I will make sure that TV can tell people. …"

"We rely on every type of social media you can. Peer notification is a major player."

"We try to use the radio, we are starting to use a little bit of social media, it is a real struggle getting that through the City, to agree to social media, but it is going to be effective."

To alert every person at risk, EMs have to navigate various obstacles, which the NWS could not overcome. This underlines the importance of the EMs as messengers of the NWS.

- Long chain of communication:

  "We go to [our] boss, who is the public safety director and we say, "This is the information; what do you want us to do?" [The public safety director] is the one that is final, for anything that is major … because he goes and talks to the mayor."

- Unsuccessful communication to key decision makers:

  "All you could see was a window of the attic, where these people had fled to and some were on the roof. 100 some people. One of them was our mayor. "

- Difficulty reaching minorities:

  " 25% of my population would be considered Spanish-speaking. And the radio and TV stations are now carrying both Mexican-speaking channels. But they don't give local weather. … I have no circulated Spanish paper to put the information in."

Given these obstacles, **the NWS needs to make sure that their forecast is formatted in such a way as to assure that it does not get diminished, distorted or becomes straight-out wrong as it travels through a variety of channels** from the forecasting center, along the Weather Office, through the emergency manager to the decision-maker.

Particularly, **the formatting of forecasts should not only ensure that users understand them correctly, but also that they disseminate them correctly as well.** This is especially salient because ideally EMs should convey the risk to the public in *non-technical* terms (Perry & Nigg 1985). Consequently, EMs do not only have to understand the rather technical forecast and its uncertainty, but also translate it into non-technical language. This will be especially difficult once uncertainty information is included in the forecasts.

When confronted with uncertainty in probabilistic forecasts, EMs are likely to pass the uncertainty information on to the public. Several studies have shown that unless it is done well uncertainty and complexity can be a major obstacle for effective risk communication (e.g., Covello & Sandman 2001). **It is therefore crucial that EMs learn in FEMA courses such as "G-272 Warning Coordination" and "G-289 Public Information Officer Awareness" how to communicate uncertainty to the general public**.

4.2.Interpreting the Forecast

Second, EMs add local information to the forecasts that the NWS cannot provide. EMs often know the town well enough to translate river stage heights into which places will flood. Additionally, many EMs have to interpolate between river gages because there is no gage close to their jurisdiction. The quotes below show two examples of assessing the situation.

> *"When the foam is on the middle of the river, it is coming up … If I see big trees coming down, I know it is a lot water. If I don't see big trees, well, it is not a lot water."*
>
> *"[If River A] isn't up and the [River B] is down … and the gates at the] Lake [are] closed, then I think it is going to go away. Or I tell [the public], see [River A] is continuing to flood, it crests up there, 12 hours later it crests down here. Same for [River B] … If I get both crests here at the same time, I have a problem. If this one is gone before this one gets there, then there is not a lot of current."*

As described above, based on additional information and experience, EMs will make an intuitive assessment if the river forecast is likely to be correct. While the interviewed EMs share their assessment of the situation and their concerns regarding the official forecast with the other people in the emergency operations center, they will only repeat the official, published forecast information to the citizens. The following quotes demonstrate these two types of dissemination.

- Communication to people in the emergency operations center

  "So I put ... all together and then I have another briefing [for my staff]. 'Guys, this is the information I have: It appears as if it is going to come up another couple of feet. But it might not.' That is a guessing game for floods."

  "Last thing that you want is that your boss, the city manager, is being caught off guard. … And so it is like if I know this and I give it to you, then you know it, too. So it is accountability."

- Communication to citizens and businesses

  "Whatever they [at NWS] tell me, I take what they say and act accordingly. I don't take a chance of saying maybe it's wrong. I can't do that."

  "If I start putting out my own forecasts, if I start telling people what I think… and then I'm wrong. And I… is the city responsible or liable?"

  "Do you tell them, 'The forecast is 19 feet, but in the past it's been 10 feet above that?'" – *"I'm not going to. No. … Even though I know. … The less I say, the better off I am, because the media will come to haunt me."*

There are three possible explanations as to why EMs are hesitant to share their assessment of the situation with the public: 1) They do not want to risk liability issues for themselves or their employer; 2) they fear that they will be held accountable, for example by the media or their superiors; or 3) they might not feel sufficiently confident enough about their judgment.

The EM's reasoning seems to be that, if the EMs only pass on the official NWS information, then the citizens and businesses will ultimately be responsible. However, if the EM interprets the forecast (i.e., assesses its uncertainties) and people act accordingly, the EM feels responsible for some or all of the damage. Consistently, it has been claimed that the NWS simply provides the users with what they want: a point estimate without any information of uncertainty (Morss & Wahl 2007) that – at least superficially – frees the decision-maker of the uncertainty inherent in the forecast.

When asked, most EMs did not know what the legal situation exactly was, but as the quotes above illustrate, they were worried about it. It seems unlikely but not impossible that EMs or their employers could be held liable for voicing interpretations of NWS forecasts in public. Describing lawsuits against weather forecasters, Klein and Pielke (2002a) explain that the Federal Tort Claim Act (FTCA) protects the federal government against such claims, because forecasting falls under the discretionary function exception.[14] Most states have a similar immunity statue (Swanson 2000). Discussing private sector forecasts, Klein and Pielke (2002b) cannot find any published court decisions where a weather forecaster was held liable for publishing inaccurate forecasts. Based on court decision in other fields, Klein and Pielke (2002b) conclude that forecasters are unlikely to be held liable, but that "predictions containing false statements of fact that do not have a reasonable basis and that were made with an intent to deceive, manipulate, or defraud, or with reckless indifference, will result in liability". Swanson (2000) writes that "in virtually most statues granting immunity, immunity is not available if the death, injury, and damages are the result of willful misconduct, gross negligence, wanton disregard, or bad faith on the part of the actor."

---

[14] To establish if this exception applies, the U.S. Supreme Court's decision in *U.S. v. Gaubert* established a two-part test in 1991. According to Klein and Pielke (2002a): "The first part examines whether the challenged conduct involves an element of judgment or choice. If so, the second part of the test examines whether the judgment is the type meant to be shielded by the discretionary function exception which depends on whether the challenged conduct is susceptible to policy analysis."

The authors could only find one lawsuit where a county emergency management was sued in connection with weather forecasts. In 1989, 30 students were injured or killed when a wall at an elementary school collapsed because of a tornado or strong winds. The relatives of the deceased sued among others the county for failing to warn the school, even though NWS had published a tornado watch (not warning). In 1992, the Appellate Division of the Supreme Court of the State of New York ruled in *Litchhult v. Reiss*[15] that the county was not liable because "the County's plan in this respect falls squarely within the definition of a discretionary act because it requires an analysis of a situation and a decision about whether notification is warranted. … The County had to exercise its discretion in determining whether evacuation or other extreme measures would be taken in a similar manner as if a tornado warning had been issued."

Thus, the case history does not give the EMs much reason to worry about liability. However, Klein and Pielke (2002b) warn that there is a large gray area in cases when "the forecast may have been made in good faith, [but] it strayed from established professional standards" that define reasonable care. **As long as professional standards regarding dissemination of weather forecasts have not been defined for EMs, they will remain reluctant to voice their own assessment of the situation in public. Additionally, it is important that the NWS and FEMA together evaluate the legal aspects of using forecasts. Most importantly, EMs need to be informed regarding their legal situation to reduce their current concerns.**[16]

EMs and NWS would benefit if NWS published uncertainty information as part of the forecast. Not only would that decrease the (already low) probability of liability claims against NWS.

---

[15] Case Citation: *Lichult v. Reiss* 183 A.D.2d 1067 (N.Y. App Div. 1992). URL: http://scholar.google.com/scholar_case?case=9093875182234136223&q=weather+%22emergency+manager%22&hl=en&as_sdt=6,39 (Accessed 03/04/2014)

[16] According to Nicholson (2006), the liability issues tend to be neglected in emergency management. Regulations differ between jurisdictions and type of employment (professional or volunteer). The legal background gets little attention in the education of EMs (Nicholson 2006), contributing to their risk-averse behavior (i.e., assuming the forecasts are correct at all times) in decision-making.

If uncertainty information was published by a federal authority such as NWS, EMs would feel more confident in a legal sense to communicate the uncertainty to the decision-makers, such as home-owners.

4.3.Motivating People to Take Action

Third, the EMs motivate people to take action. Even though EMs might hold back their own interpretation of the situation in public, the EMs do try to suggest actions to the public. Indirectly they assist people to cope with uncertainty this way:

> *"We may make phone calls to certain hospitals and ball parks and things like that. 'This is the time you probably need to pull people off the field'; things like that."*

False positives – a consequence of uncertainty – are a major reason why EMs often have a hard time waking people from their lethargy:

> *"They have heard [a severe weather warning] so much, that it doesn't really… they are expecting me as a government official to take them by the hand and take them to the cellar when they need to go to the cellar."*

In this context, it is most important that EMs enjoy authority, especially in closely-knit communities. Through their profession and mindset, they often have a strong and dense personal network. In tornado alley, the EMs and the communities have been through many critical situations together so that there is a high degree of respect for the EMs. The following quotes illustrate how EMs use their reputation to make people move.

> *"I just call. 'Hey, we are going to flash flood.' … That is all I have to say to them. That is called the personal touch. "*
>
> *"What I normally tell them is, … 'I want to get a good look of you before I leave, because later on when I am called to identify you, I want to know who you are.' And most times they leave when you tell them that."*

Through their connectedness to the people exposed to flooding, the EMs have the ability to prevent decision makers to become paralyzed when faced with uncertainty.

4.4. Summary

The three tasks EMs perform (alerting, interpreting and motivating) add substantial value to the river forecast produced by the NWS. Particularly, EMs already and with some additional measures could even more facilitate decision makers in coping with uncertainty. This article points to many ways in which NWS can assist their clients to make the most out of their river forecasts, particularly once they include uncertainty information. **Most importantly, NWS and FEMA would both benefit from integrating the valuable expertise of EMs in forecast dissemination. It would increase the impact of the forecasts and provide emergency responses with better information.**

5. Conclusions

One academic commented somewhat indignantly on our findings that emergency managers seemed to be very unprofessional. Yes, they are. In small-town America, the EMs are often volunteers. Indeed, emergency management is not the main profession of most EMs. Certainly, for all of them interpreting weather information is only one of innumerous tasks. Many emergencies are not even weather-related. But EMs are experts of the local circumstances and in a position to add great value to the NWS products before they reach the decision makers. NWS could reap substantial benefits from getting to know EMs better.

This study clearly shows that NWS is taking the right steps by planning to include uncertainty information in short-term river forecasts. Without it, decision-makers are unlikely to make decision-makers aware of uncertainties, which has led to insufficient emergency responses in the past. But as it is already the case with deterministic forecasts, simply providing (uncertainty) information to decision-makers does not automatically result in adequate decisions in emergency management. There are a number of things NWS in cooperation with FEMA and state EMAs can do to significantly increase the utility of its current and future river forecast products:

- Consider what impact publishing uncertainty information will have on the function of a river forecast in an emergency operation. As of today, river forecasts mainly play a role for accountability but are not really used to plan emergency operations; largely because of potentially large, unpredictable forecast errors.
- Clarify the impact of river forecasts on accountability and liability considerations of local EMs. Potentially, a (locally perceived) shift of accountability could occur if the NWS started publishing probabilistic short-term river forecasts.
- Consider the EMs as an integral part of the forecast dissemination. The EMs are the ones who bring the forecast to the attention of decision-makers, often interpret it for them, and motivate them to take action.
- Format forecasts so that people not only understand the forecast, but also share information with others correctly. Sharing information on uncertainty is likely to be more difficult than sharing a single-point forecast.
- Provide personal assistance in cases where EMs have trouble interpreting uncertainties. If experts are not available for clarification of the forecast when faced with uncertainty, this gap will undoubtedly be filled in by other, more *ad hoc* sources of information.
- Amend EMs' trainings, workshops and conference presentations on NWS products with explanations on how to utilize those to in crisis situations. Additionally, provide training on how to make decisions under uncertainty and how to communicate uncertainties to the public. Otherwise, uncertainty can lead to inaction. Interactive small-scale case studies, e.g., a local ballgame, seem to work best.

Acknowledgements: Most importantly, I wish to thank the emergency managers who made the interviews possible. I was deeply impressed by their passion for their communities.

6. References

Arkansas Department of Emergency Management (2014): *Training Plan. January 2014 Update.* ADEM Training Plan V 2.5 Rv. NOV 2006, updated January 2014.

Baumgart, L.A., Bass, E.J., Philips, B., & Kloesel, K., 2006: Emergency management decision-making during severe weather. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* **50**(3), 381-385.

Covello, V., & Sandman, P. M., 2001: Risk communication: evolution and revolution. *Solutions to an Environment in Peril*, A. Wolbarst (ed.), John Hopkins University Press, 164-178.

Federal Emergency Management Agency (2014): *Fiscal Year 2014 Training Catalog. Train, Exercise, Educate.* Emergency Management Institute, Emmitsburg, Maryland. URL: http://training.fema.gov.

International Association of Emergency Managers, 1990. *Existing Standards for Local Emergency Management Coordinators.*

Klein, R., & Pielke Jr, R. A., 2002a: Bad weather? Then sue the weatherman! Part I: Legal Liability for Public Sector Forecasts. *Bulletin of the American Meteorological Society*, ***83***(12), 1791-1799.

Klein, R., & Pielke Jr, R. A., 2002b: Bad weather? Then sue the weatherman! Part II: Legal Liability for Private Sector Forecasts. *Bulletin of the American Meteorological Society*, ***83***(12), 1801-1807.

Morss R.E., and E. Wahl, 2007: An ethical analysis of hydrometeorological prediction and decision making: The case of the 1997 Red River Flood. *Environmental Hazards,* **7**, 342-352.

Morss R.E., Lazo J.K., and J.L. Demuth JL, 2010a: Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication. *Meteorological Applications,* **17**, 149-162.

Morss R.E., 2010b: Interactions among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases. *Natural Hazards Review*, August 2010, 83-96.

Morss RE, Demuth J, and J.K. Lazo, 2008. Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Weather and Forecasting,* **23**, 974–991.

National Hydrologic Warning Council, 2002. *Use and Benefits of the National Weather Service River and Flood Forecasts*. Prepared by EASPE, Inc. URL [accessed 02/19/2014]: http://www.nws.noaa.gov/oh/ahps/AHPS%20Benefits.pdf

NOAA National Weather Service, 2011: *NWS River Forecasting.* Backgrounder, October 2011. URL [accessed January 24th, 2014]: http://www.crh.noaa.gov/images/dvn/downloads/backgrounder_DVN_River_Forecasting.pdf

Nicholson, W.C., 2006, Chapter 14: Legal Issues in Emergency Management. *Fundamentals of Emergency Management,* M.K. Lindell, C.S. Prater, R.W. Perry, Federal Emergency Management Organization. URL [accessed January 23rd, 2014]: http://training.fema.gov/EMIWeb/edu/fem.asp

Oklahoma Department of Emergency Management (2009): *Training Requirements. Newly Appointed Oklahoma Emergency Management Directors.* URL [accessed 02/10/2014]: http://www.ok.gov/OEM/Programs_&_Services/Training_Information/2009newdirector.html

Parker, D. J. and J.W. Handmer, J. W., 1998: The Role of Unofficial Flood Warning Systems. *Journal of Contingencies and Crisis Management,* **6**, 45–60.

Pennsylvania Emergency Management Association (2011): *PEMA Directive 2011-02*. Subjective: Emergency Management Certification and Training Requirements. Effective Date November 7, 2011.

Perry R.W. and J.M. Nigg, 1985: Emergency Management for Communicating Hazard Information. *Public Administrative Review,* **45**, Special Issue: Emergency Management: A Challenge for Public Administration, 72-77.

Pielke, R.A., 1999: Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood. *Applied Behavioral Science Review,* **7(2),** 83–101.

Rayner, S., Lach, D. and H. Ingram, 2005: Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. *Climate Change,* **69**, 197–227.

Swanson, H. D., 2000: Delicate Art of Practicing Municipal Law under Conditions of Hell and High Water, The. *NDL Rev.*, *76*, 487.

Welles, E., Sorooshian, S., Carter, G. and B. Olsen, 2007. Hydrologic Verification: A Call for Action and Collaboration. *Bulletin of the American Meteorological Society*, **88(4),** 503–511.

STUDY 2

Use of Weather Forecasts in Emergency Management: An Application of the Theory of Planned Behavior[17]

Abstract

Many factors affect the extent to whcih forecasts inform emergency responses. In a survey based on the Theory of Planned Behavior (Ajzen, 2006), we asked 207 U.S. American emergency managers (EMs) about (1) their past and intended future use of short-term weather forecasts and recorded weather data; (2) the perceived limitations of the forecasts and recorded weather data; (3a) their attitude towards the usefulness of such weather information; (3b) their attitude towards their job and toward uncertainty; (4) perceived social norms; and (5) self-assessed numeracy. We find that work experience is the best predictor of the extent to which an emergency manager relies on forecasts and recorded weather data. Those with a positive attitude towards the weather information and who had received instructions on using weather information relied to a larger extent on forecasts. Perceived unavailability for the area, insufficient information and irregular release times, and a mindset favoring prevention over reaction tend to decrease the reliance on recorded weather data.

Keywords: Emergency managers, forecasts, weather, Theory of Planned Behavior

[17] In fall 2014, a version of this chapter will be submitted to the Bulletin of the American Meteorological Society (BAMS).

1. Introduction

The purpose of forecasts is to provide their users with lead time to prepare for the events to come. In the case of emergency managers (EMs), weather forecasts can give the opportunity to evacuate hazardous areas, prepare equipment to be able to respond quickly or to protect or move valuable property before a severe weather event hits.

Studying the use of forecasts in emergency management is difficult: no town is like another; different regions are subject to different weather hazards, and each state has different regulations. Natural emergencies themselves are difficult to study, because they are usually unique with no control group. Additionally, the reasoning behind decisions during an emergency is usually not recorded, because it takes valuable time away from preventing damage and bringing people to safety. Consequently, it is often difficult to determine afterwards what consideration led to which decision, and whether that was a good decision. Unsurprisingly, the literature on forecast use in decision-making mainly consists of case studies and a few surveys.

1.1. Forecast Use

A much discussed case study is that of Grand Forks, ND, during the Red River Flood in 1997, where despite a good forecast there was ~$1 to 2 billion in flood damage, because the decision-makers did not realize how much uncertainty was associated with the forecast. Even though emergency responders took into account some uncertainty, the people of Grand Forks could not imagine such a devastating flood. Instead, they anchored their expectations on the much more manageable, predicted crest level which was more in accordance with their past experiences (Pielke, 1999; Morss and Wahl, 2007).

Besides the Red River flood in 1997, Morss (2010) describes two more cases: a flash flood in Fort Collins, CO in July 1997, and a flash flood in Pescadero Creek Basin, CA in February 1998.

Comparing these cases, it can be concluded that experience greatly drives expectations and therefore preparations. If the event deviates from previous ones, people are often caught off-guard (e.g., representative, availability, and anchoring heuristics; Tversky, Kahnemann, 1974). Forecasts seem to reinforce this process, by not providing what-if scenarios to make people think outside their experience. A successful forecast can help to manage expectations and to guide further information collection by smartly placing monitoring crews (Morss, 2010).

There are a few empirical studies on forecast use in decision-making. Morss et al. (2010) find in a survey of the general U.S. public, that most people associate uncertainty with weather forecasts and can sufficiently interpret probabilistic forecasts. However, people decide to take action at different probability thresholds. Empirically, Demuth et al. (2011) show that the frequency with which the general public obtains weather forecasts depends on what they are used for, their attitude towards the forecast (i.e., importance of and confidence in the information), the accuracy of the forecast,[18] gender, age, and how long they have lived in the area. In a laboratory study in which emergency managers from Oklahoma worked through a simulation of a severe storm, Baumgart et al. (2007) find that emergency managers highly value many kinds of weather information during all stages of a storm. In the simulation, radar products and (storm) spotters' observations were used frequently and evaluated as very valuable. In their descriptive framework, the authors illustrate that emergency managers interact with the NWS at different stages of an emergency; this interaction is two-way in many cases (Baumgart et al., 2007). Additionally, Baumgart et al. (2007) find that more experienced emergency managers use weather forecast products more frequently.

In a previous interview-based study (Study 1), we find that forecast use is driven by more factors than the desire to have the adequate information to make good decisions, i.e., to make use of

[18] When the error in Probability of Precipitation (PoP) forecasts was higher, people tended to consult it *more* often. The error in the maximum temperature forecast was not a significant predictor (Demuth et al., 2011).

the lead time that a forecast potentially provides. For actual decision-making, emergency managers seem to rely much more on radar and observations on the ground, because they were acutely aware of the uncertainty associated with NWS forecasts. Forecasts seem to be used to communicate with the public, because they represent the official assessment of the situation.

1.2. Theory of Planned Behavior

When using forecasts to prepare for weather-related events, emergency managers consider a number of things: their evaluation of the forecast, the social repercussions they might face, if they do or do not use them, etc. (Study 1). The Theory of Planned Behavior (TPB) captures all of these components.

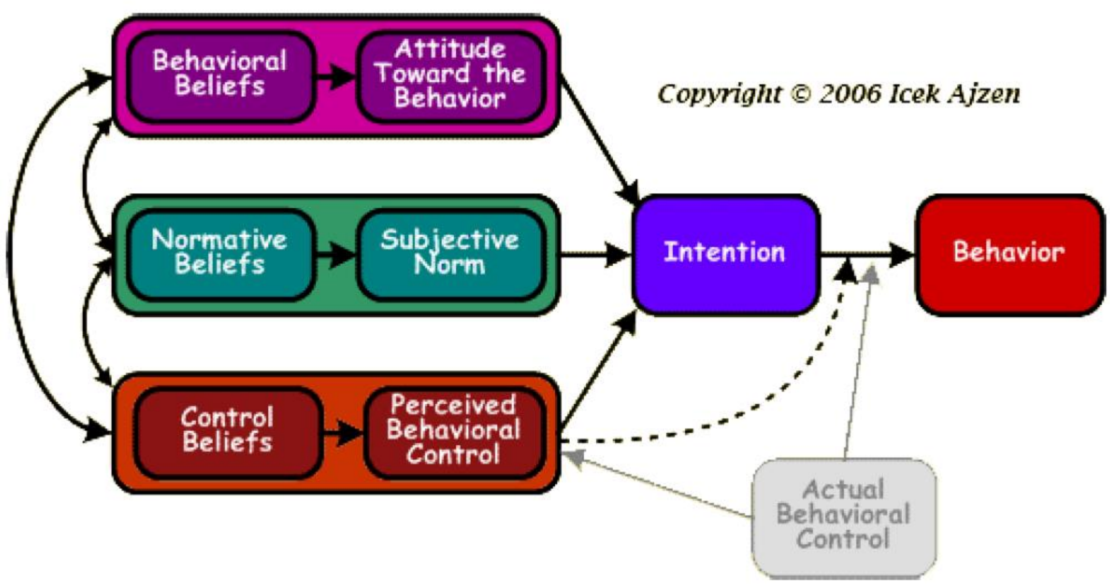


Figure 2: Framework of the Theory of Planned Behavior (Ajzen 2006)

Ajzen (2006), who originally proposed TPB, summarized its components as shown in Figure 2. The dependent variable is a human behavior, in this case the use of forecasts and recorded weather data to prepare for weather-related events. According to TPB, three types of considerations combine to influence the agent's behavioral intention. First, there are behavioral beliefs that inform the agent's attitude towards a behavior, e.g., whether it is beneficial for them. Second, the agent considers normative beliefs held by others, informing the agent's subjective norm on whether (s)he should behave in a certain way. Third, the agent accounts for all factors that might help or hinder the considered behavior, determining whether the agent believes that (s)he has enough control to carry out the planned actions. Depending on the true extent to which the agent has control over his/her actions, these intentions result in the studied behavior (Ajzen, 2006). In a more mathematical notation, TPB can be summarized as shown in Equation 1.

Equation 1: Theory of Planned Behavior

$$Behavior \sim Intention = f(Attitude, Social\ Norms, Perceived\ Control)$$

TPB has been widely used in public health and social science and found to be valid for a range of actual behaviors and intentions, e.g., to undergo genetic tests (Wolff et al., 2011), to buy a home (Cohen et al., 2009), and pedestrians violating traffic rules (Moyano Diaz, 2002). By the end of 1997, Armitrage and Connor (2001) were already able to base their meta-analysis on 185 independent TPB studies.

TPB has also been applied to weather forecasts. Together, Artikov et al. (2006) and Hu et al. (2006) investigated, why U.S. farmers do not use climate forecasts more despite the continuous improvement of such forecasts. This study investigated multiple dimensions by asking the same questions side-by-side for current and recent-past weather/climate conditions, short-term weather forecasts and long-term forecasts. Giving the descriptive results of this study, Hu et al. (2006) report that the perceived utility of the forecasts was low (attitude) and that some social groups – especially crop consultants and spouses – are perceived to support the farmers' use forecasts (social norms). Additionally, farmers worried about the reliability of the forecast provider, which constitutes a perceived obstacle (control beliefs).

Artikov et al. (2006) ran four regressions on the data described by Hu et al. (2006). The regressions revealed that the attitude indicating the perceived utility of the forecasts was the best predictor of actual forecast use. Social norms also were a significant predictor of forecast use. Furthermore, the authors suggest that changing the farmers' attitude is likely to increase forecast use more than addressing technical limitations of the forecast. In addition to the traditional TPB model, financial ability and motivation was also included in the model as a measure of "actual control" (see Figure 2). This variable clarified the effect of attitude and norms on the behavior.

2. Method

In the current study, a survey for emergency managers was devised to assess the four components of the TPB for two types of weather information: 1) observed weather conditions of the past few days and hours (recorded weather data), and 2) short-term weather forecasts for the next few days and hours; similar to the survey among farmers by to Hu et al. (2006) and Artikov et al. (2006). Furthermore, the questions to assess numeracy on the Subjective Numeracy Scale constructed by Fagerlin et al. (2007) were added to the survey. Each component is discussed in more detail in the following sections. The complete questionnaire can be found in the appendix.

2.1. Dependent Variable – Past Behavior/Intentions

"How much did you rely on NWS data to make any of the following decisions or carry out any of the following actions?"

The dependent variable in this study is the use of recorded weather data and short-term weather forecasts by emergency managers in the past and future. In this study, it is not possible to measure the intentions first and the actual behavior a few months later, as is done in most TPB studies. Consequently, the intentions cannot serve as an independent variable in the regression model with the actual behavior as the dependent variable, as is conventionally done.

To make the survey salient to more emergency managers, six hazards were included in the study: Flash flooding, river flooding, hurricanes, tornadoes, heat waves, and snow and ice storms. The participating emergency managers were asked to select the one hazard that they found most difficult to respond to in the last ten years and subsequently answered the questions about forecast use for that hazard.

Use of weather information was operationalized by asking how much the emergency manager relied on recorded weather data and short-term weather forecasts for a range of decisions and actions, such as alerting citizens, alerting colleagues, and opening shelters.

2.2. Perceived Control: Limitations of Weather data

"Please rate how much the following factors have limited your reliance on NWS data when responding to events."

The variable pertaining to perceived control in this study deviates somewhat from those in other TPB studies. In the case of weather information, it did not make much sense to ask how much control the EM feels (s)he has over actually relying on such data. It seemed unlikely that anyone would order the EM what to do, just as it seems impossible for an EM to estimate how much the unique circumstances of each emergency have influenced or will influence his reliance on forecasts. Instead, the survey assessed the perceived limitations of the weather information. Potentially, this could lead to an overlap with the attitude towards the data.

To assess perceived behavioral control, the emergency managers were asked to judge the extent to which characteristics such as the accuracy of the recorded weather data and short-term weather forecasts, the availability for their area, irregular release times, etc. limited the reliance on such products.

A distinction can be made between self-efficacy (i.e., one's own ability to do something) and behavioral control (i.e., external factors hindering the agent). In their meta-analysis, Armitrage and Connor (2001) found evidence that self-efficacy performs as well in in predicting intentions and behavior as perceived behavioral control. Therefore, we additionally asked the participants assess their own ability to understand the weather information and to apply it to the situation at hand limited their use of weather information (self-efficacy).

2.3. Social Norms

"Please rate how much the following groups expect you to rely on NWS data when responding to an event."

"Please rate how much you worry about criticism from the following groups when responding to an event."

In an earlier, interview-based study (Study 1), we find that emergency managers consider the reactions of the public and the media and other types of accountability when responding to weather-related events. In such considerations, the weather forecast is seen as the official interpretation of the situation and therefore more readily shared than the emergency manager's own assessment, even though the latter likely includes much valuable local information.

To assess these sorts of social pressures, the emergency managers were asked to what extent a variety of groups – such as colleagues, elected officials, and residents – would want them to rely on recorded weather data and short-term weather forecasts respectively. Additionally, it was asked to what extent the emergency managers worry about being criticized by each of these groups.

If groups, whose criticism matters to emergency managers, would want the emergency manager to rely on recorded weather data and/or short-term weather forecasts, it is expected that the emergency manager would use such weather information more.

2.4. Attitude

The emergency managers' attitude was assessed in two ways. First, their attitude towards the weather data and forecasts themselves was assessed. The emergency managers were asked to what extent they agreed or disagreed with statements such as *"I would recommend other emergency managers to rely on weather data."* or *"Relying on weather data has resulted in better decisions."*

Second, the participants were questioned about their job attitude. The underlying thought is that many emergency managers are or have been fire fighters or paramedics at some point. In those jobs, they were primarily drilled to respond to ,rather than to prepare for, crises. It is thus conceivable that they would make less use of forecasts because they are not as familiar with the opportunities that lead time provided by forecasts presents. The participants were asked, to what extent they agreed or disagreed with statements such as *"It is my job to respond to weather-related events rather than to prepare for them."* and *"It is my job to rescue rather than to protect people."*

Compared to responding to an emergency, preparing for it requires coping with uncertainty. After all, beforehand it is still unknown how an emergency will evolve. To capture this, three questions covering the EMs' attitude towards uncertainty are in the attitude section. These include statements such as *"I routinely think in what-if scenarios when responding to events."* and *"A critical situation can develop in so many different ways, it is difficult to determine appropriate actions."*

2.5. Subjective Numeracy Scale

In surveys, numeracy is usually assessed by asking people mathematical questions. When they encounter such questions, people often feel judged and tend to quit the survey. Additionally, it cannot be assured that participants do not use calculators or seek help from others when conducting a survey online or over the telephone. To address this problem, Fagerlin et al. (2007) developed and Zikmund-Fisher et al. (2007) validated the Subjective Numeracy Scale (SNS). In an empirical study, they selected eight of originally 49 questions that best correlate with established objective numeracy scales. Four of those eight questions assess the ability to conduct mathematical operations, e.g., calculating a 15% tip. The other four questions ask participants, whether they prefer numbers or percentages over words, etc. SNS has been applied in a number of studies (e.g., Hess et al., 2011; Paolacci et al., 2010; Hawley et al., 2008), mainly in the health sector and the risk communication field. Thus, it is possible to compare emergency managers' subjective numeracy to that of other (professional) groups.

2.6. Demographics

Lastly, the survey assessed a number of demographic factors pertaining to their job as emergency manager, their professional background, their education and training. Cohen et al. (2009) mentions that demographic variables are seldom included in TPB studies,[19] but they find them to be among the best predictors for the intention to buy a home and the actual purchase.

2.7. Hypothesis

It is hypothesized that EMs rely more on the recorded weather data and short-term weather forecasts, if – compared to their colleagues – they think the following:

- Evaluate the studied weather information as more helpful for decision-making (attitude)
- Perceive those products to have less limitations and who think to have the ability to use them (perceived behavior control, self-efficacy)
- Put a greater emphasis on protecting than rescuing and are more confident in coping with uncertainty (job attitude)

[19] According to Cohen et al. (2009), the main reason for not including demographic variables in TPB studies is the fact that many studies are done among college undergraduates who exhibit little demographic variety.

- Experience more social pressure to rely on weather information (social norms)
- Have a higher self-assessed numeracy

2.8. Sample

The components of the Theory of Planned Behavior, as described above, were assessed in a survey distributed online[20] in April and May 2014 to emergency managers throughout the United States. Please see the appendix for the full questionnaire.

The survey was sent out through a variety of channels. The International Association of Emergency Managers (IAEM) published the survey call in their weekly newsletter that reaches ~10,000 emergency management professionals. Additionally, the survey call was posted to the discussion boards of the six LinkedIn-groups for emergency management professionals.[21] Furthermore, the survey call was sent to all email addresses of emergency managers available on the websites of each state's emergency management associations (EMA), the emergency management departments of each state's administration and – if existing – the professional organization in each state. The number of email addresses retrieved on those websites was ca. 2,000.

It is difficult to estimate how many emergency managers received the survey calls. The above mentioned LinkedIn groups and newsletter are also frequented by emergency management professionals that are not emergency manager responsible for an U.S. American jurisdiction. Thus, they do not belong to the target group of this survey. Additionally, it is unknown how many of the email addresses online are up-to-date and/or are checked regularly. Consequently, a survey response rate is difficult to calculate.

[20] Qualtrics was used to build and distribute the online survey.
[21] LinkedIn groups to which the survey call has been posted: Crisis, Emergency & Disaster Recovery Professionals; Emergency Management Professionals; Disaster & Emergency Management; Disaster Researchers and Disaster Management Professionals; Emergency Management and Homeland Security Professionals; International Association of Emergency Managers.

Only U.S. emergency managers who are older than 18 years could participate. As an incentive a raffle in which four participants won a $50 voucher for their favorite store was added to the survey.

3. Results

In the following, the answers to the various survey components are first described. Second, the results of the regression models based on the Theory of Planned Behavior are reported.

3.1. Descriptive Results

3.1.1. Demographics

In total, 363 emergency managers (EMs) started and 207 completed the survey.[22] Most participants were county EMs (127) working full-time (152), and had 5-10 years of work experience as an EM. Given that most were between 51 and 60 years old, it is likely that this job is their second career. Indeed, 101 of the participants are either professional fire fighters (65), paramedics (23), policemen (10) or in the military (3). Another 67 are volunteers in one of these functions (11, 8, 20, and 28 respectively). Figure 4 shows the distributions of these characteristics. EMs from all over the U.S (Figure 3). participated, with especially many from Florida (30), Kansas (24), and North Dakota (14). Twelve states are not represented.[23] Three-quarters of the participants (152) were male.

[22] Most participants dropped out when encountering the first large matrix table asking them to how much the relied on recorded weather data and short-term weather forecasts for a number of decisions.
[23] The twelve states that are not represented are: Connecticut, Delaware, Idaho, Maine, Maryland, Montana, New Hampshire, New Mexico, North Carolina, Rhode Island, Vermont, and West Virginia.

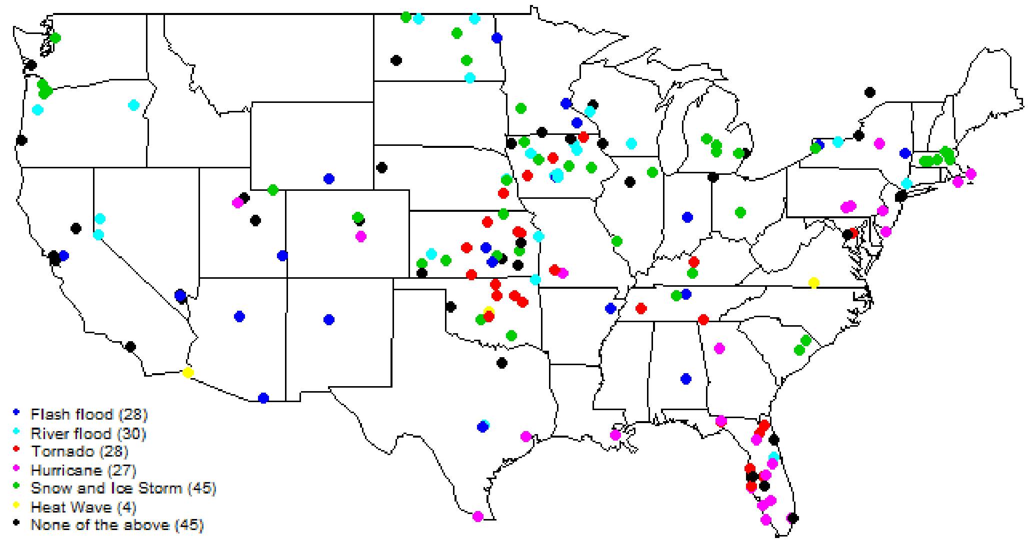


Figure 3: Geographic position of survey participants, color-coded by hazard.

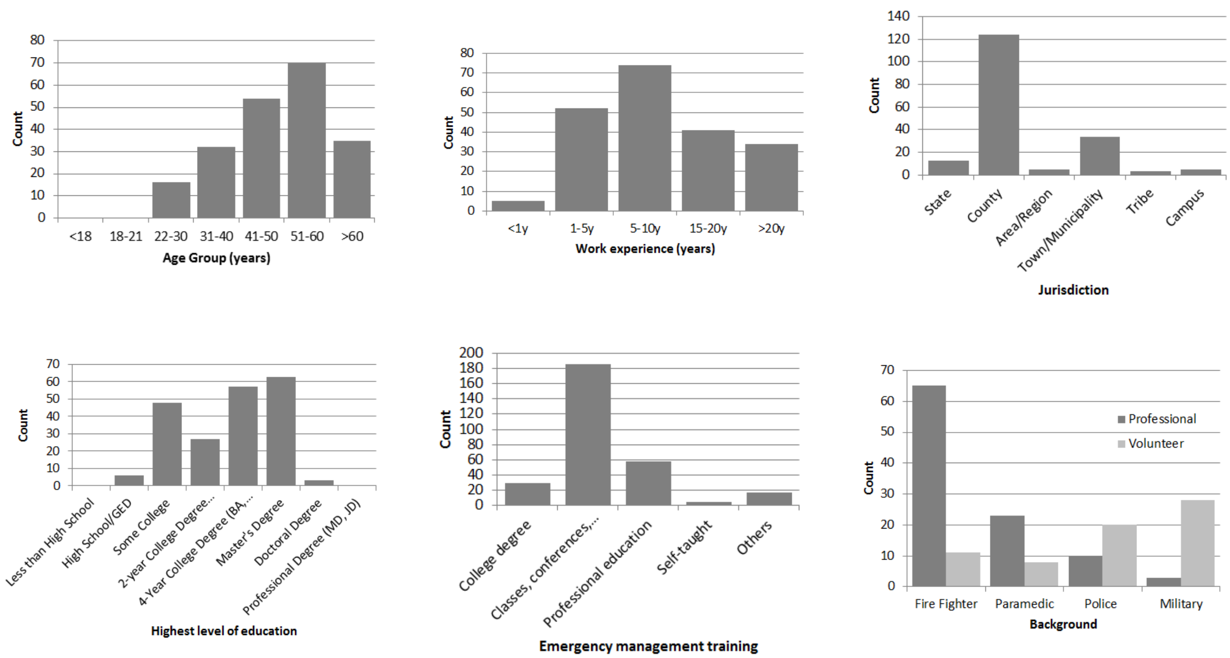


Figure 4: Histograms showing the distribution of the participants' age, work experience, jurisdiction, highest completed level of education, kind of emergency management training and professional background.

Of the 207 participating emergency managers, 48 have attended college, 57 graduated from college and 63 earned a Master degree. To become an EM, the vast majority has followed courses, trainings and workshops (186). Thirty have a college degree related to emergency management; 58 have professional degree; five were self-taught. While our finding that 84% have had weather instructions at some point sounds reassuring, this has to be qualified by the fact that only 31% of courses that EMs take pertain to weather-related events, while 78% of the emergencies faced by EMs are actually caused by weather (Weaver et al*.,* 2014).

When compared to the demographic study of emergency managers by Weaver et al. (2014), our convenience sample seems to be representative of the profession. Of Weaver et al.'s 1,058 participants, 81% were male and 72% older than 45 years. Fischer (1996) reports that the typical local emergency management association director participating in his study was 50 years old and had 14 years of work experience in the field. As in our sample, a high percentage (78%) of Weaver's respondents was college-educated. There seems to be a clear trend to more education. According to Fischer (1996), EMs' highest education was typically a few completed colleges courses. In our and Weaver's (2014) sample, there were much more EMs with college or even graduate degrees. Additionally, Weaver et al. (2014) found that younger EMs tend to have a higher education. In contrast, EMs with a lower education tend to have more experience in other emergency response professions (Weaver et al*.,* 2014). In our sample, 49% had such prior experience, compared to 69% in theirs.

The survey asked which type of weather event had led to the most difficult situation for the EMs. Weaver et al. (2014) report that 78% of disasters that EMs faced in the past were weather-related, while EMs anticipate that to be 63% in the future. With 20 to 25 EMs selecting each hazard, our survey answers are uniformly distributed across flash floods, river floods, tornadoes and hurricanes. Only snow and ice storms and heat wave stand out for being experienced as very difficult respectively very often and not often. Only four EMs selected heat waves, and as many as 45 selected winter storms as the most difficult hazard (see Figure 5). Of the respondents, 22% did not regard any of these six natural hazards as the most difficult and selected "None of the Above". Those EMs only answered the questions on job attitude, subjective numeracy and demographics. Accordingly, 162 EMs were asked the questions regarding weather information.

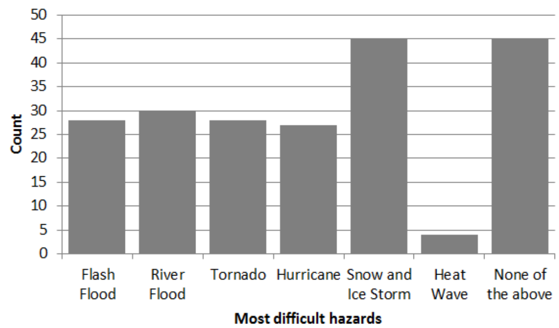


Figure 5: Distribution of hazards that were perceived most difficult to respond to in the past ten years.

3.1.2. Dependent Variable – Past Behavior/Intentions

"How much did you rely on NWS data to make any of the following decisions or carry out any of the following actions?"

The dependent variable in this study is the usefulness of the forecast to make different decisions. We asked how much EMs relied on recorded weather data or short-term weather forecast for different decisions, and how much they intended to rely on those in the future.

As Table 2 shows, significant percentages reported either recorded weather data or short-term weather forecast to be inapplicable to determine where and when to deploy storm spotters, when to initiate evacuation and when to open shelters. EMs used this kind of information least to decide when to deploying storm spotters, which is an activity to gather more data. Of the EMs that did use the data to support their decisions, more than half relied relatively much on forecasts (scores

4-5). Again, the only exception is deploying storm spotters, where 40% of the participants rely much on recorded weather data and forecasts to make decisions.In all cases, EMs intend to rely on recorded weather data and forecasts in the future.

A Multivariate Analysis of Variance (MANOVA) was used to determine, whether the reliance differs across hazards and the type of data. As expected, that perceived usefulness depends on the hazard ($F = 46.2$, $p < 0.001$), whether recorded weather data or forecasts are concerned ($F = 57.9$, $p < 0.001$), and whether the past or intended behavior is studied ($F = 28.3$, $p < 0.001$). Since this is the dependent variable, the correlations with demographic data will be discussed later on.

Additionally, it was asked which type of data EMs would prefer, if they only could have one. Overwhelmingly, across all hazards, the participants opted for the forecast (Table 1). Presumably, they reckon that they could get to know the weather of the past few days somehow, but would have a much harder time making projections for the future.

Table 1: Percentage of EMs who prefer either recorded weather data or short-term weather forecasts, if they only could have one.

	Recorded weather data	Short-term weather forecast	N
Flash Flood	10.7	89.3	28
River Flood	16.7	83.3	30
Tornado	7.1	92.9	28
Hurricane	0.0	100.0	27
Winter	8.9	91.1	45
Heat	0.0	100.0	4
All	8.6	91.4	162

Table 2: Extent to which survey participants relied on the four types of weather data: Past behavior and intended behavior regarding recorded weather data and short-term weather forecasts. In percentages.

	Past Use – Recorded weather data					Intended Use – Recorded weather data					Past Use – Short-term weather forecast					Intended Use – Short-term weather forecast				
%	NA	Low (1-2)	Neut. (3)	High (4-5)	N	NA	Low (1-2)	Neut. (3)	High (4-5)	N	NA	Low (1-2)	Neut. (3)	High (4-5)	N	NA	Low (1-2)	Neut. (3)	High (4-5)	N
Deploy storm spotters	43.8	17.5	8.8	30.0	160	28.3	16.4	19.5	35.8	159	36.5	12.8	10.3	40.4	156	28.2	12.8	15.4	43.6	**156**
Activate EOC	5.0	16.9	16.9	61.3	160	3.8	11.4	16.5	68.4	158	5.7	8.9	5.7	79.6	157	3.2	4.5	5.1	87.3	157
Determine event location	1.9	10.1	20.8	67.3	159	1.9	11.3	16.4	70.4	159	1.3	3.2	9.6	86.0	157	0.6	3.8	8.9	86.6	157
Warn public	3.8	13.9	15.8	66.5	158	4.5	10.2	12.7	72.6	157	5.1	1.3	8.3	85.3	156	4.5	3.2	5.7	86.6	157
Track the event progress	0.6	9.4	17.6	72.3	159	1.9	8.2	16.4	73.6	159	1.9	5.1	11.5	81.5	157	1.3	3.8	9.6	85.3	156
Initiate evacuat.	25.3	13.9	13.9	46.8	158	12.6	14.5	20.1	52.8	159	25.6	10.3	10.9	53.2	156	14.1	8.3	12.2	65.4	156
Open shelters	14.6	16.6	21.0	47.8	157	9.6	15.3	17.8	57.3	157	13.4	11.5	14.6	60.5	157	10.2	7.6	12.1	70.1	157

3.1.3. Perceived Behavior Control: Limitations of Weather data

"Please rate how much the following factors have limited your reliance on NWS data when responding to events."

In this survey, the component of the Theory of Planned Behavior called Perceived Behavior Control is represented by the EMs' perceived limitations of the recorded weather data and the short-term weather forecast.

Table 3: Survey answers how much various issues limited the extent to which EMs relied on weather data in percent. Values in bold were pre-set to be not applicable.

%	Recorded weather data					Short-term weather forecast				
	Not applicable	Little (1-2)	Somewhat (3)	Much (4-5)	N	Not applicable	Little (1-2)	Somewhat (3)	Much (4-5)	N
Inaccurate forecast of event magnitude						9.9	41.7	23.2	25.2	151
Frequent changes to forecast of event magnitude						10.8	44.6	18.9	25.7	148
Inaccurate forecast of event timing						10.7	45.3	19.3	24.7	150
Frequent changes to forecast of event timing						11.5	48.0	16.9	23.6	148
Receiving information too late to be useful	6.9	43.8	30.0	19.4	160	4.5	55.4	14.6	25.5	157
Information unavailable for your area	14.5	45.3	20.1	20.1	159	12.0	48.1	15.8	24.1	158
Insufficient/irrelevant information	9.4	54.1	18.9	17.6	157	7.1	55.8	18.6	18.6	156
Irregular information release times	8.9	55.4	22.3	13.4	162	7.6	55.7	20.3	16.5	158
Your ability to apply the information to the emergency response	6.5	51.6	20.3	21.6	153	3.8	54.8	15.9	25.5	157
Your ability to understand the information	8.6	58.6	15.1	17.8	152	5.2	61.9	12.3	20.6	155

More than half of the respondents felt little to no limitations for most of the characteristics. They were slightly less positive about receiving the information in a timely fashion and it being available for their area. This is true for both recorded weather data and short-term weather forecasts. Regarding recorded weather data, EMs felt that their ability to apply the information to the emergency response and the unavailability of the information for their area were the most hindering characteristics. These issues were important for relying on forecasts, too. But the frequent changes of the forecasted event magnitude, the inaccuracy of the forecast magnitude, and the inaccurate forecasted timing were perceived to be most hindering when wanting to rely on short-term weather forecasts. Unsurprisingly, the participating EMs felt that the characteristics of the short-term weather forecast were somewhat more limiting than those of the recorded weather data.

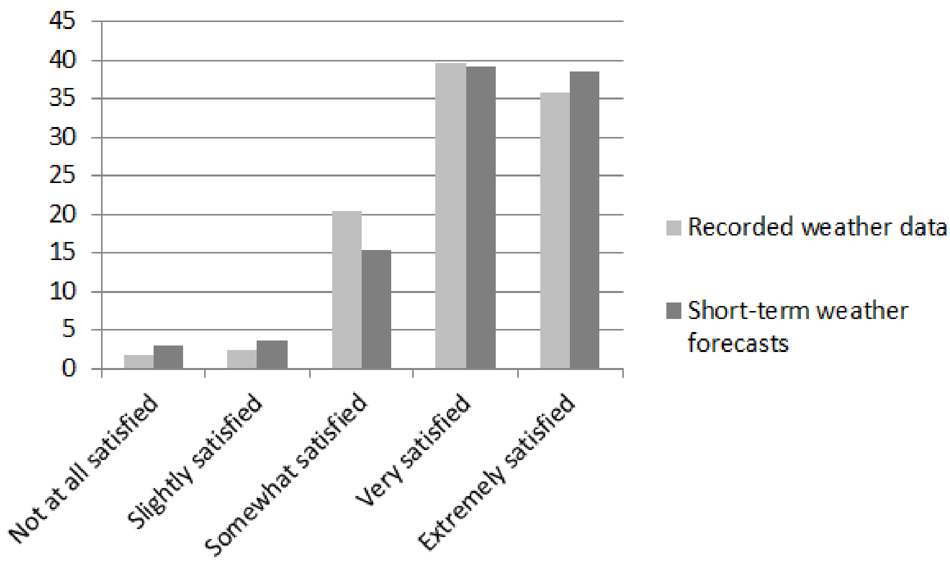


Figure 6: Survey answers to what extent EMs are satisfied with the recorded weather data and short-term weather forecasts by the NWS.

Additionally, we asked how satisfied the participants are with the data from the NWS. As Figure 6 indicates, ca. 75% EMs were very or extremely satisfied. A larger percentage was extremely satisfied with the short-term weather forecasts than with the recorded weather data, even though forecasts come with much more uncertainty than recorded weather data. Equally many people were more satisfied with the forecast than with the recorded data (see Figure 7).

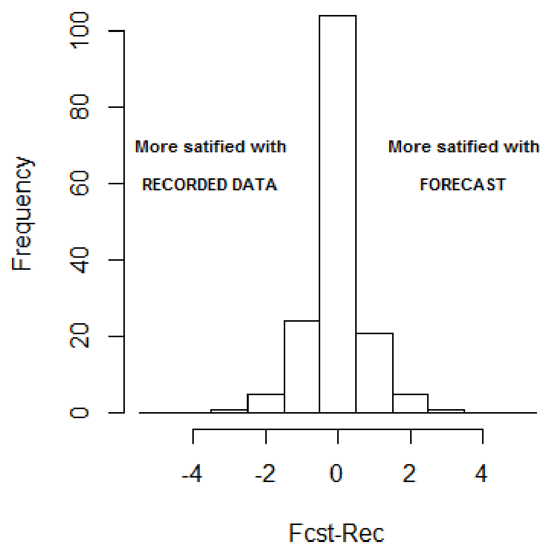


Figure 7: Distribution of participants that were more satisfied with one type of weather informationt than with the other.

A MANOVA revealed that the perceived limitation of the weather forecasts differs by hazard (F=4.4, p=0.001), jurisdiction (F=6.9, p<0.001), work experience (F=8.3, p<0.001), the highest completed level of education (F=22.9, p<0.001), and whether recorded weather data or forecasts are concerned (F=105.2, p<0.001). Additionally, having a background as fireman (F=13.0, p<0.001) or paramedic (F=4.0, p=0.007) also makes a statistical difference.

3.1.4. Social Norms

"Please rate how much the following groups expect you to rely on NWS data when responding to an event."

"Please rate how much you worry about criticism from the following groups when responding to an event."

The questions about social norms assess the extent to which the EMs feel pressure from their surroundings to rely on recorded weather data and short-term weather forecasts. Table 4 shows that, EMs seem to overwhelmingly think that the people around them expect them to rely on both forecasts and recorded weather data. For forecasts, EMs think that locals such as EM colleagues, city employees, elected officials, and residents want them to rely on forecasts more than others. For recorded data, that distribution is relatively flat. In both cases, EMs most often judge the training and workshop instructors, government officials and the NWS as not applicable; presumably, because they miss local knowledge. Nonetheless, these groups have the most knowledge about forecasts.

Table 4: Survey answers to what extent EMs think that a variety of social groups wants them to rely on either of the two types of weather data. In percent.

%	**Recorded weather data**					**Short-term weather forecast**				
	Not applicable	Little (1-2)	Somewhat (3)	Much (4-5)	N	Not applicable	Little (1-2)	Somewhat (3)	Much (4-5)	N
Local and regional emergency management colleagues	1.2	2.1	16.9	78.8	160	0.6	1.9	9.5	88.0	158
City employees	3.8	6.3	15.7	74.2	159	3.2	3.8	8.9	84.2	158
Elected officials	3.8	5.7	15.1	75.5	159	3.2	3.2	9.5	84.2	158
NWS employees	14.5	3.8	15.7	66.0	159	15.5	3.2	11.6	69.7	155
Employees of government agencies	11.3	8.8	23.9	56.0	159	10.7	5.7	21.4	62.3	159
Training/workshop instructors	20.8	11.9	18.9	48.4	159	21.0	12.1	16.6	50.3	157
Residents	2.5	4.4	16.4	76.7	159	3.8	3.2	11.5	81.4	156
The Media	3.7	5.6	21.7	68.9	161	3.8	6.3	17.7	72.2	158
The general public	2.5	6.2	13.7	77.6	161	2.5	3.8	11.3	82.4	159
Family & Friends	4.4	3.8	18.1	73.8	160	5.0	4.4	12.6	78.0	159

The second set of questions on social norms asked how much EMs worried about being criticized by each social group (Table 5). The opinions of their emergency management colleagues, employees of the NWS and of government agencies are most valuable to EMs; probably, EMs judge these groups to be most competent in working with such data. The answers regarding all other groups have bimodal distributions. That bimodal distribution is mirrored for groups representing officials as compared to groups representing the public.

Table 5: Survey answers to what extent EMs worry about being criticized by a variety of social groups. In percent. (N=162)

	Not applicable	Little (1-2)	Somewhat (3)	Much (4-5)
Local and regional emergency management colleagues	3.1	24.7	26.5	45.7
City employees	1.2	39.5	17.3	42.0
Elected officials	1.2	53.1	13.0	32.7
NWS employees	4.9	17.9	35.2	42.0
Employees of government agencies	3.7	24.1	29.6	42.6
Training/workshop instructors	11.1	16.7	38.3	34.0
Residents	0.6	58.6	9.3	31.5
The Media	0.6	44.4	15.4	39.5
The general public	0.6	59.3	9.3	30.9
Family & Friends	2.5	42.0	20.4	35.2

Multiplying the expectations with the values associated with groups results in the overall impact of each social group. Table 6 ranks the groups by their influence. Training and workshop instructors seem to have the least influence on the EMs' perception whether they should rely on recorded weather data and forecasts. The opinions of their colleagues at the emergency department and in the town hall in general are most relevant to the EMs. This is followed by NWS employees, the media, and family and friends. Therefore, managing the use of weather data is probably most effectively done through peer education and with direct contacts between EMs and the NWS. Probably, it is the *local* knowledge that makes these opinions more valuable than the views of workshop and training instructors. Additionally, Morss et al. (2005) mention that practitioners prefer to talk to people with whom they have had long-term relationships when seeking scientific information.

Table 6: Ranking of most influential groups for recorded weather data and short-term weather forecasts

Overall Impact Recorded weather data	Overall Impact Short-term weather forecast
Local and regional emergency management colleagues	Local and regional emergency management colleagues
City employees (for example, city departments, fire, police, EMS)	City employees (for example, city departments, fire, police, EMS)
Employees of the National Weather Service	Employees of the National Weather Service
The Media	Family & Friends
Family & Friends	The Media
Employees of government agencies (for example, Army Corps of Engineers, FEMA)	Employees of government agencies (for example, Army Corps of Engineers, FEMA)
Elected Officials	Elected Officials
Residents	The general public
The general public	Residents
Training/workshop instructors	Training/workshop instructors

Last, the survey asked to what extent EMs value the opinion of others in general (Table 7). Most EMs, 60.7%, agreed that taking into account the expectations of others is useful. Additionally, it was asked whether the participants worry about liability when responding to emergencies. The vote is less clear, featuring a bimodal distribution. While 17.9% are undecided, 34.0% disagree and 29.0% agree that they are concerned about liability. In any case, a substantial 37.6% worry about liability.

Table 7: Survey answers to what extent EMs agree or disagree with two statements about valuing the opinion of others and worrying about liability. (N=162)

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
In general, taking into account the expectations of others is useful when responding to an event.	3.1	9.9	9.9	60.5	16.7
I worry about potential or possible liability claims made against me or my employer when responding to an event.	10.5	34.0	17.9	29.0	8.6

A MANOVA indicates that the type of hazard (F=11.3, p<0.001), the jurisdiction (F=5.8, p<0.001), the work experience (F=3.0, p=0.019), the highest completed level of education (F=2.5, p=0.023), and a background in the military (F=7.0, p=0.008), as a fireman (F=2.3, p=0.050) and paramedic (F=3.1, p=0.079) make a statistical difference in the overall impact of the social groups.

3.1.5. Attitude: Weather Information

The questions about the benefits using NWS weather data received overwhelmingly positive answers, for both recorded weather data and short-term weather forecasts (see Table 8). Across all questions, more than 80% agreed or strongly agreed that using NWS was beneficial in various ways. The one exception is the question about whether these data were the most important source of information. In that case, EMs agreed or strongly agreed with that statement regarding recorded weather data (60%) and forecasts (70%).

A MANOVA indicated that the beneficial perception of the weather information depends on hazard (F=7.1, p<0.001), the EMs' jurisdiction (F=8.1, p<0.001), the work experience (F=3.3, p=0.011), the highest completed level of education (F=4.3, p<0.001), and whether recorded weather data or forecasts were concerned (F=12.0, p<0.001). Additionally, the professional background has a statistically significant impact: Fire fighter (p=9.5, p<0.001), policeman (F=4.5, p=0.004), paramedic (F=3.5, p=0.016), and military (F=0.5, p<0.001).

Table 8: Answers to questions about the benefits of using weather data in percent. The right side of the table (agree/strongly agree) corresponds with a positive attitude towards using weather data. For the question with the minus in front, the scale is reversed. (N=153)

%	**Recorded weather data**				**Short-term weather forecast**			
	Disagree (1-2)	Undecided (3)	Agree (4-5)	N	Disagree (1-2)	Undecided (3)	Agree (4-5)	N
Using NWS data to respond to events has resulted in better decisions/actions.	5.0	11.2	83.9	161	2.6	2.6	94.9	156
NWS data is the most important source of information for responding to events.	15.6	24.4	60.0	160	12.9	16.8	70.3	155
(-) Relying to NWS data to respond to events can be harmful.	78.9	11.2	9.9	161	79.5	7.7	12.8	156
I would recommend other emergency managers to rely on NWS data to respond to events.	1.9	16.8	81.4	161	0.6	11.5	87.8	156
NWS data gives me confidence in my decisions/actions during responses to events.	5.0	13.0	82.0	161	11.3	12.3	86.4	154

3.1.6. Attitude: Job

The job attitude was assessed to figure out, whether emergency managers had a more preventive or responsive mindset. It is hypothesized that intensive forecast use does not match well with a responsive mindset. A majority (65.7%) views their job to protect rather than to rescue citizens. They are divided on the question, whether they are supposed to prevent or to respond to hazardous situations. A slight majority votes for the latter (51.7%), see Table 9. EMs are divided in their attitude towards uncertainty. Those questions feature bipolar distributions. Between 57% and 72%

(strongly) disagrees that uncertainty hinders them in their work. Morss et al. (2005) explain that "practitioners often deal with uncertainty by finding the best information they can quickly and easily obtain and interpret, making the decision require for the moment, *and moving on*." [Emphasis added.] EMs might simply experience uncertainty not as a problem but as a daily given.

Table 9: Answers to job attitude questions in percent. For the questions with the minus in front, the left side corresponds with a preventive mindset. For the other questions, the scale is reversed. (N=207)

%	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
It is my job to prevent rather than to respond to hazardous situations.	13.5	38.2	13.0	29.0	6.3
It is my job to protect rather than to rescue citizens.	5.3	17.9	11.1	50.2	15.5
(-) I pay more attention to the current situation and less to possible sequences of events when making a plan to respond to an event.	15.0	57.0	12.6	14.5	1.0
I routinely think in what-if scenarios when responding to events.	0.0	2.9	3.4	57.0	36.7
(-) Not knowing what will happen during an event makes it difficult for me to respond.	10.6	46.4	10.6	27.1	5.3
(-) A critical situation can develop in so many different ways, it is difficult to determine appropriate actions.	6.8	53.6	13.0	25.6	1.0

Table 10: Binary question to assess whether EMs react intuitively to emergencies. (N=207)

When faced with a critical situation, I consider all facts, figures, and different scenarios and weigh my options before I take action.	71%
When faced with a critical situation, I know what to do, when I see what is going on around me, and hear what is happening.	29%

Additionally, we posed the binary question, whether EMs react rather intuitively or whether they deliberately consider the information and options they have before taking action (Table 10).Of the 207 participants, 29% assessed themselves as responding intuitively.

Using a hypothetic tornado case study, Weaver et al. (2014) found that many EMs decide to react very late. For example, 22.2% would only come into action by the time that the tornado has already caused damage, while 30.3% would initiate a full response after the Doppler radar has confirmed a tornado 30 minutes away (Weaver et al., 2014). Combining these results with ours, it seems that EMs do distinguish themselves from professions like fire fighters and paramedics in the sense that they rather protect than rescue people, i.e., EMs try to get people out of harm's way beforehand. However, they do not see themselves as having to prevent hazardous situations. That might explain the late responses found by Weaver et al. (2014). Maybe, EMs simply regard 15 minutes as enough to reach a tornado shelter. Any damage prevention beyond that might fall outside the perceived scope of the EMs' task. Given that EMs do not seem to experience uncertainty as seriously hindering their work, the late response might also be meant to reduce the number of false alarms. In a previous study (Study 1), we found that EMs do choose the weather information with the least uncertainty in them. But apparently, a majority does not think that uncertainty prevents them from acting.

A MANOVA indicated, that the most concerning hazards make a difference as to how EMs view their job ($F=6.7$, $p<0.001$). It was hypothesized that a background in some other emergency response profession might impact the understanding of the job as well. This is indeed the case for policemen ($F=8.1$, $p<0.001$), fire fighters ($F=12.2$, $p<0.001$), and EMs associated with the military ($F=8.4$, $p<0.001$), but not for paramedics.

3.1.7. Subjective Numeracy

The first four subjective numeracy questions asked for the participants' confidence in working with fractions and percentages. This was measured on a scale from 1 (not at all confident) to 6 (very confident). The remaining four questions asked for the preference for words or numbers. The scale goes from 1 (preference for words) to 6 (preference for numbers).

The results are summarized in Table 11. The participants were confident that they can work with fractions and percentages (average 4.8) and clearly preferred numbers (average 4.7). The only question pertaining to fractions ("How good are you at working with fractions?") received the lowest average score (4.2), indicating that that EMs feel more comfortable working with percentages than with fractions. While EMs generally seem to prefer numbers over words, they do slightly less so when talking to people directly (rather than reading the information.) That question received the lowest average score for that set of questions (4.4). The highest average score was for the most general question, whether numerical information is generally useful.

When comparing these scores with the values that Fagerlin et al. (2007) report in their paper that introduces the subjective numeracy scale (SNS), EMs score considerably higher than the 287 respondents recruited from a Veterans Affairs hospital waiting room (see comparison in Table 11). Fagerlin et al. (2007) report that these participants had lower numeracy skills than their university participants (4.1 vs. 4.6), but they do not report the latter score per question. Another study (N=318) – comparing the recruiting techniques on Mechanical Turk, at university and on internet boards – found an average subjective numeracy of 4.35, 4.17, and 4.25 respectively (Paolacci et al., 2010). Zikmund-Fisher et al. (2007) reported a median subject numeracy of 4.2 and 4.5 in studies of 996 and 1270 people recruited online by a professional company. The large percentage of male EMs is the most prominent difference between our participants and those in the literature. In any case, EMs' subjective numeracy is well above average.

An MANOVA analysis was done to identify any demographic factors correlated to numeracy. Work experience, age, emergency management education, and background in other emergency-related profession all did not correlate with subjective numeracy. Only the highest level of education did (F=5.12, p<0.001), as illustrated by Figure 8.

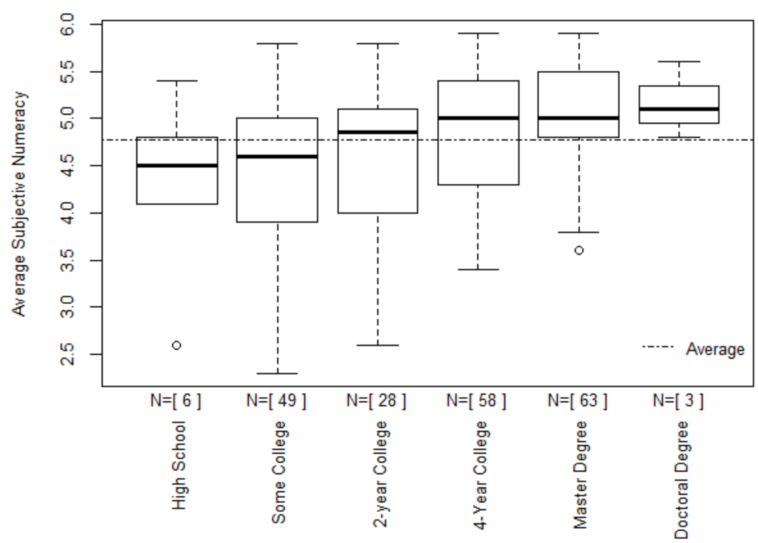


Figure 8: Relationship between subjective numeracy and the highest completed level of education.

Table 11: Survey answers numeracy including the correlation to objective numeracy and comparison to results reported in Fagerlin et al. **(2007). (N=207)**

Fraction & Percentages				Numbers vs. words			
Question	**Average/ St.Dev**	**Average Veterans***	**Corr. obj. num.***	**Question**	**Average/ St.Dev**	**Average Veterans***	**Corr. obj. num.***
1. How good are you at figuring out how much a shirt will cost, if it is 25% off?	5.1 / 1.0	4.6	0.73	5. When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?	4.6 / 1.1	3.8	0.57
2. How good are you at working with percentages?	4.9 / 1.1	3.6	0.80	6. When people tell you the chance of something happening, do you prefer that they use words (for example, "it rarely happens") or numbers (for example, "there's a 1% chance")?	4.4 / 1.3	3.5	0.56
3. How good are you at calculating a 15% tip?	5.0 / 1.0	4.2	0.76	7. When you hear a weather forecast, do you prefer predictions using percentages (for example, "there will be a 20% chance of rain today") or predictions using only words (e.g., "there is a small chance of rain today")?	4.8 / 1.2	3.1	0.61
4. How good are you at working with fractions?	4.2 / 1.4	3.7	0.72	8. How often do you find numerical information to be useful in your daily life?	5.1 / 0.9	4.2	0.51
Average	4.8			**Average**	4.7		
Scale: 1 (Not at all good) – 6 (Very good)				**Scale**: 1 (Prefer words/Numbers not useful) – 6 (Prefer numbers/Numbers are useful)			
***As reported by Fagerlin** et al. **(2007)**							

3.2. Inferential Results

3.2.1. Principal Component Analyses

To reduce the number of variables, principle component analysis (PCA) was run for each of the components of the Theory of Planned Behavior in order to aggregate their items into single variables. Based on the screeplot, the eigenvalues and non-trivial factor loadings, the best number of components to be included in the regressions was determined. Prior to the PCA, the variables were normalized. The question "*Relying to NWS data to respond to events can be harmful.*" was excluded from the PCA for attitude towards the weather information, because it correlates poorly with the other questions about attitude towards the weather information. It was the only one in that set of questions that was worded negatively.

Table 12: Results of Principal Component Analysis for each of the variables of the Theory of Planned Behavior.

		Number of questions	Number of components	% cumulative variance
Dependent Variable	**Past Behavior – Recorded weather data**	7	1	53.9
	Past Behavior – Forecasts	7	1	44.2
	Intentions – Recorded weather data	7	1	50.1
	Intentions - Forecasts	7	1	55.1
Social Norm	**Recorded weather data**	10	2	58.3
	Forecasts	10	2	57.5
Perceived Behavioral Control	**Recorded weather data**	10	3	81.4
	Forecasts	10	3	80.5
Data Attitude	**Recorded weather data**	4	1	69.3
	Forecasts	4	1	67.6
Job Attitude	**-**	6	2	51.2
Numeracy	**-**	8	2	63.6
Total Indep.			10	

Table 12 shows the result. The chosen components cover only relatively small amounts of the variance. Only for perceived behavior control (i.e., the limitations of the weather data) do the components include more than 80% of the variance.

In all conducted PCAs, the first component (called "General Component" hereafter) is a combination of the underlying variables with almost equal weights. As can happen with highly correlated variables, the weights of the first component are negative; the first component of job attitude being the only exception. Therefore, in the regression, a negative coefficient for job attitude and demographic variables indicates a stronger reliance on the type of weather data in question.

Social norms, planned behavioral control and job attitude have more than one component. The second component of social norms "Officials vs. Public" contrasts EM colleagues, NWS and government employees, and training instructors against the public, such as residents and the media. For perceived behavioral control, the second component "Information quality" has positive weights for inaccuracy of and often changes to event magnitude and timing. The third PBC component stresses the two questions pertaining to self-efficacy for recorded data "REC_User Abilities" and forecast dissemination for short-term forecasts "FCST_Dissemination". The second component of the job attitude "Mindset vs. uncertainty" contrasts the questions on a preventive versus a responsive mindset with those on coping with uncertainty. The second component of numeracy "Preference vs. Confidence" sets the preference for numbers or words apart from the confidence in working with fractions and percentages. The loadings for all components can be found in the appendix.

3.2.2. Correlations among TPB variables

As the TPB suggests, past behavior and intended future behavior are indeed strongly correlated; r=0.77 and r=0.72 for recorded weather data and forecasts respectively.

The relatively low correlations between the variables themselves, at most exceeding r=0.3, indicate that the TPB variables do measure very different aspects of recorded weather data and forecast use. However, some interesting correlations do exist. Overall perceived limitation (Perceived Limitations General) is enhanced by lower subjective numeracy (Numeracy General). Data attitude and overall subjective numeracy (Numeracy General) correlates with the overall social norms (Social Norms General, Officials vs. Public), indicating that the EMs might be reflecting some of their own perception onto others, e.g., they might (unconsciously) assume that others think similar to themselves. The attitude towards the data (Data Attitude) is clearly driven by the ability to understand and apply the data, its timeliness, availability for the area, regular releases and sufficient information (Information Quality) in the case of recorded weather data and perceived limitations overall (Perceived Limitations General) for forecasts.

For forecasts, a more preventive mindset (Job Attitude General, Mindset vs. Uncertainty) correlates with more pressure by social norms (Social Norms General). Possibly, the perceived expectations of others to rely on forecasts lead to the EMs adopting such a mindset.

For recorded weather data, EMs with a high subjective numeracy (Numeracy General) perceived fewer limitations (Perceived Limitations General). Using recorded weather data for emergency response requires some extrapolation of those data somewhat into the future, as it only describes the past. Possibly, EMs with a higher subjective numeracy feel more confident in doing this. Additionally, preferring numbers over of words (or having less confidence in working with fraction and percentages) (Preference vs. Mindset) correlates with a higher perception of the data being unavailable for the area, receiving it too late, having insufficient information and irregular release times (or having more confidence in the ability to understand and apply the information) (User Ability).

Table 13: Intercorrelations between TPB variables for recorded weather data.

RECORDED WEATHER DATA	**Past Behavi.**	**Intent.**	**Gen. Soc. N.**	**Officials vs. Publ.**	**Gen. Data Att.**	**Gen. Perc. L.**	**Info. Quality**	**User Abilities**	**Gen. Num.**	**Pref. vs. Confid.**	**Gen. Job. Att.**	**Mind. vs. Uncert.**
Past Behav.	1.0											
Intent.	0.77***	1.0										
Gen. Soc. N	**0.32*****	**0.28*****	1.0									
Officials vs. Public	0.01	0.06	-0.02	1.0								
Gen. Data Att.	**0.26*****	**0.32*****	**0.31*****	-0.01	1.0							
Gen. Perc. Lim.	-0.05	-0.06	0.10	**-0.23*****	-0.09	1.0						
Information Quality	-0.07	-0.06	**-0.14***	0.00	**-0.22*****	-0.01	1.0					
User Abilities	0.07	0.09	-0.06	0.06	0.09	-0.02	-0.03	1.0				
Gen. Num.	**0.00**	-0.04	-0.11	0.07	0.01	**-0.21*****	-0.12	-0.03	1.0			
Preference vs. Confidence	-0.04	-0.05	0.12	-0.09	0.03	-0.10	-0.02	**0.15***	0.00	1.0		
Gen. Job. Att.	0.06	0.17	0.12	0.08	0.10	-0.07	0.04	0.01	0.08	-0.02	1.0	
Mindset vs. Uncertainty	-0.01	0.05	-0.06	-0.02	**0.16****	0.07	0.13	**0.14***	-0.12	-0.08	0.00	1.0

P-value: *** - <0.01; ** - 0.05; * - 0.1

Table 14: Intercorrelations between TPB variables for short-term weather forecasts.

WEATHER FORECASTS	Past Behavi.	Intent.	Gen. Soc. N.	Officials vs. Public	Gen. Data Att.	Gen. Perc. L.	Info. Quality	Dis-semin.	Gen. Num.	Pref. vs. Confid.	Gen. Job. Att.	Mind. vs. Uncert.
Past Behav.	1.0											
Intent.	0.72***	1.0										
Gen. Soc. N	**0.25*** **	**0.27*** **	1.0									
Officials vs. Public	-0.12	-0.05	0.01	1.0								
Gen. Data Att.	**0.30*** **	**0.41*** **	**0.17** **	-0.04	1.0							
Gen. Perc. Lim.	-0.06	-0.05	0.03	**-0.16***	**-0.25*** **	1.0						
Info. Quality	0.04	0.09	-0.01	-0.09	0.01	0.01	1.0					
Dissemination	0.11	0.07	-0.09	0.08	0.09	-0.02	-0.01	1.0				
Gen. Num.	-0.07	-0.08	-0.07	0.11	0.09	-0.14	-0.10	0.04	1.0			
Preference vs. Confidence	-0.01	-0.11	0.09	-0.10	-0.02	-0.09	006	0.01	0.00	1.0		
Gen. Job. Att.	-0.07	0.10	**0.13***	0.06	0.05	-0.12	0.03	-0.04	0.08	-0.02	1.0	
Mindset vs. Uncertainty	-0.05	0.10	**0.14***	0.02	0.11	0.01	0.09	0.09	-0.02	-0.08	0.00	1.0

P-value: *** - <0.01; ** - 0.05; * - 0.1

For recorded weather data, feeling confident about coping with uncertainty (Mindset vs. Uncertainty) is positively related to feeling more able to understand and apply the data (User Ability).

3.2.3. Regressions

In total, four regression models were built: one for past and intended future behavior for each recorded weather data and short-term weather forecasts. To reduce the numbers of variables, in a previous step, we tested which demographic variables are relevant. This turned out to be work experience, the type of hazard and EM education "Other." For the short-term weather forecasts, weather instruction and a police background were also added to the models. All other demographic variables were excluded from the four main models. Multicollinearity was not a problem, because the variance inflation factors for all independent variables in the regression models are between 1.0 and 1.5.

The regression results in Table 16 and Table 17 show that, generally, the models for short-term weather forecasts were able to explain more variance than those for the recorded weather data. Similarly, the models perform better for intended future behavior than for past behavior.

Work experience and type of hazard

Work experience and the hazard "Snow and Ice Storm" are the dominant variables for recorded weather data. Apparently, EMs rely less on recorded weather data for winter storms than for other hazards. Supporting this finding, Steward et al. (2004) found that forecasts do not include pavement temperature, which is a critical variable for decisions about icing and snow removal. Hence, thruway supervisers have to make their own estimate or act intuitively (Stewart et al., 2004).

Confirming the findings of Baumgart et al. (2007), people with less than one year of work experience rely very much less on weather information. Especially those with 5-10 years or more than 20 years tend to rely more on recorded weather data than others. This supports Pielke's and Conant's (2003) opinion that "weather forecasts have value not because they are by any means perfect, but because the vast experience of users of those predictions fosters the incorporation of them into the decision routines." While work experience would be regarded positive in this context, Morss (2010) reports that expectations are often anchored to past experiences which can lead to inadequate emergency responses.

TPB variables

For past behavior the perceived limitations – as opposed to self-efficacy – (User Ability), and for intended future behavior the confidence to cope with uncertainty (Mindset vs. Uncertainty) are significantly positively correlated with reliance on recorded weather data. However, compared to the demographic variables their weight is rather small. A possible explanation is that using recorded weather data requires the ability to extrapolate the data somewhat and the confidence to do so, given the uncertainty of the future.

For forecast data, having received weather instructions and being a policeman by profession weigh into the regression considerably besides work experience and type of hazard. Work experience has a heavier influence on past behavior than on intention for the future. Having work experience apparently leads to a heavier reliance on forecasts. People with work experience of 5-10 years rely on forecasts most. Not having received weather instructions is related with much less reliance on the forecast. Last, EMs who are simultaneously policemen by profession, i.e., wear two hats in their job, tend to rely less on forecasts.

For forecasts, more variables of the Theory of Planned Behavior are statistically significant than for recorded weather data, most notably the data attitude. Having a positive attitude leads to a stronger reliance on forecasts. For past behavior, perceived higher social expectations (Social Norm

General) resulted in higher reliance on the forecast. Additionally, EMs who perceived unavailability for the area, insufficient information, irregular lead times and receiving the forecast too late (Dissemination) as more of a problem, relied on forecasts less. This confirms the finding that perceived limitations of the forecast – rather than self-efficacy – influenced the use of recorded weather data and forecasts in the past, but does not inform intentions for the future.

For intended future behavior, additionally, the overall subjective numeracy (Numeracy General) is statistically significant. Feeling more confident in this regard seems to be related with a stronger intention to rely on forecasts in the future.

The picture changes considerably, when demographic variables are not included in the regressions. Then, social norms (Social Norm General) and data attitude are the driving factors for both recorded weather data and forecasts (Figure 9). These two variables are precisely the ones that Artikov et al. (2006) found to be significant predictors for the use of climate forecasts by farmers. A possible interpretation is, that work experience translates to the EMs having learned to listen to their social surroundings and having experienced the benefits of using such data to inform their decisions. Supporting this finding, Golden and Adams (2000) report that "close coordination among Weather Forecast Offices, local media, and local/state emergency management offices" is a major contributor to successful tornado warnings. Artikov et al. (2006) offer a slightly different interpretation. They think that farmers might decide whether to use climate forecasts after consulting others or might simply do what everyone else around them does. This could also be the case for EMs. For example, Morss and Ralph (2007) found that EMs find personal relationships with forecasters very important, when faced with a critical situation.

Studying the correlations among the TPB variables showed that both social norms and data attitude correlate with each other and the perceived limitations of the weather information (Figure

9). Given that that perceived limitations are related to both social norms and data attitude, it seems plausible that EMs reflect their own perception of the weather information onto their social surroundings, when answering the survey questions. At the end of the causal chain, subjective numeracy is correlated to the perceived limitations of recorded weather data and more weakly with those of forecasts (Figure 9). But it is also possible that a positive attitude of the social surrounding influence the EMs' attitude towards and perception of the limitation of the data. Not taking into account social norms and numeracy, Demuth et al. (2011) found that perceptions of the forecasts (i.e., the perceived importance of NWS information, and confidence in the forecasts) are significant predictors of the frequency with which the public consults weather forecasts. Additionally, the general public seems to obtain forecasts *more* often when the forecast error (i.e., the actual error, rather than the perceived limitations as measured in our study) is larger (Demuth et al., 2011).

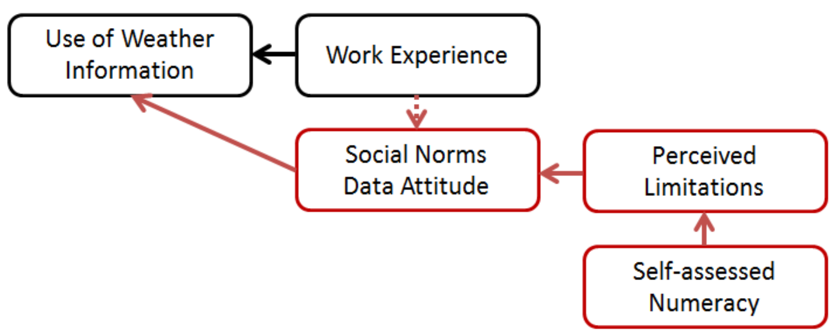
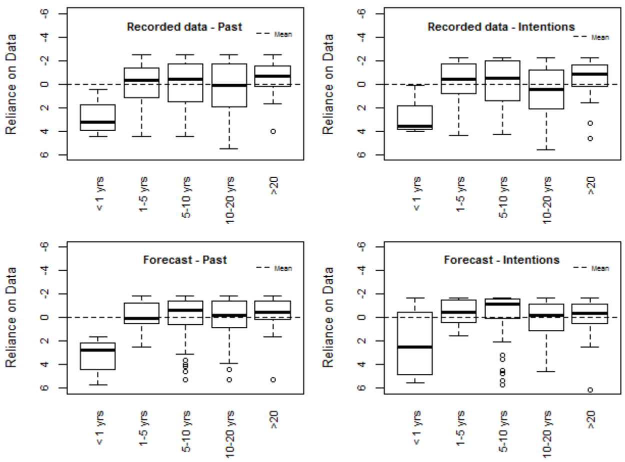


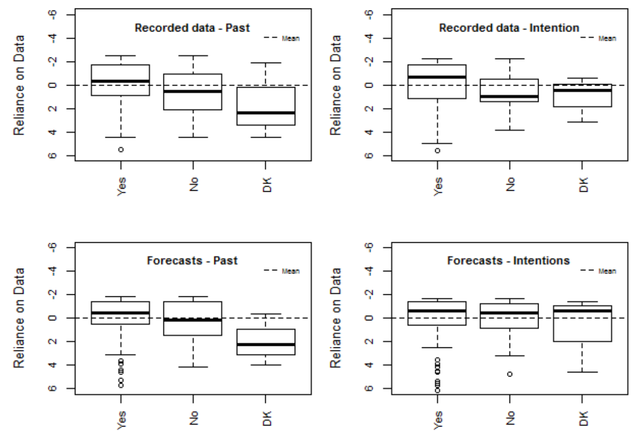
Figure 9: Correlations in the Theory of Planned Behavior model when including when including demographic data (black) and excluding it (red).

Table 15: Boxplot and scatterplots of the most relevant variables in the four regression models: Work experience, having received weather instructions, social norms, perceived limitations of and attitude towards recorded weather data and short-term weather forecasts, job attitude, type of hazard, and numeracy.

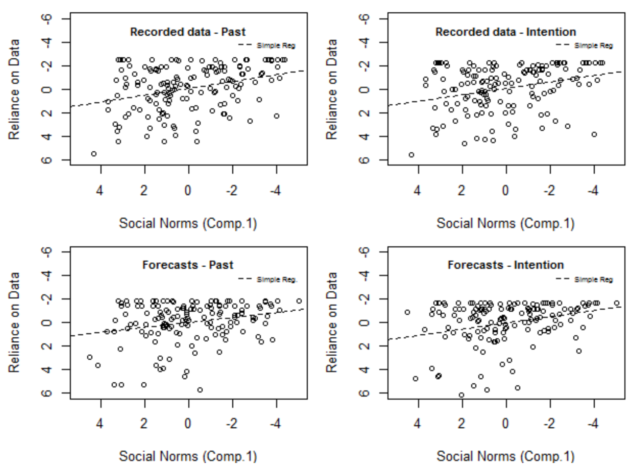
Work Experience



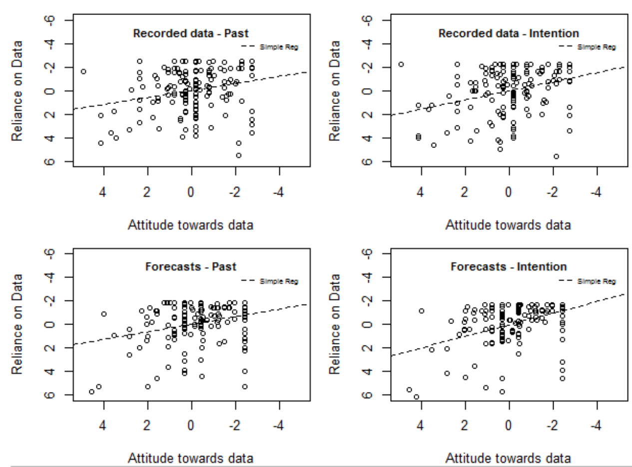
Having received weather instructions



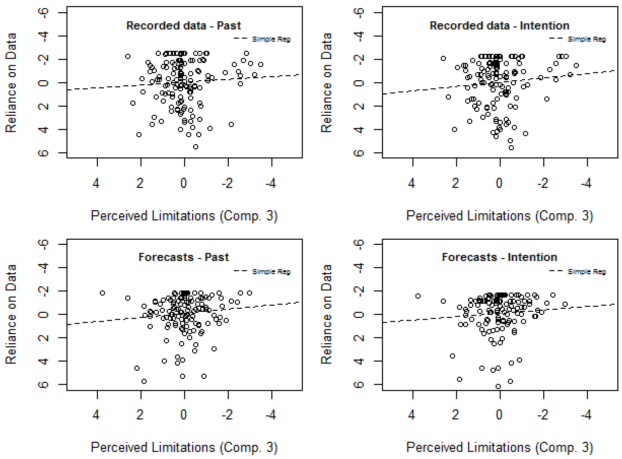
Social Norms: General



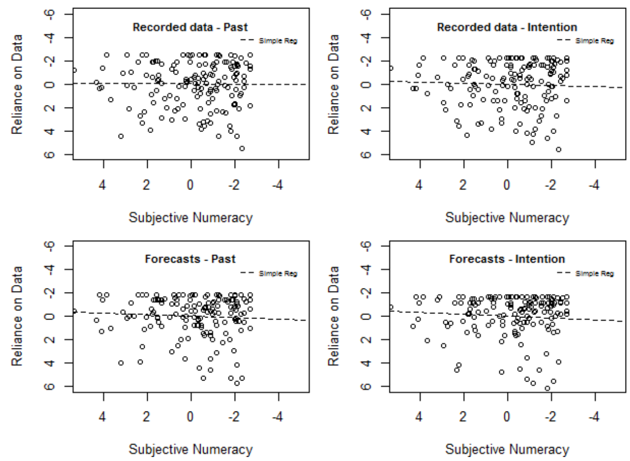
Data Attitude



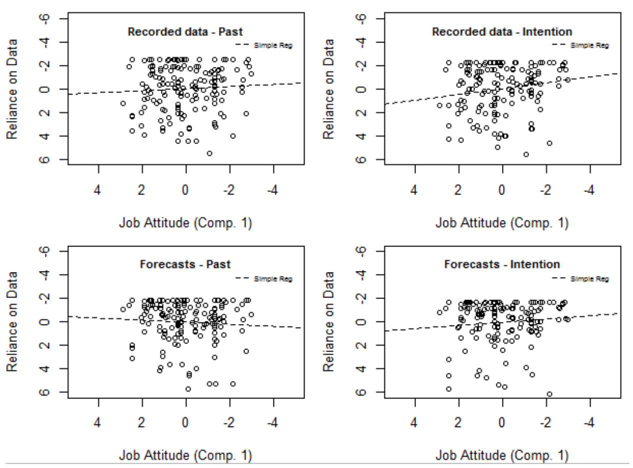
Perceived Limitations:
REC_User ability, FCST_Dissemination



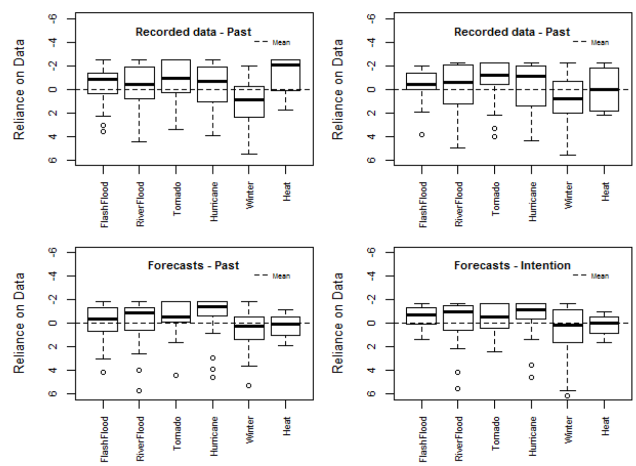
Numeracy General



Job Attitude: General



Hazard



Police Background

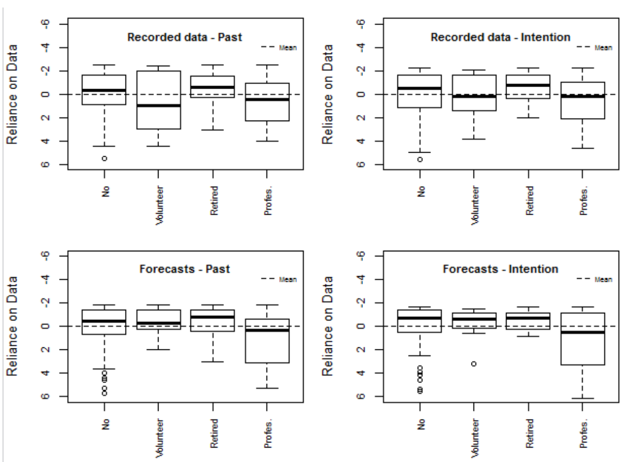


Table 16: Regression results based on PCA for <u>recorded weather data</u> with and without demographic variables.

	Past Behavior			Intended Future Behavior			Past Behavior			Intended Future Behavior		
	Coef	SE		Coef	SE		Coef	SE		Coef	SE	
Intercept	2.43	1.10	*	2.64	1.47	.	-0.10	0.16		-0.10	0.17	
PBC												
General	-0.08	0.08		-0.08	0.08		-0.05	0.08		-0.05	0.09	
Info. Quality	-0.10	0.11		-0.07	0.11		-0.06	0.11		-0.04	0.12	
User Ability	**0.17**	**0.15**	*	0.16	0.16		-0.16	0.15		0.16	0.16	
Social Norms												
General	0.20	0.09		0.14	0.09		**0.28**	**0.09**	**	**0.22**	**0.09**	*
Officials vs. Public	0.07	0.14		0.16	0.15		0.06	0.14		0.12	0.15	
Data Attitude	0.07	0.11		0.18	0.11		0.15	0.11		**0.24**	**0.11**	*
Job Attitude												
General	0.10	0.12		**0.26**	**0.12**	*	0.06	0.12		0.21	0.12	.
Mindset vs. Uncertainty	-0.05	0.15		-0.004	0.16		-0.02	0.15		0.11	0.16	
Numeracy												
General	-0.02	0.04		-0.11	0.09		-0.002	0.09		-0.05	0.09	
Prefer. vs. Confidence	-0.04	0.14		-0.01	0.14		-0.15	0.14		-0.08	0.15	
Hazard												
River Flood	0.08	0.51		0.13	0.56							
Tornado	-0.20	0.55		-01.8	0.59							
Hurricane	0.30	0.57		0.52	0.61							
Winter Storm	**1.41**	**0.50**	**	**1.11**	**0.54**	*						
Heat Wave	-1.0	0.99		0.26	1.03							
Work Exp.												
1-5 years	**-2.72**	**1.11**	*	**-2.97**	**1.42**	*						
5-10 years	**-3.04**	**1.11**	**	**-3.63**	**1.43**	*						
10-20 years	**-3.03**	**1.15**	**	**-3.02**	**1.44**	*						
> 20 years	**-3.24**	**1.14**	**	**-3.59**	**1.45**	*						
EM Educ. Other	0.57	0.59		**1.83**	**0.61**	**						
R^2		0.32			0.35			0.16			0.19	
Adjusted R^2		0.20			0.27			0.09			0.13	
P-Value		<0.001			<0.001			0.013			0.003	

Significance: *** - 0.001; ** - 0.01: * - 0.05; . - 0.1

Table 17: Regression results based on PCA for <u>short-term forecasts</u> with and without demographic variables.

	Past Behavior			Intended Future Behavior			Past Behavior			Intended Future Behavior		
	Coef	SE		Coef	SE		Coef	SE		Coef	SE	
Intercept	2.87	0.89	**	1.05	0.97		-0.10	0.15		-0.04	0.17	
PBC												
General	-0.06	0.06		-0.06	0.07		-0.03	0.07		0.02	0.07	
Info. Quality	0.04	0.13		0.02	0.14		0.03	0.13		0.06	0.15	
User Ability	**0.27**	**0.15**	.	-0.01	0.17		0.21	0.15		0.02	0.18	
Social Norms												
General	**0.12**	**0.07**	.	0.12	0.08		**0.17**	**0.07**	*	**0.19**	**0.09**	*
Officials vs. Public	-0.06	0.12		0.17	0.12		-0.10	0.12		0.02	0.14	
Data Attitude	**0.19**	**0.09**	*	**0.43**	**0.10**	***	**0.27**	**0.09**	**	**0.46**	**0.10**	***
Job Attitude												
General	-0.15	0.10		0.15	0.11		-0.16	0.11		0.08	0.12	
Mindset vs. Uncertainty	-0.05	0.13		0.02	0.14		-0.11	0.13		0.05	0.16	
Numeracy												
General	-0.06	0.07		**-0.18**	**0.08**	*	-0.04	0.07		-0.08	0.08	
Prefer. vs. Confidence	-0.10	0.12		-0.21	0.13		-0.11	0.12		**-0.27**	**0.14**	.
Hazard												
River Flood	0.01	0.44		**1.10**	**0.48**	*						
Tornado	-0.18	0.45		0.41	0.49							
Hurricane	-0.31	0.51		0.53	0.54							
Winter Storm	0.48	0.43		**1.14**	**0.47**	*						
Heat Wave												
Work Exp.												
1-5 years	**-3.41**	**0.92**	***	**-2.24**	**1.00**	*						
5-10 years	**-3.67**	**0.91**	***	**-2.5**	**0.99**	*						
10-20 years	**-3.03**	**0.94**	**	-1.6	1.03							
> 20 years	**-3.37**	**0.94**	***	-1.82	1.02	.						
EM Education Other	0.50	0.53		**1.92**	**0.57**	*						
Weather Instructions												
No	-0.21	0.46		-0.90	0.51	.						
Cannot remember	2.81	**0.93**	**	**2.23**	**1.02**	*						
Police												
Volunteer	-0.10	0.63		-0.82	0.8							
Retired	0.23	0.47		0.61	0.51							
Professional	**1.47**	**0.42**	***	**1.43**	**0.47**	**						
R^2 (Adj. R^2)	0.44 (0.30)			0.54 (0.43)			0.18 (0.11)			0.26 (0.20)		
P-Value	<0.001			<0.001			0.011			<0.001		

Significance		
***	-	0.001
**	-	0.01
*	-	0.05
.	-	0.10

4. Discussion

The most consistent predictor of how heavily an EM relies on recorded weather data and short-term weather forecasts is work experience. The strongest difference in data used exists between EMs having less than a year of work experience and those having more than that. Similarly, Baumgart et al. (2007) found that EMs with more work experience tend to make use of weather information more frequently.

Compared to work experience, the variables of the Theory of Planned Behavior (TPB) have a much weaker correlation to the reliance on weather information. Armitrage and Connor (2001) found in their analysis of 185 studies that TPB accounted on average for 21% of the variance in self-reported behavior and 39% in intention, but those studies did not include demographic variables. Applying TPB to reliance on recorded weather data and forecasts yielded lower values.[24] When trying to explain the frequency with which the general public obtains forecasts using slightly different variables, Demuth et al. (2011) were likewise able to only explain 22% of the variance. There are several reasons why our TPB model does not explain more of the variance.

Most importantly, the variables have been constructed by applying principal component analysis (PCA) to the original survey questions. PCA was necessary to reduce the number of variables, given the limited amount of data points. However, the choice of components cumulatively did not add up to much more than 50% of the variance for all TPB variables, except for perceived limitations (see Table 12). Consequently, a lot of the information inherent to the survey data would only be available with a much larger sample size.

Second, conventional linear regression might not be the best mathematical method. Generally, some question, whether survey answers on a Likert Scale can be treated as interval data in

[24] The intercorrelations between the variables are low as well when compared to the averages reported by Armitrage and Connor (2001) in their meta-analysis.

order to subject them to principal component analyses and linear regressions. Given that those survey answers are really ordinal data, correspondence analysis and logit regression should be used instead. In those analyses, each answer option is treated as a binary variable. Thus, a question with an answer scale from one to five becomes four variables instead of one. However, such a large number of variables cannot be accommodated by a dataset consisting of 207 participants. Therefore, in this case, the only choice is to revert to principal component analysis and conventional linear regression.

The limited amount of variance that the regression models are able to explain is most likely due to the highly variable nature of emergencies. As mentioned in the introduction, each emergency response is unique and the response highly dependent on the context. In a previous study, we found that often cascading events, such as collapsing infrastructure during a flood, are the most trying challenges for emergency responders (Study 1).

In the traditional configuration of the TBP model – without demographic variables – social norms and the attitude towards the weather information are the main predictors. In contrast, Armitrage & Connor (2001) report that in other TPB studies social norms are the weakest predictor of intentions, and thus behavior. They attribute this to poor measurement of the variable and advise a multi-item measurement. Additionally, including measures of attitude strength has been suggested to increase the predictive power of attitudes (Armitrage, Connor, 2001). Both was done for social norms in this study. We asked how much EMs worry about being criticized by social groups, and a variety of social groups were included. This could explain the relative success of social norm as a predictor in the regression models excluding demographic factors. The strong performance of social norms and attitude as predictors could also explain the weak performance of the perceived limitations (Armitrage, Connor, 2001). This seems indeed to be the case, because perceived limitations correlate with social norms and data attitude in our study.

An important limitation of this study is that we do not exactly know, which specific weather information EMs were thinking of when answering the survey. For example, does an EM facing a flood understand a "short-term weather forecast" to be the predictions by the NWS river forecast center or the radar? To make the survey relevant for as large a number of EMs as possible, it was necessary to include different hazards in the survey. While it is interesting to know, how forecast use varies across hazards, it was not possible to specifically ask for the weather information that are available for each hazard. The study would then have become too multi-dimensional. Therefore, only conclusions on a level higher than specific NWS products are possible.

5. Conclusion

Work experience was found to be the best predictor of whether an emergency manager relied on recorded weather data and short-term weather forecasts in the past or intends to do so in the future. Additionally, those products were relied on less for ice and snow storm than for other weather-related hazards. If work experience really is the main factor driving the reliance on recorded weather data and forecasts, emergency management training leaves needs to be improved significantly. Especially, since 84% of the participants had received instructions on how to use these products at some point.

Comparatively, the TPB variables weigh in significantly less. For the use of the recorded weather data and forecasts in the past, perceived limitations of those products – rather than the ability of the EM to understand and apply them – lead to less reliance on those types of data. For forecasts, the data attitude is the strongest TPB predictor for both past and future. For past behavior social norms, and for intentions for the future numeracy is slightly relevant. EMs who feel confident in dealing with uncertainty (job attitude) have stronger intentions to use recorded weather data in the future.

When demographic variables are excluded, as is done in most TPB studies (Cohen et al., 2009), social norms and data attitude become the two dominant predictors for recorded weather data and short-term weather forecasts for both past behavior and future intentions. It is possible that work experience results in learning to appreciate these types of weather information for decision-making and to listen to the people around the EM. Social norms and data attitude are correlated to the perceived limitations of the forecast, which is turn are correlated to subjective numeracy. However, it cannot be determined which direction the causality goes.

Turning to the TPB variables themselves, it was found that EMs rely little on recorded weather data and forecasts to determine when and where to deploy storm spotters and on recorded weather data when to initiate evacuations. Storm spotters are a way to gather more data early on. But apparently, it is not so much a worrying forecasts that informs this decision.

To use recorded weather data, the inability to apply information to the emergency response, the unavailability of the data for the area, and receiving the data too late are perceived to be most limiting. The former could be addressed by emphasizing the use of recorded weather data more in instruction sessions. For forecasts, the inaccuracy of and frequent changes to event timing and magnitude are perceived to be the greatest obstacles. This confirms that uncertainty limits the use of forecasts in practice.

Social pressure to rely on the studied weather information seems to come mainly from peers and those who possess local knowledge: EM colleagues, other city employees and NWS employees. Training and workshop instructors are least influential, probably because of the limited relevance to local circumstances. Therefore, localized training and peer education seem to be the most promising way to increase the value of weather information for decision-making. The EMs would appreciate it, because they have an overwhelmingly positive attitude towards those products. The fact that 75% were very or extremely satisfied indicates that EMs are unlikely to take initiative themselves.

One-third of the EMs thought that they tend to act intuitively rather than considering all facts and figures extensively. A majority of 65.7% prioritized protecting citizens. This distinguishes EMs from professions such as fire fighters and paramedics. But slightly more than half thought that it was not their task to prevent hazardous situations. This is not surprising since three-quarters of the participants have a background in another emergency response profession. Consequently, there could be much room for improvement to make the most use of the lead time that forecasts provide. Especially, since uncertainty was not considered much of a problem for planning an emergency response by about two-thirds of the participants (57-72%).

Compared to the participants in previous studies that included the subjective numeracy scale by Fagerlin et al. (2007), the EMs assessed their numeracy much higher. Generally, they feel more confident in working with percentages then with fractions. Additionally, they tend to prefer numbers over text, unless it is in a conversation with someone.

Reflecting on the Theory of Planned Behavior itself, a distinction between perceived behavioral control (i.e., external factors limiting a behavior) and self-efficacy (i.e., internal factors limiting a behavior) was found to be valuable.

***Acknowledgments*:** Most importantly, I would like to thank all the emergency managers that took the time to complete the survey. Special thanks to those, who send me additional comments. I also want to express my gratitude to the people who helped distributing the survey; most notably the International Association of Emergency Managers who published the survey call in its newsletter. Last, I would like to thank Alex Davis for patiently discussing the appropriate statistical methods with me.

6. References

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology*, *40*(4), 471-499.

Artikov, I., Hoffman, S. J., Lynne, G. D., Zillig, L. M. P., Hu, Q., Tomkins, A. J., ... & Waltman, W. (2006). Understanding the influence of climate forecasts on farmer decisions as planned behavior. *Journal of Applied Meteorology & Climatology*, *45*(9).

Ajzen, I. (2006): "Constructing a Theory of Planned Behavior Questionnaire", URL [accessed 04/22/2014]: http://people.umass.edu/aizen/pdf/tpb.measurement.pdf

Baumgart, L.A., Bass, E.J., Philips, B., Kloesel, K. (2007): "Emergency Management Decision Making during Severe Weather." *Weather and Forecasting* Vol. 23: pp.1268-1279.

Beyth-Marom, R., Austin, L., Fischhoff, B., Palmgren, C., & Jacobs-Quadrel, M. (1993). Perceived consequences of risky behaviors: Adults and adolescents. *Developmental psychology*, *29*(3), 549.

Cohen, T. R., Lindblad, M. R., Paik, J. G., & Quercia, R. G. (2009). Renting to owning: An exploration of the theory of planned behavior in the homeownership domain. *Basic and Applied Social Psychology*, *31*(4), 376-389.

Cooper, M. L., Shapiro, C. M., & Powers, A. M. (1998). Motivations for sex and risky sexual behavior among adolescents and young adults: a functional perspective. *Journal of personality and social psychology*, *75*(6), 1528.

Demuth, J. L., Lazo, J. K., & Morss, R. E. (2011). Exploring variations in people's sources, uses, and perceptions of weather forecasts. *Weather, Climate, and Society*, *3*(3), 177-192.

Fagerlin, A., Zikmund-Fisher, B.J., Ubel, P.A., Jankovic, A., Derry, H.A., & Smith, D.M. Measuring numeracy without a math test: Development of the Subjective Numeracy Scale (SNS). *Medical Decision Making*, 2007: 27: 672-680.

Fischer, H. W. (1996). What emergency management officials should know to enhance mitigation and effective disaster response. *Journal of Contingencies and Crisis Management*, *4*(4), 208-217.

Golden, J. H., & Adams, C. R. (2000). The tornado problem: Forecast, warning, and response. *Natural Hazards Review*, *1*(2), 107-118.

Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., & Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, *73*(3), 448-455.

Hess, R., Visschers, V. H., Siegrist, M., & Keller, C. (2011). How do people perceive graphical risk communication? The role of subjective numeracy. *Journal of Risk Research*, *14*(1), 47-61.

Hu, Q., Zillig, L. M. P., Lynne, G. D., Tomkins, A. J., Waltman, W. J., Hayes, M. J., ... & Wilhite, D. A. (2006). Understanding Farmers' Forecast Use from Their Beliefs, Values, Social Norms, and Perceived Obstacles. *Journal of Applied Meteorology & Climatology*, *45*(9).

Morss, R.E., Lazo, J.K., Demuth, J.L. (2010): "Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication." *Meteorological Applications* Vol. 17: pp. 149-162.

Morss, R.E. (2010): "Interaction among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases." *Natural Hazard Review,* August 2010: 83-96.

Morss R.E., Wahl, E. (2007): „An ethical analysis of hydrometeorological prediction and decision making: The case of the 1997 Red River flood." *Environmental Hazards* Vol. 7: pp. 342-352.

Morss, R. E., Wilhelmi, O. V., Downton, M. W., & Gruntfest, E. (2005). Flood risk, uncertainty, and scientific information for decision making: lessons from an interdisciplinary project. *Bulletin of the American Meteorological Society*, *86*(11), 1593-1601.

Moyano Díaz, E. (2002). Theory of planned behavior and pedestrians' intentions to violate traffic regulations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(3), 169-175.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, *5*(5), 411-419.

Pielke, R.A. (1999): "Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood", *Applied Behavioral Science Review*, 7(2), pp.83–101.

Pielke Jr, R. A., & Conant, R. T. (2003). Best practices in prediction for decision-making: lessons from the atmospheric and earth sciences. *Ecology*, *84*(6), 1351-1358.

Stewart, T. R., Pielke Jr, R., & Nath, R. (2004). Understanding user decision making and the value of improved precipitation forecasts: lessons from a case study. *Bulletin of the American Meteorological Society*, *85*(2), 223-235.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Weaver, John, Lindsey C. Harkabus, Jeffry Braun, Steven Miller, Rob Cox, John Griffith, Rebecca J. Mazur, 2014: An Overview of a Demographic Study of United States Emergency Managers. *Bull. Amer. Meteor. Soc.*, **95**, 199–203.

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, *27*(5), 663-671.

STUDY 3

Performance and Robustness of Probabilistic River Forecasts Computed with Quantile Regression based on Multiple Independent Variables in the North Central U.S.A.[25]

Abstract

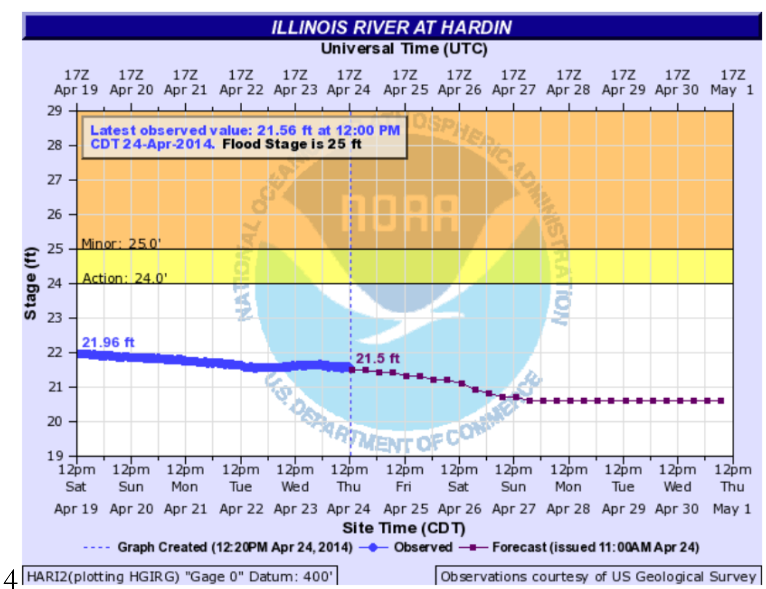
This study further develops the method of quantile regression (QR) to predict exceedance probabilities of flood stages by post-processing forecasts. Using data from the 82 river gages, for which the National Weather Service's North Central River Forecast Center issues forecasts daily, this is the first QR application to U.S. American river gages. Archived forecasts for lead times up to six days from 2001-2013 were analyzed. Earlier implementations of QR used the forecast itself as the only independent variable (Weerts et al., 2011; López et al., 2014). This study adds the rise rate of the river stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago to the QR model. Including those four variables significantly improved the forecasts, as measured by the Brier Skill Score (BSS). Mainly, the resolution increases, as the original QR implementation already delivered high reliability. Combining the forecast with the other four variables results in much less favorable BSSs. Lastly, the forecast performance does not depend on the size of the training dataset, but on the year, the river gage, lead time and event threshold that are being forecast. We find that each event threshold requires a separate model configuration or at least calibration.

Keywords: River forecasts, quantile regression, probabilistic forecasts, robustness

[25] In August 2014, a version of this chapter was submitted to Hydrology and Earth System Sciences, an Open Access journal.

1. Introduction

River-stage forecasts are inherently uncertain. The past has shown that unfortunate decisions have been made in ignorance of the potential forecast errors (e.g., Pielke, 1999; Morss, 2010). As of today, the National Weather Service does not routinely publish uncertainty information along with their short-term river-stage forecast (Figure 10). Only two long-term NWS river forecasts are probabilistic, i.e., quantify uncertainty: an exceedance curve for a period of three month and bar plots for each week of a three months period, see Figure 11 and Figure 12. These graphs can be used to determine with which probability each river stage will be exceeded in those weeks or three-months period. Although the short-term weather forecasts for the next few days are much used to prepare for flood events, they have remained deterministic,[26] as shown in Figure 10.[27]



4

Figure 10: Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014.

Source:http://water.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2.

[26] Two forecasts are published. The first (non-QPF) does not take any precipitation into account. The second (QPF) includes the precipitation forecast for the next 12 hours. This is the only way for the user to infer some sort of uncertainty.

[27] The deterministic forecasts are also available as text or tables.

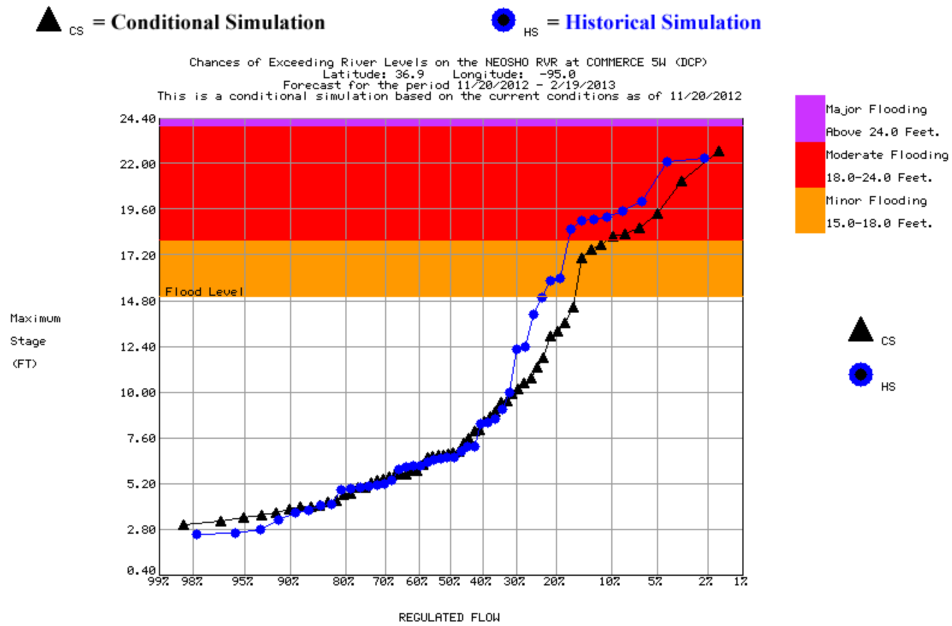


Figure 11: Probabilistic long-term forecast as published by the NWS for Commerce, OK on December 14th, 2012: Exceedance curve for three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2

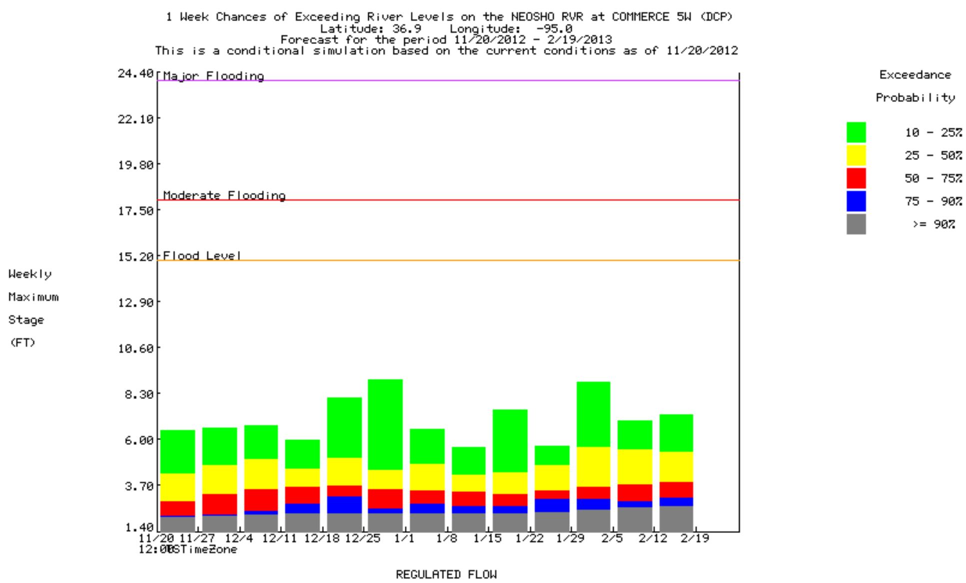


Figure 12: Probabilistic long-term forecast as published by the NWS for Commerce, OK on December 14th, 2012: Bar plot for each week of a three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2

NWS is currently developing and implementing ensemble forecasting to quantify some of the uncertainty of river-stage forecasts probabilistically (NWS, 2012; NOAA, 2001). The kinds of short-term probabilistic forecast products and visualizations that NWS envisions have not been made public yet.

This paper further develops the method by Weerts et al. (2011). An important difference is that we predict the probabilities that flood stages are exceeded rather than uncertainty bounds, because the former are more relevant to decision-making. In an attempt to balance missed alarms and false alarms, decision-makers are likely to resort to the best estimate (i.e., the deterministic forecast) rather than basing actions on the 75th or 90th confidence interval. Additionally, predicting the probability of an event corresponds with other forecasts with which users have much experience, e.g., the probability of precipitation. Morss et al. (2010) found in a survey of the general U.S. public that most people are able to base decisions on those forecasts.

Weerts et al. (2011) achieved impressive results in estimating the 50% and 90% confidence interval of river-stage forecasts in England and Wales using QR based on two years of archived river-stage forecasts. To our knowledge, this paper is the first application of this method to the U.S. American context. Additionally, we are fortunate to have much larger dataset, consisting of archived forecasts for 82 river gages covering 11 years available.

In this paper, the QR method is applied to the U.S. American context using the 82 river gages of the North Central River Forecast Center (NCRFC).[28] The method is further developed by demonstrating the benefit – measured by an increase in Brier Skill Score (BSS) – of including the rise rates of water levels in past hours and the past forecast errors as independent variables into the quantile regression. For extremely high water levels the variable combination has to be customized

[28] As of spring 2014, the NCRFC does not publish any sort of probabilistic forecasts.

for each river gage. For those, sets of few independent variables work best. Variable combinations for other event thresholds should include as many dependent variables as possible. Using the same combination for all of them works satisfactorily. Furthermore, it is found that the forecast – the only independent variable in the original QR method – is difficult to combine with the other dependent variables. Last, the method is shown to be robust to the size of the training dataset. However, the forecast performance does vary significantly across locations, lead times, water levels, and forecast year.

1.1. NWS River Forecasting

The National Weather Service (NWS) issues river-stage forecasts for ~4,000 river gages every day. In emergency management, which is the focus of this paper, usually short-term river-stage forecasts are used.[29] Such daily published forecasts predict the stage height in six-hour intervals for up to five days ahead (20 6-hour intervals).[30] When floods occur and increased information is needed, the local river forecast center (RFC) can decide to publish river-stage forecasts more frequently and for more locations.

The models used by the NWS to predict stage heights are numerical hydraulic models that simulate the river discharges in each watershed. Each river-stage forecast is the product of at least as many models as there are watersheds upstream; often in the hundreds. Welles et al. (2007) provides a detailed description of the forecasting process.

---

[29] Like in Grand Forks, ND long-term forecasts are only used for emergency management along rivers where the high discharges are dominated by snowmelt.

[30] The river-stage forecasts are produced by one of NWS' thirteen river forecasts centers (RFCs). Every morning the forecasts are forwarded to one of NWS's 122 local weather forecast offices (WFOs), who then disseminate the information to the public through a variety of media channels or by issuing warnings.

The published short-term weather forecasts are deterministic, see Figure 10. The only way that users can get a sense of the uncertainty is by comparing the quantitative precipitation forecast (QPF) with the non-QPF forecast. The QPF-forecast includes the precipitation predicted for the next 12 hours and zero precipitation for the forecasts beyond 12 hours[31]. The non-QPF forecast assumes no precipitation. Combined, these two forecasts give an idea of how much difference (a short period of) precipitation would make for the stage height in the river. The non-QPF serves as a reasonable lower bound; however, the QPF forecast is not an upper bound (i.e., precipitation could exceed the forecast values).

1.2. Uncertainty and Error in River-Stage Forecasts

The error framework for river-stage forecasts by Leahy et al. (2007) summarizes the major sources of error for river-stage forecasts, see Figure 13. Much of the uncertainty is exogenous to the models. The main model inputs are precipitation forecasts and observations. Those forecasts are significantly uncertain themselves, and only the forecast for the next 12 hours is taken into account. Additionally, the grid size of those prediction models does not match the watersheds, but is much larger. Another major source of uncertainty stems from the rating curves needed to translate the output of the river forecast models discharge to water level (Leahy et al., 2007). Especially in the case of high stage heights, those rating curves add much uncertainty. Sedimentation decreases the accuracy of the rating curves as well. The RFCs often receive only half-yearly rating curve updates from institutions like U.S. Geological Survey (USGS). Furthermore, Arkansas-Red Basin River Forecast Center (ABRFC) estimates that dam operators stick to their stated operations schedules only 65-70% of the time. This means that that RFC effectively lacks correct input data for their models one third of the

[31] This practice differs from RFC to RFC and also over time. For the ABRFC Welles et al. (2007) report: ~1993-1994: zero QPF; ~1995-2000 24hr QPF for first 24hrs, zero QPF beyond 24hrs; ~2001-2003 12hr QPF for first 12hrs, zero QPF beyond 12hrs.

time. Finally, high water levels only occur infrequently. The part of the model used for high water levels has thus been built on much fewer data points than the one for low stage heights. With rising stage heights the models are thus increasingly "flying blind," resulting in increasing uncertainty.

Ensemble forecasting, being developed and implemented by the NWS, considers the uncertainty in the input data. Post-processing the data, as is done in this paper, takes into account all sources of uncertainty.

For users, forecasts are most important in extreme situations, such as droughts and floods. Due to their infrequency and the subsequent scarcity of data, forecasts have larger errors where accuracy has the most value. Additionally, users might only experience such an event once or twice in their lifetime, so that they have no experience to what extent they can rely on deterministic forecasts in such situations. Given the many sources and complexity of uncertainty and the lacking user experience, it is easy to see how forecast users find it difficult to estimate the forecast error.

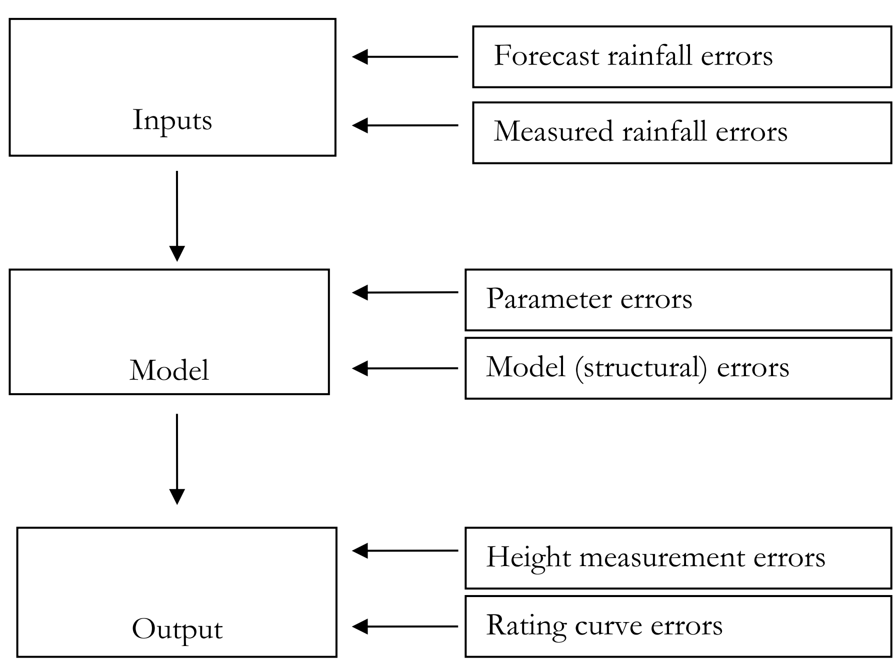


Figure 13: Error framework for river-stage forecasting by Leahy et al. (2007).

2. Method

The use of quantile regression to quantify the error distribution of river-stage forecasts has first been presented by Weerts et al. (2011) for river catchments in the England and Wales. In this paper, we use substantially more data and further develop Weerts' original method in three ways: a) by including additional variables instead of using only the forecast itself as an independent variable; b) by testing the robustness of the method across locations, lead times, event thresholds, forecast years, and the size of training dataset; c) by estimating the more decision-relevant probability of exceeding flood stages rather than confidence bounds. To develop the different configurations of quantile regression and to compare their performance, the Brier Skill Score (BSS) is used.

In the following, the quantile regression itself, the proposed addition to the method, and the undertaken computations are explained.

2.1. Quantile Regression

In the context of river forecasts, linear quantile regression has been used to estimate the distribution of forecast errors as a function of the forecast itself. Weerts et al. (2011) summarize this stochastic approach as follows:

> *"[It] estimates effective uncertainty due to all uncertainty sources. The approach is implemented as a post-processor on a deterministic forecast. [It] estimates the probability distribution of the forecast error at different lead times, by conditioning the forecast error on the predicted value itself. Once this distribution is known, it can be efficiently imposed on forecast values."*

Quantile Regression was first introduced by Koenker (2005; 1978). It is different from ordinary least square regression in that it predicts percentiles rather than the mean of a dataset. Koenker and Machado (1999, p. 1305) and Alexander et al. (2011) demonstrate that studying the coefficients and their uncertainty for different percentiles generates new insights, especially for nonnormally distributed data. For example, using quantile regression to analyze the drivers of international economic growths, Koenker and Machado (1999) find that benefits of improving the terms of trade show a monotonously increasing trend across percentiles, thus benefitting faster-growing countries proportionally more.

In its original application to river forecasts by Weerts et al. (2011) the forecast values and the corresponding forecast errors are transformed into the Gaussian domain using Normal Quantile Transformation (NQT), as instructed by Bogner et al. (2011), to account for heteroscedasticity. Building on this study, López et al. (2014) compare different configurations of QR with the forecast as the only independent variable, including configurations omitting NQT. They find that no configuration was consistently superior for a range of forecast quality metrics (López et al., 2014). To be able to combine variables of different nature, we build a model based on untransformed variables. The reason to do so will be discussed and illustrated later (see Figure 21 and Figure 22).

Using the transformed data, a quantile regression is run for each lead time and desired percentile with the forecast error as the dependent variable and the forecast and other variables as the independent variables.[32] To prevent the quantile regression lines from crossing each other, a fixed effects model is implemented below a certain forecast value. Weerts et al. (2011) give a detailed mathematical description for applying QR to river forecasts. Mathematically, the approach is formulated as follows:

Equation 2: Original QR implementation with NQT, with percentiles of the forecast error as the dependent variable and the only independent variable being the forecast itself, bot transformed into the normal domain.

$$F_\tau(t) = f(t) + NQT^{-1}[a_\tau * V_{NQT}(t) + b_\tau]$$

[32] As mentioned in Weerts et al. (2011), our quantile regression models have likewise a higher predictive capacity, if the forecast error rather than the forecast itself is used as the dependent variable.

Equation 3: QR implementation without NQT, with percentiles of the forecast error as the dependent variable and multiple independent variables.

$$F_\tau(t) = f(t) + \sum_{i}^{I} a_{i,\tau} * V_i(t) + b_\tau$$

with $F_\tau(t)$ – estimated forecast associated with percentile τ and time t
$f(t)$ – original forecast at time t
$V_i(t)$ – the independent variable i (e.g., the original forecast) at time t
$V_{i;NQT}(t)$ – the independent variable I transformed by NQT at time t
$a_{i,\tau}$, b_τ – model coefficients

The second part of the equations stands for the error estimate based on the quantile regression model for each percentile τ and lead time. In Equation 2, that was used in the original QR method proposed by Weerts et al. (2011), this estimation was executed in the Gaussian domain using only the forecast as independent variable.[33]

2.2. Brier Skill Score

The original QR implementation by Weerts et al. (2011) was evaluated by determining the fraction of observations that fell into the confidence intervals predicted by the QR model; i.e., ideally, 90% of the observations should be larger than the predicted 10th percentile for that day, and smaller than the predicted 90th percentile. López et al. (2014) used a number of metrics to assess model performance, e.g., the Brier Skill Score (BSS), the mean continuous ranked probability (skill) score (RPSS), the relative operating characteristic (ROC), and reliability diagrams to compare QR configurations.

We use the Brier Skill Score to compare the different versions of the QR model proposed in this paper. We chose to optimize our QR models based on the BSS, first introduced by Brier (1950)

[33] All quantile regressions were done using the command *rq()* in the R-package "quantreg" (Koenker, 2013).

for two reasons. First, for decision-making the probability with which a certain water level, e.g., a flood stage, is exceeded is more useful than confidence intervals. Second, the Brier Score can be decomposed into two different measures of forecast quality (see Equation 4): Reliability and resolution. The third component is uncertainty, which is a hydrological characteristic inherent to the river gage. Thus, it is not subject to the forecast quality. Equation 4 gives the definition of the (de-composed) Brier Score (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009).[34]

Equation 4: Brier Score; de-composed into three terms: reliability, resolution and uncertainty.

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}) = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

with	BS	– Brier Score
	N	– number of forecasts
	K	– the number of bins for forecast probability of binary event occurring on each day
	n_k	– the number of forecasts falling into each bin
	$\bar{o}_k$	– the frequency of binary event occurring on days in which forecast falls into bin k
	f_k	– forecast probability
	$\bar{o}$	– frequency of binary event occurring
	f_t	– forecast probability at time t
	o_t	– observed event at time t (binary: 0 – event did not happen, 1 – event happened)

The Brier Score pertains to binary events, e.g., the exceedance of a certain river stage or flood stage. Reliability compares the estimated probability of such an event with its actual frequency. For example, perfect reliability means that on 60% of all days for which it was predicted that the

[34] Bröcker (2012) showed that the conventional decomposition of the Brier Score is biased for finite sample sizes. It systematically overestimates reliability, under- or overestimates resolution, and underestimates uncertainty. Several authors proposed less biased decompositions (e.g., Bröcker, 2012; Ferro and Fricker, 2012). Additionally, Stephenson et al. (2008) proved that the Brier Score has two additional components when it is computed based on bins, as is usually done. Nonetheless, we chose to stick to the conventional decomposition and using bins, as implemented in the R-package "verification" (NCAR-Research Applications Laboratory, 2014; Wilks, 1995) to ensure that our results can be readily compared to other studies like López et al. (2014). After all, the Score is mainly used to compare model configurations, rather than establishing the absolute performance of each model.

water level would exceed flood stage with a 60% probability, it actually does so. A forecast with perfect reliability would follow the diagonal in Figure 14, i.e., the area in Figure 14a representing reliability would equal zero (e.g., Jolliffe, Stephenson, 2012; "Brier Score," 2014; WWRP/WGNE, 2009). The configuration by López et al. (2014) performs well in terms of reliability. When estimating confidence intervals, Weerts et al. (2011) achieved good results especially for the more extreme percentiles (i.e., 10th and 90th).

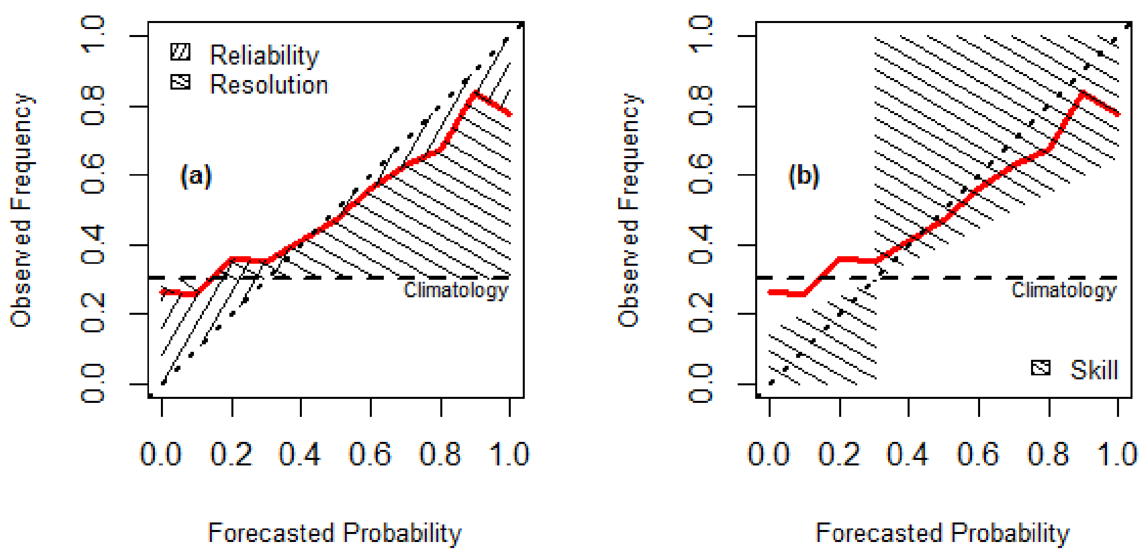


Figure 14: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill (BSS > 0), i.e. performs better than random guessing, if it is inside the shaded area in the figure b. Ideally, the forecast would follow the diagonal (BSS=1). (Adapted from Wilson, n.d.; and Hsu, Murphy, 1986).

Resolution pertains to how much better the forecast performs than taking the historical frequency (climatology) as a forecast. For example, for a gage where flood stage is exceeded on 5% of the days in a year, simply using the historical frequency as the forecast would mean forecasting that the probability of the water level exceeding flood stage is 5% on any given day (e.g., Jolliffe, Stephenson, 2012; "Brier Score," 2014; WWRP/WGNE, 2009). In Figure 14, a forecast with good resolution would be steeper than the dashed line that represents climatology, i.e., the area in Figure 14a representing resolution would be maximized. In absolute terms, the resolution can never exceed the third term in Equation 3 representing the uncertainty inherent to the river gage. Through the resolution component, the Brier Score is related to the area under the relative operating characteristic (ROC) curve (for more detail, see Ideka et al., 2002). The latter likewise quantifies how much better a forecast is than random guessing in detecting a binary event; though unlike the Brier Score it focuses on the ratios of false and missed alarms (e.g., Jolliffe, Stephenson, 2012; "Brier Score," 2014; WWRP/WGNE, 2009).

A forecast possesses skill, i.e., performs better than random guessing or climatology, if it is inside the shaded area in Figure 14b. The Brier *Skill* Score (BSS) equals the Brier Score normalized by climatology to make the score comparable across gages with different frequencies of a binary event.[35] The BSS can range from minus infinity to one. A BSS below zero indicates no skill; the perfect score is one (e.g. Jolliffe, Stephenson, 2012; "Brier Score," 2014; WWRP/WGNE, 2009).

2.3. QR with more than one variable

Intuitively, more information should lead to better prediction of the distribution of the forecast error, because the regression models would be based on more data. The most obvious variables to include besides the forecast itself are the observed water level 24 and 48 hours ago, the observed rise in water level in the last 24 and 48 hours (called rise rate hereafter), the forecast error 24 and 48 hours ago, or the time of the year, e.g., month or season. Other potential variables are the water levels observed up- and downstream at various times, the precipitation upstream of the catchment

[35] All measures of forecast quality were computed using the R-package "verification" (NCAR, 2014).

area, and the precipitation forecast. However, these latter variables are much more difficult to gather because of the way data is archived at the National Climatic Data Center.[36]

Table 18: Variable Combinations

Combi	fcst	err24	err48	rr24	rr48
1	●				
2		●			
3			●		
4				●	
5					●
6	●	●			
7	●		●		
8	●			●	
9	●				●
10		●	●		
11		●		●	
12		●			●
13			●	●	
14			●		●
15				●	●
16	●	●	●		
17	●	●		●	
18	●	●			●
19	●		●	●	
20	●		●		●
21	●			●	●
22		●	●	●	
23		●	●		●
24		●		●	●
25			●	●	●
26	●	●	●	●	
27	●	●	●		●
28	●	●		●	●
29	●		●	●	●
30		●	●	●	●
31	●	●	●	●	●

fcst = forecast; rr24, rr48 = rise rate in the past 24 and 48 hours;
err24, err 48 = forecast error 24 and 48 hours ago

In preliminary trials on two case studies (gages HARI2 and HYNI2), it was found that season and months are not significant in quantile regression models to predict the quantiles of the forecast error. It was also found that the rise rates and the forecast errors are better predictors than the water levels observed in previous days. After all, the observed water levels are used to compute

[36] For the North Central River Forecast Center (NCRFC), the river forecast and the observed water levels are saved in the same text product available at [accessed 6/4/2014]: http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelect?datasetname=9957ANX. (Station ID: KMSR, Bulletin ID: FGUS5). Requesting the corresponding precipitation and precipitation forecast requires an extensive effort or direct access to the database.

the rise rates and forecast errors, so that these latter variables include the information of the former variable.

To determine which set of variables preforms best in generating probabilistic forecasts, all 31 possible combinations of the forecast (fcst), the rise rate in the last 24 and 48 hours (rr24, rr48), and the forecast error 24 and 48 hours ago (err24, err48) were tested for 82 gages that the NCRFC issues forecasts for every morning (Table 18). Based on the Bier Skill Score, a metric of forecast quality explained below, it was determined which variable combination on average and most often leads to the best out-of-sample results for various lead times and water levels.

2.4. Computations

For each river gage and lead time, one QR model was trained on the first half of the dataset for each the following quantiles: [0.05, 0.1, 0.15, … , 0.85, 0.90, 0.95]. For the forecast published on each day in the verification dataset, i.e., the second half of the dataset, the water levels corresponding with these quantiles were predicted. Effectively, for each day, a probability distribution of water levels to expect is computed. Using linear interpolation,[37] the probability with which various water levels, called event thresholds hereafter, will be exceeded was then calculated. After having repeated this procedure to determine the probabilities of exceeding those event thresholds for each day in the verification dataset, the BSS for each threshold was computed.[38]

Analyses have been done for two sets of thresholds. First, the computation was done for the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile of observed water levels to study whether the various

[37] Using the command *approx(x, y, xout, yleft=1,yright=0,ties=mean)* in the R-package "stats" (R Core Team, 2014).

[38] In this study, an event was defined as the river water level exceeding a certain water level. This is different from the conventional definition of an event, as the probability that the water level will be between for example the moderate and the major flood stage. The more binary definition of exceeding or not exceeding a certain water level seemed to be more relevant to real-life decision-making.

combinations of variables perform equally well for high and low thresholds. Second, the analysis was repeated for the four flood stages (action stage, and minor, moderate, and major flood stage) of each gage, because those play an important role in flood response planning.

This procedure is repeated for the 31 variable combinations (Table 18). The result is 31 BSSs for 82 river gages for four different lead times (one to four days) and for different event thresholds (i.e., flood stages or percentiles of the observed water level). This set-up allows a thorough analysis of which sets of variables perform best under the various circumstances.

2.5. Data

All forecasts made by the North Central River Forecast Center (NCRFC) between May 1st, 2001 and December 31st, 2013 were available for analysis. In total, the NCRFC produces forecasts for 525 gages (Figure 15). For 82 of those gages, forecasts have been published daily for a sufficient number of years, and are not inflow forecasts. The latter have been excluded from the forecast error analysis because they forecast discharge rather than water level. About half of the analyzed gages are along the Mississippi River. The Illinois River and the Des Moines River are two other prominent rivers in the region. The drainage areas of the 82 river gages average 61,500 square miles (minimum 200 sq.miles; maximum 708,600 sq.miles).

Two river gages serve as an illustration for the points made throughout this paper. Hardin, IL is just upstream the confluence of the Illinois River and the Mississippi River (Figure 15). Therefore, it probably experiences high water levels through backwatering, when the high water levels in the Mississippi River prevent the Illinois River from draining. Henry, IL is located ~200 miles (~320 km) upstream of Hardin, having a difference in elevation of ~25 feet (~7.6 m). The

Illinois River is ~330 miles (~530 km) long,[39] draining an area of ~13,500 square miles (~35,000 km^2) at Henry[40] and ~28,700 square miles (~72,000 km^2) at Hardin.[41]

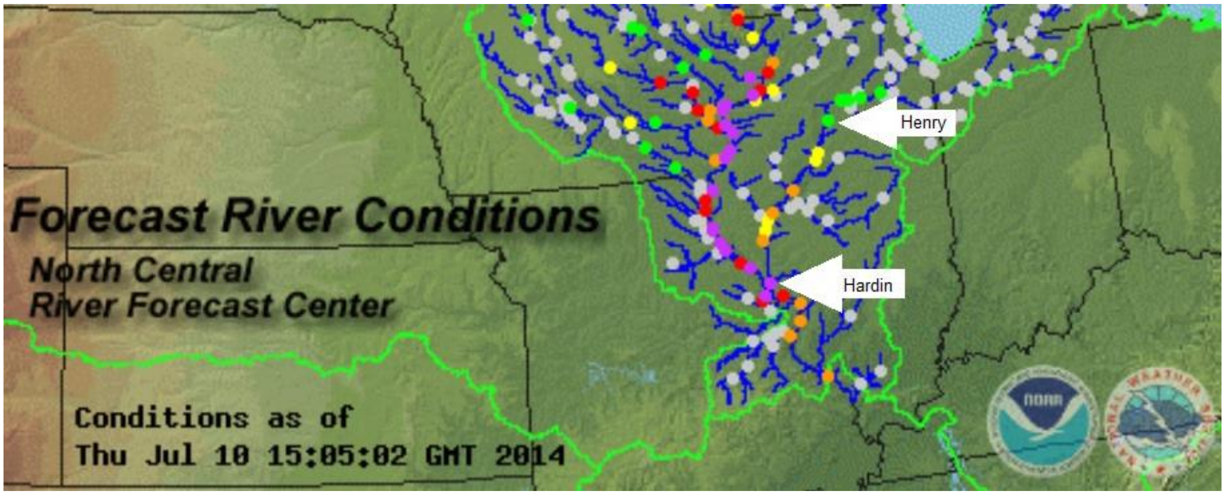


Figure 15: Portion of the North Central River Forecast Centers river gages with Henry (HYNI2) and Hardin (HARI2) indicated by the upper and lower red arrow respectively. Source: http://www.crh.noaa.gov/ncrfc/

3. Results

3.1. Forecast error at NCRFC's gages

In general, the NCRFC's forecasts are well calibrated across the entire dataset. The average error, defined as observation minus the forecast, is zero for most gages. For lead times longer than three days, a slight underestimation by the forecast is noticeable. By a lead time of 6 days this underestimation averages 0.41 feet only (Figure 16a, Table 19a). Extremely low water levels, defined as below the 10th percentile of observed water levels, are also well calibrated (Figure 16b, Table 19b).

[39] Illinois Environmental Protection Agency: "Illinois River and Lakes Fact Sheets", URL [accessed 04/24/2014]: http://dnr.state.il.us/education/aquatic/aquaticillinoisrivlakefactshts.pdf
[40] Source: http://waterdata.usgs.gov/nwis/nwisman/?site_no=05558300&agency_cd=USGS
[41] Source: http://waterdata.usgs.gov/nwis/nwisman/?site_no=05587060&agency_cd=USGS

However, when considering higher water levels the picture changes.[42] The underestimation becomes more pronounced, averaging 0.29 feet for three days of lead time and 1.14 feet for six days of lead time, when only observations exceeding the 90th percentile of all observations are considered (Figure 16c, Table 19c).

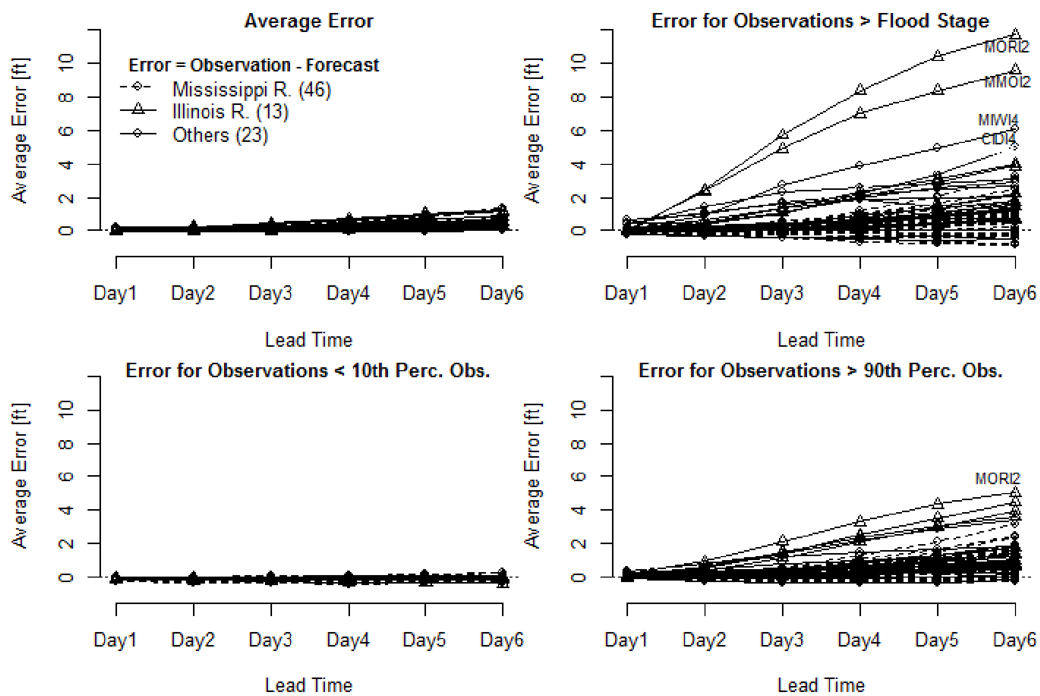


Figure 16: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the 90th percentile of observations; (d) error on days that the water level exceeded minor flood stage.

[42] The gages MORI2 and MMOI2 are upstream of a dam. It is likely that the forecasts performed so poorly there, because the dam operators deviated from the schedules that they provide the river forecast centers to base their calculations on.

Table 19: Error statistics for the forecast error a) of the whole dataset; b) on days that the water level did not exceed the 10th percentile of observations; c) on days that the water level exceeded the 90th percentile of observations; d) on days that the water level exceeded minor flood stage.

Average errors	Lead Time					
of 82 gages	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**
a) ALL OBSERVATIONS						
Minimum	-0.21	-0.08	-0.09	-0.07	-0.04	0.02
Median	0.01	0.02	0.06	0.13	0.22	0.30
Mean	**0.01**	**0.04**	**0.10**	**0.18**	**0.30**	**0.41**
Maximum	0.19	0.21	0.76	1.65	2.62	3.47
b) OBSERVATIONS < 10th PERCENTILE						
Minimum	-1.2	-0.35	-0.38	-0.41	-0.38	-0.39
Median	-0.03	-0.04	-0.05	-0.05	-0.04	-0.04
Mean	**-0.06**	**-0.06**	**-0.06**	**-0.06**	**-0.05**	**-0.04**
Maximum	0.03	0.04	0.05	0.12	0.17	0.25
c) OBSERVATIONS > 90th PERCENTILE						
Minimum	-0.11	-0.23	-0.31	-0.38	-0.38	-0.27
Median	-0.01	0.02	0.15	0.32	0.55	0.81
Mean	**0.01**	**0.09**	**0.29**	**0.55**	**0.82**	**1.14**
Maximum	0.34	1.01	3.12	5.13	6.81	8.56
d) OBSERVATIONS > FLOOD STAGE						
Minimum	-0.20	-0.30	-0.44	-0.63	-0.78	-0.80
Median	-0.02	-0.03	0.22	0.45	0.78	1.10
Mean	**0.01**	**0.17**	**0.45**	**0.80**	**1.14**	**1.51**
Maximum	0.65	2.44	5.70	8.37	10.40	11.74

When only looking at observations that exceeded the minor flood stages corresponding to each gage,[43] the underestimation averages 0.45 feet for three days of lead time and 1.51 feet for 6

[43] Flood stages are based on the damage done by previous floods. It depends on the context, e.g., the shape of the river bed and the development of the river shores, which water levels cause damage. Therefore, it depends on the river gage which percentiles of observed water levels the flood stages correspond with.

days of lead time (Figure 16d, Table 19d). However, some gages, such as Morris (MORI2), Marseilles Lock/Dam (MMOI2) – both on the Illinois River – and Marshall Town on the Iowa River (MIWI4) experience *average* errors of 5 to 12 feet for water levels higher than minor flood stage.

3.2. Including more variables

In total, the Brier Skill Score (BSS) for 31 variable combinations (Table 18) across various lead times and event threshold have been compared. Across 82 river gages, it has been analyzed (a) which combinations perform best and worst most often, and (b) which sets of variables deliver the best BSSs on average.

3.2.1. Frequency Analysis

For each lead time (i.e., one to four days) and various event thresholds (i.e., 10^{th}, 25^{th}, 75^{th}, 90^{th} percentiles as well as the four flood stages), we counted how often each variable combination resulted in the highest and the lowest BSS across the 82 river gages. Figure 17 shows that for water levels below the 50^{th} percentile variable combinations with four or more variables return the best BSSs most often, while those with one and two variables perform worst most often. For thresholds higher than the 50^{th} percentile the distributions gradually become more flat. For the 90^{th} percentile, a clear trend is no longer detectable. The same set of histograms for the four flood stages (i.e., action, minor, moderate, and major) confirms this (Figure 18). Across lead times, there is a slight trend noticeable that single variables tend to be the worst combination more often for longer lead times. Thus, the further out one is forecasting, the more important it becomes to include more data in the model.

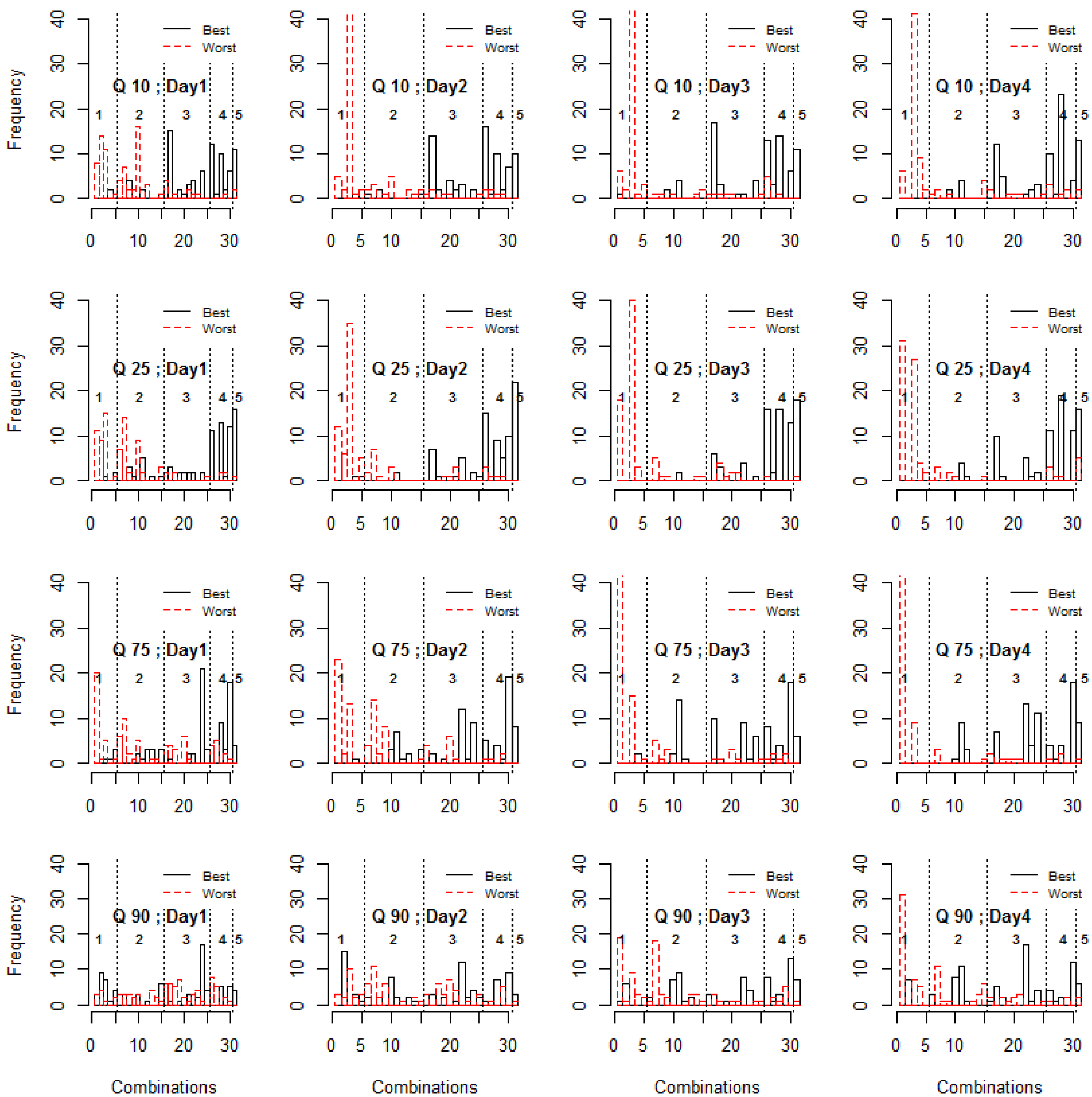


Figure 17: Histograms of variable combinations returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to an event threshold defined as a percentile of the observed water levels, and each column to a lead time. The dotted vertical lines in the histograms distinguish variable combinations with different numbers of variables.

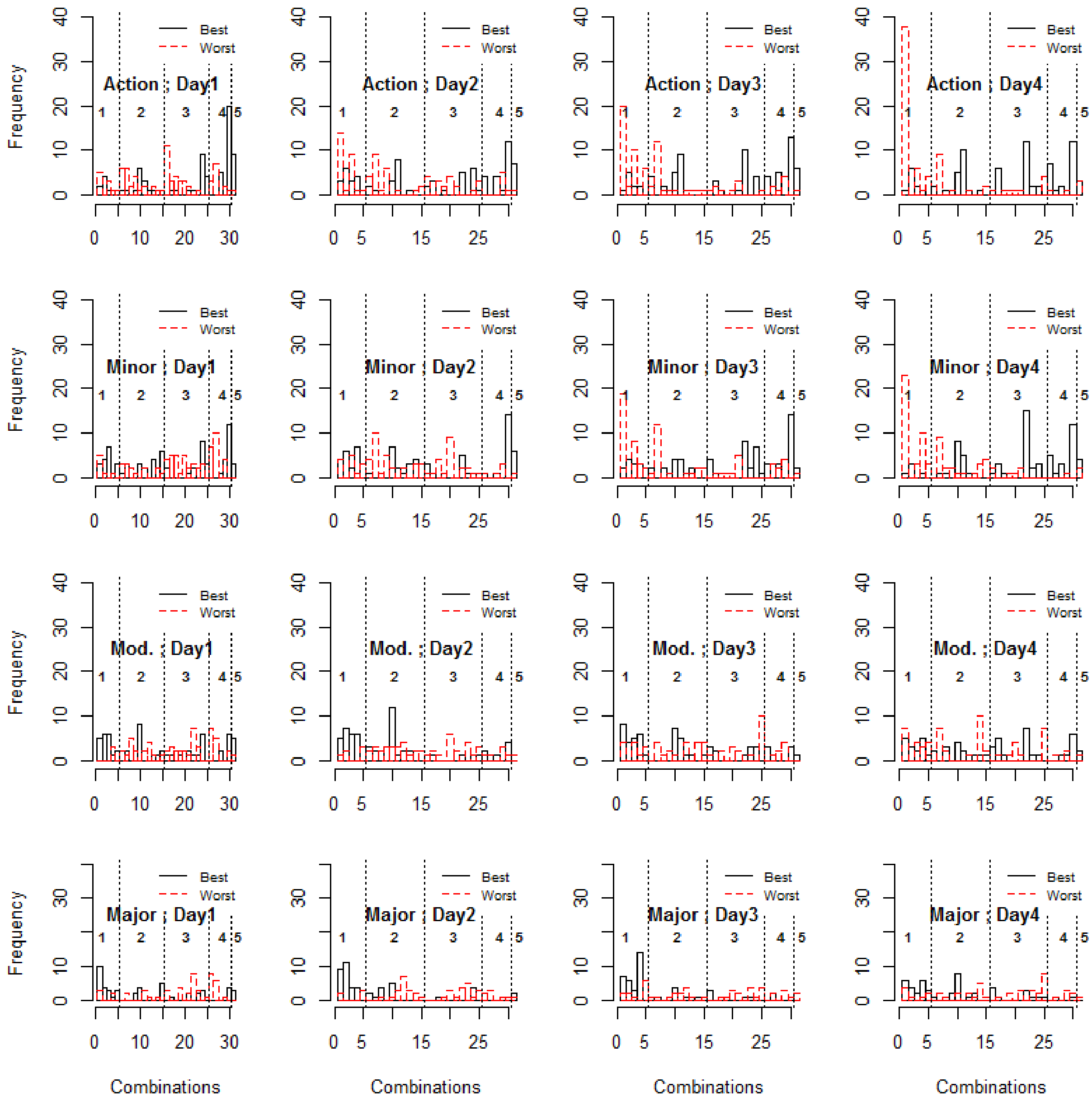


Figure 18: Histograms of variable combinations returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to a flood stage, and each column to a lead time. The dotted vertical lines in the histograms distinguish variable combinations with different numbers of variables.

3.2.2. Best performing combinations on average

For each river gage, the combinations have been ranked by BSSs. It was found that the more variables are included in a set, the higher that set of variables will rank on average (Figure 19). However, for extremely high water levels, this trend gradually reverses (Figure 20). For action stage[44] and minor flood stage,[45] a slightly increasing trend is still visible. For moderate[46] and major flood stage,[47] combinations with fewer variables rank higher on average.

Considering these findings and those of the frequency analysis earlier, the models for the various river gages can generally be based on the same variable combinations of four or more variables. But for extremely high water levels, a model specific to each river gage has to be built in order to achieve high BSSs.

The combinations including the forecast (indicated by gray vertical lines in Figure 19 and Figure 20) perform less well than those that exclude it. Plotting the independent variables against the forecast error as the dependent variable makes the reason visible (Figure 21, Figure 22). Without a transformation into the normal domain, the forecast does not provide a lot of information for the QR model. In contrast, the other four variables do not lend themselves for linear quantile regression after performing NQT. Further research is necessary to reconcile these two types of variables. A possible solution could be to build QR models for subsets of the transformed dependent and independent variable.

---

[44] Across the 82 stations, action stage corresponds with water levels between the 60th and 100th percentile.

[45] Across the 82 stations, minor flood stage corresponds with water levels between the 70th and 100th percentile.

[46] Across the 82 stations, moderate flood stage corresponds with water levels between the 80th and 100th percentile.

[47] Across the 82 stations, major flood stage corresponds with water levels between the 90th and 100th percentile.

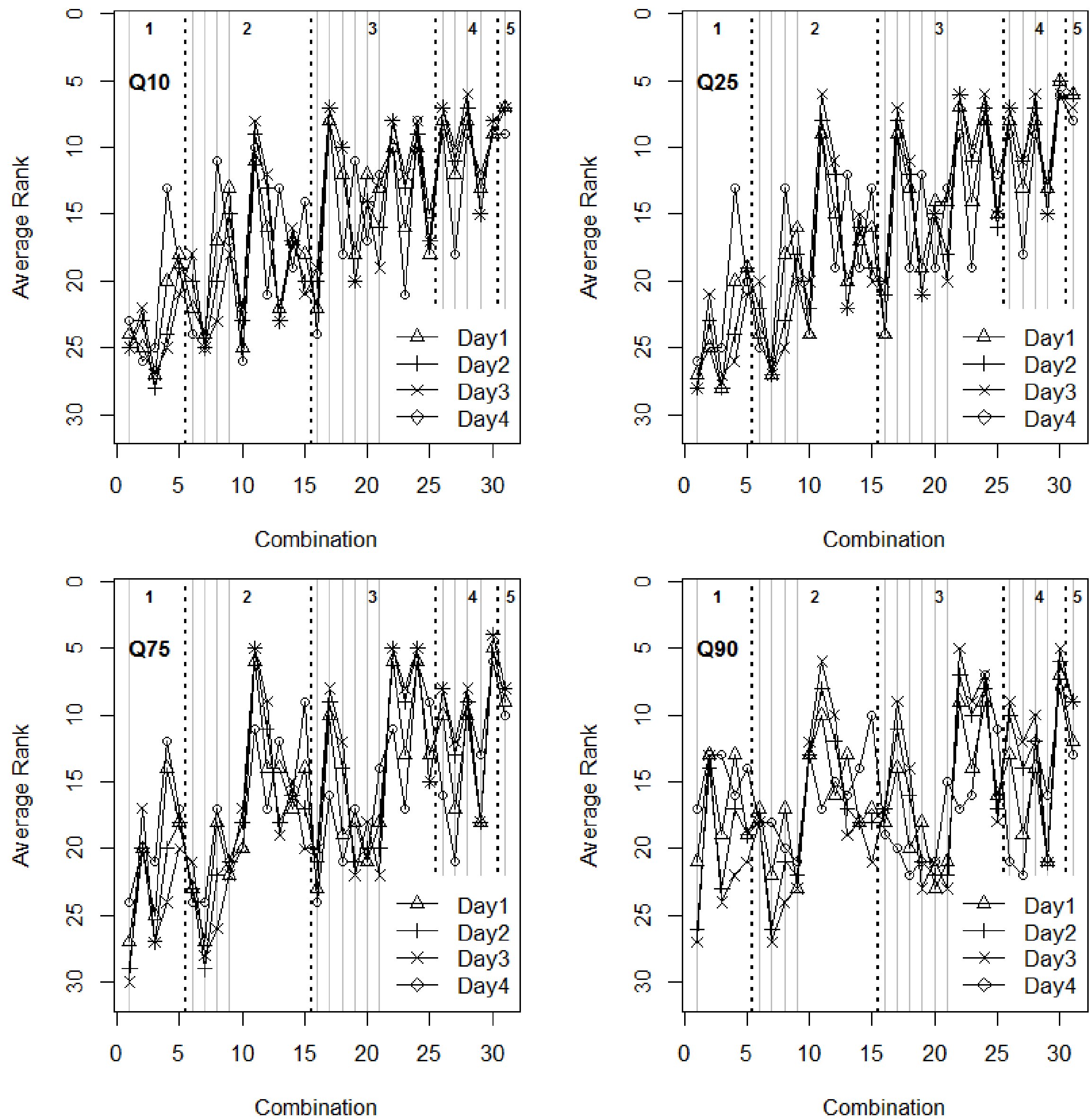


Figure 19: Average rank for each variable combination for one to four days of lead time and four percentiles of observed water levels. Vertical gray lines indicate variable combinations including the forecast.

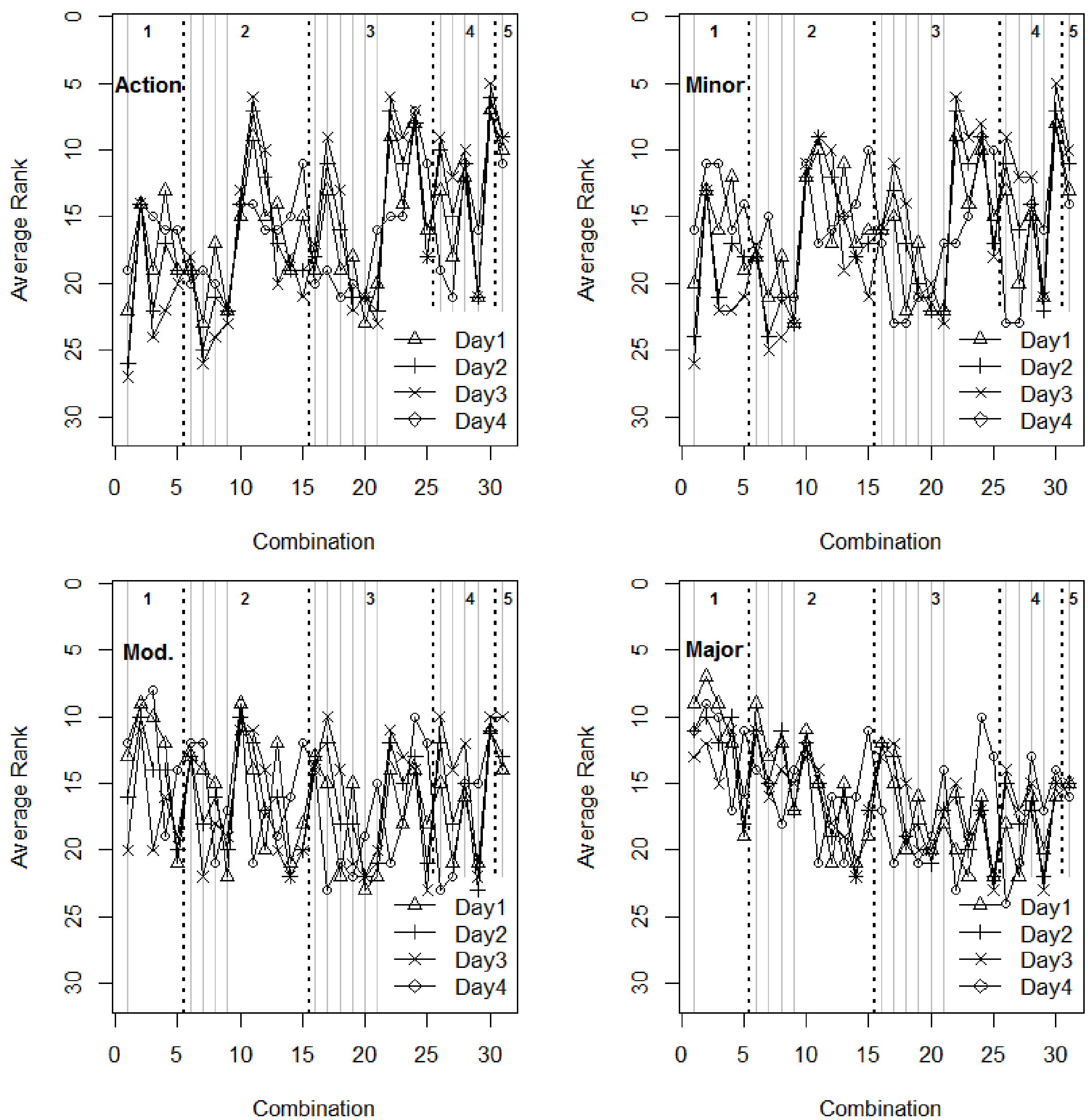


Figure 20: Average rank for each variable combination for one to four days of lead time and four flood stages. Vertical gray lines indicate variable combinations including the forecast.

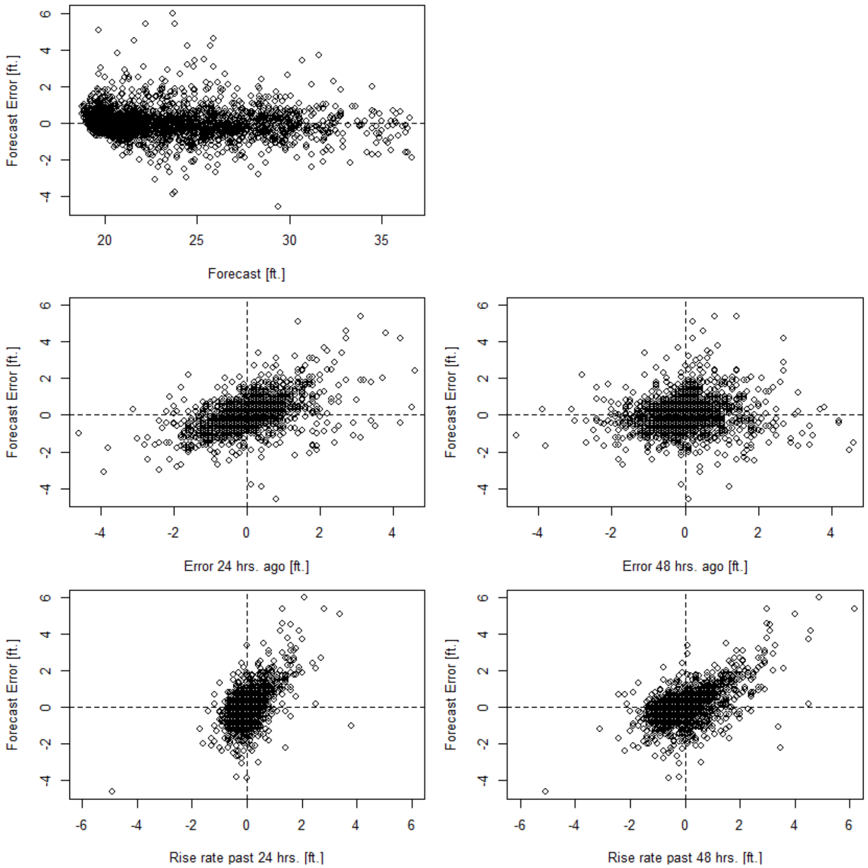


Figure 21: Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rise rates.

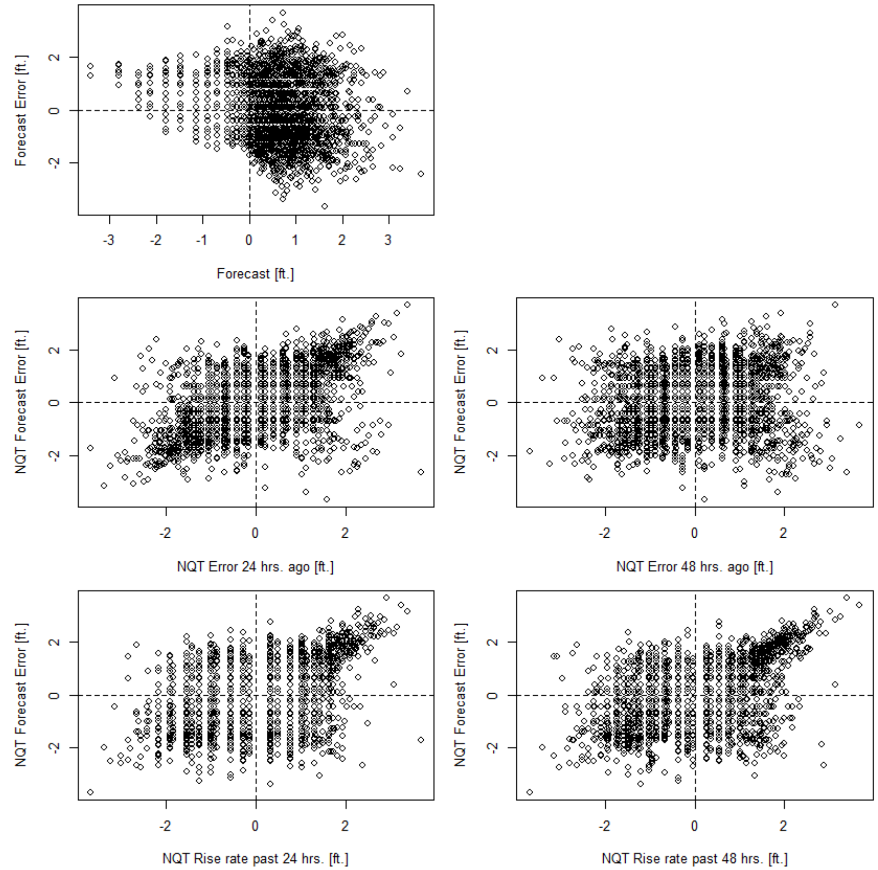


Figure 22: Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rise rates.

3.2.3. Brier Skill Score

Including the rise rate and forecasts errors as independent variables into the QR model improves the Brier Skill Score (BSS) significantly. Figure 23 illustrates the BSS when using the model as originally introduced by Weerts et al. (2011). Using the best performing variable combination instead, gives an upper bound of the BSSs that can be achieved at best. This configuration increases the mean and decreases the standard deviation (Table 20, Figure 24). The performance improves most where all model configurations perform worst: at the 10^{th} percentile.[48] The decrease of the BSSs with lead time also becomes considerably less with this configuration. Additionally, an one-size-fits-all approach was tested to investigate, whether customizing the QR model to each river gage would be worth it. In this configuration, the rise rates in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago serve as the independent variables (combination 30). It was found that this approach returns only slightly worse results than working with the best performing configuration for each river gage (Figure 25, Table 20). Accordingly, the same variable combination can be used for all river gages.

As shown in Figure 18, this last conclusion is not true for extremely high water levels. Including more variables does improve the BSSs considerably (Figure 26 and Figure 27; Table 20). However, for each river gage the best combination of variables needs to be identified separately. Because data to build models is scarce for extreme levels, the QR models all perform less well for each increase in flood stage.

[48] Possibly, the models do not perform well for low percentiles, because the dependent variable – the forecast error – exhibits very little variance at those water levels, i.e., the average error is very small (Table 19).

Table 20: Mean and standard deviation three QR configurations: the original using the transformed forecast only as independent variable; the best performing combination for each river gage (upper performance limit); rise rates in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution).

	Q10	Q25	Q75	Q90	Q10	Q25	Q75	Q90
	Day 1				**Day 2**			
NQT-fcst	0.34 (0.52)	0.65 (0.36)	0.90 (0.07)	0.88 (0.08)	0.24 (0.57)	0.59 (0.35)	0.85 (0.10)	0.82 (0.12)
Best combi.s	0.54 (0.34)	0.78 (0.18)	0.93 (0.05)	0.91 (0.06)	0.49 (0.36)	0.74 (0.19)	0.90 (0.05)	0.87 (0.07)
Rise rate 24/48 +error 24/48*	0.49 (0.41)	0.77 (0.18)	0.92 (0.05)	0.93 (0.06)	0.42 (0.44)	0.73 (0.19)	0.90 (0.06)	0.86 (0.09)
	Day 3				**Day 4**			
NQT-fcst	0.20 (0.61)	0.56 (0.33)	0.81 (0.10)	0.75 (0.15)	0.19 (0.55)	0.55 (0.31)	0.77 (0.13)	0.69 (0.18)
Best combi.s	0.47 (0.37)	0.74 (0.17)	0.89 (0.05)	0.85 (0.09)	0.46 (0.37)	0.73 (0.18)	0.89 (0.05)	0.84 (0.09)
Rise rate 24/48 +error 24/48*	0.40 (0.44)	0.72 (0.19)	0.88 (0.06)	0.84 (0.11)	0.39 (0.43)	0.71 (0.20)	0.88 (0.05)	0.82 (0.20)
	Action	**Minor**	**Moderate**	**Major**	**Action**	**Minor**	**Moderate**	**Major**
	Day 1				**Day 2**			
NQT-fcst	0.81 (0.27)	0.42 (1.12)	0.38 (1.02)	-0.80 (2.07)	0.68 (0.59)	0.41 (0.90)	0.25 (1.2)	-1.30 (1.96)
Best combi.s	0.86 (0.26)	0.78 (0.27)	0.73 (0.24)	0.36 (0.66)	0.82 (0.29)	0.73 (0.28)	0.68 (0.24)	0.26 (0.67)
	Day 3				**Day 4**			
NQT-fcst	0.67 (0.37)	0.37 (0.87)	-0.09 (1.42)	-1.69 (2.24)	0.62 (0.35)	0.22 (1.00)	-0.07 (1.05)	-1.52 (1.96)
Best combi.s	0.81 (0.26)	0.71 (0.31)	0.64 (0.23)	0.19 (0.76)	0.79 (0.26)	0.69 (0.30)	0.60 (0.23)	0.13 (0.72)

* Combination 30

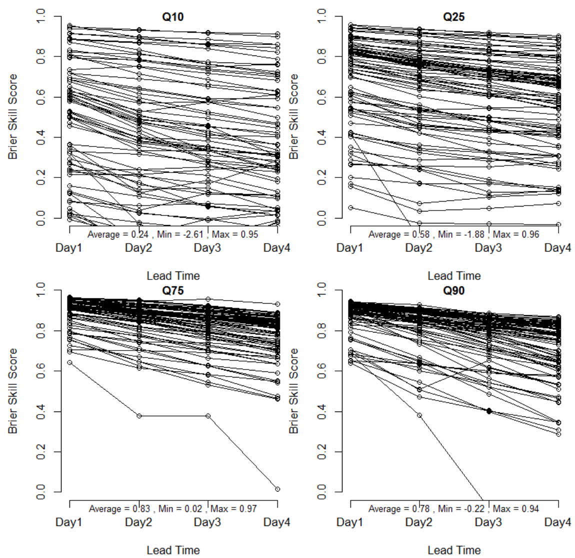


Figure 23: Brier Skill Scores of the original QR model (i.e., using the transformed forecast as the only independent variable) for four lead times and percentiles of observed water levels.

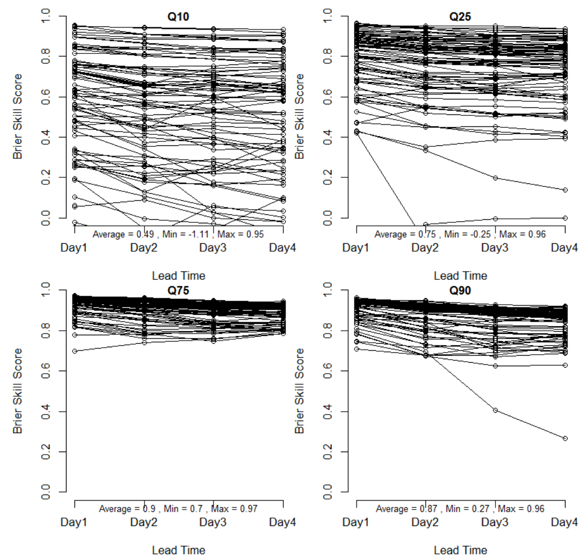


Figure 24: Brier Skill Scores for four lead times and percentiles of observed water levels using the best variable combination for each river gage as independent variables in the QR model.

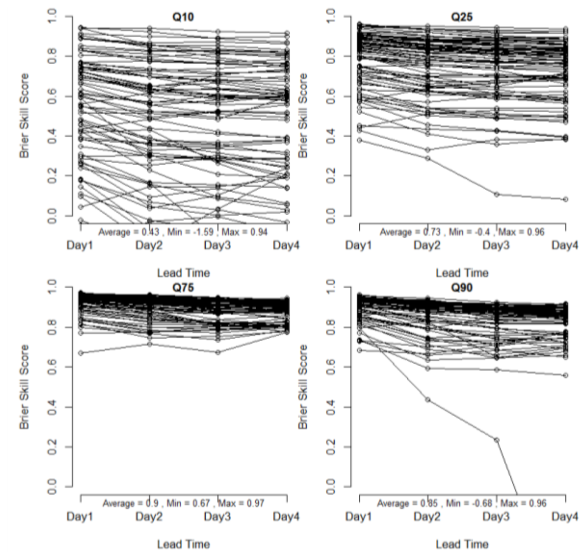
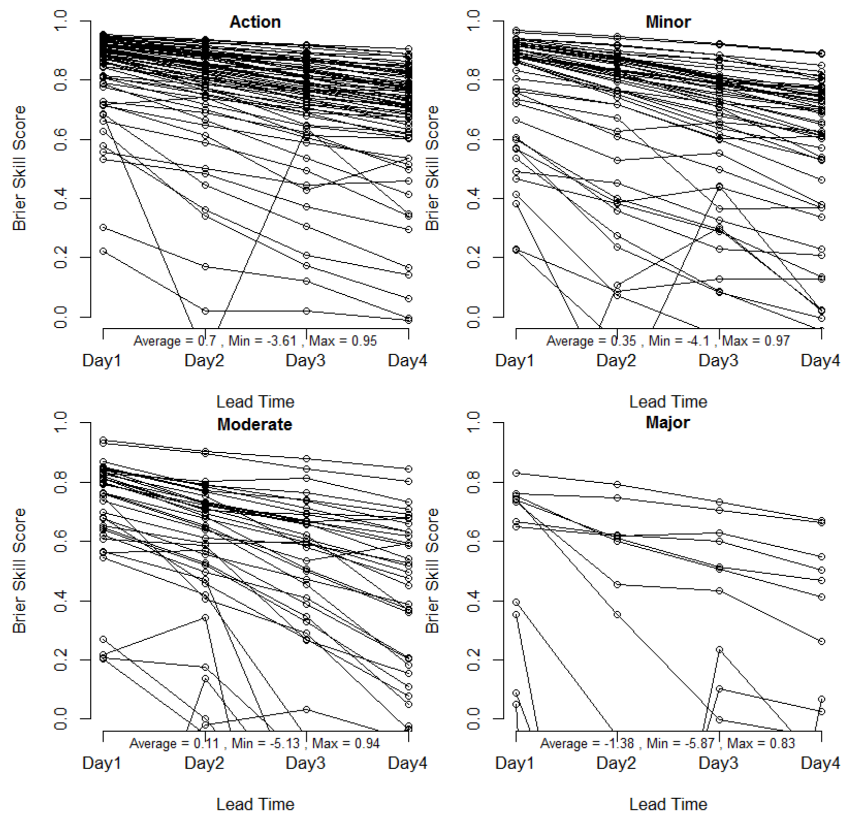


Figure 25: Brier Skill Scores for four lead times and percentiles of observed water levels using a one-size-fits-all approach (i.e., rr24, rr48, err24, err48) for the independent variables in the QR model.



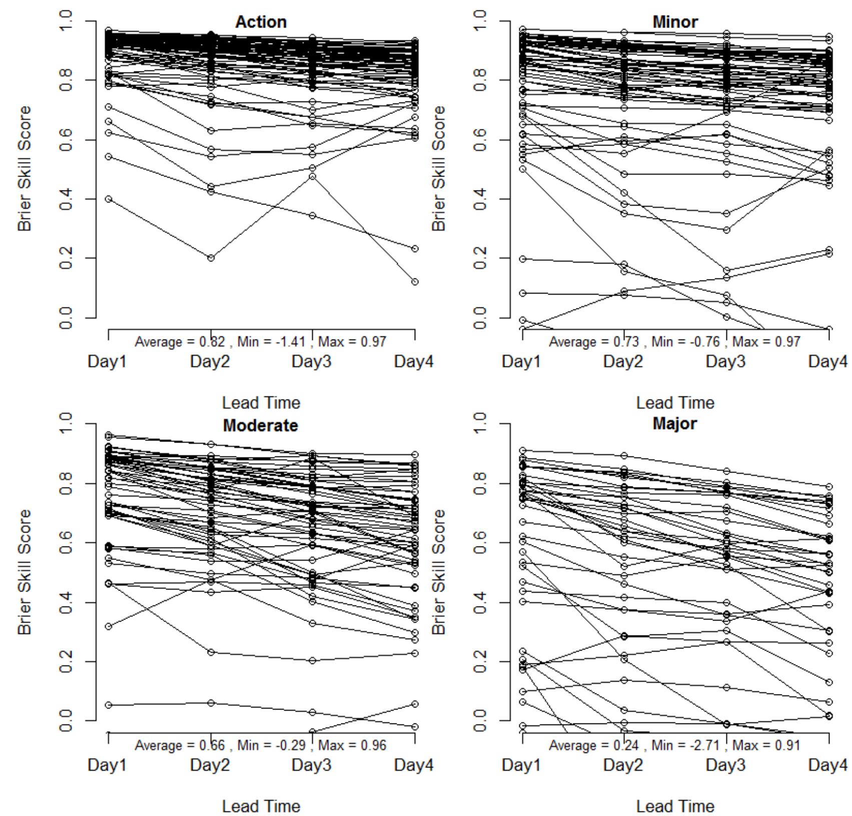


Figure 26: Brier Skill Scores of the original QR model (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

Figure 27: Brier Skill Scores for four lead times and flood stages of observed water levels using the best variable combination for each river gage as independent variables in the QR model.

The fact that the Brier Score can be de-composed into reliability, resolution and uncertainty allows a closer look at which improvements are being achieved by including more variables. Figure 28 shows that the original QR model configuration by Weerts et al. (2011) has high reliability (i.e., the reliability is close to zero). The Brier Score and the Brier Skill Score mainly improve when using rise rates and forecast errors as independent variables, because the resolution increases. The forecast quality improves along other dimensions as well, i.e., the areas under the ROC curves and the ranked probability skill score (RPSS) increase. The first weighs missed alarms against false alarms and has a perfect score equal to one. The latter is a version of the Brier Skill Score. While the Brier Skill Score pertains to a binary event, the RPSS can take into account various event categories. Its perfect score equals one (e.g., WWRP/WGNE, 2009).

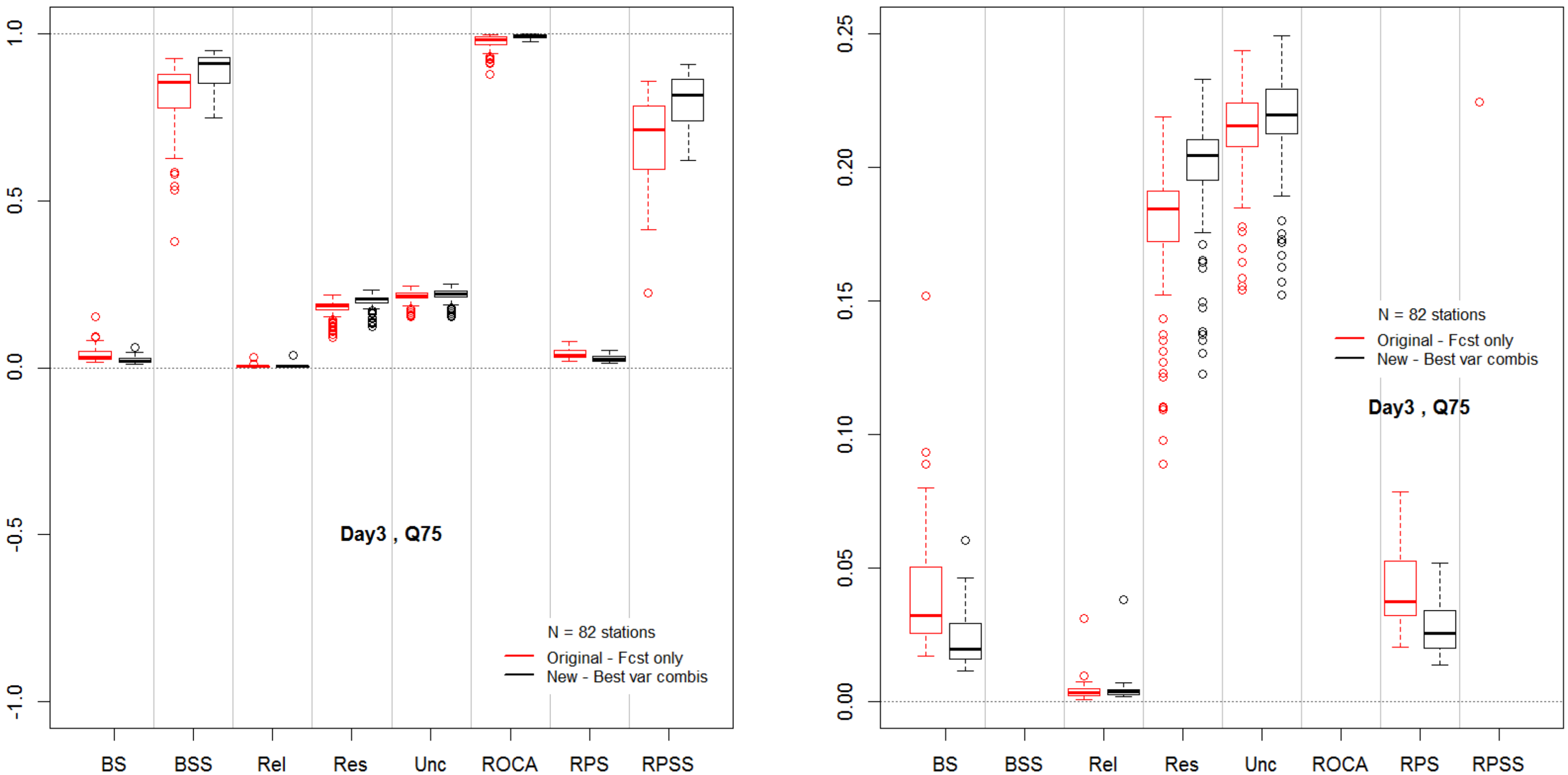


Figure 28: Comparison of the original QR model (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., rise rates and forecast errors as independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), Uncertainty (Unc), Area under the ROC curve (ROCA), ranked probability score (RPS), ranked probability skill score (RPSS). Lead time: 3 days; 75th percentile of observation levels as threshold. The left figure zooms in on the right figure to make changes in reliability and resolution better visible.

3.3. Robustness

The impact of the length of the training dataset on the model's performance measured by the Brier Skill Score (BSS) was assessed for the one-size-fits-all QR model (i.e., rise rates and forecast errors as independent variables for all gages) for Hardin and Henry on the Illinois River. Each year between 2003 and 2013 was forecast by models trained on one year up to however many years of archived forecasts were available. Figure 29 and Figure 30 show that for those gages, it does not matter for the BSS how many years are included in the training dataset. That is good news, if stationarity cannot be assumed (Milly et al., 2007), a step-change in river regime has occurred, or forecast data have not been archived in the past. In those cases, only short training datasets are available. However, the BSS varies considerably for what year is being forecast. The forecast performance varies greatly, especially for the 10th and 25th percentile of observed water levels. It is likely, that a very large dataset, including more infrequent events, would improve these results. However, most river forecast centers only recently started archiving forecasts in a text-format, so that even having ten years' worth of data is an exception.[49]

[49] To illustrate that point, the National Climatic Data Center has archived data from 2001 onwards available in their HDSS Access System.

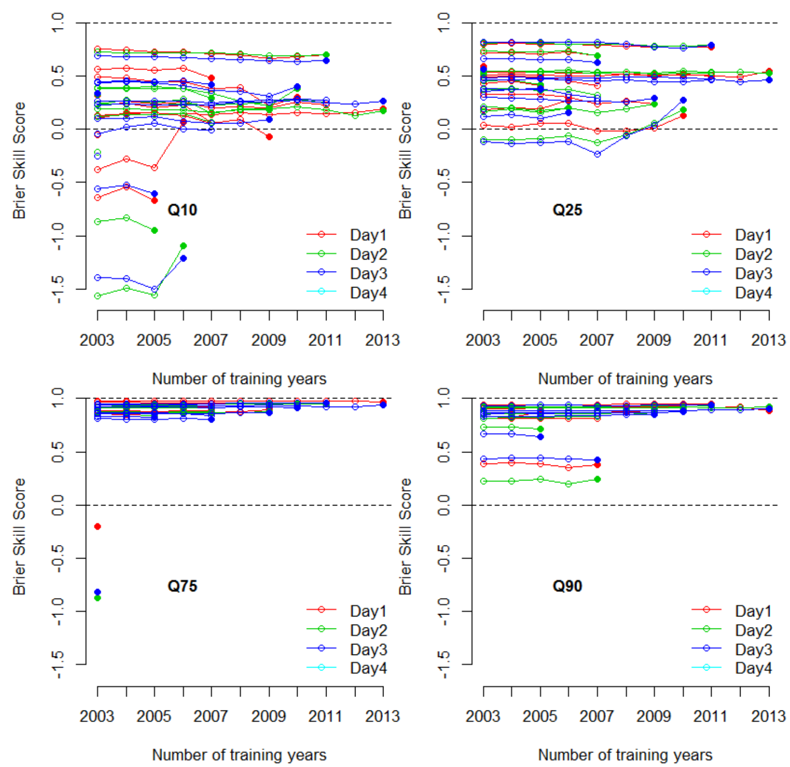


Figure 29: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

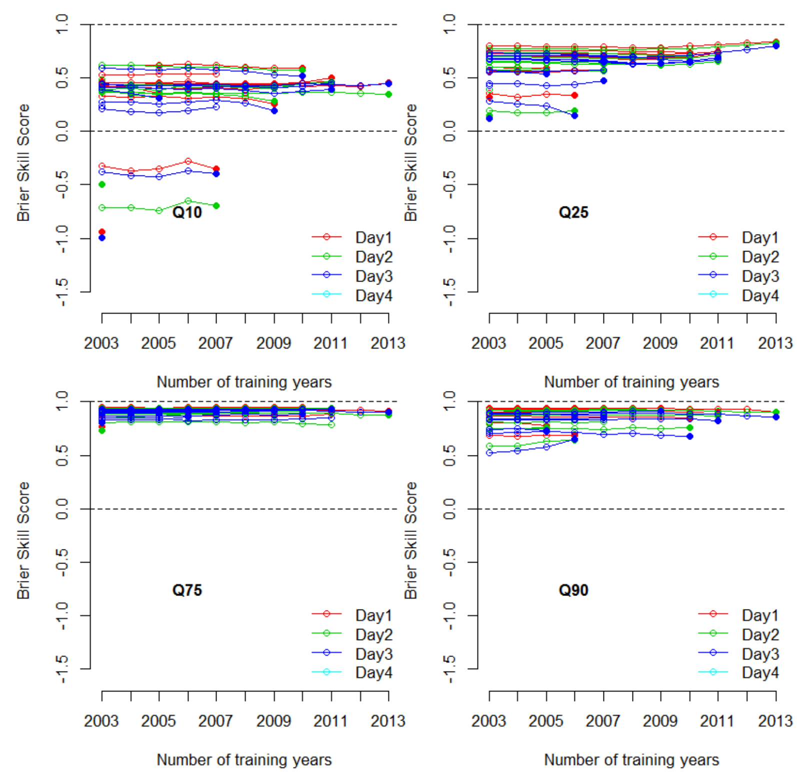


Figure 30: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

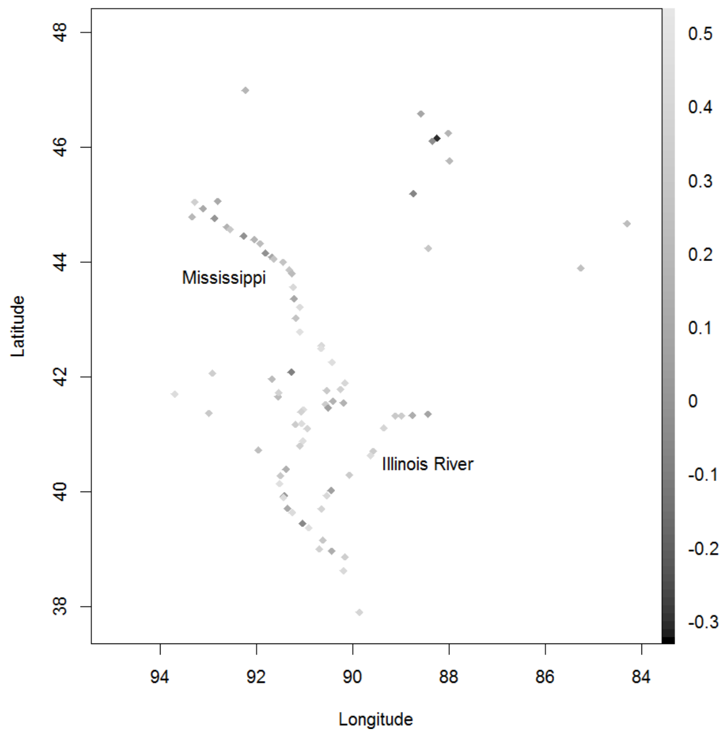


Figure 31: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

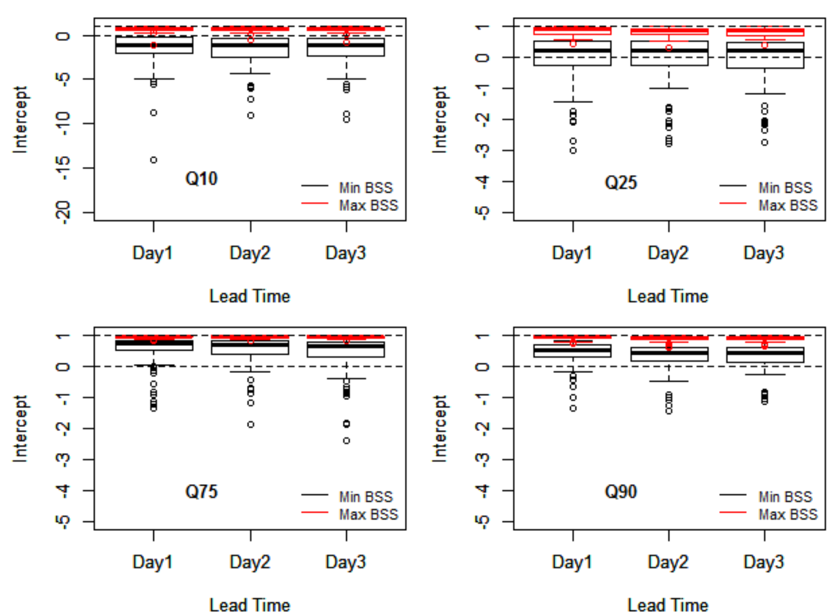


Figure 32: Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.

4. Conclusion

In this study, quantile regression (QR) has been applied to estimate the probability of the river water level exceeding various event thresholds (i.e., 10^{th}, 25^{th}, 75^{th}, 90^{th} percentiles of observed water levels as well as the four flood stages of each river gage). This is the first study applying this method to the U.S. American context. Additionally, it further develops the method by including more independent variables and testing the method's robustness across locations, lead times, event thresholds, forecast years and sizes of training dataset.

Most importantly, it was found that including rise rates in the past 24 and 48 hours and the forecast errors of 24 and 48 hours ago as independent variables improves the performance of the QR model, as measured by the Brier Skill Score. Since the reliability was already high, the original QR method as proposed by Weerts et al. (2011), the new configuration mainly increases the resolution.

For extremely high water levels, the combinations of independent variables that perform best vary across stations. On those days, combinations of fewer variables perform better than those that include more. In contrast to these extremely high event thresholds, larger sets of variables work better than smaller ones for non-extreme and low event thresholds. Additionally, a one-size-fits-all approach (i.e. the rise rates and forecasts errors as independent variables) performs satisfactorily for those cases.

The new independent variables – rise rates and forecast errors – do not combine well with forecast itself. The latter was the only variable included in the original QR configuration as studied by Weerts et al. (2011) and López et al. (2014). To account for heteroscedasticity, the forecast was transformed into the Gaussian domain. However, the rise rates and the forecast errors do not lend themselves for linear quantile regression after such a transformation. Therefore, it is difficult to combine these two variables. A possible solution could be to build regression models for subsets of the transformed data. However, such an approach drastically decreases the amount of data available for each model.

The proposed QR method is robust to the size of training dataset, which is convenient if stationarity cannot be assumed (Milly et al., 2007), a step-change in the river regime has occurred, or – as is the case for most river forecast centers – only recent forecast data have been archived. However, the performance of the method does depend on the river gage, the lead time, event threshold and year that are being forecast. This results in a very wide range of Brier Skill Scores. For the user, it is particularly difficult to know, how much to trust a forecast, if the performance depends so much on context. Likewise, this is case for the original QR configuration.

As measured by the Brier Skill Score, the year 2012 was a relatively easy year to forecast, the year 2008 a particularly difficult one. The proposed approach performs less well for longer lead times, for gages far upstream a river or close to confluences, for low event thresholds and extremely high ones. The model might be performing less well for low event thresholds, because the variance in the dependent variable – the forecast error – is smaller. After all, river forecasts have much smaller errors for lower water levels. In turn, for extremely high water levels, the scarcity of data decreases the model performance.

Future Work

The methods can be further developed in several ways to achieve higher Brier Skill Scores and more robustness. First, more independent variables can be added. Trials with a different method, classification trees, showed that the observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels significantly improve models. Presumably, this is the case, because the QPF-forecast includes the precipitation forecast only for the next 12 hours. However, currently, the precipitation data and forecasts can only be requested in chunks of a month, three chunks per day, from the National Climatic Data Center's (NCDC) HDSS Access System.[50] For a period of 12 years, requesting such data for several weather stations[51] is obviously time-consuming. Upstream water levels can easily be included after manually determining the upstream gage(s) for each of the 82 NCRFC gages. To improve model performance at gages close to river confluences, the upstream water level of the gages on the joining river should be included as well.

Different approaches of sub-setting the data to improve models results also warrant consideration. Particularly, clustering the data by variability seems promising. However, early trials indicated that this method is very sensitive to the training dataset.

As mentioned above, the QR method works less well for low than for high event thresholds. Further study should investigate, why that is the case, and identify possible solutions. The current study focused on extremely high event thresholds, i.e., flood stages, but not on lower ones, i.e., below the 50th percentile of observed water levels.

Last, the proposed method would need to be verified for gages for which the NCRFC does not publish daily forecasts. Ignorance of the uncertainty inherent in river forecasts have had some of the most unfortunate impacts on decision-making in Grand Forks, ND and Fargo, ND (Pielke, 1999; Morss, 2010). Both of those stages are discontinuously forecast NCRFC gages.

Acknowledgements: Many thanks to Grant Weller who suggested looking into quantile regression to predict forecast errors. As to funding, Frauke Hoss is supported by an ERP fellowship of the

[50] URL [accessed July 2014]:
http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelect?datasetname=9957ANX

[51] The geographical units of the weather forecasts bulletins do not correspond with those of the river forecast bulletins.

German National Academic Foundation and by the Center of Climate and Energy Decision Making (SES-0949710), through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (CMU).

5. References

Alexander, M., Harding, M., Lamarche, C. (2011): "Quantile Regression for Time-Series-Cross-Section Data." *International Journal of Statistics and Management System,* 4(1-2), pp. 47-72.

Brier Score. (n.d.). In *Wikipedia*. Retrieved July, 2014, from http://en.wikipedia.org/wiki/Brier_score

Bogner, K., Pappenberger, F., Cloke, H. L., & de Michele, C. (2012). Technical Note: The normal quantile transformation and its application in a flood forecasting system. *Hydrology & Earth System Sciences*, *16*(4).

Bröcker, J. (2012). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, *39*(3-4), 655-667.

Ferro, C. A., & Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, *138*(668), 1954-1960.

Hsu, W. R., & Murphy, A. H. (1986). The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, *2*(3), 285-293.

Ikeda, M., Ishigaki, T., & Yamauchi, K. (2002). Relationship between Brier score and area under the binormal ROC curve. *Computer methods and programs in biomedicine*, *67*(3), 187-194.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Koenker, R. (2013). *quantreg: Quantile Regression.* R package version 5.05, http://CRAN.R-project.org/package=quantreg.

Koenker, R. (2005). *Quantile regression* (No. 38). Cambridge university press.

Koenker, R., & Machado, J. A. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, *94*(448), 1296-1310.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Leahy, C.P. et al., 2007. Objective Assessment and Communication of Uncertainty in Flood Warnings. In 5th Flood Management Conference Warrnambool.

López, P., Verkade, J. S., Weerts, A. H., & Solomatine, D. P. (2014). Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. *Hydrology and Earth System Sciences Discussions*, *11*(4), 3811-3855.

Milly, P. C. D., Julio, B., Malin, F., Robert, M., Zbigniew, W., Dennis, P., & Ronald, J. (2007). Stationarity is dead. *Ground Water News & Views*, *4*(1), 6-8.

Morss, R.E. (2010): "Interactions among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases." *Natural Hazards Review*, 11(3), p.83.

Morss, R.E., Lazo, J.K., Demuth, J.L. (2010): "Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication." *Meteorological Applications* Vol. 17: pp. 149-162.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*(4), 595-600.

National Oceanic and Atmospheric Administration (2001): *Implementation of Advanced Hydrologic Prediction Service*. National Weather Service, August 2001. URL [accessed 4/25/2014]: http://www.nws.noaa.gov/oh/ahps/Implementation_Plan.pdf

NCAR – Research Applications Laboratory (2014). *verification: Weather Forecast Verification Utilities*. R package version 1.39. http.//CRAN.R-project.org/package=verification.

National Weather Service (2012): *Overview of the Hydrologic Ensemble Forecast Service (HEFS)*, Release 0.1.2, June 2012, Office for Hydrologic Development. URL [accessed 4/25/2014]: http://www.nws.noaa.gov/os/water/RFC_support/HEFS_doc/HEFS_Overview_0.1.2.pdf

Pielke, R.A. (1999): "Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood", *Applied Behavioral Science Review*, 7(2), pp.83–101.

R-Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://ww.R-project.org/.

Stephenson, D. B., Coelho, C. A., & Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, *23*(4), 752-757.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S. (2011): "Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)", *Hydrol. Earth Syst. Sci*., 15, 255-265, doi:10.5194/hess-15-255-2011, 2011.

Wilks, D.S. (1995): *Statistical Methods in the Atmospheric Sciences*. San Diego: Academic Press.

Wilson, L.J. (n.d.). *Verification of probability and ensemble forecasts*. Atmospheric Science and Technology Branch, Environment Canada. URL [accessed July 2014]: http://www.swpc.noaa.gov/forecast_verification/Assets/Tutorials/Ensemble%20Forecast%20Verification.pdf

WWRP/WGNE (2009). Methods for probabilistic forecasts. *Forecast Verification – Issues, Methods and FAQ*. URL [accessed July 2014]: http://www.cawcr.gov.au/projects/verification/verif_web_page.html#BSS

STUDY 4

Modeling the Effectiveness of Urban Best Management Practices for Stormwater Treatment as a Function of the Volume of Runoff[52]

Abstract

Impervious surfaces, such as roads and parking lots, can increase the volume of runoff and lead to more pollution reaching streams, rivers, and lakes. Best Management Practices (BMPs) reduce the peak discharge into the storm sewer system and remove pollutants such as sediments, phosphorus and nitrogen from the stormwater runoff. Bioretentions, dry and wet ponds, porous pavement, and many other methods collect the runoff and reduce the concentrations of sediments and nutrients with varying effectiveness. Numeric models to develop and assess plans to implement BMPs need to take into account the variation of BMP effectiveness with the volume of runoff (inch-acre) in order to estimate their performance under a range of different climate scenarios. Circumventing the lack of monitoring data, this study uses the Environmental Protection Agency's (EPA) System for Urban Stormwater Treatment and Analysis IntegratioN Model (SUSTAIN) to run 22 years of weather data from the Patuxent River (Maryland) through six types of BMPs. It is found that BMP effectiveness decreases sooner, steeper and deeper with increasing sizes of storm events than assumed in current models. At a minimum, the resulting performance curves should differentiate between BMPs; ideally, also between landuses.

Keywords: Best Management Practices, Stormwater, Climate Change, Storm Size, Return Frequency, Effectiveness

---

[52] In fall 2014, a version of this chapter will be submitted to the journal Water Research.

1. Introduction

Best Management Practices (BMPs) reduce the quantity and increase the quality of stormwater before it reaches larger water bodies such as rivers. Climate change is hypothesized to alter precipitation frequencies and intensities. BMP implementations plans need to be robust to this uncertain future; wisely chosen and dimensioned BMPs should not be overwhelmed by possible changes in precipitation event frequency and intensities.

The Environmental Protection Agency (EPA) uses large numerical hydrological models to develop and assess BMP implementation plans (e.g., CBP, 2012). However, due to sparse monitoring data, the relationship between BMP effectiveness and storm size is only tentatively defined. Our understanding of effect of storm size on BMP effectiveness needs improvement before we can assess the impact of changing precipitation frequencies and intensities. This paper presents a numerical study of this relationship for six types of BMPs in order to improve the capability of hydrologic models to study the robustness of BMP implementation plans to climate change. The Patuxent watershed in Maryland with the hydrological model used there serves as a case study.

1.1 Pollution in Stormwater Runoff as a Consequence of Urbanization

Some watersheds, like the Patuxent in Maryland, experience increasing urbanization. While this happens on a small scale – development project by development project – the accumulated effects are large. One third of the pollutants in the largest estuary in the US, the Chesapeake Bay, comes from urban stormwater runoff (Maryland, 2007).

Urbanization results in large impervious surfaces, such as roads and roofs, that reduce stormwater ground infiltration (Maryland, n.d.). When the stormwater cannot infiltrate in the soil, this has two effects. First, the peak discharge increases because more water flows into the sewage system quicker, possibly overwhelming the storm sewers. The return frequency of peak runoff rates can increase by 6-18 times when an area undergoes development (Roesner et al., 2001). Second, the flowing water picks up many urban pollutants. Fertilizer and pesticides from lawns, oil and rubber from cars, trash and sediment are washed off with the runoff (Maryland, n.d.). Without measures being taken, more water with a higher concentration of pollutants will arrive in neighboring water bodies causing pollution that accumulates towards the downstream end of a watershed.

1.2. Best Management Practices (BMPs)

BMPs were first introduced to reduce the peak discharge into the storm sewers. Dry ponds were built to retain the water and release it gradually into the sewage system. For this reason, such ponds are one of the most common BMPs in the U.S. today. Later, environmental concerns led to the idea of Low Impact Development (LID). In LID, a building project should impose as little change as possible to the environment, including the hydrology (Dietz, 2008). Therefore, it became important to not only decrease the peak discharge of stormwater runoff, but to make sure that the runoff does not carry pollutants with it. This popularized "green" BMPs such as bioretentions, bio swales, infiltration trenches, etc., both reduce the volume and offer sedimentation, plant uptake and filtration as means to treat stormwater (e.g., the water is filtered when it infiltrates into the soil).

Monitoring BMPs poses many challenges. Measuring the inflow, the amount of pollutants in the inflowing runoff, etc. is labor-intensive and the measurement technologies are vulnerable to weather and other circumstances. When data are collected, they often carry many caveats. Very small or large storms might fall outside the range of measuring equipment. Sample collection to determine the pollutant concentration in in- and outflow is tricky and the pollutants are not uniformly distributed throughout a sample. Lastly, BMP performance is heavily dependent on context and design, so that monitoring results are difficult to generalize (Jones et al., 2004). Given these difficulties, it is not surprising that the BMP effectiveness data from the monitoring studies collected in the International Stormwater BMP Database (2013) are characterized by large uncertainties and only allow rough estimates, such as that a bioretention removes 80% of the phosphorus from the influent.

1.3. BMP effectiveness as a function of storm size

If climate change alters the frequency and intensity of precipitation events, then existing BMPs will have to cope with different amounts of water in the future. Intuitively, it is likely that the effectiveness of BMPs to remove pollutants from the runoff varies with the amount of inflowing runoff. BMPs like ponds might overflow, the vegetation that takes up pollutants in bioretentions might become saturated or filter layers might become clogged more rapidly.

There are very few studies of BMP effectiveness as a function of storm size. To our knowledge, there is only one study that modeled pollutant removal in BMPs (Ackerman & Stein 2008) and several others studying the discharge reduction for various storm sizes (Damodaram et al. 2010, Heitz et al. 2000).

Ackerman & Stein (2008) modeled a retention facility, a flow-through swale and a combination of those two with the predecessor of the software used in this study. Running the model with ten years of runoff data for two creeks in California, they found BMP effectiveness suffered during large storms and wet years. BMP effectiveness proved to be most sensitive to volumetric changes and loss rates. Changing the infiltration rate had a large effect, while altering the decay rate had a modest effect (Ackerman, Stein, 2008). Damodaram et al. (2010) implemented several scenarios of LIDs and conventional BMPs in a watershed model of their campus in Texas. Modeled LIDs and conventional BMPs included permeable pavements, rainwater harvesting, green roofs and detention ponds. They conclude that the infiltration-based LIDs are more successful than storage-based conventional BMPs in peak flow reduction during small storms. But for large storms the conventional BMPs deliver better results. The authors recommend implementing a combination of LIDs and conventional BMPs (Damodaram, 2013). Heitz et al. (2000) used a computer model to determine rules for dimension ponds so that they achieve a pre-determined effectiveness. His study shows that a pond relative to its drainage area (a proxy for the volume of runoff reaching the pond) needs to increase to boost volume capture efficiency. Heitz was unable to specify the effect on pollutant removal but expects it to be non-linear (Heitz et al., 2000).

There are many studies reporting monitoring results from BMPs all over the nation (e.g., as collected in the International Stormwater BMP database). However, only very few mention the relationship between storm size and BMP effectiveness. Barrett et al. (1998) monitored the removal of suspended solids in two vegetated swales along highways in Texas. Evaluating 34 storms, they did not find a relationship between storm size and removal rate. They hypothesize that either longer, larger storms result in similar water depths like small, intense storms or that pollutant removal occurs in the shallow parts of the swale where the water level does not vary much with storm size (Barrett et al., 1998). Li et al. (2009) monitored six bioretentions in Maryland and North Carolina for 22-60 events over period of 10-15 months. They focused on the reduction of peak discharge rate and total amount of discharge. Based on a regression analysis, they concluded that peak rate reduction decreases for increasing rainfall depths and durations, while the reduction of the total amount of discharge is only affected by the rainfall depth ($R^2 > 0.41$). Average daily temperature, rainfall intensity and antecedent dry weather period did not impact either of those two metrics (Carpenter, Kaluvakolanu, 2010). Hunt & Winston (2010) found a strong correlation between storm size and percent peak flow reduction and percent volume reduction when monitoring level spreader – vegetated filter strips in North Carolina. They estimate that for storm events larger than 1.25 cm, the BMP effectiveness starts to decrease (Hunt et al., 2009).

Barber et al. (2003) ran tests in a physical model of an ecology ditch and used those test results to calibrate a computer model. They found that percent peak reduction and peak attenuation decrease with larger storm size. For storms larger than 1.27 cm (0.5 inch), the decrease seems to flatten. They suspect that soil water content drives this process.

This literature review suggests that there is an agreement that BMP effectiveness decreases non-linearly with storm size. However, the exact nature of this relationship is unknown for most BMPs and pollutants.

2. Case Study: Upper Patuxent, Maryland

The case study presented in this paper is the Upper Patuxent watershed (Figure 33). The Patuxent River is part of the Chesapeake Bay watershed. The Chesapeake Bay is the largest estuary in the U.S. Half of the water in the bay is freshwater coming from the rivers in the watershed, while the other half is saltwater from the Atlantic Ocean. Between 1950 and 2008, the population in the watershed doubled to 16.8 million. It is estimated to grow by 157,000 people per year (EPA Region 3, 2010).

With its roughly 650,000 inhabitants – including those of Washington D.C. and Baltimore – the watershed of the Patuxent River is an especially densely populated part of the Chesapeake Bay watershed. In Maryland, the Patuxent is the largest and longest river. This case study is based on data from four land segments in the Upper Patuxent Watershed.

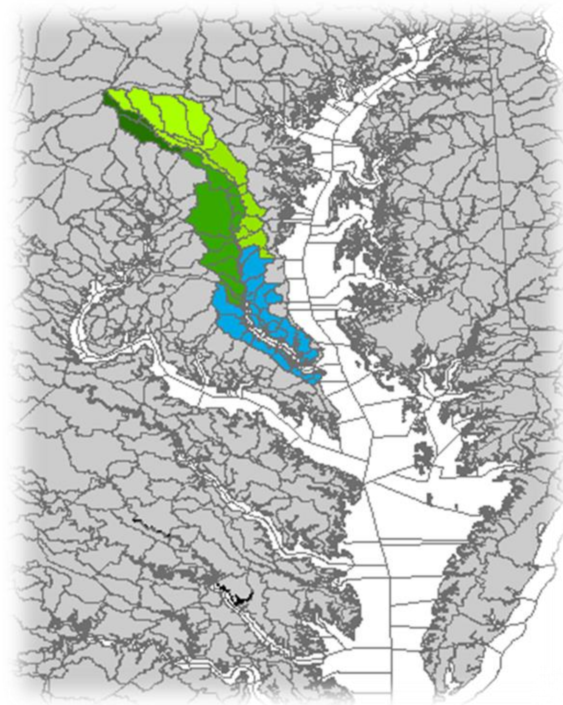


Figure 33: Chesapeake Bay Watershed (left) and studied land segments in the Patuxent watershed (right). In the Phase 5 Watershed model, these land segments are known as A24003, A24027, A24031, and A24033. ***Source left picture:*** **EPA(2010):** ***A Method to Assess Climate-***

Even though the pollutants loads in the Patuxent have been decreasing in the last several decades, 3.7 million pounds of nitrogen, ca. 280,000 pounds of phosphorus and ca. 130,000 tons of sediment reached the Patuxent in 2005. For nitrogen and phosphorus, urban areas are the major source (34% and 38% respectively), followed by agriculture (22% and 23%) and point sources (16% and 25%). Agriculture is the largest source of sediment, contributing 54%; 28% of the sediments originate from urban land uses (Maryland, 2007).

2.1. Watershed Implementation Plans

In the Chesapeake Bay, BMPs are implemented as part of Watershed Implementation Plans (WIPs). These plans are developed by states and districts to ensure that the water quality standards as required by the Clean Water Act and particularly the Total Mean Daily Loads (TMDLs) as set by Environmental Protection Agency (EPA) are met. TMDLs specify how much pollutants can be added to the water, before the water standard is compromised. WIPs determine which Best Management Practice will be constructed where to reduce the pollution in those runoffs. EPA reviews the draft WIPs and monitors the implementation of the finalized WIPs using two-year mile stones (EPA, 2013).

The current TMDL for the Chesapeake Bay issued in December 2010 aims for a 25% load reduction in nitrogen to an annual load of 185.9 million pounds, 24% in phosphorus (12.5 million pounds) and 20% of sediments (6.45 billion pounds). Because urban runoff contributes a large share to the overall pollution, this translates to a reduction by 35% of the total nitrogen 36% of the total phosphorus in urban runoff (Geosyntec Consultants, 2012). To achieve the TMDL for the Chesapeake Bay jurisdictions across six states have established Watershed Implementation Plans (WIPs) that describe through which measure the "pollution diet" will be carried out (EPA, 2013).

2.2. Models to develop and assess plans for BMP implementation

As part of the Chesapeake Bay Program (CBP), EPA uses a hydrological model called Phase Watershed 5 Model to determine TMDL and BMP requirements and to assess WIPs. That model is based on the Hydrologic Simulation Program-Fortran (HSPF) model, first introduced in the 1960s. Wurbs (1997) succinctly summarizes the HSPF model developed by the E.P.A. as follows:

> *"HSPF is a comprehensive package for simulation of watershed hydrology and water quality for both conventional and toxic organic pollutants. … The model uses information such as: the time history of rainfall, temperature and solar radiation; land surface characteristics such as land use patterns and soil properties; and land management practices to simulate the processes that occur in a watershed. Flow rates, sediment loads, and nutrient and pesticide concentrations are predicted for the watershed runoff. The model*

uses these results, along with the input data characterizing the stream network and point source discharges to simulate instream processes. Model out includes a time history of water quantity and quality at all pertinent locations in the watershed/stream system. …"

HSPF models are divided into land segments that determine the weather and runoff and river segments that collect the runoff from the land segments. Each land segment has different landuses, such as developed, agriculture, and forest. It is a categorically segmented model, meaning that acres with the same land use in the same land segment are treated as one unit; even if they are topographically detached.

In the Phase 5 Watershed model, the BMP effect is modeled as an end-of-pipe percent reduction of the pollutants in the runoff. The effectiveness of a BMP is specified (e.g. a bioretention removes 80% of the sediment in the runoff) as is the extent to which it is to be implemented in different land uses and land segment. After the runoff for each land segment's land use has been computed and before it is forwarded to a river segment, the amount of pollutants in the runoff is reduced by the pre-determined percentage.

As described above, there is reason to assume that the BMP effectiveness changes with the volume of runoff. In the CBP Phase 5 Watershed model that is defined by a Michaelis-Menten (MM) function. This function computes BMP effectiveness as a percentage of its given effectiveness as a function of return frequency. As an example, if a BMP has a target value of 80%, its actual effectiveness for any given day will be between 0% and 80% depending on the return frequency of the volume of runoff. In the model, this works as follows. The MM function computes a multiplier with the range [0, 1] as a function of the volume of runoff. The target value of the BMP (e.g. 80%) is then multiplied by that multiplier (resulting in actual effectiveness between 1*80%=80% and

0*80%=0%). The MM function is the one piece in the Phase 5 Watershed model to capture the effect of climate change on the performance of BMP implementation plans.

The three parameters of the MM function (see Appendix for formula) have an easy-to-understand meaning and have been assigned rather optimistic values in the current Phase 5 Watershed model. The first parameter is the asymptote (Asymp), which stands for the effectiveness of the BMP in the worst case, i.e. during the largest storm. In the current CBP Phase 5 Watershed model that is assumed to be 20% of the given effectiveness. BMPs can thus never completely fail. Additionally, this asymptote is not reached until an volume of runoff reoccurring with a return period of much larger than 200 years; i.e. for an volume of runoff with the return period of 200 years, the effectiveness is still 25% of the given BMP effectiveness.

The second parameter (Thres)[53] describes the threshold from which volume of runoff onwards, the BMP effectiveness starts to deteriorate. The default assumption in the Phase 5 Watershed model for threshold is volume of runoff reoccurring with a frequency of five years. More than 99.9% of all days experience less rainfall, so that the MM function only applies to <0.01% of the data.

The third parameter (HalfSat) is a sort of half time value that determines how steeply the curve drops towards the asymptote after the threshold. The default assumption is 17 years, meaning that the effectiveness is 60% of the given effectiveness during a storm with the return frequency of 17 years. That is half way between maximum effectiveness and the asymptote.

[53] Called MinFreq in the CBP Phase 5 Watershed model's Hydro-Effect method.

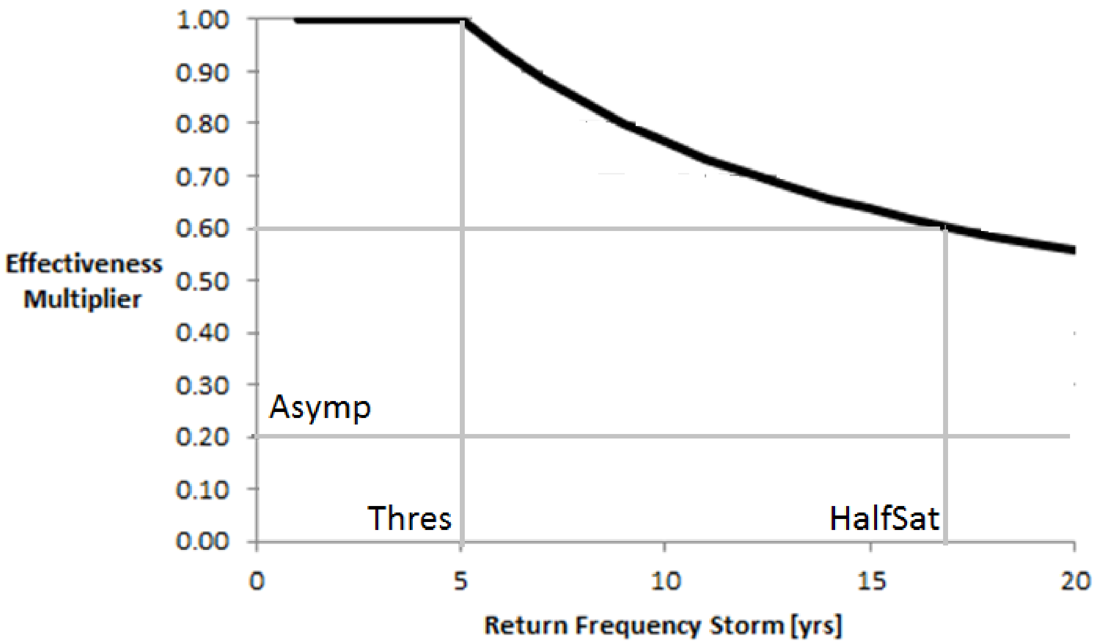


Figure 34: Plot of Michaelis-Menten curve with the default values of its three parameters in the Phase 5 Watershed model: Asymp – Aysmptote (20%), Thres – Threshold (5 years), HalfSat – Half time value (17 years)

2.3. Research Questions

The research presented in this paper seeks to answer the following question: **What is the best definition of the relationship between BMP effectiveness and the volume of runoff to be used in hydrological models such as the Chesapeake Bay Phase 5 Watershed model?**

A software package that numerically models the physical processes inside a BMP was used together with weather data from the Patuxent watershed to study the effect of storm size on the effectiveness of seven BMPs: Bioretention, porous pavement, wet pond, dry pond, extended dry pond, vegetated swale and infiltration trench.

3. Method

This research was carried out in three steps. First, the EPA software package SUSTAIN was used to model the physical processes inside the BMP. Second, the Michaelis-Menten (MM) curve was fit to the SUSTAIN output to refine the assumptions in the Phase 5 Watershed model. Third, we investigated how the parameters of the MM curve change across different BMPs, land segments and pollutants. Besides these steps, the following proposes changes to the MM curve to better capture the full spectrum of BMP effectiveness.

3.1. Modeling and Optimizing BMPs

The System for Urban Stormwater Treatment and Analysis IntegratioN Model (SUSTAIN) developed by EPA and TetraTech was used to model the physical process inside the BMPs. SUSTAIN's Process Module "can evaluate management practices under an individual storm event and/or continuous storm conditions. Available BMP processes include weir and orifice functions to define surface capacity and control, swale characteristics, hydraulic transport, infiltration and saturation, underdrain, evapotranspiration, general pollutant removal and filtration through soil media" (Lai et al., 2007).

For the years 1984-2005, the hourly surface runoff, the groundwater recharge volume and the pollutant loads were extracted from the CBP Phase 5 Watershed model to serve as input in the SUSTAIN model.

The following seven BMPs were modeled using SUSTAIN's templates: Bioretention, porous pavement, dry pond, extended dry pond, [54] wet pond, vegetated swale, and infiltration trench. Each BMP has been modeled for each time series separately. Put differently, for each model run, it has

[54] The extended dry pond was modeled with the same BMP template as the dry pond, but with higher target values for removal efficiency.

been assumed that there is only one land use and one BMP. For the lot-scale BMPs (bioretention, porous pavement) the size of that imaginary watershed was one acre, for the community-scale BMPs (remaining five) the imaginary watershed covered five acres.

The dimensions of the BMPs were optimized to the runoff provided by each time series. In SUSTAIN's optimization module each BMP was assigned target values for three pollutants: sediment, nitrogen and phosphorus. The target values are the average annual effectiveness of a BMP and have been taken from the CBP Phase 5 Watershed model. For each of the 14 time series, this module was then used to find the minimum length so that all three pollutant removal targets were met. The module minimizes the cost, which in this case effectively means minimizing the volume of the BMP (or surface in some cases, see Table A1).

Afterwards, the pollutant loads in the water flowing in and out of the BMP were estimated using the SUSTAIN output. The hourly data were aggregated from midnight to midnight to get daily data points, because the Phase 5 Watershed model operates with a time step of one day. To compute the daily effectiveness for each pollutant the mass in the outflow was divided by the mass in the inflow.

The return period of runoff was computed using the Log Pearson Type 3 distribution (Bobee, 1975). For each of the 22 years, the maximum daily runoff was determined. For the log of those 22 data points, the mean, standard deviation and skewness were calculated. Using those three statistical measures as the parameters for the Pearson Type 3 distribution, the return period for the log of each daily runoff was computed.

Finally, a revised version of the Michaelis-Menten function (described below) was fit to the SUSTAIN output. To facilitate curve fitting, it was desirable to have as many data points as possible for each fit. Therefore, pervious (rpd, npd) and impervious (nid, rid) landuses each have been lumped together. Additionally, the data for phosphorus and nitrogen removal were combined. Compared to sediment, those two pollutants undergo rather similar processes. Consequently, the data was stratified along those four dimensions into 112 subsets (2 land uses * 4 land segments * 6 BMPs * 2 pollutants). The MM curve was fit to each of these subsets. The parameters of each fit can then be compared to the default assumptions in the Phase 5 Watershed model.

3.2. Changes to the Michaelis-Menten curve as used in the Phase 5 Watershed model

In comparison to the conventional MM function, several changes were made to capture as much information from the SUSTAIN output as possible. These changes are described in the following.

3.2.1. Performance curve as a function of runoff rather than return period

In the CBP Phase 5 Watershed model, a MM curve describes the relationship between return frequency of the volume of runoff and the BMP effectiveness (called performance curve hereafter). As described above, the Log Pearson Type 3 distribution is used to compute the return frequency of various amounts of runoff. A mathematical constraint of this distribution is that the return period cannot be smaller than 1 year. However, in the Patuxent watershed more than 97.5% of the storm events are smaller than that. To not lose this information, it was decided to define the effectiveness as a function of the volume of runoff rather than the return period. Additionally, the runoff better reflects the characteristics of the land use and other climate characteristics such as temperature.

3.2.2. Adjusted Multiplier

In the current CBP Phase 5 Watershed model the threshold after which the BMP effectiveness starts dropping off is so high (runoffs with a return period of more than five years) that the maximum effectiveness approximately equals the average effectiveness. There are so few days with an volume of runoff larger than the one occurring once in five years that they barely weigh into the average. In this case, it does not matter if the given target value is the maximum or average effectiveness. However, it is expected that effectiveness falls off sooner than the current threshold. Then the maximum effectiveness will not equal the average effectiveness anymore, especially for BMPs whose effectiveness drops off soon and steeply. This causes the following problem.

When BMPs are monitored usually the average effectiveness is reported. It is therefore assumed that the given target values represent the average rather than the maximum effectiveness of a BMP. In the current CBP Phase 5 Watershed model, the multiplier computed by the MM function is constrained to a range [0; 1]. However, in cases where the maximum BMP effectiveness is higher than the average effectiveness, the multiplier will have to be larger than one. To cope with that, the multiplier has to be re-scaled by dividing it by the maximum of the multiplier, so that the adjusted multiplier never exceeds one.[55]

$$Adjust.Multiplier = \frac{Eff}{given\ annual\ average\ effectiveness} * \frac{1}{\max(\frac{Eff}{given\ annual\ average\ effectiveness})}$$

$$Eff = (1 - \frac{Pollutant\ mass\ in\ BMP\ 24\ hr\ Outflow}{Pollutant\ mass\ in\ BMP\ 24\ hr\ Inflow})$$

for $0 \le Eff \le 1$

3.2.3. Revised definition of Performance Curve

The changes to the original MM curve described above result in the following revised definition of the MM curve:

[55] Note, the adjusted multiplier in this study is different to the multiplier used in the Hydro-Effect Method in the current CBP Phase 5 Watershed model. That multiplier is the first half of the adjusted multiplier formula.

$$Adj.Multiplier = (1 - \frac{(1 - Asymp) * (Runoff\ - Thres)}{Runoff - Thres + HalfSat - Thres})$$

$$For\ (Runoff < Thres): Eff = 1$$

$$0 \leq Asymp \leq 1; HalfSat > Thres$$

for $0 \leq$ *Adj. Multiplier* ≤ 1

With:

(Asymp, Thres, Halfsat)	– function parameters
Runoff	– Volume of runoff (SURO) in 24 hours [in-acre]
Multiplier	– Re-scaled multiplier of for BMP effectiveness [%]

4. Results

Here we discuss results, including the empirical relationship between BMP effectiveness and the volume of runoff and the parameters of that relationship.

4.1. Effectiveness as a function of runoff

When optimizing the BMP dimensions to achieve the given effectiveness, sediment removal was in all cases the determining factor. For nitrogen and phosphorus the achieved removal efficiencies are much larger than the target value.

Table 21 shows the given target and the ranges of effectiveness in the SUSTAIN output for each BMP and pollutant. Either the modeling of processes in SUSTAIN does not capture reality well or the assumed target values exported from the CBP Phase 5 Watershed model are unrealistic for the Patuxent watershed.

Figure 35 shows the scatterplots for four BMPs. As expected, the BMP effectiveness decreases for higher amounts of runoff. It is interesting to note, that some days record negative

BMP effectiveness. This could be an artifact of aggregating the data from midnight to midnight. If it rains late in the day, the pollutants stay in the BMP past midnight and are carried away the following day. On that day, it looks as if the effectiveness is negative because more pollutants are coming out of the BMP than entering it. When aggregating the data over a week (168 hours), the negative values indeed disappear (Figure 36, row 2). Additionally, it is noticeable that the steepness of the performance curve decreases. When the data are aggregated over a year, the performance curve becomes flat (Figure 36, row 3). Therefore, the BMP effectiveness is assessed to be independent of how much rain falls in a year.

Table 21: Effectiveness target values and optimization results [Low, Mean, High]for each Best Management Practice and pollutant. The values denote the average annual effectiveness, i.e. the fraction of pollutants removed from the storm water.

BMP	Sediment	Total Nitrogen	Total Phosphorus
	Given average annual effectiveness SUSTAIN output: Effectiveness [Low, Mean, High]		
Bioretention	**0.80** **[0.81, 0.83, 0.90]**	0.75 [0.84, 0.91, 0.96]	0.70 [0.83, 0.89, 0.94]
Porous Pavement	**0.70** **[0.70, 0.73, 0.78]**	0.50 [0.75, 0.85, 0.90]	0.50 [0.73, 0.79, 0.83]
Extended Dry Pond	**0.60** **[0.60, 0.61, 0.62]**	0.20 [0.67, 0.77, 0.84]	0.20 [0.64, 0.70, 0.75]
Dry Pond	**0.10** **[0.10, 0.11, 0.13]**	0.05 [0.22, 0.27, 0.31]	0.10 [0.14, 0.18, 0.22]
Infiltration Trench	**0.95** **[0.95, 0.95, 0.97]**	0.8 [0.96, 0.98, 0.99]	0.85 [0.95, 0.97, 0.98]
Vegetated Swale	**0.70** **[0.70, 0.72, 0.78]**	0.45 [0.78, 0.83, 0.86]	0.45 [0.74, 0.81, 0.78]
Wet Pond	**0.60** **[0.61, 0.66, 0.84]**	0.20 [0.64, 0.78, 0.86]	0.45 [0.63, 0.72, 0.85]

SUSTAIN output: $Effectiveness = (1 - \frac{Pollutant\ mass\ in\ BMP\ Outflow}{Pollutant\ mass\ in\ BMP\ Inflow})$
Given average annual effectiveness taken from CBP Phase 5 Watershed model.
Pollutant that determined BMP size is printed in bold.

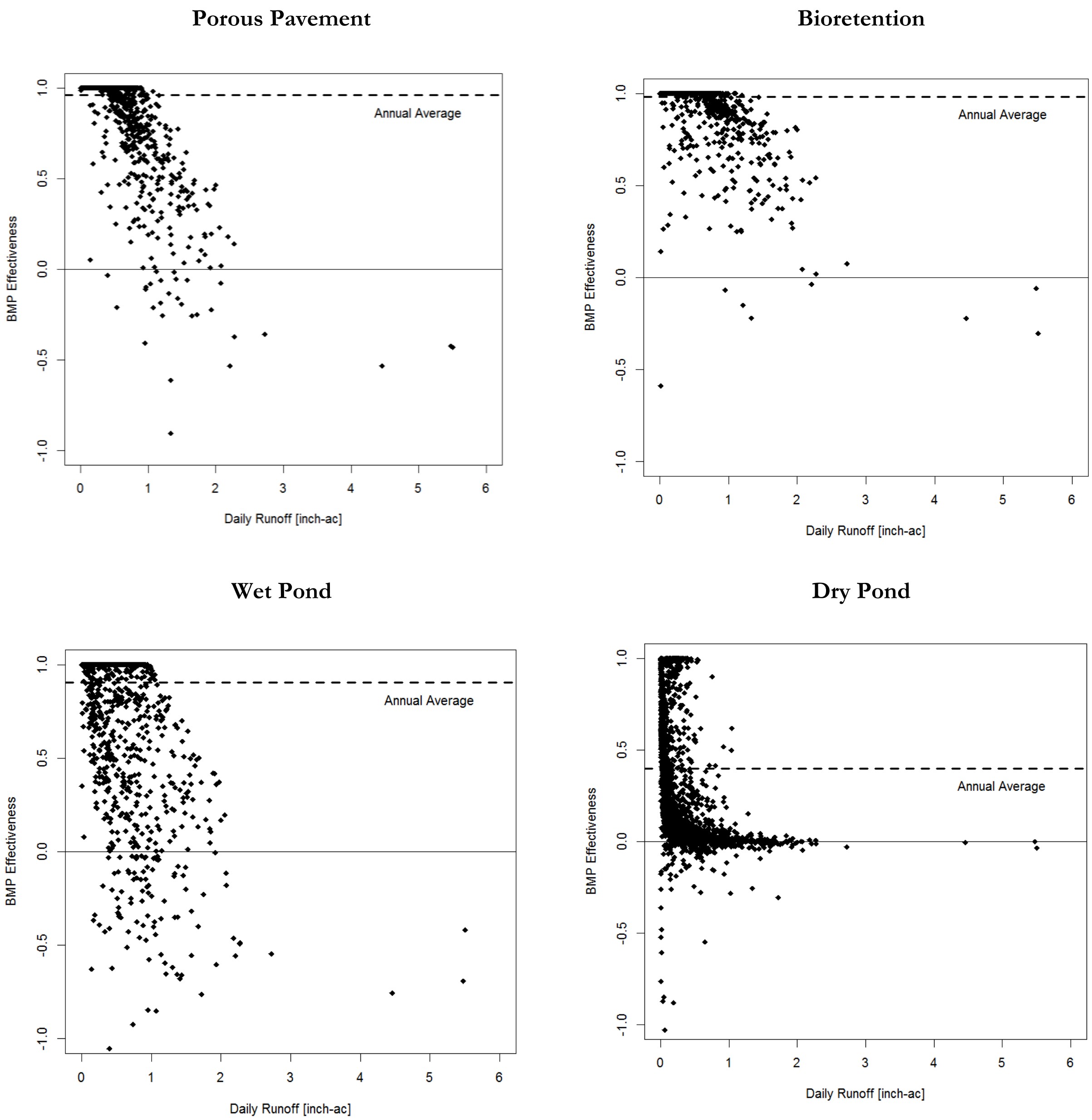


Figure 35: Sediment removal for bio retention, porous pavement, wet and dry pond for impervious land use A24003, aggregated for 24 hours (1 day).

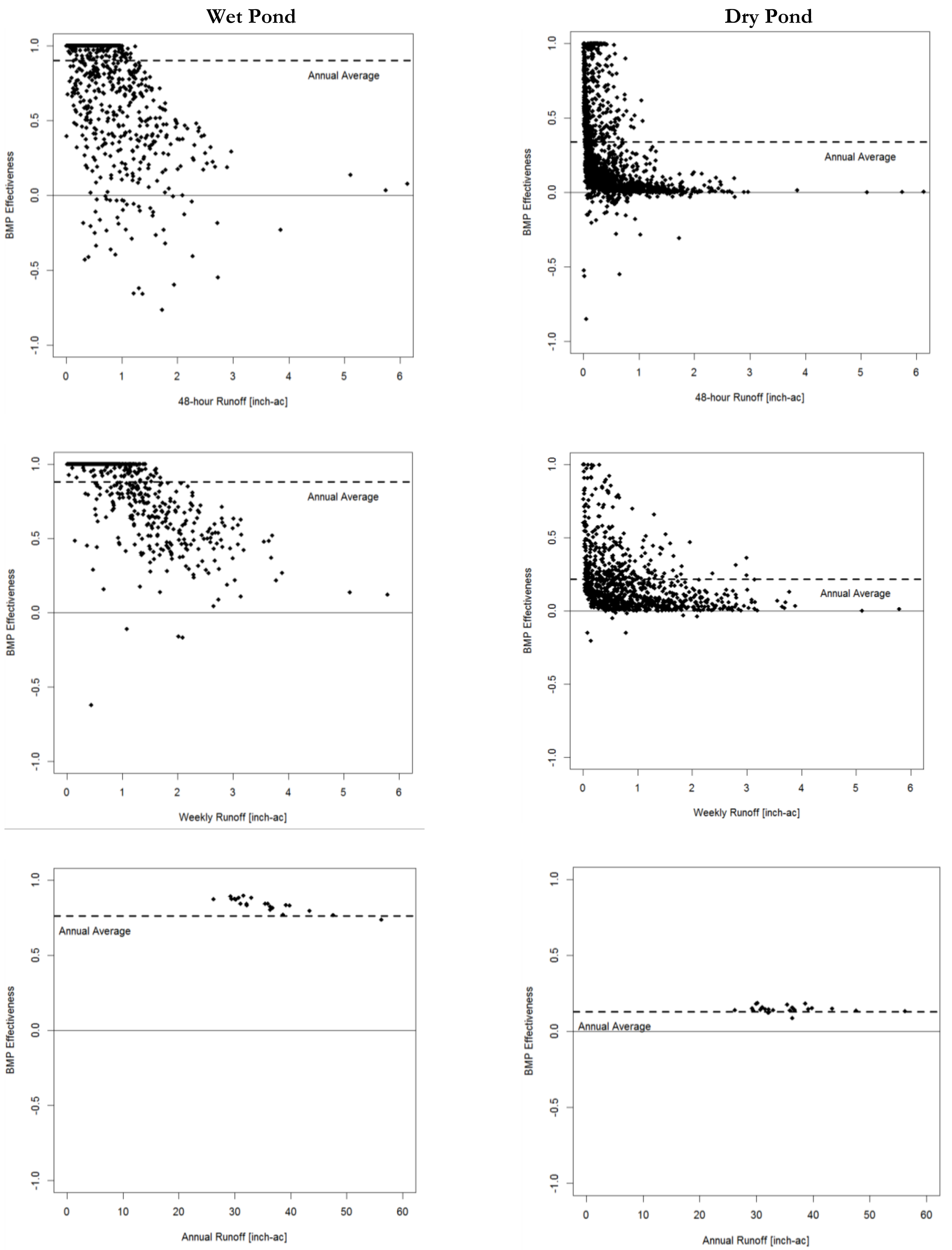


Figure 36: Sediment removal for wet and dry pond on regulated impervious developed land use A24003, aggregated for 48 hours (2 days), 168 hours (1 week), and 8760 hours (1 year)

4.2 Refined Michaelis-Menten curve

The MM curve was fit to the SUSTAIN output per BMP, pollutant, land segment and land use. Figure 37 shows the revised MM curves for the developed impervious land use in one of the land segments. To be able to compare it to the default MM curve in the Phase 5 Watershed model, it has been plotted against the return period rather than the volume of runoff. Due the rescaling described earlier, the multiplier of the revised MM curve can exceed 1, so that the average annual effectiveness equals the BMP effectiveness observed in the field. In the following, each parameter of the MM-curve will be discussed.

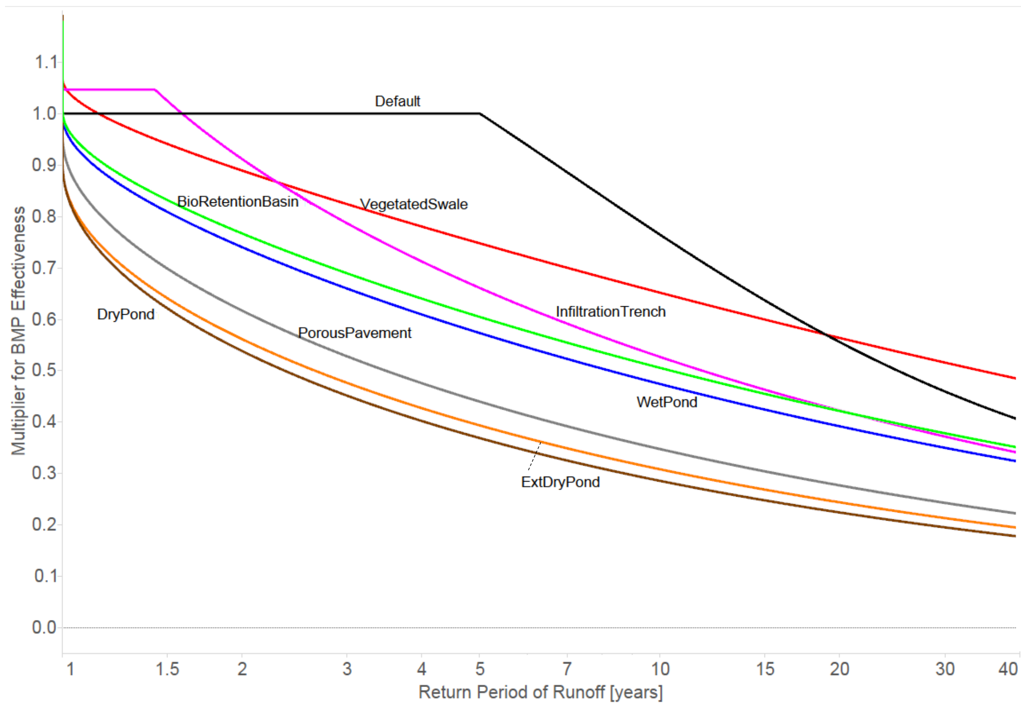


Figure 37: Revised MM curves for several BMPs plotted against the return period. The "default" curve denotes the MM curve currently assumed in the Phase 5 Watershed model. The figure shows the MM curves for the regulated impervious developed land uses in land segment A24003 in the Upper Patuxent watershed.

4.1.1. Threshold

In reality, the size of BMPs is usually chosen to be able to cope with a certain volume of runoff, e.g. for amounts of runoff with return periods of one to two years.[56] The threshold volume, up to which which the BMP functions perfectly, should thus equal that design return period. In this study, the BMPs were dimensioned to achieve pollutant removal rates that have been observed in the field. The effectiveness of the modeled BMPs drops off much faster than assumed in the Phase 5 Watershed model. For example, in land segment A24003, only the bioretention, the porous pavement, and the extended dry pond do not decline for runoffs occurring more often than once a year.[57] In that land segment, all BMPs have started losing effectiveness for runoffs with a return period greater than 1.5 years. For impervious land uses, only infiltration ponds have a threshold larger than one year. That is also true, when averaging over both land uses, see Table 32 in the appendix.[58] It is possible that the BMPs that were monitored in the field could cope with larger *quantities* of water, e.g. those occurring every one or two years. The *quality* of the effluent seems to start deteriorating earlier.

4.1.2. HalfSat (Steepness of MM-curve)

The rate at which the BMP effectiveness decreases with an increasing return period is important information, if new BMP requirements are meant to be robust to climate change. In that case, BMPs whose effectiveness drops off more gradually with increasing amounts of runoff are the better

[56] In Maryland, a BMP's capacity has to be large enough to cope with 0.9-1 inch-acre of rainfall encompassing ca. 90% of the annual rainfall. Schueler, T.R. and R.A. Claytor, *Maryland Stormwater Design Manual.* Maryland Department of the Environment. Baltimore, MD, 2000.
[57] That is probably also true for infiltration trench. Due to their high effectiveness, it was not possible to fit performance curves for infiltration trench in pervious land uses. They release effluents on too few days.
[58] For some land uses and BMPs the scatter plots were repeatedly too sparse to fit a performance curve, see Table 32. This skews the average.

choice. In this study, infiltration-based BMPs such as the bioretention, vegetated swale and infiltration trench are on average more robust to different storm sizes. They are followed by wet pond, extended dry pond and porous pavement. Conventional pond-based BMPs are least robust on average. They were originally not intended to improve water quality. So it is not surprising that they fare worst in terms of their effectiveness in improving water quality.

4.1.3. Asymptote

The asymptote is 0% in all cases. There is very little uncertainty around the asymptote, with three BMPs having a maximum asymptote as low as 10%. Consequently, the default assumption of a minimum effectiveness of 20% of the target value is too optimistic.

4.2. Discussion Revised Michaelis-Menten Curve

This case study shows that performance curves of BMPs are sensitive to context. How well the MM-curve fits the data differs considerably for different BMPs, land uses and land segments.

First, the steepness of the performance curves is heavily dependent on the annual average effectiveness of the BMPs. Due to their given target values, dry pond and infiltration trench were the least and most effective BMP respectively. Not by coincidence, their performance curves are the steepest and least steep respectively. For pervious land uses, it was not even possible to fit a MM-curve through the infiltration trench data, because it functions so well that most runoffs do not lead to an effluent. This could be explained as follows. Larger storms with more runoff and pollutants will be a larger impact relative to its size on a small or not so effective BMP than on a larger or very effective BMP. Therefore, the performance curve of a small or ineffective BMP decreases quicker with larger amounts of runoff than the performance curve of a large or effective BMP.

Additionally, the MM-curves fit more or less well to different BMPs. It was not possible to fit a BMP curve through data for the dry pond for some impervious land uses. Its effectiveness declines too rapidly. When giving the dry pond less stringent effectiveness targets (Extended Dry Pond), its performance curve blends in with those of the other BMPs. But curve fitting for all three pond-based BMPs results on average in the largest root mean squared errors (RSMEs), suggesting that the MM-curve leaves space for improvement. The performance curves of the pond-based BMPs have different shapes, too. The wet pond's performance, for example, looks much less convex than the performance curves of bioretention, porous pavement or infiltration trench. This could be due to the fact that the later three are infiltration-based BMPs. In contrast, the dry pond displays a sharply curved performance curve.

The data are especially sparse in the tail with high amounts of runoff. The vegetated swale and also the bioretention have relatively wide tails, where few data points cover a large band width. For those BMPs, it is much less certain what the performance curve actually looks like in the tail.

Second, the land segment matters as well. Out of the four modeled land segments, A24027 seems to produce the least runoff. Therefore, in many model runs the BMPs did not release any effluent. Occasionally, this also happens for the pervious land uses in A24031. The resulting sparse scatterplots made it in many cases impossible to fit a performance curve through the data points. As a result, the RMSE for those land uses also have the larger average values than for the other two land uses.

Third, it makes a difference for the shape of a performance curve, whether a BMP is placed in a previous or impervious land use. Pervious land uses produce less runoff because the stormwater gets the chance to infiltrate into the ground on its way to the next water body. The threshold after which BMP effectiveness starts dropping off is much larger in terms of return frequency on pervious land uses than on impervious ones. The effectiveness drops off much less rapidly on impervious land uses, probably because an impervious land use produces proportionally more runoff for events with a larger return period. The larger threshold and steeper decline give performance curves on pervious land uses a more convex look.

In reality, BMPs are built to serve the estimated percentage of the surface that is impervious. The rationale is that development converts pervious surfaces to impervious surfaces. In that rationale, pervious surfaces are assumed not to contribute to an increase in runoff or pollutant load. Obviously, that assumption is more valid for the quantity of runoff, but much less so for the quality of runoff. E.g., a lawn might be pervious, but due to fertilizer the runoff will contain much more pollutants. Additionally, perviousness comes in many degrees, so that the quantity of runoff from different pervious land uses is variable as well. This study shows that performance curves are dependent on the land use. It follows that it is important to study how much runoff of what quality will be captured when dimensioning a BMP.

In its basic form, the Phase 5 Watershed model, cannot distinguish between all these different contexts. However, it does give the option to distinguish by BMP. This option should definitely be made use of. It would be impractical to assign performance curves to each land segment. In this case, it is wise to make conservative assumptions. As to land use, it would be worth to consider making the performance curve a function of runoff. Figure 38 demonstrates why. Like mentioned before, BMPs dimensions are based on the quantity of runoff rather than the return period. The BMPs are mainly dimensioned for the very frequent small runoffs, but are much less optimized to meet the demands of large runoffs. Yet, that is where the frequency distribution of the runoff from different land uses increasingly diverges. When performance curves are plotted as a function of return frequency without distinguishing between different land uses, this information is lost.

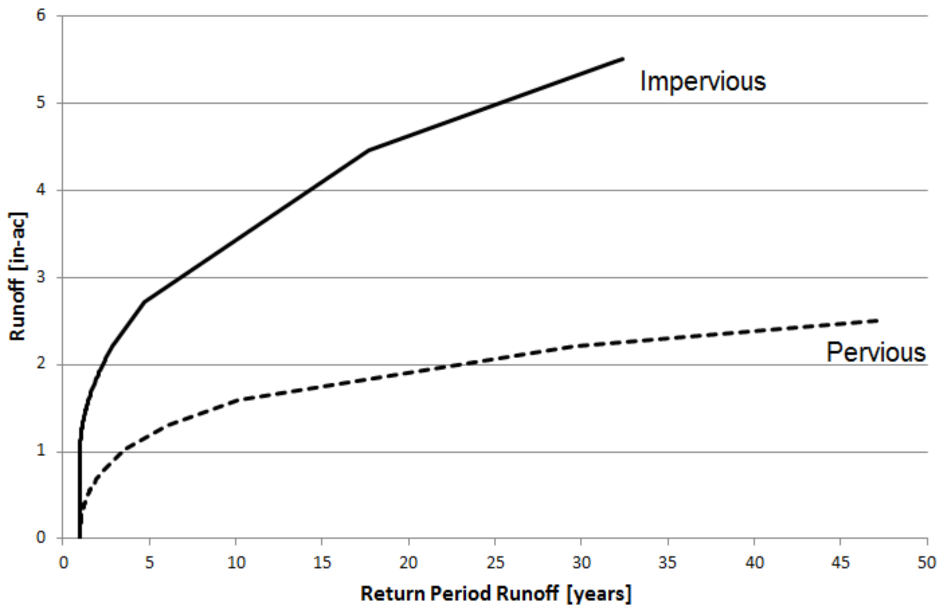


Figure 38: Difference in return period of stormwater runoff on pervious (rpd03) and impervious (rid) for developed land uses in the A24003 land segment in the Upper Patuxent watershed.

A performance curve as a function of the volume of runoff makes the performance curve independent of the land use and dependent on the size of the BMP. Figure 39 shows that the threshold in terms of runoff is much higher for BMPs on impervious land uses, because the BMPs there are optimized to hold more water than those on pervious land uses.[59] It is much easier to define the BMP effectiveness as a function of runoff and distinguish by BMP than accounting for all those different land uses (and land segments) by assigning them individual parameters for the MM curve.

[59] Note, for the performance curves as a function of return frequency, the threshold for BMPs on impervious surfaces is *lower* than for the ones on pervious surfaces. Contrarily, when the performance curves is plotted as a function of the volume of runoff, the threshold for BMPs on impervious surfaces is *higher* than for the ones on pervious land uses.

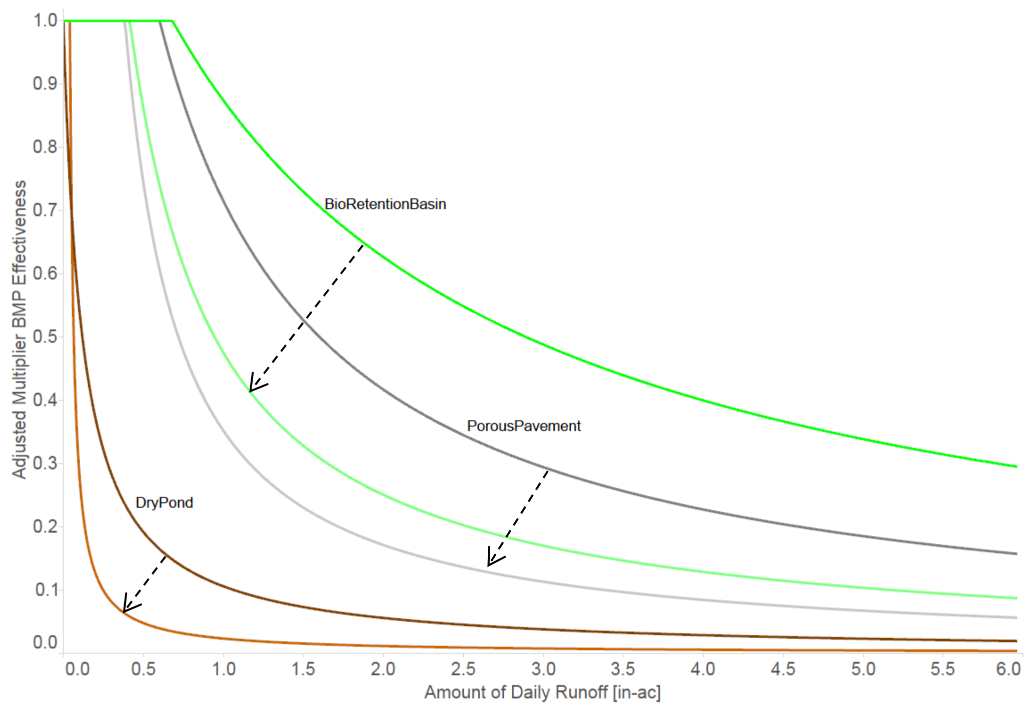


Figure 39: Revised MM curves for several BMPs plotted against the volume of runoff. The figure shows the MM curves for the regulated impervious and pervious (vague lines) developed land uses in land segment A24003.

5. Caveats and Discussion

5.1. Caveats

Several differences between the BMPs modeled in this study and those implemented in reality exist. First, BMPs are almost always implemented in treatment trains. For example, there will first be a sedimentation pond to let the sediment settle before the runoff travels on to a BMP that excels in removing other pollutants, e.g. a bio retention. This optimizes efficiency, increases retention time and makes BMP maintenance easier. With the exception of BMPs solely meant to decrease the peak discharge into the stormsewer system, stand-alone BMPs as modeled in SUSTAIN are uncommon. Second, in SUSTAIN, the BMPs were sized to remove certain amounts of pollutants to replicate the given annual average effectiveness (target values) exported from the Phase 5 Watershed model. It turned out that SUSTAIN could not replicate the given target values. Sediment was dominant for the dimensioning of all BMPs. Consequently, the BMPs' effectiveness for phosphorus and nitrogen was (much) more favorable than the target value. Either the assumed target values are not valid for these land segments and land uses, or the SUSTAIN model does not capture all natural processes inside the BMP well. Third, this study and the Phase 5 Watershed model use 24 hours as time step rather than storm event as is done in many monitoring studies.

5.2. Discussion

Climate change affects the effectiveness of BMPs in much more ways than just the changing amounts of runoff. The effectiveness of "green" BMPs, for example, depends on the vegetation inside the BMP that slow down the flow and take up some of the pollutants. With a changing climate, the type of plants in the BMPs might change and with them the BMP's effectiveness. Climate change might also affect how much water the soil can absorb, for example by changing soil properties such as the initial soil moisture content. BMPs depending on infiltration would then become more or less effective.

The BMP effectiveness in this study and the Phase 5 Watershed model is quantified as a percentage reduction of pollutant load (lbs.). It depends on the amount of inflow how many pounds of pollutant mass such a percent reduction equals to. However, the capacity of a BMP to remove pollutants is a function of mass. It would be much more realistic to quantify the effectiveness as the amount of removed pollutant mass in pounds per square foot or cubic foot of BMP, so that it becomes independent of the amount of inflow. Strecker et al. (2008) list 15 detailed reasons why percent removal rates should not be used. They additionally suggest to take into account the effect on peak discharge rate and total volume of discharge. After all, BMPs do not only remove pollutants but also prevent urban flood problems (Strecker et al., 2000). EPA suggests taking basing BMP effectiveness on three metrics: volume, total load and concentration (EPA, 2009). Even though the problems with using percent pollutant removal rates have been discussed and alternatives have been suggested in numerous newsletters, technical bulletins, journals and on websites (e.g., Singelis & Kosco, 2008; Geosyntec Consultants, 2012; Schueler, 2011, Jones et al., 2007; Bahr et al., 2012), the conventions in the modeling world have apparently not changed yet. An influential model such as the HSPF-based Chesapeake Bay Program Phase 5 Watershed model still uses percent removal rates for BMP effectiveness.

6. Conclusions

A number of findings and recommendations come forth from this study:

General

- The BMP effectiveness should be expressed as a function of the volume of runoff rather than the return frequency thereof, as is done in the CBP Phase 5 Watershed model. This captures the differences between pervious and impervious land uses and between land segments with different rainfall regimes much better. Additionally, data for storms occurring more often than once a year can then be included in the model which is currently lost due to mathematical constraints.
- The performance curves should at the very least be differentiated by BMP, ideally also by land use. The study clearly shows that each BMP performs differently under varying storm sizes and in different land segments and landuses. The fact that the rate at which the effectiveness decreases with larger amounts of runoff is a function of the BMP's average annual effectiveness complicates matters.
- Changing frequencies and intensities of storm events make it necessary to choose BMPs that are robust to those changes. The BMPs with the least steep decrease in effectiveness with

storm size are the infiltration-based BMPs. Pond-based BMPs fare less well. This is not surprising. Dry ponds, for example, were originally intended to reduce the *quantity* of stormwater flowing into the storm sewer, rather than improving the *quality* of the runoff.

Upper Patuxent Watershed/Phase 5 Watershed model

- BMP effectiveness drops sooner, deeper and on pervious land uses also steeper than assumed in the Phase 5 Watershed model. With the given target values, all BMPs except the infiltration trench are not maximally functioning for amounts of runoff occurring more often than once a year. Initially, the effectiveness drops off steeper than assumed in the Phase 5 Watershed model. But due to its convex character, the performance curve is less steep than assumed in the tail. There is very little uncertainty around the asymptote that the BMP effectiveness drops down to. For all BMPs, it is equal or close to 0% rather than 20% assumed in the Phase 5 Watershed model.
- The MM curve, which is the current default function in the CBP Phase 5 Watershed model leaves much room for improvement. Especially, the MM-curve does not fit pond-based BMP very well.
- There is a gap between the target values observed in the field and in the model. The BMPs in the model are much more efficient in removing nitrogen in phosphorus. It warrants further research, whether either the field monitoring and/or the model need improvement.

Acknowledgements: This paper was written during a summer associateship at the RAND Corporation, and is part of a project for the U.S. EPA Office of Research and Development. Jordan Fischbach and Edmundo Molina-Perez at the RAND Corporation advised the project. Gopal Bhatt, Gary Shenk, James Goodrich, Susan Julius and Tom Johnson at EPA provided valuable help along the way. At the RAND Corporation, Rob Lempert, Abdul Tariq, and David Johnson contributed advice. Thank you to all!

7. References

Ackerman, D., & Stein, E. D. (2008). Evaluating the effectiveness of best management practices using dynamic modeling. *Journal of Environmental Engineering*, *134*(8), 628-639.

Bahr, R. et al. (2012). *Recommendations of the Expert Panel to Define Removal Rates for Urban Stormwater Retrofit Projects.* URL [accessed 9/26/2013]: http://www.chesapeakebay.net/documents/Final_CBP_Approved_Expert_Panel_Report_on_Stormwater_Retrofits--_short.pdf

Barber, M.E. et al. (2003): "Ecology Ditch: A Best Management Practice for Storm Water Runoff Mitigation", *Journal of Hydrologic Engineering*, Vol. 8 No. 3, May/June 2003, p. 111-122.

Barrett, M. E., Walsh, P. M., Jr, J. F. M., & Charbeneau, R. J. (1998). Performance of vegetative controls for treating highway runoff. *Journal of environmental engineering*, *124*(11), 1121-1128.

Bobee, B. (1975). The log Pearson type 3 distribution and its application in hydrology. *Water resources research*, *11*(5), 681-689.

Carpenter, D. D., & Kaluvakolanu, P. (2010). Effect of roof surface type on storm-water runoff from full-scale roofs in a temperate climate. *Journal of Irrigation and Drainage Engineering*, *137*(3), 161-169.

Chesapeake Bay Program (2012): *Phase 5.3 Watershed Model.* Website. URL [accessed July 2014]: http://www.chesapeakebay.net/about/programs/modeling/53

Damodaram, C. (2010). *Simulating and Optimizing Storm Water Management Strategies in an Urban Watershed* (Doctoral dissertation, Texas A&M University).

Dietz, M. E. (2007). Low impact development practices: A review of current research and recommendations for future directions. *Water, air, and soil pollution*, *186*(1-4), 351-363.

Environmental Protection Agency (n.d.): *Chesapeake Bay TMDL. How does it work? Ensuring Results.* Website. URL [accessed 09/26/2013]: http://www.epa.gov/reg3wapd/tmdl/ChesapeakeBay/EnsuringResults.html?tab2=1

Environmental Protection Agency (2009): "Three Keys to BMP Performance – Concentration, Volume and Total Load", *National Pollutant Discharge Elimination System (NPDES)*, URL [accessed July 2014]: http://cfpub.epa.gov/npdes/stormwater/urbanbmp/bmptopic.cfm

Fisher, S. G., Welter, J., Schade, J., & Henry, J. (2001). Landscape challenges to ecosystem thinking: creative flood and drought in the American Southwest. *Scientia Marina*, *65*(S2), 181-192.

Geosyntec Consultants, W.W.E., (2012). "BMP Performance Summary: Chesapeake Bay and Related Areas." In *International Stormwater Best Management Practices (BMP) Database.*

Heitz, L. F., Khosrowpanah, S., & Nelson, J. (2000). "Sizing of surface water runoff detention ponds for water quality improvement." *JAWRA Journal of the American Water Resources Association* 36(3), 541-548.

Hunt, W. F., Hathaway, J. M., Winston, R. J., & Jadlocki, S. J. (2009). Runoff volume reduction by a level spreader–vegetated filter strip system in suburban Charlotte, NC. *Journal of Hydrologic Engineering*, *15*(6), 499-503.

Ice, G. (2004). History of innovative best management practice development and its role in addressing water quality limited waterbodies. *Journal of Environmental Engineering*, *130*(6), 684-689.

International Stormwater BMP Database (2013). Developed by Wright Water Engineers, Inc. and Geosyntec Consultants for the Water Environment Research Foundation (WERF), the American Society of Civil Engineers (ASCE)/Environmental and Water Resources Institute (EWRI), the American Public Works Association (APWA), the Federal Highway Administration (FHWA), and U.S. Environmental Protection Agency (EPA). URL [accessed July 2014]: http://www.bmpdatabase.org/index.htm

Jones, J., et al., (2008). "Reasons you should think twice before using percent removal to assess BMP performance." *Stormwater Magazine* January-February 2008. URL [accessed July 2014]: http://www.stormh2o.com/SW/Articles/15_Reasons_You_Should_Think_Twice_Before_Using_Per_16496.aspx

Jones, J.E., et al. (2004). "Why is it so difficult to acquire reliable BMP performance data." In *Proceedings of the 2004 World Water and Environmental Resources Congress: Critical Transitions in Water and Environmental Resources Management, Salt Lake City, UT*. 2004.

Lai, F. H., Dai, T., Zhen, J., Riverson, J., Alvi, K., & Shoemaker, L. (2007). SUSTAIN-AN EPA BMP process and placement tool for urban watersheds. *Proceedings of the Water Environment Federation*, *2007*(5), 946-968.

Li, H., Sharkey, L. J., Hunt, W. F., & Davis, A. P. (2009). Mitigation of impervious surface hydrology using bioretention in North Carolina and Maryland. *Journal of Hydrologic Engineering*, *14*(4), 407-415.

Li, Y., Lau, S. L., Kayhanian, M., & Stenstrom, M. K. (2006). Dynamic characteristics of particle size distribution in highway runoff: Implications for settling tank design. *Journal of Environmental Engineering*, *132*(8), 852-861.

Maryland, D.o.N.R., (2007). *Maryland Tributary Strategy Patuxent River Basin Summary Report for 1985-2005 Data*, Maryland, Department of Natural Resources.

Maryland, T.D.o.t.E., (n.d.). *Maryland's Stormwater Management Program. Why Stormwater Matters: Impacts of Runoff on Maryland's Watersheds*. URL [accessed July 2014]: http://www.mde.state.md.us/programs/Water/StormwaterManagementProgram/SedimentandStormwaterHome/Pages/programs/waterprograms/sedimentandstormwater/home/index.aspx.

Roesner, L. A., Bledsoe, B. P., & Brashear, R. W. (2001). Are best-management-practice criteria really environmentally friendly?. *Journal of Water Resources Planning and Management*, *127*(3), 150-154.

Simpson, Thomas & Weammert, Sarah (2009): *Developing Best Management Practice Definitions and Effectiveness Estimates for Nitrogen, Phosphorus and Sediment in the Chesapeake Bay Watershed. Final Report.* December 2009, University of Maryland Mid-Atlantic Water Program.

Singelis, N. & Kosco, J. (2008): "How to choose better stormwater BMPs", *ceNEWS*, July 2008, URL [accessed 9/26/2013]: http://www.cenews.com/magazine-article-cenews.com-7-2008-how_to_choose_better_stormwater_bmps-5880.html

Strecker, E. et al. (2000): "15 Reasons You Should Think Twice Before Using Percent Removal to Assess BMP Performance", *Stormwater,* January-February 2008. URL [accessed 9/26/2013]: http://www.stormh2o.com/SW/Articles/16496.aspx

Schueler, T.R. and R.A. Claytor, (2000). *Maryland Stormwater Design Manual.* Maryland Department of the Environment. Baltimore, MD.

Schueler , T.R. (2011): *Nutrient Accounting Methods to Document Local Stormwater Load Reductions in the Chesapeake Bay Watershed.* CSN Technical Bulletin No. 9, URL [accessed 9/26/2013]: http://chesapeakestormwater.net/2012/03/technical-bulletin-no-9-nutrient-accounting-methods-to-document-local-stormwater-load-reductions/

U.S. Environment Protection Agency Region 3 (2010). *Chesapeake Bay Total Maximum Daily Load For Nitrogen, Phosphorus and Sediment.*

Wurbs, R. A. (1997). *Computer Models for Water-Resources Planning and Management: National Study of Water Management During Drought.* DIANE Publishing, p. 183.

CONCLUSION

There is no doubt that adding uncertainty information to (river) forecasts will improve the product both technically and for decision-making. In the past, there have been cases, when decision-makers were not sufficiently aware of the uncertainty in forecasts. The most prominent example is the devastating flood in Grand Forks, ND in 1997 (Pielke, 1999). Precipitation forecasts, used by the general public, are already probabilistic. Morss et al. (2010) has shown that people are capable of interpreting those correctly.

In my thesis, I find that emergency managers are usually aware of the uncertainty in forecasts. They do not perceive this uncertainty as a major problem. Instead, emergency managers worry much more about the cascading events of critical situations. For example, failing air conditioning in a shelter during a heat wave or a skateboard stuck in a stormsewer during heavy rainfall make a tense situation a crisis.

Work experience is the best predictor of how much emergency managers rely on forecasts and recorded weather data. Therefore, it must have a benefit for them. At least partially, it is a lack of early training that emergency managers learn of the benefits of forecasts on the job. I found some evidence that subjective numeracy influences of how useful emergency managers find the forecast in terms of accuracy, timeliness, etc. They then project that attitude onto the people around them. Work experience in turn seems to teach emergency managers the benefits of taking into account the the opinions of their emergency management and city hall colleagues. Peer-education seems to be a promising option to improve decision-making.

Emergency managers cope with the uncertainty in forecasts by collecting large amounts of information and using weather information, such as radar, that are less uncertain. While the forecast is only one of many pieces of information, it does serve as a standard that is also communicated to the public. The emergency managers may add local interpretation and instructions to the forecast, but they are reluctant to publish their judgment of the forecast uncertainty. They are concerned about what happens when their judgment is wrong. The history of law cases shows that it is unlikely that an emergency manager could (successfully) be sued for interpreting forecast uncertainty incorrectly. However, emergency managers are either not aware of this or are worried about the criticsm in tidely knit communities.

I conclude that forecast uncertainty is only likely to be communicated to the broader public, if it is included in the official forecast. This is important, because it is not the emergency manager who actually carries out preventive measures. Instead, the public and the city departments take action, e.g., evacuate, close off areas, and secure property.

While probabilisitic forecasts are beneficial, it is illusionary to think that they will automatically lead to better decision-making. People are often caught off guard when a crisis develops differently than those in the past. Heuristics like the anchoring, representativeness and availability heuristic help people making decisions in a complex world, but often serve them poorly in extreme situations (Tversky, Kahneman, 1974). This will remain the case with probabilistic forecasts. Therefore, emergency managers need to be trained in a “boots-on-the-ground” manner not only how to read probabilistic forecasts, but also how to base decisions one them. Then some of the mathematical benefits as computed by Verkade and Werner (2011) have a chance to materialize. I find that currently decision-making based on weather forecasts is insufficiently part of the standard curriculum for emergency managers. It also needs to be considered how the end users of any forecasts, the general public, will use probabilistic forecasts, when asked to heed the advice of authorities in critical situations. While I have not studied this in my research yet, I think, that it makes a difference whether the uncertainty is communicated as a probability of exceeding some threshold, as confidence bounds, or average forecast error. Last, attention has to be paid to accountability issues when substituting deterministic forecasts with probabilistic ones.

Turning to the technical aspects of river forecasts, I found that emergency managers perceive the inaccuracy of and frequent changes to event magnitude and timing as the most important obstacle for relying on forecasts (rather than their ability to interpret the forecast or the timely availability for the area). Probabilistic forecast thus really do address a problem. My experiments with quantile regression confirm for the U.S. American context, that probabilistic forecasts of good quality can be generated fairly easily by post-processing published forecasts. A major benefit of post-processing forecasts is that it captures uncertainty from all sources: exogenous and endogenous to the forecast model. Thus, an expensive overhaul of the entire forecasting system is unnecessary. Solutions as small as a smartphone application are conceivable, because the computational load is low. I found that a single model is sufficient to predict the probability of exceeding water levels, except extremely high ones. For flood stages, the model needs to be customized to the river gage in question.

There is still much room for improving the quantile regression approach to probabilistic river forecasts. The model does not perform well for lower water levels, and its performance is sensitive to river gage, forecast year, lead time, and event thresholds. River gages close to river confluences and far upstream are especially difficult to predict. Including more variables such as observed and forecast precipitation is an obvious way to improve the model. However, some variables are not easily combined in the linear version of the quantile regression model.

For a user of probabilistic forecasts it is difficult to base decisions on it, if the forecast performance is so senstivitive to the context. It is therefore necessary, that the performance record of those forecasts is published alongside them; i.e., the uncertainty of probabilistic forecasts needs to be quantified as well.

References

Pielke, R.A. (1999): "Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood", *Applied Behavioral Science Review*, 7(2), pp.83–101.

Verkade, J. S., & Werner, M. G. F. (2011). Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology & Earth System Sciences Discussions*, *8*(4).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Morss, R.E., Lazo, J.K., Demuth, J.L. (2010): "Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication." *Meteorological Applications* Vol. 17: pp. 149-162.

Appendix – Study 2

1. Questionnaire

In the following, the survey questions for each component of the Theory of Planned Behavior are listed. For the sake of brevity, some of the survey infrastructure, e.g., the introduction text, has been omitted.[60] The question texts are indicated by cursive font.

1.1. Hazard Type

The questionnaire starts off by making the participant choose one type of hazard. (S)he answers the questions for the chosen hazard.
Which of the following hazards did you find most difficult to respond to as an emergency manager in the past 10 years (2004-2014)? (Select one answer.)

- Flash flood
- River flood
- Tornado
- Hurricane
- Snow and Ice Storm
- Heat Wave
- None of the above[61]

In which year, did this event occur?
2014-2000 (drop-down box)

In the following, [event] will denote the combination of selected weather event and year. For example if "Heat Wave" and "2009" have been selected in the above questions: [event] = the 2009 Heat Wave

1.2. Intention and Past Behavior

The TPB components "Intention" and "Past Behavior" have the same answer options, but different questions.

Past behavior: *During the [event], how much did you rely on National Weather Service data to make any of the following decisions or carry out any of the following actions?*

Intentions: *If an event like the[event] happened again in the near future, how much do you think would you rely on National Weather Service data to make any of the following decisions or carry out any of the following actions?*

[60] Please contact the author for a copy of the full questionnaire.
[61] If "None of the above" was selected, the participant was not shown the sections "Past Behavior & Intention" and "Perceived Control".

Please answer the questions for two types of data from the National Weather Service (NWS):

1. *Recorded weather data from the past few hours/days (observed conditions).*
2. *Short-term weather forecasts for the next few hours/days*

The answer options are displayed in Table 22.

If you could only have one, which type of data from the National Weather Service would you prefer to have when responding to events like [event]?

- *Recorded weather data (Past few hours/days)*
- *Short-term weather forecasts (Newt few hours/days)*

1.3. Perceived Control

Please rate how much the following factors have limited your reliance on National Weather Service data when responding to events like the [event].

Please answer the questions for two types of data from the National Weather Service (NWS):

1. *Recorded weather data from the past few hours/days (observed conditions).*
2. *Short-term weather forecasts for the next few hours/days*

The answer options are displayed in Table 23.

When thinking of responding to an event like the [event], how satisfied are you with the data from the National Weather Service?

	Extremely satisfied	**Very satisfied**	**Somewhat satisfied**	**Slightly satisfied**	**Not at all satisfied**
Recorded weather data (Past few hours/days)					
Short-term weather forecasts (Next few hours/days)					

1.4. Social Norms

Please rate how much the following groups expect you to rely on the National Weather Service data when responding to an event like the [event]?

Please answer the questions for two types of data from the National Weather Service (NWS):

1. *Recorded weather data from the past few hours/days (observed conditions).*
2. *Short-term weather forecasts for the next few hours/days*

The answer options are displayed in Table 24.

Please rate how much you worry about criticism from the following groups when responding to an event like [event].

Please answer the questions for two types of data from the National Weather Service (NWS):

3. *Recorded weather data from the past few hours/days (observed conditions).*
4. *Short-term weather forecasts for the next few hours/days*

	Does not apply	**Very much - 1**	**2**	**Some-What - 3**	**4**	**Not at all - 5**
Local and regional emergency management colleagues						
City employees (for example, city departments, fire, police, EMS)						
Elected officials (for example, mayor, city council)						
Employees of the National Weather Service						
Employees of government agencies (for example, Army Corps of Engineers, FEMA)						
Training/workshop instructors						
Residents						
Friends & Family						
The media						
The general public						

As an emergency manager, to which degree do you agree or disagree with the following statements?

	Strongly Disagree	**Disagree**	**Undecided**	**Agree**	**Strongly Agree**
In general, taking into account the expectations of others is useful when responding to an event.					
I worry about potential or possible liability claims made against me or my employer when responding to an event.					

Table 22: Answer options for "Past Behavior" and "Intention" questions.

	Recorded weather data (Past few hours/days)						**Short-term weather forecasts** (Next few hours/days)					
	Does not apply	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**	**Does not apply**	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**
Determine where and when to deploy (storm) spotters												
Determine when notify other first responder or when to activate Emergency Operations Center												
Determine likely impact location of the event												
Decide when to warn public												
Track the progress of the event (for example, to protect first responders and predict damage)												
When to initiate evacuation												
When to open shelters												

Table 23: Answer options for main "Perceived Control" question.

	Recorded weather data (Past few hours/days)						**Short-term weather forecasts** (Next few hours/days)					
	Does not apply	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**	**Does not apply**	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**
Receiving information too late to be useful												
Information unavailable in your area												
Insufficient/irrelevant information												
Irregular information release times												
Inaccurate forecast of event magnitude	x											
Inaccurate forecast of event timing	x											
Frequent changes to forecast of event magnitude	x											
Frequent changes to forecast of event timing	x											
Your ability to understand the information												
Your ability to apply the information to the emergency response												

1.5. Attitude

Thinking of responding to a weather-related crisis to what extent do you agree or disagree with the following statements?

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
It is my job to prevent rather than to respond to hazardous situations.					
It is my job to protect rather than to rescue citizens.					
I pay more attention to the current situation and less to possible sequences of events when making a plan to respond to an event.					
Not knowing what will happen during an event makes it difficult for me to respond.					
I routinely think in what-if scenarios when responding to events.					
A critical situation can develop in so many different ways, it is difficult to determine appropriate actions.					

Which statement best characterizes your response to weather-related crises? (Select one.)

- *When faced with a critical situation, I consider all facts, figures, and different scenarios and weigh my options before I take action.*
- *When faced with a critical situation, I know what to do when I see what is going on around me, and hear what is happening.*

Table 24: Answer options for "Social Norms" question.

	Recorded weather data (Past few hours/days)						**Short-term weather forecasts** (Next few hours/days)					
	Does not apply	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**	**Does not apply**	**Very much 1**	**2**	**Some what 3**	**4**	**Not at all 5**
Local and regional emergency management colleagues												
City employees (for example, city departments, fire, police, EMS)												
Elected officials (for example, mayor, city council)												
Employees of the National Weather Service												
Employees of government agencies (for example, Army Corps of Engineers, FEMA)												
Training/workshop instructors												
Residents												
Friends & Family												
The media												
The general public												

Table 25: Answer options for "Attitude" component.

	Recorded weather data (Past few hours/days)					Short-term weather forecasts (Next few hours/days)				
	Strongly Disagree	**Disagree**	**Un-decided**	**Agree**	**Strongly Agree**	**Strongly Disagree**	**Disagree**	**Un-decided**	**Agree**	**Strongly Agree**
Using NWS data to respond to events has resulted in better decisions/actions.										
NWS data is the most important source of information for responding to events.										
Relying to NWS data to respond to events can be harmful.										
I would recommend other emergency managers to rely on NWS data to respond to events.										
NWS data gives me confidence in my decisions/actions during responses to events.										

Thinking of an event like [event], to what extent do you agree with the following statements?

Please answer the questions for two types of data from the National Weather Service (NWS):

1. *Recorded weather data from the past few hours/days (observed conditions).*
2. *Short-term weather forecasts for the next few hours/days*

The answer options are displayed in Table 25.

1.6. Subjective Numeracy

The Subjective Numeracy Scale has been developed by Fagerlin et al. (2007).

For each of the following questions, please check the box that best reflects how good you are at doing the following things:

	Not at all good 1	**2**	**3**	**4**	**5**	**Extremely good 6**
How good are you at working with fractions?						
How good are you at working with percentages?						
How good are you at calculating a 15% tip?						
How good are you at figuring out how much a shirt will cost, if it is 25% off?						

When reading the newspaper, how helpful do you find tables and graphs that are parts of the story?

Not at all helpful 1	**2**	**3**	**4**	**5**	**Extremely helpful 6**

When people tell you the chance of something happening, do you prefer that they use words (for example, "it rarely happens") or numbers (for example, "there is a 1% chance")?

Always prefer words 1	**2**	**3**	**4**	**5**	**Always prefer numbers 6**

When you hear a weather forecast, do you prefer predictions using percentages (for example, "there will be a 20% chance of rain today") or predictions using only words (for example, "there is a small chance of rain today")?

Always prefer words 1	**2**	**3**	**4**	**5**	**Always prefer percentages 6**

How often do you find numerical information to be useful in your daily life?

Never 1	2	3	4	5	Very often 6

1.7. Demographics

The following demographic information has been collected at the end of the survey:

- Professional or volunteer position as emergency manager
- First-responder background (e.g., policeman, fire fighter or paramedic)
- Paid/unpaid emergency management position
- Length of experience as emergency manager

- Type jurisdiction (e.g., county, municipality, tribe)
- Size jurisdiction (number of residents)
- State

- General level education
- Type of emergency management education
- Received instructions for use of forecasts
- Gender
- Age

2. Principal Component Analysis – Loadings

2.1. Dependent Variables

Table 26: Component loadings for the four dependent variables.

	Past Behavior Recorded weather data	Past Behavior Forecasts	Intentions Recorded weather data	Intentions Forecasts
Determine where and when to deploy (storm) spotters	-0.24	-0.17	-0.23	-0.19
Determine when notify other first responder or when to activate Emergency Operations Center	-0.41	-0.43	-0.41	-0.43
Determine likely impact location of the event	-0.42	-0.43	-0.41	-0.42
Decide when to warn public	-0.44	-0.44	-0.43	-0.42
Track the progress of the event (for example, to protect first responders and predict damage)	-0.41	-0.39	-0.39	0-41
When to initiate evacuation	-0.34	-0.33	-0.35	-0.35
When to open shelters	-0.35	-0.38	-0.38	-0.38
Proportion of Variance	53.9%	44.1%	59.1%	55.1%

2.2. Perceived Limitations

Table 27: Component loadings for perceived limitations.

	Recorded weather data			Short-term weather forecast		
	Component 1	**Component 2**	**Component 3**	**Component 1**	**Component 2**	**Component 3**
Inaccurate forecast of event magnitude	-0.35	**0.35**	**0.14**	-0.34	**0.33**	
Frequent changes to forecast of event magnitude	-0.34	**0.38**		-0.34	**0.35**	-0.26
Inaccurate forecast of event timing	-0.36	**0.32**	**0.13**	-0.36	**0.30**	
Frequent changes to forecast of event timing	-0.35	**0.34**		-0.35	**0.35**	-0.23
Receiving information too late to be useful	-0.28	-0.32	**0.18**	-0.29	-0.21	**0.32**
Information unavailable for your area	-0.27	-0.37	**0.24**	-0.31	-0.20	**0.37**
Insufficient/irrelevant information	-0.31	-0.35	**0.24**	-0.32	-0.19	**0.39**
Irregular information release times	-0.29	-0.30	**0.27**	-0.36	0.12	**0.24**
Your ability to apply the information to the emergency response	-0.29	-0.19	-0.19	-0.27	-0.42	-0.50
Your ability to understand the information	-0.27	-0.20	-0.20	-0.26	-0.50	-0.42
Proportion of Variance	46.6%	24.1%	11.8%	55.5%	14.9%	10.1%
Cumulative Proportion	46.6%	69.7%	81.5%	55.5%	70.4%	80.5%

2.3. Social Norms

Table 28: Component loadings for social norms.

	Recorded weather data		Short-term weather forecast	
	Component 1	**Component 2**	**Component 1**	**Component 2**
Local and regional emergency management colleagues	-0.25	**0.37**	-0.23	**0.35**
City employees	-0.32	**0.14**	-0.32	**0.16**
Elected officials	-0.37	-0.11	-0.36	-0.13
NWS employees	-0.22	**0.51**	-0.22	**0.51**
Employees of government agencies	-0.33	**0.26**	-0.33	**0.30**
Training/workshop instructors	-0.22	**0.42**	-0.21	**0.43**
Residents	-0.38	-0.36	-0.39	-0.35
The Media	-0.33	-0.16	-0.32	-0.19
The general public	-0.37	-0.36	-0.38	-0.34
Family & Friends	-0.33	-0.22	-0.34	-0.18
Proportion of Variance	44.6%	13.7%	42.9%	14.5%
Cumulative Proportion	44.6%	58.3%	42.9%	57.5%

2.4. Job Attitude

Table 29: Component loadings for job attitude.

	Component 1	Component 2
It is my job to prevent rather than to respond to hazardous situations.	0.50	**0.47**
It is my job to protect rather than to rescue citizens.	0.47	**0.46**
(-) I pay more attention to the current situation and less to possible sequences of events when making a plan to respond to an event.	0.26	-0.31
Intuitive vs. Analytical	-0.39	
(-) Not knowing what will happen during an event makes it difficult for me to respond.	0.39	-0.51
(-) A critical situation can develop in so many different ways, it is difficult to determine appropriate actions.	0.41	-0.47
Proportion of Variance	31.1%	20.1%
Cumulative Proportion	31.1%	51.3%

2.5. Attitude Weather Data

Table 30: Component loadings for attitude towards weather data.

%	Recorded weather data	Short-term weather forecast
Using NWS data to respond to events has resulted in better decisions/actions.	-0.47	-0.49
NWS data is the most important source of information for responding to events.	-0.45	-0.45
I would recommend other emergency managers to rely on NWS data to respond to events.	-0.53	-0.53
NWS data gives me confidence in my decisions/actions during responses to events.	-0.54	-0.53
Proportion of Variance	69.2%	67.6%

2.6. Subjective Numeracy

Table 31: Component loadings for subjective numeracy.

	Component 1	Component 2
1. How good are you at figuring out how much a shirt will cost, if it is 25% off?	-0.43	-0.33
2. How good are you at working with percentages?	-0.45	-0.12
3. How good are you at calculating a 15% tip?	-0.44	-0.30
4. How good are you at working with fractions?	-0.39	-0.19
5. When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?	-0.28	**0.20**
6. When people tell you the chance of something happening, do you prefer that they use words (for example, "it rarely happens") or numbers (for example, "there's a 1% chance")?	-0.23	**0.60**
7. When you hear a weather forecast, do you prefer predictions using percentages (for example, “there will be a 20% chance of rain today”) or predictions using only words (e.g., “there is a small chance of rain today”)?	-0.28	**0.54**
8. How often do you find numerical information to be useful in your daily life?	-0.25	**0.26**
Proportion of Variance	45.1%	18.5%
Cumulative Proportion	45.1%	63.6%

Appendix – Study 4

1. Formulae used to compute pollutant loads in surface runoff

For the years 1984-2005, the hourly continuous storm conditions for land segments A24003, A24027, A24031 and A24033 and the corresponding runoff for four land uses (rid, rpd, nid, npd)[62] in the Patuxent watershed were extracted from the CBP Phase 5 Watershed model. From those times series the surface outflow volume (SURO, [in-acre/time step]), the groundwater recharge volume (AGWO, [in-acre\time-step]) and the sediment load (SEDM, [tons/time step]) were directly used. The other pollutant loads were computed from the output with the following formulae:

- Total Nitrogen:

  $TOTN = NH_3D + NH_3A + NH_3I + NH_3C + NO_3D + RORN + BODA \times 0.0436 + PHYT \times 0.0863$

  $TOTN = (SNH_3+INH_3+ANH_3) + (0) + (DNH_3 \times 0.70) + (DNH_3 \times 0.30) + (SNO_3+INO_3+ANO_3) + (DRON+SRON+IRON+ARON) + ((DLON+SLON+ILON+ALON) \times 22.95) \times 0.0436 + (0) \times 0.0863$

- Total Phosphorus:

  $TOTP = PO_4D + PO_4A + PO_4I + PO_4C + RORP + BODA \times 0.00603 + PHYT \times 0.0119$

  $TOTP = (SPO_4+IPO_4+APO_4) + (0) + (DPO_4 \times 0.70) + (DPO_4 \times 0.30) + ((DRON+SRON+IRON+ARON) \times 0.01384) + ((DLON+SLON+ILON+ALON) \times 22.95) \times 0.00603 + (0) \times 0.0119$

[62] Land use definitions: rid – regulated impervious developed, nid – non-regulated impervious developed, rpd – regulated pervious developed, npd – non-regulated impervious developed. Impervious developed areas are roads, roofs, pavement, etc. Pervious developed areas are, for example, lawns. If a land use is classified as regulated, it is subject to the National Pollution Discharge Elimination System (NPDES) permits regulating stormwater discharges.

The gray components of the formulae were omitted, because only surface runoff was taken into consideration.

These five components (SURO, AGWO, SEDM, TN, TP) were combined in one input file for SUSTAIN. In total 16 input files (4 land uses for 4 land segments) were created for each BMP.

2. Derived Parameters for revised Michaelis-Mention curve

The MM curve was fit to subset of the dataset, stratified by BMP, pollutant, land use and land segment. To facilitate curve fitting, it was desirable to have as many data points as possible for each fit. Therefore, pervious (rpd, npd) and impervious (nid, rid) have been lumped together. Additionally, the data for phosphorus and nitrogen removal was combined. Compared to sediment, those two pollutants undergo rather similar processes. Table 32 summarizes the resulting parameter ranges. The default value is the average of the parameter across the impervious land uses.[63]

[63] Note that these parameters cannot be used in the Hydro-Method in the current CBP Phase 5 Watershed model, because in this case the MM function describes a *re-scaled* multiplier for BMP effectiveness as a function of the volume of runoff rather than return frequency.

Table 32: Derived parameters for revised Michaelis-Menten curve. The average is based on impervious landuses only. The minimum and maximum in brackets are based on impervious and pervious land uses.

	Number of curves		Average number of data points per curve	Average RMSE	Parameters for MM function: Default; [Low, High]		
BMP	**Imp***	**Perv***			**Asymp**	**Thres**	**Half Sat**
Bioretention	6	4	6,372	0.05	0.0; [0.0, 0.0]	0.6; [0.4, 0.7]	3.1; [0.9, 3.4]
Dry Pond	5	4	6,228	0.17	0.0; [0.0, 0.0]	0.0; [0.0, 0.1]	0.2; [0.1, 0.2]
Ext Dry Pond	8	8	5,634	0.11	0.0; [0.0, 0.0]	0.3; [0.0, 0.5]	1.9; [0.4, 3.6]
Infiltration Trench	8	0	7,863	0.02	0.0; [0.0, 0.1]	1.8; [1.6, 2.1]	3.6; [3.2, 4.1]
Porous Pavement	6	4	6,373	0.06	0.0; [0.0, 0.0]	0.5; [0.4, 0.6]	1.8; [0.6, 2.3]
Vegetated Swale	6	8	5,493	0.08	0.0; [0.0, 0.1]	0.1; [0.0, 0.1]	3.1; [0.8, 4.4]
Wet Pond	8	4	6,360	0.17	0.0; [0.0, 0.1]	0.1; [0.0, 0.3]	1.7; [0.7, 2.1]

*The maximum number is 8 curves per land use. If the number is less than eight, it was not possible to fit a curve to do either data sparsity (esp. in pervious land uses) or form constraints of the MM-curve (esp. in impervious land uses).