UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MACHINE LEARNING PREDICTIONS OF FLASH FLOODS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ROBERT ALLAN CLARK III
Norman, Oklahoma
2016

MACHINE LEARNING PREDICTIONS OF FLASH FLOODS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

_____
Dr. Robert Palmer, Chair

_____
Dr. Jonathan Gourley, Co-Chair

_____
Dr. Jeffrey Basara

_____
Dr. Yang Hong

_____
Dr. Mark Morrissey

_____
Dr. Randa Shehab

"Machines take me by surprise with great frequency."

Turing, A. M., 1950. Computing machinery and intelligence. Mind*, 49, 433-460.*

Acknowledgements

I would like to thank my academic advisors, Dr. Jonathan J. Gourley and Dr. Yang Hong, for believing in me and taking the unusual step of hiring an Oklahoma State University chemical engineering graduate as a budding meteorologist and Graduate Research Assistant way back in 2010. I thank the late director of the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS), Dr. Peter Lamb, for his decision to fund my research, and I thank his successor, Dr. Randy Peppler, and the CIMMS Executive Director of Operations and Finance, Ms. Tracy Reinke, for their decision to continue this funding through the completion of my degree. Dr. Kim Elmore of CIMMS engaged in stimulating conversations about the worlds of artificial intelligence and machine learning. Dr. Pierre-Emmanuel Kirstetter of the Advanced Radar Research Center provided invaluable career advice throughout my degree and was always there to answer statistics questions. Dr. Humberto Vergara of CIMMS acted as a frequent devil's advocate and made sure I always knew the exact implications of the assumptions and assertions I made at conferences, in papers, and at group meetings. Finally, I thank Dr. Zac Flamig for his willingness to share tips on Python syntax, good programming practices, and visionary ideas.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation contains a literature review and three studies concerned with the development, assessment, and use of machine learning (ML) algorithms to explore automatically generated predictions of flash floods. The literature review explores several relevant issues: how flash floods are defined, the organization and structure of the flash flood forecasting and alerting enterprise in the U.S., proposed methods and tools for understanding and forecasting flash floods, the statistical underpinnings of ML, and how ML techniques can be applied to a wide variety of complex scientific problems, including those of a meteorological bent.

Using an archive of numerical weather predictions (NWP) from the Global Forecast System (GFS) model and a historical archive of reports of flash floods across the U.S., I develop a set of machine learning models that output forecasts of the probability of receiving a *Storm Data* report of a flash flood given a certain set of atmospheric and hydrologic conditions as forecast by the GFS model. I explore the statistical characteristics of these predictions, including their skill, across various regions and time periods. Then I expound upon how various atmospheric fields affect the probability of receiving a report of a flash flood and discuss different methods for interpreting the results from the proposed ML models. Finally, I explore how the mooted system could be operationalized, by delving into two case studies of past impactful flash floods in the U.S., by presenting results of National Weather Service forecasters using and interacting with the proposed tools in a research-to-operations testbed environment, and by geographically extending the predictions to cover additional parts of the world's landmass via a set of case studies on the European continent.

One ML algorithm in particular, the random forest technique, is used throughout the vast majority of the dissertation, because it is quite successful at incorporating large amounts of information in a computationally-efficient manner and because it results in reasonably skillful predictions. The system is largely successful at identifying flash floods resulting from synoptically-forced events, but struggles with isolated flash floods that arise as a result of weather systems largely unresolvable by the coarse resolution of a global NWP system. The results from this collection of studies suggest that automatic probabilistic predictions of flash floods are a plausible way forward in operational forecasting, but that future research could focus upon applying these methods to finer-scale NWP guidance, to NWP ensembles, to new regions of the world, and to longer forecast lead times.

Chapter 1: Introduction

The U.S. National Weather Service (NWS) Glossary (2009) defines a "flash flood" with the following statement:

> A rapid and extreme flow of high water into a normally dry area, or a rapid water level rise in a stream or creek above a predetermined flood level, beginning within six hours of the causative event (e.g., intense rainfall, dam failure, ice jam). However, the actual time threshold may vary in different parts of the country. Ongoing flooding can intensify to flash flooding in cases where intense rainfall results in a rapid surge of rising flood waters.

This definition is not universally accepted in the scientific literature (e.g. Gaume et al. 2009, Braud et al. 2014), but serves as the starting point for the organization of operational flash flood forecasting and monitoring in the U.S. Flash floods are among the deadliest storm-related hazards in the United States and around the world from year-to-year. Ashley and Ashley (2008), analyzing NWS data, found that floods, regardless their cause, were the deadliest storm-related hazard in the U.S. through the decade ending in 2006. Like other hazardous weather phenomena, the flash flood forecasting enterprise requires a team of highly-trained meteorologists and support personnel with disparate responsibilities working together in close collaboration. Though flash floods are often largely caused by meteorological conditions, they are not solely meteorological. Doswell et al. (1996) stated that flash floods arise from a combination of two factors: heavy rainfall and hydrologic response. Therefore, success in forecasting flash floods requires both meteorological and hydrologic knowledge.

In some countries, hydrologic and meteorological services exist in separate silos. In the U.S., however, hydrometeorological hazards are intended to be brought to public attention by one agency of the federal government – the NWS. Their forecasting and alerting enterprise is mostly staffed by meteorologists, with hydrologists making up a

smaller percentage of the overall effort. One can view this enterprise as being organized along the "forecast funnel" approach (Snellman 1982), where the portion of the funnel focused on longer spatial and temporal scales is the purview of the meteorologists at the NWS Weather Prediction Center (WPC), part of the National Centers for Environmental Prediction (NCEP). WPC conducts long-range diagnosis of automated weather forecast guidance, provides 0-72 h (0-3 d) forecasts of heavy or "excessive" rainfall, and generates 0-168 h (0-7 d) quantitative precipitation forecasts (QPF). The next, finer resolution portion of the forecast funnel is the responsibility of the 13 regional River Forecast Centers (RFCs) that combine to cover all 50 states and the U.S. territories. RFCs are primarily staffed by hydrologists; they mostly focus on riverine flooding with two important exceptions: RFCs are responsible for producing Flash Flood Guidance (FFG, Clark et al. 2014) and collecting and editing the data sent to NCEP for use in the Stage IV quantitative precipitation estimates (QPEs, Lin and Mitchell 2005). At a local level (and at the finest spatial and temporal scales), the flash flood alerting enterprise is administered by 122 NWS Weather Forecast Offices (WFOs), staffed mainly by meteorologists who issue specific point-based forecasts for their local areas of responsibility. These point forecasts include probabilities of precipitation and QPFs. WFO meteorologists are also responsible for issuing Flash Flood Watches (FFAs) when there is a 50 to 80 percent chance of flooding conditions in the next 48 hours (two days, Clark 2011), and more urgently, they issue flash flood warnings (FFWs) when flooding is "imminent or likely" over a period generally less than six hours in length but up to 12 hours in length depending on the circumstance.

Clark et al. (2014) outline the history and assess the skill of the primary suite of tools used to issue FFWs – FFG. FFG is defined as the amount of precipitation required in a given time period to induce bankfull flows on small natural stream networks. When QPE from that same time period begins to approach or exceed the FFG value, a flash flood may be imminent, though Clark et al. (2014) determined that FFG is a more skillful tool when the QPE-to-FFG ratio is 1.25 or 1.5, instead of 1.0, depending on the region of the U.S. under consideration. Importantly, FFG is a monitoring tool and does *not* include any hydrologic or meteorological forecast component. Despite this noteworthy limitation, the lack of flash flood *forecasting* tools has resulted in a situation in which the original definitions of FFG have been stretched to accommodate new uses, including WPC's products identifying the probability of QPF exceeding FFG (Barthold et al. 2015). Due to the advanced age of the FFG concept, its relatively low critical success index when used to predict NWS *Storm Data* flash flood reports (< 0.2, Clark et al. 2014, Gourley et al. 2012), recent improvements in radar-derived QPE like the National Severe Storms Laboratory's (NSSL) Multi-Radar Multi-Sensor QPE project (Zhang et al. 2016), and advancements in high-resolution distributed hydrologic models (DHMs, Clark et al. 2016), FFG is slated to be augmented in NWS operations by NSSL's Flooded Locations and Simulated Hydrographs (FLASH) suite of forecasting and monitoring tools in 2016 (Gourley et al. 2016). Other major research efforts in this area include the development and proposed implementation of the NWS's National Water Model, formerly known as the Weather Research and Forecasting Model Hydrological modeling extension package (Gochis et al. 2014), which, like FLASH, will be capable of forcing a high-resolution

DHM with precipitation forecasts from convection-allowing models (Barthold et al. 2015).

Despite these important advancements, there has been a recent lack of research into how to automatically forecast combinations of hydrologic and synoptic-scale meteorological environments favorable for the outbreak of flash floods. In 1979, Maddox et al. categorized flash floods into four different categories, based upon the locations of the events and the environments in which they developed. Doswell et al. (1996) developed an "ingredients-based methodology", based on physical understanding, for forecasting heavy precipitation. In the 20 years since the publication of their methodology, of course, the physical principles underlying the development and maintenance of heavy precipitation have not changed, but numerical weather prediction (NWP) models and the QPF generated from them has advanced substantially in resolution and skill.

These advanced NWP models have dramatically enlarged the amount of information available to weather forecasters. Traditionally, recognizing patterns or relationships present in these vast amounts of data was a manual problem. The human brain is capable identifying patterns or relationships present in a few dozens to a few hundred examples arising from any sort of system. However, as the complexity of the system being studied increases, the number of examples any one person can aggregate and analyze effectively decreases, so people are trained to reduce the effective complexity of the system via rules-of-thumb, heuristics, dimensional reduction, and other techniques. Weather forecasting is an example of a highly complex system where experts are trained to reduce complex mathematical relationships and physical laws into simpler

procedures, like checklists, cookbooks, or indexes. For example, indexes are empirically-derived mathematical relationships that might relate several meteorological variables to the likelihood of a particular weather outcome (Doswell and Schultz 2006). Typically, a meteorologist uses some combination of physical understanding, training/experience, and empirical relationships (indexes, model output statistics [MOS], and others) to forecast the weather.

Computers have often been used to generate the statistical relationships necessary to come up with forecast indexes, MOS, and other tools, by directly programming the empirical relationships uncovered via analysis of several dozen or a few hundred cases. As weather forecasting matured, computer scientists were independently developing a set of techniques known as "machine learning" (ML), which is a subset of the broader field of artificial intelligence. Rather that requiring a computer be directly programmed with a checklist or a set of rules-of-thumb, ML algorithms "learn" and evolve over time, by iterating through vast amounts of data thousands, millions, or hundreds of millions of times, often identifying patterns that elude the traditional methods outlined above (Kohavi and Provost 1998). In other words, physical understanding guides the collection of data used in the ML context, but ML is often capable of identifying patterns and relationships that would not have been readibly identifiable solely via physical principles.

ML tasks can be either "supervised" or "unsupervised"; in the former, the dependent attribute, or label, is provided as part of the dataset fed to the algorithm, while in the latter, the label is not specified as part of the dataset (Kohavi and Provost 1998). When the ML algorithm is explicitly told what it is supposed to predict, the learning is supervised. When the ML algorithm is allowed to cluster cases drawn from the data into

whatever categories naturally arise from the data, the learning is unsupervised. ML algorithms themselves can be categorized into many different types, including support vector machines (Cortes and Vapnik 1995), artificial neural networks (Rojas 1996, MacKay 2005), and classification and regression trees (CART, Breiman et al. 1984, Quinlan 1986).

Quinlan (1986) noted that "expert systems" are in high demand for completing complex tasks in modern society. In what he terms the "interview approach", which humankind has used for nearly all of history, "domain specialists" and "knowledge engineers" work together to develop explicit rules outlining and defining the knowledge available about the operation of a particular expert system; such a method may result in "a few rules per man day". By contrast, computers, and thus, ML techniques, enable the rapid elucidation of the thousands of rules often required to develop new expert systems. Weather forecasting certainly qualifies as an expert system and a complex task, one that requires thousands of rules (or more) to complete. In fact, one of Quinlan's (1986) examples of a classification task is categorizing the type of weather occurring on a certain day of the week. Similarly, Brieman et al. (1984), who first came up with the CART acronym for classification and regression trees, used the prediction of ozone risk days in the Los Angeles basin as an example system from which CARTs could be useful. Although neither of these authors are atmospheric scientists, they recognized early on that meteorology is an extremely complex and data-dense field where ML has the potential to provide both additional skill in the forecasting process and physical insight into the relationships between meteorological data and weather outcomes.

More recently, ML techniques have been applied to a broad range of meteorological problems, including but not limited to convective initiation (Mecikalski et al. 2015), daily solar energy production (Martin et al. 2016), extreme rainfall (Nayak and Ghosh 2013), storm-scale ensemble probabilistic QPF (Gagne et al. 2014), mesoscale convective system initiation (Ahijevych 2016), aircraft turbulence (Williams 2014), and tornado development from mesocyclones (Trafalis et al. 2014).

The purpose of this study is to apply, for the first time, ML techniques to a large archive of NWP model output and produce automatic optimized probabilistic forecasts of flash floods at what can broadly be considered the FFA scale. From this application, I hypothesize the following:

1. ML techniques based upon NWP data will result in forecasts of flash floods with more skill than methods that rely on a single NWP model field to predict flash floods. These ML techniques can be calibrated or otherwise adjusted to generate reliable probabilistic forecasts of flash floods.
2. ML techniques provide physical insight into the atmospheric environments, as forecast by operational NWP systems, associated with flash floods across regions and over long periods of time. These physical insights build upon our current understanding, as expressed in the scientific literature, of flash floods.
3. ML techniques can be applied to the flash flood forecasting problem in a quasi-operational context, where the ML forecasts can be used as another piece of evidence by human forecasters in the Flash Flood Watch issuance

process. They can also be used to introduce additional automation into the forecasting and alerting process, particularly in applications outside the U.S.

Chapter 2 is a survey of different ML algorithms, the use of NWP in weather forecasting, and the current state of the flash flood forecasting problem in the U.S. Chapter 3 describes a study applying a type of ML – the random forest – to an archive of outputs from the Global Forecast System NWP model over the U.S. to probabilistically predict reports of flash floods. Chapter 4 presents the results of a series of tests exploring how atmospheric variables interact with one another to contribute to the development of flash floods, and additionally discusses how ML techniques can improve understanding of these interactions. Chapter 5 is concerned with the operationalization (and implications thereof) of automated ML predictions of flash floods; it contains two case studies for which the system proposed in Chapter 3 is applied – one from May 2013 and one from May 2015. Chapter 5 additionally conveys some early results from research-to-operations activities where operational forecasters used the proposed system in a testbed setting and discusses the how the proposed system might be expanded to cover the globe. Finally, Chapter 6 is devoted to conclusions and recommendations arising from this work as well as a discussion of how the studies presented herein fit within the history and future of automation in weather forecasting.

Chapter 2: Literature Review

Machine Learning

The phrase "artificial intelligence" (AI) was first used in 1956; the earliest and most basic definition of the idea was that machines would be trained to solve problems that had, until then, been solved by in the human brain (Standage 2016). Quinlan (1986) identifies machine learning (ML) as a "central research area" within artificial intelligence, because if one is to achieve true AI, it stands to reason that learning must be an integral part of that intelligence, just as learning is an integral part of natural intelligence (Weiner 1961). Although authorities disagree exactly on how to define "learning," a concrete way of doing so is to consider learning as a process of acquiring knowledge that the learner can use to develop a set of rules (Quinlan 1986). The other driving force behind ML's centrality to the study and development of AI lies in ML's easy applicability to real-world problems. As our world grows ever more populated by tremendously complicated systems, vast reams of data pour forth as a result. The result, the Information Revolution, is the successor to the earlier Industrial Revolution (Veneris 1990). These data are generated at rates so rapid and in quantities so large that traditional manual interpretations of them by the human brain are precluded. Computers, however, are well-suited for this because they are designed to rapidly complete repetitive tasks, including learning: developing sets of rules based upon big datasets.

ML algorithms can process large archives of data, identify useful patterns, and develop rules for identifying these patterns in an automatic and optimized way. An archive of data is required; these data are split into two parts: training and testing. Figure 1 is a flowsheet describing the general data flow required to develop a ML model.

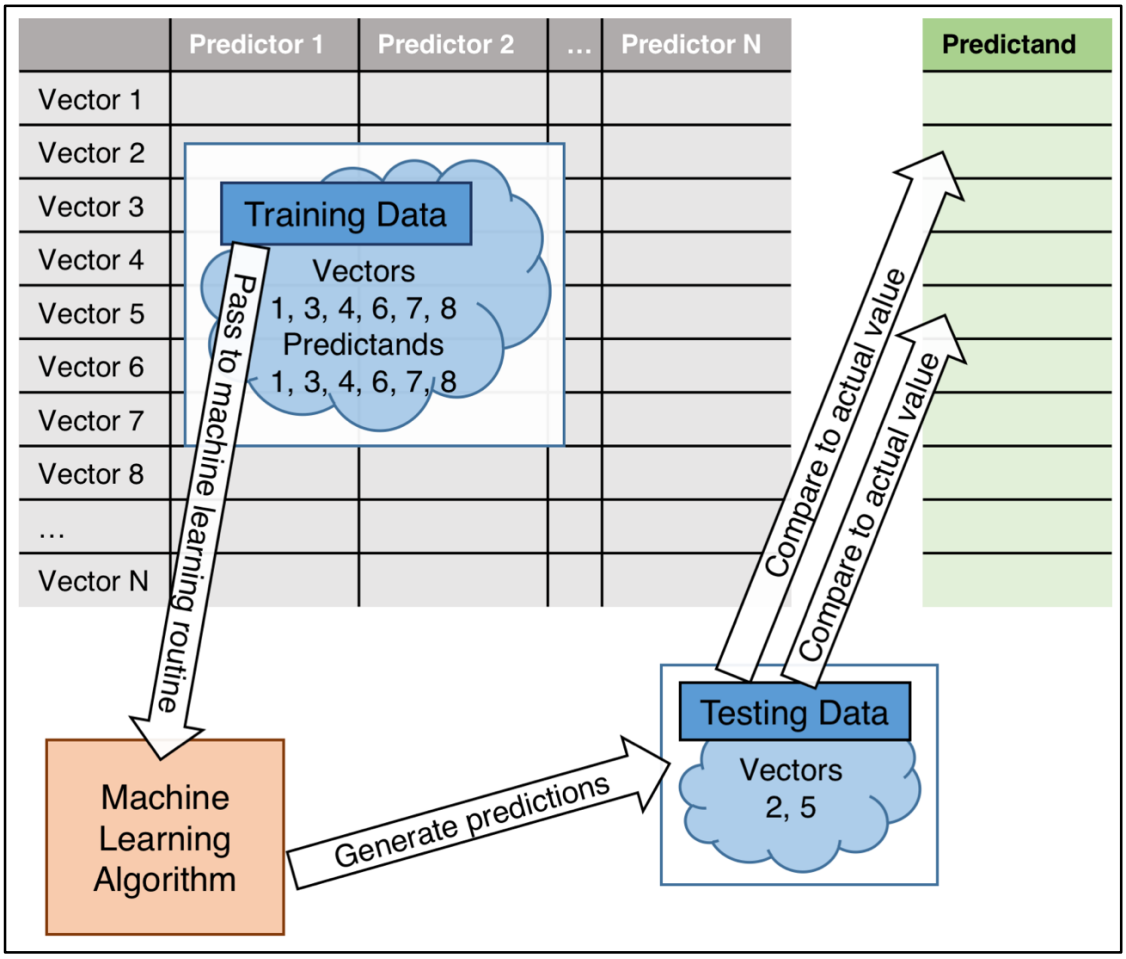

Figure 1. General flowsheet of ML process

The training data consist of vectors that contain some number of candidate predictor variables (or independent variables) and a predictand (or the dependent variable). The ML algorithm analyzes the training data and develops a mathematical model for predicting the dependent variable given a vector of candidate predictor variables. Then the ability of the mathematical model developed by the ML algorithm to skillfully predict the dependent variable can be assessed using the testing data, by running vectors of predictor variables through the model and comparing the ML predictions with the actual values of the predictands in the testing data.

ML methods are typically used to solve problems in one of two ways: categorization or regression. Categorization refers to a process in which the output of the

ML technique is a label; the algorithm develops rules from the data fed to it and then uses these rules to label individual records or instances in the dataset as belonging to a particular category. These categories are of a finite number and are separate from one another in some way. Regression proceeds similarly, but the output from the ML model is instead an estimate of some continuous response variable. One other important distinction within ML is that between supervised and unsupervised learning. Supervised learning encompasses those tasks in which the classifier (or regressor) is given the output categories (or values of the response variable) that correspond with vectors of predictor data. In unsupervised learning, the classifier is not given the output categories and instead collects the various input vectors into separate categories that arise as the data are analyzed by the algorithm (Bishop 2007, Kohavi and Provost 1998).

Archer and Kimes (2008) outline the characteristics of problems for which ML is often employed. When the number of possible predictor variables is large or when candidate predictor variables are not independent from one another, ML is superior to traditional predictive approaches, like logistic regression (LR) or multiple linear regression. A wide variety of algorithms are considered machine learners, but they all share two goals: making accurate predictions and providing insight into how a particular prediction arises from the classifier. Although most ML algorithms are widely applicable across disciplines and problems, any method has to be tested for applicability, by fitting a model to a set of data and then evaluating its skill in solving that particular problem (McDonald et al. 2014). Selecting the appropriate method therefore requires answering two questions: Can the classifier make accurate predictions given the available training and testing data? Does the classifier provide new insight into the problem?

Random Forests

First proposed by Breiman (2001), random forests (RFs) are a ML approach from the classification and regression tree family. An RF is a collection of decision tree predictors grown via bootstrap samples from a training data set "such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman 2001). RFs have become popular for classification and regression tasks and have been used to attempt solutions to a wide range of scientific and engineering problems. Ahijevych et al. (2016) explain that each tree consists of a series of nodes (or "splits"), each of which is followed by two child branches. At the end of each branch is another node with two child branches extending from each node, and so on, as shown in Figure 2.

A randomly sampled set of vectors of predictor data and their corresponding predictands (or, a set of "cases") from the training dataset start at the base of a tree within the forest. Other trees begin with other randomly sampled sets of cases from the training dataset. Then one predictor variable is selected from a random subsample (with replacement, so that the subsample of predictors is available for potential use at any other node in the tree) of all the candidate predictor variables. The selected predictor variable is the one that results in the purest (or sharpest) split between all the predictands from the subsample. For example, in a binary classification problem, where the end goal of the forest (and thus, each of the individual trees) is to decide if a case should be labeled "yes" or "no", the RF algorithm will preferentially select that individual predictor variable that results in one child branch of a node having as many "yes" votes as possible, leaving as many "no" votes as possible at the end of the other child branch.

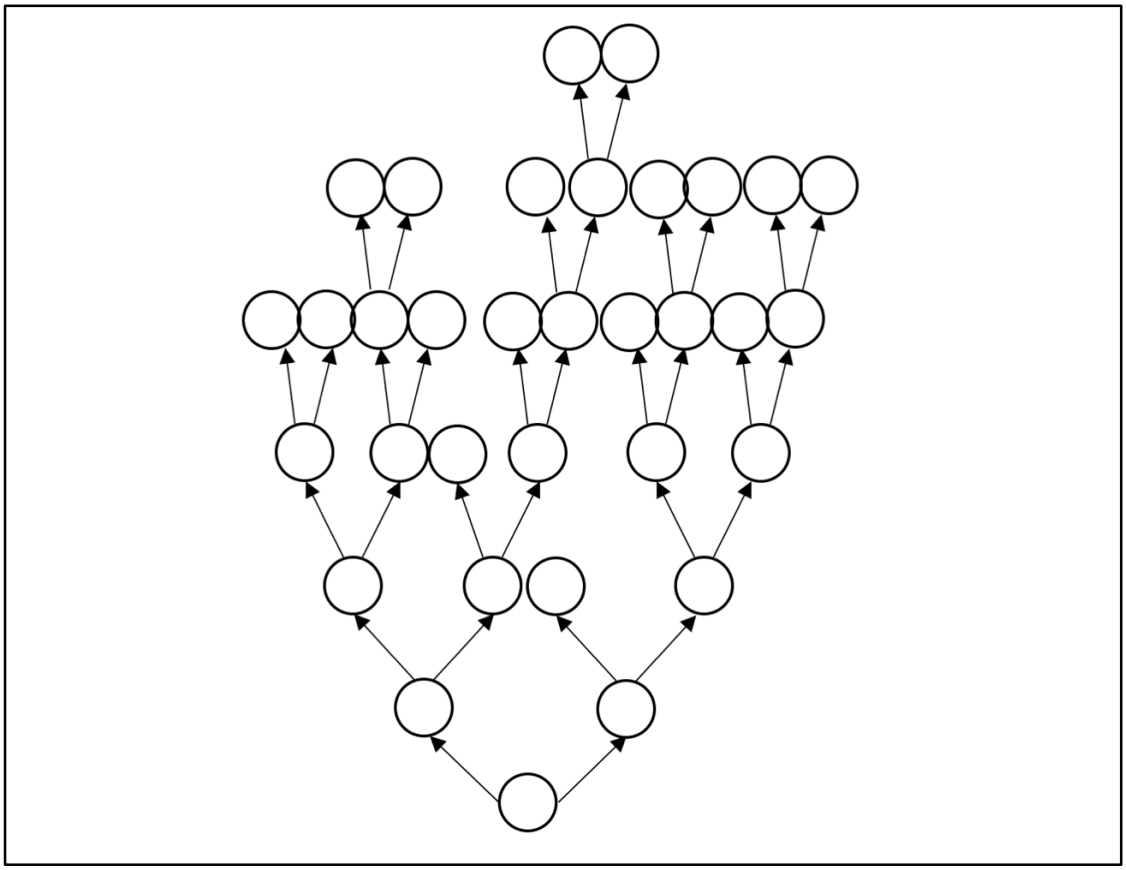

Figure 2. Schematic of a decision tree within an RF

When the selected predictor at a node results in a perfect split (i.e., all of the cases at that node are labeled "yes" or "no"), the tree stops growing from that node. Each node consists of a simple rule of the following form: when the data reach node *n*, follow branch *A* if the predictor *X* used at that node is greater than some threshold and follow branch *B* if the predictor *X* used at that node is less than or equal to the same threshold. At the end of the growth process, each tree in the forest will be different, because random sampling is involved in determining which cases from the overall training dataset will be used to grow each tree and random sampling is used to determining from what subset of predictors the splitter variable at each node will be chosen. Although each individual tree could be overfit to the particular subsample of cases used to grow that tree, the

randomness in the fitting process results in trees that are only weakly correlated with one another; thus, RFs are relatively unsusceptible to overfitting (Touw et al. 2012).

To summarize the fitting process, each tree provides an "expert" opinion about a particular case. In a binary classification problem, given a particular case, some trees will vote to label the case "yes" and others will vote to label it "no". This difference of opinion arises because of the differences in the growth process of each tree. The point of both random subsampling steps in the process is to reduce the "statistical dependence" of each tree upon its colleagues in the RF (Williams 2009).

In the fitting process, the cases not initially used in growing a given tree are retained as "out-of-bag" (OOB) examples; these are used to internally determine the importance of each predictor variable (Ahijevych et al. 2016). Williams (2009) explains that this "importance" can be determined by calculating the "permutation accuracy importance" (also called the "mean decrease in accuracy" or MDA). This metric is calculated by randomly altering the value of that predictor variable and then measuring how the error in the final predictions changes as a result. If altering the value of a particular predictor variable results in a large increase in the error of the predictions made upon data from the OOB cases, that implies that predictor variable is important.

In addition to the MDA, other metrics arising from the RF process can be used to quantify variable importance. Touw et al. (2012) describe the Gini importance of a variable, which is defined as the "sum of the Gini impurity decrease of every node in the forest for which" that variable was used as the splitter. The Gini impurity is defined as the probability that a randomly selected label from node 1 (in this case, "yes" or "no") would be incorrectly guessed by an outside observer. Therefore, if node 1 starts with an

equal mix of "yes" and "no" cases, the Gini impurity is high, because the probability of picking a case and guessing the wrong label of that case is high. If predictor *X* leads to nodes 2 and 3, and node 2 is mostly "yes" and node 3 is mostly "no", the Gini impurity is low, because now the probability of picking a random case from either node 2 or 3 and guessing its label incorrectly is low. When the Gini importance of a variable is normalized across all the nodes of all the trees in the forest, it is known as the mean decrease in Gini impurity (MDG). For example, collect all the nodes in the forest for which predictor *X* was the splitter. Now, at each of these nodes, calculate both the Gini impurity at the node itself and the Gini impurity after the cases have been split by predictor *X* (i.e., at the end of both of the child branches). If the Gini impurity decrease across all the nodes is large, that implies that predictor *X* does a good job at discriminating between "yes" and "no" cases and therefore is an important predictor. If the Gini impurity does *not* substantially decrease as a result of predictor *X* it implies that *X* is not an important predictor. Tan et al. (2005) identify other metrics, including information gain and entropy, that behave similarly to MDG.

RFs have several other appealing features that recommend their use. They are conceptually easier to explain to potential users when compared to competing methods like support vector machines (SVMs), as decision trees are a common and easily-understood human-readable way of organizing rules and criteria into a systematic decision-making framework. RFs start from the decision tree concept, fix some of its undesirable characteristics (like overfitting), and make it suitable for use with large amounts of data that would otherwise preclude the use of manually-generated decision trees. Ahijevych et al. (2016) note that RFs result in empirical models that can represent

many sorts of mathematical relationships; on the other hand, prediction techniques like linear regression require that the predictor and predictand be linearly-related. In an RF, the trees can be allowed to grow out to points that result in entirely pure terminal nodes, or "leaves" (Touw et al. 2012). Traditional decision trees must be "pruned", rather than allowed to grow to their full extent. Recall from the explanation above that nodes continue to spawn new branches until the end of the branch contains cases where all predictands are labeled "yes" or "no". One can use these fully-extended trees to classify cases because the final result of the forest is an average from an ensemble of individual weakly-correlated trees.

Another major advantage of RFs is the inherent cross-validation process that arises as a result of storing OOB cases during the tree growth process. These OOB cases are automatically used to determine the error rate (the OOB error) expected when all cases in the training set, and not just those selected for the tree growth process, are classified. In other words, as Touw et al. (2012) point out, each tree inherently possesses a training data set (the cases selected for growing the tree) and a test data set (the OOB cases). Individual decision trees, because of their tendency to overfit data, are very susceptible to changes in the training data. If the training data do not adequately encompass the entire range of possible outcomes one hopes to predict, an individual decision tree will not be capable of making competent predictions (Gagne et al. 2014). However, RFs smooth out the large error variance one would observe when using individual decision trees; this in turn means the results from the forest are more robust if the training set changes. Finally, because randomness is involved in selecting the variables to be used at a given node, the RF method is capable of identifying interactions

between variables that would elude other methods (Gagne et al. 2014, Ahijevych et al. 2016). The RF method also automatically removes any bias that might be present in a particular predictor *X* and extracts the information needed from that predictor to produce the best possible predictions (Ahijevych et al. 2016).

RFs are also untroubled by multicollinearity between explanatory variables or by the inclusion of unimportant explanatory variables. Because predictor variables are selected for use at each node of a tree based upon which of the available variables results in the best split between labels at the subsequent nodes, some variables will never (or rarely) be used in the process of growing a decision tree (Mecikalski et al. 2015). When there are duplicative predictors present in a training set, the RF method of selecting splitter variables from a random subset of all the predictors outperforms other ML techniques (Archer and Kimes 2008).

When compared to other ML techniques, RFs have fewer parameters and require less "tuning" to achieve results comparable to those achieved with other ML classifiers that require more tuning (Touw et al. 2012). The "tunable" parameters in an RF consist of *mtry*, which refers to the number of predictors in the random subsample available for splitting at each node of each tree (Touw et al. 2012, Breiman 2001), *ntree*, the number of trees in the forest (Tatsumi et al. 2015), and what this study will refer to as *dtree*, the maximum depth (or height) a tree is allowed to reach. The size of individual trees in the forest can also be regulated by limiting the number of total splits in a tree or by setting the number of cases a node must exceed in order to be allowed to split (Touw et al. 2012). Tatsumi et al. (2015) cite dueling studies on the topic of *mtry*: one explains that reducing *mtry* simultaneously weakens individual trees and improves the forest by reducing the

correlation between trees, while another states that increasing *mtry* results in better predictions and allows developers to reduce the number of predictors fed to the RF algorithm. Tatsumi et al. (2015) conclude that, given the disagreement in the literature, *mtry* is a parameter that should be optimized for the particular system to which one is applying an RF. Deeper trees result in less biased predictions, as each tree is more closely fit to the training dataset. However, because each individual tree is now a much more complex model, the variance between trees is increased. Of course, a greater number of trees (increased *ntree*) acts to reduce this variance, because a greater number of "votes" (trees) contribute to each prediction and so one or two highly variable votes have relatively little impact on the final outcome. Thus a trade-off exists between bias and variance, and given a fixed amount of time and computing power, one must balance the characteristics of the deep-but-fewer treed forest with those of the shallower-but-many treed forest. The type of forest that results in the desired predictive skill is the forest with the "correct" values of *ntree* and *dtree* for that application. Finally, note that, if computing power and time are not major concerns, trees can be grown such that all terminal nodes (leaves) are 100% pure, as explained in the text associated with Figure 2. In that case, the only RF parameters left to tune are *ntree* and *mtry*; however, increasing *ntree* under even those conditions will eventually lead to some point where the variance is minimized and the prediction accuracy has been optimized.

Prediction models based upon RFs can be developed relatively quickly compared to other ML techniques. Firstly, the requirements imposed upon the developer or programmer of the forest are relatively light: as shown, the tunable parameters are few in number and governed by simple rules. The computational requirements of developing

an RF are also relatively light (Trafalis et al. 2014), which makes RFs a suitable choice for problems where the testing dataset is large (Hardman et al. 2013). Touw et al. (2012) state that well-tuned SVMs outperform RFs for certain applications, but that RFs compare favorably to SVMs because of their ease of use and fundamental simplicity. For all these reasons, those fields in which large amounts of data are frequently generated tend to be those in which RFs (but also, many other ML techniques) are often employed.

Example Uses of Machine Learning

As the rate of growth of computer processing power, memory, and storage has accelerated over the last few decades, ML has been applied to many new problems. Large numbers of disciplines now use modern laboratory techniques, the logical outcome of which is vast amounts of data. The biological and life sciences are perhaps the most canonical example of this trend, but in the physical sciences physics and astronomy are disciplines where the same is true. In the earth sciences, remote sensing and meteorology fall into this category. A review of the extant scientific literature bears this assertion out: in particular, bioinformatics researchers and engineers were among the first to adopt RF techniques for answering questions in their fields, but more recently earth scientists have joined them among the most avid ML and RF users.

In 2007, Archer and Kimes concluded that RFs are useful in microarray studies where researchers attempt to predict the phenotype of an organism based on a large number of candidate genes. In a review article from 2012, Touw et al. identified 58 representative studies from research areas including genomics, metabolomics, proteomics, and transcriptomics (often referred to as the "-omics") where RFs were successfully applied. They determined that RFs represent an attractive and versatile solution to classification and regression problems in data-intensive sciences like

bioinformatics and its various "-omics" sub-fields. Behavioral scientists have also applied RFs to their work. Hardman et al. (2013) used RFs to predict the progression of college and university students in the U.K., while noting that RFs were originally adopted most enthusiastically in the biological sciences. McDonald et al. (2014) used RFs to predict how steering wheel angles can be used to predict drowsiness-related lane departures, and found comparable or better results, using RFs, to the previous "gold standard" prediction model used in that discipline. Electrical grid managers are at once engineers and behavioral scientists of a sort, as they must study, understand, and react to changing aggregate electrical demand that arises as a result of the individual actions and behaviors of millions of customers. In this vein, Lahouar and Slama (2015), used RFs to predict peak electrical demand in Tunisia with a day's worth of lead time. Their method is roughly analogous to a persistence forecast in meteorology; the RF is fed information about the previous day's morning and evening peak demands along with total load information from 24 and 48 hours prior to the time the next day's forecast is generated. Along with information about the time of year and the minimum and maximum forecast temperatures, they were able to generate peak load forecasts with overall mean absolute percentage error of 2.24%. What makes Lahouar and Slama's approach interesting is their use of additional decision tree logic *outside* the RF process to constrain the growth and development of the decision trees in the RF. In other words, to improve the prediction of rare demand values that arise during holidays, for example, the authors employ a rule that forces that particular day's training data for the RF to come from the previous holiday's load information, instead of the previous day's load information. This "expert input selection" plays to the strength of the RF process by allowing it to take care of

routine events represented by vast tranches of data in the training set and simultaneously allowing an outside expert (in this case, an engineer) to supplant the forest's predictions with knowledge not adequately represented in the training set. In the U.S. state of Oklahoma, gradient boosted decision trees and SVMs have been used to predict daily solar energy production (Martin et al. 2016).

Of course, vast amounts of data are also generated in the earth sciences. Space-borne remote sensing platforms have contributed heavily to this state of affairs. Geographers, agronomists, soil scientists, and other earth scientists have adopted RFs in an effort to more effectively predict and map important quantities. In British Columbia, Canada, RFs were used to effectively map various classes of soil material with minimal expense and *in situ* effort while maintaining accuracy that compared quite favorably with that achieved using traditional in-place soil surveying techniques (Heung et al. 2014). Gambill et al. (2016) were able to use RFs to accurately determine Unified Soil Classification System (USCS) soil type codes, based upon U.S. Department of Agriculture (USDA) soil classifications, at a range of military installations widely scattered across the U.S. Specifically, the authors noted that predictions from this RF were significantly superior to previous "crosswalk" methods where USDA soil classes are directly converted to USCS soil type codes via a look-up table.

RFs have also proven quite useful in land-use classification. In a region of homogenous land use in Peru, Tatsumi et al. (2015) were able to classify Landsat pixels into one of eight crop classes using the RF algorithm. Across the forested regions of northern Minnesota, Corcoran et al. (2013) used RFs to successfully 1) classify land into upland, water, and wetland areas and 2) categorize wetland pixels into different types.

The authors were also able to use RF variable importance measures to recommend how and when to employ remote sensing platforms across seasons and hydrologic regimes to achieve the best classification accuracy for the smallest cost. In southwest Oklahoma, Yan and de Beurs (2016) used the RF algorithm to classify areas dominated by sparsely vegetated cover, winter wheat cover, C_3 carbon fixating grasses, and C_4 carbon fixating grasses. Ireland et al. (2015) used ML techniques to identify flooded areas in Landsat imagery over the Mediterranean.

Of course, if RFs can be used to classify regions into wetlands, uplands, and water, or by the type of crop grown there, or by the type of grass present, it stands to reason that RFs could also be used to *predict* the risk of inundation in an area based on remote sensing datasets. Wang et al. (2015) did this for a river basin in the Guangdong Province of the People's Republic of China and showed that RFs were reasonably successful at identifying those locations within the river basin most subject to flood hazard risk, based on comparisons with reports of flood impacts from historical events. In the course of this assessment, they hit upon two critical characteristics of the RF method: its utility in solving non-linear problems and the ability of the method to provide physical insights about a phenomenon via internal metrics of variable importance. They also provide a framework by which RFs can be assessed for use in a particular prediction problem. First, the RF must be applied to the problem. Next, one must demonstrate that RF is an appropriate way to solve the proposed problem. Finally, one must assess the success of the RF output in solving the problem. These three points of the framework can easily be extended to any extant ML technique, as long as the second point, that the method is appropriate for the problem at hand, is fulfilled.

Scientists' understanding of other remote sensing topics has been improved by the use of RFs. Hutengs and Vohland (2016) used RFs to improve the resolution of land surface temperature grids from the Moderate Resolution Imaging Spectroradiometer (MODIS) from ~1 km to ~250 m. Tested over the eastern Mediterranean region, they found that fitting an RF model to data derived from the Shuttle Radar Topography Mission's digital elevation model, MODIS land cover products, and MODIS surface reflectance products resulted in accuracy improvements over the typical methods used to downscale MODIS land surface temperatures to higher resolutions. Rather than use an RF to directly model snow depth, Tinkham et al. (2014) took an interesting approach and collected LiDAR (Light Detection and Ranging) surveys of snow depth over southwestern Idaho. They then used an RF to model the spatial distribution of the LiDAR-introduced snow depth error. The authors discovered that, relative to the LiDAR-introduced errors, the RF modeling approach introduced little additional error to the snow depth estimation process.

In problems situated more firmly in the realm of atmospheric sciences, ML techniques, including RFs, have taken longer to adopt than in other fields but have nonetheless been successful in a wide variety of prediction problems, from severe weather to solar energy production to initiation and maintenance of mesoscale convective systems (MCS). Sun et al. (2016) note that predicting solar radiation across regions is critically important to constituencies in the business of generating electricity from solar energy. However, extensive air pollution in some regions of the world makes prediction of solar radiation (and thus, solar energy potential) a difficult problem. Sun and colleagues determined that RF-derived predictions of solar radiation that included an air

pollution index as a predictor outperformed all the previously-published empirical solar radiation models they tested at all locations. In a similar vein, Yu et al. (2016) used the RF algorithm to predict the air quality of urban areas. They found that RFs were more successful at classifying air pollution in Shenyang, China, than any of the other ML or empirical methods tried.

One example demonstrates the sorts of advantages ML has over physically-based alternative prediction methods. Nayak and Ghosh (2013) used SVMs to predict extreme rainfall over Mumbai, India, with lead times between six and 48 hours. They describe the "fingerprinting" method for identifying the atmospheric conditions for a location that have led to the sort of event one hopes to predict; this is the "fingerprint" of the hazard (Root et al. 2007). Then standardized anomalies in atmospheric fields are calculated and clustered in an effort to determine how closely a particular set of conditions matches the fingerprint that was derived from past events. However, Nayak and Ghosh (2013) note that this method suffers from serious limitations in comparison to ML techniques. The fingerprint does not consider false alarms, because it is derived only from those archived atmospheric conditions from which the hazard developed, and only one fingerprint is associated with the hazard, which is less than helpful in situations where the hazard may arise under many different combinations of atmospheric conditions. However, although SVMs improved upon the fingerprinting method, the authors found that the ML method still resulted in too many false alarms. The Root el al. (2007) method also considers only the "most important anomaly fields"; ML methods are designed to consider as much data as possible in a comprehensive, optimal way.

Random Forests in Meteorology

The scientific literature now contains a wealth of examples of ML being used to enhance our ability to forecast hazardous weather. Nearly all types of hazardous weather have been subject to this treatment, but in particular, RFs have been used to forecast heavy rainfall, tornadoes, aviation turbulence, convective initiation and more. Trafalis et al. (2014) applied RFs and several other ML classifiers to a dataset of mesocyclones in an effort to predict which mesocyclones go on to produce tornadoes. Although they found an SVM to be the most skillful classifier for their particular problem, the training dataset described in Trafalis et al. (2014) consists of 5,409 records, a quantity of data tiny in comparison to other ML studies, which may involve millions or tens of millions of records. They do note that all the algorithms tested, including RFs, performed similarly, and that RFs have "good accuracy and computational efficiency," which is of course a greater concern in studies dealing with orders of magnitude more data. Of particular concern to Trafalis et al. (2014) was the unbalanced nature of their dataset: only 6.7% of records were labeled in the minority class with 93.3% comprising the majority class; despite the rarity of the minority class (i.e., storms that produced tornadoes), ML methods resulted in skillful predictions.

Williams (2014) used RFs to develop an operational convectively-induced turbulence prediction system. He noted that operational concerns are critical in determining what sorts of predictor variables should or can be considered for use in the ML context. Working closely with the U.S. National Weather Service's (NWS) Aviation Weather Center (AWC) dictates that all predictors be available quickly, reliability, and freely to AWC personnel. Like Trafalis et al. (2014), Williams is required to handle an unbalanced dataset, where reportable aircraft turbulence occurred in either 0.25% or

1.33% of all available records, depending on the airline and level of the atmosphere being considered. Williams also hypothesized that RFs are able to unlock relationships between non-linearly correlated predictor variables that elude more traditional analysis methods. Perhaps most importantly, however, Williams notes that techniques like the RF become more essential year by year as the data produced in meteorology and other data intensive sciences rapidly grow; where these huge quantities of data make, for example, traditional manually-created decision trees a hopeless endeavor, ML techniques hold the promise of identifying new insights and optimizing forecast skill.

Rain and thunderstorms have been the focus of several scientific papers combining meteorology and ML. Ahijevych et al. (2016) developed an RF that provides two-hour forecasts of the probability of MCS initiation. They identify current developments in meteorological applications of ML as the latest iteration of a lengthy evolution of statistical models and weather forecasting that started in the early 1970s with the first forays in model output statistics (MOS). The authors demonstrate that the RF predictions of MCS initiation beat climatology, and conclude that nowcasting impactful events may be the area of meteorology best suited for the deployment of RFs and similar ML methods at this time. Along those lines, Mecikalski et al. (2015) used RFs and LR to produce probabilistic forecasts of 0-1h convective initiation (CI) based on satellite data and NWP output. They found that including numerical weather prediction (NWP) data in a CI algorithm (using either LR or RFs) resulted in better performance than in the operational satellite-only alternative. They also found that the performance difference between the two statistical learning methods was small and probably not significant, although for the cases tested LR slightly outperformed RFs. However, their results

showed that the RF method was able to successfully use more predictors than LR without a concomitant decrease in performance. Mecikalski et al. (2015) observe that LR's slight advantage over RF was unexpected, given the number of studies that have shown RF's suitability (and, indeed, superiority to LR) in complex meteorological nowcasting and forecasting problems. They suggest that the predictor variables in the study are monotonically associated with the probability of the predictand (CI), a situation in which LR can excel. They also suggest that the available training dataset for the study cover too small of a geographical region, such that the RF was unable to fully exploit the possible parameter space governing the true probability of CI in a wide range of conditions.

The advent of storm-scale ensemble NWP has resulted in the generation of vast amounts of fine-scale quantitative precipitation forecasts (QPFs). Gagne et al. (2014) used ML techniques to make sense of this data by improving upon deterministic QPFs from individual NWP ensemble members. They found that RFs, using deterministic ensemble member QPFs and environmental NWP data, improved the skill of precipitation forecasts (forecasters that rain would occur or that heavy rain would occur) compared to LR or uncalibrated ensemble probabilistic QPF. They note that reliance on just NWP data is a risk, because if all the ensemble members are yielding bad predictions there will be no useful information for the ML methods to key in upon. It is also possible that the RF will not be able to generalize its predictions to all possible conditions if the training dataset that grew the RF was too small or restricted; this is especially concerning when rarer heavy precipitation events are anticipated. Because the training period was relatively short, there was not enough data to fit locally-specific RF models, so the authors employed a regional approach and divided the contiguous U.S. (CONUS) into

three parts, fitting a separate RF and separate LR for each region. Despite these limitations, Gagne et al. (2014) point out the most important characteristic of any ML method: it will optimally use all the information available to it. With traditional forecasting methods, this is far from true in the overwhelming majority of cases.

Numerical Weather Prediction

NWP is a two-step mathematical process by which future states of Earth's atmosphere are predicted via complex numerical models (Lynch 2008). The diagnostic step involves using available observations to characterize the atmosphere as accurately as possible at the current time, and the prognostic step involves determining changes in this state over time using the laws of motion and the primitive equations (Bjerknes 1904 in Lynch 2008). Today, NWP is the foundation of the entire weather forecast and altering enterprise. Because NWP models are so complex, especially those that attempt to predict the state of the atmosphere across the entire globe, they have been primarily funded, developed, and implemented by the governments of (or combinations of the governments of) wealthy developed nations, though large private sector entities are beginning to develop and operationalize their own global NWP systems. Global NWP models in wide operational use include those operated by the United Kingdom Met Office, Environment Canada, the Japan Meteorological Agency, the European Centre for Medium-Range Weather Forecasts (ECMWF), and the National Centers for Environmental Prediction (NCEP), part of the U.S. NWS.

The Global Forecast System

Since 1980, the Global Forecast System (GFS) has, as its name implies, provided global forecasts of the state of earth's atmosphere (NCEP Central Operations 2016a). As a product of the U.S. federal government, all GFS outputs are part of the public domain

and thus are readily available and redistributable for any purpose. The GFS produces forecasts out to 16 days (384 hours) of lead time by assimilating millions of observations into the system every six hours; new cycles begin at 0000, 0600, 1200, and 1800 UTC each day. From analysis time to forecast hour 240 (day 10), the model runs at a horizontal resolution of 13 km at the equator, and from forecast hour 240 to forecast hour 384 the model's horizontal resolution is 55 km at the equator. The GFS has 64 vertical levels, which range from the land surface to 0.27 hPa; in addition, a four-layer land surface model is part of the system (Global Climate & Weather Modeling Branch 2016). The GFS has 1-h temporal resolution from analysis time to forecast hour 120 (day five), 3-h from forecast hour 120 to 240, and 12-h from hour 240 to 384 (McClung 2016).

Data distributed to end users are, as of July 2016, post-processed to a horizontal resolution of 0.25 degrees (~25 km at the equator) from analysis time to forecast hour 120. Before 1 May 2016, GFS data were available at a maximum horizontal resolution of 0.5 degrees (~50 km), and before January 2007, the maximum available resolution was 1.0 degrees (~100 km). Currently, post-processing reduces the number of vertical levels presented to the end user to 46 (NCEP Central Operations 2016a). When all the possible vertical levels and model outputs are accounted for, each run of the GFS results in 283 output fields at analysis time and 366 output fields at each forecast hour (NCEP Central Operations 2016b). Although convective and total QPFs are a part of these output fields, the GFS does not produce any direct forecasts of flash floods.

Because computing power and forecast skill are positively correlated with one another (Lynch 2008), upgrades to the GFS are a semi-regular occurrence and occur, on average, about once every other year (NCEP Central Operations 2016a). If more minor

changes to model components are included in this count, modifications occur once every few months, on average (NCEP Central Operations 2016c). GFS code changes and upgrades to the model physics, parameterizations, or resolution potentially affect the statistical properties of the model output fields, and this, in turn, has implications for ML models based upon those model output fields and their respective statistical characteristics.

Understanding Flash Floods

The first step in forecasting flash floods is understanding them, and the first step in understanding them requires defining them. As stated in the introduction to this work, the NWS Glossary (2009) defines a "flash flood" with the following statement:

> A rapid and extreme flow of high water into a normally dry area, or a rapid water level rise in a stream or creek above a predetermined flood level, beginning within six hours of the causative event (e.g., intense rainfall, dam failure, ice jam). However, the actual time threshold may vary in different parts of the country. Ongoing flooding can intensify to flash flooding in cases where intense rainfall results in a rapid surge of rising flood waters.

In general, defining flash floods is not a trivial problem; to do so requires setting an objective threshold dividing riverine floods from flash floods on the basis of concentration time, a term that refers to the amount of time required for water to reach the watershed outlet from the point in the watershed most distant from the outlet. In turn, concentration time is related to the contributing area of the watershed, but dozens of complicating factors prevent hydrologists from reducing concentration time and watershed area into a single relationship (Woodward 2010), though some studies have attempted this. For example, Gaume et al. (2009), when compiling a database of flash floods covering several European regions, defined "flash floods" as "extreme flood events induced by severe stationary storms". Following this definition, they determined

that flash floods should be considered to occur in catchments less than 500 km^2 in area and are the result of storms lasting less than 24 hours. Braud et al. (2014) modified the definition of Gaume et al. (2009) to include a sliding scale relating the time it takes the hydrograph to reach a peak to the size of the affected watershed. On their sliding scale, the rise time for a 1000 km^2 catchment should be less than 24 hours, while the rise time for catchments less than 100 km^2 should be shorter than a few hours. Braud et al. (2014) also require the event's peak unit discharge to exceed 0.5 $m^3 s^{-1} km^{-2}$ for the event to be classified as a flash flood. The NWS's Flash Flood Guidance (FFG) product defines the point at which bankfull conditions are occurring on small natural stream networks as a flash flood (Clark et al. 2014). The trouble in settling upon a consistent definition of flash floods is one of the factors that conspires to make forecasting them so difficult. However, all authorities agree on at least some components necessary for a flash flood. Doswell et al. (1996) identified these as heavy rainfall and hydrologic response. Llasat et al. (2010) are some of many to note that societal factors must also be considered to fully understand flash floods. Schroeder et al. (2016b) propose a "flash flood severity index" which takes into account both the physical and societal impacts of a flash flood.

Although flash floods can arise as a result of ice jams, dam breaks, and other non-heavy rainfall causative factors, for the purposes of the studies contained within, only flash floods occurring as a result of heavy rainfall are considered. On this basis, a workable flash flood definition is as a follows: a flash flood arises from high precipitation rates that persist over a location for a long enough period of time to induce a hydrologic response that poses a threat to human lives or property.

Heavy Rainfall

While Doswell et al. (1996) explicitly state they are primarily concerned with physical explanations of flash floods and not statistical connections between heavy rainfall and a set of predictor variables, their insights are still invaluable to the aims of this dissertation. They quote C. F. Chappell (1986) as follows: "the heaviest precipitation occurs where the rainfall rate is the highest for the longest time." From this statement, one can immediately recognize that flash flood forecasting requires understanding of rainfall (or storm) development, maintenance, efficiency, and movement. For rainfall to develop, rising atmospheric motion and moist air must be collocated; then rainfall efficiency relates the rainfall rate to the rising motion and the water vapor content of the air. Although Doswell et al. (1996) note that efficiency, in this context, is a complicated quantity to calculate precisely, it can be qualitatively estimated with knowledge of the microphysical properties of the storm, the environmental relative humidity, and wind shear. The problem of forecasting high rainfall rates can be somewhat simplified by primarily focusing upon convective systems, where the rainfall efficiency (and thus, the resultant rate) is usually highest (Doswell et al. 1996). The final category of ingredients to be considered concern storm size, shape, and motion. The period during which the rainfall occurs is a function of the speed and direction of the precipitating system and the length of the precipitating system measured parallel to the direction of the system motion. As with precipitation efficiency, these orientation and motion properties can be related to environmental variables, including the mean wind through the cloud-bearing layer, slow moving outflow (or other types of) boundaries in an environment where strong flow is present, or when propagation due to the evolution of individual storm cells acts to oppose the motion vector of the individual storm cells (Doswell et al. 1996). Most of the

environmental factors described in this paragraph are, today, implicitly considered in or explicitly output by NWP systems.

Hydrologic Response

Hydrologic response to heavy rainfall is a requirement for a flash flood to develop. A given rainfall rate induces different hydrologic responses depending on the degree of saturation of the underlying soil, the shape, size, and slope of the watershed subjected to the rainfall, and the land use/land cover present in the area. In general, when rainfall reaches the land surface, some portion infiltrates into the soil and the remainder runs off over the surface. Urbanized regions are more likely to experience strong hydrologic response to a given rainfall amount because of the great degree of impervious cover, which reduces water infiltration and increases surface runoff. The type of soil present in a region is also an important consideration: soils with large pore spaces in between particles have more available infiltration capacity than tightly-packed soils with less pore space. For example, sandy soils infiltrate a great deal more water than clay soils; thus, sandy soils generate less runoff from a given amount of rain. Runoff generation is also affected by the antecedent rainfall in a region, since previously-infiltrated water uses up some of the soil's available water storage space. Once runoff has been generated, the speed with which it moves and is focused into a particular area governs the timing and peak of the flood wave. These factors are, in turn, controlled by the shape and size of the watershed and by the slope of areas in the watershed. Small watersheds focus water more quickly than large ones; watersheds with steep slopes focus water more quickly than flatter watersheds.

Human Impacts

Flash floods are among the deadliest storm-related hazards in the United States (Ashley and Ashley 2008); overall, floods kill more people globally than any other natural disaster (Doocy et al. 2013). Accepting a definition of a flash flood that requires there to have been a human impact forces us to recognize that flash floods are the outcome of constant interactions between society and its physical environment. The potential human impact from a flash flood in an urbanized region is generally greater than in rural areas, not just because of the increased hydrologic response associated with urbanized regions, but also due to the increased population density and the higher likelihood that a larger number of people live, work, or recreate in the area. In regions where flash floods have struck in the past, actions like moving people and infrastructure out of flood-prone areas, increasing the amount of green space, or improving storm water management practices have been demonstrated to reduce the potential of a flash flood and the associated impacts. Therefore, the probability of a flash flood occurring, and of its impact being observed and reported, involves some human influence.

Flash Flood Forecasting Practices

Meteorological Methods

Several studies attempting to forecast the meteorological components contributing to flash floods have been undertaken in the last fifty years. One early attempt, the "K" index (George 1960), originally focused on forecasting thunderstorms, but has since been used operationally as part of various flash flood forecasting procedures. The K index is calculated from (1):

$$K = (T_{850} - T_{500}) + T_{d_{850}} - (T_{700} - T_{d_{700}}) \quad (1)$$

In (1), T refers to air temperature and T_d to the dew point temperature. The subscripted numbers are the constant-pressure (isobaric) levels (in hPa) upon which the air or dew point temperature are to be calculated. George (1960) included thresholds for the meaning of various K values: K > 35 corresponds to "numerous thunderstorms", K < 20 to "no thunderstorms", and intermediate values to various categories of areal thunderstorm coverage.

Giordano (1994) summarized the state of flash flood forecasting knowledge at the time by tabulating a series of atmospheric indexes thought to be important to severe weather and flash flood forecasting, together with thresholds of concern for each index. Richardson et al. (2011) found that, despite differences in synoptic-scale set-ups, high surface dew point temperatures, high precipitable water (PW), moderate convective available potential energy (CAPE), slow moving surface boundary features parallel to the mean wind through the 850-300-hPa layer, and other factors were common in flash flood environments. On the basis of six events, the authors developed a manual decision tree for use in flash flood forecasting. Schroeder et al. (2016a) collected 40 urban flash floods and found that anomalously high PW values were present in nearly all the events; other patterns recurring from event-to-event include relatively saturated low levels, moderate or weak wind shear, moderate CAPE, low convective inhibition, and high K index. Jessup and DeGaetano (2008) studied a series of flash floods that occurred in the Binghamton, NY county warning area (CWA) and found that the events, relative to climatology, tended to be associated with heavier precipitation, abnormally high atmospheric moisture in the upper levels, and stronger vertical motion, greater (but not extremely high) surface-based CAPE, higher K index, and more strongly negative lifted

index (LI). However, while wind direction and shear were found to help illuminate the physical processes governing the development of flooding rainfall, the precipitation, moisture, and convective parameters had to be used in concert with wind information to achieve the best predictions. Many of these factors are produced as output from NWP models.

Indeed, predictions of the environmental fields associated with high precipitation are often of a better quality than the NWP QPFs themselves. For example, research conducted at ECMWF has produced evidence that environmental variables (in this case, integrated water vapor transport) contributing to long-lasting heavy rainfall are more easily predicted than the resultant rainfall itself (Flamig 2016, personal communication). Kursinski et al. (2008) determined that the accuracy of initial PW vapor estimates in NWP had a significant impact upon the accuracy of subsequent QPFs. Antolik (2000) noted that QPF is actually not the most important predictor of observed precipitation in many cases. Perica and Foufoula-Georgiou (1996) successfully downscaled low-resolution model QPF by introducing an additional model variable: the CAPE. Other model variables also hold the promise of improving NWP QPF by including additional, less variable NWP output in the precipitation forecast process (Ganguly and Bras 2003).

QPF has improved in the twenty years since Doswell et al. (1996) were writing. QPF skill, as measured by Gilbert skill score (also called the equitable threat score), has improved steadily in the warm (JJA) and cold (DJF) seasons as well as year-round (Barthold et al. 2015). However, the performance increase in the warm season has been slower than that in the cold season, and Barthold et al. (2015) suggest that this is because NWP models experience more trouble in accurately generating QPF from the small-scale

thunderstorms that often predominate during the warm season. Though producing accurate QPF is riddled with difficulties, the advent of convection-allowing models holds great promise for doing so at the scale of individual storms (Gagne et al. 2014). Still, despite these advances, the inclusion of environmental fields from NWP in a rainfall prediction system can add additional information helpful in the quantitative precipitation forecasting process.

Hydrologic Methods

From the hydrologic perspective, differences in soil type and texture contribute to soil water capacity, which, for example, can be used alongside other factors like impervious area ratio and hydraulic conductivity as *a priori* parameters in hydrologic models suitable for flash flood forecasting (Clark et al. 2016). Jessup and DeGaetano (2008) found that, for Binghamton, NY CWA, soil moisture was a good discriminator between flood and non-flood events. Indeed, the operational tools used in NWS flash flood forecasting rely heavily upon antecedent soil moisture and the land's potential for generating surface runoff (Clark et al. 2014). Other NWS entities have developed hydrologic indexes that account for the impact of land slope, wildfire scars, and soil characteristics upon surface runoff generation (Smith 2003), but these index methods have generally been unsuccessful at flash flood forecasting when used in isolation (Clark et al. 2014).

Clark et al. (2014) and Gourley et al. (2012) demonstrated that the tools used in NWS operational flash flood forecasting drastically over-forecast flash floods compared to the NWS's *Storm Data* historical event database (MacAloney 2016). These tools, part of the FFG family, consist of gridded values that represent the amount of rain required over a given period of time to induce bankfull conditions on small natural stream

networks. FFG is produced using one of four methods; the method used for a particular location depends upon which NWS River Forecast Center (RFC) has jurisdiction over the location. Two broad concepts underlie the production of FFG: a rainfall-runoff model and a series of surveyed "threshold runoff" (or ThreshR) values. Carpenter et al. (1999) defined ThreshR as the "amount of *effective* rainfall of a given duration that is necessary to cause minor flooding" [emphasis in the original]. After accounting for infiltration into the soil, evaporation, and other losses, effective rainfall is the amount that becomes surface runoff. The absorption capacity of the soil is related to the soil moisture; wetter soil conditions act to decrease the effective rainfall. FFG is determined from a rainfall-runoff model, where the "runoff" in the model is the effective rainfall (the ThreshR as modified by the modeled soil moisture conditions), and the "rainfall" (the rainfall necessary to generate the runoff) becomes the FFG value.

Although FFG contains information related to antecedent rainfall and hydrologic response, it does little to address the meteorological side of the equation. A key issue in flash flood forecasting is knowing how heavy precipitation behaves over time periods of one hour or less (Brooks and Stensrud 2000). However, FFG is updated infrequently; at best, this process occurs every six hours but it is highly inconsistent from region to region, and even when the products are regularly updated they only address rainfall accumulation periods of one, three, and six hours (Clark et al. 2014).

Additionally, FFG has no provisions for the impact of *future* rainfall on the threat of a flash flood. Consider a situation in which heavy rainfall is forecast to occur on and off over the next three days. A forecaster wishes to use FFG to determine the risk of a flash flood. Unfortunately, the FFG guidance available to our hypothetical forecaster is

valid *only* from the current time to one, three, or six hours in the future. Forecast rain that falls between six and 72 hours from the current time will act to increase the antecedent soil moisture and thus decrease the FFG values (on the other hand, if no additional rain occurs, the soil will begin to dry out and the FFG values will increase). If our forecaster compares the current 66 to 72-h QPF to the current 6-h FFG to produce a 3-d flash flood forecast, he runs the risk of underestimating the threat of a flash flood if rain occurs between now and 66 hours, or of overestimating the risk of a flash flood if no rain occurs over that period. Despite these perils, operational entities have explored the use of FFG in this context for the purposes of issuing 24 and 48-h flash flood outlooks, because no NWP models or other available systems are able to directly forecast the potential of a flash flood any farther than six hours into the future. One promising avenue of research (Martinaitis et al. 2016) is to feed a hydrologic model with QPFs from NWP, but the location errors present in even 3-h storm scale QPFs have a deleterious effect on the skill of the hydrologic predictions. Of course, it is possible that QPF location errors could be reduced via the addition of other, more stable NWP outputs to the prediction process, as discussed above in "Meteorological Methods".

Conclusions

The NWS is the U.S. government agency responsible for issuing messages to the public of flash floods. Via its network of local Weather Forecast Offices (WFOs), regional RFCs, and the national Weather Prediction Center (WPC), the NWS monitors and forecasts the risk of flash floods across the nation. Local meteorologists (and a handful of hydrologists) at WFOs provide the public with life-saving flash flood alerts, including Flash Flood Warnings (FFWs), Flash Flood Watches (FFAs), Urban and Small

Stream Flood Advisories, Areal Flood Advisories, and more. Hydrologists at regional RFCs develop, maintain, and run the hydrologic models and help produce the precipitation estimates used to issue these local messages. In a less public role, the WPC guides the regional and local forecasters by providing insight into the differences and agreements between different NWP solutions and by producing probabilistic forecasts of exceeding FFG in the 0-24-, 24-48-, and 48-72-h ranges.

Clark et al. (2014) and others have suggested that this organizational structure introduces a possible responsibility mismatch with troubling implications for the success of the NWS flash flood program. While FFG is available for short-fuse FFWs, it is, by its very definition, of little or no utility in the FFA development process, a process for which there is currently a glaring absence of available tools. In lieu of specific flash flood forecasting tools, meteorologists use NWP to predict heavy rainfall and then use their knowledge and experience to translate this information into potential flash flood impacts. This process, though, requires the meteorologist to possess at least some understanding of the hydrologic and societal processes at work in the region of a potential flash flood. Meteorologists, inculcated in the atmospheric sciences, are often not trained to use hydrologic expertise to predict the risk of a flash flood given a particular set of hydrologic conditions. In fact, FFG is designed explicitly to "meteorology-ize" the hydrology involved in flash flood forecasting by reducing the hydrologic part of the problem down to a specific rainfall accumulation threshold (Clark et al. 2014). This inevitably results in a portion of the possible hydrologic knowledge being rendered unavailable for use in either the alert decision-making or forecast process.

However, ML techniques offer a bright ray of hope. They have now been applied to categorization/labeling and regression problems in many fields, and there is a growing recognition of the potential utility they hold when applied to meteorological forecast problems. Aircraft turbulence, mesoscale convective system initiation, hail size, severe thunderstorm maximum wind gust, and other phenomena have been predicted from NWP and observational data using these algorithms. In general, ML requires a set of predictors (like NWP output) and a predictand (like observations of interesting phenomena); these techniques inherently contain elements of automation and objectivity and they are particularly well-suited for applications where manual analysis has traditionally been used to draw conclusions from huge amounts of data. Additional advantages statistical approaches in this domain include the fact that they require far fewer computing resources than physical models and require far less time to run. Finally, these techniques are capable of sorting through, in an optimal way, all the available and relevant data for a particular problem.

However, disadvantages to statistical techniques also exist. Because these algorithms work by identifying patterns in large archives of data, it can be difficult to draw physical interpretations from statistical models, especially if the statistical model is applied to NWP output, which is only an imperfect proxy for the real atmosphere. Although research is spotty, it is also possible that statistical techniques and increasing forecast automation results in forecaster disengagement from the forecast process, which is undesirable when extreme cases appear that are not well-sampled or well-represented in archives of potential training data. Snellman, writing in 1977, called this inexorable increase in forecast automation “meteorological cancer” and warned that, as machines

take on more and more of the legwork of forecasting, humans will become less skilled at the fundamentals of forecasting. However, given the enterprise's incredible advancements over the last 40 years, the era of "forecaster as communicator", as Snellman (1977) put it, has not led to an apocalypse or, indeed, anything even close to that.

Several properties of flash flood events conspire to make forecasting them problematic, including the rapidity with which waters rise, the small spatial scales over which their impacts characteristically vary, and the difficulties inherent in objectively defining and observing these events. Fundamentally, flash floods exist at the interface between society, hydrology, and meteorology; unlike other weather phenomena with objective physical definitions, a flash flood only exists once some human impact has been observed. As NWP and other modern forecasting tools have woven their ways into the very heart of the weather enterprise, a huge archive of useful data has grown – day-by-day – on computer servers and in cold storage across the world. Over the same time, computer scientists have developed ever-more-efficient ways of milling through these vast stores of data. These techniques waste nothing, can result in physically-interpretable outputs, and hold the promise of improving forecasts and life-saving alerts to mitigate the impacts of flash floods and – ultimately – to better protect life and property.

Chapter 3: Random Forest Predictions of Flash Floods in the United States

Introduction

In this study, machine learning (ML) techniques, primarily random forests (RFs), are applied to a lengthy archive of outputs from the Global Forecast System (GFS), a numerical weather prediction (NWP) model, to automatically forecast the likelihood of a flash flood resulting from a particular set of atmospheric conditions. This prediction framework is intended to eventually improve the ability of weather forecasters to identify regions susceptible to flash floods with 24 hours or more of lead time. When 6-h NWP forecasts are considered, the method outperforms operational equivalents (namely, using individual NWP model fields to forecast flash floods) by resulting in fewer false positives and better detection of events. Unlike quantitative precipitation estimates (QPE) or quantitative precipitation forecasts (QPF), this method also provides probabilistic forecasts of flash floods. In general, it is possible to calibrate raw confidence "scores" from the RF method to produce more accurate probabilities of an event's occurring, when NWS *Storm Data* (MacAloney 2016) reports of flash floods are used as verification.

Data Preparation

Global Forecast System Data

GFS analyses have been archived by the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI) since 2 March 2004. Ideally, new GFS analyses are created every four hours as part of the daily 0000, 0600, 1200, and 1800 UTC model initialization cycles. During this period, accounting for model outages and archiving failures, NCEI stored 16,678 analyses. As of the 1800 UTC 31 December 2015 analysis and modeling cycle, the GFS post-

processor outputs 315 individual products when accounting for all combinations of field type and levels upon which each field is computed. Table 1 contains specifics regarding this list of products. In Table 1, CAPE is the convective available potential energy, CIN is the convective inhibition, and LI is the lifted index.

Model code changes and upgrades between 2004 and the end of 2015 result in some changes in this list in Table 1 through time. There have been changes in the horizontal resolution of the model output: the GFS3, a 1.0-degree x 1.0-degree version of GFS output, is available between 2 March 2004 and 31 July 2015. The GFS4, a 0.5-degree x 0.5-degree version of GFS output, is available from 1 January 2007 to 31 December 2015. (On 1 May 2016, model upgrades enabled the generation of 0.25-degree x 0.25-degree versions of the output fields.) For the present study, 16,066 GFS3 analyses from 2 March 2004 to 31 July 2015 have been downloaded and stored. An additional 612 GFS4 analyses from 1 August to 31 December 2015 have been downloaded, stored, and subsequently resampled to match the grid upon which the GFS3 analyses are computed. The GFS3 version of the model outputs was selected for this study to enable storage and processing of the longest possible archive of model data. For example, while a single GFS3 analysis (containing all model fields) requires 30 megabytes of storage, a single GFS4 analysis requires 60 megabytes of storage. The entire archive of analyses from 2 March 2004 to 31 December 2015, including all available model fields, is approximately one-half terabyte in size.

Table 1. Example GFS analysis field inventory (valid 1800 UTC 31 December 2015)

Field Name	*Units*	*# of Levels*
Atmospheric products (28 fields; 302 total products)		
Geopotential height	m	34
Relative humidity	%	36
U-component of wind	m s^{-1}	38
V-component of wind	m s^{-1}	38
Absolute vorticity	s^{-1}	26
Temperature	°C	39
Ozone mixing ratio	kg kg^{-1}	12
Vertical velocity	Pa s^{-1}	22
Cloud water mixing ratio	kg kg^{-1}	21
Pressure	Pa	8
Vertical speed shear	s^{-1}	3
CAPE	J kg^{-1}	3
CIN	J kg^{-1}	3
LI	°C	2
Specific humidity	kg kg^{-1}	3
Standard atmosphere reference height	m	2
Planetary boundary layer height	m	1
Wind speed (gust)	m s^{-1}	1
Pressure of level from which parcel was lifted	Pa	1
Potential temperature	°C	1
PW	kg m^{-2}	1
Cloud water	kg m^{-2}	1
Total ozone	Dobson	1
Percent frozen precipitation	%	1
U-component storm motion	m s^{-1}	1
V-component storm motion	m s^{-1}	1
Dew point temperature	°C	1
Storm relative helicity	J kg^{-1}	1
Field Name	*Units*	*# of Levels*
Land surface model outputs (7 fields; 13 total products)		
Soil temperature	°C	4
Soil moisture (volumetric)	fraction	4
Land/sea mask	dimensionless	1
Ice cover	fraction	1
Wilting point	fraction	1
Snow depth	m	1
Snowpack water content	kg m^{-2}	1

An important consideration in ML is balancing the desire to intelligently reduce the number of variables used to develop a prediction model with the desire to avoid eliminating any potential knowledge no matter how marginal that knowledge might be. Unfortunately, the changes in the GFS model core and post-processing routines through time force a reduction in the number of candidate predictor variables automatically. If a variable is not regularly produced (or cannot be otherwise determined or calculated) throughout the entire period over which the ML model is to be developed, it cannot be used in the ML fitting process. This eliminates several promising candidate variables in Table 1 from consideration, including the storm motion components, specific humidity at some levels, and vertical speed shear. Other fields are not used because they are believed to be irrelevant to the problem, so computational and storage concerns outweigh the need to include them in the process. In this latter category are all products on isobaric surfaces where the pressure is less than 150 hPa and the ozone mixing ratio product at all levels. The final variable reduction step involves eliminating a few isobaric surfaces from the mid- and low-levels, as well as those layers expressed in the sigma and above ground level coordinate systems.

This process results in 95 GFS analysis fields being selected for potential use in the ML model fitting process. These are summarized in Table 2. The "Additional Level Descriptions" column of Table 2 describes only those levels other than the standard isobaric levels chosen for this study, which are defined when the barometric pressure is either 150, 200, 250, 300, 400, 500, 700, 850, or 925 hPa. In Table 2, "AGL" and "BGL" stand for "above ground level" and "below ground level".

Table 2. List of GFS analysis fields used in the ML fitting process

Field Name	*Units*	*# of Levels*	*Additional Level Descriptions*
Geopotential height	m	9	
Relative humidity	%	11	2-m AGL, entire atmosphere
U-component of wind	m s^{-1}	10	10-m AGL
V-component of wind	m s^{-1}	10	10-m AGL
Absolute vorticity	s^{-1}	9	
Temperature	°C	11	2-m AGL, entire atmosphere
Vertical velocity	Pa s^{-1}	9	
Cloud water mixing ratio	kg kg^{-1}	9	
Pressure	Pa	1	mean sea level
CAPE	J kg^{-1}	1	surface
CIN	J kg^{-1}	1	surface
LI	°C	2	surface, best 4-layer
Specific humidity	kg kg^{-1}	1	2-m AGL
PW	kg m^{-2}	1	entire atmosphere
Cloud water	kg m^{-2}	1	entire atmosphere
Soil moisture	fraction	4	10-, 40-, 100-, and 200-cm BGL
Snowpack water content	kg m^{-2}	1	surface
Soil temperature	°C	4	10-, 40-, 100-, and 200-cm BGL

In addition to the analysis fields described in Tables 1 and 2, two pieces of information from the GFS 3-h forecast were downloaded and processed: 0-3 h precipitation rate and 0-3 h convective precipitation rate. This results in a total of 97 GFS fields available for use in the study.

For each GFS model cycle, all available fields are initially downloaded in GRIB (GRidded Binary, used for GFS3 analyses and forecasts) or GRIB-2 (GRidded Binary-2, used for GFS4 analyses and forecasts) format. From each GRIB or GRIB-2 file, each of the 97 desired model fields is extracted, converted to GeoTIFF (Geographical Tagged Image File Format), and stored. To develop training, validation, and testing datasets for the ML process, these series of GeoTIFFs are masked such that only grid cells in or near the contiguous U.S. (CONUS) are considered, as shown in Figure 3.

The series of GeoTIFFs for each model field is converted to a comma-separated values (.csv) file, where each line of each file corresponds to a single value of a model field at a particular time and location. Then these 97 text files are merged based on time and location information to create a .csv predictor matrix, where, for each time and location, the corresponding 97 GFS model field values are stored.

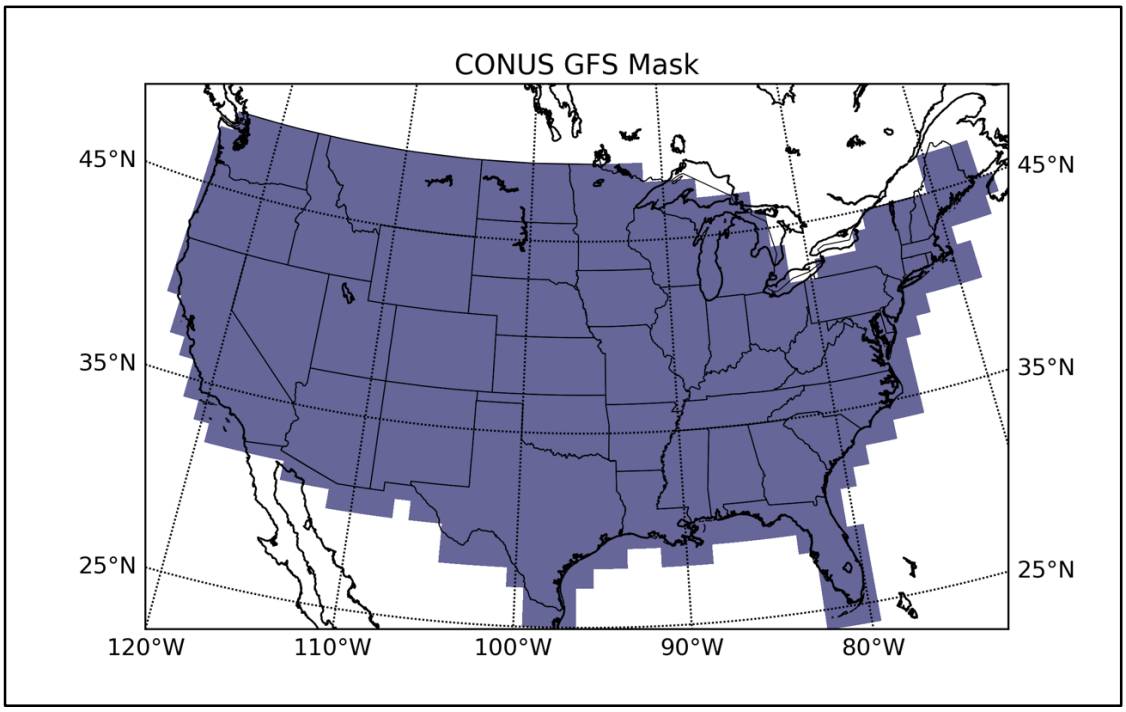

Figure 3. Mask used to process CONUS GFS data for ML predictions

Flashiness and Other Static Maps

Saharia et al. (2015) developed a metric – "flashiness" – across the U.S. to "characterize the ability of a basin to produce flash floods." Flashiness is defined as "the difference between the peak discharge and the action stage discharge normalized by the flooding rise time and basin area." Flashiness can be computed for gauged basins where the NWS has identified a stage height and discharge that correspond to a flooding impact consistent with the definition of "action stage". Then, using an empirical cumulative distribution function, Saharia et al. (2015) scale the observed flashiness values between

zero and one, where values of one represent those basins ablest to produce a flash flood. They employ statistical methods to predict flashiness at ungauged locations from a set of geomorphologic parameters, the most important of which included basin area, slope, annual precipitation, and mean temperature. Saharia et al. (2015) produced these predictions of flashiness at a 1-km horizontal resolution. For the purposes of this study, their 1-km flashiness grid was resampled to the GFS3 grid by taking the median 1-km flashiness value in each large GFS3 pixel. The result, which is Figure 4, is extracted to a text file and then merged with the GFS data already described.

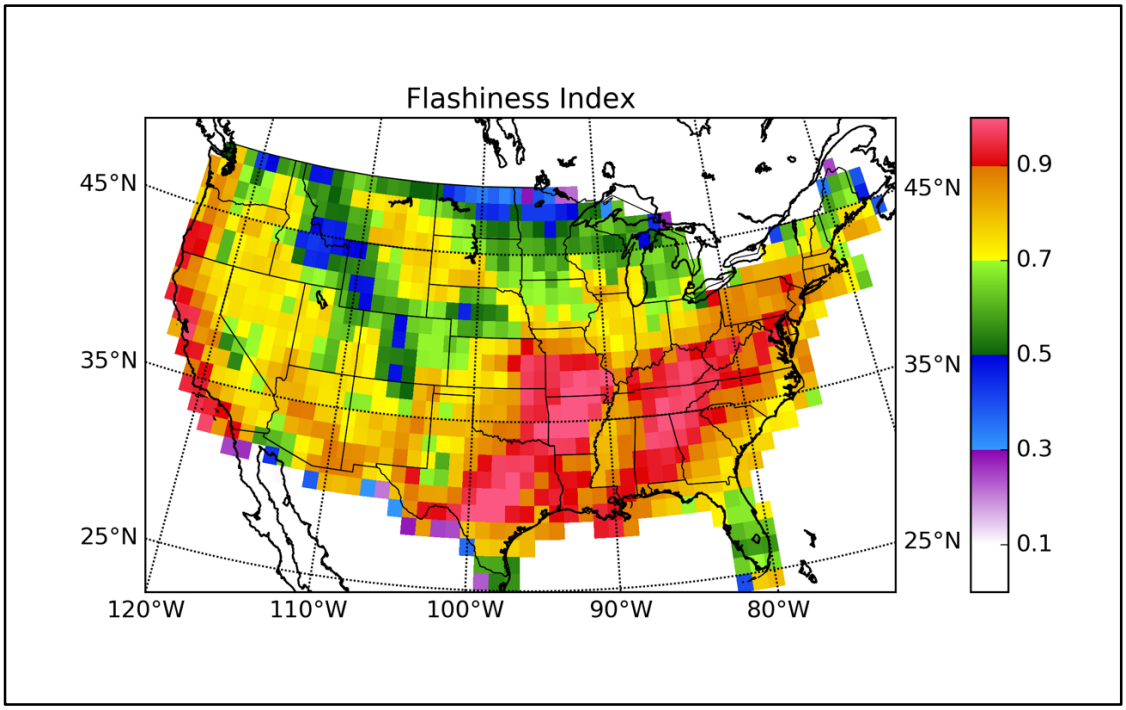

Figure 4. Median flashiness resampled to 1.0-degree x 1.0-degree resolution

Although flashiness accounts for variations in the response of the land surface to a given amount of rainfall, multiple studies have found that the nature of flash flood forecasting and reporting differs from region to region of the U.S. In particular, western flash floods have sometimes been treated separately from eastern flash floods for meteorological reasons (Maddox et al. 1979), for hydrologic reasons (Smith 2003), and

due to societal factors (Schroeder et al. 2016b). The ability of standard operational techniques to forecast western flash floods is also less than it is in the eastern CONUS (Clark et al. 2014).

Among the GFS predictors selected for use in this study are several model variables generated on the 925- and 850-hPa levels. However, many parts of the western U.S. have station surface pressures less than 925 or 850 hPa; in those regions, GFS outputs on these isobaric surfaces have been extrapolated from a combination of lower isobaric surfaces or adjacent pixels. To avoid the introduction of additional uncertainty as a result of this process, it is desirable to exclude 925-hPa or both 925- and 850-hPa grids, as appropriate, from regions with moderately high or high elevations. To achieve this exclusion and simultaneously better account for regional differences in flash flood environments and hydrologic response, three separate regional ML models, based on elevation, will be generated using the scheme shown in Figure 5.

Low elevation regions (green) are denoted with a "1" in the predictor matrix, moderate elevations (yellow) with a "2", and high elevations (blue) with a "3". This procedure is similar to that used in previous studies; for example, due to reduced Doppler radar coverage in the western U.S., Gagne et al. (2014) divided the U.S. into different regions by longitude and developed separate ML models for each region. Table 3 contains a few example records from the predictor matrix used in the study, including time and location information, values from static fields (elevation and flashiness), and values of certain GFS model fields at those times and locations.

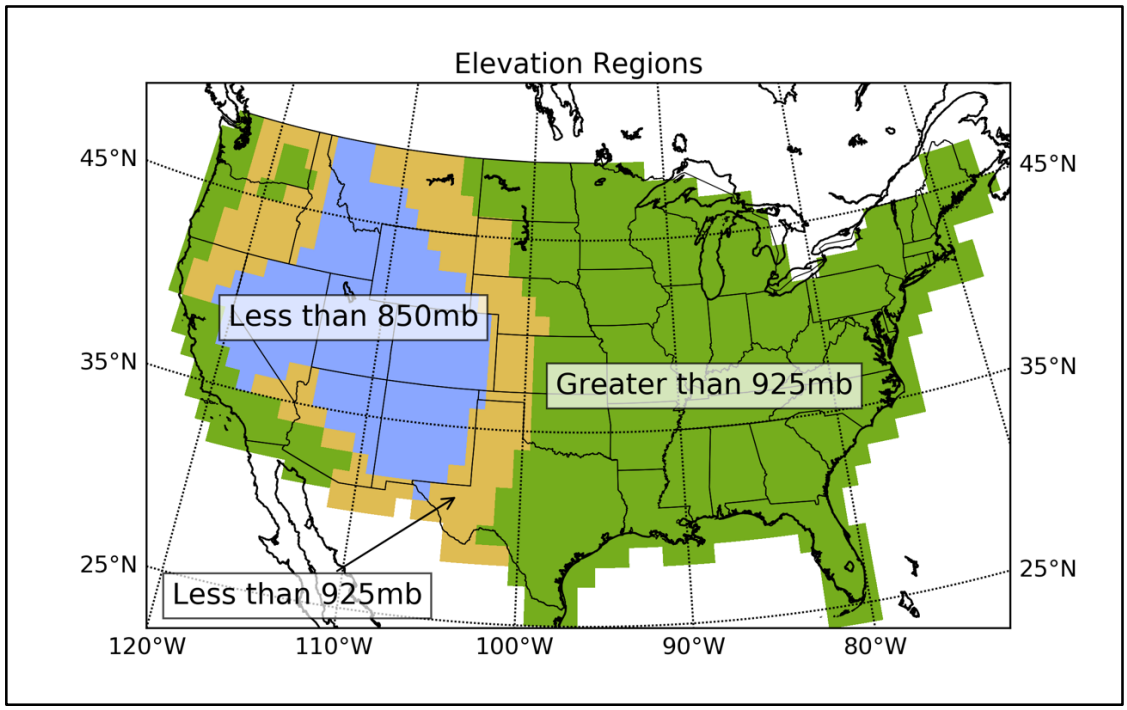

Figure 5. Elevation regions used for regionalizing flash flood forecasts and excluding extrapolated pressure fields

Table 3. Example records from the predictor matrix

time_y_x	*flashiness*	*elevation*	*10m_uwind*	*10m_vwind*	*150hgt*
20120905_00_42_55	0.8478	1	4.3	-4.7	14025.6
20120905_00_43_55	0.8268	1	4.1	-6.9	14039.5
20120905_00_44_55	0.7398	1	2.6	-9.2	14054.3
20120905_00_45_55	0.7856	1	1.1	-13	14068.9

Derived Quantities

Several studies have identified derived quantities thought to be useful in the flash flood forecasting process. Focusing mostly on convective initiation, Manacos and Schultz (2005) collect 19 studies, dating back to 1953, that use moisture flux convergence (MFC) in various sorts of forecast problems; the earliest uses of MFC were to predict the location and amounts of heavy rainfall in midlatitude cyclones. Richardson et al. (2011) identified 0-2 km AGL MFC as an important quantity in flash flood situational awareness. Junker (2008) considered 850-hPa MFC as an important factor to consider when forecasting precipitation from mesoscale convective systems. Waldstreicher (1989)

explains that stationary circular areas of MFC often lead to "areas of excessive rainfall". MFC is defined by (2); (3) is the same equation with the vector terms expanded.

$$\text{MFC} = -\nabla \cdot \left(q\vec{\mathbf{V}}_h\right) = -\vec{\mathbf{V}}_h \cdot \nabla q - q\nabla \cdot \vec{\mathbf{V}}_h \tag{2}$$

$$\text{MFC} = -u\left({}^{\partial q}/_{\partial x}\right) - v\left({}^{\partial q}/_{\partial y}\right) - q\left({}^{\partial u}/_{\partial x} + {}^{\partial v}/_{\partial y}\right) \tag{3}$$

In (3), u represents the zonal (east-west) wind component, v represents the meridional (north-south) wind component, q is the specific humidity, and x and y are the zonal and meridional grid coordinates, respectively. The first and second terms in (3) together represent the horizontal advection of q and the third term represents the horizontal convergence of q. The GFS directly provides q at 2-m AGL, but at other levels it must be computed using the temperature and relative humidity at the desired level, as partially shown by (4) (Murray 1967), which yields the saturation vapor pressure, e_s, in hPa.

$$e_s = 6.11 * \left(10^{{}^{7.5T}/_{237.3+T}}\right) \tag{4}$$

In (4), T is the air temperature in °C. The saturation mixing ratio, w_s, in g kg^{-1}, is given by (5) (Brice and Hall 2013a).

$$w_s = 621.97\left[{}^{e_s}/_{(P_{station} - e_s)}\right] \tag{5}$$

In (5), $P_{station}$ is the station pressure in hPa and e_s, in hPa, is from (4). Then (6) relates the relative humidity in %, RH, and w_s, to the mixing ratio, w, in g kg^{-1}.

$$w = \left({}^{RH}/_{100}\right)w_s \tag{6}$$

One can assume that w is roughly equivalent to q. This procedure results in grids of q (in g kg^{-1}) at 300, 400, 500, 700, 850, 925, and 1013.25 hPa (at 1013.25 hPa, the GFS-provided 2-m AGL q is used directly, and u and v come from the GFS 10-m AGL wind components). MFC and q are not calculated at 150, 200, or 250 hPa because most of the

studies referenced in Banacos and Schultz (2005) interested in the relationship between MFC and precipitation neglect these low pressure levels due to the presumed lack of moisture at these levels.

GFS data are provided to end users on unprojected latitude-longitude grids. Therefore, calculating vertical and horizontal gradients of q, u, or v on these grids actually yields, for example, $\partial q/\partial \lambda$ and $\partial q/\partial \phi$, the rates of change of q with respect to longitude, λ, and latitude, ϕ, respectively. However, $\partial \phi/\partial y$ and $\partial \lambda/\partial x$, the rates of change of latitude with respect to y and of longitude with respect to x can be used to obtain rates of change of q, u, and v with respect to x and y, like $\partial q/\partial x$ and $\partial q/\partial y$. Horizontal moisture convergence, moisture advection, and MFC are each stored in gridded form at the same levels upon which q was computed.

Another set of derived predictors considered in this study consists of wind speeds computed upon the 150-, 200-, 250-, 300-, 400-, 500-, 700-, 850-, and 925-hPa and 10-m AGL surfaces using the formula for vector length, L, which is given by (7).

$$L = \sqrt{(u^2 + v^2)} \tag{7}$$

L is intended to act as a simple proxy for the strength of flow at a particular level of the atmosphere. One can sum the u- and v-components of the winds at the 500-, 700-, 850-, and 925-hPa levels and then determine the magnitude of the subsequent vector using (7). Depending on which of the levels are included in this calculation, the result is either the 500-700-, 500-850-, or 500-925-hPa layer-mean wind. (For moderate and high elevation areas, respectively, the 500-925-hPa layer-mean and the 500-925- and 500-850-hPa layer-mean winds are neglected). Speed shear is also computed by subtracting the 700-, 850-, and or 925-hPa wind speed from the 500-hPa wind speed and normalizing the result

by the distance between the geopotential heights of the two layers (Markowski and Richardson 2006). Neither the 500-850-hPa nor the 500-925-hPa speed shears are used in the high elevation model, and the 500-925-hPa speed shear is not used in the moderate elevation model.

The K index is computed for the dataset following (1). The 850- and 700-hPa dew point temperatures, T_d, in °C, are determined from the corresponding saturation vapor pressure and relative humidity using (8) (Brice and Hall 2013b).

$$T_d = {237.3\ln\left({e_s RH}/{611}\right)}\Big/{7.5\ln(10) - \ln\left({e_s RH}/{611}\right)} \tag{8}$$

In the high elevation region, the K index is neglected because it relies upon GFS *RH* and *T* at the 850-hPa level.

Previous studies have suggested that the precipitable water (PW) anomaly is a better predictor of flash flood potential than PW itself (Giordano 1994, Schroeder et al. 2016a). Standardized PW anomalies are calculated for the entire GFS dataset (2 March 2004 to 31 December 2016) following this procedure: 1) calculate the average GFS PW value for each grid cell and each month, 2) calculate the standard deviation of GFS PW for each grid cell and each month, and 3) create standardized PW anomaly grids every six hours by comparing each GFS PW grid through time to its corresponding monthly model average and monthly model standard deviation. Comparing the model PW to the model climatology accounts for any bias relative to observations that may be present in the GFS PW fields.

The 46 derived predictors available for use in this study are summarized in Table 4. Once produced in gridded form and stored as GeoTIFFs, these gridded derived predictors can be converted to text and merged into the predictor matrix from Table 3.

Table 4. Summary of derived predictors used in the study

Field Name	*Units*	*# of Levels*
Wind magnitude	m s^{-1}	10
Specific humidity	g kg^{-1}	7
Horizontal moisture convergence	g kg^{-1} s^{-1}	7
Horizontal moisture advection	g kg^{-1} s^{-1}	7
Moisture flux convergence	g kg^{-1} s^{-1}	7
K index	dimensionless	1
Standardized PW anomaly	dimensionless (σ)	1
Speed shear	s^{-1}	3
Layer-mean wind	m s^{-1}	3

Flood Events in Storm Data

Although past efforts have been made to obtain reports of flash floods from automated systems including U.S. Geological Survey stream gauges (Gourley et al. 2013), the only comprehensive national collection of reports of flash floods is contained within the NWS *Storm Data* publication (MacAloney 2016). Despite the fact that *Storm Data* reports are collected by professional meteorologists and hydrologists, any human-augmented reporting system is subject to variations in population density, diurnal cycles of human activity, and more mundane transcription or memory errors that affect the timing and location of reports (Barthold et al. 2015). Evidence of these issues has been found in assessments of FFG skill (Clark et al. 2014); at least one study has found that the distribution of *Storm Data* reports of flash floods is affected by the distribution of the human population (Marjerison et al. 2016) and similar issues have been noted for years in the reporting of other hazardous weather events (Frisbie 2006). Errors in timing or location can be accounted for by subtracting time from the start of the report, adding time

to the end of the report, or increasing the effective size of the report (Gourley et al. 2012, Clark et al. 2014).

In the present study, *Storm Data* reports have been obtained for the period starting 1 October 2006 and ending 31 December 2015. From 1 October 2006 to 30 September 2007, the local Weather Forecast Offices (WFOs) that initially collect *Storm Data* reports had the option of delineating the location of the event via either a polygon (with up to 8 vertices) encompassing the event's impacts or a single point designed to represent the center of the observed impacts. After 1 October 2007, all WFOs were required to adopt the polygon-based reporting methodology (Gourley et al. 2013).

Each flash flood must be associated in space and time with a set of predictor variables. This is accomplished by first locating the GFS3 grid cell in which the report occurred (for the polygon reports, this is the centroid, and for the point reports, it is the report location). Figure 6 is a map of the frequency, over the entire archive of reports, with which events occurred in each GFS3 model grid cell.

Because of the large size of the GFS3 grid cells, any additional expansion in the area of individual reports is probably unnecessary, but it is possible to envision remote scenarios in which a report is close enough to the edge of a GFS3 grid cell that a location error in the *Storm Data* publication could have an adverse effect upon the ML prediction process. A report is considered "active" for a particular GFS time if the start time of the report occurs within six hours after that GFS cycle begins. For example, consider a 0000 UTC GFS run and a *Storm Data* report that begins at 0300 UTC. The hypothetical report is counted as being active at 0000 UTC because 0300 UTC is only three hours after the start of this GFS cycle. If more than one report is valid in the same grid cell at the same

time, only one of the reports is stored in the predictor matrix; the rest are neglected. If the *Storm Data* report lasts longer than six hours, it is considered "active" only for the first GFS cycle that transpires in the six hours prior to the beginning of the report. This is to ensure that – to the extent possible – the ML method is only considering *pre*-flash flood environments.

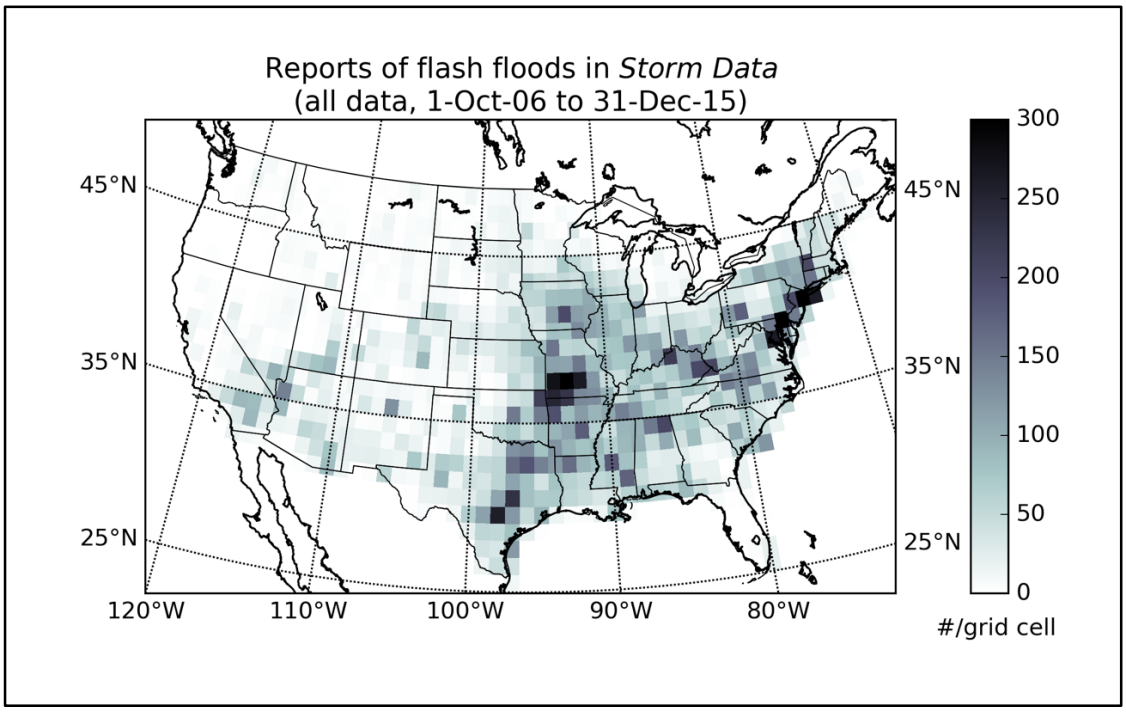

Figure 6. Number of NWS *Storm Data* reports of flash floods (N = 33,072) per grid cell over the entire archive

Finally, these processed *Storm Data* reports (predictands) are merged in space and time with the predictor matrix. This event extraction and storage process is intended to focus on the environmental variables associated with the *start* of each flash flood at the expense of those present at the *end* of each flash flood, in an effort to avoid contaminating the dataset with persistent flash flood impacts occurring after the environment that originally led to the flash flood has advected away, or been modified or displaced in some way.

Development of Training and Testing Datasets

The predictor matrix resulting from the previous section must be checked for invalid data prior to its use in an ML model. Each GFS field (i.e., each column of the predictor matrix) is checked for the presence of unrealistic values. A few GFS model analysis cycles were found to contain corrupt or unphysical data; cases (i.e. rows in the predictor matrix) containing these data have been excised from the prediction matrix. A more frequent problem is the result of the CONUS mask displayed in Figure 3 encompassing grid cells outside the domain of the GFS land surface model. Because the domain of the GFS land surface model is not consistent through time, this process must be undertaken at the beginning of each ML model fitting trial; these cases can then be removed from the predictor matrix prior to any of the activities discussed below.

There are 144 total parameters for the low elevation case: 97 GFS predictors, 46 derived predictors, and flashiness (but not elevation, since it will be used to partition the dataset prior to the testing-training split). In the literature, there is some disagreement regarding the appropriate relative sizes of the testing, training, and validation datasets for ML. For instance, Guyon (1997) proposed a relationship for the ratio between the size of the validation (or testing set) and training sets based on upon the number of predictors. In this case, the proposed relationship recommends using approximately 8% of the available cases for testing and the remaining 92% for training. When moderate elevations are considered, 19 variables defined at the 925- or 1013.25-hPa levels must be neglected, and for high elevations, an additional 15 850-hPa variables are also removed from the mix. Applying these results to Guyon (1997)'s formula results in a recommendation to use 9% and 10%, respectively, of the moderate and high elevation cases for testing. However, Williams (2009), Trafalis et al. (2014), and Ahijevych et al. (2016) each used

50% of their total data for testing; Guyon (1997)'s formula would have recommended using between 17% and 25% of the total data for testing in those studies. In the present study, using the Guyon (1997) formula (via, for example, holding out 9% to 10% of all data points by, perhaps, selecting those cases occurring on the 10^{th}, 20^{th}, or 30^{th} of each month) results in test datasets containing too few flash floods, especially when regional or temporal subsets of the data are being considered. Therefore, in the present study, a compromise between the approaches outlined in previous applications of RF to meteorological problems and the Guyon (1997) formula is employed. Approximately 20% of the total cases are held out for independent testing purposes; these cases are collected by storing all data points from the 5th, 10th, 15th, 20th, 25th, or 30th day of each month. Within the remaining 80% of cases, a 75/25 split is executed, this time randomly and without replacement, resulting in "training" and "validation" datasets, respectively. Figure 7 is a map, prior to processing the Storm Data reports into GFS3 grid cells, of all the flash floods in the archive with a start time occurring on the 5th, 10th, 15th, 20th, 25th, or 30th of any month. Figure 8 is a map of all other Storm Data reports in the archive (i.e., those that will make up the validation and training datasets).

Table 5 summarizes important characteristics of several time periods of interest in the present study. Three recent GFS model "epochs", or time periods in which no major upgrades to the GFS model core or postprocessor were implemented, are identified to test for changes in the statistical properties of GFS model fields as a result of GFS upgrades. When considered together, these three GFS epochs will be referred to as the "study period" (5 September 2012 – 31 December 2015), while the 1 October 2006 – 31 December 2015 period will be referred to as the "entire archive".

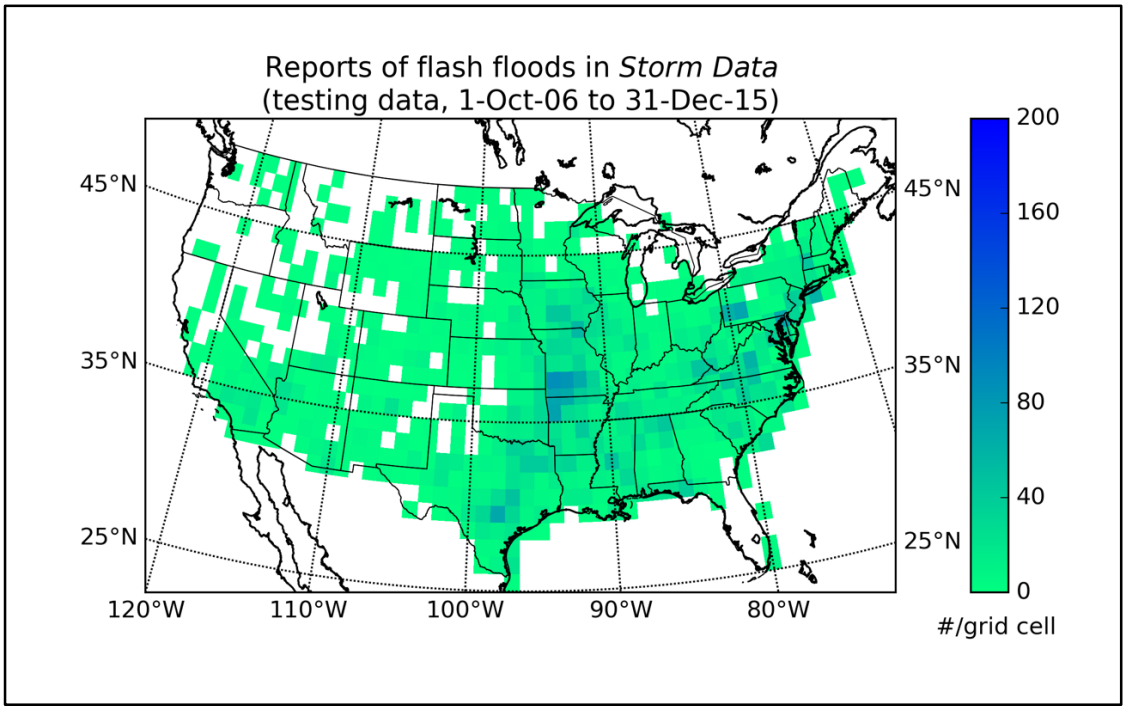

Figure 7. Number of NWS *Storm Data* reports of flash floods occurring on test days (N = 6,607) per grid cell over the entire archive

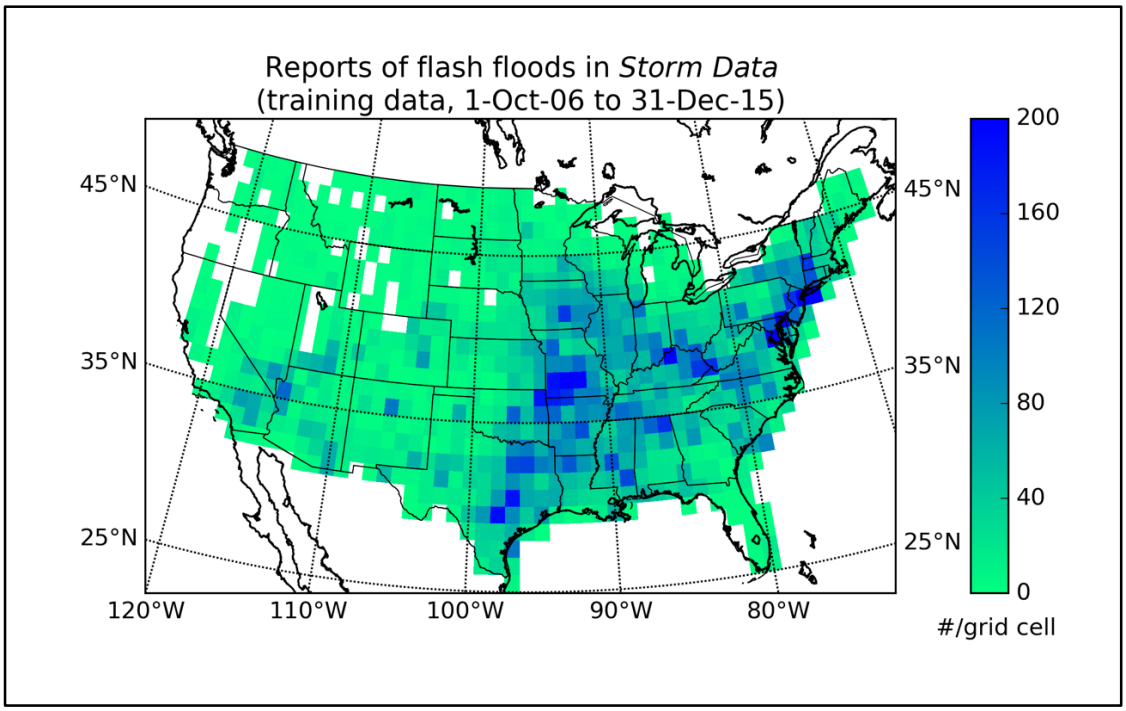

Figure 8. Number of NWS *Storm Data* reports of flash floods occurring on test days (N = 26,465) per grid cell over the entire archive

Despite the fact that the 15 January – 31 December 2015 ("2015") epoch has more than double the rate of flash floods than either the 21 August 2013 – 14 January 2015 ("2014") epoch or the 5 September 2012 – 20 August 2013 ("2013") epoch, the salient characteristic of this predictor matrix is the extreme rarity with which flash floods are observed. The ratio of flash flood-containing GFS3 grid cells to GFS3 grid cells without flash floods is 1:694 for 2013, 1:702 for 2014, 1:338 for 2015. It is 1:576 for the study period and 1:586 for the entire archive of *Storm Data* reports. Table 5 also contains information indicating how many flash flood and non-flood cases are removed from the testing, training, validation datasets as a result of unavailable GFS land surface model outputs, corrupt GFS model fields, or the process of merging multiple *Storm Data* reports into a single GFS3 grid cell.

Table 5. Summary of important characteristics of GFS model epochs

All elevations, both testing and training data

Epoch	*Cases deleted, %*	*Flash floods deleted, %*	*Usable cases*	*Usable flash floods*	*Flash floods, % of total*
2013	8.2	45	1,199,060	1,729	0.14
2014	8.2	40	1,733,248	2,469	0.14
2015	19	56	735,482	2,174	0.30
Study period	11	48	3,667,790	6,372	0.17
Entire archive	8.7	43	11,004,540	18,787	0.17

Table 6 contains the final size of each testing and training set for each elevation region and each epoch. From Table 6, a major pitfall in restricting the ML model fitting process to a particular elevation region and GFS model epoch is apparent. The number of flash floods available in some of these subsets is relatively small, reaching a low of 18 cases when the testing data for the middle elevation region of the 2013 GFS model epoch is considered. In general, the number of flash flood cases in the testing dataset for the study period is still fairly low, but likely adequate for the purposes of the present study.

The number of flash flood cases in the entire archive is more than adequate, but requires the inclusion of GFS model output from a 9.5-yr period, over which major changes were made to the GFS model.

Table 6. Sample sizes (total cases and flash flood cases) for each combination of GFS model epoch and elevation region

		Training		*Testing*	
Epoch	*Elevation*	*# cases*	*# flash floods*	*# cases*	*# flash floods*
2013	Low	583,105	1,163	140,541	259
2013	Middle	176,808	121	42,577	18
2013	High	206,314	132	49,715	36
2014	Low	843,170	1,303	201,795	301
2014	Middle	256,198	281	61,391	58
2014	High	298,946	184	31,618	46
2015	Low	344,960	1,360	82,456	362
2015	Middle	116,101	175	27,705	47
2015	High	132,642	184	31,618	46
Study period	Low	1,771,235	3,826	424,792	922
Study period	Middle	549,107	577	131,673	123
Study period	High	637,902	728	153,081	196
Entire archive	Low	5,342,292	12,163	1,284,578	3,094
Entire archive	Middle	1,627,455	1,340	391,919	295
Entire archive	High	1,900,545	1,537	457,751	358

The entire dataset and all of its subsets are extremely unbalanced between the majority (no flood) and minority (flash flood) classes. In the testing dataset, this imbalance comes with the territory; since the imbalance is the result of the true prevalence of *Storm Data* reports of flash floods, no modifications to the test dataset should be made. However, in the training dataset, these unbalanced classes are a problem. If left unremedied, they result in ML models that could forecast "no flood" every time without fear of reprisal, because the flash flood class is so rarely encountered in the fitting process.

Unbalanced datasets are a common issue in ML, particularly as they relate to meteorology and forecasting relatively rare weather hazards. Trafalis et al. (2014) dealt

with a tornado dataset in which 6.7% of their total records fell in the minority class; they set the minority to majority class ratio in the testing as well as the training and validation sets at 7:100. On the other hand, Williams (2009) undersampled the majority class in his training datasets (while maintaining the true class split in the validation and testing datasets), until he achieved a 30:70 minority-majority ratio. In his study, the true prevalence of the minority class was 0.25% in one dataset and 1.33% in the other. In the present study, the majority class is undersampled to achieve a 50:50 split in the testing data, while maintaining the true ratio between classes in the validation and testing sets. Therefore, after the 75/25 training/validation split is executed, all flash floods falling into the training set are retained, along with an equal number of randomly selected non-events from the "75" side of the 75/25 split.

Results

Performance Metrics

The implementation of RF used in the present study is that from the scikit-learn Python library (Pedregosa et al. 2011). An RF is grown on the training dataset and then the validation dataset is run through the new RF to produce predictions. These RF predictions are compared to reality (i.e., *Storm Data*), and the skill of the forecasts is quantified via the Brier score (Brier 1950). Brier called this score "P" and expressed it in the form shown here as (9).

$$P = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}\left(f_{ij} - E_{ij}\right)^2 \tag{9}$$

The outcome, E_{ij}, is 1 when the event, i, took place in the class, j (out of a total of r classes), and 0 when the event did not. The forecast, f_{ij}, is the forecast of the likelihood of i occurring in j. The total number of forecasts is represented by n. When the forecast

is answering a yes-or-no question (i.e., "Will there be a flash flood?"), $j = 2$ and the Brier score can be simplified into (10).

$$P = \frac{1}{n}\sum_{i=1}^{n}(f_i - E_i)^2 \tag{10}$$

From this simplified definition, it is readily apparent that the value of the worst possible Brier score is 1, which happens when f_i and E_i are always as far apart from one another as possible (if the forecast probability is 100% and the event does not occur or if the forecast probability is 0% and the event does occur). On the other hand, the best Brier score is 0, which happens when the forecast probability is 0% and the event does not occur, or when the forecast probability is 100% and the event does occur.

Contingency tables are frequently used to assess the quality of deterministic predictions. Table 7 is a 2x2 contingency table, used with both the forecast and the outcome are binary. Hits, *a*, occur when an event is forecast and the event is observed. False alarms, *b*, occur when an event is forecast but the event does not occur. When an event is not forecast but does occur, the result is a miss, *c*. When an event does not occur and was not forecast to occur, the case is labeled as a correct negative, *d*.

Table 7. Example contingency table

		Was the event observed?		
		Yes	No	Totals
Was the event forecast?	Yes	Hit (a)	False alarm (b)	a + b
	No	Miss (c)	Correct negative (d)	c + d
	Totals	a + c	b + d	n = a + b + c + d

The contingency table can be used to derive several metrics, including the probability of detection (POD), false alarm rate (FAR), critical success index (CSI or Gilbert Skill Score [GSS, Stephenson {2000}]), and more. The POD is given by $a/(a+c)$, the FAR by $b/(a+b)$, and the CSI by $a/(a+b+c)$. CSI, which ranges from zero to one, where zero is undesirable and one is desirable, is also called the "threat

score". CSI is non-linearly related to both the POD and the FAR. Although originally proposed for the verification of rare events, the CSI is not zero for random forecasts or climatological forecasts, so it cannot measure skill relative to either of these conditions (Stephenson 2000). POD ranges from zero to one; a score of zero means no events were detected and a score of one means all events are detected. The FAR ranges from one to zero; a score of one indicates all forecasts were false alarms and a score of zero indicates no forecasts were false alarms. Confusingly, G. K. Gilbert's name is also used in connection with a second skill measure sometimes referred to as either the "Gilbert score" (Schaefer 1990 in Hogan et al. 2010) or the "equitable threat score" (ETS, Hogan et al. 2010). Despite the common name, the ETS is not "equitable", in the sense that different random forecasting systems will return different values of the ETS (Hogan et al. 2010). The misnamed ETS is given by $(a - a_r)/(a + b + c - a_r)$ where a_r is $(a + b)(a + c)/n$, which represents the fraction of hits expected from a random forecast (Hamill and Juras 2006). The ETS ranges from -1/3 to zero, where values greater than zero indicate a forecast with more skill than a random forecast. Finally, the Peirce skill score (PSS) is given by (11).

$$PSS = POD - POFD = {(ad - bc)}\Big/{(a + c)(b + d)} \tag{11}$$

In (11), *POFD* refers to the probability of false detection. The PSS is equitable, unlike the ETS or the GSS (Hogan et al. 2010). The PSS ranges from negative one to one, where values greater than zero indicate a forecast with skill; scores closer to one have greater skill. However, the PSS has one disadvantage, namely that if the correct negatives are large, the false alarms are relatively important to the final score (Stephenson 2000).

The ROC (receiver operator characteristic) diagram is a plot of the true positive rate (TPR) as a function of the corresponding false positive rate (FPR, Trafalis et al. 2014) for various contingency tables created as a result of applying different thresholds to a classifier. The FPR is given by $b/(b+d)$ and the TPR is equivalent to the POD; the curve formed from this relationship is called the "ROC curve"; Figure 9 is an example.

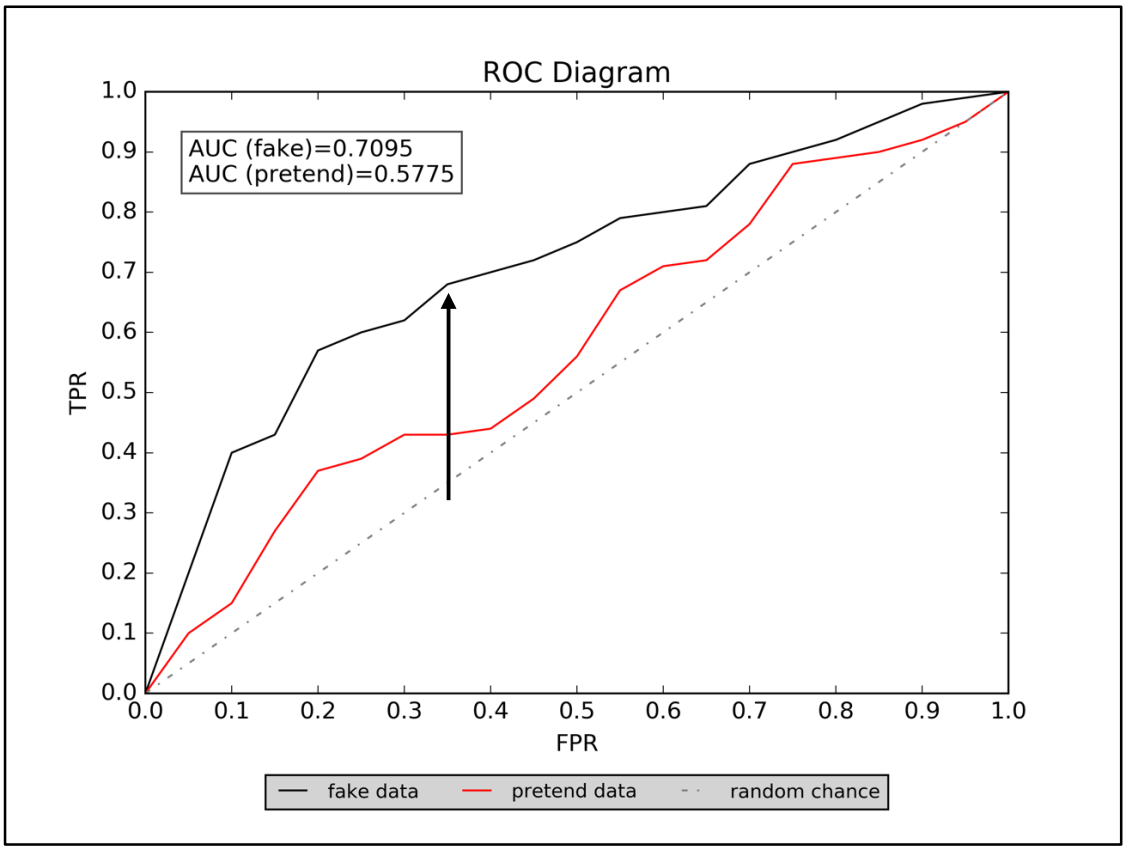

Figure 9. Example ROC diagram

A random classifier would possess a ROC curve following the dash-dotted 1:1 line; better-skilled classifiers will possess ROC curves progressively closer to the top-left of the diagram, where the TPR = 1 (i.e., the classifier detects all the events) and FPR = 0 (i.e., the classfier does not result in any false alarms). The PSS for a particular threshold is equal to the vertical distance from the no-skill line to the ROC curve (Manzato 2007), as in Figure 9, where the vertical arrow stretches from the 1:1 line to

the ROC curve of the "fake data". A ROC curve can be summarized by a single number: the AUC (area under the curve). The AUC ranges from zero to one; a perfect AUC would be achieved by a ROC curve that hugs the left and top borders of the ROC diagram while an AUC of 0.5 is achieved by a ROC curve that follows the 1:1 line.

Determining Optimal Random Forest Parameters

An important consideration in fitting an RF model is determining acceptable values of a handful of parameters that govern the performance of the RF model. Chief among these is *ntree*, the number of trees in an RF. More trees act to reduce the variance and the out-of-bag (OOB) error rate. Typically, the OOB error rate converges to some minimum level once a certain value of *ntree* is reached. A second important RF parameter is *mtry*, the number of predictors from which the RF method can select that predictor resulting in the optimal split between labels at a given node. Breiman (2001) showed that using the integer value of the square root of the total number of predictors ("sqrt") worked well as *mtry* for a number of different datasets, but noted that the final choice of *mtry* is problem-dependent. Finally, maximum tree depth is another important characteristic of the forest. This tree property is actually a function of several additional RF parameters, including the minimum number of samples required to split an internal node (two, in the present study) and the minimum number of samples (one) required for a new leaf to be generated. Those selected values do not preclude a tree from reaching its maximum possible depth (i.e., where the tree grows until each leaf is 100% pure). For that reason, the problem of maximum tree depth in this study can be reduced to a single RF parameter – *dtree*. *Dtree* is the maximum number of levels of nodes each tree is allowed to contain. Deeper trees reduce the bias but increase the variance of predictions from the forest. Due

to the interactions between these parameters, fixing the values of *ntree*, *dtree*, and *mtry* is an optimization problem, where OOB error rate is a function of the three parameters.

Figure 10 is a plot of the partial solution to this optimization problem for the low elevation regime using training data drawn from the entire archive, where the OOB error rate is the objective function to be minimized.

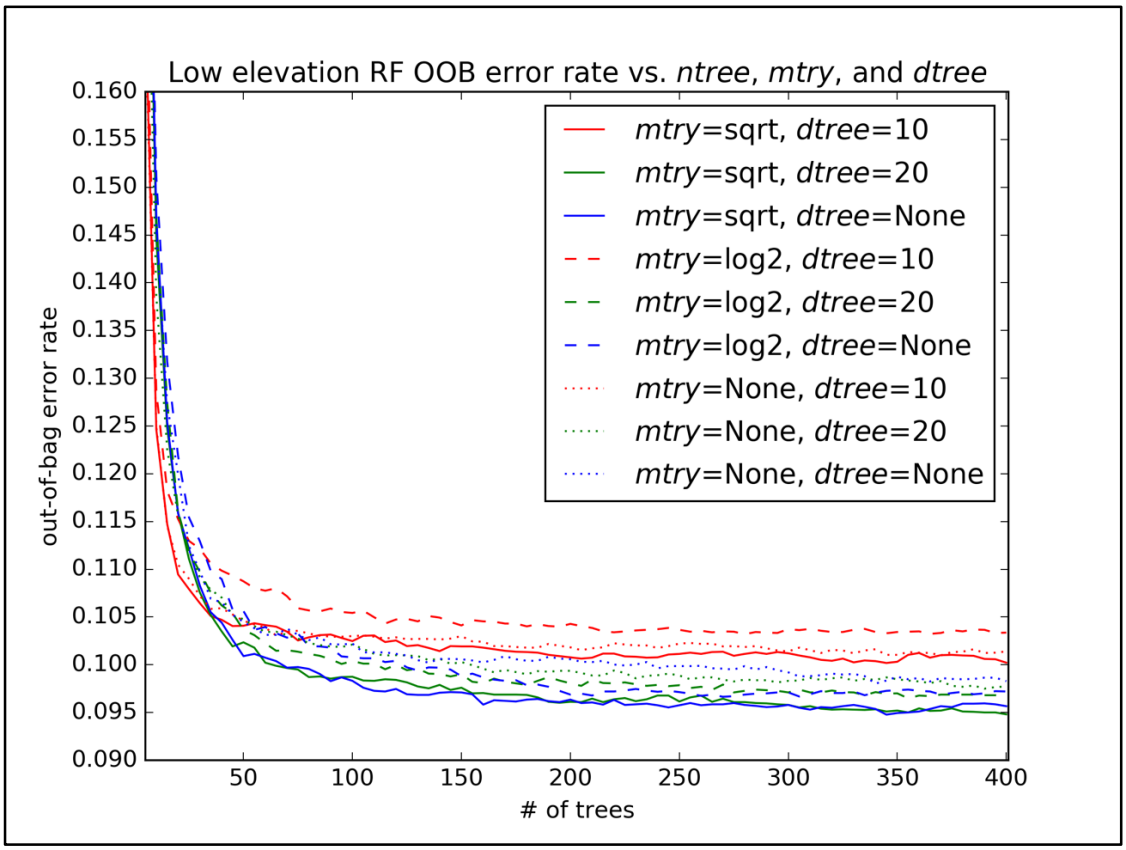

Figure 10. OOB error rate as a function of the *ntree*, *mtry*, and *dtree* RF parameters

Independent testing and validation datasets are not required for this trial because it is intended only to optimize an internal RF metric and observe the sensitivity of the OOB error rate to said parameter optimization process. All flash floods occurring in the low elevation areas and during the study period are included ($N = 15{,}257$). Following the undersampling procedure discussed previously, an equal number of non-flood cases from

the low elevation area and the study period are randomly selected; this results in a training dataset with N = 30,514, split evenly between flash floods and non-floods.

In this test, *dtree* can be either 10, 20, or "None", where "None" means that the depth of a tree is not restricted. When *dtree* is "None" (blue lines in Figure 10), trees range in depth from 18 to 40 layers, as shown in Table 8. The distributions of the unrestricted tree depths for the low and middle elevation regimes are nearly identical, but for the high elevation cases, the unrestricted trees are deeper, which implies that more predictor variables, and therefore, more tree levels, are required to completely split flash flood cases from non-flood cases. Note that *ntree* = 400 in Table 8 because the tree depths are recorded at the end of the experiment explained by Figure 10.

Table 8. Tree depths when *dtree* = "None", *ntree* = 400, and *mtry* = "sqrt"

Elevation	*Tree depth range*	*Avg. tree depth*	*Med. tree depth*	*Mode tree depth*
Low	18 to 40	25	25	23
Middle	17 to 39	24	23	23
High	22 to 49	32	31	29

Overall, the *dtree* analysis (comparing between colors in Figure 10) shows that "None" (blue) is best, 20 (green) is second, and 10 (red) is generally worst. However, the differences are small and lower *dtree* values do result in a slightly faster RF growth process. Given the minor differences in OOB error rate and compute time introduced by changes in *dtree*, *dtree* will be "None" for the remainder of the study.

In Figure 10, the *mtry* parameter can be "sqrt" (solid lines), "log2" (the integer value of the base-2 logarithm of the total number of predictors, represented by dashed lines), or "None" (the predictors are not randomly subsampled when a new node is generated, represented by dotted lines). Comparison of the *mtry* values for given values of *ntree* and *dtree* shows that the solid ("sqrt") or dashed ("log2") lines are generally

better than "None" (dotted lines), with exception of the dashed red line, which represents *dtree* = 10 and *mtry* = "log2". Regardless of the values of *mtry* or *dtree* selected, the OOB error rate quickly asymptotes to a minimum value around 0.1 once *ntree* reaches 150.

Figures 11 and 12 are equivalent to Figure 10, but for moderate (N = 3,270, split evenly between flash floods and non-flood cases) and high elevations (N = 3,790, again split evenly), respectively.

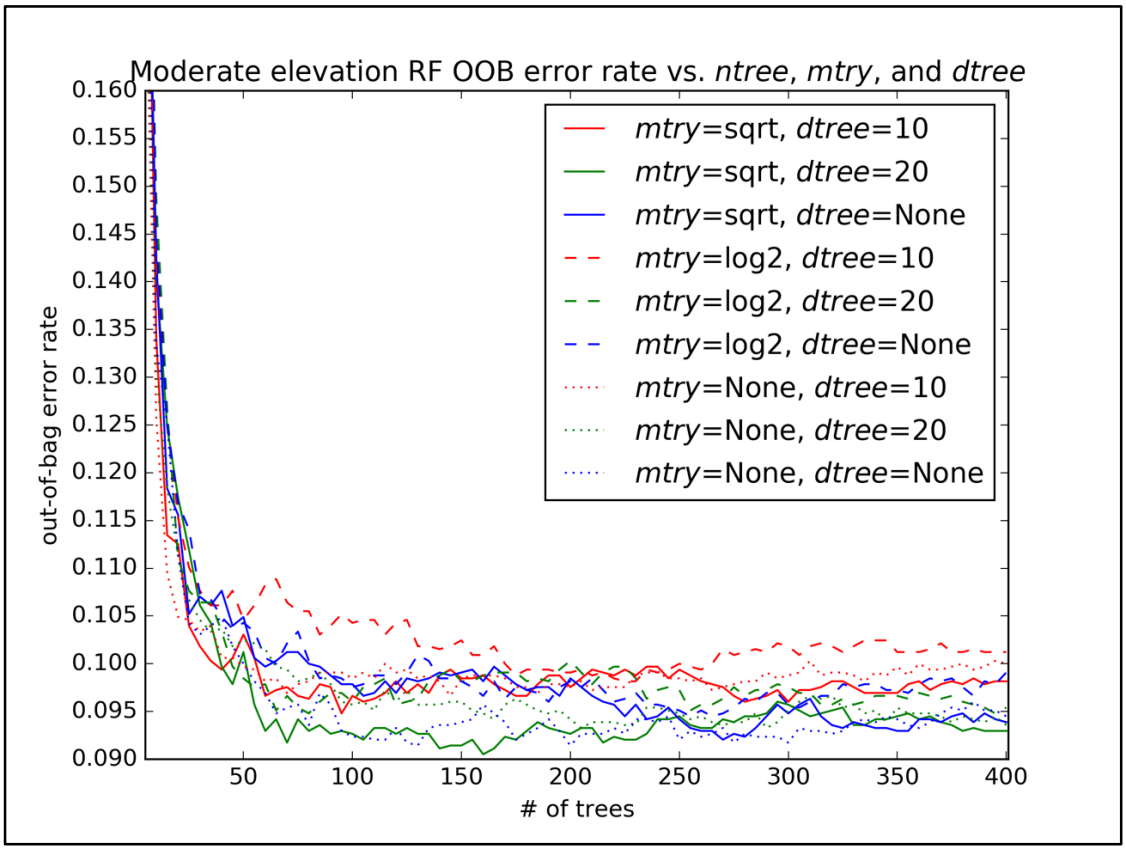

Figure 11. As in Figure 10, but for moderate elevation data

The patterns in Figure 11 are less clear than those in Figure 10; Figure 11 is also generated from a dataset with much smaller N, which results in a "noisier" appearance. The evidence from Figure 11 suggests that 100 trees are sufficient to reduce the OOB error rate to its maximum achievable level, while changes in *dtree* and *mtry* have smaller impacts upon the OOB error rate. In any case, the differences in OOB error rates between

the parameter combinations are quite small, which tends to verify past studies: the RF OOB error rate is sensitive mostly to the number of trees in the forest.

In Figure 12, 50 trees are enough to minimize the OOB error rate. Like when low elevation cases are considered, the combination of *mtry* = "sqrt" and *dtree* = "None" (the blue line) results in good performance, although setting *dtree* = 20 (the green line) is also competitive. However, the differences between these lines are minor.

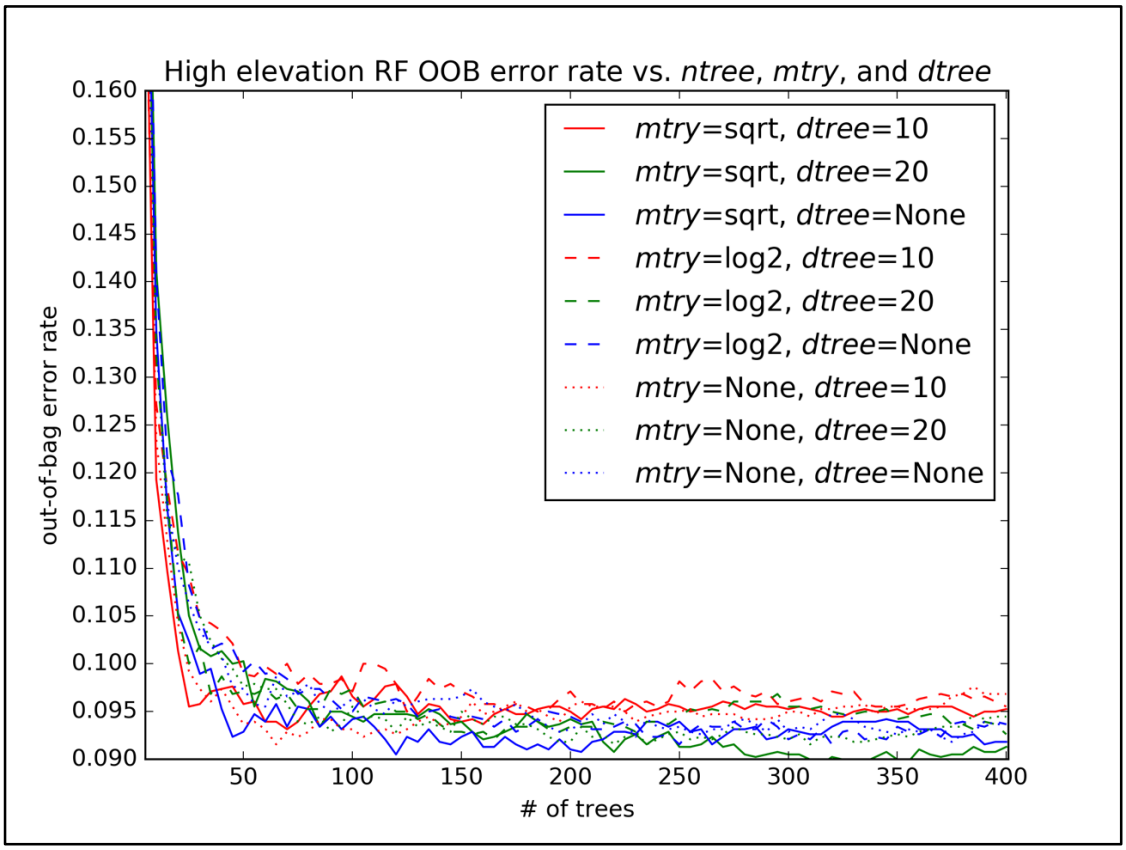

Figure 12. As in Figure 10, but for high elevation data

To be safe, *ntree* will be set to 300, which is safely within the region of lowest OOB error rate in all three elevation areas. At a value of 300, for example, any combination of the other two predictors in any elevation area is represented by an OOB error rate that varies between 0.090 and 0.105, which indicates that the OOB error rate for that number of trees is relatively immune to the choice of *mtry* or *dtree*. On this basis,

Breiman (2001)'s original recommendation is suitable in this case: set *mtry* to "sqrt" and *dtree* to "None" (solid blue line) for all three elevation conditions.

Because randomness is inherent in the RF growth process (see Chapter 2) and in the division of cases between validation and testing data, multiple iterations of random forest training, validation, and testing are advisable. These iterations help to quantify the variability introduced into the process as a result of these sampling procedures. Separate forests were grown for each of the elevation regions and the three GFS epochs, the study period, and the entire archive; this procedure was repeated 50 times for each elevation-epoch combination, which allows for the distribution of resulting Brier scores to be well-characterized. Then this procedure was repeated, but with 15 trials instead of 50. The results of this are summarized in Table 9.

Table 9. Examination of sample size and random subsampling process upon RF skill

Elevation	*Epoch*	*Mean Brier score (50 trials)*	*Mean Brier score (15 trials)*	*Std. dev of Brier score (15 trials)*	*% of 50-trial range shown by 15 trials*
Low	2013	0.096	0.094	0.005	64
Moderate	2013	0.10	0.097	0.01	63
High	2013	0.086	0.084	0.009	60.
Low	2014	0.098	0.095	0.006	62
Moderate	2014	0.097	0.091	0.01	90.
High	2014	0.095	0.097	0.006	61
Low	2015	0.10	0.11	0.006	94
Moderate	2015	0.13	0.13	0.02	1.0×10^2
High	2015	0.13	0.13	0.01	46
Low	study	0.096	0.097	0.003	77
Moderate	study	0.10	0.10	0.007	77
High	study	0.099	0.10	0.006	49
Low	entire	0.094	0.093	0.001	46
Moderate	entire	0.099	0.099	0.004	58
High	entire	0.088	0.089	0.004	87

From Table 9 is it readily apparent that 15 trials result in Brier score distributions similar to those observed when 50 trials are executed, instead. The mean Brier score

calculated from 15 trials is similar to that calculated from 50 trials. The percent of the 50-trial range in Brier scores shown when only 15 trials are conducted ranges between 46 and 100 percent, as shown in Table 9.

The contingency tables for each of the 15 or 50 trials can also be examined to observe how the 75/25 training/validation split affects the number of flash floods available in the validation set for each trial. The results of this analysis make up Table 10.

Table 10. Examination of random sampling process upon the number of flash floods available for RF validation

Elevation	*Epoch*	*Mean # of validation flash floods (50 trials)*	*Mean # of validation flash floods (15 trials)*	*Std. dev. (15 trials)*	*% of 50-trial range shown by 15 trials*
Low	2013	871	872	11	70.
Moderate	2013	91	91	4.0	67
High	2013	99	101	6.0	88
Low	2014	978	977	16	79
Moderate	2014	212	210	8.2	97
High	2014	308	307	11	1.0×10^2
Low	2015	1,021	1,015	20.	1.0×10^2
Moderate	2015	132	129	4.9	86
High	2015	138	138	4.1	75
Low	study	2,864	2,865	24	92
Moderate	study	435	438	8.3	74
High	study	546	542	7.2	63
Low	entire	9,124	9,116	39	78
Moderate	entire	1,004	1,002	15	67
High	entire	1,154	1,155	11	54

As with the above analysis of the impact of the number of trials upon Brier score, 15 trials are sufficient to capture much of the variability resulting from the training/validation dataset split, between 54 and 100 percent of the 50-trial range in the number of flash floods in the validation dataset accounted for with just 15 trials. On the basis of this evidence, 15 trials generally represent *most* of the variability in RF prediction skill introduced as a result of the RF bagging process, the random subsampling of

predictor variables available at each node of each tree, and the random splitting between validation and testing datasets in each trial.

Comparisons with Other Machine Learning and Statistical Techniques

To test the applicability of the RF method to this problem relative to other classification and regression techniques, I conducted 15 trials in which I applied the training and testing datasets for the entire archive to three other techniques implemented in the scikit-learn library. Table 11 contains the results of this set of trials. In Table 11, each mean skill score or metric shown is statistically compared to the corresponding "RF (tuned parameters)" value via a two-tailed heteroscedastic *t*-test (Sawilowsky et al. 2002). Differences significant at the 95% level are shown in color-coded cells in Table 11, where orange-shaded values represent a statistically-significant degradation in the mean value of the metric relative to the tuned RF and green-shaded values represent a significant improvement in the mean value of the metric relative to the tuned RF. In the "RF (default parameters)" trials, *ntrees* is 10, *dtrees* is "None", and *mtry* is "sqrt"; the other statistical methods are deployed with their respective default parameters, as well.

All five methods detect (POD) greater than 80% of *Storm Data* flash floods in each elevation regime. In the low elevation regime, the Naïve Bayes (NB, Hand and Yu 2001) is statistically similar to the tuned RF for this metric, while logistic regression (LR, Cox 1958), the gradient boosting classifier (GBC, Freidman 2001) and the default RF are all worse than the tuned RF. The ETS is barely positive for all five methods, which indicates that they are all more skillful, though barely, than the chance forecast. GBC has the best Brier score of the five methods. The tuned RF is next, followed by LR, the default RF, and NB. The PSS indicates that the GBC method is the most skillful, followed by the tuned RF, either the LR or the untuned RF, and then the NB method.

Table 11. Comparison of ML and statistical methods

	RF (tuned parameters)	*RF (default parameters)*	*Naïve Bayes*	*Gradient boosting classifier*	*Logistic regression*
			Low elevation		
POD, %	92.7	87.6	92.4	92.4	91.9
FAR, %	98.3	98.4	99.0	98.2	98.4
CSI	1.64×10^{-2}	1.68×10^{-2}	9.97×10^{-3}	1.78×10^{-2}	1.60×10^{-2}
ETS	1.40×10^{-2}	1.45×10^{-2}	7.58×10^{-3}	1.54×10^{-2}	1.36×10^{-2}
PSS	7.96×10^{-1}	7.44×10^{-1}	6.99×10^{-1}	8.01×10^{-1}	7.82×10^{-1}
Brier score	9.28×10^{-2}	1.02×10^{-1}	2.15×10^{-1}	8.70×10^{-2}	9.53×10^{-2}
Average time (s) per trial	253	10.3	13.4	47.6	28.5
			Moderate elevation		
POD, %	93.6	88.7	90.7	93.1	90.4
FAR, %	99.5	99.5	99.7	99.5	99.5
CSI	5.00×10^{-2}	5.16×10^{-2}	3.43×10^{-3}	5.42×10^{-3}	4.73×10^{-3}
ETS	4.25×10^{-3}	4.41×10^{-3}	2.68×10^{-3}	4.68×10^{-3}	3.98×10^{-3}
PSS	7.95×10^{-1}	7.58×10^{-1}	7.08×10^{-1}	8.02×10^{-1}	7.60×10^{-1}
Brier score	9.61×10^{-2}	1.06×10^{-1}	1.93×10^{-1}	9.67×10^{-2}	1.10×10^{-1}
Average time (s) per trial	35.1	1.70	3.40	11.1	1.70
			High elevation		
POD, %	89.3	86.2	91.5	89.1	86.4
FAR, %	99.4	99.4	99.6	99.4	99.4
CSI	5.65×10^{-3}	5.84×10^{-3}	3.65×10^{-3}	6.04×10^{-3}	5.56×10^{-3}
ETS	4.87×10^{-3}	5.07×10^{-3}	2.87×10^{-3}	5.27×10^{-3}	4.78×10^{-3}
PSS	7.70×10^{-1}	7.47×10^{-1}	7.19×10^{-1}	7.76×10^{-1}	7.43×10^{-1}
Brier score	8.69×10^{-2}	9.57×10^{-2}	1.91×10^{-1}	8.73×10^{-2}	9.23×10^{-2}
Average time (s) per trial	38.8	1.70	3.50	12.4	1.60

The untuned RF is the fastest of the five methods in the low elevation case, with 10.3 seconds required to complete an average trial, which includes fitting the model, generating deterministic predictions on the test set, and calculating probabilities on the test set. The NB method is not much slower on the low elevation cases, with the LR and GBC slower still. The tuned RF (with 300 trees) is by far the slowest method on the low elevation cases, which shows how dramatically the average time per trial depends on the

number of trees in the RF. At moderate elevations, the untuned RF is again the fastest, though, on average, it is tied with the LR for these trials. NB is therefore in third place, while the GBC and the tuned RF are the two slowest methods. At high elevations, LR is the fastest method by a nose, with the untuned RF in second place, NB in third, and GBC and tuned RF again bringing up the rear. In general, Table 11 suggests that the NB, LR, and untuned RF methods are substantially worse for this problem, while the GBC matches or outperforms the tuned RF.

It is certainly plausible that some degree of tuning applied to either the LR or NB could result in comparable performance to that achieved by the tuned RF or the GBC. Note that, although there is a drastic increase in compute time occurring as a result of increasing the number of trees in the RF, the performance improves concomitantly, especially in the skill of the probabilistic predictions as shown by the Brier score. Fitting a model in four minutes or so is a fairly minor amount of computer time in the grand scheme of the weather forecasting and altering enterprise. Given RF's past use in solving meteorological problems, its desirable statistical properties, and its internal measures of importance and physical interpretation, the choice of RF as a ML algorithm is justified for this problem. However, this preliminary evidence would indicate that other ML and statistical methods are generally applicable to this problem, as well.

Differences Between Global Forecast System Epochs

Several 300-tree RF models are fit to data drawn from each combination of five separate time periods (identified in Table 5) and three elevation regions (identified in Figure 5). Fifteen RF models are generated for each of these 15 combinations of elevation region and time period; for each of the 15 elevation-time combinations, the RF that produced the best Brier score on its corresponding validation data is stored. The

importance of the epoch identification process can now be tested via cross-validation: for example, use the low elevation 2015 RF to generate predictions on the 2014, 2013, study period, or entire archive test data, and compare the resulting skill values, as shown in Table 12.

Table 12. Results of cross-epoch testing process

	Brier score rank relative to other trials using the same test data but on different GFS epochs		
GFS epoch	*Low elevation*	*Middle elevation*	*High elevation*
2013	3rd	4th	2nd
2014	5th	2nd	4th
2015	1st	1st	1st
Study period	2nd	3rd	4th
Entire archive	4th	5th	5th

To explain the results of Table 12, let us delve into an example. Start by noting that the middle elevation data from the 2013 epoch is marked "4th". This ordinal is the result of growing an RF upon the 2013 middle elevation training set. Using this RF, I then generated predictions on the 2013, 2014, 2015, study period, and entire archive middle elevation *test* sets and then compared the resulting Brier scores from each of the five. In this example, the 2013 RF model actually yielded its second-worst predictions on the 2013 test set (as marked by the ordinal in Table 12) and its worst on the entire archive test set. The third-best results were achieved on the study period test set, the second-best on the 2014 test data, and the best on the 2015 test data (not shown). If a particular RF fit were valid only for the GFS model epoch from whose training set it was generated, one would expect to see all cells in Table 12 marked "1st". One may infer that if there are statistically significant differences in the underlying distributions of the individual GFS model fields between GFS model epochs, an RF trained on data from a given epoch should produce more skillful predictions on independent test data drawn

from that same epoch (and statistical distribution). These results indicate that an RF generated from data drawn from one model epoch can be used to generate predictions as good or better on predictor data drawn from another model epoch, at least when the GFS model upgrades implemented between August 2012 and December 2015 are considered. The relatively poorer performance of the RFs fit to training data from the entire archive suggests that there *are* important differences in GFS model fields between the beginning of the archive (2006) and the end of the archive (2012-2015). (Note also that the test cases from 2015 are easier for the RF method, in general, to classify. It is possible that this is an unlucky artifact and that the events from the 5th, 10th, 15th, 20th, 25th, and 30th of the months in 2015 just happen to be dominated by situations in which the GFS produced more skillful forecasts. Another possible explanation is that the 2014-2015 model epoch transition resulted in a better, more skillful GFS.) On the basis of these results, GFS model epochs will generally not be used to divide predictors into testing, training, and validation data through the rest of the present study.

Random Forest Probabilities and Calibration

In an RF, each predictor ("tree") votes for the class into which it will sort each set of predictor variables. These vote counts can be construed as a measure of the confidence the forest has in a particular prediction. For example, if 250 of 300 trees vote to consider a particular set of predictors as belonging to the flash flood category, the predictor is more confident in labeling the data as belonging to the minority class than if only 150 of the 300 trees vote to do so. In both cases, however, using this classifier in deterministic fashion would result in a final label of flash flood being assigned to the predictor data; by default, a vote fraction of 0.50 is the threshold used to generate contingency table statistics. However, it is possible to convert these raw votes from

ensemble members (trees) into a calibrated probability that the set of predictors falls into a particular class. Reliability diagrams derived from raw RF votes often take the shape of a sigmoid function. In an RF, inter-tree variance tends to pull the final raw RF vote totals away from the ends of the [0, 1] probability interval. For example, for an RF to predict a probability of zero or one for a given case, all trees in the forest must vote the same way. Since the trees are generated via a bagging process, this is relatively unlikely (Niculescu-Mizil and Caruana 2005). Although past research has focused upon the use of sigmoid or isotonic functions to improve the reliability of probabilities derived from ML classifiers, it is possible to fit other types of functional relationships to the raw RF scores, as long as the functional relationship chosen is extendable to one or more independent testing dataset(s) (Williams 2009).

I use the RFs generated previously for the inter-epoch comparison test to explore the relationships between the RF-generated probabilities and historical observations. Each forest is used to generate predictions on its respective independent validation set; then these predictions are compared to the actual recorded *Storm Data* reports. Figure 13 is a type of reliability diagram that summarizes these results. All the records in the test dataset drawn from the entire archive and a particular elevation region are grouped into fifty evenly-spaced bins based upon the proportion of RF trees voting in favor of classifying that particular record as a flash flood. For example, when between 295 and 300 RF trees vote to classify a case as a flash flood, that case is included in the 0.98 – 1.00 "Forecast Probability" bin in Figure 13. Then, within in each of these bins, the fraction of cases verified by actual *Storm Data* reports is calculated and plotted. There are three sets of points in Figure 13: one for each of the three elevation regions.

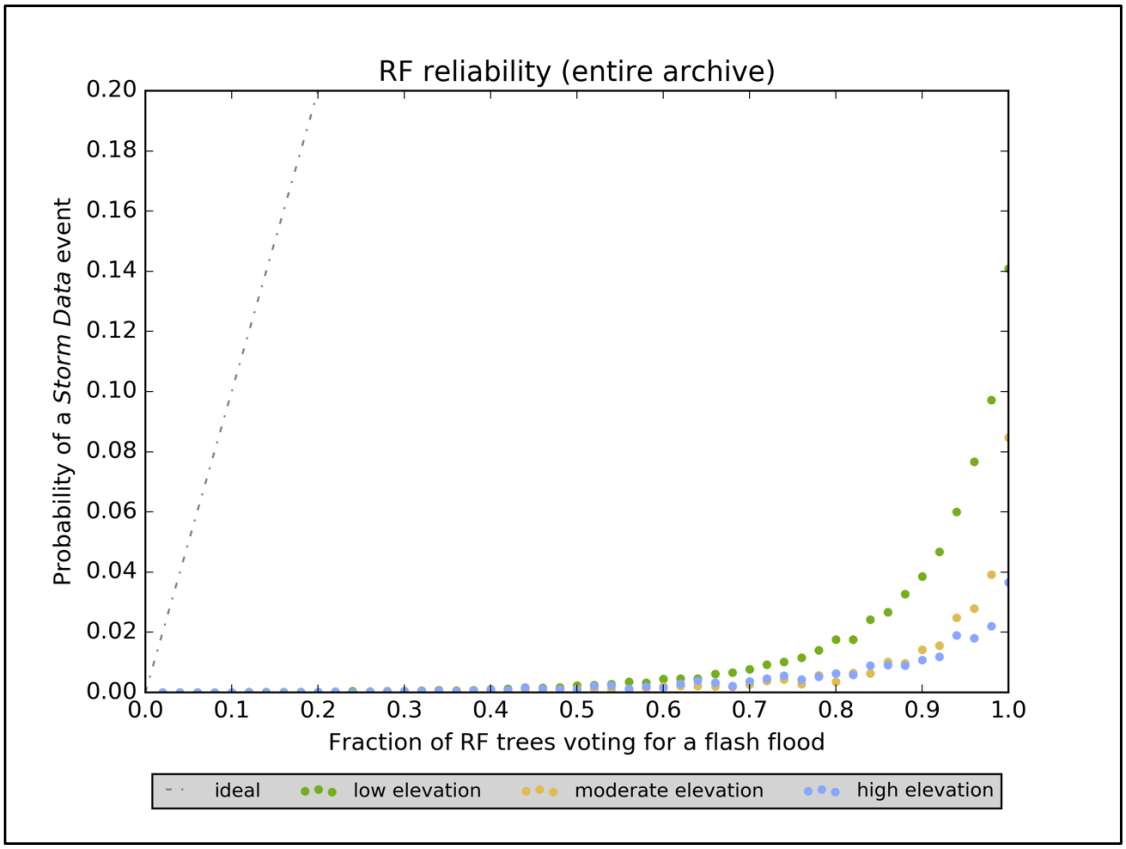

Figure 13.The historical probability of a flash flood (in *Storm Data*) as a function of the fraction of RF trees voting for the flash flood label, broken down by the elevation region from which the cases were drawn

Several properties of Figure 13 are of interest: when all trees in the RF vote for a *Storm Data* event among the low elevation cases (green dots in Figure 13), a flash flood is observed only about 14% of the time; this over-prediction is plausible for a relatively rare and transient event like a flash flood (Williams 2009). In other words, when all the GFS model parameters are together in such a way that an RF or other ML classifier believes a flash flood will occur, the verifying event is recorded in *Storm Data* only 14% of the time. This number drops to 4% for the moderate elevation cases (in orange) and rises to 9% for the high elevation cases (in blue). This difference can be explained due to the smaller number of flash floods in the moderate and high elevation test datasets, but it likely also represents the increased difficultly in accurately forecasting flash flood

environments in the High Plains and the West, where a greater proportion of reports of flash floods result from individual monsoon thunderstorms in the Southwest and isolated heavy rainfall from localized severe summertime convection in the High Plains. Both of these meteorological regimes are hard or impossible to resolve in the GFS3, which has only 1.0-by-1.0-degree resolution. Additionally, issues related to the underreporting of flash floods also help to explain the regional differences. Overall, it is likely no surprise that RFs, just like humans, find flash flood forecasting to be fundamentally more difficult in the western U.S. than it is in the eastern U.S.

As the proportion of trees voting to classify a set of predictors as a flash flood increases, the observed probability of a *Storm Data* report having been recorded increases monotonically. This suggests that there is useful confidence information contained in the percentage of trees voting for a particular label. This useful information can be extracted via a curve-fitting procedure; all three sets of points in Figure 13 are best represented by power law relationships.

Figure 14 contains three power law relationships; their associated mathematical representations are in the upper-center inset of the figure. In each of these equations, x is the fraction of RF trees voting to classify a particular case as a flash flood and y is the calibrated probability of that case being verified by a historical *Storm Data* report of flash flooding. Figure 15 is a zoomed-in reliability diagram that consists of three sets of points. These points are the calibrated RF probabilities resulting from the application of the power law relationships identified in Figure 14. The low maximum probabilities achieved by these calibration relationships are the result of using coarse-resolution NWP datasets

to forecast flash floods, which are often highly-localized phenomena; this acts to reduce the maximum confidence the RF classifier could ever hope to yield.

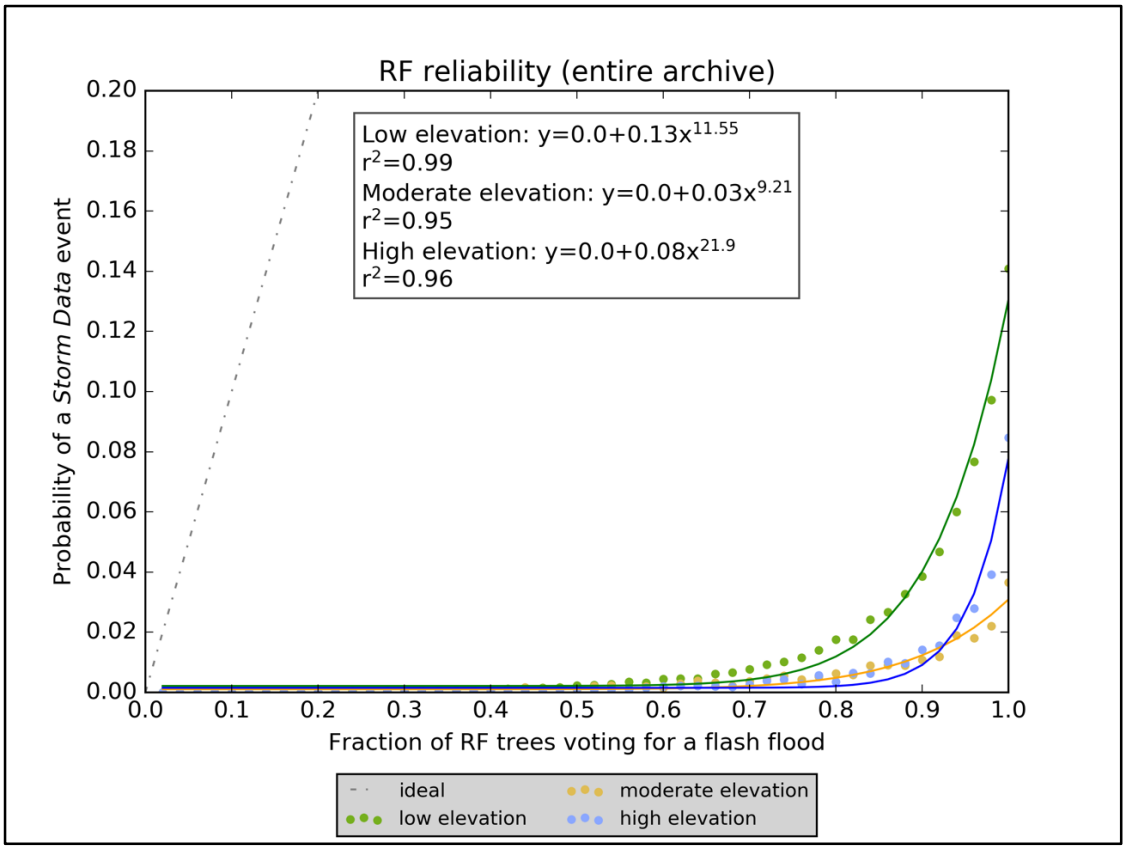

Figure 14. RF vote-to-probability transformation relationships for each of the three elevation regions

The calibrated probabilities in Figure 15 are significantly more reliable than those shown in Figure 13. However, the calibration process is replete with pitfalls. In particular, the calibrated RF probabilities are too high relative to the probability of *Storm Data* reports of flash floods occurring the independent test dataset. The fit of the moderate elevation curve in Figure 14 is less good than that of the cases drawn from the other two elevation regions. This is likely due to the smaller number of flash floods in the moderate elevation dataset ($N = 295$). Despite these issues, the fact that a power law represents the best fit regardless of elevation region suggests that this probability calibration concept

may be generalizable through time, a concept which holds important operational implications. If something close to the RF strategies outlined in this study were to be operationalized, one would likely fit an RF to the past year or so (or maybe most recent NWP epoch's worth) of available NWP output and then use that RF to generate real-time predictions going forward. For this strategy to succeed, one would need to demonstrate that the RF and its associated probability calibration continued to output reliable probabilities even in the face of NWP model code changes and upgrades. The datasets used in the present study provide two opportunities to test this hypothesis.

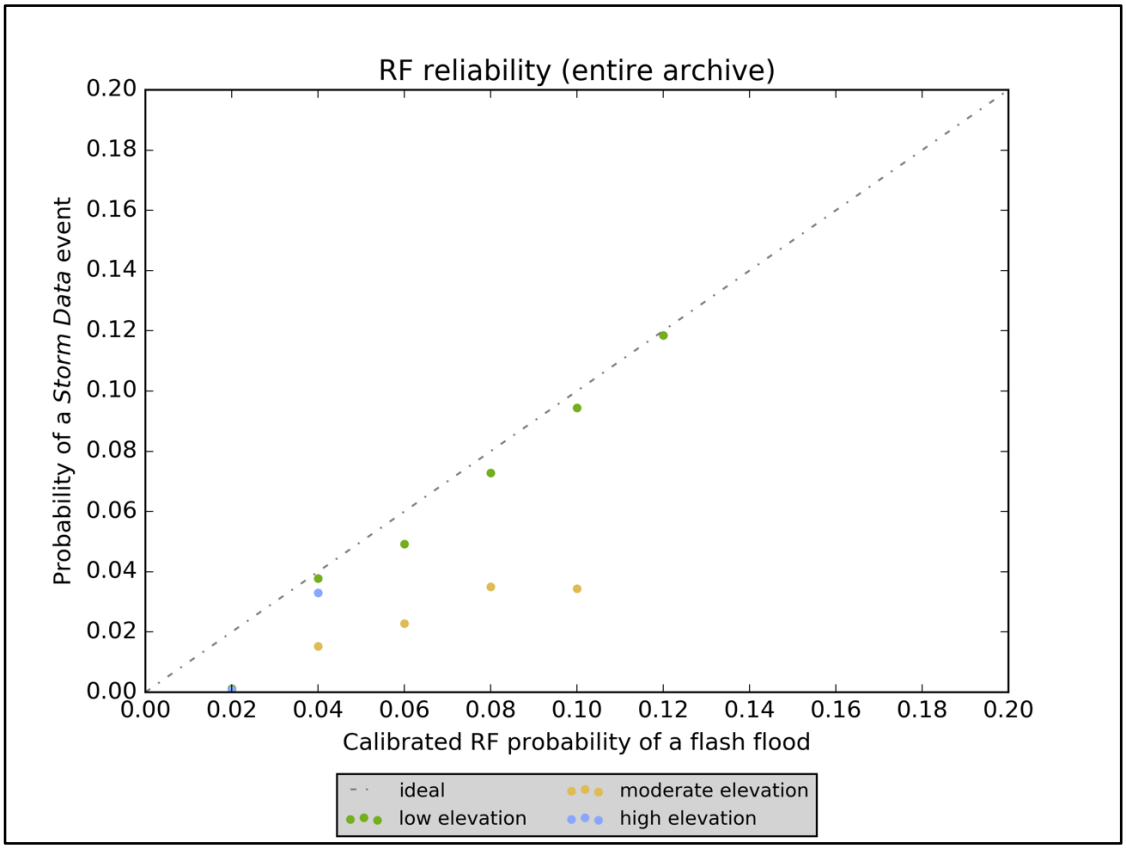

Figure 15. Zoomed-in view of RF probabilities for each elevation region after the application of power law calibration processes

Figures 16 and 17 show the results of these two opportunities. In Figure 16 the RF fitting and calibration process for each elevation takes place using the 2013 training

data and the final probabilities are generated on the 2014 testing data. In Figure 17 the datasets used are drawn from 2014 training and 2015 testing data, respectively. The results demonstrate that, although there is a degradation in performance relative to Figure 15, which arises as a result of epoch-to-epoch differences between the power law fits, and the small sample size available in the 2014 and 2015 testing datasets, the calibrated probabilities for the low elevation cases (the green dots) are still in the same neighborhood as the observed probabilities.

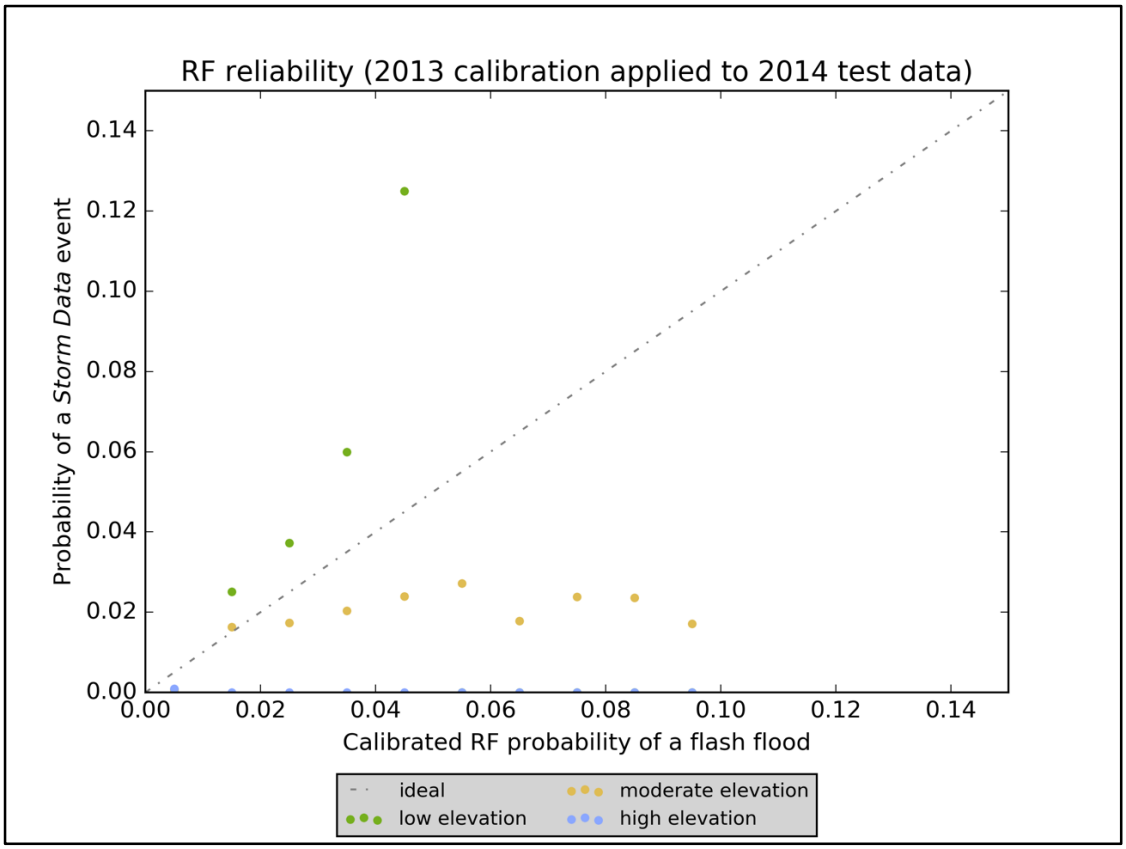

Figure 16. Application of a 2013 probability calibration relationship to 2014 testing data

This process results in overforecasting of moderate elevation events (the orange dots) and underforecasting of low elevation events. Unfortunately, there are so few elevation events in the 2014 or 2015 test data that the reliability of these (the blue dots)

cannot be accurately assessed using the calibration procedure proposed in this section. The poor performance on the moderate elevation fit in Figure 15 results in similarly poor performance on the 2014 test data. There are not enough moderate elevation flash floods in the 2015 test data to accurately plot those reliabilities.

Components of the Brier Score

The Brier score can be decomposed into constituent terms that better diagnose *how* a particular forecast system is useful and where it can be improved. Murphy (1973) proved that the Brier score can be decomposed into the reliability plus the uncertainty minus the resolution. Recall that lower Brier scores represent more skillful forecasts and are therefore more desirable.

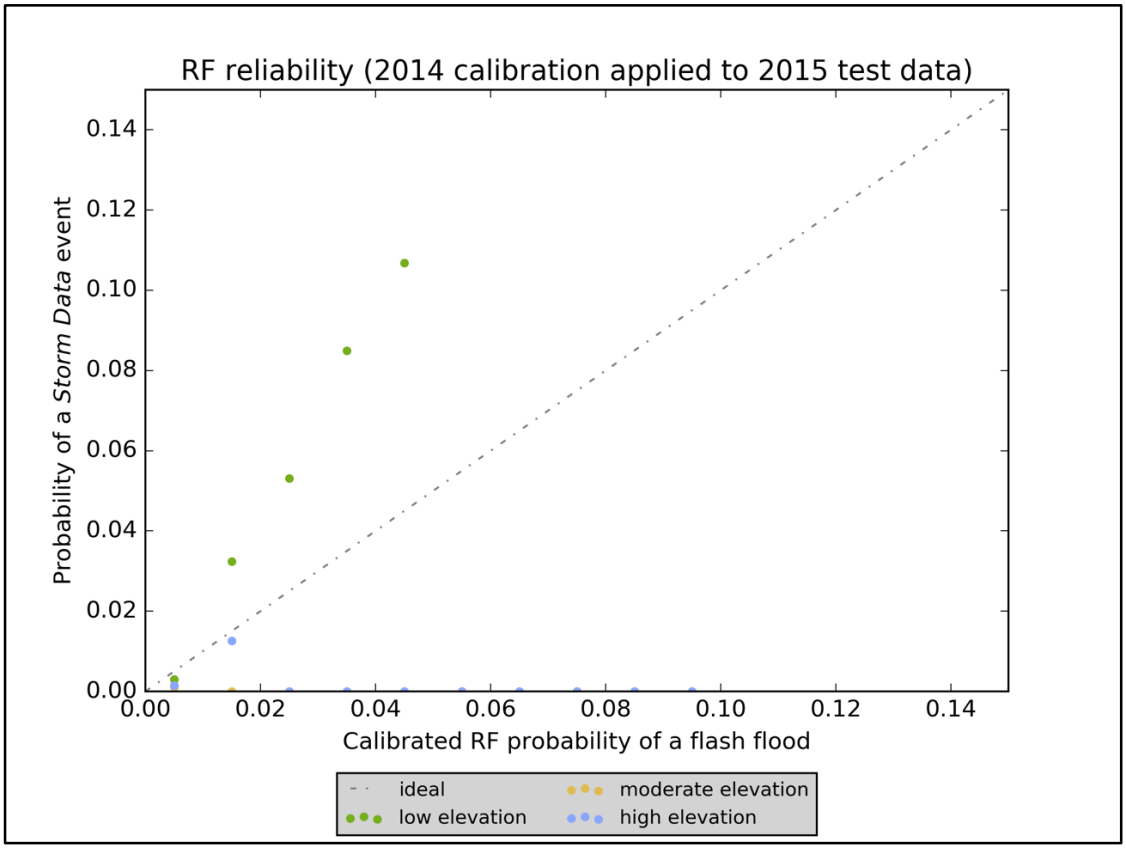

Figure 17. As in Figure 16, but with a 2014 probability calibration relationship applied to 2015 testing data

The first constituent of the Brier score is the uncertainty, which is related to how frequently a particular event actually occurs. Higher uncertainties occur when the classes one is attempting to predict are similar in climatological frequency to one another. So, for situations when the climatological probability of the class one is attempting to forecast is either very low or very high, the best possible Brier score is lower. For example, if an event never occurs, one could always issue 0% forecasts and achieve a perfect Brier score. Note, however, that this does *not* mean rare events are "easy" to forecast; just the opposite, in fact!

The second component of the Brier score is called reliability, and it is a measure of how closely the forecast probabilities track the true probabilities and can be visualized via reliability diagrams like those shown in Figures 13, 14, 15, and 16. Reliability is better when the differences between the forecast and observed probabilities are small.

Finally, the resolution measures the ability of a forecaster to separate cases (via his forecast probabilities) into classes for which there is an actual difference in the probability of the event being forecast. In other words, resolution asks: is a forecaster actually assigning probabilities that differ from the underlying climatological rate of the event? Unlike reliability or uncertainty, high resolution is desirable because it indicates the forecaster is picking up on real differences in the underlying observed probabilities as her forecast probabilities vary.

Let us begin with calculating the uncertainty. The uncertainty is the base rate (or the rate at which flash floods are observed) times the 1 minus the base rate. In this case, there are 15,257 flash floods in the low elevation region over the entire archive and there are 6,626,870 total cases for the same epoch-elevation combination. This results in a base

rate of 0.23%. In decimal form, the base rate is 0.0023, and the uncertainty is 0.0023. Assuming perfect reliability and resolution, therefore, the best possible Brier score one could expect to achieve when forecasting flash floods in the low elevation region is 0.0023. For both the moderate and high elevation cases, the uncertainty is 0.0008.

Resolution is calculated by binning the forecast probabilities and calculating the observed rates of the event in each bin ("Flash flood obs. rate" in Table 13). Then the differences between observed rates in each bin and the base rate (2.3×10^{-3}) for the whole dataset are taken (note that is slightly different from the result in the previous paragraph, as Table 13 is considered only with the training and validation datasets) and squared. The squared differences are weighted by the number of forecasts in each bin, and then divided by the total number of forecasts. The sum of these is the resolution. Using the data from Table 13, the resolution is 7.8×10^{-5}. For the moderate and high elevation cases, the resolution is 9.2×10^{-6} and 8.5×10^{-6}, respectively.

Table 13. Data required for resolution calculation for low elevation cases across the entire archive

Forecast probability	*# of forecasts,* N	*# of* Storm Data *reports*	*Flash flood obs. rate*	*(1/*N*) x (Bin obs. rate – Total obs. rate)^2*
0.000 – 0.099	2,309,518	56	2.4×10^{-5}	3.0×10^{-6}
0.100 – 0.199	430,646	78	1.8×10^{-4}	4.8×10^{-7}
0.200 – 0.299	302,693	115	3.8×10^{-4}	2.8×10^{-7}
0.300 – 0.399	224,132	167	7.5×10^{-4}	1.4×10^{-7}
0.400 – 0.499	192,893	309	1.6×10^{-3}	2.4×10^{-8}
0.500 – 0.599	162,927	530	3.3×10^{-3}	3.7×10^{-8}
0.600 – 0.699	134,894	767	5.7×10^{-3}	3.8×10^{-7}
0.700 – 0.799	110,607	1,319	1.2×10^{-2}	2.6×10^{-6}
0.800 – 0.899	87,494	2,334	2.7×10^{-2}	1.3×10^{-5}
0.900 – 1.000	50,820	3,549	7.0×10^{-2}	5.8×10^{-5}
Totals:	**4,006,624**	**9,234**	**2.3×10^{-3}**	**Resolution: 7.8×10^{-5}**

The reliability is determined by binning the forecast probabilities, as in Table 13. For each bin, the average forecast probability is compared to the rate of forecasts in that

bin that verified. The squared differences between each forecast rate and verification rate are calculated and then weighted by the number of forecasts in each bin. The sum of these is the reliability; lower numbers arise as a result of small squared differences and thus, more reliable probabilities. For the entire archive and the low elevation area of CONUS, bin the forecast probabilities into ten groups, where forecasts of greater than 0% but less than 10% are in one bin, those greater than or equal to 10% but less than 20% are in another, and so on. Table 14 contains the parameters necessary to complete the resolution calculation for the low elevation cases drawn from the entire archive, using the same data shown in Figure 13.

Table 14. Data required for reliability calculation for low elevation cases across the entire archive

Forecast probability	*# of forecasts,* N	*Flash flood obs. rate*	*Avg. forecast prob.*	*(1/*N*) x (Avg. forecast – Total obs. rate)^2*
0.000 – 0.099	2,309,518	$2.4\text{x}10^{-5}$	0.021	$2.0\text{x}10^{-4}$
0.100 – 0.199	430,646	$1.8\text{x}10^{-4}$	0.14	$2.0\text{x}10^{-3}$
0.200 – 0.299	302,693	$3.8\text{x}10^{-4}$	0.25	$4.6\text{x}10^{-3}$
0.300 – 0.399	224,132	$7.5\text{x}10^{-4}$	0.35	$6.8\text{x}10^{-3}$
0.400 – 0.499	192,893	$1.6\text{x}10^{-3}$	0.45	$9.6\text{x}10^{-3}$
0.500 – 0.599	162,927	$3.3\text{x}10^{-3}$	0.55	$1.2\text{x}10^{-2}$
0.600 – 0.699	134,894	$5.7\text{x}10^{-3}$	0.65	$1.4\text{x}10^{-2}$
0.700 – 0.799	110,607	$1.2\text{x}10^{-2}$	0.74	$1.5\text{x}10^{-2}$
0.800 – 0.899	87,494	$2.7\text{x}10^{-2}$	0.85	$1.6\text{x}10^{-2}$
0.900 – 1.000	50,820	$7.0\text{x}10^{-2}$	0.94	$1.1\text{x}10^{-2}$
Totals:	**4,006,624**	**$2.3\text{x}10^{-3}$**	**0.18**	**Reliability: $9.1\text{x}10^{-2}$**

The reliability of the low elevation forecasts for the entire archive is 9.1 x 10^{-2}. The equivalent numbers for the moderate and high elevation cases are 9.3 x 10^{-2} and 7.6 x 10^{-2}, respectively. Because flash floods are so rare, the uncertainty is very low and achieving a good Brier score is not difficult; the Brier scores reported throughout this chapter are dominated by the contribution of the reliability term of the score. (The resolution, from Table 13, is essentially nil.) However, the results of the decomposition

process and the reliability diagrams indicate that the RF is not, for example, achieving a good Brier score by always forecasting "no flood" (indeed, the average low elevation forecast probability is 18%). The number of *Storm Data* reports in Table 13 increases with the increased forecast probability of a flash flood. Unfortunately, as is often the case when dealing with extremely rare events, there is no one single metric that appropriately assesses the skill of these forecasts.

Receiver Operator Characteristic Curves and Comparisons to Other Methods

The ROC diagram can be used in concert with reliability diagrams and the Brier score decomposition to assess the prediction skill of a classifier. Figure 18 contains ROC curves (generated upon the study period test set) for each of the three elevation regions. The optimal threshold for a binary classifier corresponds to the point on the ROC curve closest to the upper-left corner of the ROC diagram. The contingency table resulting from that threshold has a PSS equal to the length of a vertical line drawn from the 1:1 line on the ROC diagram to the ROC curve.

Because there is currently no widely-accepted NWP-based flash flood *forecasting* system, comparing the proposed RF to some other baseline is not trivial. However, the results from the RF can be compared to individual NWP model fields used in heavy rainfall or flash flood forecasting, including convective precipitation rate, precipitation rate, ground-to-0.1-m-BGL soil moisture, PW, model PW anomaly, and K index. The resultant ROC curves for these model fields are plotted in Figure 19. The AUC for the RF method is 0.95 or 0.96, depending on the elevation region being considered; the six individual GFS model fields considered in Figure 19 have AUCs ranging from 0.59 for the K index to 0.90 for the shallowest soil moisture field. Based on a comparison of the AUCs in Figures 18 and 19, it is clear that the RF method, by combining a series of model

variables together, adds value compared to the individual model fields frequently used in operational flash flood monitoring or forecasting.

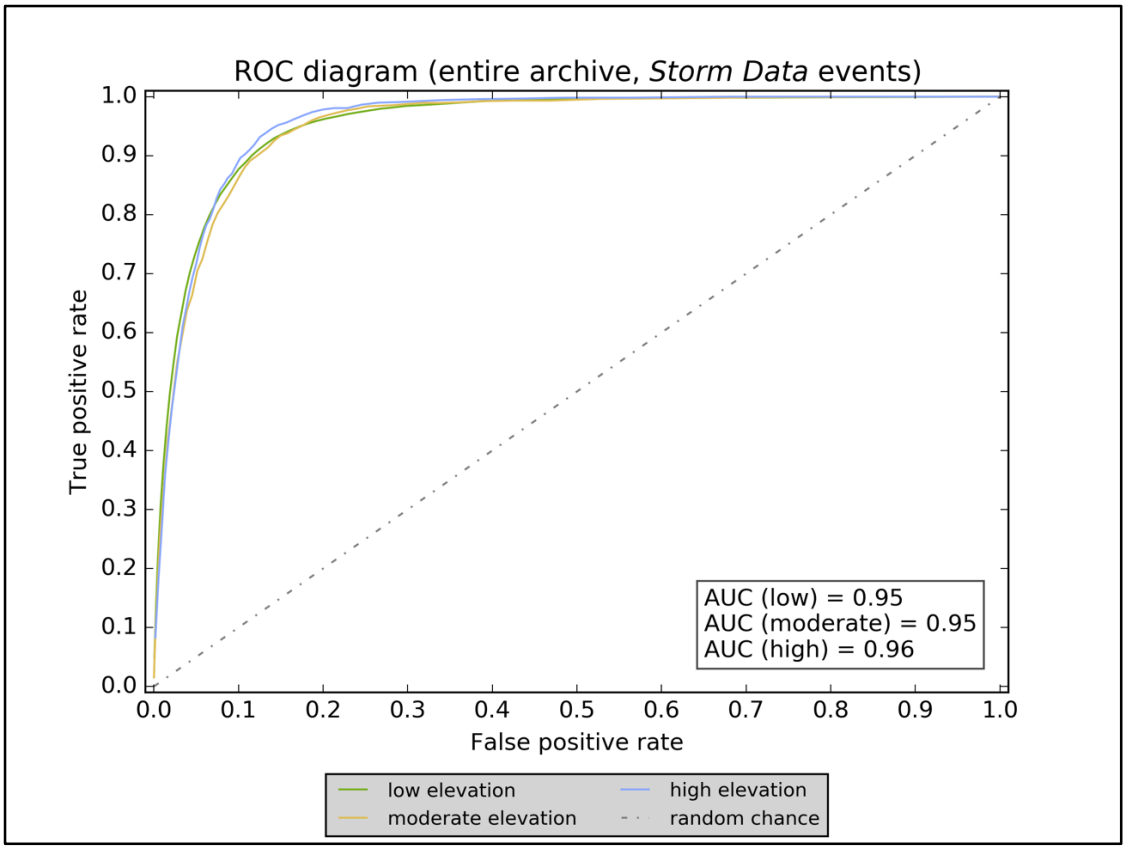

Figure 18. ROC diagram based upon data from the entire study period with curves for each of the three elevation regions

Concluding Thoughts

The results presented in this chapter demonstrate that ML holds promise for forecasting rare events like flash floods from NWP, with some major caveats. The RF yields probabilistic predictions that can be calibrated to make them more reliable, according to the Brier score, ROC diagrams, and reliability diagrams. However, this calibration process is specific to the predictions arising from a particular RF model and therefore may not be able to be extending to other regions or other time periods. Given the extreme rarity of *Storm Data* flash flood reports and the lack of operational forecast

tools, though, all the competitive methods tested in this chapter, whether statistically- or physically-based, perform about the same as or worse than the RF method.

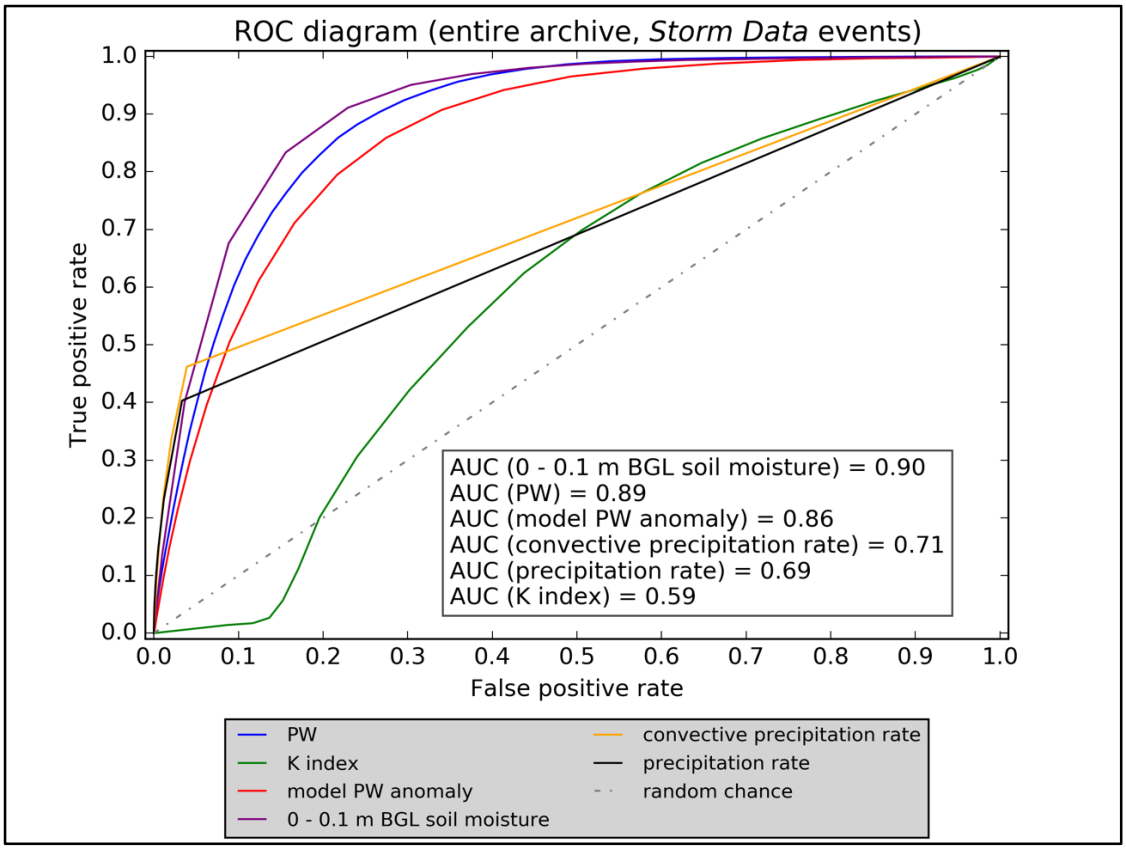

Figure 19. ROC diagram comparing the skill of various thresholds applied to a series of GFS model fields at forecasting flash floods

These results do not, on their own, provide much insight into how human forecasters could use the proposed system as a time-saving device or into the NWP model parameters that are most strongly associated with flash flooding. But from Chapter 3 we have learned the following: despite their rarity, flash floods can be probabilistically forecast in an efficient manner by using the output from a coarse-resolution NWP model in an RF. These predictions of flash flooding are more skillful than alternative methods currently in use, including model QPF, model PW anomaly, and others.

Chapter 4: Variable Selection and Physical Interpretations

In general, explaining the results of an ML algorithm or other statistical model from a physical perspective is desirable. One of the great advantages of many machine learning (ML) algorithms, and the random forest (RF) method in particular, is their ability to provide physical insights into the system upon which the ML algorithm is being applied. In combination with *external* expert knowledge and recognition of the physical laws governing a system, the insights derived from an ML algorithm are a powerful tool for better understanding and thus, just possibly, improved predictions, as well.

In the case of RFs, the process by which a forest is grown and then used to generate predictions results in a series of internal metrics that can used to explain the RF outputs. One of these internal metrics is the mean decrease in Gini impurity (often abbreviated to MDG); each candidate predictor in the RF has an MDG. Those variables with high MDGs are more skillful splitters and thus are more important when generating final classifications. The Gini impurity, G, is given by (12) (Tan et al. 2005).

$$G = 1 - \sum_{i=1}^{n_c} p_i^2 \tag{12}$$

In (12), n_c is the number of classes being predicted and p_i is the proportion of the total cases falling into a particular class at that node. For the binary prediction problem, (13) simplifies to (13).

$$G = 1 - p_{yes}^2 - p_{no}^2 \tag{13}$$

From (13), it is readily apparent that, when the samples are split evenly between "yes" and "no", $G = 0.5$ because p_{yes} is 0.5 (and of course, p_{no} is also 0.5). If $p_{no} = 0$ or 1, then $G = 0$. The RF algorithm minimizes G at each node of each tree because it searches for the predictor variable that results in the purest split between classes of the cases present

at the node. To calculate G for a particular node and its two child branches is a simple undertaking: $G = G_{node} - G_{child,1} - G_{child,2}$. The MDG for a candidate variable, considered across the entire RF, is calculated by averaging MDG over all the splits where the variable is used and then weighting this value by the number of splits in which that variable is used. Other importance metrics include the entropy and the misclassification error. Both of these have similar properties to the MDG, in that they are all minimized when $p_{yes} = 0$ or 1 and maximized when p_{yes} is 0.5 (Tan et al. 2005).

The results of an ML classifier can also be explained via more indirect methods *external* to the classifier. This involves using expert opinion, physical understanding, empirical studies, and other methods to choose which variables are initially provided to the classifier and in what format or context the variables are provided. In meteorology, a common example involves the use of derived variables known or suspected to be important to a particular forecasting or classification task. Derived variables, especially those defined by non-linear combinations of other quantities, cannot be fully accounted for using the RF or other ML methods. However, methods like RF are capable of identifying *some* non-linear interactions between predictor variables. Another example is the use of normalized values of predictor variables, as opposed to raw values of the predictor. This strategy is critically important for some ML classifiers, but is not strictly required with RFs. In any event, success in variable selection requires a balance between physical or "expert" understanding of the system to which the classifier is being applied and statistical metrics internal to the classifier.

External Evidence

Expert Variable Selection

Chapter 2 contains a summary and synthesis of several studies from the meteorological literature examining flash floods. Some studies (e.g., Doswell et al. 1996) consist of a combination of basic principles and case study evidence while others are data-driven statistical examinations of collections of flash floods (e.g., Jessup and DeGaetano 2008 2008 or Schreoder et al. 2016a). Studies from both categories agree on the basic meteorological and hydrologic requirements for the development of a flash flood, but there is disagreement among other authors on how to define a flash flood and what considerations and rules-of-thumb are most important in the flash flood forecast process. One attractive feature of the RF method – and other decision tree algorithms – is their ability to deal with collinear, uninformative, or random variables without significant degradation in the quality of the final predictions. This encourages an ML developer to give an RF extra information on the off chance that the classifier figures out how to derive some value from it. On the other hand, there are situations in which the bagging process of an RF can result in a generally uninformative predictor being "accidentally" selected for use at a particular tree node; this can have a minor impact upon the classifier's ability (Tan et al. 2005). Put another way, ensembles of decision trees are quite robust but not completely immune to the effects of overfitting.

Correlations Between Variables

Uncovering relationships between predictor variables can be useful in the context of dimensional reduction. Although RFs are not typically strongly affected by meaningless or uninformative variables (i.e., those are that are so strongly correlated with another predictor that they provide little additional information), the importance metrics

from RFs *are* affected by these correlations. For example, if precipitable water (PW) and 700-hPa specific humidity, *700q*, are correlated with one another, their correlation acts to reduce the individual MDG of both variables. Because they are correlated, they are each similarly likely to be selected as the best split at new tree nodes. They are also likely to split cases in comparable ways. This reduces the effective splitting power that either of these two variables would have had in isolation. If the correlation between the two variables is perfect, in situations where either would result in the best split, the bagging process ensures that, over many cases, each of the two identical variables would be selected 50% of the time. Even in the rare case of a perfectly-correlated pair of variables, their individual MDGs cannot be added together to yield an "effective MDG". Although there are more efficient ways to reduce the dimensions of a predictor matrix and, in turn, attempt to build a more parsimonious model, collecting and identifying correlations between predictor variables and then comparing these to our physical understanding of the atmosphere can provide a "sanity check" on the construction of the predictor matrix.

In the present study, cross-correlations between predictor variables occur frequently, because all but one of the predictor variables (flashiness) are outputs from the same physically-based numerical weather prediction (NWP) model. A common way of checking correlations between variables, especially when the variables are typically distributed over very different numerical ranges is the Spearman rank correlation (Spearman 1904). The Spearman rank correlation is the Pearson correlation coefficient between the variables when ranked (Pearson 1895). The Pearson coefficient, r, and the Spearman rank coefficient, ρ, range from negative one to one, where a value of negative one represents a perfect negative correlation between two variables (or their ranks), a

value of zero indicates no correlation between the variables (or their ranks), and a value of positive one indicates perfect positive correlation between the variables (or their ranks). Although the Pearson coefficient represents the degree of linear correlation between the variables, the Spearman coefficient represents how well any monotonic function can account for the relationship between the variables.

For the low elevation cases, there are 234 variable pairs (out of 20,592 total possible variable pairs) with ρ greater than 0.8 or less than -0.8, and 227 variable pairs have r values that meet the same criteria. Most of these occur when the same *type* of Global Forecast System (GFS) field is compared between two different levels (e.g., 150- and 200-hPa geopotential height, *150hgt* and *200hgt*, are highly correlated with one another [ρ = 0.99 and r = 0.99]). Other common correlations occur amongst those variable pairs related to one another via well-known meteorological relationships (e.g., temperature and geopotential height are frequently highly-correlated, PW and q, or various wind components and the wind speed or speed shear). The geopotential height at 150 hPa has Spearman correlations >= 0.8 with 17 other candidate predictors. Let us remove each of these 17 other predictors and test the impact of that change upon the skill of the RF and upon the MDG value of *150hgt*. Table 15 summarizes the results of this test.

With all original predictor variables included, the average Brier score of the predictions from the forest was 0.094. In this case, removing correlated predictors does *not* improve the quality of the final predictions. In fact, individual removal of one-half of a particular correlated pair resulted in the Brier score of the forest's predictions remaining the same or declining in skill by up to 3%. Of course, there were still many correlated

predictor pairs present in the predictor matrix even when all of the 17 of these pairs were accounted for, so it is possible that some *other* correlated pair might be harming the quality of the predictions, but in this case, the likely answer, based upon the known statistical properties of the RF algorithm, is that highly-correlated predictor pairs and the presence of excess or uninformative variables do not result in worse predictions. Indeed, this behavior has been observed over and over again in many studies where the RF method was applied to predictor matrices with a high degree of collinearity.

Table 15. Results of a test for the effect of Spearman correlation coefficient upon MDG values of 150-hPa and the Brier score of flash flood predictions from the RF

Variable removed	ρ *relative to* 150hgt	*Change in* 150hgt *MDG*	*Relative change in* 150hgt *MDG rank*	*Resulting Brier score*
200hgt	0.99	11 %	+5	0.097
250hgt	0.98	20 %	+10	0.094
300hgt	0.98	11 %	+4	0.096
400hgt	0.97	9.0%	-1	0.096
500hgt	0.95	21 %	+8	0.095
700hgt	0.88	4.0%	-2	0.095
250temp	0.81	13 %	+3	0.096
300temp	0.94	17 %	+7	0.095
400temp	0.96	8.7%	+3	0.096
500temp	0.96	28 %	+18	0.096
700temp	0.93	14 %	+6	0.096
850temp	0.89	23 %	+10	0.096
925temp	0.88	7.7%	-4	0.096
2m_temp	0.85	15 %	+6	0.096
sfctemp	0.81	12 %	+5	0.096
2m_q	0.82	1.7%	-5	0.097
1013.25q	0.82	13 %	+2	0.096
(remove all)	N/A	160 %	+31	0.097

However, the evidence is quite strong that correlated predictors do change the importance metrics arising from the RF. On average, as each correlated predictor in Table 15 was removed, the new MDG score of *150hgt* was 13% higher than before the variable removal process. The new MDG rank of *150hgt* after the variable removal process was, on average, four spots higher in the list of candidate predictor variables, accounting for

the fact that, in each case, after variable removal there was one less candidate predictor to rank. When all 17 predictors highly-correlated with *150hgt* were removed, the resultant MDG score of *150hgt* was 160% higher than before, and *150hgt* was ranked 31 spots higher than before in the list of important variables, even accounting for the fact that there were 17 fewer predictors to rank after the removals were completed.

Historical Distribution of Global Forecast System Model Fields

One way of determining variable importance, and thus, discrimination power, is via visualization of the distribution of the predictor variables across a large number of cases, particularly by plotting these predictor variables with the flash flood and non-flash flood cases shown separately from one another. However, due to the extremely low historical prevalence of *Storm Data* reports of flash floods, the respective flash flood and non-flash flood portions of the distributions must be normalized such that areas of the bins of each histogram sum to one. The result of this procedure applied to the GFS PW analyses is shown as Figure 20, while Figure 21 is the result of the same procedure for the PW anomalies.

Many past studies have found that flash floods are associated with high PW values, but these studies were based on observations, not upon NWP model analyses. The distribution of all PWs in this study roughly follows a log-normal distribution, with a large degree of skewing to the right (higher PW values), as shown in the blue bars of Figure 20. On the other hand, the distribution of the standardized anomalies of the GFS-analyzed PW values (relative to historical GFS-analyzed PW values at each grid and for each month) is much closer to normal, as shown in the blue bars of Figure 21. In Figure 21, it is quickly apparent that, on a normalized basis, higher values of GFS-analyzed PW

(or more positively-anomalous values of GFS-analyzed PW) are indeed associated with an increased probability of there being a collocated *Storm Data* report of a flash flood.

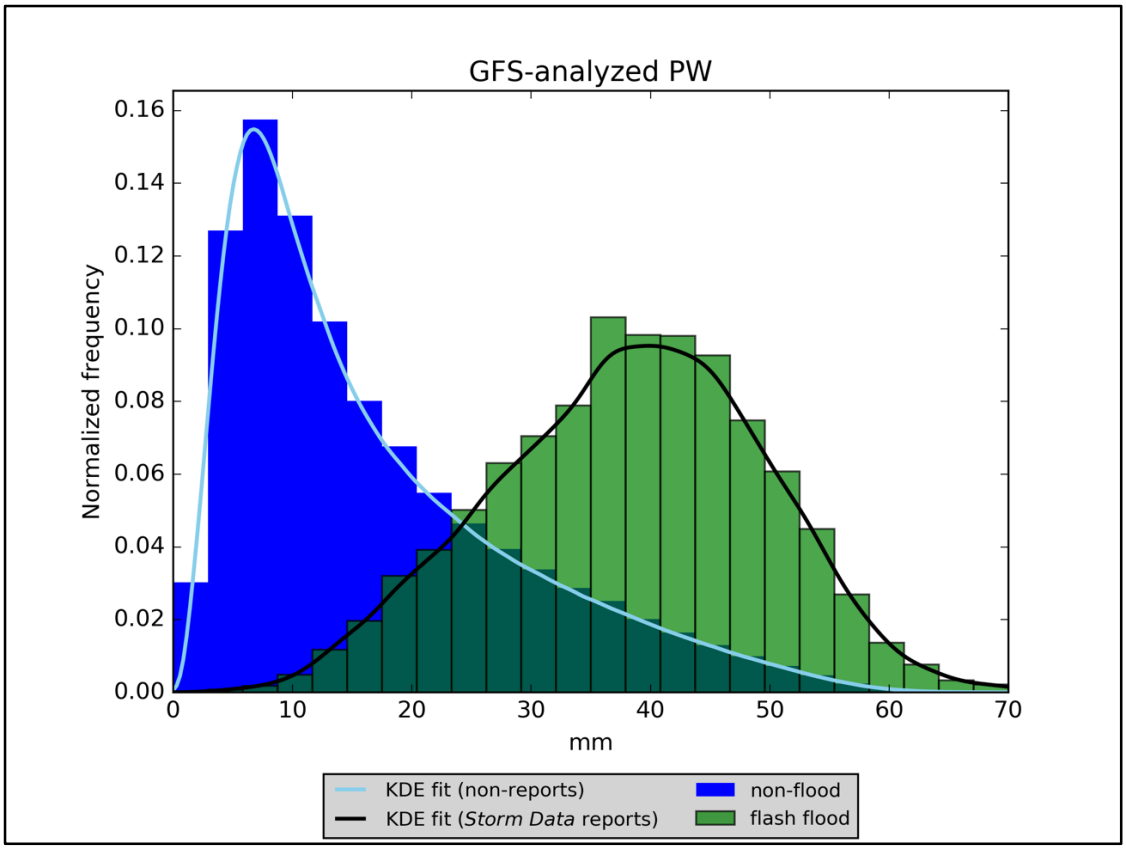

Figure 20. Normalized histogram and best KDE fit of the GFS-analyzed PW of all cases from the entire archive, comparing flash floods to non-flash flood cases

These normalized histograms have been created for all of the 146 predictor variables and all three of elevation regions in the present study. Although histograms are helpful in comparing distributions of variables, they are dependent upon the width of the bins used to generate them. They are also not automatically able to be represented by simple functional relationships, but a procedure called kernel density estimation (KDE) can be used to create a non-parametric estimate of the probability density function (PDF) of the variable in a histogram (Scott 1992). The KDE process smooths out a histogram by calculating the value of this PDF at each point from the contributions from the

neighborhood of values around each point. The size of this neighborhood is called the bandwidth, and it has a strong impact upon how smooth the KDE function is and upon the quality of the KDE fit. Scott (1992) proposed a rule, given by $n^{(-1/(d+4))}$, for setting the bandwidth used in the KDE process, where *n* is the number of data points and *d* is the number of dimensions. This rule is implemented in the SciPy Python library (Jones et al. 2001), which is used to generate all the KDE fits to the histograms shown in this chapter. It is critical to remember that Figures 20 and 21 are *normalized* histograms; this normalization is just a way to visually compare the differences between the flash flood and non-flood distributions of PW and other GFS model variables.

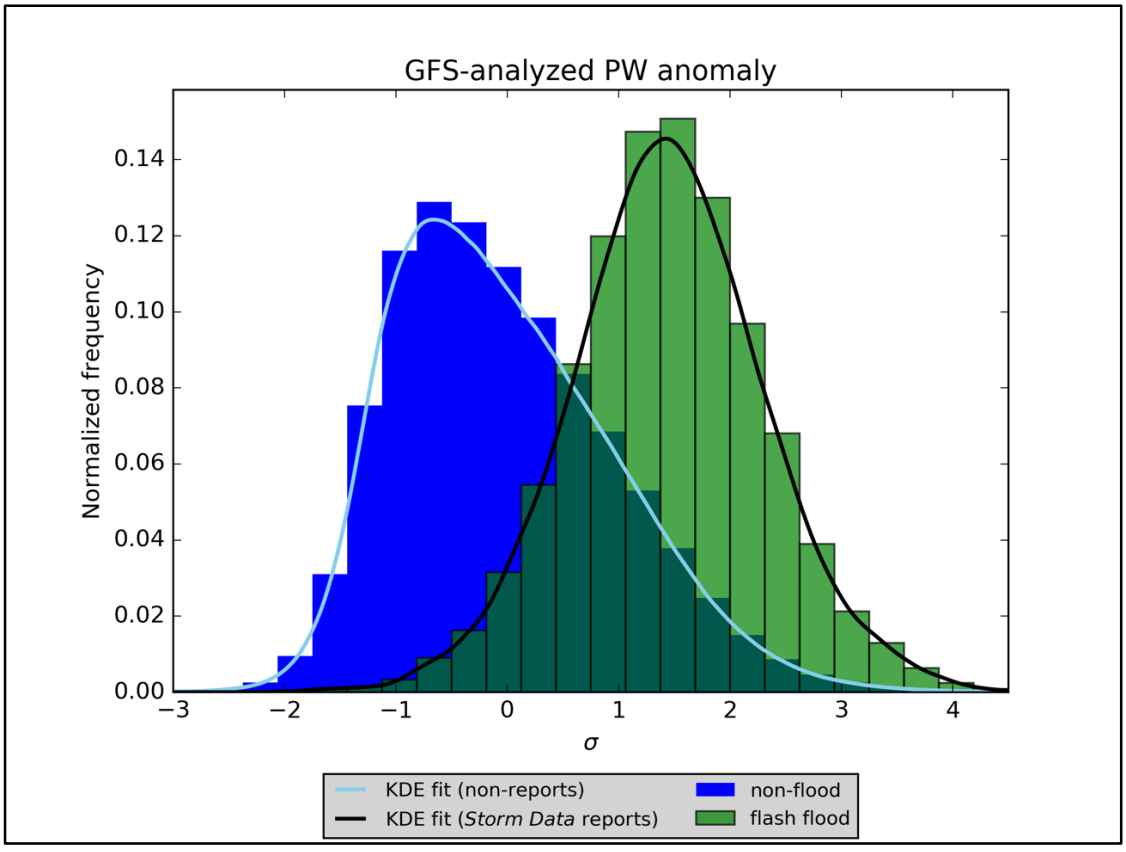

Figure 21. Normalized histogram and best KDE fit of the GFS-analyzed PW anomaly of all cases from the entire archive, comparing flash floods to non-flash flood cases

The normalized histogram for a particular variable can provide qualitative evidence of the importance of a variable in predicting flash floods. This qualitative evidence can be combined with information provided directly from the RF, like the MDG values associated with a particular variable. When two completely independent methods agree that a certain predictor is important or unimportant in generating skillful predictions, the evidence is strong that the RF is using the right predictors in an appropriate way and that the MDG is providing physically-grounded interpretations of the RF model output. Indeed, examination of the histograms of variables that appear frequently atop the MDG league tables bears out the fact that these are the histograms with the largest "splits" between their flash flood and non-flood KDEs. Another example is Figure 22, a normalized histogram comparing the model K index for flash floods and for non-floods.

Although these normalized histograms provide some physical insight into the flash flood forecasting problem, the extreme rarity of flash floods precludes the introduction of simple thresholds applied to a specific GFS model field (i.e., PW) or field derived from GFS model fields (i.e., K index), as shown in the analysis at the end of Chapter 3 (Figure 19). Even when one considers K index, the "best" predictor according to the split observed between the KDE fits to its normalized histogram, no one threshold applied to *K* can skillfully divide non-floods from flash floods. For example, implement a simple algorithm under which K index > 35 results in a flash flood and K index <= 35 does not. The resulting contingency table contains 9,074 hits, 10,002 misses, 562,264 false alarms, and 11,550,752 correct negatives. From this contingency table, the probability of detection (POD) is 47.6%, the false alarm rate (FAR) is 98.4%, and the

Peirce skill score (PSS) is 0.429. For comparison, a contingency table based upon RF forecasts (where the threshold for labeling a case a "flash flood" is reached when 50% or more of the trees in the forest vote for that label) produced on low elevation cases from the entire archive yielded a contingency table in which there were 8,471 hits, 642 misses, 550,346 false alarms, and 3,447,260 correct negatives. From this, the POD is 92.9%, the FAR is 98.5%, and the PSS is 0.792, a huge improvement in PSS. Over 50 trials, the PSS ranged from 0.791 to 0.797, demonstrating that this particular trial was no outlier.

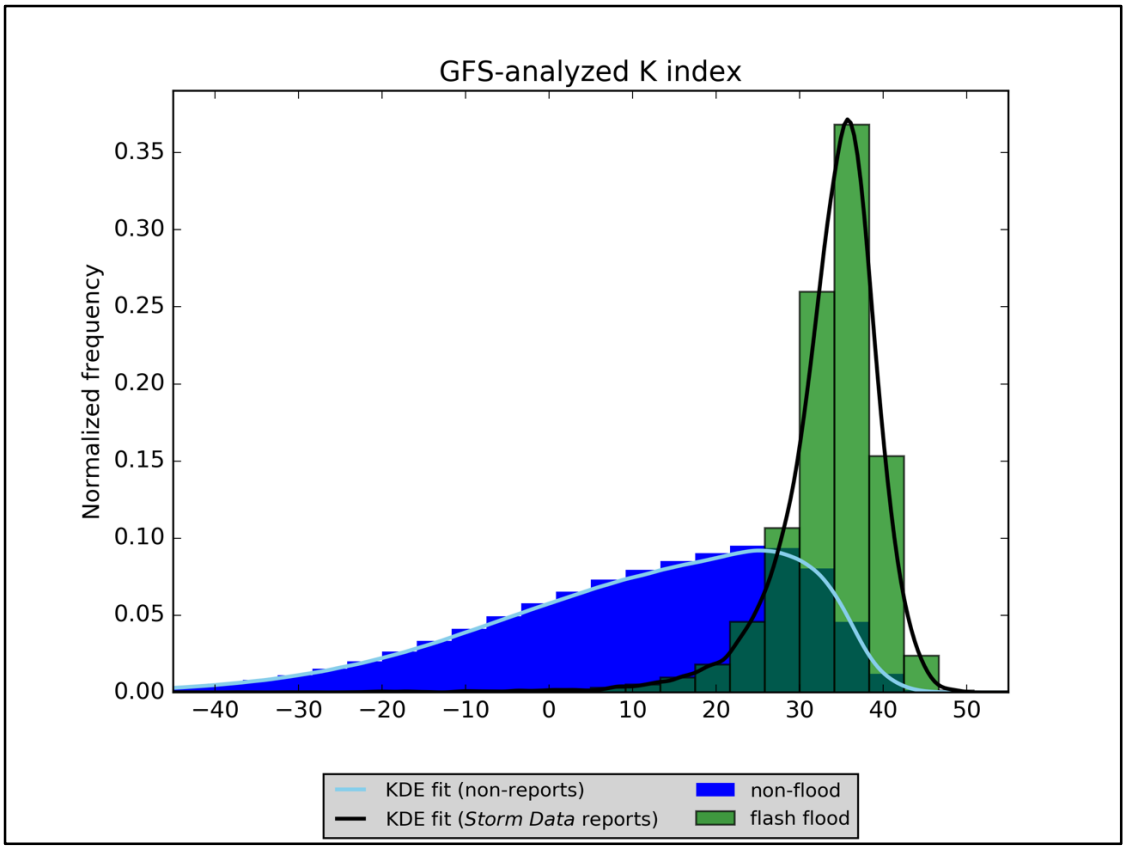

Figure 22. Normalized histogram and best KDE fit of the GFS-analyzed K index of all cases from the entire archive, comparing flash floods to non-flood cases

However, not all predictors are as useful as splitters between flash floods and non-floods. For example, the GFS 700-hPa v-component of winds, shown in Figure 23,

is relatively unable to distinguish between those grid cells and times where a *Storm Data* report was recorded and those where one was not.

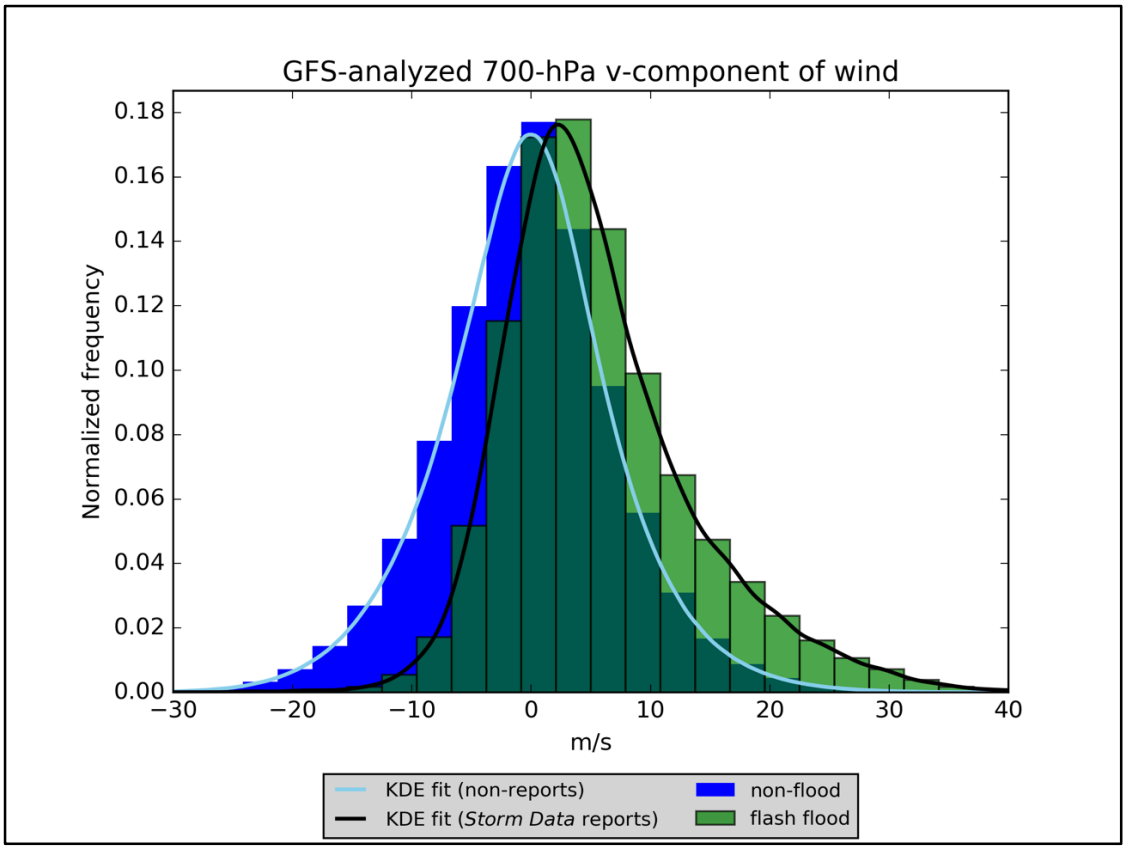

Figure 23. Normalized histogram and best KDE fit of the GFS-analyzed 700-hPa v-component of wind of all cases from the entire archive, comparing flash floods to non-flood cases

Machine Learning Evidence

Mean Decrease in Gini Impurity

One of the most important pieces of evidence for physical interpretation of the results of an RF comes from rankings of internal variable importance, namely, the MDG. For each elevation region, cases not occurring on the test days (i.e. the 5th, 10th, 15th, 20th, 25th, and 30th days of each month) were set aside as training and validation data. The training data are drawn from these cases as follows: 25% of the cases are randomly-sampled and held out as validation data, and of the remaining 75%, all flash floods are

retained along with a randomly-sampled equal number of non-flood cases. Then, following this procedure, 50 trials are completed in which an RF is fit to the training cases and the MDG of each predictor variable is recorded.

Table 16 contains the results of these trials broken apart by the elevation region of the cases used to grow the RF. Some of the candidate predictor variables with strong splits between flash floods and non-floods in their normalized histograms appear in Table 16 as having relatively high mean values of MDG. In particular, the K index is the most important predictor, on average, in the low and moderate elevation regions (this quantity is not defined for the high elevation cases, as it relies upon model fields taken from the 850-hPa level, which is effectively below ground in the high elevation region). The PW is the second-most-important variable for the low and high elevation cases, and ranks third for the moderate elevation cases. The best 4-layer lifted index (LI) appears in the top ten for all three elevations, as does the specific humidity, *q*, on multiple isobaric or above ground level (AGL) levels. The integrated relative humidity, *rh*, appears in the low and high elevation cases. It barely missed the top ten for moderate elevation cases, ranking between 8th- and 19th-most-important there.

However, there are differences between the elevation regions. Note that the precipitation (*preciprate)* and convective precipitation (*cpreciprate*) rates contribute heavily to the RF predictions for low elevations, but are ranked much lower in the moderate (*preciprate* is between 10th- and 26th-most-important, except for one trial in which it ranked 38th, and *cpreciprate* between 9th- and 22nd-most-important, except for two trials in which it ranked 38th and 28th) cases. In the high elevation cases, *preciprate* ranked between 16th- and 52nd-most-important, and *cpreciprate* ranked between 10th-

and 39th-most-important among all predictors. This difference suggests that the synoptically-forced precipitation typically resolved by the GFS3 is either less predictable in the western U.S. or that the precipitation leading to flash floods in the western U.S. is just more isolated and sporadic and thus less resolvable by the GFS3, not less predictable.

Table 16. Predictor variables with the greatest mean MDG scores across a series of RF trials

Mean MDG rank	*Variable name*	*Range of MDG ranks*	*Std. dev. of ranks*	*Mean MDG score*	*Std. dev. of mean*
		Low elevation			
1	k	1 to 4	0.7	0.066	0.007
2	pw	1 to 8	1	0.053	0.008
3	preciprate	1 to 9	2	0.046	0.007
5	cpreciprate	2 to 8	2	0.043	0.006
5	700q	1 to 10	2	0.043	0.006
6	850q	2 to 10	2	0.038	0.006
7	rh	2 to 10	2	0.035	0.006
8	pw_anom	3 to 11	2	0.034	0.006
9	4layer_li	4 to 11	2	0.033	0.005
10	925q	7 to 15	2	0.025	0.003
		Moderate elevation			
1	k	1 to 3	0.8	0.078	0.01
2	4layer_li	1 to 7	1	0.061	0.01
3	pw	1 to 8	2	0.059	0.01
4	700q	1 to 7	2	0.057	0.01
5	850q	1 to 9	2	0.050	0.01
6	sfccape	4 to 10	1	0.036	0.007
7	pw_anom	4 to 14	2	0.032	0.007
8	sfccin	6 to 19	2	0.029	0.007
9	sfctemp	7 to 18	2	0.024	0.005
10	2m_temp	6 to 17	2	0.022	0.006
		High elevation			
1	700q	1 to 4	0.8	0.088	0.01
2	pw	1 to 5	0.9	0.075	0.01
3	2m_q	1 to 5	0.9	0.073	0.01
4	4layer_li	1 to 6	0.9	0.059	0.009
5	sfccape	4 to 11	2	0.044	0.009
6	rh	3 to 14	2	0.036	0.008
7	sfctemp	4 to 15	3	0.032	0.009
8	sfccin	5 to 14	2	0.031	0.007
9	2m_temp	5 to 19	4	0.027	0.009
10	300temp	6 to 16	3	0.024	0.006

The surface-based convective available potential energy (CAPE, *sfccape*) and surface-based convective inhibition (CIN, *sfccin*) are both among the most-important predictors in the moderate and high elevation cases. In the low elevation cases, *sfccin* never appears in the top 30 of all predictors, and *sfccape* ranks between 13th and 24th among all variables. It is possible that CAPE and CIN are acting as more skillful synoptic-scale proxies for heavy rainfall over the western U.S.; in the eastern U.S., the RF more frequently directly uses model quantitative precipitation forecasts (QPFs) as alternatives.

Model PW anomalies, *pw_anom*, appear in the top ten for both the low and moderate elevation cases. For the high elevation cases, *pw_anom* ranks as the 8th- to 26th-most-important predictor. Finally, low-level air temperatures (*2m_temp*, *sfctemp*) appear in the high and moderate elevation cases, along with upper-air temperature (*300temp*) in the high elevation cases. These variables may be associated with the heavy seasonal dependence of flash floods in the moderate and high elevation regions, where most flash floods occur as the result of the southwest summer monsoon and summertime severe convection in the High Plains. In other words, relatively high summertime temperatures (when compared with those from the other three seasons) in the West are associated with an increased likelihood of a flash flood.

Another important result of Table 16 involves the stability of the MDG values and rankings from trial-to-trial. Although there is some variability in MDG rank, variables at the top of the list tend to be consistently highly-ranked from trial to trial. This has important implications for our ability to use MDG for physical interpretation of the RF results – in other words, because MDG does not drastically change from trial to trial,

there is a high degree of likelihood that these results are not simply some sort of statistical artifact.

Cross-Validation of Elevation Regimes

Regional differences in the development of environments favorable for flash floods are accounted for throughout this study via the division of the contiguous U.S. (CONUS) into three regions based upon the average surface station pressure in each GFS3 model grid cell. As shown by Table 16 and its associated discussion, there is evidence that different variables are important to the success of the RF predictions depending on the region over which the RF is being applied. This evidence can be further tested in two ways: by splitting histograms like those in Figures 20, 21, 22, and 23 (though without normalization) by the elevation region associated with each case and by testing the RF models generated for a specific region on another region. Figure 24 is a histogram of the GFS-analyzed PW for all cases in the entire archive, separated by the elevation region of the cases. The low elevation cases (dark blue) tend to have the highest PW, followed by the moderate elevation cases (light blue), and then the high elevation cases (medium blue). However, the differences in the distribution of PW between elevation regions are fairly minor. Note that, although flash flood cases are included in Figure 24, they are so rare in absolute terms that they are essentially invisible.

We can repeat the procedure that led to Figure 20, but restrict it such that only low elevation cases are considered. The result is given in Figure 25, the normalized flash flood vs. non-flood histogram for all low elevation cases in the entire archive. The flash floods have GFS-analyzed PW values between 20 and 60 mm, while the non-floods have PW values between 0 and 50 mm. Non-floods most frequently have a PW less than 10 mm, while flash floods are most frequently associated with PWs around 40 mm.

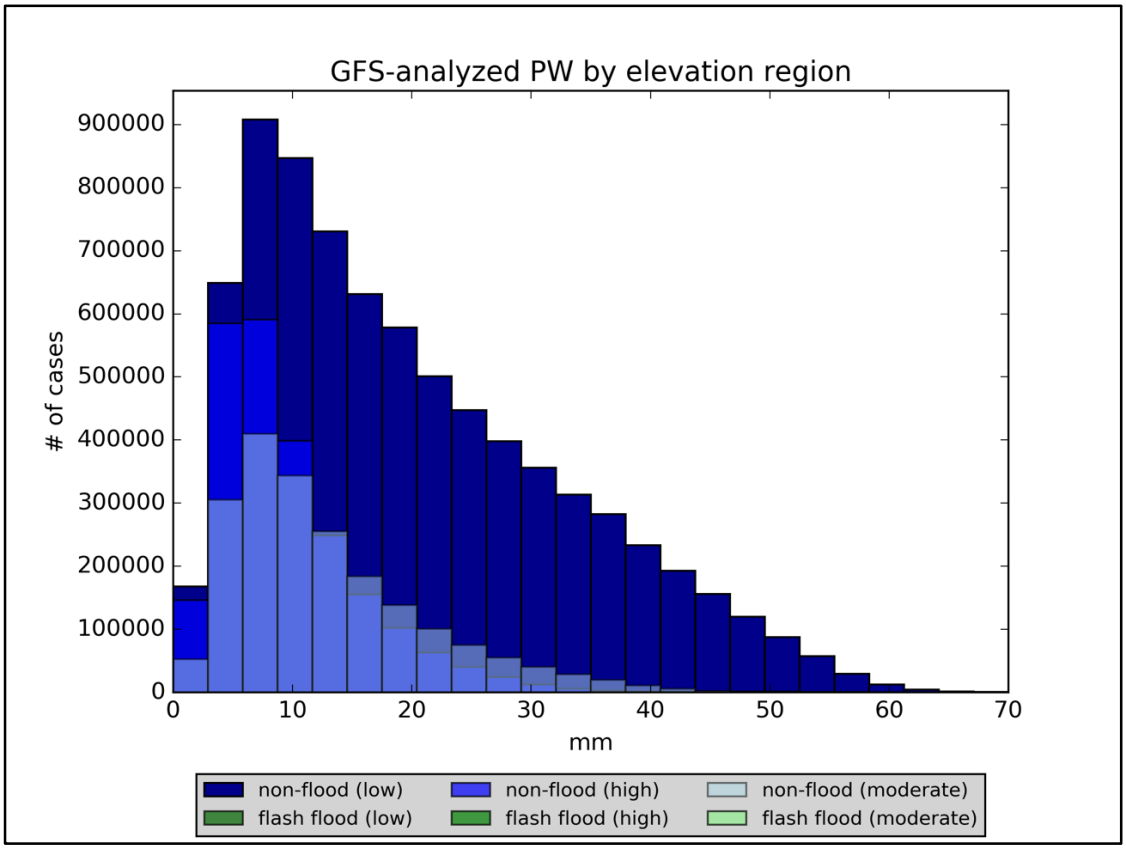

Figure 24. Histogram of GFS-analyzed PW from all cases in the entire archive, separated by elevation region (flash flood reports are so rare that they are essentially invisible)

However, in the moderate elevation cases (Figure 26), flash floods usually have PWs between 10 and 50 mm, less than those in the low elevation area. They are most commonly associated with PW values of just under 30 mm, also less than that seen for the low elevation cases. Non-floods in the moderate elevation region range between 0 and 40 mm, with a strong peak in frequency detected at between 7 and 10 mm.

The high elevation non-flood cases (Figure 27) more closely resemble the moderate elevation non-flood cases than the low elevation non-flood cases. In high elevations, the non-flood PW is most frequently ~8 mm, the same as observed for the moderate elevation cases. However, in the high elevation region, non-flood PWs greater than 15 mm are less-frequently observed than in the other two elevation regimes, and the

range of non-flood PWs for the high elevation cases is fairly narrow, at 0 to 30 mm. Reports of flash floods in the high elevation area are generally associated most frequently with PWs of 25 mm or so, but they range from 10 to 40 mm, a lower (and narrower) range than in either of the other two elevation regions.

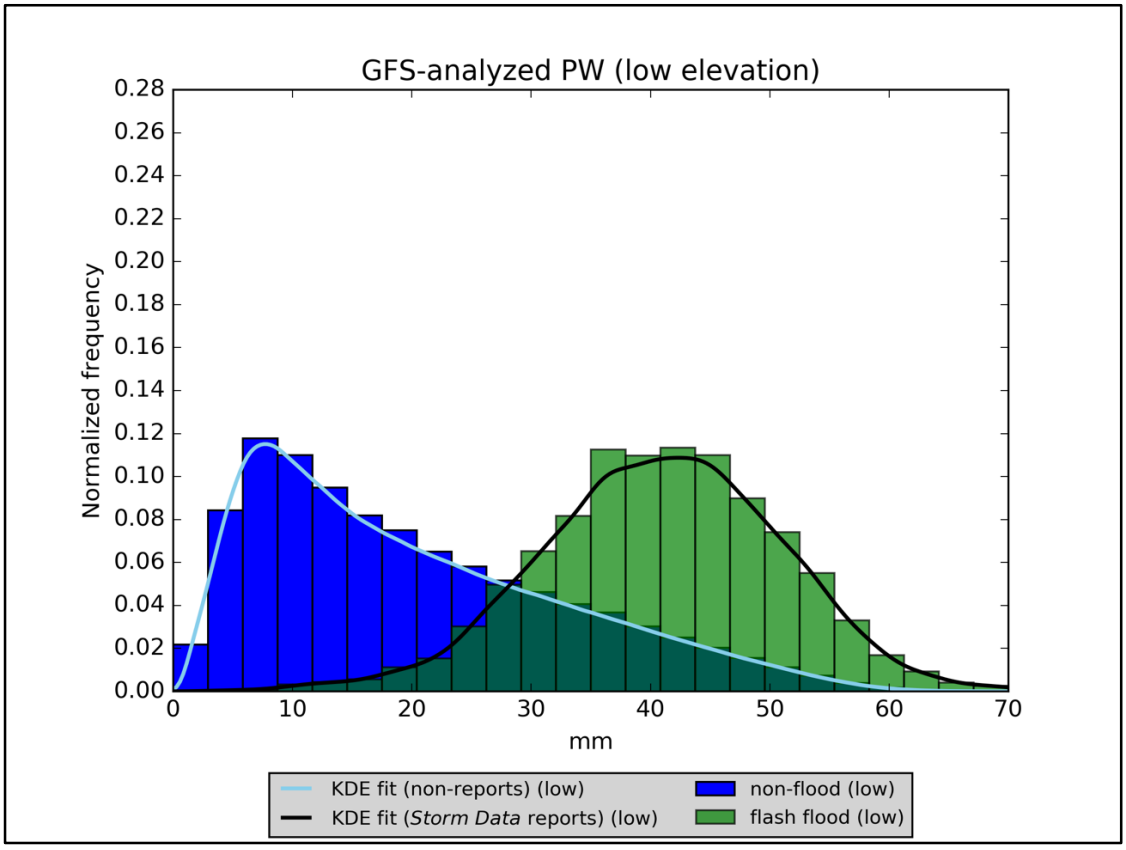

Figure 25. Normalized histogram of GFS-analyzed PW from all low elevation cases in the archive

When raw distance between the KDE fit peaks (in mm) is considered, PW is most effective as a splitter in the low elevation cases; it is slightly less effective for the moderate elevation cases and is the least effective for the high elevation cases. However, if the split is measured by the *shared* area under each of the KDE fits (i.e., the dark green area in Figures 25, 26, and 27), PW is most effective in the high elevation cases, with moderate and low elevation cases bringing up the rear. This same order is observed in

Table 16, where PW has the highest mean MDG value for the high elevation cases, followed by moderate and low elevation cases.

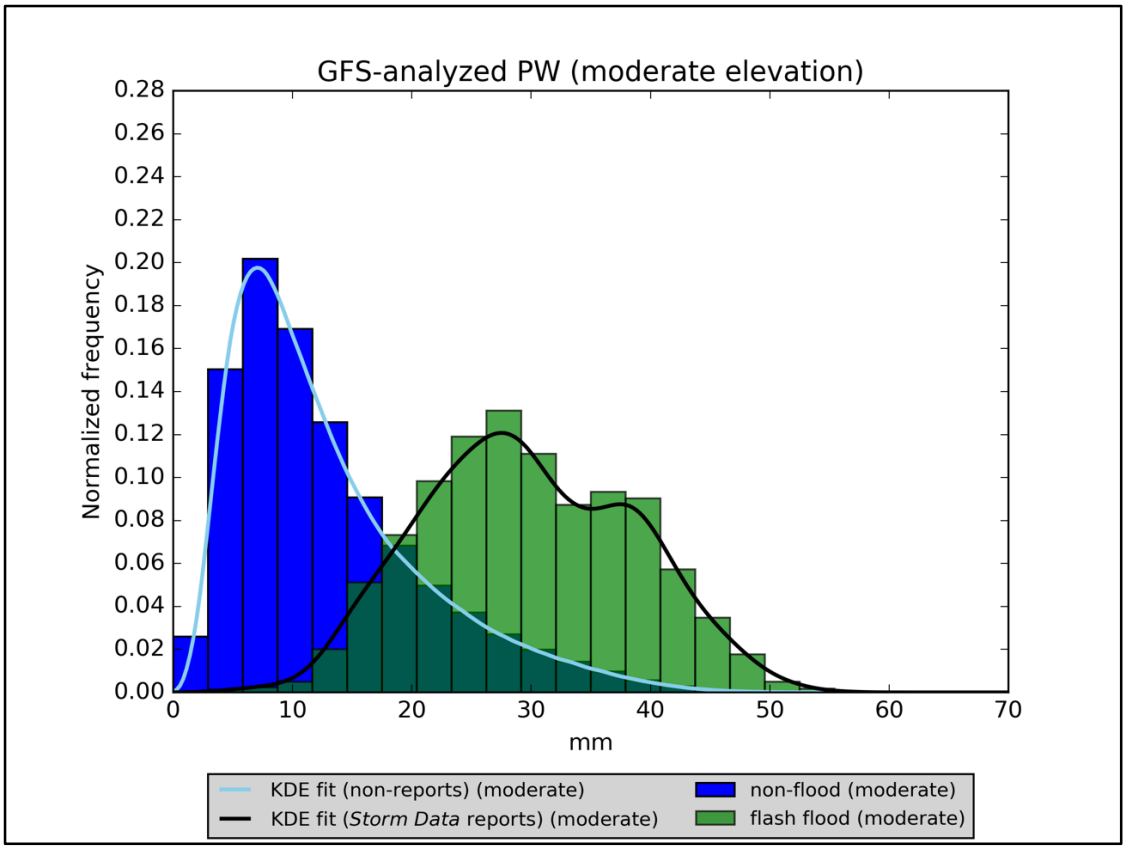

Figure 26. Normalized histogram of GFS-analyzed PW from all moderate elevation cases in the archive

The differences between regions can also be tested via RF cross-validation. I took low elevation test data from the 2013, 2014, and 2015 GFS model epochs and ran the data through RFs fit to the high elevation cases from those years; the resultant Brier scores were between 27% and 108% worse as a result. The reverse was also tested, where the high elevation test data from the same three GFS model epochs were run through RFs fit to the low elevation cases from those years. These Brier scores were 34 to 71% better, which suggests that additional model variables made available in the low elevation cases result in more skillful RFs. In any event, the regionally-specific fitting process proposed

in this dissertation has important implications for the skill of the predictions of flash floods.

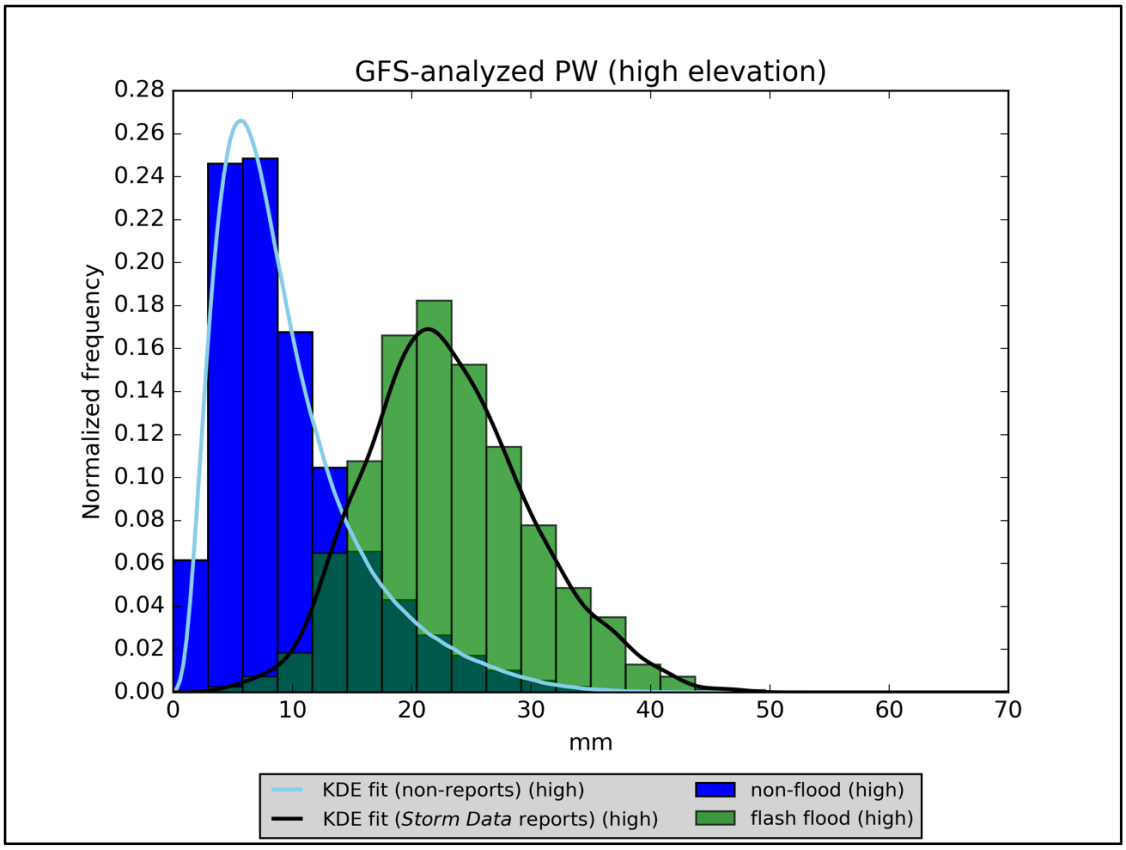

Figure 27. Normalized histogram of GFS-analyzed PW from all high elevation cases in the archive

Forward Selection and Backward Elimination

Table 17 contains the first five predictor variables selected as a result of ten forward selection/backward elimination trials conducted for each elevation regime. To produce Table 17, a stepwise variable selection process was implemented following the procedure in Ahijevych et al. (2016). In this process, each predictor variable was tested in an RF and that variable resulting in an RF with the best Brier score was stored. Then, all the remaining variables were added, one-by-one, to the original variable and an RF generated from each of these predictor pairs; the predictor pair with the best Brier score was stored. Then a third variable was added, choosing from all remaining predictors (one-

by-one) yet to be selected; the triplet of variables producing the best skill in an RF was stored. Then each of the three variables in this triplet was removed in a stepwise fashion and the Brier score of an RF generated on the remaining two variables was recorded.

Table 17. First five predictor variables selected in each of ten forward selection/backward elimation trials for each elevation region

			Low elevation		
	1st variable	*2nd variable*	*3rd variable*	*4th variable*	*5th variable*
1	cpreciprate	pw	soilm_shallow	300omega	925temp
2	cpreciprate	pw	soilm_shallow	400omega	300uwind
3	cpreciprate	1013.25q	soilm_shallow	700q	200hgt
4	cpreciprate	soilm_shallow	pw	500omega	925vwind
5	cpreciprate	soilm_shallow	700q	500omega	850temp
6	cpreciprate	400omega	2m_q	soilm_shallow	850temp
7	cpreciprate	300omega	2m_q	soilm_shallow	soilt_200cm
8	cpreciprate	2m_q	soilm_shallow	300vwind	500omega
9	cpreciprate	pw	soilm_shallow	200vwind	300omega
10	cpreciprate	pw	soilm_shallow	500omega	soilt_shallow
			Moderate elevation		
1	sfccape	sfccin	preciprate	soilt_100cm	2m_temp
2	4layer_li	flashiness	preciprate	2m_temp	250hgt
3	4layer_li	sfctemp	700rh	10m_uwind	850temp
4	4layer_li	700rh	sfctemp	250uwind	soilm_200cm
5	4layer_li	250vwind	sfctemp	2m_q	preciprate
6	soilt_100cm	250vwind	200uwind	sfccin	4layer_li
7	4layer_li	sfctemp	soilm_shallow	400div_q	cpreciprate
8	sfccape	500q	flashiness	850uwind	cpreciprate
9	sfccape	flashiness	250vwind	soilm_shallow	pw
10	sfccape	cpreciprate	sfctemp	200vwind	soilm_200cm
			High elevation		
1	700q	preciprate	sfctemp	300div_q	500vwind
2	pw	sfccin	sfccape	500vwind	soilt_100cm
3	sfctemp	700rh	4layer_li	150hgt	soilt_shallow
4	2m_temp	400vwind	700rh	250uwind	150temp
5	4layer_li	sfctemp	rh	700rh	soilt_100cm
6	700q	cpreciprate	sfctemp	500rh	200magnitude
7	700q	flashiness	soilm_shallow	400vwind	sfctemp
8	sfctemp	preciprate	flashiness	700adv_q	2m_rh
9	4layer_li	pw	sfccin	200omega	flashiness
10	rh	sfctemp	300rh	cpreciprate	150omega

The variable whose removal resulted in the best Brier score was removed. Then the double forward selection process was repeated, followed by a single backward elimination process until each RF was utilizing ten predictor variables. The first five selected for each of the ten trials in each elevation region are shown in Table 17. Table 18 shows how frequently each predictor variable was chosen in the top ten during these trials. If a variable was selected ten times in its elevation regime, that means the variable was chosen to be used in the top ten in every trial. The results show that convective precipitation rate, the PW, and shallow soil moisture are critical to the success of the low elevation RFs, while specific humidity, vertical velocity ("omega"), and a combination of wind components and air temperatures at various levels are also helpful. In the moderate elevation trials, surface-based CAPE and best 4-layer LI are necessary, while surface air temperatures, precipitation rates, soil moisture, and flashiness help considerably. In the high elevation trials, best 4-layer LI, the surface air temperature, relative humidity, and flashiness appear most frequently. Note that, in Tables 17 and 18, speed shear, K index, mean-layer wind, and model PW anomaly were not available for selection by the RFs.

There is one primary difference of interest between the results of these forward selection-backward elimination trials and the MDG analysis shown in Table 16: land-surface variables (soil moisture, soil temperature, and flashiness) are much more prominent in the forward selection-backward elimination process. In this process, the RF is allowed to select the predictor variable that, on its own, produces the best Brier score relative to the *Storm Data* report archive. Then, in the next iteration of the process, a second predictor variable is selected that achieves the best Brier score when used in

combination with that selected in the first step. This results in the RF preferentially selecting as its second predictor something that provides *skillful yet distinct* information from the predictor selected in step one. Therefore, the RF selects variables drawn from different areas of the overall NWP parameter space. One example of this is the difference between land surface and atmospheric moisture information. For example, PW and shallow soil moisture are not highly correlated with one another, but each provides valuable information and is correlated with the occurrence of a flash flood. Therefore, both appear prominently in Table 18, because when the RF has only ten variables to use instead of 146, land surface information is at a relative premium. On the other hand, the MDG analysis in Table 16 ends up not containing the land surface variables for at least two reasons: 1) soil moisture and temperature are highly-correlated with one another throughout the model archive, which acts to artificially suppress their overall MDG values, since each land surface quantity is roughly equally as likely as any other land surface quantity to be selected as a splitter variable at a tree node and 2) the land surface model in the GFS uses a definition of soil moisture that depends heavily upon the texture of the soil particles at each grid cell, and so localized thresholds of soil saturation, not adequately resolved by this study's division of the conterminous U.S. into three regions, would be needed to optimally use the soil moisture variables from the model.

Derived Variables

Several candidate predictor variables were derived from base GFS fields for this study. Although these fields are relatively easy to compute and do not require any data outside that already available from the GFS, it is reasonable to ask whether these derived predictors improve the quality of the RF predictions. Based on other pieces of evidence, including the normalized histograms discussed in the previous section and the MDG

importance rankings, some of the derived variables, including K index, PW anomaly, and specific humidity at certain levels, are probably important and others like speed shear and wind speed (not shown) seem relatively unimportant. For the entire study period and each individual elevation region, an RF was generated without any derived predictor variable that appears in Table 4.

Table 18. Frequency with which predictor variables were selected in the forward selection/backward elimination process

Low elevation		*Moderate elevation*		*High elevation*	
Variable	*# selections*	*Variable*	*# selections*	*Variable*	*# selections*
cpreciprate	10	sfctemp	8	sfctemp	7
soilm_shallow	10	4layer_li	6	4layer_li	6
pw	5	cpreciprate	6	rh	5
400omega	4	sfccape	5	sfccin	4
500omega	4	soilm_shallow	5	flashiness	4
300omega	3	flashiness	5	300uwind	4
850temp	3	preciprate	4	700temp	3
250vwind	3	250vwind	4	700q	3
400uwind	3	sfccin	3	700rh	3
2m_q	3	700rh	3	500vwind	3
soilt_40cm	3	500rh	3	2m_rh	2
200vwind	2	soilm_200cm	2	400omega	2
mslpres	2	250uwind	2	soilt_200cm	2
150uwind	2	200uwind	2	pw	2
925hgt	2	250hgt	2	cpreciprate	2
700q	2	10m_uwind	2	preciprate	2
200hgt	2	pw	2	sfccape	2
700hgt	2	rh	2	300div_q	2
850vort	2	150cloud	2	soilt_100cm	2
925uwind	2	300omega	2	400vwind	2
300vwind	2	2m_temp	2	500rh	2
(30 others)	1	soilt_100cm	2	soilm_shallow	2
		soilt_40cm	2	700uwind	2
		(24 others)	1	250uwind	2
				(30 others)	1

This RF was used to produce predictions on its independent validation dataset. The Brier score of these predictions was compared to the Brier score of predictions generated on the independent validation dataset of an RF trained using all the candidate

predictors. Over 15 trials on the low elevation dataset, the RF without any of the derived predictors resulted an average Brier score of 0.095, while the RF that included all of the derived predictors resulted in an average Brier score of 0.096. In the moderate elevation cases, the average Brier score with all variables was 0.102, and 0.103 when the derived variables were excluded. For the high elevation cases, the mean Brier score was 0.094 when derived variables were excluded and 0.099 when they were included. These differences are statistically-insignificant at a 95% confidence level, which indicates that derived predictors are not required for the RF to generate skillful forecasts of flash floods. With the exception of the model PW anomaly, K index and specific humidity at various levels, none of these derived predictors are observed to have high MDG values or rankings, so it is possible that future manifestations of this method should include those three types of derived predictor while excluding all the others.

Optimal Number of Predictors

The optimal number of predictor variables to be fed to an RF model is the smallest number that results in the best prediction quality, the least amount of computer time to generate and utilize the forest, and the easiest physical interpretation of the internal RF importance metrics. In the present study, a simple experiment is conducted to determine an optimal number of predictor variables. The experiment proceeds thusly: for each elevation region, generate a 300-tree RF using all available predictor variables and test on the RF's independent validation dataset. Rank all the predictor variables using the MDG metric and record the Brier score of the RF's predictions on the validation set. Now fit a new RF to the original training set, this time excluding the predictor variable with the lowest-ranked MDG value and record the new Brier score. Proceed, excluding the predictor with the new lowest-ranked MDG value each time, until only one variable is

left. Figure 28 is a plot the Brier score of the predictions from the validation set as a function of the number of available predictor variables.

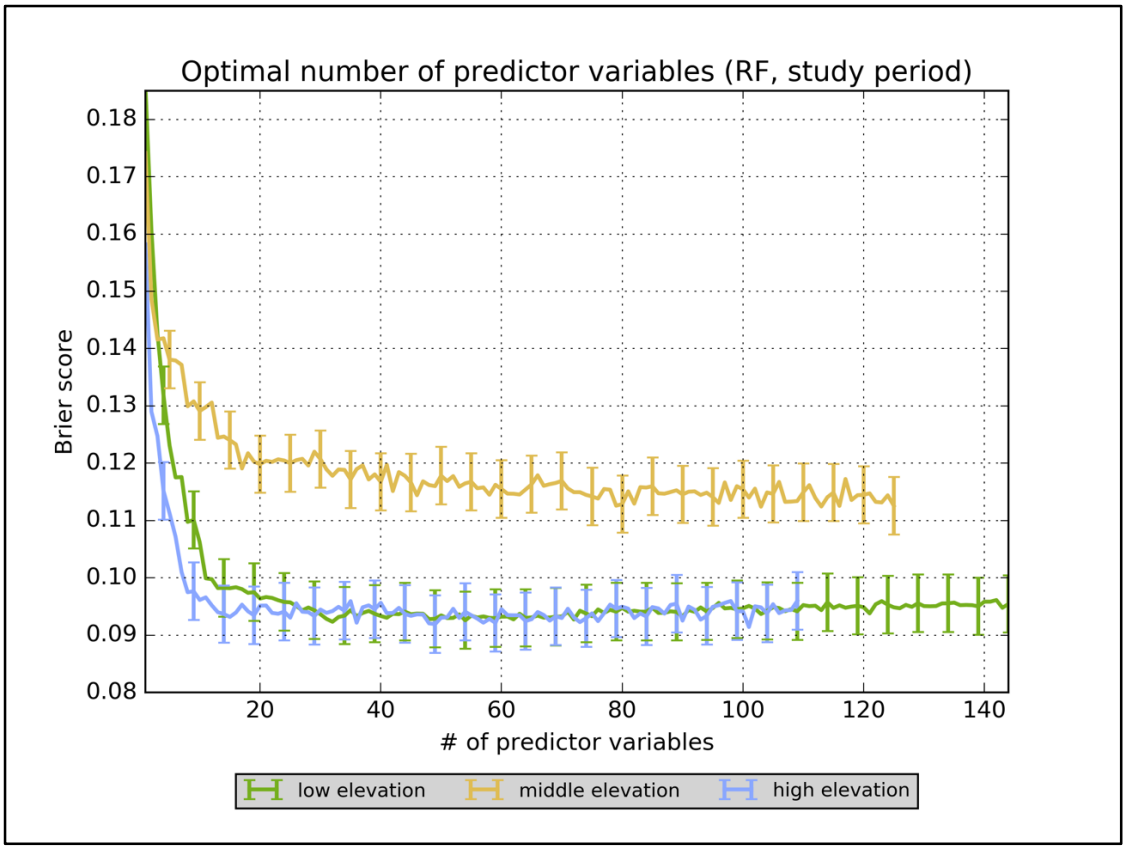

Figure 28. Plot of Brier score of RF predictions as a function of the number of predictor variables used to generate the RF

The error bars in Figure 28 encompass the distribution of Brier scores expected when 15 independent fitting and validation trials are run at each step of the test. The Brier scores range from 0.092 (when 32 variables are used) to 0.185 (when one variable is used) for the low elevation cases, from 0.113 (when 79 variables are used) to 0.174 (when one variable is used) for the middle elevation cases, and from 0.092 (when 36 variables are used) to 0.158 (when one variable is used) for the high elevation cases. Figure 28 shows a pattern similar to that observed from the RF meta-parameter (*ntree*, *dtree*, and *mtry)* analysis. The Brier score quickly improves as the first ten to 20 predictors are added

to the RF, with diminishing returns kicking in quickly after about the 20th predictor variable is added. There is a slight, though statistically-insignificant, decrease in performance once the optimum number of predictors is exceeded. This observation matches up with previously-reported statistical properties of the RF method regarding overfitting, collinear candidate predictors, and large numbers of candidate predictors.

Table 19 contains the 20 predictors used in a parsimonious RF model resulting from this predictor optimization test. Most of these predictors have been associated with flash flooding and/or heavy rainfall in the literature. K index is the most important predictor for the low elevation cases and the second-most-important for the middle elevation cases (the K index is not defined for the high elevation cases and thus is not available for use there). The specific humidity at a particular level appears five times in the low elevation list, four times in the middle elevation list, and four times in the high elevation list. PW appears in the top five for all three sets of cases, and the model PW anomaly, though slightly less important than the raw PWs, appears in the top ten of all three lists. PW, specific humidity, and PW anomaly all speak to the amount of moisture available for precipitation in the atmosphere.

Other commonalities between regions include surface-based CAPE, the best 4-layer LI, omega at multiple levels, and the integrated relative humidity. CAPE and LI are measures of buoyancy and are predictors of moist convection, which is required for most flash floods (Doswell et al. 1996). There are also intriguing differences between the three lists. For the low elevation cases, soil moisture from 0 – 10 cm and from 10 – 40 cm BGL appears in the list, but no soil moisture quantity appears in the middle or high elevation cases. The middle and high elevation regions are mostly in the western third to half of

CONUS; previous studies have suggested that soil moisture is less important to flash flood forecasting in the western U.S. (Smith 2003).

Table 19. Most important predictors (MDG) as determined by the variable elimination process

Low elevation		*Middle elevation*		*High elevation*	
MDG rank	*Variable*	*MDG rank*	*Variable*	*MDG rank*	*Variable*
1	k	1	4layer_li	1	700q
2	700q	2	k	2	4layer_li
3	pw	3	pw	3	2m_q
4	cpreciprate	4	700q	4	sfccape
5	pw_anom	5	850q	5	pw
6	preciprate	6	250vwind	6	sfccin
7	850q	7	2m_q	7	sfctemp
8	400omega	8	pw_anom	8	500uwind
9	4layer_li	9	sfccape	9	pw_anom
10	500omega	10	500q	10	2m_temp
11	soilm_shallow	11	200vwind	11	500vwind
12	300omega	12	300vwind	12	500q
13	rh	13	sfctemp	13	rh
14	925q	14	sfccin	14	400omega
15	250omega	15	400vwind	15	400q
16	sfccape	16	2m_temp	16	250temp
17	500q	17	500omega	17	150hgt
18	250vwind	18	150vwind	18	300uwind
19	2m_q	19	500vwind	19	10m_vwind
20	soilm_40cm	20	rh	20	500omega

The v-components of the winds at various levels are in all three lists, but the properties of the wind fields certainly appear more frequently in the middle and high elevation cases. The majority of the flash floods in the middle elevation dataset occurred in the High Plains and the southwest monsoon regions. Heavy rainfall and flash floods in both these regions are associated with southerly winds and associated northbound moisture transport (i.e. positive v-components) from the Gulf of Mexico and the Gulf of California, respectively. The cases from the high elevation dataset also frequently arise from the southwest monsoon pattern. Low-level air temperatures appear in the middle

and high elevation lists, which is likely also associated with the seasonal cycle of flash floods observed in the southwest monsoon region and with the spring and summer severe weather season in the High Plains. The appearance of the 250-hPa meridional wind in the low elevation list is likely associated with jet streaks and large-scale forcing for ascent (which is why direct forecasts of ascent in the omega fields also appear in the list). Note that the GFS QPF appears *only* in the low elevation list. Both are absent from the other two lists (each quantity is ranked between 20th- and 40th-most-important by MDG in each list); there are several possible explanations for this. One is that the GFS QPF is those regions of the U.S. is less skillful than it is the lower elevation area. Another likely component is that western flash floods occur as a result of small-scale individual storm cells, while eastern flash floods are associated with large-scale mesoscale convective complexes and other weather systems that occur with characteristic length scales that can be adequately resolved by 1-degree GFS data. Strong synoptic-scale forcing for ascent is clearly critical to the overall success of this method as applied to GFS data.

No derived variables appear in the top 20 of any of the three lists with the exception of the K index and the model PW anomaly. Many past studies have found that moisture flux convergence (MFC) is an important predictor in flash flood forecasting, both at individual atmospheric levels and in a vertically-integrated form. The vertically-integrated form was not tested in the present study, but the MFC calculated on any individual level appears far down the list of variable importances in all three regions. Other derived variables, including wind speed and speed shear, also appear far down the three lists. While many of these quantities have been associated with flash floods, they have not been used to *distinguish* between flash floods and non-floods. Therefore, it is

likely that low to moderate speed shear or weak layer-mean winds are necessary but not sufficient conditions for the development of a flash flood. It is also possible that directional wind shear, not considered in the present study, may be more applicable to this problem than speed shear or the mean wind through an atmospheric layer.

Based upon the results of this trial, the RF should continue to be used with all available predictor variables as long as they are available from the GFS post-processor. The process of calculating wind speeds, speed shears, PW anomaly, K index, and the other derived variables is extremely fast and adds little to no overhead to the process of generating predictions from GFS data in real-time. Less important raw GFS model fields should continue to be included in any RF, as well, because the GFS model data are distributed in a single bundle, so inclusion of relatively unimportant GFS model fields requires only a trivial additional amount of required processing and storage.

However, there is one major advantage to variable reduction: interpretability. When used in a context in which diagnosis of the pattern and magnitude of probabilistic predictions is critical, a 20-variable RF will be easier to understand and explain. Because MDG is calculated on the basis of the change in Gini impurity at each node, MDG variable importances can be automatically provided to end-users of the RF predictions associated with each run of the GFS.

Summary

The experiments outlined in this chapter demonstrate that the RF algorithm can successfully be applied to GFS output to skillfully forecast flash floods and simultaneously improve our understanding of the atmospheric and hydrologic factors that contribute to them. Expert variable selection was used initially to restrict the number of

GFS model fields provided to the RF algorithm. Then, by aggregating the results of past case studies and larger-scale analyses of flash flood events across the U.S., a series of additional candidate predictors were derived and also used in the RF algorithm, although most of these additional predictors do not result in statistically-significant improvements in the skill of the RF-generated predictions.

Several tests outlined in this chapter resulted in RFs that use various combinations of candidate predictors and cases from different regions of the U.S. Results from these show that plausible physical interpretations of the RF output can be made. However, additional research, including generating a larger number of regional forests instead of just three, could yield even better physical understanding of how environments favorable for flash floods should be characterized.

Chapter 5: Case Studies and Research-to-Operations Activities

This chapter consists of four short examples for which the proposed machine learning (ML) flash flood prediction system was put to use. Two of these examples are from archived case studies (May 31, 2013 in Oklahoma City, Oklahoma [OKC, OK] and May 2015 across the U.S. Southern Plains) and are designed to serve as examples for how ML predictions of flash floods would look in real-time use. The third example describes the set-up, use, and evaluation of ML flash flood predictions based upon Global Forecast System (GFS) model data during the 2016 Multi-Radar/Multi-Sensor (MRMS) Hydrometeorological Testbed Experiment (HMT-Hydro) conducted as part of the 2016 Experimental Warning Program (EWP) at the Hazardous Weather Testbed (HWT) at the National Weather Center (NWC) in Norman, OK. Finally, the fourth example consists of geographical cross-validation of the proposed ML model, via application of the proposed model to an archive of reports of flash floods from the European continent, collected by the European Severe Storms Laboratory (ESSL). The fourth example demonstrates that ML predictions of flash floods can be generated globally and in real time using the same strategies outlined in Chapters 3 and 4.

31 May 2013: Oklahoma City, Oklahoma

From the evening of Friday, May 31, 2013 into the morning of Saturday, June 1, 2013, an area of extremely heavy rainfall developed over central OK and the OKC, OK metropolitan area. Earlier on that Friday afternoon, a tornadic supercell developed just west of the built-up area of OKC; the tornado and its associated impacts received continuous coverage from all major broadcast television outlets in OKC between 2200 UTC (5 pm Central Daylight Time [CDT]) 31 May 2013 and 0500 UTC (midnight CDT)

1 June 2013. Just 11 days before, a devastating tornado had torn through the heart of Moore, OK, a suburban community abutting the southern edge of OKC. The combination of public fear of tornado impacts and wall-to-wall television coverage of severe weather throughout the second half of May 2013 resulted in tragedy on the evening of May 31 and the morning of June 1, when 14 people lost their lives due to flash flooding (Suffern et al. 2014).

As the event unfolded, I closely monitored broadcast television and social media outlets, and in the immediate aftermath of the event monitored print media, as well. As a result of this monitoring process, I collected several dozen highly-specific locations of reported impacts of the flash flood. I subsequently geocoded these to latitude-longitude pairs collocated with susceptible infrastructure using aerial imagery to identify relevant bridges, culverts, and other structures. Later, I interviewed a representative of the City of Oklahoma City Police Department about the evolution of the flash flood and the municipal government's response to it. As a result of this interview, I was provided with detailed information about the flash flood's impact to neighborhoods across OKC and the surrounding communities of central OK (F. Barnes, personal communication, November 6, 2013). Because there is evidence that the impacts of the flash flood were made worse by abnormally large numbers of vehicles on the highways and roads of central Oklahoma during the event, I obtained and analyzed vehicle counts from several dozen points across the OKC metropolitan area from the Oklahoma Department of Transportation (ODOT) for May 31, 2013 (M. Folsom, personal communication, April 25, 2014). These processes, though labor-intensive, resulted in significantly more

information about the impacts of the flash flood than appear in official National Weather Service (NWS) sources like *Storm Data* or Suffern et al. (2014).

Meteorological Synopsis

According to the National Severe Storms Laboratory's (NSSL) MRMS radar quantitative precipitation estimate (QPE) product, a series of supercell thunderstorms produced widespread rainfall amounts between 150 and 300 mm over central OK between 1200 UTC 31 May 2013 and 1200 UTC 1 June 2013, as shown in Figure 29.

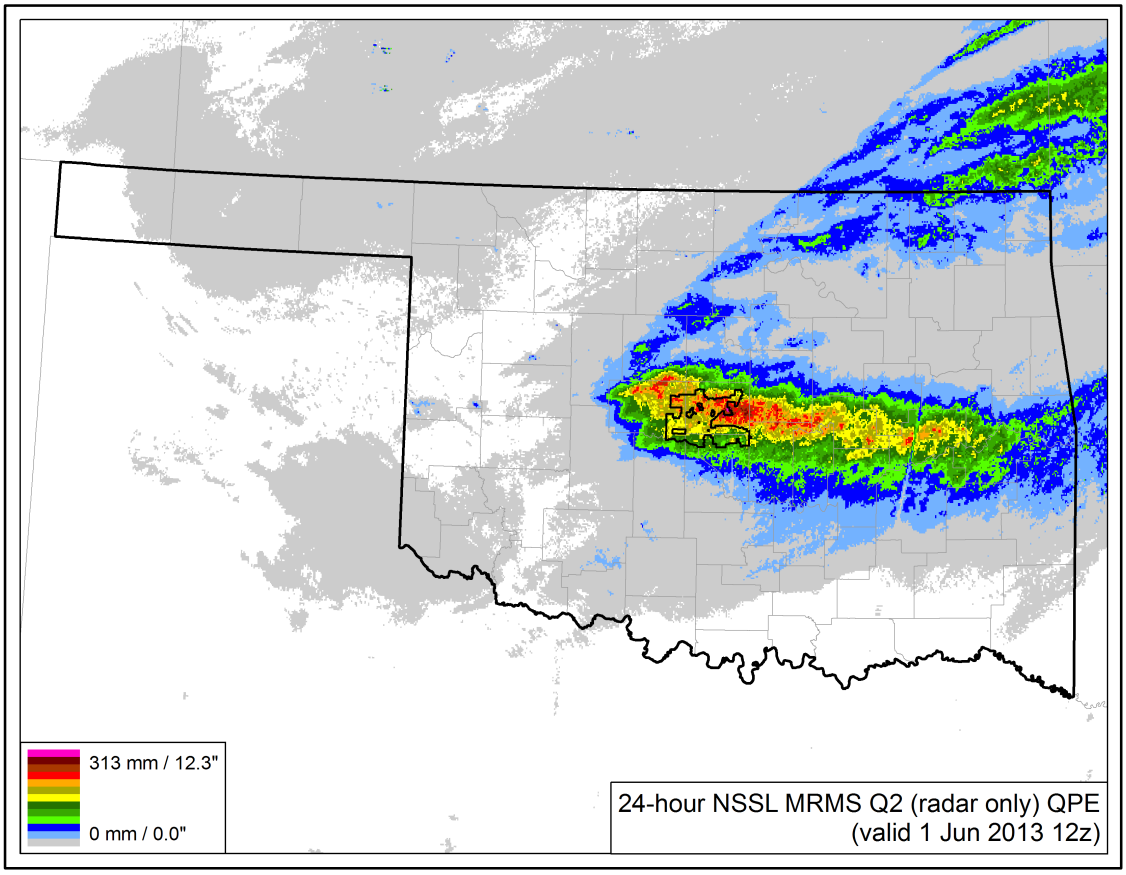

Figure 29. 24-h NSSL MRMS Q2 (radar only) QPE valid 1200 UTC 1 June 2013, with the municipal boundaries of the City of OKC marked with the black line at the center of the state of Oklahoma

As evidenced by Figure 29, the greatest QPE was extremely focused in a zonally-oriented two-county wide band directly centered over the OKC metropolitan area. Comparison of the QPE product in Figure 29 with corresponding rain gauge data from

the Oklahoma Mesonet indicates generally fair agreement between the radar-based estimates and "ground-truth" data, but there is some slight overestimation (as shown in Figure 30) of rainfall by the radar-only QPE product.

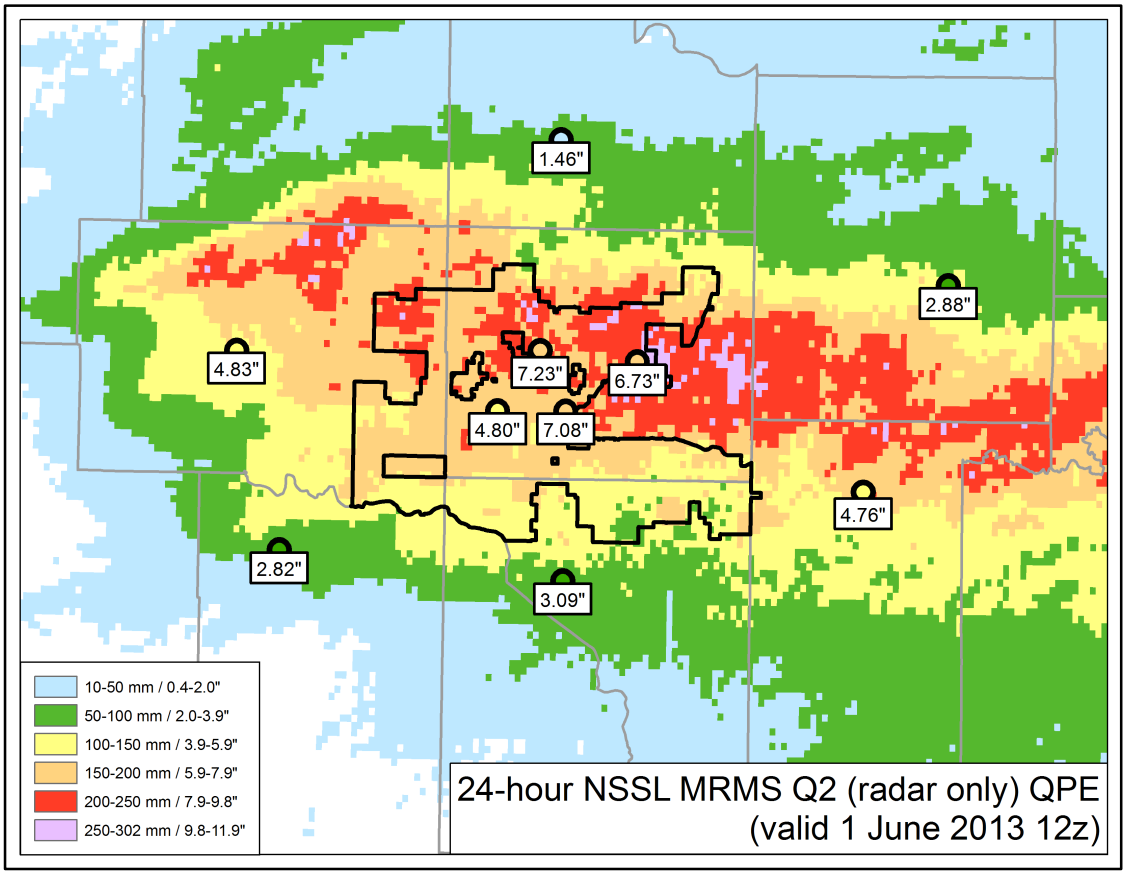

Figure 30. 24-h rainfall totals from Oklahoma Mesonet rain gauges (in inches) overlaid on corresponding NSSL MRMS Q2 QPE, valid 1200 UTC 1 June 2013

Data from the Oklahoma Mesonet's "Oklahoma City North" station, (shown with a "7.23"" in Figure 30), shows that rainfall began at 2330 UTC on May 31 and ended at 0940 UTC on June 1. Of the 184 mm recorded in ten hours and ten minutes at this gauge (18 mm h^{-1}), 80 mm fell during the first 90 minutes of the event (53 mm h^{-1}) and another 100 mm fell during a 5-h period near the end of the event (20 mm h^{-1}). The remaining 4 mm of rainfall occurred sporadically over the last three hours and forty minutes of the event. Therefore, the rainfall forcing to this flash flood was characterized by two distinct

periods: one short period with very heavy rainfall rates and a second, longer period, with just heavy rainfall rates.

The environment in which these rain totals occurred was characterized by high moisture and instability. The closest upper-air sounding site to the flash flood is located in Norman, OK (OUN), near the dot labeled “3.09"” in Figure 30. In the 1200 UTC 31 May 2013 sounding from OUN, mean specific humidity, q, was observed to be 16.6 g/kg, precipitable water (PW) was 37.1 mm, and the K index was 29. A special sounding was conducted six hours later, at 1800 UTC 31 May 2013, and at this time, the PW had increased to 38.1 mm, but q was still 16.6 g/kg and the K index decreased to 26. These three parameters, according to sounding data, were each maximized in the 0000 UTC sounding from 1 June 2013, where PW increased to 42.2 mm, q was 18.2 g/kg, and the K index was 33. Relative to the historical climatology of observed PW at OUN, the 1200 and 1800 UTC values were greater than the 30-d moving average of 90th percentile PWs and the 0000 UTC PW was the highest ever recorded for that date and time.

Forecast tools were in general agreement in the days before the event that a flash flood was possible. However, these tools often disagreed upon on the location of potential flash flood impacts. The NWS Weather Prediction Center (WPC) issued a quantitative precipitation forecast (QPF) at 0949 UTC on May 31, valid for the 24 hours ending 1200 UTC on June 1, which showed 101 mm of rain falling in far northeastern OK, with amounts ranging from 19 to 38 mm over the area in which the flash flood was later observed. The Norman, OK Weather Forecast Office (OUN WFO) noted the potential for heavy rainfall and an associated flash flood beginning on the afternoon of Tuesday May 28; this concerned continued in the area forecast discussions issued on Wednesday,

Thursday, and Friday afternoons. A Flash Flood Watch (FFA) for central and north-central OK was issued on May 30 at 1730 UTC, as shown in Figure 31.

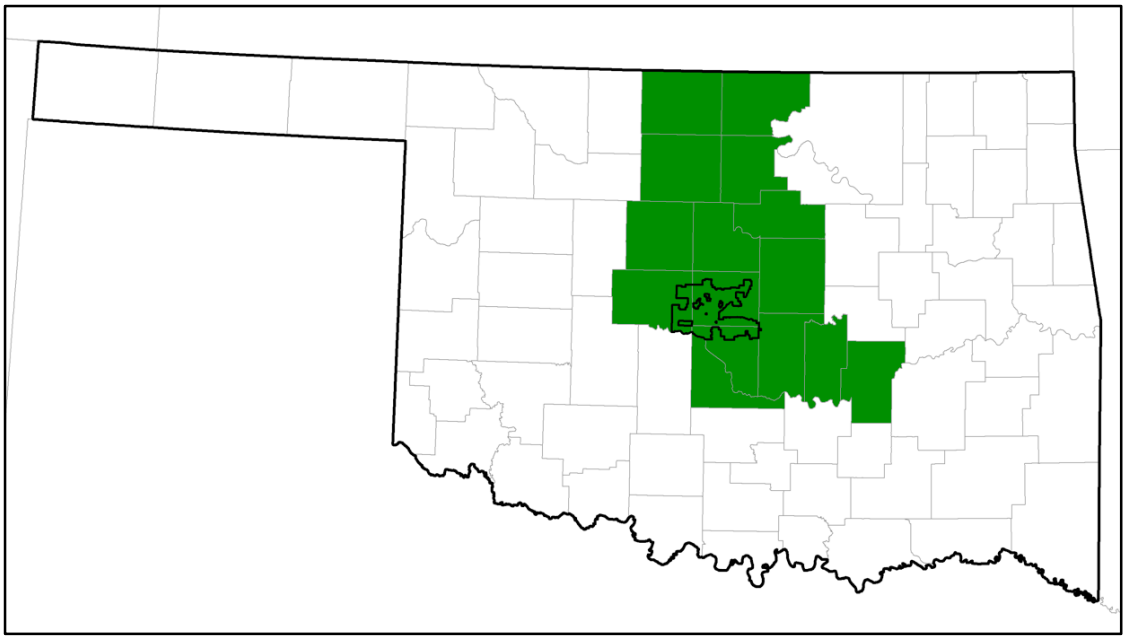

Figure 31. FFA issued 1730 UTC 30 May 2013 for potential flash flood between the evening of 31 May 2013 and the morning of 1 June 2013

As the event approached, convection-allowing numerical weather prediction (NWP) models (CAMs) highlighted northern OK for heavy rainfall potential. QPF from the High Resolution Rapid Refresh (HRRR) model in particular matched the shape and time of the resultant QPE (Figures 20 and 30), but with large location errors, shown in Figure 32. The HRRR QPF is centered approximately 150 km northeast of where the greatest observed rainfall fell.

At 2221 UTC 31 May 2013, prior to the development of the flash flood, the WPC issued a Mesoscale Precipitation Discussion (MPD) highlighting several ingredients favorable for a flash flood over parts of the southern plains, including moist southerly inflow, surface dew point temperatures in the low 70s °F (low 20s °C), PW just under 40 mm, 8 °C km^{-1} lapse rates from 700 to 500 hPa, and 3,000 to 5,000 J kg^{-1} of convective

available potential energy (CAPE). Four-and-a-half hours later, at 0257 UTC 1 June 2013, a second MPD suggested that "1-2 inches of rain per hour [25-50 mm h^{-1}] is expected with flooding possible where local training allows activity to last a couple of hours." This second MPD referenced 850-hPa flow running largely parallel to a stationary boundary draped across northwestern OK; this boundary had helped to initiate the tornadic convection that later evolved into the heavy rainfall producer. What was not forecast by this second MPD was the redevelopment of convection and associated heavy rain rates that dropped 100 mm of rain over 5 hours in areas that had just received 80 mm of rain in 90 minutes.

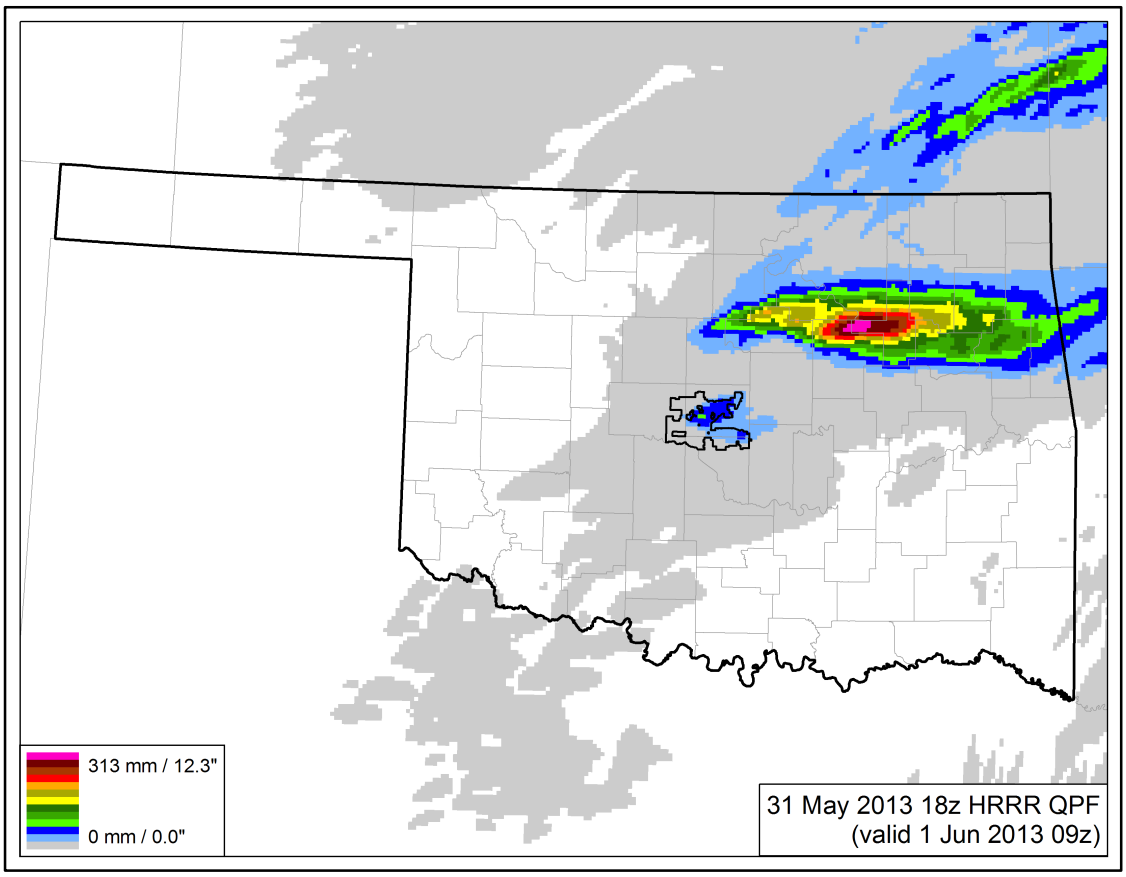

Figure 32. HRRR 15-h QPF initialized 1800 UTC 31 May 2013 and valid 0900 UTC 1 June 2013 (K. Mahoney, personal communication, January 9, 2014)

Hydrological Synopsis

Throughout 2012, Will Rogers World Airport in OKC (KOKC) received 81% of its 1980-2010 normal annual precipitation. In January 2013, KOKC received 82% of its 1980-2010 normal January precipitation. Lake Hefner, one of OKC's primary reservoirs, reached its lowest level since the 1970s; boats at a popular marina were entirely cut off from water that January as the water level dropped that winter (Crum 2013). On January 30 and continuing through February 22, the City of OKC exercised their rights to water from Lake Canton, located on the North Canadian River 100 km upstream of OKC, due to worries about the City's water supply for the upcoming spring and summer seasons. Over half the water volume stored in Canton was released into the North Canadian River during this time. That amount of water (equivalent approximately 33% of Lake Hefner's normal capacity) caused Hefner's level to jump by 3.5 m in February 2013; Lake Canton did not return to normal levels until over three years later, on April 22, 2016. After above-normal rains at KOKC in February 2013, and below normal rains in March, all of Oklahoma was experiencing severe (or worse) drought at the end of March. In particular, the area eventually affected by the flash flood at the end of May was in severe drought at the end of March.

In April, KOKC recorded 192 mm of precipitation – 247% of its 1980-2010 normal April precipitation, and more precipitation in a single month than had fallen in the entire period from October 15, 2012 to March 31, 2013. As a result of this rain, and that which fell in May 2013 prior to the 31st, most of central OK had exited drought by May 28, and only the far western and northern parts of the OKC metropolitan area were considered "abnormally dry". Additionally, most of the Oklahoma Mesonet stations in the OKC metropolitan area recorded greater than 90% soil saturation at 5 and 25 cm

below ground level in the weeks prior to the flash flood. May 2013 ended up being the wettest May, and the second-wettest month ever, in OKC history, as 312% of the city's normal May precipitation was recorded. (Since that time, May 2015 has obliterated both marks; OKC received nearly 450% of its normal May rainfall in May 2015, a full 33% more than that observed in May 2013).

OKC is drained by the North Canadian River and the Deep Fork Creek. The North Canadian (officially named the Oklahoma River within the boundaries of Oklahoma County, for which OKC is the county seat) flows through the heart of the city and drains much of the city's southern side, as well as its near northern side. The Deep Fork Creek drains the vast majority of northern OKC, including some areas quite close to the northern bank of the North Canadian. Brock Creek and Lightning Creek, tributaries of the North Canadian, drain two small, heavily-urbanized catchments on the city's near south side. Although both creeks are surrounded by a fair amount of parkland, many low-income homes lie in the watersheds of these two creeks. Through central OKC, the North Canadian has been controlled by three low-water dams since the early 2000s. An additional upstream dam at Lake Overholser, approximately 10 km upstream of downtown OKC, serves recreational, water storage, and flood control purposes.

On the night of May 31, 2013, heavy rains overwhelmed the Lightning Creek watershed; hundreds of homes were affected by the waters. On the north side of town, waters from the Deep Fork Creek flooded low-lying spots on Interstate Highway 44 (I-44) and Interstate Highway 235 (I-235). Areas in the North Canadian flood plain, including downtown OKC, had not been seriously affected by floodwaters in decades,

but basements and ground floors alike throughout the central business district were subject to flooding during the event.

Impacts

The impacts from the flash flood in OKC were significant. Traffic was snarled, homes and businesses were flooded, and 14 people lost their lives (F. Barnes, personal communication, November 6, 2013). Figure 33 shows areas where roads were closed or damaged due to the flash flood; it also contains the locations of the reported fatalities resulting from the event.

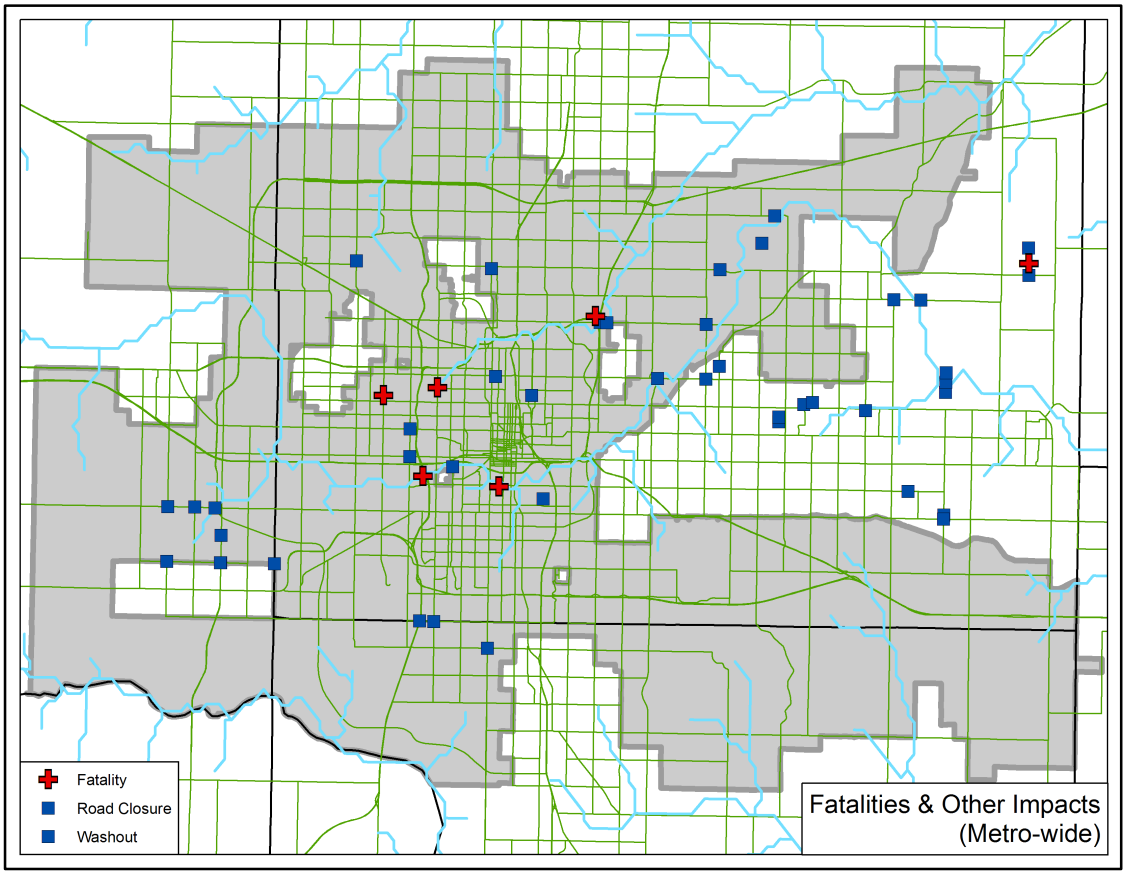

Figure 33. Fatalities and major infrastructure impacts observed as a result of the May 31, 2013 flash flood in central OK

Of the 14 fatalities, seven occurred when a family took shelter from what they believed to be a violent tornado in a culvert near NW 30th St. and N. Meridian Ave.; this

culvert, though often dry, contains the headwaters of the Deep Fork Creek. Days after the event, one body from this group was recovered in the Deep Fork Creek at NW 36th St. and N. May Ave., just over 3 km downstream, and another body was recovered from the Deep Fork Creek just south of the junction between I-35 and I-44, 13.5 km downstream of the culvert. This family primarily spoke Spanish at home; in central OK, Spanish-language severe weather and flash flood television and radio coverage is harder to obtain than English-language coverage (Suffern et al. 2014).

Another five fatalities were reported when a group of 11 people took shelter from the tornado in a culvert just south of the N. Canadian River at I-44. Four bodies were recovered near the culvert, while another was recovered in the days after the event atop the low-water dam between S. Walker Ave. and S. Western Ave. Another two fatalities were reported the morning after the event; one victim drove his vehicle into the swollen N. Canadian River in rural NE Oklahoma County and the second victim was reported to have died from driving into floodwaters near the town of Clearview, OK, in Okfuskee County in east-central OK.

Figure 34 is a map of structures damaged in central OK as a result of the flash flood. The City of OKC estimated that the flash flood caused $17 million in infrastructure damage across central OK, primarily in OKC proper, although damage to schools, parks, and other structures was reported in Edmond, Luther, Jones, Choctaw, Midwest City, and Del City, as well. Roads were the most-frequently impacted type of structure in the event, with over 40 washouts and closures reported. Between 0014 and 1400 UTC 1 June 2013, the City of OKC dispatched first responders to 114 separate calls for flood rescue or flood assistance, mainly involving motorists.

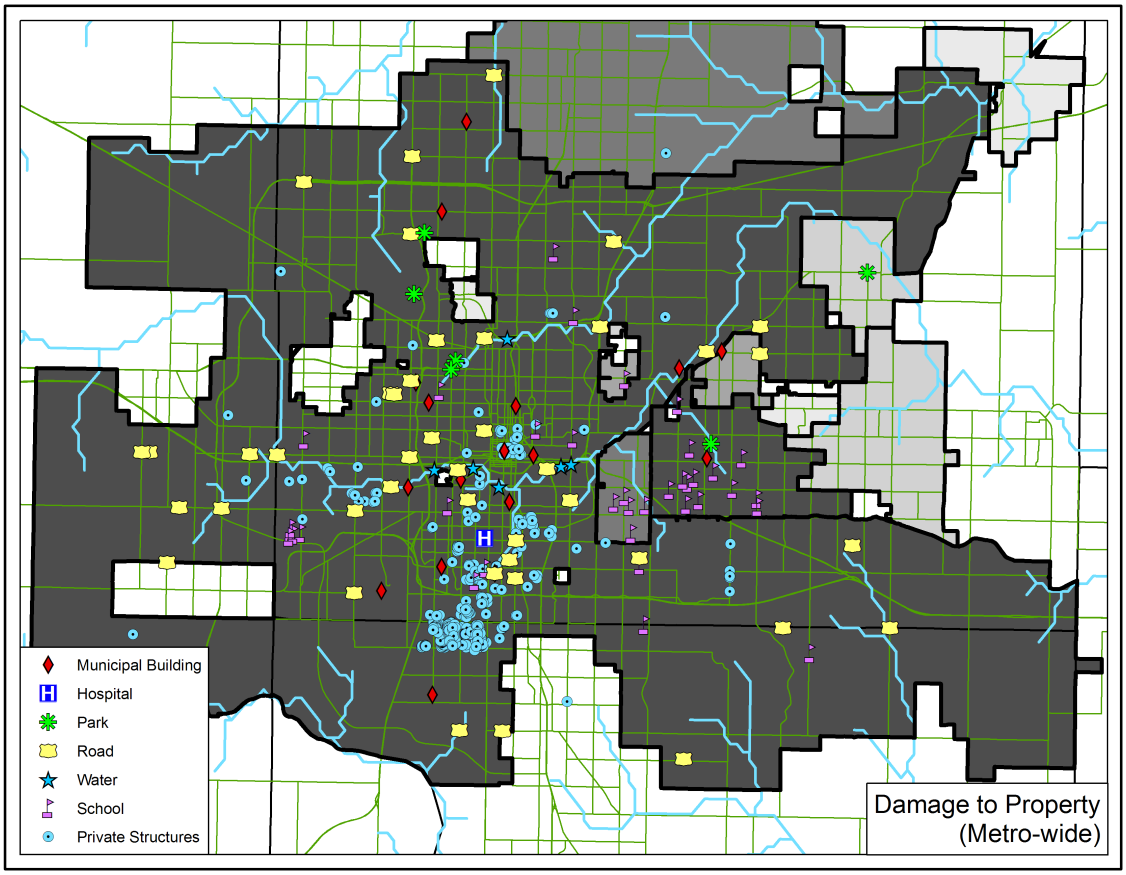

Figure 34. Map of structures damaged as a result of the May 31, 2013 flash flood in central OK, with dark shading corresponding to increasing monetary impacts by municipality

Traffic was abnormally high during the event, according to ODOT data shown in Figure 35 (M. Folsom, personal communication, April 25, 2014). An EF-3 tornado, at 2.6 km, the widest ever recorded in official NWS statistics, affected central Canadian County, just to the west of OKC, prior to the flash flood. In response, many people to the east of this, in the OKC metropolitan area, fled their homes via automobile. In Union City, OK, traffic on May 31, 2013 was 88% higher than average just to the south of the tornado, as residents (and likely, storm chasers, too) filled U.S. Highway 81 (US-81). In between El Reno and Yukon, OK, traffic on I-40, which was closed for hours after the tornado, was 13% lower than average. Many residents of central OKC drove south or east, away from the tornado and the flooding rainfall. Traffic on US-77, I-35, and State

Highway 9 (SH-9), all in Norman, OK just to the south of OKC, was 46%, 25%, and 17% higher than average, respectively. In Mustang, OK, on OKC's southwest side, traffic on SH-152 was 25% higher than normal, while I-35 in Moore, OK, between OKC and Norman, was 16% busier than the 2012 average.

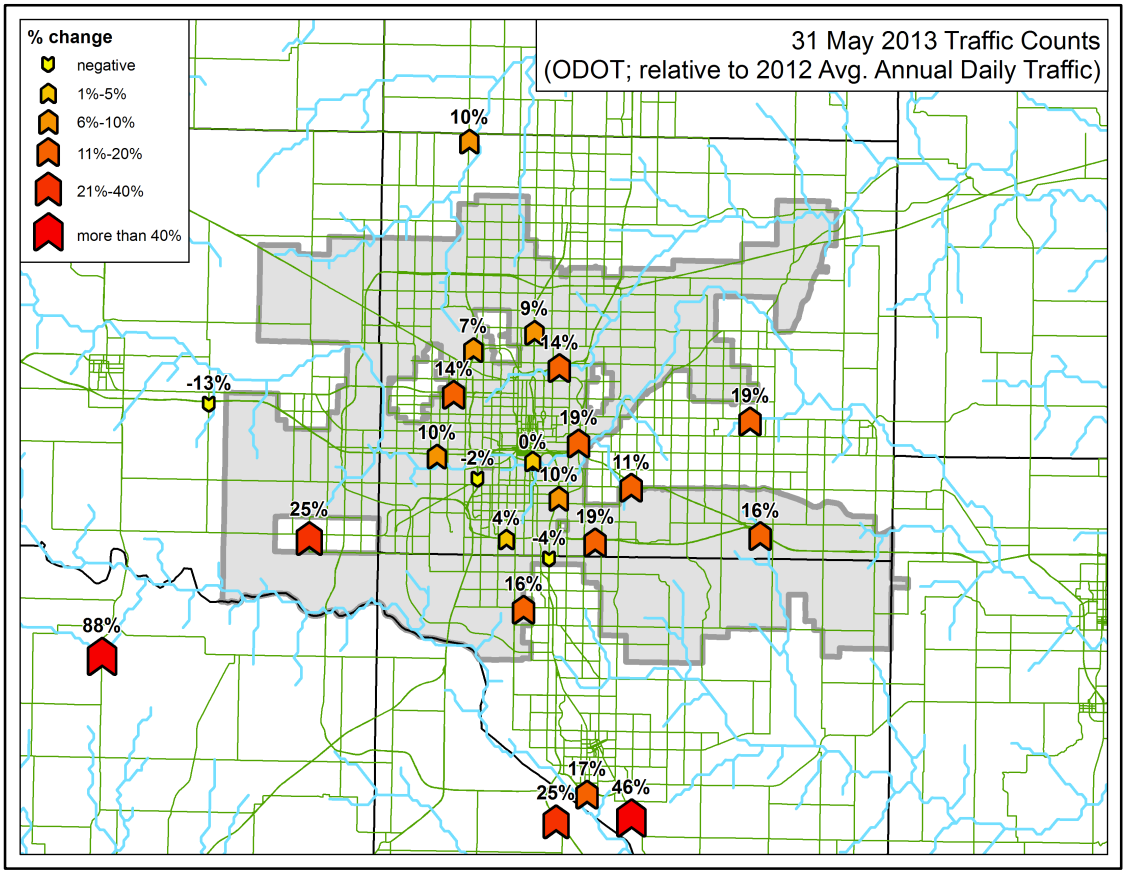

Figure 35. Percent change in daily traffic (compared to the previous year's average annual daily traffic) on May 31, 2013 at selected ODOT traffic monitoring stations in central OK

Results of Random Forest Predictions

The random forest (RF) model proposed in Chapter 3 was applied to GFS forecasts valid between 0000 UTC 24 May 2013 to 1800 UTC 1 June 2013. In this case, the RF model generally performed well, correctly identifying that there was a threat of receiving *Storm Data* reports of flash floods from OK to Missouri (MO), Illinois (IL), and Indiana (IN) on the night of May 31 and the morning of June 1. Figure 36 contains

the result of applying the RF model, fit to data drawn from the entire archive, to a 120-h (5-d) GFS forecast from 0600 UTC 27 May 2013, valid at 0600 UTC on 1 June 2013. In Figure 36, the RF probabilities have been calibrated using the power-law relationship derived from the entire archive of cases. Figure 36 (and 37, 38, 39, 40, 41, and 42) displays probabilities from 0-7% to improve the visualization of individual pixel probabilities. From the model calibrations discussed in Chapter 3, the maximum confidence the low-elevation RF can have in a flash flood occurring at a specific grid cell is 14%. Verifying *Storm Data* reports of flash floods are shown with gray dots.

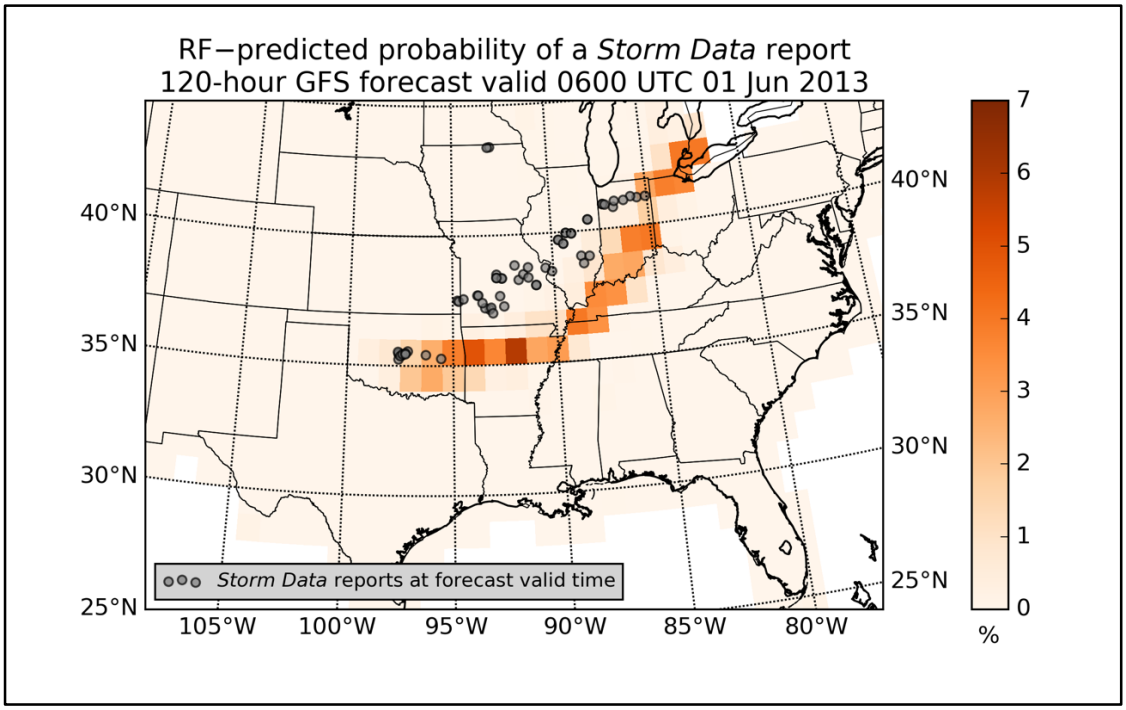

Figure 36. RF 120-h forecast probability of a report of a flash flood, valid 0600 UTC 1 June 2013

Although the forecast in Figure 36 missed the *Storm Data* report in northern Iowa (IA), it did correctly identify a line of potential flash floods extending from central OK to northern Ohio (OH) and southeastern Michigan. The 120-h forecast was slightly offset

to the south and east of the eventual confirmatory reports. The greatest probability of a *Storm Data* report was forecast in northern Arkansas (AR) and east central OK.

Figure 37 is a 48-h (2-d) GFS forecast from the same model initialization in Figure 36.

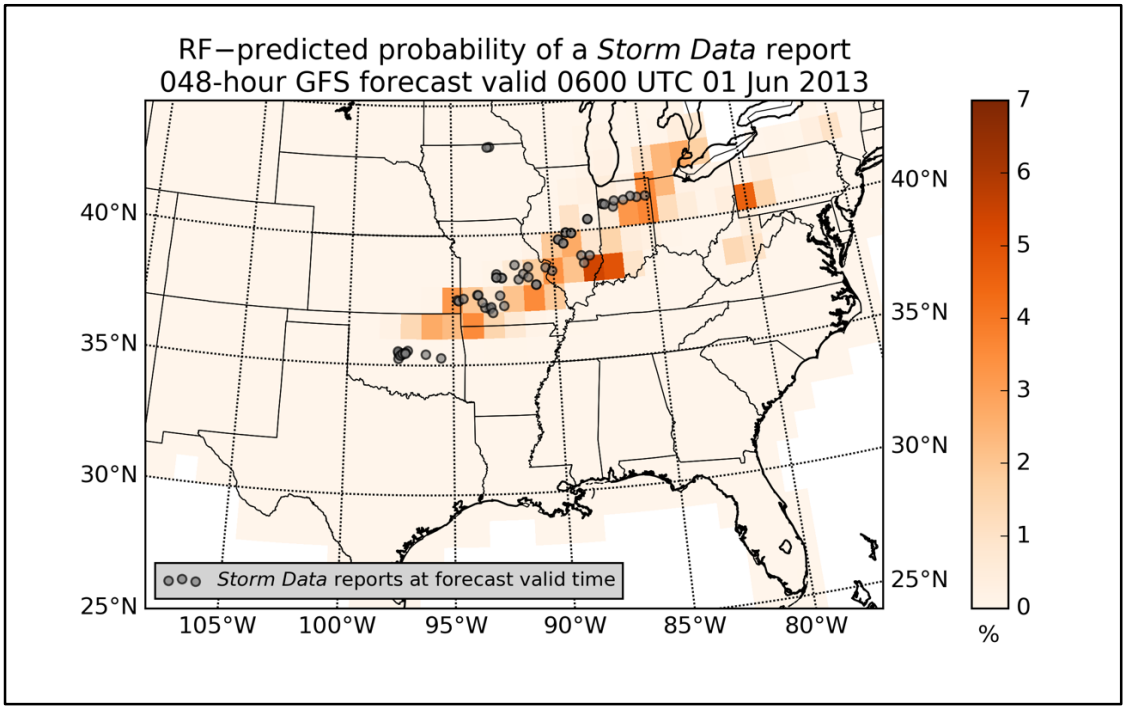

Figure 37. RF 48-h forecast probability of a report of a flash flood, valid 0600 UTC 1 June 2013

The 48-h forecast is generally better-aligned with the swathe of *Storm Data* reports across the Midwest and mid-Mississippi River valley. However, it does not identify central OK as a potential hotspot and also incorrectly predicts a moderately-high probability of a *Storm Data* report in southwestern Pennsylvania. The 12-h (0.5-d) forecast, shown in Figure 38, correctly keys in on central OK, and better-forecasts the reports in MO, IL, and IN. In general, the 12-h forecast is shifted slightly too far to the south and east, but performed fairly well overall. Although the probabilities resulting from the calibration process seem low, it is important to remember that the base rate of a

flash flood for the entire archive is 0.17%, so a probability of even 5% means a flash flood is nearly 30 times more likely than it would be on a "typical day".

May 2015: U.S. Southern Plains

In May 2015, heavy rain affected the Southern Plains, especially OK and Texas (TX). In OKC, as stated in the previous section, the all-time record for the wettest month was shattered by May 2015, as the city received 450% of its 1980-2010 normal May rainfall. Wichita Falls, TX received 543% of its normal May rainfall during May 2015. Major flood and flash flood impacts were felt in the states of Kansas (KS), AR, and Louisiana (LA) (Breslin 2015).

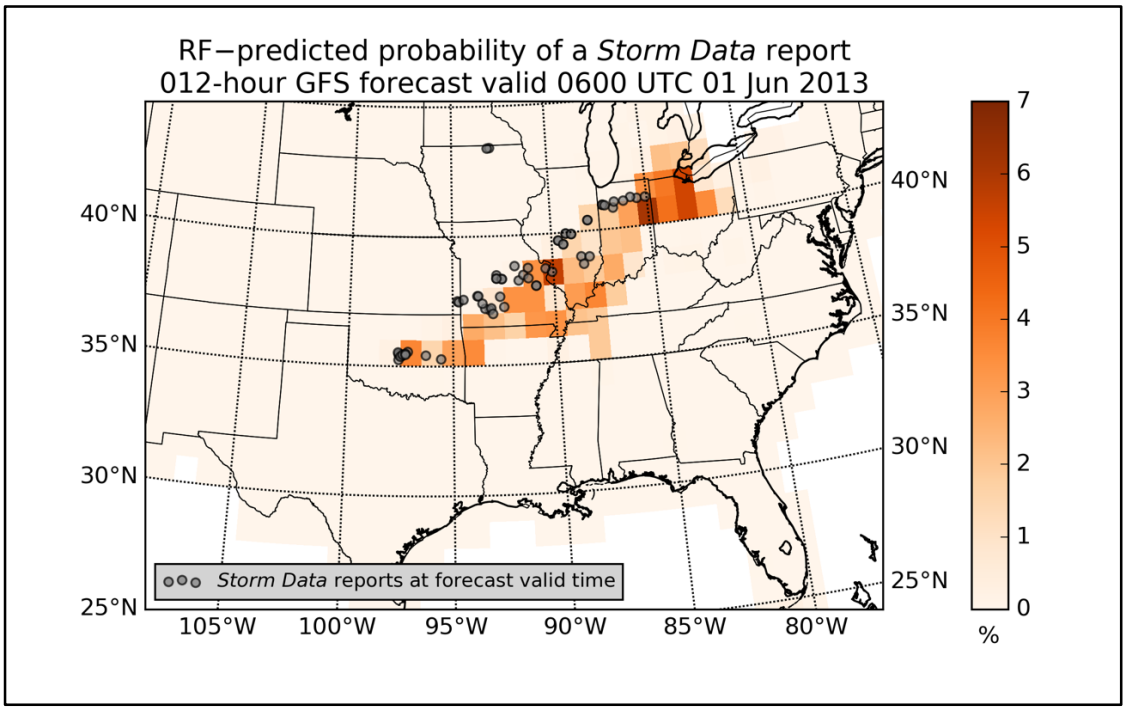

Figure 38. RF 12-h forecast probability of a report of a flash flood, valid 0600 UTC 1 June 2013

On May 18, 2015, highways in northeast LA were flooded, along with nearly a dozen homes. In northwest LA, roads were closed in several parishes, many homes were flooded, a child was killed, and another child and an adult were both injured when a car

was washed away by swift water. In TX, San Angelo Regional Airport was closed and cars were washed away by floodwaters. Major highways and city streets were closed in dozens of communities across eastern, southeastern, and south central TX. Multiple rescues and one automobile-related fatality were also reported in TX during the event (NCEI 2015).

Results of Random Forest Predictions

The proposed RF model was applied to GFS forecasts from 2300 UTC 11 May 2015 to 1200 UTC 18 May 2015. The 156-h (6.5-d) forecast, valid 1200 UTC 18 May 2015, is shown in Figure 39.

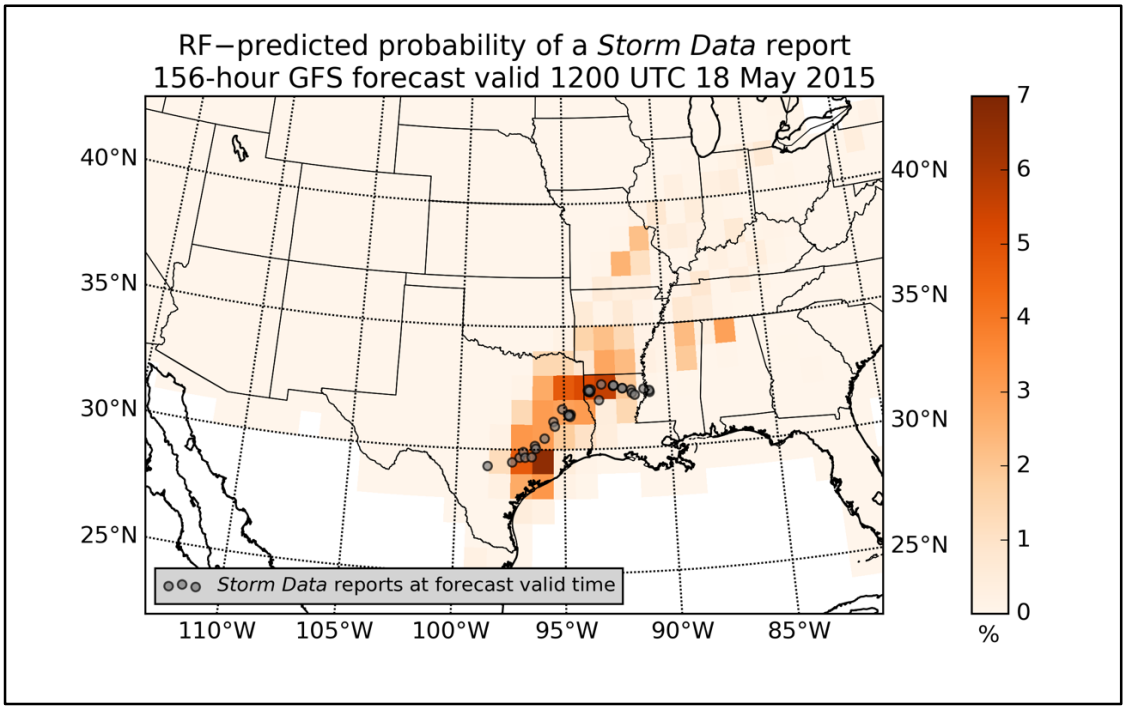

Figure 39. RF 156-h forecast probability of a report of a flash flood, valid 1200 UTC 18 May 2015

Although the RF product shows some false alarming in MO, southern AR, and northern Mississippi (MS) and Alabama, it correctly identifies the majority of the impacts of the eventual flash floods across northern LA and eastern and southern TX. Figure 40

shows the 60-h (2.5-d) forecast valid at the same time, while Figure 41 contains the corresponding 12-h forecast. In the 60-h forecast there are significant false alarms from AR all the way northeast to OH. (There were *Storm Data* reports of flash floods in AR, but they occurred three hours prior to the valid time of these forecasts.) In the 12-h forecast, these false alarms have been reduced in area and magnitude, but are still present in MS, Tennessee (TN), and Kentucky (KY).

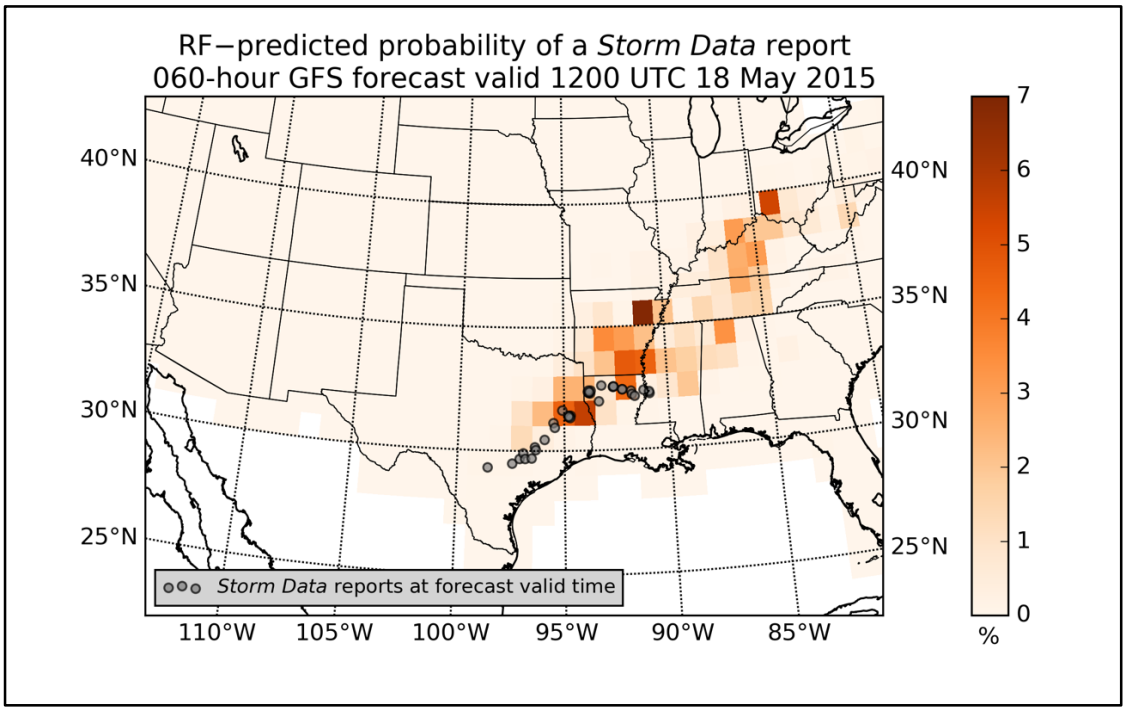

Figure 40. RF 60-h forecast probability of a report of a flash flood, valid 1200 UTC 18 May 2015

Hydrometeorological Testbed – Hydrology 2016 Experiment

The 2016 HMT-Hydro experiment ran from June 20, 2016 to July 15, 2016, and involved 16 NWS forecasters from WFOs and River Forecast Centers (RFCs) across the U.S. Forecasters were asked to use a series of experimental flash flood monitoring and forecasting tools to issue experimental FFAs and flash flood warnings (FFWs). For details on the monitoring tools available to the forecasters, see Gourley et al. (2016).

Martinaitis et al. (2016) explain the experimental setup used in the 2015 iteration of the HMT-Hydro Experiment, while Clark and Gourley (2015) explain and describe the 2014 iteration of what was, at that point, called the HWT-Hydro Experiment. HMT-Hydro 2016 was conducted in cooperation with the Flash Flood and Intense Rainfall Experiment (FFaIR) at WPC. Information about FFaIR's experimental setup was given in Barthold et al. (2015).

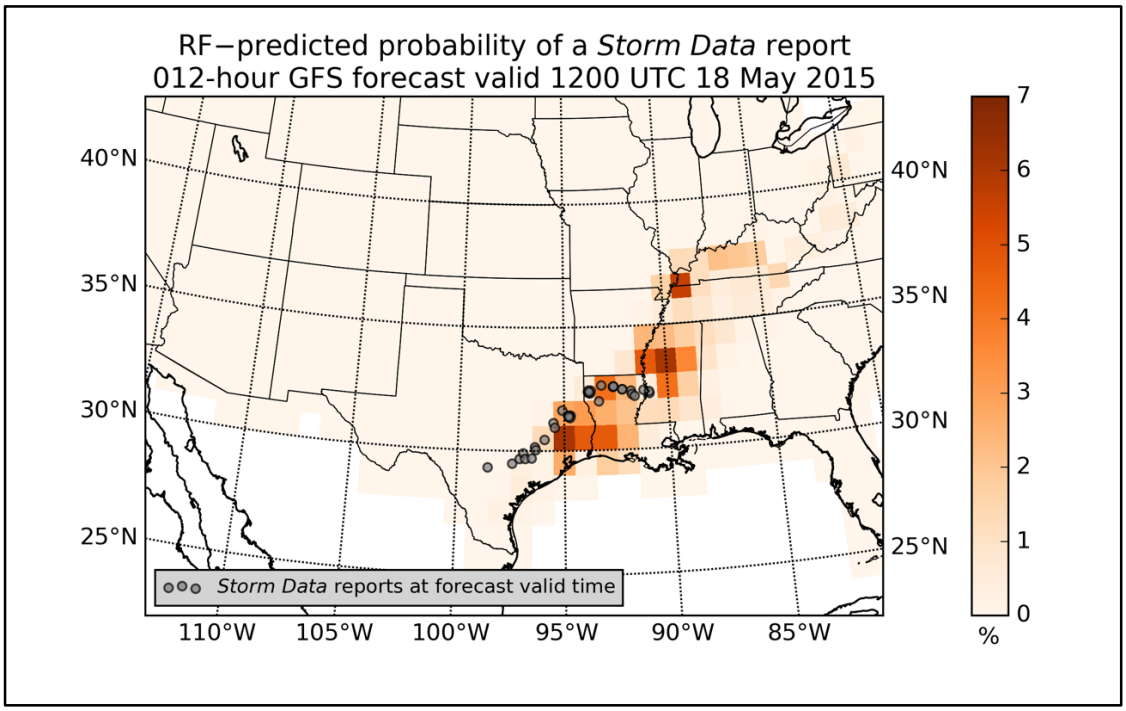

Figure 41. RF 12-h forecast probability of a report of a flash flood, valid 1200 UTC 18 May 2015

During HMT-Hydro 2016, forecasters had access to a real-time, quasi-operational version of the RF model described in and proposed by Chapters 3 and 4 of this dissertation. This RF was fit to the study period and each elevation region identified in Chapter 3. It included all candidate predictor variables from Chapter 3 except for those related to speed shear, mean layer wind, the K index, and model PW anomaly. Additionally, the training, testing, and validation datasets used to create this RF included

hourly linearly-interpolated grids of all GFS model fields and derived predictors; in the rest of this dissertation, RFs were grown from original GFS model data and derived predictors arising from original GFS model data, with no hourly linear interpolation. The HMT-Hydro 2016 RF was *fit* to GFS3 model data, at a 1.0-degree x 1.0-degree resolution, as in the rest of the dissertation, but the RF was *applied* to the newly-available 0.25-degree x 0.25-degree resolution GFS to generate probabilistic predictions of receiving a report of a flash flood in a given grid cell. Each day, the RF was applied to 6-, 12-, 18-, and 24-h forecasts from each 1200 UTC cycle of the GFS, which yielded probabilities valid at 1800 UTC on day 1, and 0000, 0600, and 1200 UTC on day 2.

Results of National Weather Service Forecaster Surveys

The experimental FFAs and FFWs resulting from each forecasting shift during the experiment were evaluated via a series of survey questions. Two questions asked in each forecast evaluation session concerned the performance of the RF predictions available to the forecasters. Figure 42 summarizes the results of these two questions. In the first question, each forecaster was asked, for each experimental forecast shift, to rate his or her agreement with the statement "The spatial accuracy of the GFS prediction probability forecast for the previous day was skillful," via a five-segment Likert scale (Likert 1932). If we assign a value of "1" to "Strongly Disagree" and "5" to "Strongly Agree", as shown in Figure 42, the average score on this question over the 12 experimental forecast shifts was 2.6, between "Neutral" and "Disagree". When asked to rate their degree of agreement with "The probability values of the GFS prediction probability forecast for the previous day were accurate," the average score was 2.8, also between "Neutral" and "Disagree".

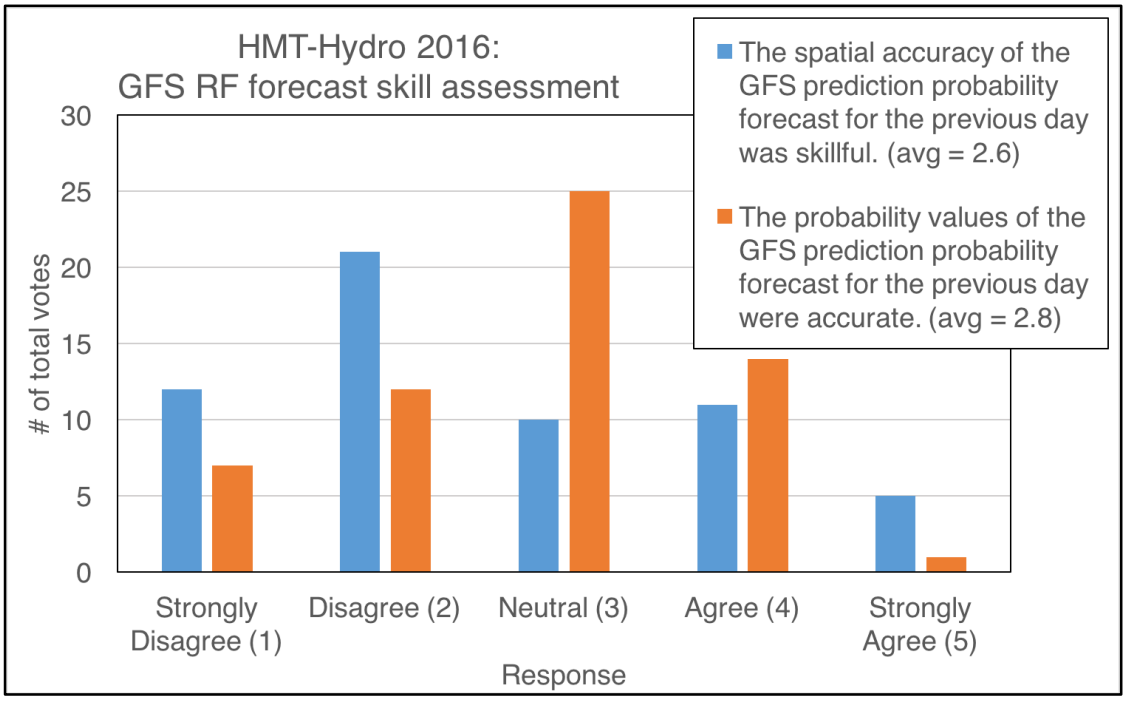

Figure 42. Results of 2016 HMT-Hydro survey questions on the use of GFS RF predictions of flash floods in a testbed environment

In association with each question, forecasters were allowed to provide comments recorded by the personnel facilitating each evaluation session. These comments are reproduced verbatim in Table 20. In general, the comments are focused on the limitations imposed upon the RF method by the use of GFS NWP output, which is generally better at forecast synoptic-scale conditions due to its relatively coarse resolution. Of particular interest are the comments of June 29, 2016 and June 30, 2016. On June 29, 2016, a forecaster noted that the RF tool would likely be of less utility in western U.S. flash flood forecasting, an assertion borne out by the results presented in Chapters 3 and 4 of this dissertation. On June 30, 2016 widespread flash floods were reported in the Las Vegas, Nevada area. All available operational NWP guidance was largely unsuccessful at forecasting environments and QPF supportive of an outbreak of flash floods, and the GFS was no exception to this rule.

Table 20. Comments made by forecasters regarding GFS RF flash flood predictions during the 2016 HMT-Hydro Experiment

Date	*Spatial coverage*	*Magnitude*
20-Jun-16	No data available	No data available
21-Jun-16	No comments	No comments
22-Jun-16	No comments	Would be better if values were higher
23-Jun-16	No comments	No comments
27-Jun-16	Difficult event to pinpoint. The line was pushing thin and fast…do not expect GFS to pick up on that.	Just seemed to be of no value here.
28-Jun-16	As a whole, it did fairly well. They were not high probability values, but it basically got the areas correct. One said he would look at it as an area of interest…situational awareness tool. It is a good way to highlight areas of interest a couple days out, but maybe not so great for 0-6 hr.	No comments
29-Jun-16	This product will not likely work out west because it is the GFS. One forecaster said neutral because there were low values and not much happened…so perhaps it gave some decent info.	Probabilities 0 to near 0 and not much happened.
30-Jun-16	This was a big event, and GFS totally missed it.	Ditto.
11-July-16	It was off spatially but in terms of the general region of the country it was okay.	The model is useful as a potential red flag even if its spatial accuracy is off. Can be good to use along with the GFS.
12-July-16		By product of the resolution that the model missed a storm event over a town. Race won't be held accountable for that :-)
13-July-16	It had an idea that this area would have something, it's just displaced. Majority says it missed a little bit.	There should have been something there bc it's saying zero and yet roads were washed away. Limitations of the GFS are well understood.
14-July-16	It hinted at something. It was a lot better than previous days.	General consensus is meh

However, on other days, the RF tool provided some value to the forecast process, including on June 28, 2016 and July 11, 2016. At the end of the experiment's first week, the group of forecasters chose to highlight the RF tool during a webinar designed to transmit their findings from the testbed experiment to an operational NWS audience. They felt that the RF tool was reasonably useful in synoptically-forced situations, but was unable to adequately resolve mesoscale and storm-scale features associated with more isolated flash floods.

Case Study: Fatal Flash Flood in West Virginia

On June 23, 2016, heavy rainfall and the consequent flash flood resulted in a tragic loss of life in West Virginia (WV). Twenty-four people lost their lives during the event, the deadliest flash flood in the U.S. since flash floods claimed 27 lives in TN, MS, and KY (Sterling et al. 2016). As a part of the FFaIR Experiment at WPC, HMT-Hydro 2016 forecasters were provided each day with national 1500 UTC Day 1 to 1200 UTC Day 2 outlooks identifying the probability of excessive rainfall in an area (defined as the probability of QPE exceeding flash flood guidance). Figure 43 contains this WPC outlook (valid from 1500 UTC 22 June 2016 to 1200 UTC 23 June 2016, shown as colored contours), the 24-h RF forecast probability of a *Storm Data* report of a flash flood (valid 1200 UTC 23 June 2016, shown as gridded data), and the associated confirmatory NWS local storm reports (LSRs) of flash floods (valid from 1200 UTC 22 June 2016 to 1200 UTC 23 June 2016).

In this example, the GFS RF probabilities correctly ignored the contoured areas over southern Colorado and northern IL and IN. However, both the human forecasters and the ML tool missed the LSR in northern IA. The ML correctly identified low probabilities associated with the LSR in northern KY, but overforecast probabilities in

Pennsylvania and North Carolina. Both methods correctly zeroed in on the greatest impacts, which occurred in OH and WV. Note that the WPC forecasters participating in the FFaIR Experiment did not have access to the GFS-derived RF probability product when drawing their forecast contours.

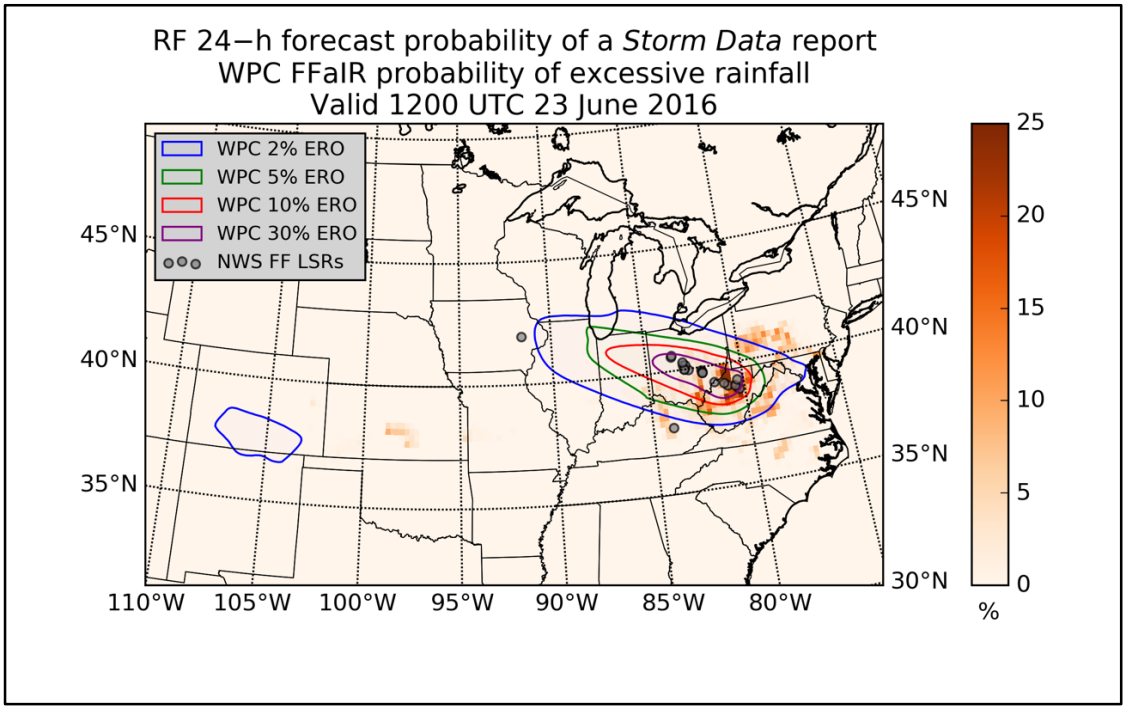

Figure 43. RF 24-h forecast probability of a report of a flash flood and experimental WPC probability of excessive rainfall, valid 1200 UTC 23 June 2016

Global Flash Flood Prediction

European Severe Storms Laboratory Report Archive

A database of flash floods across Europe is available from the European Severe Storms Laboratory, or ESSL. This European Severe Weather Database (ESWD) contains important weather hazards occurring across the European continent and in regions adjacent to the European continent (Dotzek et al. 2009). All ESWD reports tagged as flash floods were downloaded for the same period as the entire archive used in the rest of this dissertation. Figure 44 is a map of these reports (N = 14,013), restricted to those

that passed ESSL's "QC1" or "QC2" quality checks, which mean that the report was "confirmed by a reliable source" or that "extraordinary work has been performed to verify the validity of the all pieces of information given in a certain report", respectively. Using the procedures outlined in Chapter 3, these reports have been collocated with GFS model fields and predictors derived therefrom in space and time. Then, this European predictor-and-predictand matrix was used to grow an RF, but in this test, elevation was not accounted for in any way; in other words, only one forest is grown for all of Europe and its surrounding regions.

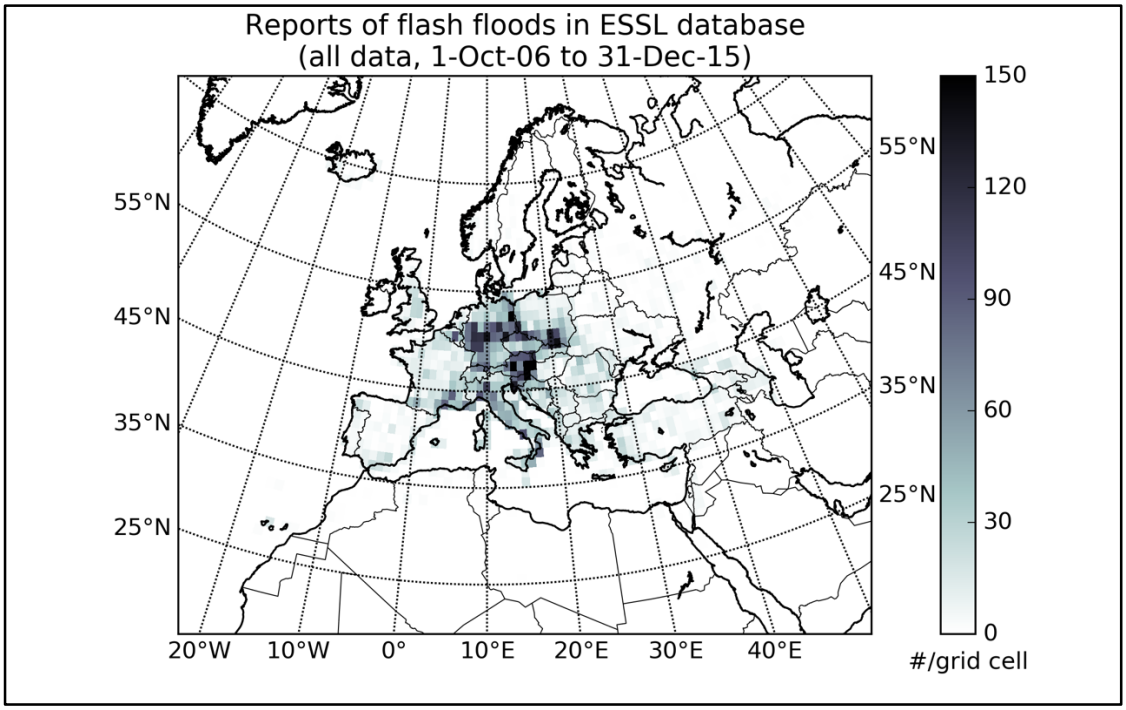

Figure 44. Number of ESSL reports of flash floods (N = 14,013) per grid cell over the entire archive

When applied to its independent validation datasets over 50 different sampling, fitting, and validation trials, this European RF had a Brier score between 0.104 and 0.088, comparable to that achieved with the U.S. low-elevation RF, which ranged from 0.090 to 0.099. The mean Brier score of the European RF was 0.093, while the mean Brier

score of the U.S. RF (over low elevation cases) was 0.094. Table 21 summarizes the most important variables in the European RF by their mean decrease in Gini impurity (MDG) scores over 50 trials. The most important variable, on average, is the K index, just as it is in the U.S. Specific humidity at low levels is quite important, as well, as is the precipitation rate, convective precipitation rate, best 4-layer LI, surface-based CAPE, and PW.

Table 21. Most important prediction variables by MDG for the European flash flood RF fitting process

Mean MDG rank	*Variable name*	*Range of MDG ranks*	*Std. dev.*	*Mean MDG score*	*Std. dev.*
1	k	1 to 6	0.9	0.062	0.008
2	4layer_li	1 to 7	1	0.053	0.01
3	925q	2 to 10	2	0.041	0.008
4	850q	2 to 12	2	0.040	0.008
5	2m_q	3 to 10	2	0.037	0.006
6	700q	2 to 11	3	0.037	0.007
7	sfccape	2 to 11	2	0.032	0.007
8	1013.25q	2 to 13	2	0.031	0.005
9	pw	4 to 13	2	0.028	0.005
10	cpreciprate	7 to 15	2	0.021	0.004

The consistency in these results across continents is promising for the development of additional RF models and the application of them to areas of the world presently underserved by the flash flood forecasting enterprise. Unfortunately, because flashiness is not defined over Europe, the U.S. and European RFs cannot be directly cross-validated against one another without modifications. Still, the commonalities in skill and MDG suggest an RF fit to one mid-latitude continent might be applicable to another mid-latitude continent.

Summary

This chapter presented a collection of small studies demonstrating that the RF technique can be applied to the flash flood forecasting problem in a quasi-operational

way. In a testbed context, NWS forecasters found output from the method useful in a handful of cases over a 3-week period during the summer of 2016. Additionally, in one highly impactful case from the testbed experiment, the RF predictions yielded a 24-h forecast that fairly closely tracked the equivalent excessive rainfall forecast issued by human forecasters. In one case from May 2013 and another from May 2015, the RF method was able to skillfully forecast the probability of a *Storm Data* report starting between five and seven days prior to the event. Finally, the method was successfully applied to the European continent, where it performed similarly to its U.S. application.

Chapter 6: Conclusions and Implications

A score or more of studies, especially over the last two years, have explored applications of machine learning (ML) methods and techniques to forecast problems in meteorology. Despite the recent wave of ML studies propagating throughout the literature, ML is not just a fad. Rather, it is the latest incarnation of a strain of thought that has long been fundamental to weather forecasting, the idea that the enterprise exists only because man and machine work in concert with one another.

Vilhelm Bjerknes warrants a significant share of the credit for expanding our understanding of atmospheric physics and laying down the foundational theories of how this understanding could apply to scientific weather forecasting. After Lewis Fry Richardson's first attempts to apply Bjerknes' ideas operationally were unsuccessful, it became clear that any weather forecasting beyond simple advection of observed conditions was going to require a massive paradigm shift of some sort (Lynch 2008). The invention of the electronic computer, it turns out, was that paradigm shift. The earliest of these massive machines, the most complex objects ever conceived of by the human race, were used for two primary purposes: to protect the interests of the West in the nascent Cold War and to achieve numerical weather prediction (NWP). John von Neumann, Jule Charney, and others began the long process of realizing Bjerknes' and Richardson's ideas in 1950 (Charney et al. 1950).

As computers quickly grew in power, weather forecasting grew up. But there was a backlash in the 1970s. NWP got better and scientists worried: were meteorologists forgetting how to be meteorologists? This "meteorological cancer" took hold quickly (Snellman 1979) as NWP continued to advance. Since that time, weather forecasts have

gotten better, and NWP has overwhelmingly been the driving force underlying that improvement. As a result, NWP has become a bigger and bigger part of the forecasting process. Of course, any operational forecaster would lament that NWP is not perfect – the meteorologist remains, and will continue to remain, a critical part of the forecasting process. The meteorologist uses her expertise and her experience to correct (or ignore) the NWP model output when appropriate. This correction process is complex; personal rules-of-thumb, on-the-job training, mentorship from experienced forecasters, and other factors are important in explaining how or why a weather forecaster adds significant value to NWP.

Early users of NWP quickly realized that this correction process could be augmented by statistical techniques. The first of these still makes up a core part of the National Weather Service's (NWS) operational capability: model output statistics, or MOS. MOS combines NWP output and outside pieces of information to correct known deficiencies in NWP that arise as a result of the numerical techniques used to solve the primitive equations (Glahn and Lowry 1972). ML algorithms can be used to do this same thing, but with vastly greater numbers of predictor variables and cases. MOS uses multiple linear regression to relate a set of predictor variables to a predictand. ML algorithms additionally account for non-linear relationships between predictor and predictand and for complex interactions between candidate predictor variables. In the 1990's, ML methods, particularly artificial neural networks, were used in an array of studies that essentially sought to improve upon or supplant MOS (e.g., Hall et al. 1999). Like MOS, these studies often focused upon a small area or a particular station and were based upon a generally small number of cases.

The current wave of ML-in-meteorology studies are novel in three important ways: 1) they often seek to address forecasting problems in more comprehensive and generalizable ways, 2) they use larger numbers of candidate predictors and a greater number of cases due to technological improvements over the intervening years, and 3) they frequently pursue predictions of ever rarer events. As yet, however, ML has not been applied to the flash flood forecast problem. The studies that make up this dissertation have laid out the theoretical foundations for this application, have presented some of the statistical properties of the forecasts generated from such application, have used this application to provide physical insights into how flash flood forecasts can be made and optimized, and finally, have given examples of ML methods in use in case study and research-to-operations contexts.

Caveats of Machine Learning and Automation

Despite their promise, ML methods are not a panacea, and there is a conceivable danger that too much in the way of statistical postprocessing will lead to overconfidence in automated methods and, thus, forecaster disengagement. Like any other statistical technique, ML methods work better when they are not asked to extrapolate over, for example, previously-unseen conditions. This concern can be somewhat mitigated via the use of very large datasets for training, validation, and testing, but even senior weather forecasters can be heard to admit that the atmosphere is an oft-surprising beast.

Specifically, ML classifiers like the random forest (RF) method used so prominently in these studies are generally quite robust – but not entirely immune – to common statistical complications like overfitting, collinearity, and random noise. Additionally, the RF probably cannot identify every possible nonlinear interaction

between candidate predictors. There is still a larger wealth of research in the literature (and a wealth of understanding in operational forecasting) than that which has been exploited in the present studies. This body of knowledge includes various plausible predictors that could be derived from base NWP model fields (or, in turn, from other derived predictors) but have not been adopted in this dissertation, including, for example, vertical integrations of some of my candidate predictors.

Using primarily NWP data to forecast flash floods is advantageous in many ways – it means the RF model is robust, from an operational point-of-view. In other words, the RF forecasting process does not depend on a range of disparate datasets, and thus, this process would not fail to run unless the Global Forecast System (GFS) itself were to fail. However, this brings to mind a major pitfall of using these data: if a particular GFS forecast is inadequate or unreliable, the RF probabilities derived therefrom will suffer the same fate. Another major related caveat involves resolution. The present study is designed to answer the question: "Is there enough available information in coarse-resolution NWP to statistically forecast flash floods?" However, because coarse-resolution NWP is unable to resolve many of the salient characteristics of flash flood environments, and because flash floods are generally rare events, the RF method is only able to reliability provide flash flood forecast probabilities between 0 and 14%.

Therefore, we must dispose of the hypotheses outlined in Chapter 1 as follows:

1. As shown at the end of Chapter 3, it is clear that the RF technique provides additional forecast skill over the imposition of simple thresholds upon single NWP model outputs associated with flash floods in the literature. It is also true

that ML techniques can be calibrated to yield reliable probabilistic forecasts of flash floods.

2. Secondly, Chapter 4 demonstrated that the RF technique provided insight into the atmospheric environments associated with flash floods. These insights match up with past studies in this area, though it is interesting to note that the RF method suggests a greater importance for PW over model PW anomaly, for instance, and that quantities like mean layer wind and speed shear were not frequently used by the RF in the production of flash flood forecasts. The RF's preference for raw model PW suggests that the division of the conterminous U.S. into three separate model regions was sufficient to account for regional differences in the PW required to induce a flash flood. The absence of speed shear and mean-layer winds from the RF importance analysis suggests that winds favorable for flash floods occur quite frequently when a flash flood is not observed. In other worse, they are useful parameters for characterizing flash flood environments but not for distinguishing between flash flood and non-flash flood environments. These new findings should be reassessed using observational data, not just NWP data, and if reconfirmed, made a part of the training and education of operational forecasters.
3. Finally, the RF technique was successfully applied to the flash flood forecasting problem in the Hazardous Weather Testbed during the summer of 2016. Although the aggregate impression of the NWS forecasters using the tool was slightly more negative than neutral, there were events in which the forecasters found the RF probabilities useful. Additionally, Chapter 5 showed the RF method to be

applicable outside the U.S. Finally, the method was shown to be useful in longer-range forecast contexts, of up to six days, in two historical events over the U.S.

Future Work

The present studies have demonstrated that NWP can be automatically used to forecast flash floods, and can do so more skillfully than via the use of just quantitative precipitation forecasts. However, there is a large amount of work to be done, and many questions raised by this dissertation to be answered. Most obviously, perhaps, is the question of how additional NWP models besides the GFS fit into this process. A multi-model ensemble of RF flash flood forecasts would help to capture the range of possible uncertainties inherent in a particular forecast. Along these same lines, this method could be applied to internal model ensembles, like the Global Ensemble Forecast System (GEFS), which is part of the GFS family of models. Other possible studies involve the inclusion of additional derived variables, particularly those vertically-integrated through different atmospheric layers.

Ideally, separate RFs should be fit to GFS model output by forecast hour, even though reasonably-good results can be achieved by applying a forest fit to GFS analyses directly to GFS forecasts, as shown for three cases in Chapter 5. Longer-range GFS forecasts are archived by the National Centers for Environmental Information but only in "cold storage", which makes obtaining long enough archives of them for research purposes a difficult, though not impossible, endeavor. More simply, other flavors of GFS output could also be tested more rigorously. The studies in this dissertation have primarily drawn upon GFS3 data, which are stored at 1.0-degree x 1.0-degree resolution. However, the results from the subjective evaluation of RF predictions available during

the 2016 Hydrometeorological Testbed are based upon 0.25-degree x 0.25-degree resolution GFS data, which have only been available since May 2016. These higher-resolution data likely allow the NWP to better (though not completely) resolve mesoscale atmospheric features favorable for the development of flash floods.

The main advantage of the GFS and GEFS family of NWP models is their global coverage and free availability. The results from the European continent presented in this dissertation are evidence that global flash flood prediction is feasible with the RF method. However, lack of observations of flash floods (and thus, training datasets) in most of the world outside Europe and the U.S. is a serious impediment to these efforts. Global versions of flash flood guidance (FFG) have been proposed and implemented (Georgakakos et al. 2013), but these suffer from the same limitations as the FFG system used operationally in the U.S. Direct statistical simulation of the flash flood threat from global NWP, as proposed in these studies, could serve as an important complement to or replacement for these FFG systems, if local or regional records of past flash floods can be obtained.

Most critically, though, this method must be applied to higher-resolution NWP guidance, including convection-allowing models, and high-resolution hydrologic model output if it is to be adopted for wider use in the U.S. NWS flash flood forecasting and alerting enterprise.

Final Thought

Based upon the results presented from the studies in this dissertation, ML forecasts of flash floods are plausible. Rare events are difficult to forecast whether statistical methods or human forecasters are doing the majority of the heavy lifting, and

the fact that we have come this far is promising. Whether used as a time-saver and situational awareness tool by human forecasters or as part of a quasi-global automated flash flood alert dashboard, the RF algorithm holds great potential to improve our ability to monitor, forecast, assess, and understand flash floods and the dangerous impacts they all-too-frequently bring to bear upon vulnerable communities around the world, year after year after year.

References

Ahijevych, D., J. Pinto, J. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecast.*, **31**, 581-599.

Antolik, M., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.*, **239**, 306-337.

Archer, K., and R. Kimes, 2007: Empirical characterization of random forest variable importance measures. *Comput. Stat. Data. An.*, **52**, 2249-2260.

Ashley, S., and W. Ashley, 2008: Flood Fatalities in the United States. *J. Appl. Meteorol. Clim.*, **47**, 805-818.

Barthold, F., T. Workoff, B. Cosgrove, J. Gourley, D. Novak, and K. Mahoney, 2015: Improving Flash Flood Forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bull Amer. Met. Soc.*, **96**, 1859-1866.

Bishop, C., 2007: *Pattern Recognition and Machine Learning*. Springer, 738 pp.

Bjerknes, V., 1904: Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteor. Zeit.*, **21**, 1-7. Translated by Y. Mintz, 1954: The problem of weather forecasting as a problem in mechanics and physics.

Braud, I., and Coauthors, 2014: Multi-scale hydrometeorological observation and modelling for flash flood understanding. *Hydrol. Earth Syst. Sc.*, **18**, 3733-3761.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984: *Classification and regression trees*. Chapman and Hall, 368 pp.

Breiman, L., 2001: Random Forests. *Mach. Learn.*, **45**, 5-31.

Breslin, S., 2015: Southern plains flooding: All-time rainfall record broken in Oklahoma City; evacuations underway in Oklahoma, Texas. *Weather.com*. Accessed 22 July 2016. [Available online at https://weather.com/storms/severe/news/southern-plains-flooding-texas-arkansas-oklahoma.]

Brice, T. and T. Hall, 2013a: Mixing Ratio. NOAA/NWS/WFO El Paso. Accessed 7 July 2016. [Available online at http://www.srh.noaa.gov/images/epz/wxcalc/mixingRatio.pdf.]

-----, 2013b: Wet-bulb Temperature and Dewpoint Temperature from Air Temperature and Relative Humidity. NOAA/NWS/WFO El Paso. Accessed 8 July 2016. [Available online at http://www.srh.noaa.gov/images/epz/wxcalc/wetBulbTdFromRh.pdf.]

Brier, G., 1950: Verification of forecasts expressed in terms of probabilities. *Month. Weather Rev.*, **78**, 1-3.

Brooks, H., and D. Stensrud, 2000: Climatology of Heavy Rain Events in the United States from Hourly Precipitation Observations. *Mon. Weather Rev.*, **128**, 1194-1201.

Carpenter, T., J. Sperfslage, K. Georgakakos, T. Sweeney, and D. Fread, 1999: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *J. Hydrol.*, **224**, 21-44.

Chappell, C., 1986: Quasi-stationary convective events. *Mesoscale Meteorology and Forecasting*, P. Ray, Ed., Amer. Meteor. Soc., 289-310.

Charney, J., R. Fjörtoft, R., and J. von Neumann, 1950: Numerical Integration of the Barotropic Vorticity Equation. *Tellus*, **2**, 237-254.

Clark, E., 2011: Weather Forecast Office Hydrologic Products Specification. National Weather Service Instruction 10-922, 84 pp. Accessed 3 July 2016. [Available online at http://www.nws.noaa.gov/directives/sym/pd01009022curr.pdf.]

Clark, R., and J. Gourley, 2015: The 2014 Multi-Radar/Multi-Sensor (MRMS) HWT-Hydro Testbed Experiment: Final Report. NOAA/OAR/NSSL. Accessed 22 July 2016. [Available online at http://blog.nssl.noaa.gov/flash/wp-content/uploads/sites/7/2014/06/hwt-hydro_final_report_2014.pdf.]

Clark, R., J. Gourley, Z. Flamig, Y. Hong, and E. Clark, 2014: CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products. *Weather Forecast.*, **29**, 377-392.

Clark, R, Z. Flamig, H. Vergara, Y. Hong, J. Gourley, D. Mandl, S. Frye, M. Handy, and M. Patterson, 2016: Hydrological Modeling and Capacity Building in the Republic of Namibia. *Bull. Amer. Met. Soc.* (in review).

Corcoran, J., J. Knight, A. Gallant, 2013: Influence of Multi-Source and Multi-Temporal Remotely Sensed and Ancillary Data on the Accuracy of Random Forest Classification of Wetlands in Northern Minnesota. *Remote Sens.*, **5**, 3212-3238.

Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.*, **20**, 273-297.

Cox, D., 1958: The regression analysis of binary sequences (with discussion). *J. Roy. Stat. Soc. B*, **20**, 215-242.

Crum, W., 2013: Drought puts Lake Hefner boating season in jeopardy. *The Oklahoman,* 5 February 2013. [Available online at http://newsok.com/article/3752345.]

Dotzek, N., P. Groenemeijer, B. Feuerstein, and A. Holzer, 2009: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmos. Res.*, **93**, 575-586.

Doocy, S., A. Daniels, S. Murray, and T. Kirsch, 2013: The Human Impact of Floods: a Historical Review of Events 1980-2009 and Systematic Literature Review. *PLOS Currents Disasters*.

Doswell, C., and D. Schultz, 2006: On the Use of Indices and Parameters in Forecasting. *Electronic. J. Severe Storms Meteorol.*, **1**, 1-14.

Doswell, C., H. Brooks, and R. Maddox, 1996: Flash Flood Forecasting: An Ingredients-Based Methodology. *Weather Forecast.*, **11**, 560-581.

Friedman, J., 2001: Greedy Function Approximation: A Gradien Boosting Machine. *Ann. Stat.*, **29**, 1189-1232.

Frisbie, P., 2006: The population bias of severe weather reports west of the Continental Divide. *Natl. Weather Dig.*, **30**, 11-16.

Gambill, D., W. Wall, A. Fulton, and H. Howard, 2016: Predicting USCS soil classification from soil property variables using Random Forest. *J. Terramechanics*, **65**, 85-92.

Ganguly, A., and R. Bras, 2003: Distributed Quantitative Precipitation Forecasting Using Information from Radar and Numerical Weather Prediction Models. *J. Hydrometeor.*, **4**, 1168-1180.

Gaume, E., and Coauthors, 2009: A compilation of data on European flash floods. *J. Hydrol.*, **367**, 70-78.

Georgakakos, K., R. Graham, R. Jubach, T. Modrick, E. Shamir, C. Spencer, and J. Sperfslage, 2013: Global Flash Flood Guidance System, Phase 1. Hydrologic Research Center Technical Report No 9, 134 pp. Accessed 14 July 2016. [Available online at http://www.hrc-lab.org/projects/projectpdfs/HRC%20Technical%20Report%20No%209.pdf.]

George, J., 1960: *Weather Forecasting for Aeronautics*. Academic Press, 673 pp.

Giordano, L., 1994: A fingertip guide to key upper air index values used in evaluating severe weather and flash flood potential. NOAA/NWS/WFO Pittsburgh. Accessed 6 July 2016. [Available online at http://www.weather.gov/pbz/svrffwpot.]

Glahn, H., and D. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasts. *J. Appl. Meteorol.*, **11**, 1203-1211.

Global Climate & Weather Modeling Branch, 2016: The Global Forecast System (GFS) – Global Spectral Model (GSM) (GSM Version 13.0.2). NOAA/NWS/NCEP/EMC. Accessed 5 July 2016. [Available online at http://www.emc.ncep.noaa.gov/GFS/doc.php.]

Gochis, D., W. Yu, and D. Yates, 2014: The WRF-Hydro model technical description and user's guide, version 2.0. NCAR Technical Document, 120 pp.

Gourley, J., J. Erlingis, Y. Hong, and E. Wells, 2012: Evaluation of Tools Used for Monitoring and Forecasting Flash Floods in the United States. *Weather Forecast.*, **27**, 158-173.

Gourley, J., and Coauthors, 2013: A unified flash flood database across the United States. *Bull. Amer. Met. Soc.*, **94**, 799-805.

-----, 2016: The Flooded Locations and Simulated Hydrographs (FLASH) project: improving the tools for flash flood prediction across the United States. *Bull. Amer. Met. Soc.* (accepted).

Guyon, I., 1997: A Scaling Law for the Validation-Set Training-Set Size Ratio. Unpublished Technical Report, AT&T Bell Laboratories. Accessed 7 July 2016. [Available online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf.]

Hall, T., H. Brooks, and C. Doswell, 1999: Precipitation Forecasting Using a Neural Network. *Weather Forecast.*, **14**, 338-345.

Hamill, T., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. Roy. Meteor. Soc.*, **132**, 2905-2923.

Hand, D., and K. Yu, 2001: Idiot's Bayes – Not So Stupid After All? *Int. Stat. Rev.*, **69**, 385-398.

Hardman, J., A. Paucar-Caceres, and A. Fielding, 2013: Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Syst. Res. Behav. Sci.*, **30**, 194-203.

Heung, B., C. Bulmer, and M. Schmidt, 2014: Predictive soil parent material mapping at a regional-scale: A Random Forest approach. *Geoderma*, **214-215**, 141-154.

Hogan, R., C. Ferro, I. Joliffe, and D. Stephenson, 2010: Equitability Revisited: Why the "Equitable Threat Score" Is Not Equitable. *Weather Forecast.*, **25**, 710-726.

Hutengs, C. and M. Vohland, 2016: Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sens. Environ.*, **178**, 127-141.

Ireland, G., M. Volpi, and G. Petropoulos, 2015: Examining the Capability of Supervised Machine Learning Classifiers in Extracting Flooded Areas from Landsat TM Imagery: A Case Study from a Mediterranean Flood. *Remote Sens.*, **7**, 3372-3399.

Jessup, S., and A. DeGaetano, 2008: A Statistcal Comparison of the Properties of Flash Flooding and Nonflooding Precipitation Events in Portions of New York and Pennsylvania. *Weather Forecast.*, **23**, 114-130.

Jones, E., E. Oliphant, P. Peterson, et al., 2001: SciPy: Open Source Scientific Tools for Python. Accessed 16 July 2016. [Available online at http://www.scipy.org.]

Junker, W., 2008: Heavy rainfall forecasting training manual: An ingredients based methodology for forecasting precipitation associated with MCS's. NOAA/NWS/NCEP/WPC. Accessed 7 July 2016. [Available online at http://www.wpc.ncep.noaa.gov/research/mcs_web_test_test_files/Page882.htm.]

Kohavi, R., and F. Provost, 1998: Glossary of Terms: Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Mach. Learn.*, **30**, 271-274.

Kursinski, E., D. Adams, and M. Leuthold, 2008: GPS Observations of Precipitable Water and Implications for the Predictability of Precipitation during the North American Monsoon. *CLIVAR Exchanges*, **45**, 13-21.

Lahouar, A. and J. Ben Hadj Slama, 2015: Day-ahead load forecast using random forest and expert input selection. *Energ. Convers. Manage.*, **103**, 1040-1051.

Likert, R., 1932: A technique for the measurement of attitudes. *Arch. Psychol.*, **140**, 1-55.

Lin, Y., and K. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc. Accessed 3 July 2016. [Available online at http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/refs/stage2-4.19hydro.pdf.]

Llasat, M., M. Llasat-Botija, A. Rodriguez, and S. Lindbergh, 2010: Flash floods in Catalonia: a recurrent situation. *Adv. Geosci.*, **26**, 105-111.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *J. Comput. Phys.*, **227**, 3431-3444.

MacAloney, B., 2016: Storm Data Preparation. National Weather Service Instruction 10-1605. 110 pp. Accessed 6 July 2016. [Available online at http://www.nws.noaa.gov/directives/sym/pd01016005curr.pdf.]

McClung, T., 2016: Technical Implementation Notice 16-11 Amended. NOAA/NWS/Headquarters. Accessed 5 July 2016. [Available online at http://www.nws.noaa.gov/os/notification/tin16-11gfs_gdasaaa.htm.]

McDonald, A., J. Lee, C. Schwarz, and T. Brown, 2014: Steering in a Random Forest: Ensemble Learning for Detecting Drowsiness-Related Lane Departures. *Hum. Factors*, **56**, 986-998.

MacKay, D., 2005: *Information Theory, Inference, and Learning Algorithms, Version 7.2*. Cambridge University Press, 628 pp.

Maddox, R., C. Chappell, and L. Hoxit, 1979: Synoptic and Meso-α Scale Aspects of Flash Flood Events. *Bull. Amer. Met. Soc.*, **60**, 115-123.

Manacos, P., and D. Schultz, 2005: The Use of Moisture Flux Convergence in Forecasting Convective Initiation: Historical and Operational Perspectives. *Weather Forecast.*, **20**, 351-366.

Manzato, A., 2008: A Note on the Maximum Peirce Skill Score. *Weather Forecast.*, **22** 1148-1154.

Marjerison, R., M. Walter, P. Sullivan, and S. Colucci, 2016: Does population affect the location of flash flood reports? *J. Appl. Meteorol. Clim.* (accepted).

Markowski, P., and Y. Richardson, 2006: On the Classification of Vertical Wind Shear as Directional Shear versus Speed Shear. *Weather Forecast.*, **21**, 242-247.

Martin, R., R. Aler, J. Valls, and I. Galvan, 2016: Machine learning techniques for daily solar energy prediction and interpolation using numerical weather models. *Concurr. Comp.-Pract. E.*, **28**, 1261-1274.

Martinaitis, S., and Coauthors, 2016: The HMT Multi-Radar Multi-Sensor Hydro Experiment. *Bull. Amer. Met. Soc.* (in review).

Mecikalski, J., J. Williams, C. Jewett, D. Ahijevych, A. LeRoy, and J. Walker, 2015: Probabilistic 0-1-h Convective Initiation Nowcasts that Combine Geostationary Satellite Observations and Numerical Weather Prediction Model Data. *J. Appl. Meteorol. Clim.*, **54**, 1039-1059.

Murphy, A., 1973: A New Vector Partition of the Probability Score. *J. Appl. Meteorol.,* **12**, 595-600.

Murray, F., 1967: On the computation of saturation vapour pressure. *J. Appl. Meteorol.*, **6**, 203-204.

National Weather Service, 2009: NWS Glossary. NOAA/NWS. Accessed 2 July 2016. [Available online at http://w1.weather.gov/glossary/.]

Nayak, M., and S. Ghosh, 2013: Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. *Theor. Appl. Climatol.*, **114**, 583-603.

NCEI, 2015: Storm Data: May 2015. *Storm Data*, **57**, 5. NOAA/NESDIS/NCEI.

NCEP Central Operations, 2016a: GFS documentation. NOAA/NWS/NCEP/NCO. Accessed 5 July 2016. [Available online at http://nomads.ncep.noaa.gov/txt_descriptions/GFS_doc.shtml.]

-----, 2016c: NCEP Products Inventory: Global Products. NOAA/NWS/NCEP/NCO. Accessed 5 July 2016. [Available online at http://www.nco.ncep.noaa.gov/pmb/products/gfs/.]

-----, 2016c: NCO PMB – Upcoming Changes. NOAA/NWS/NCEP/NCO. Accessed 5 July 2016. [Available online at http://www.nco.ncep.noaa.gov/pmb/changes/.]

Niculescu-Mizil, A., and R. Caruana, 2005: Predicting Good Probabilities with Supervised Learning. *ICML '05 Proceedings of the 22nd Int. Conf. on Mach. Learn*, **1**, 625-632.

Pearson, K., 1895: Notes on regression and inheritance in the case of two parents. *P. R. Soc. London*, **58**, 240-242.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825-2830.

Perica, S., and E. Foufoula-Georgiou, 1996: Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. *J. Geophys. Res. D*, **101**, 26,347-26,361.

Quinlan, J., 1986: Induction of Decision Trees. *Mach. Learn.*, **1**, 81-106.

Richardson, B., J. Hansford, and K. Falk, 2011: Common environmental parameters associated with heavy precipitation and flash flood events over southwest Arkansas, east Texas, and north Louisiana. *25th Conf. on Hydrology*, Seattle, WA, Amer. Meteor. Soc. Accessed 6 July 2016. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Manuscript/Paper182313/Richardson.B_WFOSHV_AMS_Hydro_Paper.pdf.]

Rojas, R., 1996: *Neural Networks – A Systematic Introduction*. Springer, 502 pp.

Root, B., P. Knight, G. Young, S. Greybush, R. Grumm, R. Holmes, and J. Ross, 2007: A fingerprinting technique for major weather events. *J. Appl. Meteorol. Clim.*, **46**, 1053-1066.

Saharia, M., P.-E. Kirstetter, H. Vergara, J. Gourley, Y. Hong, and M. Giroud, 2016: Mapping Flash Flood Severity in the United States. *J. Hydrometeor.* (in review).

Sawilowky, S., F, 2002: Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When $\sigma_1^2 \neq \sigma_2^2$. *J. Mod. App. Stat. Meth.*, **1**, 461-472.

Scott, D. W., 1992: *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 317 pp.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecast.*, **5**, 570-575.

Schroeder, A., J. Basara, J. Sheperd, and S. Nelson, 2016a: Insights into Atmospheric Contributors to Urban Flash Flooding across the United States Using an Analysis of Rawinsonde Data and Associated Calculated Parameters. *J. Appl. Meteorol. Clim.,* **55**, 313-323.

Schroeder, A., and Coauthors, 2016b: The development of a flash flood severity index. *J. Hydrol.* (in press).

Smith, G., 2003: Flash flood potential: Determining the hydrologic response of FFMP basins to heavy rain by analyzing their physiographic characteristics. NOAA/NWS/Colorado Basin River Forecast Center. [Available online at http://www.cbrfc.noaa.gov/papers/ffp_wpap.pdf.]

Snellman, L., 1979: Operational Forecasting Using Automated Guidance. *Bull. Amer. Met. Soc.*, **58**, 1036-1044.

-----, 1982: Impact of AFOS on operational forecasting. *9th Conf. on Weather Forecasting and Analysis*, Seattle, WA, Amer. Meteor. Soc., 13-16.

Spearman, C., 1904: The proof and measurement of association between two things. *Amer. J. Psychol.*, **15**, 72-101.

Standage, T., 2016: Special Report: Artificial Intelligence. *The Economist*, 25 June 2016, 3-16.

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecast.*, **15**, 221-232.

Sterling, J., R. Ellis, F. Fawzy, and J. Imam, 2016: West Virginia flooding leaves at least 24 dead. *CNN*. Accessed 22 July 2016. [Available online at http://www.cnn.com/2016/06/25/us/west-virginia-flooding-deaths/.]

Suffern, P., K. Harding, V. Brown, J. Keeney, K. Stammer, and J. Stefkovich, 2014: May 2013 Oklahoma Tornadoes and Flash Flooding. NOAA/NWS Service Assessment. Accessed 13 July 2016. [Available online at http://www.nws.noaa.gov/om/assessments/pdfs/13oklahoma_tornadoes.pdf.]

Sun, H., D. Gui, B. Yan, Y. Liu, W. Liao, Y. Zhu, and C. Lu, 2016: Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energ. Convers. Manage.*, **119**, 121-129.

Tan, P.-N., M. Steinbach, and V. Kumar, 2005: *Introduction to Data Mining*. Pearson, 769 pp.

Tatsumi, K., Y. Yamashiki, M. Torres, and C. Taipe, 2015: Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. *Comput. Electron. Agr.*, **115**, 171-179.

Tinkham W., A. Smith, H.-P. Marshall, T. Link, M. Falkowski, and A. Winstral, 2014: Quantifying spatial distribution of snow depth errors from LiDAR using Random Forest. *Remote Sens. Environ.*, **141**, 105-115.

Touw, W., J. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. van Hijum, 2012: Data mining in the Life Sciences with Random Forest: a walk in the park of lost in the jungle? *Brief. Bioinform.*, **14**, 315-326.

Trafalis, T., I Adrianto, M. Richman, and S. Lakshmivarahan, 2014: Machine-learning classifiers for imbalanced tornado data. *Comput. Manag. Sci.*, **11**, 403-418.

Veneris, Y., 1990: Modeling the transition from the Industrial to the Informational Revolution. *Environ. Plann. A.*, **22**, 399-416.

Yan, D., and K. de Beurs, 2016: Mapping the distributions of C_3 and C_4 grasses in the mixed-grass prairies of southwest Oklahoma using the Random Forest classification algorithm. *Int. J. Appl. Earth Obs.*, **47**, 125-138.

Waldstriecher, J., A guide to utilizing moisture flux convergence as a predictor of convection. *Natl. Weather Dig.*, **14**, 20-35.

Wang, Z., C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, 2015: Flood hazard risk assessment model based on random forest. *J. Hydrol.*, **527**, 1130-1141.

Weiner, N., 1961: *Cybernetics, 2nd Edition*. MIT Press, 212 pp.

Woodward, D., 2010: Chapter 15 - Time of Concentration, *Part 630 Hydrology, National Engineering Handbook*. U. S. Dept. of Agriculture, 15-1 – 15-15.

Williams, J., 2014: Using random forests to diagnose aircraft turbulence. *Mach. Learn.*, **95**, 51-70.

Yu, R., Y. Yang, L. Yang, G. Han, and O. Move, 2016: RAQ—A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors*, **16**, 86.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Amer. Met. Soc.*, **97**, 621-638.