

Chapter 1

Introduction

The weather community sails on a turbulent sea of noisy data. The accuracy of weather forecasts is driven by two main factors: the quality of the observational data and model guidance available to forecasters and the ability of the forecasters to properly interpret that information (Stewart 2001). The effect of improving guidance quality is evident in tornado warning verification trends (Fig. 1.1) with the biggest recent increases in probability of detection occurring with the deployment of the NEXRAD network in the early 1990s (Brooks 2004). As numerical weather prediction models increase in resolution and morph from deterministic runs to ensembles, forecasters will face an ever increasing number of possibilities for what may happen. Radar and satellite observations are also increasing in temporal and spatial resolution, revealing new weather features but also requiring more time to analyze. Smart phones, crowdsourcing, and personal weather stations are conveying all kinds of potentially relevant data that has shown the ability to improve forecasts (Mass and Madaus 2014), but they also bring a lot of noise due to poor calibration, siting, and human biases. Even with all this change, weather forecaster displays and methods have not been evolving fast enough to keep pace with the explosion of data. Poor

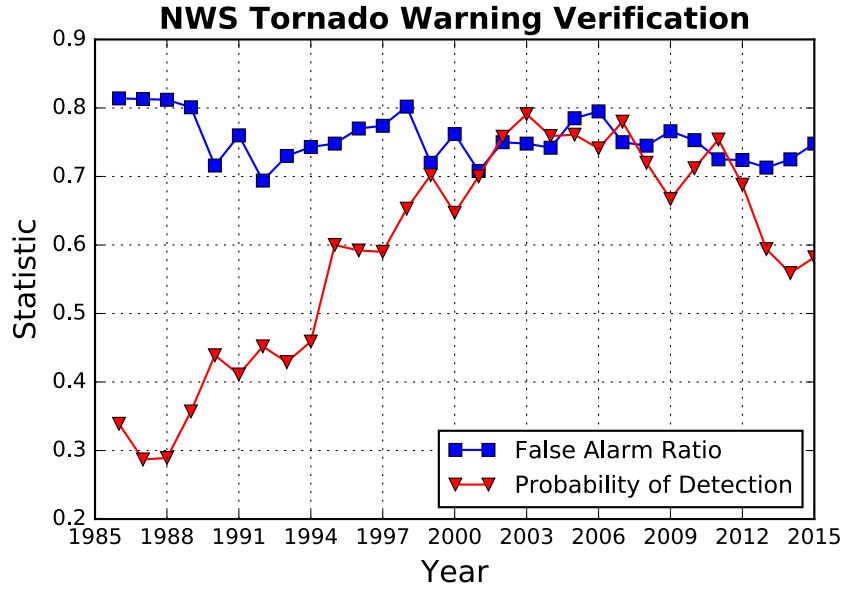


Figure 1.1: Yearly National Weather Service tornado warning verification statistics for the US. Source: NOAA.

integration of new data sources with forecasters could result in static or decreasing performance along with greater incentive for more forecaster duties to be replaced with automation (Snellman 1977).

The heuristics and biases present in human cognition have a limiting effect on the ability of using more information to improve forecasts (Stewart et al. 1992; Doswell III 2004). While providing forecasters more information does have a slightly positive impact on accuracy, it can lead to negative impacts on reliability as forecasters become overconfident in their decision (Stewart et al. 1992, 1997). Through education, training, and experience, forecasters develop a set of rules and heuristics for evaluating weather guidance and generating forecasts (Doswell III 2004). Forecasters may not benefit from additional guidance if it does not match with their conceptual model of the situation (representativeness bias), if it is a tool or situation outside of the previous experience of

the forecaster (availability bias), or if the subsequent information differs significantly from initial guidance (anchoring) (Tversky and Kahneman 1974; Doswell III 2004). The anchoring effect can lead to confirmation bias when forecasters accept additional information that supports their initial instinct while rejecting anything contrary. In order to minimize the effects of these biases, the guidance used should be as relevant, reliable, and unbiased as possible (Stewart 2001).

Statistically-corrected consensus ensemble prediction systems would then be the logical choice for initial guidance, but forecasters have expressed distrust of ensembles and statistical correction methods (Novak et al. 2008). Instead they prefer to use deterministic numerical weather prediction (NWP) models or a subset of individual ensemble members to predict high impact events. Consensus ensemble guidance in the form of ensemble mean or probability tends to smooth gradients and decrease the amplitude of extremes. Deterministic guidance is more physically consistent and can show extreme events, but it may overstate the likelihood of an extreme event occurring, especially if forecasters are biased toward looking at the extreme solutions. Existing linear bias-correction techniques like Model Output Statistics (Glahn and Lowry 1972) can correct for some issues with ensembles but have a limited ability to improve the representation of extreme events. Forecasters have expressed that they are more likely to adopt a new technique if it demonstrates a significant improvement over existing approaches and gives more direct guidance for the phenomenon being forecast (Morss and Ralph 2007).

Techniques from the data science community have the potential to address the needs of high-impact weather event forecasters by integrating the large amounts of information available into reliable hazard forecasts. Advances in computing and storage have allowed us to amass vast archives of data and

process it in real-time. Image processing techniques can filter gridded data to identify salient features, which can make the analysis of data over long periods of time easier and more consistent (Lakshmanan and Smith 2009). Machine learning methods can generate highly accurate predictive models from complex, multidimensional datasets by discovering underlying structures in the data with less reliance on theoretical assumptions about its origins (Breiman 2001b). The best machine learning methods balance predictive accuracy with robustness to noisy data and provide some level of interpretability. Ensemble decision tree methods, including random forest (Breiman 2001a) and gradient boosted regression (Friedman 2001), use interpretable base models while providing consistently high performance across many datasets (Caruana and Niculescu-Mizil 2006; McGovern et al. 2015). To produce these significant performance gains, however, machine learning models require one crucial item: a large set of forecast data paired with observations. High-impact weather events tend to be both extreme and rare, so a large amount of data is generally required to capture them well. Physical understanding is also helpful in constraining the data sources needed for prediction and ensuring that the predictions are physically consistent.

How much of a role should automated guidance play in the forecast process? For forecasts of standard weather variables, such as temperature and precipitation, the National Weather Service (NWS) currently operates with a human-in-the-loop paradigm in which forecasters subjectively blend and adjust multiple sources of guidance to create a final forecast. The NWS approach is very time and labor intensive, and now performs significantly worse than downscaled, bias-corrected ensemble forecasts for temperature and precipitation (Novak et al. 2014). Official NWS track forecasts of hurricanes, a major form of high impact weather, also perform worse than weighted ensemble consensus forecasts

(Cangialosi and Franklin 2015). There are also issues with spatial discontinuities in forecasts and warnings between the domains of different forecast offices (Gilbert et al. 2015). Many private weather firms, including the Weather Company, operate in a human-over-the-loop paradigm in which an optimal blend of bias-corrected model output is generated as needed by users, and human forecasters can adjust blending weights to account for observed short-term biases or data quality issues (Williams et al. 2016). This approach scales easily and only requires a small team of meteorologists to oversee a mostly automated system. The downside of a heavily automated approach is that forecasters may become disengaged from the forecast process (Pliske et al. 2004) and struggle to take appropriate corrective action when automation fails (Skitka et al. 1999; Pagano et al. 2016).

Given the large stakes associated with high impact weather, humans should remain a key part of the forecasting process, but their role and workflow should be designed to keep them engaged in the forecast process and contribute their time to areas where automation struggles (Pagano et al. 2016). Automation should be integrated in such a way to be complementary to forecaster skills and not be antagonistic. Forecasters should be able to interact with different sources of guidance as needed to evaluate the quality of their representation of the atmosphere (Doswell III 1992). Forecasters should be able to identify which scenarios presented by the automated system are more realistic and deemphasize those that are not in the final product. More time should be spent on constructing narratives and recommendations for end users instead of on forecast entry, although being more removed from the forecast generation process may make in-depth communication more difficult (Pagano et al. 2016). Visual analytics systems that enable realtime interactive with forecast data could help

with this. Forecasters should not be evaluated only on their skill against the automated guidance but on how effectively they communicate their forecasts to the public.

In this dissertation, I developed data-science-based approaches to improve predictions of two high-impact weather domains: hail and solar irradiance. While the two areas may seem to share little in common at first glance, both phenomena produce large economic impacts on a frequent basis. Hail causes billions of dollars in property and crop damage in the United States each year (Changnon 2009) and is a major yearly liability for insurance companies (Brown et al. 2015) with \$850 million in average annual claims. Urban sprawl and population growth in large cities such as Dallas/Fort Worth, St. Louis, Chicago, and Denver have made large amounts of property damage from hail events more likely (Rosencrants and Ashley 2015). Solar energy is a rapidly growing source of electricity whose variability needs to be predicted accurately, so electricity loads can be properly balanced and operational costs can be minimized. While solar irradiance itself does not cause disaster, poor forecasts and underestimation of the uncertainty could lead to major monetary losses for electric utilities and energy trading firms, and brownouts and blackouts due to an inadequate electricity supply could occur.

Both hail and solar irradiance exhibit non-Gaussian distributions in their range of values (Fig. 1.2). Hail size follows a gamma distribution, and clearness index, a scaled measure of observed versus idealized irradiance, exhibits a beta distribution (Falls 1974) with a larger peak near the maximum (Jurado et al. 1995). Large hail occurs rarely at a given location but happens almost daily somewhere in the contiguous US during the months from April through

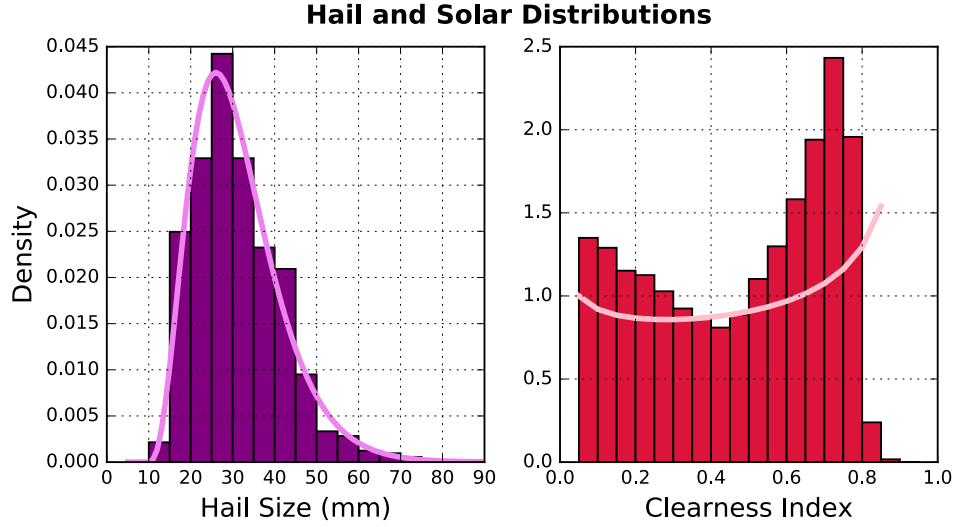


Figure 1.2: Observed distributions of hail sizes and clearness index during May and June 2015. A gamma distribution is fitted to the hail size values, and a beta distribution is fitted to the clearness index values.

July. While partly cloudy skies are rarer than mostly clear or cloudy days, they still occur fairly frequently at most locations. Both are rare events but common enough that collecting forecasts and observations over the period of a few months and a wide area can capture much of the variability associated with each phenomenon. This level of relative rareness makes these phenomena more amenable to prediction using statistical and machine learning methods.

The primary hypothesis of this dissertation is that properly configured decision tree ensemble machine learning models produce day-ahead predictions of hail size and solar irradiance that show significantly more skill than raw NWP model output, physics-based diagnostic models, and linear regression. In Chapter 3, I evaluate machine learning regression models that directly predict maximum hail size from convection-allowing model output against a physics-based diagnostic method and determine whether any of the approaches can produce

both low size errors and identify extreme hail events accurately. Utilizing the experiences from Chapter 3, in Chapter 4 I develop machine learning models with constraints in their pre-processing and training procedures aimed at producing more accurate and physically-realistic hail size forecasts. The machine learning hail forecasts are evaluated against other physics-based methods to determine how well each method detects hail events and how likely false alarms are for a given probability of detection. In Chapter 5, I create different solar irradiance forecasting systems with varied machine learning model configurations. I evaluate the prediction errors to determine which models and configurations have a significant impact on performance. I also stratify errors by forecast value and site to identify the major sources of error in the predictions.