The Pennsylvania State University
The Graduate School
College of Earth and Mineral Sciences

FIELD CAMPAIGN DECISION MAKING IN ATMOSPHERIC SCIENCE USING AN AUTOMATED DECISION ALGORITHM

A Dissertation in
Meteorology
by
Christopher Hanlon

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2015

The dissertation of Christopher Hanlon was reviewed and approved* by the following:

George S. Young
Professor of Meteorology
Dissertation Advisor, Co-Chair of Committee

Johannes Verlinde
Professor of Meteorology
Co-Chair of Committee

Martin Tingley
Assistant Professor of Meteorology

Paul Griffin
Professor of Industrial Engineering

David Stensrud
Professor of Meteorology
Department Head

*Signatures are on file in the Graduate School.

Abstract

An automated decision algorithm was developed for resource deployment decisions under weather uncertainty for an atmospheric science field campaign. Scientists on the Deep Convective Clouds and Chemistry (DC3) field campaign were tasked with using aircraft to gather in-situ measurements of isolated deep convection over three separate study regions in the United States during spring-summer 2012. The DC3 campaign was budgeted a finite number of flight hours with which to sample convection and was faced with a fixed start date and end date for the field campaign, forcing them to make difficult decisions each day about whether to fly their aircraft or whether to save their flight hours for a more promising future day. To guide decision recommendations, a quantitative definition of atmospheric conditions denoting a "successful" flight and a function defining field campaign utility as a function of "successful" flights were developed through communication with DC3 principal investigators. Utility-maximizing automated decision recommendations were generated using a dynamic programming-based decision algorithm with automated forecasts of the likelihood of "successful" conditions generated by a system employing a logistic regression with parameters tuned by a genetic algorithm. The forecasts generated by the automated forecasting system showed better skill than those produced concurrently by human forecasters, and the decisions generated by the automated decision algorithm would have improved field campaign utility relative to the decisions made by human decision-makers.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My adviser, George Young, was the perfect thesis adviser for me. George was quick to help with any problems I encountered and continues to be a seemingly endless fountain of wisdom. George is the multi-disciplinary thinker that I hope to be someday. My writing ability is an order of magnitude better than it was when I started graduate school, thanks to George's constructive criticism. Hans Verlinde was my undergraduate thesis adviser and a co-adviser throughout graduate school. Hans had a great knack for being blunt when I needed someone to be blunt and being compassionate when I needed someone to be compassionate. The input from the other committee members, Paul Griffin and Martin Tingley, has made this dissertation much stronger than it would have been without them.

Arthur Small taught me to think big about research problems and to assume that any problem can be solved. I strive for the same tenacity and creativity. Arthur's intuition about decision-making under weather uncertainty is uncanny and I'd like to think some of that has rubbed off on me over the years.

Thanks to the DC3 field campaign, and especially PIs Mary Barth, Bill Brune, Chris Cantrell, and Steve Rutledge, for letting me provide decision recommendations and bother them in the middle of their very busy days with questions about utility functions and hurdle probabilities.

My parents, Charles and Nancy Hanlon, have challenged and supported me through my academic career from pre-school through graduate school.

Finally, thank you to Kelsey, whose role in my life has changed from girlfriend to fiancée to wife during the writing of this dissertation. She has sacrificed countless nights and weekends so that I could stay home with my dissertation. Without her love, this dissertation never would have gotten finished.

I acknowledge financial support from National Science Foundation Atmospheric and Geospace Science Grant AGS-1063692.

Chapter 1
Introduction

1.1 Introduction

Risk managers in a wide variety of fields are concerned with decision making under weather uncertainty. My research has delved into one particular type of decision making under uncertainty: the resource deployment decisions faced by investigators on field campaigns in the atmospheric sciences. Atmospheric field experiments typically are tasked with collecting data for specific goals under budget and time constraints. For example, a typical field experiment might be assigned a budget of flight hours and a fixed number of days to sample the atmosphere under specific, low-probability, imperfectly predictable conditions. On each day of the field experiment, investigators must decide whether to spend some flight hours, conditional on a skilled-but-uncertain forecast of the likelihood of suitable data-collection conditions. The optimal decision is the one that maximizes, in expectation, the scientific value obtained using the scarce flight hours.

The scientific value of the field campaign is sensitive to the decision process used to deploy resources. A sub-optimal decision process may yield a significantly smaller amount of usable data, limiting the value of the field campaign. Most field campaigns in atmospheric sciences use some combination of weather forecasting heuristics and the judgment of human experts to make resource deployment decisions. An alternate method employing automated probabilistic weather forecasts of specific conditions and an automated decision recommendation algorithm was first implemented by Small et al. [2011] in a retrospective analysis of the RACORO campaign [Vogelmann et al., 2012]. The Small et al. method was adapted for the SPartICus campaign

[Mace et al., 2009] by Hanlon et al. [2013]. Most recently, an automated decision recommendation algorithm was developed for the DC3 campaign [Barth et al., 2012] by Hanlon et al. [2014b, chapter 3 of this dissertation]. For all three field campaigns, the automated decision recommendation algorithm outperformed the human forecasters, demonstrating the value of the methodology.

1.2 The general problem

The general problem addressed by the research in this area is that of launching aircraft under atmospheric uncertainty such that the scientific value of the obtained data is maximized in expectation. In one common form, field campaigns are tasked with maximizing the number of flights launched under specific atmospheric conditions. In particular, three characteristics are common to many field campaigns of this form: (1) the conditions of interest are imperfectly predictable on decision-relevant timescales, (2) aircraft flight hours are scarce, and (3) the calendar length of the field campaign is limited. Characteristic (1) dictates that at the time the fly/no-fly decision is made, the likelihood of suitable data-collection conditions can be forecast with some skill, but is not precisely known. If forecasts were no more skillful than climatology, no decision process would perform better than random guessing. If forecasts were able to perfectly distinguish good and bad days, the decision process would be trivial, and flights should be launched only on good days. Characteristic (2) demands flight hours be distributed carefully. If flight hours were not scarce, the decision process would be trivial, and flights should be launched every day. Characteristic (3) demands that field campaigns sometimes fly on relatively low-probability days. If the length of the field campaign were not limited, the decision process would be trivial, and flights should only be launched when conditions appear perfect.

While the three characteristics are strict, many field campaigns meet all three criteria. Such field campaigns face a non-trivial decision problem. On each day, given an uncertain forecast probability of good conditions, investigators must assess whether some of the flight hours ought to be deployed. Decisions should be made such that the scientific value of the field campaign, in terms of its portfolio of launched flights, is maximized. In general, resources should only be expended if the expected value associated with the use of those resources today exceeds the

expected value associated with the use of those resources on some future day.

This complicated decision problem should depend on several variables. Most obviously, the daily fly/no-fly decision depends on the forecast probability of suitable conditions: if suitable conditions are more likely, investigators should be more inclined to fly. Perhaps less obviously, the expected value of saving flight hours for a later day is a function of the climatologically expected conditions on future days, the number of days remaining in the field campaign, and the number of flights remaining in the field campaign. If there are 10 days and 9 flights remaining in the field campaign, the optimal decision process will be less picky about expending flight hours than if there are 10 days and 1 flight remaining in the field campaign. An automated system may be better suited to resolve this multi-dimensionality than a human decision-maker. The automated decision algorithm offers a way to optimally make complex decisions in the field campaign environment, allowing for more value to be extracted from a given field campaign.

1.3 A review of decision-making in meteorology and probabilistic weather forecasting

1.3.1 Background: decision-making in meteorology

Researchers have long observed that the societal value of weather forecasts is inseparable from those forecasts' application to decision problems. A perfect forecasting system applied to a problem where even a perfect forecast cannot be used to improve outcomes provides no value, whereas a relatively weak forecasting system may provide some value if applied properly to a decision problem. Therefore, when assessing a forecast system, evaluating the decisions made is a more complete measure of value than evaluating the forecasts. This dissertation is presented as a study in decision-making under weather uncertainty; one that includes tools used for weather forecasting but whose ultimate value lies in their application to decision-making. Therefore, while forecast systems often are assessed by their forecast skill, in this study, forecast evaluation is viewed as secondary to decision evaluation.

One early study on decision-making under weather uncertainty employed a straightforward approach using a 2x2 income matrix [Savage, 1951]. Imagine a

	Rain	No rain
Umbrella	4	5
No umbrella	-10	10

Table 1.1. An example of an income matrix, showing a user's preferences as a function of decision (carry an umbrella or do not carry an umbrella) vs. outcome (rain or no rain). The table is a modified version of a table from Savage [1951].

forecasting problem with two possible outcomes and a decision-maker with two possible choices. In the example used by Savage, the two outcomes are "rain" or "no rain" and the two choices are "carry an umbrella" or "don't carry an umbrella". Based on her own cost-benefit analysis, the forecast user (i.e., the decision-maker) should define a utility value associated with each combination of outcome and choice, where utility is simply defined as how much something is valued [Berger, 1993]. This set of conditional utility values constitutes a utility function. The utility function should reflect the forecast user's own preferences. For example, perhaps carrying an umbrella on a non-rainy day is somewhat bothersome, but not carrying an umbrella on a rainy day is a disaster; the utility function should account for this difference. Savage proposed organizing the end-user's preferences into an income matrix (Table 1.1). Savage observes that if probabilistic forecasts of the states are not available, the income matrix can be used to calculate a minimax solution, defined as the set of decisions that minimizes the maximum possible loss. However, Savage also notes that if the probability of rain is known, then it is trivial to calculate the expected-value maximizing decision from the income matrix. For meteorologists interested in making decisions under weather uncertainty, then, the goal is to develop well-calibrated forecast probabilities that make this two-state decision problem trivial. The Savage two-outcome, two-decision income matrix can be expanded to an arbitrary number of outcomes and decisions without loss of generality.

From the relatively generalizable form of Savage's 2x2 income matrix, several other researchers incorporated decision theory into forecast systems using weather forecasts as tools that optimize utility functions. Gringorten [1950] considered the decision problem faced on a foggy day by an airport dispatcher, who has to weigh the prospect of the fog lifting against the costs of keeping on duty a flight crew who are not allowed to fly. Perhaps if a forecaster predicted a 30% chance of fog lifting, the dispatcher would keep a flight crew on duty, but if there were only

a 15% chance of fog lifting, she would dismiss the flight crew. The dispatcher's preferences suggest there is some critical probability between 15% and 30% at which the dispatcher changes her decision. Gringorten argued that this decision is best handled if the forecaster indicates whether the chance of the fog lifting is significantly above or below the dispatcher's critical probability rather than providing a deterministic weather forecast. The critical probability was calculated as a function the profit resulting from a completed flight and the loss resulting from keeping the crew on duty. Gringorten's critical probability, therefore, implicitly connected a forecast probability to a utility function. Gleeson [1960] expanded the application of decision theory to forecast problems with a non-dichotomous outcome space, showing that for certain utility functions, the optimal strategy requires elements of game theory, where the decision maker and "nature", or the atmosphere, can be viewed as competitors in a game of strategy. Glahn [1964] predicted cloud ceiling height categories at an airport using contingency tables, with combinations of meteorological predictors chosen to maximize utility rather than to maximize the accuracy of the forecast probability. Either implicitly or explicitly, each of these methods used weather forecasts as tools that optimized a utility function, where utility optimization did not necessarily overlap with forecast skill optimization.

The research on decision theory in meteorology demonstrated that despite much research focus on improving weather forecasts, the ultimate economic value to society of weather forecasts lies in their improvement of weather *decisions*. Thompson [1962] examined a dichotomous case, where a decision-maker faced with an economic loss caused by adverse weather is allowed two possible decisions: "protect" or "do not protect". For this general decision problem, Thompson separated the potential value of incremental scientific advances (i.e. better forecasts) from the potential value of incremental operational improvements (i.e. better decisions). Thompson argued that under the then-current level of weather forecast skill, a marginal improvement in decision-making would provide economic gains of the same order of magnitude as a marginal improvement in forecast skill.

In keeping with the findings of Thompson [1962], throughout this dissertation care is taken to emphasize that the most relevant measure of forecast quality is the value of the resultant decisions, not the accuracy of the forecasts themselves. Pesaran and Skouras [2002] commend meteorology as a field where forecast systems

are sometimes scored based on utility rather than statistical skill, noting that in most fields forecast evaluation is done using purely statistical measures, separate from decision evaluation. Pesaran and Skouras cite several reasons for the widespread separation between forecast evaluation and decision evaluation in most fields: (1) forecasters and decision-makers often are separate groups of people, and so the forecasters don't know the decision-makers' utility functions, (2) decision space is highly information-intensive compared to forecast space, (3) it is has not been clear until recently that there is a significant difference between forecast evaluation and decision evaluation, and (4) decision evaluation is more technically difficult than forecast evaluation. Because statistical measures of forecast verification such as the Brier skill score are still widely accepted in meteorology, those are provided here, but alongside a decision-relevant forecast verification scoring procedure.

1.3.2 Background: probabilistic weather forecasting

Weather forecasts can be categorized as either *deterministic* or *probabilistic*. A deterministic forecast is a single point estimate of some weather conditions (e.g. "high temperature of 84 degrees"), while a probabilistic forecast consists of either a probability distribution (e.g. "high temperature of 84 degrees +/- 3 degrees") or a single probability (e.g. "40% chance of measurable precipitation"). Deterministic forecasts are popular in meteorology, due in part to their ease of interpretation. However, because probabilistic weather forecasts include a measure of forecast uncertainty, they convey more information to the optimizing decision-maker than deterministic weather forecasts. An example from Thompson [1950] illustrates the value that probabilistic forecasts provide. Thompson framed a hypothetical problem faced by a contractor in Los Angeles using a two-state income matrix similar to the one used by Savage [1951]. The hypothetical contractor has to decide each night whether to take measures to protect poured concrete overnight (at a small cost) or risk overnight rainfall ruining the concrete (at a large cost). The optimal decision is to protect the concrete when the probability of rain exceeds the ratio of the protection cost to the cost of ruined concrete. Thompson showed for one winter's worth of decisions that while the decision-maker's costs were lower when she was provided with deterministic forecasts than with no forecasts, her costs would have been markedly lower with probabilistic forecasts. For a proper

decision-making algorithm that accounts for forecast uncertainty, probabilistic forecasts are preferred to deterministic forecasts.

The most widely researched application of probabilistic weather forecasting is ensemble weather forecasting, a Monte Carlo modification [Leith, 1974] to traditional numerical weather prediction. Ensemble forecasting is primarily an attempt to account for the uncertainty in forecast model inputs [Epstein, 1969]. While the output of a standard deterministic forecast model is one estimate of the physical state of the atmosphere for each forecast time, the output of an ensemble forecast model is a set of such estimates, typically each generated using slightly different initial conditions [Gneiting and Raftery, 2005]. For many forecast applications, ensemble forecast models measure forecast uncertainty better than deterministic models, and so they naturally yield more accurate probabilistic forecasts, but care still must be taken when using ensemble forecasts for probabilistic forecasting. It is unfortunately a widespread practice in meteorology to conflate "the number of ensemble members in which conditions X occur" with "the probability of conditions X occurring". Such unmodified use of ensemble forecasts as a probabilistic forecasting system is naive; as Hamill and Colucci [1997] demonstrated, ensembles tend to be underdispersed, and so probabilistic forecasts from ensembles need to be calibrated. Several methods of post-processing ensemble forecasts have yielded skilled and well-calibrated probabilistic forecasting systems, such as the Bayesian methods employed by Raftery et al. [2005] and Katz and Ehrendorfer [2006].

While research in probabilistic weather forecasting has largely focused on ensemble forecasting techniques, the needs of the DC3 campaign required probabilistic forecasting from a deterministic forecast model. The decision was made during the development stage of the decision algorithm that to maximize credibility with DC3 forecasters, it would be best to use the NCAR WRF [Weisman et al., 2008] as the basis of the DC3 decision algorithm, because the model was so heavily used by DC3 forecasters in the pre-experiment planning stage. The NCAR WRF is a non-ensemble forecast model. While the use of a non-ensemble forecast model posed a challenge for probabilistic forecasting, ultimately DC3 principal investigators seemed to give the automated decision algorithm more credibility because its probabilistic forecasts qualitatively matched the convection forecasts from the model.

Research on probabilistic weather forecasting from deterministic forecast models

is sparser than research on ensemble forecasts. Klein et al. [1959] developed a technique commonly called the "perfect prog" method, using linear regression to establish a statistical relationship between current weather observations as predictors and future weather observations as predictands. A similar linear regression technique is used in model output statistics [MOS, Glahn and Lowry 1972], but instead of using current observations as predictors, MOS uses model forecasts as predictors. The set of variables forecast by MOS includes some probabilistic variables, including probability of precipitation. MOS continues to be the industry standard in operational objective weather forecasting, despite evidence that superior methods of objective weather forecasting are available. Applequist et al. [2002] tested the sensitivity of MOS-inspired methodology to the choice of predictive model, finding that logistic regression significantly out-performed linear regression, while Marzban et al. [2006] and Vislocky and Young [1989] have suggested that a combination of the "perfect prog" methodology and the MOS methodology is more skillful than either method alone. However, because MOS forecasts are used operationally by National Weather Service forecasters, stability and speed of computation are of utmost importance, limiting the ability for the NWS to expand MOS methods beyond linear regression.

Aside from MOS, much of the non-ensemble work in probabilistic forecasting has been aimed at generating a probability distribution of a physical quantity such as precipitation. Krzysztofowicz [1999] proposed a Bayesian forecast system that forecast hydrologic variables by separating input uncertainty from hydrologic uncertainty, yielding well-calibrated revisions of prior climatological distributions. Theis et al. [2005], proposing an intentionally simple, low-budget model post-processor, found skill in point precipitation forecasting from a deterministic model could be improved by using the area coverage of modeled precipitation around a point as the probability of precipitation at that point, even without calibrating the model forecasts to observations. Sobash et al. [2011] converted deterministic forecasts of convection into skillful probabilistic forecasts of severe weather using the spatial relationship between modeled updraft helicity maxima and observed storm reports. Marsh et al. [2012] used a kernel density function to calibrate deterministic forecasts of rare convective events, defined as those with quantitative precipitation forecasts (QPFs) above some threshold. Marsh's method converted a deterministic forecast of 6-hour QPF from a convection-allowing model into a

probability of exceeding some amount of precipitation at a given point.

Each of these probabilistic forecasting methods from deterministic models allows for the estimation of a probability of a physical variable exceeding a certain threshold at a point. DC3 operational needs, however, demanded a forecast of the probability of some conditions defined over an area. The only research to my knowledge that estimates areal probability is Young et al. [2015], which built linear and logistic regression models using MOS point probabilities of precipitation to predict the regional chance of occurrence of precipitation. The lack of research in this area forced the development of the new methods presented in this dissertation.

1.4 Other field campaigns

The first implementation of the automated decision algorithm method of field campaign resource deployment was a retrospective application of the method to the Routine Atmospheric Research Measurement (ARM) Aerial facility (AAF) Clouds with Low Optical Water Depths (CLOWD) Optical Radiative Observations (RACORO) campaign [Vogelmann et al., 2012] during 2009 [Small et al., 2011]. RACORO investigators sought to sample boundary layer clouds above the ARM Southern Great Plains (SGP) site using aircraft. Small et al. used model-forecast relative humidity (RH) profiles as the sole predictor of the presence of boundary layer clouds. Using model forecasts of RH as training data and ARM cloud fraction data [Xie et al., 2012] as verification data, an automated forecasting system used self-organizing maps, a neural network dimension reduction technique, to generate the probability of good conditions, conditional on some model forecast.

Given the historical distribution of forecasts the forecasting system would have generated, the fly/no-fly decision was guided by dynamic programming [Bellman, 1957]. Via backward induction, dynamic programming allows for the calculation of the "value" at each node of a tree, where each node represents each possible field campaign state, in terms of number of flights and number of days in the budget. After calculating the values of each node of such a tree, one can calculate the minimum probability of good conditions necessary to justify use of a flight, defined as the "hurdle probability". The fly/no-fly decision is made by comparing the forecast probability of good conditions to the hurdle probability. Retrospectively, the decision algorithm used for the RACORO campaign could have increased the

field campaign's data yield by 21%.

The SPartICus campaign during spring 2010 sought sampling of cirrus clouds with no boundary layer clouds below them. An automated decision algorithm analogous to the one used by Small et al. was developed for the SPartICus campaign [Hanlon et al. 2013]. Recommendations from this decision algorithm were provided operationally during the field campaign but were not used by SPartICus investigators. If the recommendations generated by the automated decision algorithm had been followed, the field campaign data yield would have been improved by 11% while reducing the length of the field campaign by 21%.

1.5 Deep Clouds and Convective Chemistry (DC3) campaign

The Deep Clouds and Convective Chemistry (DC3) field campaign during spring 2012 introduced a related but more complicated decision problem. DC3 investigators sought to sample "isolated, deep convection" in three study regions: northeast Colorado, northern Alabama, and a larger region covering central Oklahoma and northwest Texas. The structure of DC3 was similar to the RACORO and SPartICus campaigns: investigators sought to sample the atmosphere under specific conditions using aircraft, the number of available flights and days were constrained, and decisions had to be made under weather uncertainty.

Following the successful application of automated decision recommendation algorithms to the decision problems faced by investigators on the RACORO and SPartICus campaigns, we sought to construct a similar algorithm for DC3 decision support. The algorithm would provide a decision recommendation during the morning planning meeting each day during the field campaign. The recommendation, valid the same day, would consist of one of four options: "do not fly", "fly to Colorado", "fly to Alabama", or "fly to Oklahoma/Texas". The decision recommended would be the one that maximized in expectation the value of the end-of-season "portfolio" of data obtained by aircraft.

The DC3 campaign presented two challenges beyond those presented by the RACORO and SPartICus campaigns. First, the phenomenon of interest was isolated thunderstorms, offering a more difficult forecasting problem than those

faced by RACORO and SPartICus. In order to achieve DC3's goals while keeping investigators safe, the sampled thunderstorms had to be distinct from other storms, large enough to cause significant convective transport, long-lived enough to produce a large anvil, and small enough to safely sample. The quantitative definition of "good" conditions was thus more complicated than those employed for the earlier field campaigns, demanding the implementation of a more sophisticated forecasting system. Second, DC3 investigators sought to sample thunderstorms in three different regions, each with different climatology and different underlying science questions. The launch of a successful flight to one region was not interchangeable in value with one to another region. With respect to the three different study regions, investigators had two potentially competing objectives: maximize the total amount of data obtained and maintain balance in the data "portfolio" across the three regions. The automated decision recommendation needed to represent the preferences of investigators and estimate their willingness to substitute successful flights between the three regions. For example, where $\langle s_1, s_2, s_3 \rangle$ represent the number of successful flights in three arbitrary regions, we knew that $\langle 3, 3, 3 \rangle$ successful flights across the three regions was a more valuable portfolio than $\langle 4, 3, 2 \rangle$, but we had to quantify this difference in value in order to capture the preferences of DC3 investigators.

Aside from the trickier forecast problem and the multiple objectives, the essential tradeoff faced by DC3 investigators each day is the same as the one faced by RACORO and SPartICus investigators: flight hours can either be deployed today or saved for some future day. To address this problem using an automated system, three major sub-systems are required. An automated forecasting system is needed to generate a well-calibrated estimated probability of suitable conditions each day. A stochastic model of climatology is needed to estimate the prospects on future days beyond the forecast horizon. Finally, an optimizing decision module is needed to compare the expected value of a flight today with the expected value of a flight sometime in the future, estimating the optimal decision.

Chapter 2
Probabilistic forecasting for isolated thunderstorms using a genetic algorithm: the DC3 campaign

Part of the challenge in field campaign decision making is the challenge of forecasting whatever phenomenon scientists are interested in studying. For the DC3 campaign, the challenge was to forecast isolated, deep convection. In developing a forecasting system suitable for integration with an automated decision algorithm, the forecasting system had to be well-calibrated and probabilistic: rather than issue a yes/no forecast of suitable flight conditions, the system issued a probability corresponding to the likelihood of suitable flight conditions occurring. The methodology and results from the DC3 automated forecasting system are presented in this section, along with a discussion of some of the challenges, both those specific to this atmospheric forecasting problem and those broadly relevant to any automated forecasting system. This chapter appeared in the Journal of Geophysical Research: Atmospheres with the title *Probabilistic forecasting for isolated thunderstorms using a genetic algorithm: the DC3 campaign* and authors Christopher J. Hanlon, George S. Young, Johannes Verlinde, Arthur A. Small, and Satyajit Bose.

Abstract

Researchers on the Deep Convective Clouds and Chemistry (DC3) field campaign in summer 2012 sought airborne *in-situ* measurements of isolated thunderstorms in three different study regions: northeast Colorado, north Alabama, and a larger

region extending from central Oklahoma through northwest Texas. Experiment objectives required thunderstorms that met four criteria. To sample thunderstorm outflow, storms had to be large enough to transport boundary-layer air to the upper troposphere and have a lifetime long enough to produce a large anvil. The storms had to be small enough to sample safely and isolated enough that experimenters could distinguish the impact of a particular thunderstorm from other convection in the area.

To aid in the optimization of daily flight decisions, an algorithmic forecasting system was developed that produced probabilistic forecasts of suitable flight conditions for each of the three regions. Atmospheric variables forecast by a high-resolution numerical weather prediction model for each region were converted to probabilistic forecasts of suitable conditions using fuzzy logic trapezoids, which quantified the favorability of each variable. In parallel, the trapezoid parameters were tuned using a genetic algorithm and the favorability values of each of the atmospheric variables were weighted using a logistic regression. Results indicate that the probabilistic forecasting system shows predictive skill over climatology in each region, with Brier skill scores of 16% to 45%. Averaged over all regions, the forecasting system showed a Brier skill score of 32%, compared to the 24% Brier skill score shown by human forecast teams.

2.1 Introduction: mission and background

Field campaigns in the atmospheric sciences typically require the deployment of limited resources under conditions of uncertainty about the evolving atmospheric state. In most cases, human forecasters use experience and heuristics to forecast the state of the atmosphere and convey this information to decision makers. Algorithmic decision recommendation systems using probabilistic forecasts have shown promise in improving upon traditional heuristic forecasting and decision-making methods for field campaigns studying boundary layer clouds [Small et al., 2011] and cirrus clouds [Hanlon et al., 2013].

The Deep Convective Clouds and Chemistry (DC3) project during late spring and early summer 2012 sought to sample isolated thunderstorms in three study regions, each region with different climatology. The data-collection stage of the DC3 project began on 16 May 2012, continuing through 30 June 2012. DC3 investigators

deployed extensively instrumented aircraft to three study regions defined by the coverage of research-grade ground-based facilities to gather observations to improve understanding of the role of convective clouds in determining the composition and chemistry of the upper troposphere and lower stratosphere [Barth et al., 2012]. Observations were taken in three regions chosen for their coverage by ground-based facilities: northeast Colorado, north Alabama, and a larger region extending from central Oklahoma through northwest Texas which could be covered by mobile radars.

In order to build a decision recommendation system analogous to those implemented by Small et al. [2011] and Hanlon et al. [2013], a calibrated probabilistic forecasting system was required. The forecasting system needed to provide two major inputs for the decision recommendation system. First, for each day during the field experiment, the forecasting system needed to supply an estimated probability of regional weather conditions suitable for data collection using aircraft, conditional on the modeled state of the atmosphere. Second, the forecasting system had to provide a historical probability distribution of forecasts. To meet the requirements of the decision recommendation system, we developed an automated forecasting system rather than a forecasting system fed by human forecasts.

While an automated forecasting system offers less nuance than human forecasts, the automated forecasting system offers the advantages of calibration and historical applicability. By calibrating forecasts to outcomes using historical model and radar data, an automated forecasting system removes systematic biases. In contrast, because field experiment forecasters often lack the opportunity to calibrate their forecasts to the particular problem's climatology, they may systematically over- or under-forecast the probability of suitable conditions. The development of the automated forecasting system also yields a historical probability distribution of forecasts, which provides context to the forecasting system that is essential for the decision recommendation system. Obtaining such a historical distribution of human forecasts is impossible or impractical for most applications. For these reasons, an automated forecasting system was developed for the DC3 campaign as input to a decision recommendation system. The forecasting system is the subject of this paper.

2.2 Objectives and constraints

2.2.1 Definition of "good" conditions

A quantitative assessment of the probability of suitable data-collection conditions for a given time period necessarily requires a precise definition of suitable conditions. While an experienced human forecaster may be able to "eyeball" good conditions, an automated forecasting system requires an exact definition, in advance, applied consistently. A precise definition allows for the generation of pre-experiment statistical analysis, but precludes potentially helpful tinkering with the definition during the experiment. Creating such a precise definition for the DC3 campaign required an interpretation of investigators' pre-experiment documentation, interviews with principal investigators, and results of test flights. Having a definition of suitable conditions that matches the working definition used by researchers is critical to the value of the decision recommendation system.

The DC3 campaign sought isolated, deep convection [Barth et al., 2012]. To build a training dataset for the forecasting system, we need a way to quantitatively identify historical "good" conditions. For the purposes of this forecasting system, five sub-regions were considered. The Oklahoma-Texas region was represented by three sub-regions, each sub-region defined by the horizontal extent of a National Weather Service (NWS) Doppler radar site: central Oklahoma (Twin Lake, KTLX), southwest Oklahoma (Frederick, KFDR), and northwest Texas (Lubbock, KLBB). The north Alabama sub-region was defined by an approximation of the dual-Doppler coverage area from three radars: the Advanced Radar for Meteorological and Operational Research (ARMOR) located at the Huntsville airport, the University of Alabama at Huntsville Mobile Alabama X-band dual-polarimetric (MAX) radar, and the Hytop, AL (KHTX) NWS Doppler radar. This area was entirely covered by the KHTX radar, which was used to verify thunderstorm conditions in this sub-region. The northeast Colorado sub-region was defined by an approximation of the dual-Doppler coverage area from the CSU-CHILL and PAWNEE radars, modified by the assumption that planes could not fly west of the longitude of Boulder, CO due to topography. This area was entirely covered by the KFTG radar site, which was used to verify thunderstorm conditions in this sub-region. A map of the horizontal extent of the five sub-regions is shown in Figure 2.1. The historical

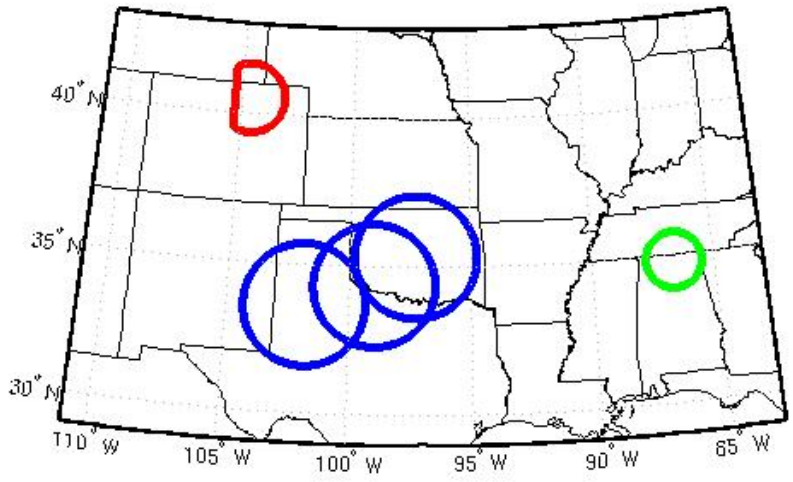

Figure 2.1. The three study regions for which probabilistic forecasts were generated: Alabama, Colorado, and Oklahoma/Texas.

set of complete volume scans at the National Climatic Data Center for all the sites was incomplete, therefore only base reflectivity data were used to characterize the state of convection.

The definition of suitable data-collection conditions conformed to experiment objectives, which required aircraft sampling of thunderstorms that met several criteria. In order to sample thunderstorm outflow, storms had to be large enough to transport boundary-layer air to the upper troposphere and have a lifetime long enough to produce a large anvil. The storms had to be small enough to sample safely and isolated enough that experimenters could distinguish the impact of a particular thunderstorm from that of other convection in the area. Isolated thunderstorms and supercell thunderstorms were deemed to be ideal targets for the DC3 campaign. Larger-scale thunderstorm systems were considered to be too large during their mature stages, but could be viable targets earlier in their development. Table 2.1 summarizes the quantitative criteria used to define "good" conditions for a particular radar volume scan. Criteria 1 and 2 ensured that convection was deep, while Criteria 3 and 4 ensured that convection was isolated and not too large. A

Criteria for "good" conditions	
Criterion	
1	Contiguous area of 50 dBZ reflectivity
2	Contiguous 50 dBZ area > 20 km^2 (40 km^2) in OK/TX region
3	80 km by 80 km box centered on area centroid has < 250 km^2 of 50 dBZ coverage
4	80 km by 80 km box centered on area centroid has < 1200 km^2 of 30 dBZ coverage
Criteria must be met for 80% of hourly radar scans	

Table 2.1. The criteria required for "good" conditions during the DC3 campaign, using base reflectivity data. For an hour to be considered "good", these four criteria must be met for 80% of the radar volume scans in the hour.

"good" hour is defined as one during which at least 80% of radar scans are "good", while a "good" day is defined as one with at least one good hour between 15Z and 00Z.

2.2.2 Calibrating conditional probability

For a well-calibrated forecasting system, the forecast probability for each region at each hour is a best estimate of the probability of good conditions based on the modeled state of the atmosphere at that hour. Rather than using the reflectivity output from the model directly, the forecasting system uses a post-processor to convert certain model output variables to a probability of good conditions. The post-processor is trained on historical model data and concurrent historical realizations of the desired conditions as determined from the NWS radars at each site. Because it is trained on past model data and radar data, this conditional probability of good conditions is well-calibrated: the forecasting system will not, in the long run, over- or under-forecast the probability of suitable conditions.

2.2.3 Converting from hourly forecasts to daily forecasts

Experience has shown that in mesoscale-forced situations such as those relevant to the DC3 campaign, the mission suitability of a day can change rapidly, on timescales of minutes to hours. In order to represent this rapid evolution properly, the forecasting system generates forecast probabilities of good conditions for each

hour of each afternoon. Sets of hour-by-hour forecast probabilities are converted to the probability of suitable conditions occurring at some time during the day using a logistic regression that fits historical hour-by-hour forecast probabilities and historical day-by-day conditions. This fitted daily forecast probability is needed by the decision recommendation system, which makes decision recommendations on the day-by-day timescale as required by the DC3 decision cycle.

This calibration is possible with the forecasting system because we have a record of what the system would have predicted in the past. For human forecasts, we have no such historical record available for most field experiments, so calibrating hourly forecasts to daily forecasts is difficult.

2.2.4 Limited availability of historical data

In order to capture the forecast climatology of the field campaign period, training data were limited to afternoon conditions during May and June. Because model and radar data were scarce, training data were limited to approximately 3000 hourly cases for each domain yielding approximately 300 daily cases. The use of a high-resolution numerical weather prediction model [Weisman et al., 2008] was necessary in order to resolve the isolated convection sought by DC3 investigators. This requirement for output from a research-grade model presented sample-size issues; however, an operational model such as the GFS would have offered a larger sample but exhibited less skill in resolving the relevant meteorology. The limited amount of data forces a simplifying assumption that all hours in the training data are used to train the same model, regardless of time of day and day of season. This assumption implies that the diurnal and seasonal variation in the probability of isolated thunderstorm formation is explained entirely by the diurnal and seasonal variation of the model predictors.

2.3 Forecasting system design and implementation

2.3.1 Forecasting system methodology

The forecasting system is inspired by the requisite conditions for thunderstorm development outlined by Fawbush and Miller [1953]. Fawbush and Miller's four

conditions required simultaneously for tornadic thunderstorm development were: convective instability, relatively dry air aloft, wind shear, and the presence of a mechanism to trigger convection. While the forecasting system was not seeking tornado development, we used these conditions as a proxy for the conditions required for supercell development. Our system represents these four conditions with four model-forecast meteorological predictors for each region. Convective instability is approximated by domain median mixed-layer convective available potential energy (MLCAPE), moisture aloft is approximated by domain median 700 mb relative humidity (RH) (500 mb RH used for CO and TX domains to account for higher elevation), wind shear is approximated by domain median bulk Richardson number (BRN), and the presence of a "trigger" is approximated by domain maximum 850 mb vertical velocity (700 mb vertical velocity used for CO and TX domains to account for higher elevation). The use of low-level vertical velocity as a predictor is also intended to account for convective inhibition. If the modeled convective inhibition is stronger than the modeled lifting mechanisms, the domain will not have high values of vertical velocity. The forecasting system was trained on these predictors rather than model-forecast radar reflectivity because the model reflectivity is sensitive to the model-resolved microphysics, introducing an unnecessary source of error.

Predictors for the forecasting system were drawn from the 0000 UTC run of the National Center for Atmospheric Research (NCAR) 3km Weather Research and Forecasting (WRF) model [Weisman et al., 2008]. Each hour between 1500 UTC and 0000 UTC during May and June during the period of record of the NCAR 3km WRF was treated as an independent case of training data for the forecasting system. We have also assumed that year-over-year changes in the physics and parameterizations of the NCAR 3km WRF did not significantly affect the forecast predictors. Corresponding radar data from each hour for each domain were converted to a binary response variable as described in Table 2.1, denoting good hours and bad hours. Concurrent model data and radar data were available for approximately 3000 hours for each domain, constituting approximately 3000 cases of training data. The five regions were treated separately in the forecasting system development because the convective climatologies differ for each.

As inspired by systems used by NCAR for other atmospheric applications [Williams et al., 2008], fuzzy logic trapezoids were used to transform the raw values

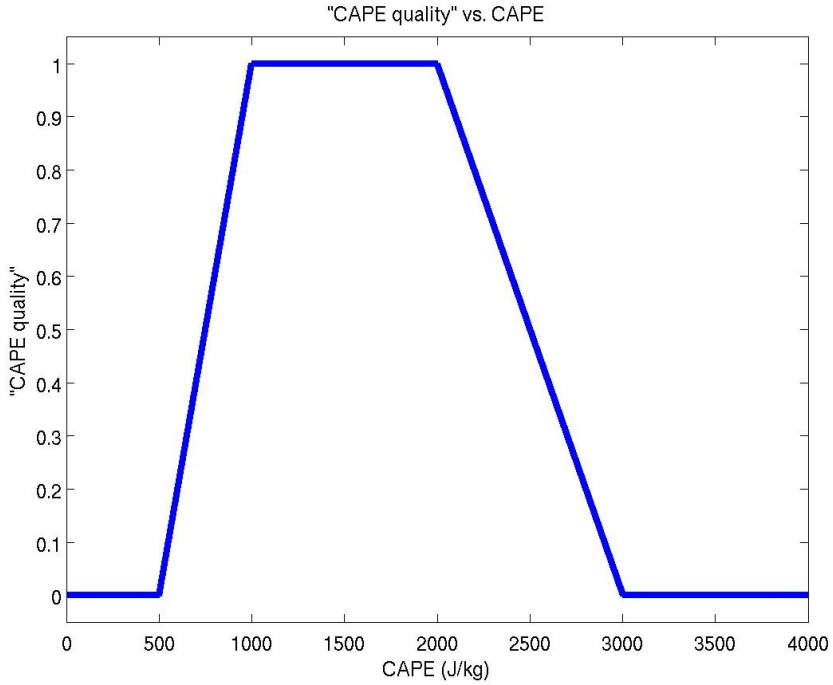

Figure 2.2. A sample fuzzy logic trapezoid. In this idealized example, the quality of CAPE is 1 for CAPE values between 1000 and 2000 J/kg, 0 for CAPE values below 500 J/kg and above 3000 J/kg, and varies linearly along the sloped portions of the trapezoid. The trapezoid can be defined by its four vertices, at 500, 1000, 2000, and 3000 J/kg.

of each predictor. For each predictor, the "suitability" of that variable is assumed to be expressible on a scale from 0 to 1. While in traditional fuzzy logic, this "suitability" value is treated as a probability, we use this value as a measure of parameter suitability in order to combine the suitability of multiple predictors in a calibrated fashion. The "variable suitability" as a function of the variable is defined by a trapezoid function, for example, as shown in Figure 2.2. Figure 2.2 shows CAPE "suitability" vs. CAPE. In this idealized example, domain median CAPE below 500 J/kg or above 3000 J/kg is assigned a "suitability" value of 0: perhaps too little CAPE produces no storms and too much CAPE produces storms that are too vigorous or too numerous to safely sample. In this example, CAPE between 1000 J/kg and 2000 J/kg is ideal and assigned a value of 1. On the sloped parts of the trapezoid, CAPE "suitability" varies linearly with CAPE. This trapezoid is defined by 4 parameters: its 4 vertices. Each of the 4 predictors has a corresponding trapezoid function, giving a total of 16 parameters for each of the five domains.

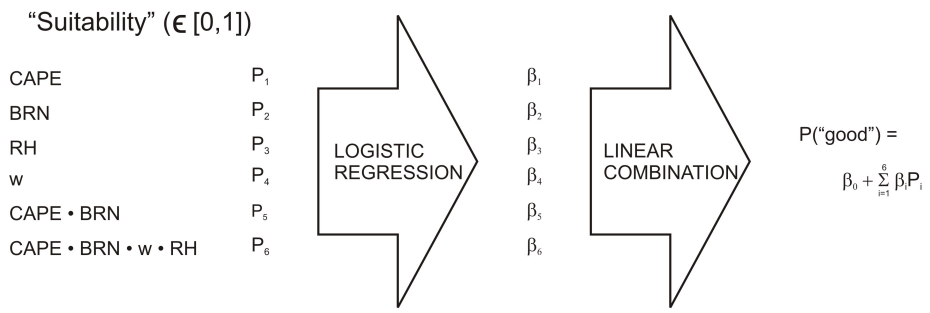

Figure 2.3. A diagram explaining the conversion of model-forecast predictors to forecast probabilities. The trapezoids fit by the genetic algorithm determine the values of $P_i, i = 1, \ldots, 6$, which serves as a measure of predictor "suitability" where 1 is ideal and 0 is unsuitable. Based on historical forecast and verification data, a logistic regression is used with predictors P_i, yielding coefficients $\beta_i, i = 0, \ldots, 6$. The coefficients are then combined with the predictor suitability values, giving a probability of "good" conditions $P_{good} = \beta_0 + \sum_{i=1}^{6} \beta_i P_i$.

Logistic regression was used to convert any set of "suitability" of predictors into a forecast probability. Six predictors, each with values between 0 and 1, are input into the logistic regression: CAPE suitability, BRN suitability, mid-level RH suitability, low-level vertical velocity suitability, CAPE suitability × BRN suitability, and the product of all four "suitability" values. The value of these six predictors aggregated using a tuned set of six regression coefficients produces a value between 0 and 1 corresponding to the probability of suitable conditions in a given hour. Figure 2.3 offers a visual demonstration of the progression from variable "suitability" to forecast probability.

The tuning of the 16 parameters defining the trapezoids and the tuning of the six logistic regression coefficients occur in parallel using a genetic algorithm, a nonlinear optimization tool from the field of artificial intelligence [Haupt and Haupt, 2004]. The genetic algorithm settings are displayed in Table 2.2. The genetic algorithm solves for the 16 trapezoid parameters such that the fit of the 6 logistic regression coefficients minimizes the Brier score [Brier, 1950] of the set of forecasts on the hourly training data. Ten instances of the genetic algorithm are run for each region as a genetic algorithm ensemble. The median of each parameter from the genetic algorithm ensemble is used in the forecasting system for that region.

Figure 2.4 shows the trapezoids generated for each predictor in each region.

CAPE lower bound	0
BRN lower bound	0
RH lower bound	0
CAPE upper bound	5000
BRN upper bound	1000
RH upper bound	1
W upper bound	18.5
Initial CAPE trapezoid parameters	unif(0,4000)
Initial BRN trapezoid parameters	unif(0,80)
Initial RH trapezoid parameters	unif(0,1)
Initial W trapezoid parameters	unif(0,15)
Generations	100
EliteCount	0
HybridFcn	@fminsearch
PopulationSize	80
FitnessScalingFcn	@fitscalingrank_4th_root

Table 2.2. The settings used by the genetic algorithm. Lower bounds of the parameters for CAPE, BRN, and RH prevent negative values of parameters for those variables. Upper bounds of the parameters for CAPE, BRN, RH, and W are set on the order of the highest model-forecast values of those variables. Initial parameters are drawn from a uniform distribution and sorted from smallest to largest. "Generations" is the maximum number of iterations before the genetic algorithm stops. "EliteCount" is the number of individuals that survive to the next generation. "HybridFcn" is the function that continues optimization after the genetic algorithm terminates. "PopSize" is the number of individuals in the population. "FitnessScalingFcn" is the function that scales values of the fitness function. All other genetic algorithm settings are default settings from the MATLAB Global Optimization toolbox.

The forecasting system prefers moderate values of MLCAPE. The BRN values are consistent with the range for supercell thunderstorms (10–40) given by Weisman and Klemp [1982, 1986]. The system prefers high values of mid-level relative humidity, consistent with the K-index criteria for airmass thunderstorms [Reap and Foster, 1979]. In the two regions (KFTG and KLBB) where 700 mb vertical velocity was used instead of 850 mb vertical velocity to account for differences in elevation, the system allows higher values of vertical velocity. In the other three regions, the forecast system seeks to avoid more violent updrafts.

A second logistic regression is used to convert a set of 10 hourly probabilities from 15Z through 00Z inclusive into a single daily probability. The hourly probabilities offer an upper bound and lower bound on the daily probability. If all forecast hours

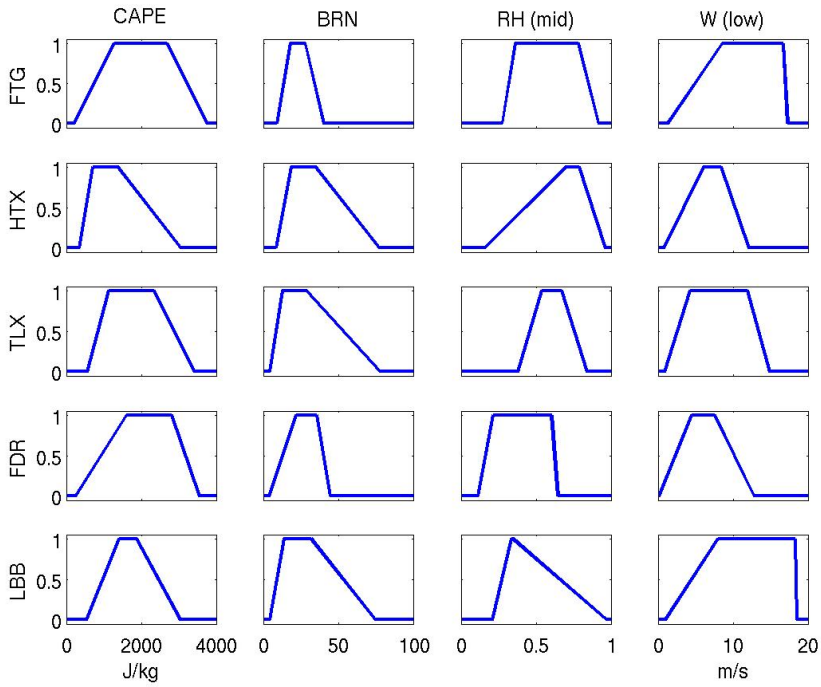

Figure 2.4. The trapezoids from each region, as fit by the genetic algorithm. The five sub-regions are represented by the corresponding NEXRAD site: northeast Colorado (FTG), north Alabama (HTX), central Oklahoma (TLX), southwest Oklahoma (FDR), and northwest Texas (LBB).

were independent, the daily probability of good conditions would be

$$fx_{UB} = [1 - \prod_{i=1}^{10}(1 - P_i)] \tag{2.1}$$

where P_i is the hourly forecast for hour i. Likewise, the daily probability can be no lower than the highest hourly probability,

$$fx_{LB} = max(P_i). \tag{2.2}$$

The logistic regression uses fx_{LB} and fx_{UB} as two predictors for each day, yielding a daily forecast probability. This allows the actual degree of serial correlation in the hourly probabilities from a single afternoon to be accounted for using the historical data.

The Oklahoma region daily probability is defined as the maximum of the 3 daily

probabilities for the three subregions,

$$P_{OK} = max(P_s), \tag{2.3}$$

where $s = 1, 2, 3$ denote the three subregions. This definition is based on the in-flight mobility of the aircraft and the statement from DC3 principal investigators that a successful flight to any of those subregions is equally acceptable to meet experiment objectives for the larger Oklahoma region.

Because only approximately 300 days of concurrent model and radar data were available in each region, the forecasting system used all available training data, leaving no independent test data. The lack of independent test data increases the risk that the forecasting system is overly optimistic due to overfitting [Witten and Frank, 2005]. With the number of "good" days in each region on the order of tens, we conjectured that the cost of not using all training data outweighed the benefit of having independent test data. The performance of the forecasting system during the DC3 campaign serves as independent test data for the forecasting system. For future forecasting problems with more data available, the forecasting system can be cross-validated on testing data prior to the operational implementation of the system, reducing the risk of a model overfit to training data.

2.4 Results

On each day of the DC3 field campaign, a morning weather briefing occurred at 0830 CDT (1330 UTC) to discuss the expected weather conditions for the next several days. Regional forecast teams comprised of human forecasters with significant expertise in each of the three regions issued a probabilistic forecast of thunderstorms in their region in three-hour time increments and 20% probability increments for the upcoming two days. The probabilistic forecast was presented as a percent chance of thunderstorms and the most probable storm mode: isolated, scattered, supercell, squall line, or mesoscale convective system. An example of a regional forecast is shown in Figure 2.5.

At the same time, model output from the NCAR WRF model, available to all forecasting teams, was used by the forecasting system to generate a probabilistic forecast for each of the three regions. Hourly probabilistic forecasts were aggregated

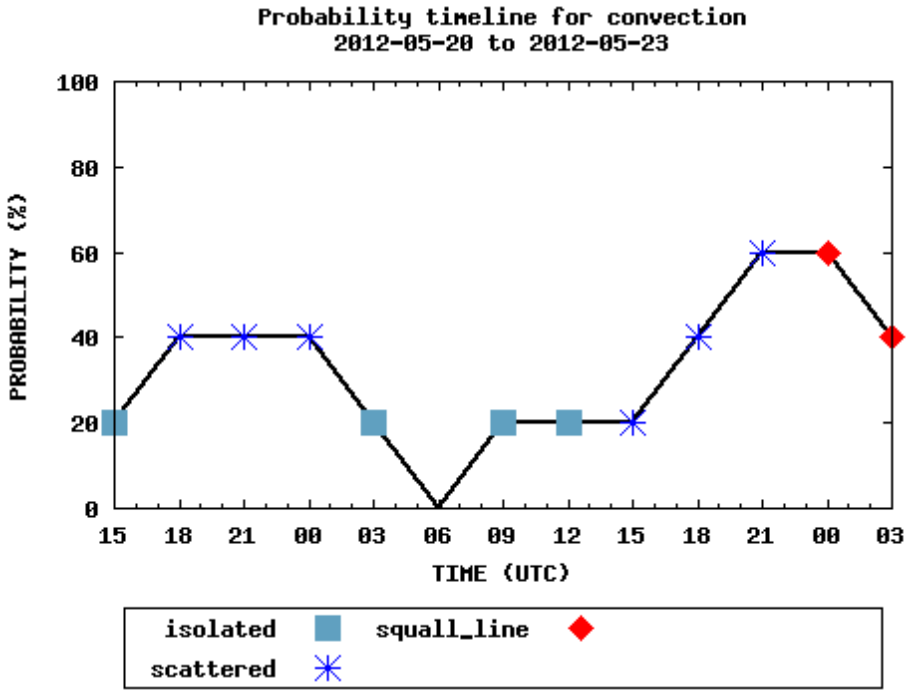

Figure 2.5. A sample forecast from the Alabama regional forecast team, issued on the morning of 20 May.

into a probability of suitable thunderstorms during the afternoon for each of the three research regions and displayed, along with a recommendation of which region to sample, if any. At each daily weather briefing a representative of the decision recommendation team provided the probability of suitable conditions in each subregion for the current day and the accompanying decision recommendation.

The forecasting system showed skill over climatology in each of the three regions. The forecasts issued by the forecasting system showed a 45% Brier skill score improvement over climatology in Colorado, a 36% improvement over climatology in Alabama, and a 16% improvement over climatology in Oklahoma.

2.5 Comparison with human forecast teams

Evaluating the skill of the automated forecasting system against the regional forecast teams was challenging, because the forecast teams produced forecasts in three-hour increments, while the forecasting system produced forecasts in hourly increments aggregated to a daily forecast, as required by the decision makers. Another challenge

is the ambiguity of the precise meaning of the forecast probabilities from the forecast teams. A three-hour increment forecast probability of 40% at 1800 UTC could be interpreted several ways: 40% probability of thunderstorms at any time between 1800 and 2100 UTC, 40% probability of thunderstorms exactly at 1800 UTC, 40% probability of thunderstorms at any time between 1730 and 1830 UTC, 40% probability of thunderstorms at any time between 1630 and 1930 UTC, or 40% probability of thunderstorms at any time this afternoon if the 1800 UTC conditions were to hold all afternoon. Furthermore, the forecasters for different regions and even different forecasters for the same region may have interpreted the forecast probabilities differently. By comparison, the forecasting system offers a precise, unambiguous definition of its forecast probabilities for the day, although in the aggregation process, information to guide the decision for the optimal flight time was lost.

As a first attempt for comparison, the forecasts from the regional forecast teams were linearly interpolated to hourly forecasts. For example, as shown in Figure 2.5, the forecast probability of thunderstorms at 1500 UTC is 20% and the forecast probability of thunderstorms at 1800 UTC is 40%. This forecast was interpolated to a 1600 UTC forecast probability of 27% and a 1700 UTC forecast probability of 33%. Based on the line-graph presentation of the forecast probabilities in Figure 2.5, linear interpolation of forecast probabilities is a logical interpretation; the method of forecast presentation will influence the thought process of the decision-maker. These hourly forecast probabilities were then verified against hourly radar data and compared to an hourly forecast consisting of the climatological hourly probability of suitable conditions. Using this method of comparison, the regional forecast teams performed worse than a climatological forecast, forecasting thunderstorms to occur much more often than climatology. We concluded that this method of aggregating the human forecast teams' probabilities was poorly representing what the forecast teams meant.

An alternate method of assessing the regional forecast teams' skill was aggregating their interpolated hourly forecast probabilities into a daily forecast probability, using the same logistic regression coefficients that were used to aggregate the forecasting system's hourly forecasts into daily forecasts. This method of assessment showed a similar result: worse performance than climatology for the regional forecast teams.

A final method of verifying the skill of the regional forecast teams was to use the maximum hourly probability predicted by the regional forecast teams during the 1500 UTC – 0000 UTC period as the forecast probability of suitable thunderstorms occurring at some time during that period. This method adjusts for the tendency for the regional forecast teams' hourly forecast probabilities to be too high. Using this method, the regional forecast teams demonstrated skill on average, with Brier skill scores showing a 8% skill reduction from climatology in Colorado, a 52% improvement over climatology in Alabama, and a 28% improvement over climatology in Oklahoma.

After adjusting for the bias in their forecasts, the human forecasters performed better than climatology. Averaged over all regions, the forecasting system (BSS = 32%) showed a small advantage over the human regional forecasters (BSS = 24%). Figure 2.6 shows the reliability diagram for the forecasting system, aggregated over all regions, while Figure 2.7 shows the reliability diagram for the human forecasters, aggregated over all regions. The human forecasters issue more forecasts for probabilities above the 0% probability bin, but are less reliable on these forecasts than the forecasting system.

2.6 Discussion

2.6.1 Difficulty in quantifying the definition of "good"

The forecasting system requires a specific definition of "good" conditions. The process of interpreting a quantitative definition from the DC3 operations plan required several iterations.

Our first definition of suitable conditions attempted to include deep convection while excluding multi-cell or supercell thunderstorms that would be more difficult to sample safely with the aircraft. Feedback following a conference with DC3 principal investigators suggested that our upper bound on the size of convection was too restrictive: they were willing to fly near larger and more severe convection than our definition allowed. Supercell thunderstorms, which investigators considered to be ideal, were being excluded by our definition, resulting in systematically low probabilities of suitable conditions.

A second definition was crafted to allow for larger isolated and supercell thun-

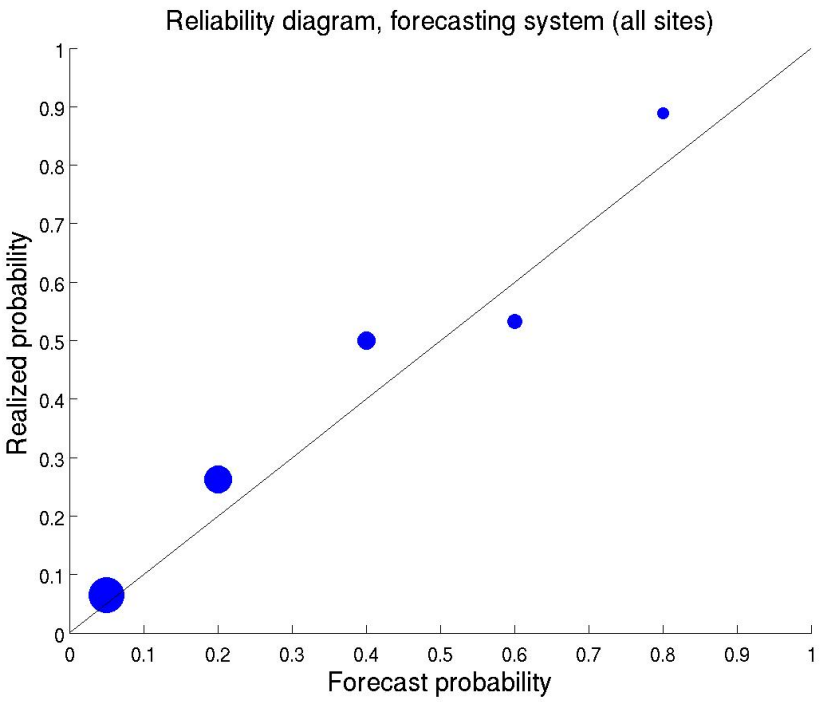

Figure 2.6. A diagram showing the reliability of the automated forecasting system. In this diagram, binned forecast probability is plotted on the abscissa while the corresponding realized probability on all days in that bin is plotted on the ordinate. A more reliable forecasting system will more closely follow the $x = y$ diagonal than a less reliable forecasting system. The size of points are area-weighted by the number of forecasts in each bin. A forecasting system with more resolution will have more weight in the extreme bins (forecasts closer to 0% or 100%) than a forecasting system with less resolution.

derstorms while still excluding mesoscale convective systems, which were deemed to be too complicated to allow reliable attribution of sampled outflow to a particular portion of the inflow boundary layer. During the instrument testing phase before the DC3 campaign, probabilities tuned on this definition were communicated to DC3 investigators. Investigators indicated that this second definition was not restrictive enough on the lower end of convective intensity. This definition was too generous in defining as suitable shallow, "popcorn" convection, which was not a viable target for the DC3 campaign because of its lack of upper tropospheric outflow.

To eliminate shallow convection, the final definition introduced a minimum horizontal area coverage of 50 dBZ reflectivity. In the Colorado and Alabama regions, a thunderstorm needed to have at least 20 km^2 of contiguous 50 dBZ reflectivity to

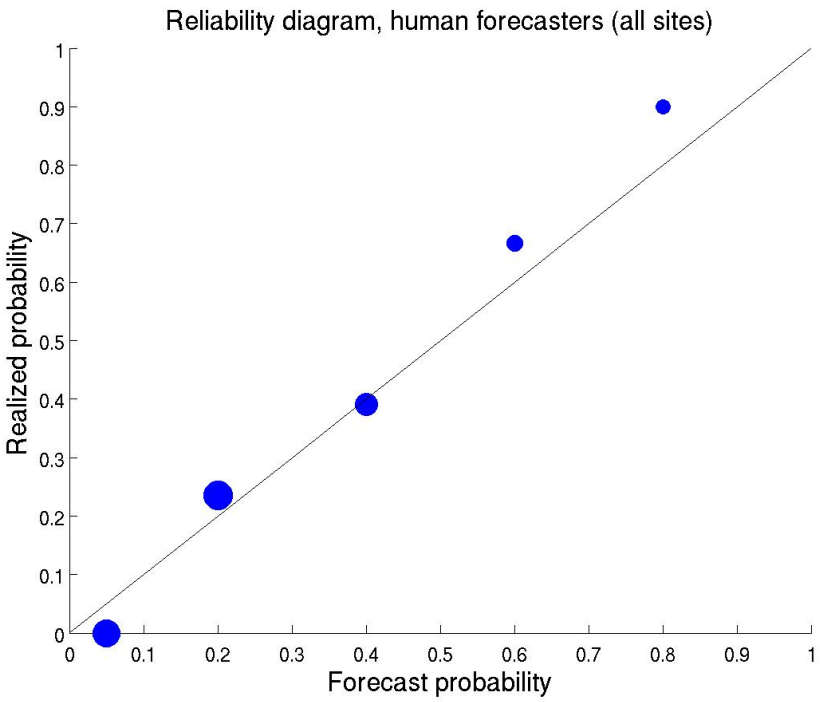

Figure 2.7. A diagram showing the reliability of the human forecasters. This figure is the same as Figure 2.6, but showing the forecasts from the humans rather than the automated system.

be defined as deep convection. In the Oklahoma region, this minimum horizontal area was 40 km^2. This minimum area was used as a proxy for thunderstorm depth, preventing unsuitable shallow convection from being considered suitable by the forecasting system. This third definition was accepted by the DC3 investigators, who found the probabilities tuned on this definition seemed appropriate based on their knowledge of atmospheric conditions.

Effective forecasting and decision making is hampered without a clear definition of the desired environmental conditions. The problems with our initial definitions were an example of the difficulties of cross-disciplinary communication: the chemists and meteorologists in the field had a clear vision of what they wanted, but we struggled to correctly interpret their terminology. It is imperative that the team developing an automated forecasting system interact with the experiment principal investigators to assure that the definition meet the experimental requirements. Moreover, care should be given in the design of the system to allow fast reset of the forecasting and decision making system should the investigation team decide

		Reliability	Resolution	Uncertainty	Brier score
System	C Oklahoma	0.024	0.122	0.349	0.251
	Colorado	0.024	0.099	0.214	0.139
	Alabama	0.065	0.221	0.394	0.238
	SW Oklahoma	0.061	0.109	0.450	0.402
	NW Texas	0.019	0.065	0.420	0.374
Humans	OK/TX	0.237	0.297	0.498	0.439
	CO	0.108	0.079	0.219	0.248
	AL	0.052	0.260	0.401	0.193

Table 2.3. The results of a Murphy decomposition of Brier skill scores for each regional forecast team and the automated forecasting system in all three regions. Low values of "reliability" indicate a more reliable forecasting system, while high values of "resolution" indicate a forecasting system with better resolution. "Uncertainty" is a measure of sample climatology: higher values of uncertainty indicate climatology closer to 50%. (The slight differences between the Uncertainty values in the AL and CO region between the automated forecasting system and the human forecasters is due to a few days where forecasts were available from one system but not the other.) The Brier score is $Reliability - Resolution + Uncertainty$. Low values of Brier score indicate greater skill.

to modify their definition based on conditions encountered during the deployment. Having agreement on such a quantitative definition of suitable conditions from principal investigators before the experiment will not only help inform the development of an algorithmic decision system, but also provide the investigator team access to a statistical analysis of the events they seek to study.

2.6.2 Murphy decomposition of forecasting systems

From a forecasting perspective, the relative strengths and weaknesses of the forecasting methods can be shown by a Murphy decomposition [Murphy, 1973, Table 2] of each forecasting method's Brier score. The humans forecasting in the Colorado and Alabama regions rotated during the field season, while the Oklahoma region employed the same forecasters throughout the field season. The decompositions are shown in Table 2.3. In the Colorado region, the resolution of the human forecast team is similar to that of the automated forecasting system, but the automated forecasting system is more reliable and thus scores better. In the Alabama region, the human forecasters were roughly as reliable as the automated forecasting system, but their forecasts showed greater resolution, leading to a better overall skill score.

The results from the Oklahoma human forecasts offer a particularly striking

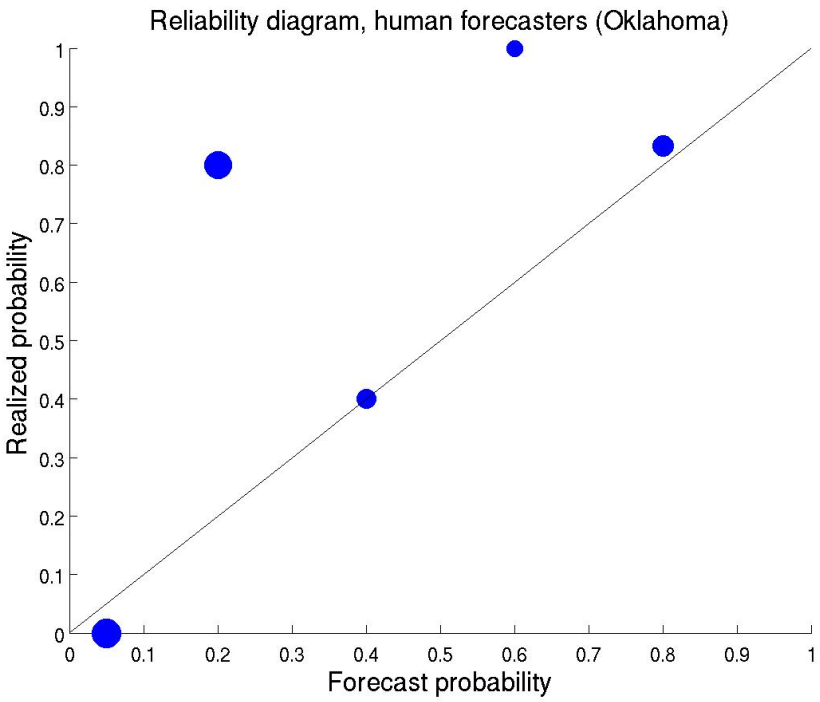

Figure 2.8. A diagram showing the reliability of the Oklahoma human forecasters. This figure is the same as Figure 2.6 and Figure 2.7, but only showing the forecasts from the Oklahoma human forecasters.

example of the difference between Brier resolution and Brier reliability. Figure 2.8 shows the reliability diagram for the Oklahoma forecast team. The human forecasts are unreliable: for example, 8 of 10 days when the 20% category was forecast by the Oklahoma forecasters verified as "good". However, the diagram suggests that the Oklahoma forecasters were quite skillful at distinguishing "good" days from "bad" days. The 0% category was forecast by the Oklahoma forecasters 11 times; of these 11 days, 0 were "good". Some other category was forecast by the Oklahoma forecasters 25 times; of these 25 days, 19 were "good". This indicates what one might expect: that the experts in Oklahoma are excellent forecasters. However, the probabilities submitted are systematically too low. The degree of miscalibration shown by the Oklahoma human forecast team is unlikely to occur in an automated system anchored to historical data.

2.6.3 Ambiguities in human probabilistic forecasting

Every day, field campaigns require making expensive decisions under probabilistic information, a process that demands exactitude in the specification of forecasts and the definition of suitable conditions. This precision in probabilistic forecasting is inherent to an automated forecasting system but is difficult for human forecasts to achieve due to the number of potential ambiguities associated with a human forecast.

The interpretation of the regional forecast teams' probabilistic forecasts of convection presents a source of possible ambiguity. Each morning during DC3, the regional forecast teams issued probabilistic forecasts for each three-hour period for the next two days, as shown in Figure 2.5. In this format, the presentation of a forecast probability at, for example, 18Z, allows for several alternative plausible interpretations, including the probability of conditions being present at any time between 1800 UTC and 2100 UTC, the probability of conditions being present any time between 1630 UTC and 1930 UTC, the probability of conditions being present any time between 1730 UTC and 1830 UTC, and the probability of conditions being present at exactly 1800 UTC. Discussions with operational forecasters, both involved with and independent of DC3, indicated that there is no clear standard for interpreting the "valid time" of such a probabilistic forecast. The evaluation of the forecasters' skill is sensitive to the interpretation of the forecasts' time window. An automated forecasting system, however, is developed such that only a decision-relevant probabilistic forecast is issued, removing any ambiguity in the interpretation of the forecast time window.

While the probabilistic forecasts from the regional forecast teams were issued in three-hour increments, the DC3 decision cycle requires that flight decisions be made on daily intervals, as only one flight to one region can be undertaken per day. Calculating a daily forecast probability from a set of sub-probabilities contained in the same time period requires a covariance matrix of the sub-probabilities. Facing the same issue, rather than calculate a covariance matrix of hourly probabilities from historical data, the automated forecasting system used a logistic regression anchored to historical forecast and verification data to convert sub-daily probability forecasts to daily probability forecasts. No such record of historical forecast data exists for the human forecasters, making the aggregation of the sub-daily probability

forecasts into a decision-relevant daily forecast a difficult task.

Ambiguity in the human probabilistic forecasts complicates the comparison of the human forecasters to the automated forecasting system by preventing the one-to-one comparison of the human forecasts to the automated forecasts. While this ambiguity is inconvenient for us when evaluating the performance of our forecasting system, it also could be inconvenient to decision-makers relying on these forecasts. The nature of the automated forecasting system allows for these ambiguities to be resolved and translated into a probabilistic forecast on a decision-relevant timescale using quantitative decision-relevant criteria.

2.7 Conclusion

The particular forecasting problem faced by DC3 investigators was many-dimensional. A forecast of suitable data collection conditions needed to resolve, as a function of time throughout an afternoon, the location of a storm with respect to the spatial coverage of the ground-based facilities, the size of a storm, the presence of nearby convection, and the time-duration of suitable conditions. Years of experience allow human forecasters to provide tremendous forecasting insight that may never be possible for an automated forecasting system, but an automated system may be better suited for the many-dimensional forecasting problems faced by DC3 and other atmospheric field campaigns. In most regions, using reasonable interpretations of the human forecasts, the human forecasters showed Brier resolution better than that of the automated forecasting scheme. However, the skill advantage from the human forecasters' better Brier resolution was offset by the skill advantage from the automated forecasting system's better Brier reliability. The authors suggest that due to the numerous possible ambiguities at various stages of the probabilistic forecasting process, relatively poor Brier reliability is likely to be a consistent problem for human forecasters in field campaign decision-making applications.

While the automated forecasting system implemented for the DC3 campaign showed skill in all regions comparable to skill shown by teams of expert human forecasters, for most atmospheric field campaigns, an automated forecasting system cannot replace forecasters. In the case of DC3, the automated forecasting system offers no information on spatial scales below the region level, nor can such a system provide the real-time forecasting support needed for flight decisions. Human

forecasters offer the ability to anticipate a rapidly unfolding weather scenario and react to atmospheric conditions that can change on timescales on the order of minutes. A skilled human forecaster can draw on experience and physical intuition to forecast the possible timeline of events in a way that our automated forecasting system is unable to emulate.

The advantage of an automated forecasting system is that it will produce unambiguous, consistent, calibrated probabilities for the desired events. The availability of these forecasts will allow the human forecasters to focus all their attention on forecast situations where they thrive, taking advantage of the detailed physics and situational knowledge not available to the automated system. An automated forecasting system issuing accurate forecasts on a day- and region-scale allows for the use of an optimizing decision recommendation system, which has been shown promise in optimizing field campaign resource deployment. We suggest that automating the day- and region-scale part of the forecast and decision for field campaigns will allow human forecasters to focus on other crucial forecast challenges that this system cannot handle while maximizing field campaign data yield.

Acknowledgments

The authors acknowledge financial support from National Science Foundation Atmospheric and Geospace Science Grant AGS-1063692. We are grateful to DC3 principal investigators Mary Barth, William Brune, Chris Cantrell, and Steven Rutledge for allowing us to apply our methodology to their field campaign. We would like to especially acknowledge Morris Weisman for making model data available to us and offering invaluable comments on our manuscript.

Chapter 3
Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign

While the automated forecasting system is a necessary input to an automated decision algorithm, the most direct measure of the algorithm's value is the actual decisions it recommends based on those forecasts. In the next chapter, we present the methodology and results from the whole automated decision algorithm and compare its results to those achieved by human decision-makers, the DC3 scientists making decisions in real-time during the field campaign. This chapter appeared in the Journal of Geophysical Research: Atmospheres as *Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign* with authors Christopher J. Hanlon, Arthur A. Small, Satyajit Bose, George S. Young, Johannes Verlinde.

Abstract

Automated decision systems have shown the potential to increase data yields from field experiments in atmospheric science. The present paper describes the construction and performance of a flight decision system designed for a case in which investigators pursued multiple, potentially competing objectives. The Deep Convective Clouds and Chemistry (DC3) campaign in 2012 sought in situ airborne measurements of isolated deep convection in three study regions: northeast Colorado, north Alabama, and a larger region extending from central Oklahoma through

northwest Texas. As they confronted daily flight launch decisions, campaign investigators sought to achieve two mission objectives that stood in potential tension to each other: to maximize the total amount of data collected while also collecting approximately equal amounts of data from each of the three study regions. Creating an automated decision system involved understanding how investigators would themselves negotiate the tradeoffs between these potentially competing goals, and representing those preferences formally using a utility function that served to rank-order the perceived value of alternative data portfolios. The decision system incorporated a custom-built method for generating probabilistic forecasts of isolated deep convection, and estimated climatologies calibrated to historical observations. Monte Carlo simulations of alternative future conditions were used to generate flight decision recommendations dynamically consistent with the expected future progress of the campaign. Results show that a strict adherence to the recommendations generated by the automated system would have boosted the data yield of the campaign by between 10–57%, depending on the metrics used to score success, while improving portfolio balance.

3.1 Introduction

Field campaigns in the atmospheric sciences typically are constrained by some exhaustible resource (e.g. flight hours) used to collect data under specific atmospheric conditions. During field campaigns of this form, principal investigators confront each day a multi-objective decision: whether to allocate a portion of the available resource budget toward data collection activities for that day. In making this decision, investigators account for numerous factors including the forecast state of the atmosphere, the amount of data already collected, the quantity of resources already expended, and the amount of time remaining in the field campaign. During the typical atmospheric science field campaign, these complicated, expensive resource-deployment decisions are made heuristically using a combination of weather forecast guidance and the judgment of expert investigators.

Automated decision systems offer an alternative approach to aid investigators with resource-deployment decisions. An automated decision system generates recommendations algorithmically by means of software that integrates statistical forecasts with tools for optimization. Retrospective analyses of the RACORO

campaign [Small et al., 2011] and the SPartICus campaign [Hanlon et al., 2013] suggest that algorithmic approaches to day-to-day decision-making could optimize the data collection process while reducing the amount of human time and energy spent on forecasting and decision-making.

The present paper reports on the results of an automated decision system created for the Deep Convective Clouds and Chemistry (DC3) campaign, which ran from May–June 2012. The DC3 campaign sought in-situ measurements of isolated, deep convection using aircraft. While the objectives for the earlier RACORO and SPartICus campaigns required in-situ aircraft measurements in only one region, the DC3 campaign sought to sample isolated thunderstorms in three different regions (named henceforth as "Alabama", "Colorado", and "Oklahoma/Texas"). While the earlier field campaigns each focused on a single objective, DC3 investigators faced a more complicated multi-objective decision problem. As an example of a multi-objective decision problem, consider a consumer purchasing a microwave oven: the consumer must balance price, size, features, efficiency, and reliability without having necessarily a single one-dimensional scale that defines value unambiguously. Analogously, DC3 investigators were tasked with balancing the desire to collect data in each of three regions, without having in advance an unambiguous definition of scientific value that would determine how to negotiate tradeoffs among objectives.

Typically, the negotiation between such tradeoffs is handled by the intuition of decision-makers. The present paper describes an alternative approach, in which decision recommendations are generated using multi-criteria optimization. The implementation of a multi-criteria optimization system to assist decision makers requires the elicitation of information about the preferences of decision-makers, including the relative importance of each objective and the degree to which decision-makers require balance among all objectives. The most common approach to constructing such a multi-criteria optimization system, adopted here, is to (1) elicit decision-maker preferences; (2) construct a utility function to rank-order options, implicitly defining tradeoffs; and (3) generate recommendations guided by the principle of maximizing utility in statistical expectation.

The decision recommendation system developed for DC3 used automated probabilistic forecasts of isolated, deep convection generated using the forecasting method of Hanlon et al. [2014]. Aside from the more complicated phenomena of interest and the multiple objectives, the decision problem faced by DC3 investigators was

structurally similar to that faced by RACORO and SPartICus investigators: investigators needed to decide daily whether to expend some flight hours for research objectives or save those flight hours for a later day. The automated forecasting system offers a progressive resolution of uncertainty. Relatively sharp information about uncertainty is available in the short term from the automated forecasting system. At longer horizons, uncertainty information is less sharp and grounded in climatology. Like RACORO and SPartICus, DC3 flight decisions are constrained by resources, specifically the number of days in the field season and the amount of flight hours available for data collection, and by side constraints, including limits on flight crew availability. This paper discusses the challenges faced in the decision recommendation algorithm development, describes the algorithm, and presents the results that would have been achieved in a counterfactual field campaign where the algorithm was entirely trusted with day-scale, region-scale flight decisions.

3.2 A flight decision algorithm for the DC3 campaign

3.2.1 Summary of the DC3 campaign

The Deep Convective Clouds and Chemistry (DC3) project, which collected data between 16 May 2012 and 30 June 2012, sought to sample isolated, deep convection in three study regions defined by the coverage of research-grade ground-based facilities, shown in Figure 3.1. DC3 investigators deployed extensively instrumented aircraft to gather observations to improve understanding of the role of convective clouds in determining the composition and chemistry of the upper troposphere and lower stratosphere [Barth et al., 2012]. While the nature of the desired conditions to sample (isolated, deep convection) was the same across all regions, the scientific questions of interest in each region varied such that a successful data-collection flight to one region was not interchangeable with a successful data-collection flight to another region. Investigators wanted to launch as many successful data-collection flights as possible, but also preferred their set of successful flights to be evenly distributed among the three regions.

DC3 investigators thus sought to achieve multiple objectives: maximize the number of successful flights and minimize imbalance among the portfolio of collected data. Along with the primary objective of sampling isolated, deep convection,

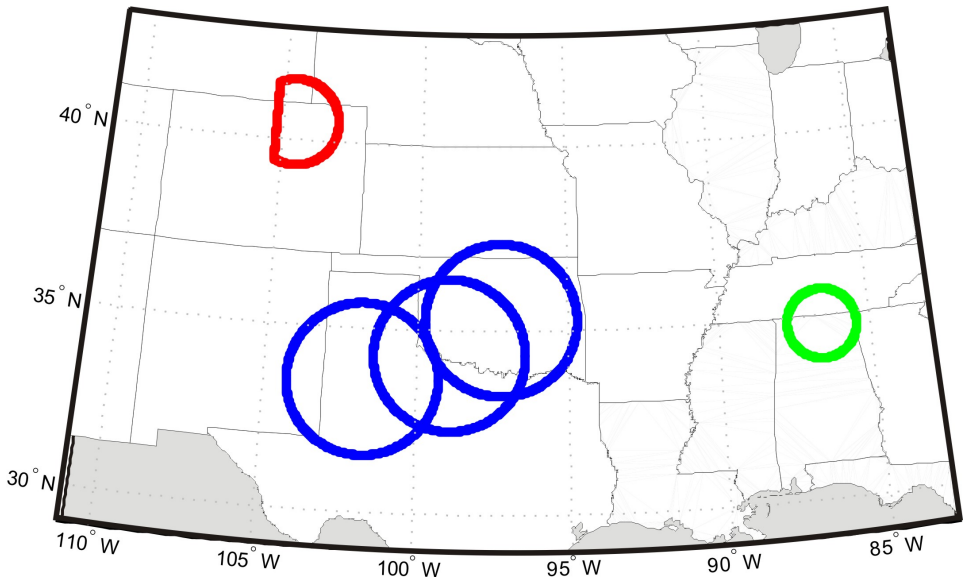

Figure 3.1. The three DC3 study regions, as approximated by the automated forecasting system and decision algorithm: Alabama, Colorado, and Oklahoma/Texas. Each study region was represented by part of one or more NEXRAD radar domains: KHTX (AL), KFTG (CO), and KLBB, KFDR, and KTLX (OK/TX). The regions were intended to correspond with ground-based research facilities. Areas west of Boulder's longitude, where topography prevented research flights, were excluded from the CO region.

investigators had related-but-distinct side objectives. In addition to the primary objective of measurements of isolated, deep convection, investigators desired measurements of "aged convective outflow." After a successful flight day sampling convection in one of the three regions, investigators assessed the likelihood of being able to track outflow from the sampled thunderstorm downstream for resampling the following day. If the upper tropospheric outflow from the sampled thunderstorm was forecast to be within range of the aircraft the next day and uncontaminated by other convective outflow, a flight would be taken to study overnight changes in the chemistry of the outflow.

Data collection by DC3 investigators was constrained by calendar time, a resource budget, flight crew restrictions, and other "soft" constraints. The calendar constraint prevented investigators from being too picky with flight decisions, as all flight hours had to be consumed by the end of the experiment, which was fixed at 30 June. The limited number of available budgeted flight hours prevented investigators from flying too frequently.

Forecasting of weather conditions for the DC3 campaign was conducted by expert

human forecasters in each of the three study regions. Each day, the forecasters issued probabilistic forecasts of deep convection to inform field campaign decision-making. Human forecasters can draw on experience and detailed physical knowledge to generate nuanced forecasts, but are saddled with a major disadvantage: obtaining a climatology of human forecasts is difficult or impossible. Without a climatology of human forecasts, there is no quantitative measure of systematic forecast bias or ability to interpret the relative rank of a given forecast compared to historical climatology.

DC3 investigators faced a complicated set of decisions each day. After an analysis of weather forecast information, investigators had to decide whether to fly to a particular study region while considering the current state of the experiment's resource budget, the number and distribution of past successful flights, and the expected climatological likelihood of future flight opportunities. If the decision was made to fly to one of the regions, investigators had to coordinate with ground-based facilities in that region, verify that aircraft and instruments were ready for operations, and plan a specific flight plan with timing and location approved by air traffic control, while apprising instrument scientists of planned operations. During flight operations, investigators had to monitor aircraft and instruments while also closely observing rapidly changing weather conditions. The difficulty of the end-to-end decision problem places a tremendous burden on human investigators, suggesting that an automated algorithm that could improve and simplify part of the process would be of great value to the field campaign.

3.2.2 Formal modeling of the DC3 decision challenge

Devising a decision recommendation algorithm to assist in the DC3 campaign demands a formalization of the DC3 decision process. The basic daily fly/no-fly decision faced by DC3 investigators is particularly well-suited to formalization. The DC3 decision cycle requires repeated, daily decisions with a regular underlying cycle. The phenomenon of interest is forecastable and training data are available for the development of a forecasting system. However, some key aspects of the decision process are poorly suited to formalization. The specific challenge of planning a flight path, including when and where to fly airplanes, is too high-dimensional to be effectively handled by this automated algorithm. The algorithmic approach also

has difficulty accounting for tacit objectives and constraints not explicitly outlined before the experiment in the science plan.

The algorithm focuses on a relatively macro-scale decision problem which is still substantial enough to provide value to field campaign investigators, allowing them to focus on the micro-scale decision problems for which the algorithm is poorly suited. The algorithm answers the macro-scale question: "should we fly today, and if so, to which region?". The four possible recommendations issued by the algorithm were: "do not fly", "fly to Alabama", "fly to Colorado", or "fly to Oklahoma/Texas". Recommendations issued by the algorithm each morning were valid for the same afternoon and were issued early enough to be incorporated into the decision process.

When modeling the DC3 decision problem, the decision recommendation algorithm considered only the primary objective of sampling isolated, deep convection within the defined study regions. Because it was deemed too difficult for an automated system to forecast, the algorithm did not offer any advice relating to the secondary objective of sampling downstream, next-day, "aged" convection. The algorithm also did not account for the unstated but real preference of investigators to retain the option of using flights for scientifically interesting "targets of opportunity" not specified in the science plan. Finally, while data collection flights meeting only some of the desired conditions could in practice be viewed as "partial successes," the algorithm defined realized conditions as binary: "good" or "bad."

The formalization of the DC3 decision problem requires a strict, quantitative definition of suitable data-collection conditions. The automated forecasting system used the definition of isolated, deep convection described in Hanlon et al. [2014] and reprinted in Table 3.1. In addition to enabling formalization and statistical analysis, precisely defining suitable data-collection conditions ensures that all experiment participants agree on the experiment's goals.

3.3 Decision algorithm methodology

3.3.1 Flowchart

Figure 3.2 shows the structure of the decision recommendation algorithm developed for use during the DC3 campaign. A numerical weather prediction (NWP) model

Objective criteria for "good" conditions in a sub-region	
Criterion	
1	Area of 50 dBZ reflectivity in sub-region
2	Contiguous 50 dBZ area $> 20\ km^2$ ($40\ km^2$ in OK/TX region)
3	80 km by 80 km box centered on area centroid has $< 250\ km^2$ of 50 dBZ coverage
4	80 km by 80 km box centered on area centroid has $< 1200\ km^2$ of 30 dBZ coverage
Criteria must be met for 80% of radar scans during hour	

Table 3.1. The criteria required for "good" conditions during the DC3 campaign, using base reflectivity data. For an hour to be considered "good", these four criteria must be met for 80% of the radar volume scans in the hour. This table is adapted from Hanlon et al. [2014a].

initially generated a physical, high-dimensional, deterministic forecast of the state of the atmosphere, which was input into a calibrated custom-built post-processor. The post-processor functioned as a coarse-but-well-calibrated analog to a human forecaster: it took weather forecasts in physical terms as input and produced as output a decision-relevant probabilistic forecast of suitable flight conditions. The post-processor also served as a tool to statistically simulate climatology. Given past NWP model output, the post-processor produced the empirical distribution of the forecast probability of suitable conditions for all past days in the training data, which was used to generate a statistical model of climatology. The post-processor thus generated the conditional same-day probability of good conditions, $P(good\,same\,day\,|\,model\,forecast)$, and the unconditional following-days probability of good conditions, $P(good\ conditions\ on\ following\ days)$.

The forecasts from the post-processor for both the same day and the subsequent days were input into an optimization module, which accounted for field campaign resource constraints, the current state of the field campaign, and the preferences of field campaign principal investigators. The optimization module used dynamic programming and Monte Carlo simulations to quantitatively account for future data-collection possibilities. The module employed a utility function to quantify the relative value to investigators of varying portfolios of successful flights, which enabled the estimation of expected utility for any future decision path. The module produced a regional-scale daily flight recommendation. Each recommendation consisted of

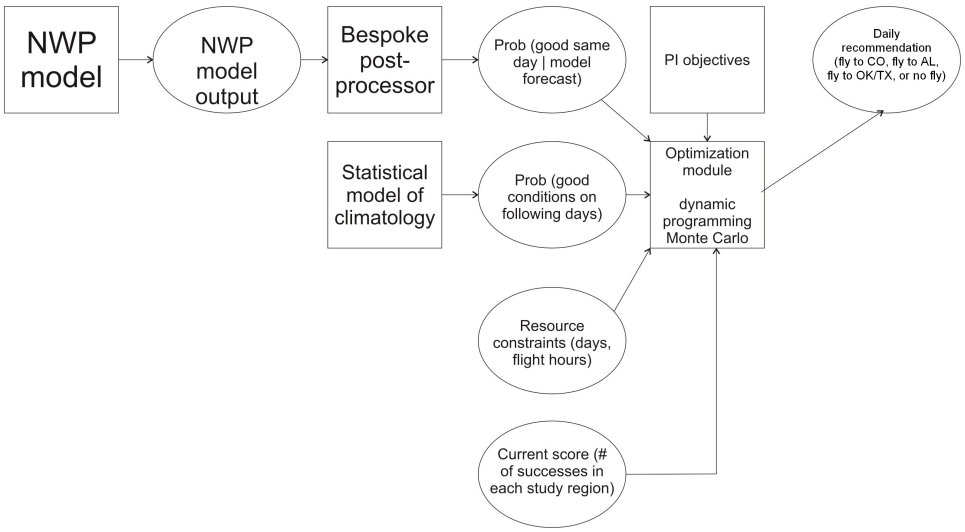

Figure 3.2. A schematic of the decision recommendation algorithm used to make daily flight recommendations during the DC3 campaign. Forecasts from a numerical weather prediction model are post-processed into probabilistic forecasts of specific conditions. These forecasts, along with climatological information and the current state of the field campaign, are input into an optimization module, which yields a decision recommendation for any day of the field campaign. Decision recommendations consist of a recommendation whether or not to fly today, and if so, to which region.

one of four choices: "fly to Alabama," "fly to Colorado," "fly to Oklahoma/Texas," or "do not fly." The recommended decision was the one estimated to maximize field campaign utility.

3.3.2 Forecasting system

The implementation of an automated decision recommendation algorithm requires a forecasting system to distinguish the likelihood of suitable conditions being present on each day during the field campaign. Specifically, DC3 investigators sought thunderstorms that met four criteria. To sample thunderstorm outflow, storms had to be large enough to transport boundary-layer air to the upper troposphere and have a lifetime long enough to produce a large anvil. The storms had to be small enough to sample safely and isolated enough that experimenters could distinguish the impact of a particular thunderstorm from other convection in the area. The definition of suitable conditions is shown in Table 3.1. An automated forecasting system was developed that produced probabilistic forecasts of suitable flight conditions for each of five sub-regions: the Colorado and Alabama regions and

three sub-regions that comprised the larger Oklahoma/Texas region. Atmospheric variables forecast by a high-resolution NWP model [Weisman et al., 2008; Romine et al., 2013] for each region were converted to probabilistic forecasts of suitable conditions using fuzzy logic trapezoids. The trapezoids quantified the favorability of each of four variables from the model: convective available potential energy (CAPE), mid-level relative humidity, bulk Richardson number, and low-level vertical velocity. In parallel, the trapezoid parameters were tuned using a genetic algorithm and the favorability values of each of the atmospheric variables were weighted using a logistic regression. Hanlon et al. [2014] provide a more complete description of the forecasting system. This forecasting system showed skill comparable to skill exhibited by human forecasters [Hanlon et al., 2014].

The same forecasting system used to generate the daily probability of suitable data-collection conditions was also used to estimate the prospects of future days for which no weather forecast is available. The forecasting system produces a well-calibrated probability of suitable conditions for each day for which past model data is available, yielding an empirical climatological distribution of forecast probabilities. For each region, a beta distribution was fit to the empirical distribution of forecast probabilities (as shown in Figure 3.3). A Monte Carlo simulation [Metropolis and Ulam, 1949] of 10000 field campaign seasons was drawn from these beta distributions. Because the distributions are unique to each region, they handle both the varying climatology for each region and the varying skill of the forecasting system across the regions. These simulations allow the decision recommendation algorithm to quantify the expected future atmospheric conditions at any stage of the field campaign.

3.3.3 Utility function and preference elicitation

As noted, DC3 PIs' objectives embraced multiple criteria: (i) to collect as much data, in aggregate, as possible ("more data"), and (ii) to collect approximately equal values of data from each of the three regions ("equal data"). Creating a decision tool requires representing these preferences formally. Essentially, the forecasting system generates a set of lotteries over alternative potential futures. The algorithm must, somehow, select one lottery from among the set of four available. The four lotteries correspond to the four possible decisions at each day's decision point: "do

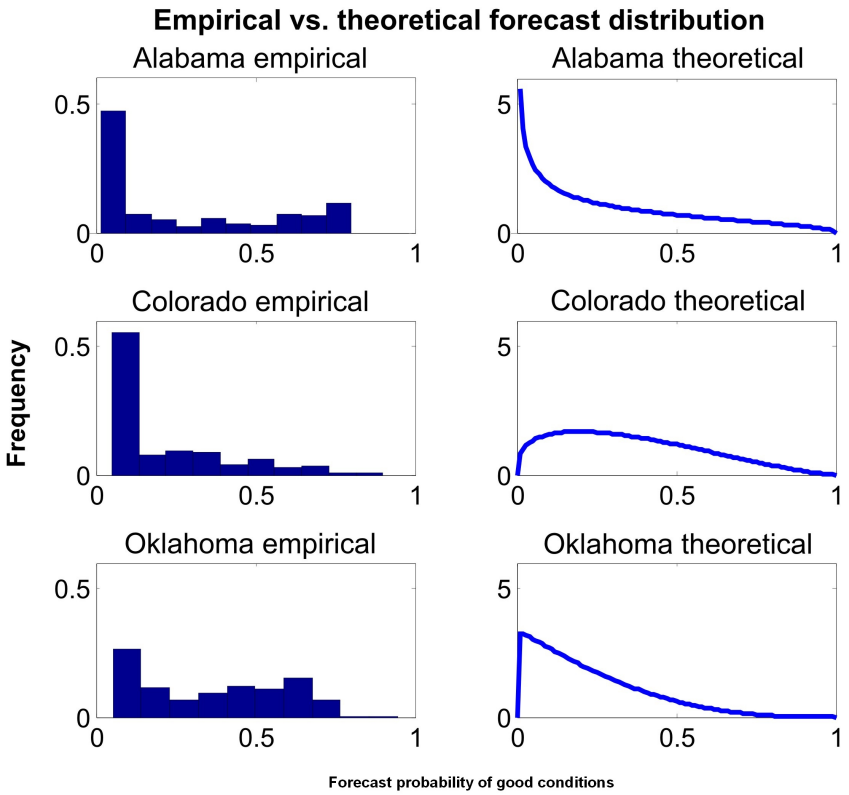

Figure 3.3. The empirical and theoretical distributions of forecast probability for all days during the training period for each region.

not fly", "fly to Alabama", "fly to Colorado", or "fly to Oklahoma/Texas". The mandate is to select the lottery that will, in expectation, generate the highest end-of-season utility for the PIs. That is, the algorithm should select, from the four available options, the one likely to give the most desirable portfolio of data by the time the field campaign concludes. In order for the algorithm to do so, it must somehow encode the desires of the campaign investigators in a machine-readable form.

PI preferences were modeled by means of a utility function $U(\vec{s})$ that assigns a numerical score to each possible end-of-season data portfolio $\vec{s} = \langle s_1, s_2, s_3 \rangle$, where s_i is the number of successful flights to region i. In order to represent the preferences of DC3 PIs, a utility function needed to satisfy certain criteria. The view that more data is always preferred to less data corresponds to a requirement that $U(\vec{s} + \vec{s'}) \geq U(\vec{s})$ for all non-negative vectors $\vec{s}$ and $\vec{s'}$ with strict equality holding if and only if $\vec{s'} = \langle 0, 0, 0 \rangle$, where $\vec{s'}$ is some additional set of successful

flights. The requirement that balanced portfolios are preferred to imbalanced ones corresponds to a requirement that $U(\vec{s}+\vec{v}) \leq U(\vec{s})$ for vectors $\vec{v}$ of the form $\vec{v} = \langle v_1, v_2, v_3 \rangle$ with $\sum v_i = 0$, where $\vec{s} = \langle n, n, n \rangle$ for some non-negative integer n, with strict equality holding if and only if $\vec{v} = \langle 0, 0, 0 \rangle$.

There are many functional forms that satisfy these requirements. The selection of a particular functional form is then guided by a desire for parsimony and clarity. One functional form used widely [McFadden, 1963] in economics to model tradeoffs is the constant elasticity of substitution (CES) utility function [Arrow et al., 1961][Uzawa, 1962]. The CES Utility Function can be written in one variation with the form

$$U(\vec{s}) = a[\sum_i (s_i + 1)^{1-\sigma}]^{\frac{1}{1-\sigma}} - K \tag{3.1}$$

where a, K, and σ are constants. The parameter σ governs the distaste of PIs for imbalanced portfolios: $\sigma = 0$ corresponds to no distaste, with larger values corresponding to a more stringent requirement for balance. The constants a and K are convenience parameters chosen to rescale the utility function for ease of interpretation. Ease of interpretation is facilitated by choosing a and K so that

$$\sum_i (s_i) - U(\vec{s}) \geq 0 \tag{3.2}$$

with equality holding if and only if $s_1 = s_2 = s_3$. (3.2) will hold if we set $a = 3^{(\frac{-\sigma}{1-\sigma})}$ and $K = 3$. Such a linear transformation of a utility function has no effect on the rank ordering of the underlying end-of-season portfolios. The difference $\sum_i (s_i) - U(\vec{s})$ can then be interpreted as a penalty the algorithm assesses against imbalanced portfolios, a penalty that is larger for larger values of σ.

The particular value of σ chosen for use during the DC3 campaign was informed through interviews with investigators eliciting their preferences. In a meeting with the DC3 principal investigators [Barth, Brune, Cantrell, Rutledge, 2012; personal communication], all investigators agreed that a portfolio of $\langle 3, 4, 3 \rangle$ (11 total successes) is preferable to a portfolio of $\langle 2, 7, 3 \rangle$ (12 total successes). According to (3.1) and (3.2), these two portfolios are equivalent when $\sigma = 1.4$, implying that the true value of σ is greater than 1.4. Likewise, all investigators agreed that a portfolio of $\langle 6, 3, 4 \rangle$ (13 total successes) is preferable to a portfolio of $\langle 4, 4, 4 \rangle$ (12 total successes), implying that the true value of σ is less than 2.0. Given these

preferences, $\sigma = 1.7$ was used to inform the automated decision recommendations.

3.3.4 Optimization module

The decision faced by investigators on each day of the DC3 campaign requires comparing the cost of expending a flight to the expected benefit achieved by flying. A model was created to estimate the optimal flight decision, given some probabilistic forecast of good conditions, for each possible future state of the field campaign (number of days remaining, number of flights remaining, and number of successful flights collected in each region). The optimization module employed an adapted version of the optimization method employed by Small et al. [2011] and Hanlon et al. [2013] for the RACORO and SPartICus campaigns, respectively.

To maximize utility in expectation, we employed dynamic programming [Bellman, 1957], a technique that solves complex stochastic optimization problems by breaking them into simpler, solvable sub-problems [Dasgupta et al., 2006]. Dynamic programming has been used to solve decision problems under weather uncertainty in fields including power dispatch [Hable et al., 2002], irrigation [Wilks and Wolfe, 1998], and air traffic management [Nilim et al., 2001]. Using dynamic programming, the implications of any flight decision can be broken into two parts: the implications for today and the implications for the rest of the experiment [Hanlon et al., 2013]. On any specific day in the experiment, the optimal decision can be described as a cost-benefit comparison: given the forecast, does the benefit of flying today and using one flight outweigh the expected cost as measured in terms of opportunity lost for future successes? To answer this question, one must estimate the expected value associated with any field campaign state. As a simple example, to make an optimal decision when there are 10 days and 3 flights remaining in the field campaign budget, we should know the expected value associated with states "9 days and 3 flights" and "9 days and 2 flights" to estimate the cost of spending a flight from the budget. These calculations are taken by means of an iterative process of backward induction. One starts with the knowledge that a field campaign state with 0 flights has 0 value and a field campaign state with as many flights as days will fly every day, with the expected value determined by the climatological average probability of suitable conditions. From these "boundary conditions", one can recursively calculate the expected value associated with any combination of

days and flights.

Using dynamic programming and backward induction, a value score V is assigned to each possible future state of the field campaign. The Monte Carlo simulations of 10000 field seasons provide a probabilistic representation of expected future conditions. By identifying paths that maximize the end-of-season utility score in expectation, the optimization module recommends the expected best current period flight decision among "do not fly", "fly to Alabama", "fly to Colorado", or "fly to Oklahoma/Texas". The technical details of the optimization module can be found in Appendix A.

3.4 Discussion/challenges

3.4.1 Logistical constraints to flight decisions

One challenge to the decision recommendation algorithm is handling the logistical constraints inherent to any field campaign. Logistical constraints are particularly extensive for a field campaign using aircraft, like DC3. These constraints can be categorized as hard constraints and soft constraints. The hard constraints can conceivably be incorporated into the algorithm before the experiment. Soft constraints on resource deployment, however, impose a more onerous challenge on the algorithm.

Field campaigns using aircraft typically require frequent rest days for pilots and flight crews. For the DC3 campaign, regulations demanded that no more than six consecutive days pass without the declaration of a "hard-down day." On a hard-down day, flight crews are granted a complete day off from field campaign activities, preventing any access to the aircraft for flight or maintenance purposes. This hard constraint is known in advance before the field campaign, allowing it to be incorporated into the decision recommendation algorithm.

Our experience during DC3 informed us that a number of soft constraints also exist. These soft constraints are more difficult to model and present a significant challenge to automated decision recommendation algorithms. During the field campaign, investigators responsible for data-collection instruments on the aircraft expressed a desire for frequent "maintenance days," where scientists have access to the aircraft to perform maintenance on their instruments. Because of the time

needed for instrument maintenance, a research flight typically could not occur on a maintenance day. Communication with instrument investigators during the field campaign confirmed that maintenance days were deemed especially necessary under certain conditions. For example, after a hard-down day, instrument scientists liked to have a day for maintenance before flying. After flying on back-to-back days, instrument scientists were also averse to flying on a third consecutive day without a day for maintenance. Near the end of the DC3 campaign, with a glut of flight hours available, instrument scientists softened their insistence on a non-flight maintenance day under these conditions. While the explicit specification of these soft constraints before the field experiment can allow the creators of a decision recommendation algorithm to account for them in the decision process, we acknowledge that the nature of such soft constraints may not be fully understood a priori.

3.4.2 Alternate objectives, unplanned opportunities, and unplanned setbacks

An automated decision recommendation algorithm can handle multiple research objectives, provided an automated forecasting system is available that can effectively forecast the future states which affect these objectives. When such future states are difficult or impossible to forecast, the alternative research objective may fall outside the purview of a decision recommendation algorithm. In addition to sampling isolated, deep convection, the DC3 campaign sought to sample "aged", previously sampled thunderstorm outflow the day after a successful flight to isolated convection. Investigators used models to track the movement of previously sampled air, and if the air was forecast to be within range of the aircraft the next day, a research flight would be taken to attempt to measure the same air to study overnight changes in its chemistry. The feasibility of an "aged" flight was a function of upper tropospheric wind patterns and the presence or absence of other thunderstorms that would contaminate the "aged" outflow and was conditional on the previous day's flight being successful. Aged flights were considered by investigators to be elusive but valuable. Because aged flights presented a forecasting problem that our forecasting system could not handle, the decision recommendation algorithm did not consider this alternate objective, but reduced the resources available for regular data collection to accommodate this objective.

Other challenges for an automated decision recommendation algorithm arise from unexpected opportunities or setbacks. Inevitably, field campaigns are riddled with events that fall outside the typical scope of pre-experiment planning, presenting a challenge to both human decision-makers and automated decision recommendation algorithms. Unplanned opportunities may allow for data-collection under conditions unspecified before the field campaign that provide great scientific value to investigators. On the other hand, unplanned setbacks threaten to negatively alter future field campaign expectations, changing optimal decision-making and undermining pre-experiment statistical analysis.

One potential weakness of the decision recommendation algorithm is its reliance on statistical analysis that assumes that all of the budgeted resources are used to address the objectives specified prior to the experiment. Field campaign investigators may decide to spend some resources on objectives not specified a priori, but which are considered to offer great scientific value to the field campaign. Investigators were presented with such an opportunity near the end of the DC3 field campaign. While the primary focus of DC3 was isolated convection, principal investigators agreed that it would be useful to sample a mesoscale convective system (MCS) that formed overnight and follow it as it decays through the morning. On 21 June 2012, DC3 used some of its flight hours to track such an MCS, rendering those flight hours and that day unusable for the experiment's primary objective. DC3 principal investigators agreed that the data collected on 21 June was valuable, despite not precisely meeting experiment objectives. We acknowledge that human decision-makers are better able to seize the option value presented by such alternative objectives than a decision recommendation algorithm. We propose that the option value associated with such unplanned opportunities can be modeled by a decision recommendation algorithm if principal investigators set aside some amount of resources in the experiment planning stage for "unexpected" objectives, rather than using resources from the general pool for these unspecified goals. The retrospective algorithm operated under the assumption that only 15 of the 22 total operational flights were available for primary data collection, essentially leaving 7 flights for aged flights and other objectives.

Finally, unplanned setbacks threaten the planning of both human decision-makers and automated decision recommendation algorithms. The algorithm is designed to simulate future possibilities under the assumption that the uncertainty

in the weather forecast is the only future uncertainty. In reality, uncertainty exists in other parts of the decision-making process, as illustrated by two examples from the DC3 campaign. On 29 May 2012, conditions seemed ideal for isolated convection in the Alabama region. Alabama regional forecasters forecast an 80% probability of convection by 0Z, while the automated forecasting system estimated an 85% probability of suitable data-collection conditions in the Alabama region. The high probability issued by the automated forecasting system led to a recommendation from the decision recommendation algorithm to fly to Alabama. Principal investigators planned to fly to Alabama on this day. However, because of the presence of Tropical Storm Beryl on the East Coast on 29 May, commercial air traffic was being redirected through Alabama, resulting in DC3 being denied airspace in the Alabama region. DC3 investigators had to change plans to fly to the Oklahoma/Texas region instead on 29 May. Later in the experiment, with 6 days left in the field campaign, engine trouble struck one of the aircraft used to sample thunderstorms, preventing its participation in the remainder of the experiment. If human decision-makers or an automated decision recommendation algorithm are systematically too optimistic about future data-collection opportunities because they did not appropriately consider the likelihood of unplanned setbacks, they may be too reluctant to use resources early in the field experiment.

3.5 Results

During the DC3 campaign, 22 flights were made on 45 operational field campaign days (May 16 through June 30). Of these 22 flights, 5 were second-day "down-wind" flights taken to sample aged convection, 1 was an overnight flight taken to sample an MCS in the northern plains, and 1 was a survey flight to convection in the Gulf of Mexico, leaving 15 flights used for the primary-objective sampling handled by the decision recommendation algorithm. With the benefit of retrospect, we can determine which daily decisions would have been recommended by the algorithm, assuming that 45 days and 15 flights are available for the primary objectives. The utility score achieved by the algorithm retrospectively can then be compared to the utility score achieved by the DC3 decision team during the field campaign.

Because investigators prefer a balanced portfolio of successful flights across the three research regions, daily decisions from the decision recommendation algorithm

are a strong function of the current "score" of the field campaign: the number of successful research flights already taken to each region. If a number of flight successes are accrued in one region, the algorithm disincentivizes flights to that region relative to regions with fewer flight successes. Operationally, the algorithm defined the current score according to the assessment of principal investigators. If a flight to one of the research regions was deemed by principal investigators to be successful, our system treated it as a success, regardless of our independent classification of the day as "good" or "bad". During our retrospective analysis, however, days were classified as "good" or "bad" using the radar-based objective classification procedure used in the training of our forecasting system. Other than its different scoring of good and bad days, the decision algorithm used in the retrospective analysis was methodologically the same as the one used during the field campaign.

The performance of the algorithm is evaluated using a retrospective analysis rather than a day-by-day piecemeal analysis of the in-the-field decision recommendations for two reasons. First, the algorithm makes its decisions holistically. On a given day, the recommended decision is not the one that achieves the best outcome that day, but the one that achieves the best expected end-of-experiment outcome. For this reason, the algorithm's performance cannot be separated into each day's performance. Second, simply comparing the algorithm's decisions in the field to the decision team's decisions in the field could produce misleading results. As a contrived example, suppose there are 10 days left in the field campaign and 1 flight left. Suppose the algorithm recommends a flight, the decision team doesn't fly, and conditions turn out to be good. Presumably, this day should be scored as a success for the algorithm. The next day, however, there are 9 days left in the field campaign and still 1 flight left in the decision team's budget. At this point, in order to provide value to the field campaign, the algorithm generates a decision recommendation based on the real field campaign flight budget, not the algorithm's counterfactual flight budget. Suppose the algorithm recommends flying again, and the team doesn't fly again, and conditions are good again. It would be possible, using this method of scoring, for the algorithm to achieve two or more successful flights from a budget that only contained one flight. The only way to avoid such a nonsense result is to score the entire field campaign based on a retrospective, counterfactual set of decisions.

The retrospectively simulated day-to-day decisions from the decision recommendation algorithm would have used the allotted 15 flights to collect 4 successful flights in Colorado, 3 successful flights in Oklahoma/Texas, and 4 successful flights in Alabama ($\langle 4, 3, 4 \rangle$), according to the objective classification procedure. Using the same objective classification procedure, the day-to-day decisions made by DC3 principal investigators yielded 2 successful flights in Colorado, 3 successful flights in Oklahoma/Texas, and 2 successful flights in Alabama ($\langle 2, 3, 2 \rangle$). Therefore, using this objective classification procedure, the algorithm would have increased the number of successful flights from 7 to 11, an increase of 57%. The subjective daily classifications made by DC3 PIs for the same days were more generous, attributing to the DC3 team 3 successful flights in Colorado, 5 successful flights in Oklahoma/Texas, and 2 successful flights in Alabama ($\langle 3, 5, 2 \rangle$). Even with the more optimistic scoring procedure used by DC3 PIs, the automated algorithm would have increased the number of successful flights from 10 to 11, an increase of 10%.

The portfolio of successful flights that would have been collected by the decision algorithm, ($\langle 4, 3, 4 \rangle$), also maintains balance across the three study regions. According to the utility function used by the decision algorithm to account for the field campaign's multiple objectives, the score $\langle 4, 3, 4 \rangle$ corresponds to a utility score of 10.87. The portfolio of successful flights attained by DC3 investigators yields a utility score of 6.84 or 9.14. Based on this scoring method which penalizes less balanced portfolios, the algorithm improved the utility score by 19–59%. However, utility theory argues that utility function values are important only in their rank. The percent improvement in utility is meaningless because it is sensitive to the particular utility function: an affine transformation of the utility function could have arbitrarily changed the percent improvement to any value [Barnett, 2003]. While the quantitative improvement in "utility" is difficult to determine, qualitatively, according to a function that accounted for investigators' desire for a balanced portfolio of successful flights, the algorithm was able to collect a portfolio of higher utility than that collected by DC3 investigators.

Of particular interest is the temporal distribution of the successful flights recommended by the decision algorithm. Despite being designed to collect a set of flights balanced across the three regions, the decision algorithm's first four recommended flights were all to the same region, Alabama. The first three flights

were successful. Therefore, facing a situation with 3 successful flights in Alabama and no successful flights in the other two regions, the decision algorithm still recommended flying to Alabama again, collecting a fourth successful flight in that region. We argue that this represents exactly the sort of decision challenge that is better handled by an algorithm than human decision makers. We suspect that in such a situation, humans would be averse to expending a flight in a region that already had three successful flights while having no successful flights in the other two regions. Using its dynamic programming scheme to assess expected future opportunities, the algorithm was able to conclude that ending the field season with a balanced portfolio was still possible, even with such an imbalanced start to the season. The algorithm was right, as its final portfolio was $\langle 4, 3, 4 \rangle$. The ability to quantitatively account for its own likely future decisions and outcomes enabled the algorithm to amass a useful portfolio of successful flights in a counter-intuitive way.

3.6 Alternative decision recommendation algorithms

3.6.1 Autocorrelation

One potential issue affecting both the forecasting system and decision recommendation algorithm is autocorrelation in the distribution of possible days. Meteorological intuition suggests that there is likely to be autocorrelation: good days and bad days are likely to non-randomly occur consecutively, because conditions related to the probability of isolated, deep convection depend largely on synoptic-scale weather patterns that vary on multi-day time scales. We have assumed that in the forecasting system, any autocorrelation is accounted for by the predictors. This amounts to an assumption that a string of good or bad days will be accompanied by a string of good or bad values of the predictors (model-forecast CAPE, model-forecast vertical velocity, etc.) which will lead to a string of high or low forecast probabilities, so no further adjustment is needed for the forecasting system. However, autocorrelation could threaten the decision recommendation algorithm, especially near the end of a field campaign. With only a few days remaining in the field campaign, if autocorrelation is present in the temporal distribution of "good" and "bad" days, the probability of an unfavorable synoptic-scale weather pattern persisting for the remainder of the experiment will be higher than one would expect

	$P(good \mid previous\ day\ good)$	$P(good \mid previous\ day\ bad)$
CO	0.3411	0.2029
OK	0.4709	0.3207
AL	0.4160	0.2251

Table 3.2. Probability of good conditions in each region, conditional on the objective verification of the previous day. In each region, good days are more likely following good days than following bad days, suggesting that the probability of good conditions is governed in part by processes with timescales greater than 1 day. A decision system that accounts for this effect would expect more strings of consecutive days with good or bad conditions than one that assumes each day is independent. $P(bad \mid previous\ day\ bad)$, not shown, is $1 - P(good \mid previous\ day\ bad)$

if "good" and "bad" days are not correlated.

To model the autocorrelation in the decision recommendation algorithm, we have considered a two-state Markov process [Ross, 2010]. Days from the historical dataset are divided into two partitions: days after good days, and days after bad days. The forecasts produced by the forecasting system for those days are used to generate two beta distributions of expected forecasts: one distribution of forecasts conditional on the previous day being good, and one distribution of forecasts conditional on the previous day being bad. As we expect from intuition, in all regions, $P(good \mid previous\ day\ good) > P(good)$ and correspondingly, $P(bad \mid previous\ day\ bad) > P(bad)$, as shown by Table 3.2. These conditional distributions allow for Monte Carlo simulations of future field seasons that include strings of good or bad days in a more realistic manner. Theoretically, this model could be extended to allow for multi-day autocorrelation, but the limited amount of training data makes this infeasible.

The field campaign was simulated using the original forecasting system but using the alternative decision recommendation algorithm which accounted for autocorrelation. This alternative algorithm was less successful than the original algorithm, collecting 4 good flights in Colorado, 2 good flights in Oklahoma, and 4 good flights in Alabama for a utility score of 9.39, compared to the utility score of 10.87 achieved by the original algorithm. This degraded performance could be caused by overfitting of the training data due to splitting the original sample into smaller sub-samples, or could be a reflection of worse luck on marginal days.

	$P(good \mid May)$	$P(good \mid June)$
CO	0.1924	0.2594
OK	0.3199	0.4048
AL	0.2211	0.3381

Table 3.3. Climatological variation in the probability of good conditions between May and June. In each region, good days are climatologically more likely in June than in May. A decision system that accounts for this effect would be less aggressive in expending resources during May than one that does not account for this effect.

3.6.2 Seasonal variation

Concerns were raised during the field campaign about non-stationarity in the probability of good conditions during the field season. Table 3.3 shows the climatological probability of "good" conditions in May and June. Given the different probabilities of success for each month, another alternative decision recommendation algorithm was developed. Just as the alternative algorithm accounting for auto-correlation employed a different beta distribution of possible forecasts based on the simulated results of the previous day, this alternative algorithm employed a different beta distribution based on whether the simulated day was a May day or a June day. This alternative algorithm would have collected 5 successful flights in Colorado, 2 successful flights in Oklahoma/Texas, and 3 successful flights in Alabama, for a utility score of 9.14, compared to the utility score of 10.87 achieved by the original algorithm. As shown in Table 3.3, climatology suggests that days in June are generally more promising than days in May, which the system should account for by being more picky in May, knowing that June should bring better conditions. In reality, during the period of the DC3 field season, May conditions were better than June conditions in 3 of the 5 sub-regions, as shown in Table 3.4. With only 8 field seasons worth of training data, splitting the training data into two separate pools to account for seasonal differences in climatology likely degraded the algorithm by causing over-fitting of the training data.

3.6.3 Sensitivity of beta distribution of forecasts

When simulating field campaigns in the training of the decision system, a beta curve of best fit was generated for each region from the training data. Figure 3.3 shows some discrepancies between the best-fit beta distribution curves and the

	May 2012 realized good days	June 2012 realized good days
Colorado	0.0625	0.1786
NW Texas	0.3125	0.2500
SW Oklahoma	0.2667	0.3333
C Oklahoma	0.2667	0.1724
Alabama	0.4375	0.1200

Table 3.4. Actual frequency of good days realized in each region during May and June 2012. The actual frequency of good days is computed by dividing the number of good days by the total number of days. While all three regions show higher climatological probability during June, (Table 3.3), three of the five sub-regions experienced better conditions during May 2012 than during June 2012. This deviation from climatology could explain the degraded performance from the alternate decision system that accounted for the effect of seasonal climatology variation.

empirical histogram of forecast probabilities. In the Colorado region, a peak at low probabilities is missed. In the Oklahoma/Texas region, a secondary peak at high probabilities is missed. The differences between the empirical and theoretical distributions suggest that perhaps the beta distribution is a poor choice of distribution to fit to the observed data. However, it appears that the decision system succeeded in spite of the limitations of the beta distribution for simulating the distribution of future days.

To test the sensitivity of the decision system to the quality of the theoretical distribution of forecasts, an alternate decision system was run where the Monte Carlo simulations of field seasons drew forecasts randomly from the empirical distribution of forecast probabilities rather than the theoretical distribution. This alternate decision system achieved 5 successful flights in Colorado, 2 successful flights in Oklahoma/Texas, and 4 successful flights in Alabama for a utility score of 10.03, compared to the score of 10.87 achieved by the original algorithm.

The similar results suggest that the overall decision algorithm is not very sensitive to the estimated distribution of climatology. The beta distribution offers simplicity and convenience compared to a more sophisticated method of simulating climatology.

3.6.4 Sensitivity of the elasticity of substitution parameter

Another sensitivity test was performed on the elasticity of substitution parameter, σ, in the utility function, which represented the degree to which DC3 investigators

were willing to trade portfolio balance for total amount of collected data. Based on discussion with DC3 principal investigators, σ was set to 1.7 in the original decision algorithm. Higher values of σ imply less willingness by investigators to compromise on the requirement for cross-region balance in their portfolio of data. Lower values of σ imply less stringency toward the requirement for balance; at $\sigma = 0$, flights to each region are considered interchangeable. Figure 3.4 displays the results of several alternate decision algorithms, each run with a different value of σ. On 35 of the 45 days, the same decision was recommended regardless of the chosen value of σ. Interestingly, the total number of flights collected was higher for higher values of σ, contrary to what is expected. This discrepancy illustrates that results are more sensitive to small-sample idiosyncrasies, such as marginal shifts in decisions on borderline days, than to the particular choice of the σ parameter.

3.6.5 Selection of over-forecasts in the Oklahoma/Texas region

In the original algorithm, three forecasts are generated for the three sub-regions of the larger Oklahoma/Texas region. The distance between these sub-regions and the nature of mobile radar deployment demanded that a flight only be taken to one of the three sub-regions on a given day, with the particular sub-region chosen hours before takeoff. Investigators considered a successful flight to any of the three sub-regions to be interchangeable in value and score-able as "a success in the Oklahoma/Texas region." Each day, the automated forecasting system generates a forecast for each of the three sub-regions. The sub-region with the highest forecast probability is chosen as the decision-relevant sub-region, and the decision system uses its forecast as the decision-relevant forecast probability.

The daily selection of the Oklahoma/Texas sub-region with the highest forecast probability introduces a bias, assuming the forecasting system is imperfect. In selecting the highest of the three forecast probabilities, we are drawing more "over-forecasts" than "under-forecasts," leading to forecast probabilities that are systematically too high for the Oklahoma/Texas region. We attempted two potential fixes to this issue. First, we used the forecast system's training data to scale all of the Oklahoma/Texas best-sub-region forecasts such that they match climatology. Alternatively, we employed a bin-based rescaling method, wherein forecasts in a bin were rescaled to climatology by a factor determined only by the forecasts in

		sigma						DC3
		0	0.2	0.7	1.2	1.7	3.2	team
day	3	AL	AL	AL	AL	AL	AL	CO
	5	AL	AL	AL	AL	AL	AL	no
	6	AL	AL	AL	AL	AL	AL	AL
	13	AL	AL	AL	no	no	no	no
	14	AL	AL	AL	AL	AL	AL	OK
	18	no	CO	CO	CO	CO	CO	CO
	19	no	no	CO	CO	CO	CO	no
	21	OK	OK	OK	OK	OK	OK	CO
	22	OK	CO	CO	CO	CO	CO	CO
	23	CO	CO	CO	CO	CO	CO	OTH
	26	AL	no	no	no	no	no	no
	27	AL	AL	AL	AL	no	no	AL
	28	no	no	no	no	OK	OK	no
	29	OK	OK	OK	OK	OK	OK	no
	30	no	no	no	no	OK	OK	no
	31	CO	CO	CO	CO	CO	CO	CO
	37	no	OK	no	OK	OK	OK	OTH
	43	CO	no	no	no	no	no	OTH
	44	CO	CO	CO	CO	no	no	OTH
	45	CO	CO	CO	CO	CO	CO	no
total	CO	3	3	4	4	4	4	2
	OK	2	2	1	2	3	3	3
	AL	5	5	5	5	4	4	2
	all	10	10	10	11	11	11	7

Figure 3.4. The decisions suggested by the decision algorithm as a function of the value of σ used in the algorithm's utility function. "CO," "OK," or "AL" in a cell indicates that a flight was taken to that region on that day for that decision algorithm, while "no" indicates no flight was taken that day. Good days are marked by yellow fill and bad days are marked by red fill. Days where no flight was taken regardless of the σ value (25 of 45) are omitted from the figure. Alternate decision algorithms with $\sigma = 2.2, 2.7$ (not shown) issued identical decisions to the decision algorithms with $\sigma = 1.7, 3.2$. For reference, the decisions made by the DC3 team for the same days are also shown, where "OTH" indicates a flight taken for a secondary objective.

the bin. Under each of these attempts to account for selection bias over-forecasts, the alternate method achieved 4 successful flights at Colorado, only 1 successful flight in Oklahoma/Texas, and 4 successful flights in Alabama, for a utility score of 7.31, suggesting that the attempts to correct for this bias degraded the forecast.

3.7 Summary and conclusion

Many field campaigns in the atmospheric sciences take the form where budgets include fixed amounts of resources and time and resource deployment decisions must be made under weather uncertainty before the sought-after phenomenon is present in the atmosphere. Seeking to optimize the amount of data collected from their fixed budget of resources, investigators must expend resources under weather uncertainty, with the knowledge that some resource expenditure today precludes a resource expenditure on some future day. Traditionally, all resource deployment decisions are made heuristically: investigators discuss the weather on a particular day and decide whether to expend resources. We propose an alternative method of deploying resources, relying on an automated decision recommendation algorithm to make first-order resource deployment decisions.

For the DC3 campaign, the first-order decision faced by investigators on each morning was whether or not to fly that afternoon, and if so, to which of three regions. During the field campaign, decision recommendations were offered by the decision recommendation algorithm. A retrospective re-simulation of the algorithm, assuming that the algorithm's recommendations were followed verbatim, demonstrated that the algorithm would have collected more data than the heuristic decisions made in the field. The decision recommendation algorithm offered no decision guidance below the region-scale spatially or below the day-scale temporally. For this reason, algorithmic decision-making cannot replace human decision-making. For first-order decisions, however, algorithmic decision-making has shown promise as an improvement over traditional human decision-making.

Some of the success of the decision recommendation algorithm relative to decision-makers in the field can likely be attributed to the use of a quantitative definition of "good" conditions. The definition was an approximation of the conditions sought after by DC3 investigators, but it is likely that decision-makers were looking for conditions merely close to those specified by the algorithm, while the algorithm was looking for exactly those conditions. This discrepancy is borne out by the several days during DC3 where decision-makers scored conditions in a region differently from the algorithm. The algorithm forces an explicit, quantitative definition of the conditions of interest. We argue that because an explicit definition removes ambiguity and allows for specific pre-experiment statistical analysis, its

benefits outweigh any benefits of using a heuristic, “we-know-it-when-we-see-it" definition.

Algorithmic decision-making offers particular promise for certain types of field campaigns. Field campaigns seeking rare events are especially well-suited for decision algorithms, because such campaigns demand counter-intuitive decision making. Experiments whose phenomenon of interest is complicated might find a decision algorithm particularly useful; in such field campaigns, forecasting and climatology from human forecasters might be murky. Finally, for long-duration field experiments, a decision algorithm can save a substantial amount of travel and lodging costs for decision-makers.

Following Small et al. [2011] and Hanlon et al. [2013], this study represents a third demonstration of the power of algorithmic field campaign decision-making under weather uncertainty. Algorithmic decision-making offers a slew of benefits: it reduces stress on field campaign PIs, reduces time spent on daily decision-making, reduces demand on operational weather forecasters, and forces investigators to agree on a quantitative, unambiguous, pre-experiment statement of goals. Most significantly, the resource deployment decisions recommended by a decision recommendation algorithm would have yielded more data than the decisions made using traditional heuristic decision-making in three distinct field campaigns with varying goals. We suggest that algorithmic decision-making should be considered for all field campaigns in atmospheric sciences to maximize field campaign efficiency and ensure that the greatest possible amount of data is collected for some level of funding.

Acknowledgments

The authors acknowledge financial support from National Science Foundation Atmospheric and Geospace Science Grants AGS-1063692 and AGS-1063733. We are grateful to DC3 principal investigators Mary Barth, William Brune, Chris Cantrell, and Steven Rutledge for allowing us to apply our methodology to their field campaign.

Radar data were obtained from NCDC. WRF data were obtained courtesy of Morris Weisman and Kevin Manning of the National Center for Atmospheric Research (NCAR) Mesoscale and Microscale Meteorology (MMM) Division. All data are available upon request.

Chapter 4
The automated decision algorithm for field campaign decision making: forecasting system sensitivities

4.1 Introduction

While DC3 offers the most recent example of the automated decision algorithm's applicability to weather-related decision challenges, the algorithm used for DC3 benefited from inputs that may not be available for most field campaigns. The algorithm incorporated both real-time forecast data and archived training data from a research-grade, high-resolution WRF model whose domain covered the field campaign's research area of interest. The typical field campaign may not have the luxury of such state-of-the-art forecast guidance. For example, many field campaigns conduct their missions far from the data-rich continental United States, in locations where the only available forecast guidance is relatively coarse-resolution global forecast models with lower forecast skill than what was available for the development of the automated decision algorithm used for DC3. To explore the sensitivity of the automated decision algorithm methodology to the quality of forecast information, we repeated the methods used in Hanlon et al. [2014a, 2014b] for the DC3 campaign, but substituted a low-resolution global model for the high-resolution regional model.

The NOAA Earth System Research Laboratory (ESRL) Physical Sciences Division (PSD) has generated historical reforecasts of National Center for Environmental Prediction (NCEP) Global Ensemble Forecasting System (GEFS) forecasts dating back to 1984 with 1-degree resolution [Hamill et al., 2013]. This dataset facilitates an exploration of the sensitivity of the automated forecasting system for isolated thunderstorms used for DC3. Because its resolution is much lower than the 3km WRF used in the original forecasting system, the data allows for a measure of the skill reduction in the forecasting system caused by reducing the quality of the model forecast inputs. While assessing any change in forecast skill associated with substituting the lower-resolution model, we also assess the change in decision skill. Instead of having WRF forecasts and training available to generate decision recommendations, if the automated decision algorithm only had GEFS forecasts and training available, how would its decisions change and affect field experiment outcomes?

In addition to facilitating a test of the sensitivity of the automated decision algorithm to forecast model quality, the GEFS forecasts differ from the WRF forecasts in other ways, enabling a broader exploration of the automated decision algorithm. The longer period of record of the GEFS reforecast data allows for a measure of the importance of the amount of training data to the forecasting system. While the WRF forecasts began being produced only in 2004, the period of record with both GEFS reforecasts and radar verification data dates back to 1997, approximately doubling the amount of training data available to the forecasting system. Separately, because each GEFS forecast includes ten ensemble members in addition to the control forecast, the size of the training set can be artificially increased by incorporating all of the ensemble forecasts.

One valuable feature of the automated decision algorithm is its ability to frame short-timescale decisions (day-to-day) in terms of their expected contribution to long-timescale goals (the end-of-field-experiment set of flights). The GEFS' longer-range forecasts might facilitate this ability better than the WRF's shorter-range forecasts. While WRF forecast information only included forecasts at one day's lead-time, the GEFS' forecasts extended eight days into the future. If skilled, the longer-range forecasts could improve decisions on the first day, incorporating sharper information about the expected future conditions.

As demonstrated in Hanlon et al. [2014b], the automated decision algorithm

using the 3-km WRF forecasts generated decision value greater than that generated by human forecasts and decisions. This study offers an assessment of the decision value that could be expected by a similar automated decision algorithm if the available forecast information were less accurate. This assessment will help inform field campaigners seeking to recreate the DC3 automated decision algorithm for another field campaign of the most important elements of the forecasting system. Finally, the assessment of the GEFS as applied to this methodology sets a worst-case scenario: if all an automated decision algorithm has available is a weak forecast system, how valuable should one expect its decisions to be?

4.2 Methods

4.2.1 Impact of a cruder forecast model

As a first test on the relative importance of the forecast model to the automated decision algorithm's forecast and decision skill, the same methodology was used as in Hanlon et al. [2014b], but with the GEFS' forecasts used as input to train the algorithm instead of the WRF's forecasts. To isolate the effect on the forecasting system from changing the forecast model from the high-resolution WRF to the low-resolution GEFS, the amount of training data used for the GEFS was limited to the years for which training data was available for the WRF, amounting to eight years of May and June training data, between 2004 and 2011. The same predictors were used for each model and the same logistic regression scheme, tuned by a genetic algorithm, was used to convert the predictors to probabilistic forecasts. The radar data used to verify the probabilistic predictions made by the forecast model were the same, and as in the original analysis, 15-hour to 24-hour model forecasts from 00 UTC were used. 45 days from May and June 2012, corresponding to the DC3 campaign, were used to test each forecasting system. The 2012 data were not used for training the system, and thus constitute a set of independent verification data.

4.2.2 Impact of more training data

Dating back to 1985, the GEFS offered the advantage of a larger dataset compared to the WRF. We expect that low resolution forecast models typically will have more

historical data than high resolution research-grade models, which may help offset the decreased resolution. In this case, the GEFS training data extended back to year 1985: farther back than the available radar data for verification of conditions, which only extended back to 1997–2000, depending on the radar site. Even using the data back to 1997 increased the amount of training data to 15 years of May and June data, compared to the only 8 years from the WRF. To measure the value of this extra training data, the same forecasting system methodology was used as in section 4.2.1, but incorporating all 15 years of GEFS training data, rather than only the eight years that overlapped with the WRF model.

As another test on the importance of the amount of training data, the GEFS ensemble runs were used to increase further the effective size of the training set. Each GEFS forecast offers one control run and 10 ensemble runs, using the same physics but different initial conditions. A forecasting system was developed using the same methodology as before, but using the control run and each of the ensemble runs as if each was a unique piece of training data. While the extra model runs for each day offer some additional forecast information, this is not quite the same as multiplying the amount of training data by 10 for two reasons. First, for all 10 ensemble runs from the same day, the verification data are exactly the same, yielding less information than 10 unique days of training data. Second, because the ten ensemble runs are highly correlated, the forecast data are not independent. Nonetheless, using all of the ensemble runs increases the amount of information available to the forecasting system. Instead of having $15\,years\,x\,61\,days\,=\,915$ training cases, the forecasting system has $915\,x\,11\,=\,10065$ training cases. The increased amount of training data may offer the potential for improved forecasting and decision-making.

4.2.3 Impact of the longer forecast horizon

While the original decision algorithm only provided a forecast and decision for one day and actual flight decisions only needed to be made one day in advance, the human decision-makers on the DC3 campaign were taking into account the longer-range weather forecast in their decision-making process. If skilled forecasts are available beyond the first day, incorporating these forecasts could affect day-1 decision making. For example, suppose today's forecast guidance indicates that

flying to Colorado and flying to Oklahoma are equally promising from an expected utility perspective, but tomorrow is expected to be a particularly promising day in Colorado. Because DC3 decision-makers wanted to have a balance of successful flights among the three study regions, the added information that tomorrow looks promising in Colorado could nudge today's decision towards Oklahoma. Because the WRF only offered 24 hours' worth of forecast information, the original automated decision algorithm assumed any days past day 1 would follow climatology; the algorithm added no skill to the forecasts beyond day 1. The GEFS reforecasts, however, offered several days of forecasts, allowing for extended forecasts to be incorporated into the decision algorithm.

To include the extended forecasts in decision-making, an alternate decision algorithm was developed. The original methodology was used for forecasting, but was applied to model forecasts for days 2 and 3 in addition to day 1. The optimization module was modified to recommend a flight decision for each of the upcoming three days. Instead of choosing the decision today that maximizes the expected end-of-season utility, the system chooses the set of three decisions for the next three days that maximizes the end-of-season utility. Because there are four possible options each day (fly to CO, fly to OK, fly to AL, or do not fly), the number of possible three-day decisions is $4^3 = 64$. While three days' worth of decisions are recommended on the first day, only the first day's decision is binding. On day two, the process repeats, so that another three-day set of decisions is made using the current forecast data and the previous day's decision recommendations are no longer considered.

4.2.4 Skill assessment

The Brier skill score (BSS) [Brier, 1950] is used to assess the skill of the probabilistic forecasts. The BSS can be interpreted as a percent improvement in skill compared to a baseline unskilled forecast of the mean squared error of a set of forecasts. The climatological probability of good conditions in each region was used as a BSS baseline. A positive BSS indicates a set of forecasts was more skilled than climatology, while a negative BSS indicates a set of forecasts was less skilled than climatology.

While the BSS assesses forecast skill, the associated decision skill is a more

direct measure of the value of the forecasting system to the end-user. Analogous to the BSS, we introduce a crude method to assess decision skill relative to climatology. Imagine a no-skill decision system with the following format: of the 15 research flights allotted to scientists, 5 are used in each of the three regions of interest on randomly chosen days. For a field campaign with no forecast or decision support, such a "Clueless Decision System" making decisions indiscriminately might be the best available decision system. How many successful flights, and in which regions, would such a system have achieved during the 2012 field season? A Monte Carlo simulation was conducted to randomly distribute flights throughout the field season. The "utility score" of any set of flights was measured according to equation 1 from Hanlon et al. [2014b], a custom-built equation that approximates the specific goals of DC3 decision-makers. Over a large number of simulations, the median "Clueless Decision Maker" would have achieved a utility score of 2.07, corresponding to 2 successful flights in one region and 1 successful flight in another region. Any decision system that generates a utility score greater than 2.07 increases decision value over an indiscriminate decision-maker.

4.3 Results

4.3.1 Impact of a cruder forecast model

Leaving all else equal and using eight years of training data but switching the WRF with the GEFS lowers the BSS from 0.29 to 0.12, averaged over all regions. As expected, using the coarser forecast model as input yields worse forecasts. The coarse model, however, still exhibits forecast skill better than climatology, suggesting that it offers some useful information for decision-making. According to Hanlon et al. [2014a], using one method of converting human forecasts to the same scale as the automated forecasts, the human forecasts averaged a 0.24 BSS. The GEFS-based forecasting system, therefore, was worse than human forecasters but better than climatology.

The GEFS-based automated decision algorithm would have collected 3 successful flights in Colorado, 5 successful flights in Oklahoma, and 0 successful flights in Alabama, yielding a utility score of 3.96, compared to the 10.87 utility score achieved by the WRF-based decision algorithm. While the decision algorithm using the

GEFS performed much worse than the algorithm using the WRF, the algorithm greatly outperformed the "Clueless Decision Maker." The utility score achieved by humans was either 6.84 or 9.14, depending on the scoring method [Hanlon et al., 2014a]. Therefore, the recommendations offered by the GEFS-based decision algorithm, using only coarse forecast information, would have been markedly less valuable than those made by human experts, but markedly more valuable to a field campaign than a no-skill decision algorithm. A hypothetical field campaign with no skilled forecasting or decision support available would be markedly improved using even this rudimentary system.

4.3.2 Impact of more training data

Including the extra seven years of training data from 1997 through 2003 actually slightly degraded the forecasting, lowering the BSS from 0.12 to 0.07. The automated decision algorithm with the extra training data collected 3 successful flights in CO, 3 successful flights in OK, and 1 successful flight in AL. Compared to the system with less training data from section 4.3.1, fewer successful flights are achieved but because the set of successful flights is better balanced across the three regions, the utility score increased from 3.96 to 6.16.

The forecasting system that incorporated each of the ensemble forecasts as a distinct case of training data performed similarly to the forecasting system that only incorporated the control run, achieving a BSS of 0.08, compared to 0.07 for the control. The decisions generated by this version of the automated decision algorithm would have yielded 3 successful flights in CO, 3 in OK, and 0 in AL for a utility score of 3.44. The utility score is markedly lower than the other variations of the automated decision algorithm because of the steep penalty associated with getting 0 successful flights in AL instead of 1.

4.3.3 Impact of longer forecast horizon

The forecast days beyond forecast day 1 actually were slightly more skillful than the day-1 forecasts. Both day-2 and day-3 forecasts had an average BSS of 0.10, compared to 0.07 for the day-1 forecasts. The decisions generated by this version of the automated decision algorithm would have yielded 3 successful flights in CO, 3 in OK, and 1 in AL for a utility score of 6.16. This set of successful flights was

the same set of successful flights achieved by the automated decision algorithm without ensembles in section 4.3.2. Despite the skilled forecasts extending multiple days, the decisions did not improve, suggesting that the value offered by the extra forecast days was marginal.

4.4 Discussion and conclusion

The differences between each of the variations of the GEFS-based automated decision algorithm amount to a few days' worth of different decisions over the course of the field season. These differences could be explained by noise: one or two different decisions on borderline days could affect the final utility score. Increasing the amount of training data improved the decision algorithm but made the forecasting system slightly worse on independent testing data. While skilled forecasts are available beyond the first day, they do not markedly improve the decision algorithm. The overall picture is consistent: each variation of the automated decision algorithm using the GEFS data as forecast input performs markedly worse than human decision makers but markedly better than an unskilled decision system.

The message from the original application of the automated decision algorithm to the DC3 field campaign was that a carefully constructed algorithm with a good forecasting model as input can make better decisions than human decision-makers. This study has a different message: that a carefully constructed algorithm with a poor forecasting model as input can still make passable decisions. Experimentation to alter the amount of training data and the length of the forecast horizon all yielded broadly similar results: that while a coarse forecast model did not forecast as well as a high-resolution model, it could still be trained to forecast with skill for phenomena much smaller than the resolution of the model. More importantly, the forecasts generated by such a model showed the ability to make decisions yielding much more value than those made randomly.

Human intuition can offer great value to the decision-making during a field campaign, especially one with a forecast-and-decision problem related to thunderstorms, where the field of meteorology is flush with expert knowledge. DC3 weather forecasters included some of the world's experts on deep convection forecasting in regions with which they were intimately familiar and the DC3 decision team was a cohesive unit of experienced field campaigners. However, one can imagine field

campaigns with forecast and decision problems where human forecasters do not add much value. Perhaps the forecast problem is too esoteric to attract the interest of the best human forecasters, or the forecast problem is sufficiently complicated that humans struggle to calibrate probabilistic forecasts. For such forecast problems, there may be no human experts to make decisions during a field campaign interested in studying that phenomenon, making these field campaigns difficult to operate efficiently. By offering even limited forecast-and-decision skill, this alternative methodology could improve the efficiency of such field campaigns, enabling more atmospheric research to be conducted.

Chapter 5
Conclusion: future applications of the automated decision algorithm

Following the successful development of automated decision algorithms for the RACORO and SPartICus field campaigns, the automated decision algorithm developed for the DC3 campaign further demonstrates the usefulness of such algorithms in atmospheric science field campaigns. Despite being faced with a complex decision problem and competing against skilled human forecasters, the automated decision algorithm recommended decisions that would have offered more scientific value to the field campaign than the decisions made by DC3 human decision-makers. Atmospheric scientists planning a field campaign can incorporate this methodology into their decision process to optimize their data collection.

While the methodology has been shown to be valuable to atmospheric science field campaigns, the next step for the automated decision algorithm should be to further explore its applicability to other decision problems under uncertainty, both in meteorology and in other geophysical fields. The general method employed by the automated decision algorithm is to define a utility function, develop a forecasting system, and recommend utility-maximizing decisions conditional on the output from the forecasting system. This method could be applied to geophysical decision problems outside of the relatively niche area of atmospheric science field campaigns.

Because development of a sophisticated decision algorithm requires time, money and expertise, it is not suitable for every decision under geophysical uncertainty. However, for individuals, groups, or businesses with high-stakes decisions contingent

on the outcomes of geophysical events with some degree of predictability, the development costs of a decision algorithm might be worthwhile. Some degree of predictability is possible for events in many geophysical fields, including meteorology, oceanography, volcanology, hydrology, seismology, or others. Practitioners exposed to uncertainty from events in these fields should note the success demonstrated by the automated decision algorithms developed for the DC3 campaign and consider whether they can improve the efficiency of their own decision processes by following the same formula.

To build an algorithm similar to the one implemented for decision recommendations in the DC3 campaign, decision-makers need to define a utility function, develop a forecasting system tuned to the utility function, and make decisions conditional on the forecast such that utility is maximized. This general method has been demonstrated to add value to three distinct atmospheric science field campaigns, and its value to a broader set of decision problems should be explored.

Appendix
Optimization module details

This technical description of the optimization module originally appeared as an appendix in the Journal of Geophysical Research: Atmospheres as Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign *with authors Christopher J. Hanlon, Arthur A. Small, Satyajit Bose, George S. Young, Johannes Verlinde (Chapter 3 of this dissertation)*

The decision faced by investigators on each day of the DC3 campaign requires comparing the cost of expending a flight to the expected benefit achieved by flying. Let the current state of the experiment be represented by

$$J(d, f, \langle s_1, s_2, s_3 \rangle) = U(\langle s_1, s_2, s_3 \rangle) + V(d, f, \langle s_1, s_2, s_3 \rangle) \tag{A.1}$$

where U is the utility function that assigns a utility value to any combination of flight successes across the three regions, V is the expected future utility starting from state $(d, f, \langle s_1, s_2, s_3 \rangle)$, d is the number of days remaining in the experiment, f is the number of flights remaining in the experiment, and $s = \langle s_1, s_2, s_3 \rangle$ is the current number of successful data-collection flights to regions 1, 2, and 3. The cost of flying is a decrease in V. V is always lower if there are $f - 1$ flights in the budget rather than f flights in the budget because having fewer flights always makes future data-collection prospects less promising. The benefit of flying is an expected increase in U. If no flight occurs today, there will be no change in U; provided the probability of flight success is greater than zero, the expenditure of a flight increases U in expectation. The optimal flight decision on each day, either flying to one of the three regions or not flying, is the one that maximizes J. If the probability of flight success in some region is high enough that the expected

increase in U exceeds the expected decrease in V, then flying to that region is preferable to not flying.

Informing flight decisions requires knowing the expected value of V for any combination of $(d, f, \langle s_1, s_2, s_3 \rangle)$. Given an expected value of V for any possible experiment state and the probability of flight success, we can calculate exactly how much expected future utility is decreased by any decision. We employ dynamic programming [Bellman, 1957] to solve for all possible values of V, arranged in the form of a three-dimensional lattice, where the three dimensions are time (number of days remaining), resources (number of flights remaining), and successes (the current set of flight successes). Figure A.1 shows an example of the lattice that ignores, for clarity of presentation, the "successes" dimension. The value of the bottom nodes of the lattice are given by a boundary condition:

$$V(f = 0) = 0. \tag{A.2}$$

If the number of flights is equal to zero, then V is equal to zero, no more flights will be taken, and the state will move horizontally across the bottom row of the lattice until the end of the experiment.

To incorporate the "successes" dimension of the lattice and fill in the lattice above the bottom row, we use a Monte Carlo simulation of 10000 field seasons, where each day's simulated meteorology can be expressed by three probabilities: the probability of optimal flight conditions in each of three regions, (P_1, P_2, P_3). Filling in the lattice starts with the node marked by a diamond at the bottom right of the lattice, where $(d, f) = (1, 1)$. From the boundary condition (A.2), we know that

$$V(d = 0, f = 0, s) = 0 \tag{A.3}$$

for all $s = \langle s_1, s_2, s_3 \rangle$.

We assume also that

$$V(d = 0, f = 1, s) = 0 \tag{A.4}$$

for all s.

The implication of equation (A.4) is intuitive: if there is one flight and one day remaining in the experiment, there is no benefit to saving the flight because there will be no future opportunities. The flight should always be taken because the state

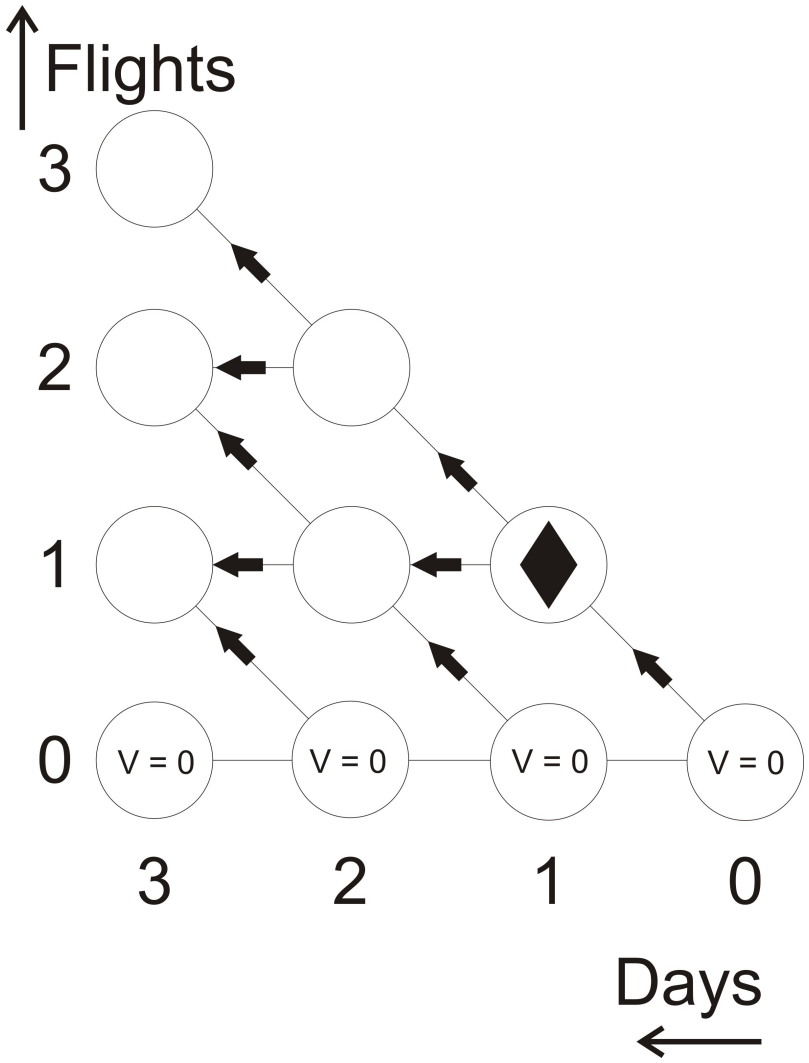

Figure A.1. A figure illustrating the dynamic programming procedure used to estimate the expected value associated with any experiment state. The value at the nodes on the bottom row is zero, because the number of flights in the budget is zero. Arrows in the figure represent dependency: nodes dependent on other nodes will have arrows pointing to them from those nodes. The value at any node on the figure can be calculated recursively if the value of all nodes "pointing" to it are known. The recursive filling of the decision tree begins with the node marked with the diamond. The third dimension of this figure, corresponding to the number and distribution of successes, is omitted for clarity but is analogous to the two dimensions displayed here.

with 0 days and 1 flight has no value. Therefore, under the condition where $d = f$, a flight will be undertaken to the region that provides the greatest expected increase to the current utility score, U. Because a flight will always occur if the number of flights equals the number of days remaining, any experiment state (d, f, s) where $f > d$ is unreachable. From state $(d = 1, f = 1)$, the region that provides the greatest expected increase to the utility score is a function of the current set of successes s and the probability of optimal conditions in each of the three regions. By simulating the probability of optimal conditions 10000 times, we can simulate the decision made at $(d = 1, f = 1)$ 10000 times for each possible value of s. These 10000 simulations allow for the estimation of $P(no\,fly)$; $P(fly\,to\,i), i = 1, 2, 3$; and $P(good|fly\,to\,i), i = 1, 2, 3$; all for each possible state $(d = 1, f = 1, s)$, assuming optimal decision-making.

The estimated probabilities from the Monte Carlo simulations yield a general equation, which recursively estimates V for every node in the lattice in reverse time. For any node $(d, f, \langle s_1, s_2, s_3 \rangle)$, V can be partitioned into a first-day portion and a second-day-and-forward portion:

$$V(d, f, s) = V_t + V_f. \tag{A.5}$$

The first-day portion of V_t is given by

$$\begin{aligned} V_t = \frac{1}{10000} \sum_{i=1}^{10000} & a_{1,i} P_{1,i} [U(< s_1 + 1, s_2, s_3 >) - U(< s_1, s_2, s_3 >)] + \\ & a_{2,i} P_{2,i} [U(< s_1, s_2 + 1, s_3 >) - U(< s_1, s_2, s_3 >)] + \\ & a_{3,i} P_{3,i} [U(< s_1, s_2, s_3 + 1 >) - U(< s_1, s_2, s_3 >)] \end{aligned} \tag{A.6}$$

where $P_{j,i}$ is the probability of suitable data-collection conditions in region j in simulation i and $a_{j,i}$ is a binary variable taking a value of 1 if a flight is made to region j in simulation i and a value of 0 if a flight is not made to region j in simulation i. The second-day-and-forward portion V_f is given by

$$
\begin{aligned}
V_f = P(fly\,to\,1)P(success|fly\,to\,1)V(d-1, f-1, <s_1+1, s_2, s_3>)+ \\
P(fly\,to\,1)P(failure|fly\,to\,1)V(d-1, f-1, <s_1, s_2, s_3>)+ \\
P(fly\,to\,2)P(success|fly\,to\,2)V(d-1, f-1, <s_1, s_2+1, s_3>)+ \\
P(fly\,to\,2)P(failure|fly\,to\,2)V(d-1, f-1, <s_1, s_2, s_3>)+ \\
P(fly\,to\,3)P(success|fly\,to\,3)V(d-1, f-1, <s_1, s_2, s_3+1>)+ \\
P(fly\,to\,3)P(failure|fly\,to\,3)V(d-1, f-1, <s_1, s_2, s_3>)+ \\
P(no\,fly)V(d-1, f, <s_1, s_2, s_3>).
\end{aligned}
\tag{A.7}
$$

This general V equation can recursively calculate V at any node meeting the following conditions: (a) V has already been calculated for the node on the diagonal to its right and down (i.e., $V(d-1, f-1)$ is known for all s); and (b) either V has already been calculated for the node to its right (i.e., $V(d-1, f)$ is known for all s) or $d = f$, so that $P(no\,fly) = 0$ and so the last term in equation (A.7) is 0. Starting with the boundary conditions in equations (A.2), (A.3), and (A.4), V can therefore be calculated recursively using equations (A.5), (A.6), and (A.7) for each node in the lattice.

Once V has been calculated for all d, f, and s, daily flight recommendations can be made by evaluating the expected impact on J associated with each of the four possible decisions (fly to region 1, fly to region 2, fly to region 3, or do not fly). Given a set of operational forecast probabilities $P_{op} = \langle P_{op1}, P_{op2}, P_{op3} \rangle$ for a state $(d, f, \langle s_1, s_2, s_3 \rangle)$, the expected ΔJ for a flight to region 1 is

$$
E[\Delta J_1] = E[\Delta U_1] + E[\Delta V_1] \tag{A.8}
$$

where

$$
E[\Delta U_1] = (P_{op1})[U(\langle s_1 + 1, s_2, s_3 \rangle) - U(\langle s_1, s_2, s_3 \rangle)] \tag{A.9}
$$

and

$$
\begin{aligned}
E[\Delta V_1] = (P_{op1})[V(d-1, f-1, \langle s_1+1, s_2, s_3 \rangle) - V(d, f, \langle s_1, s_2, s_3 \rangle)]+ \\
(1 - P_{op1})[V(d-1, f-1, \langle s_1, s_2, s_3 \rangle) - V(d, f, \langle s_1, s_2, s_3 \rangle)].
\end{aligned}
\tag{A.10}
$$

Equations for $E[\Delta J_2]$ and $E[\Delta J_3]$ follow the same form, but with $\langle s_1+1, s_2, s_3\rangle$ replaced with $\langle s_1, s_2+1, s_3\rangle$ and $\langle s_1, s_2, s_3+1\rangle$, respectively. Finally, the expected ΔJ associated with not flying is

$$E[\Delta J_{no}] = V(d-1, f, \langle s_1, s_2, s_3\rangle) - V(d, f, \langle s_1, s_2, s_3\rangle). \tag{A.11}$$

On any day during the field campaign, the optimal decision recommendation is to choose the option that yields the greatest $E[\Delta J]$. During the DC3 campaign, decision-makers were presented each day with the probabilities of optimal conditions generated by our forecast system for each region (i.e., P_{op}); the *hurdle probability* for each region, defined as the break-even probability P_{opi} at which decision-makers should be indifferent between a flight to region i and not flying; and the recommended decision. The hurdle probability HP_i is calculated for each region i by setting $E[\Delta J_i] = E[\Delta J_{no}]$ and solving for P_{opi}. For example,

$$HP_1 = \frac{V(d-1,f,\langle s_1,s_2,s_3\rangle) - V(d-1,f-1,\langle s_1,s_2,s_3\rangle)}{U(\langle s_1+1,s_2,s_3\rangle) - U(\langle s_1,s_2,s_3\rangle) + V(d-1,f-1,\langle s_1+1,s_2,s_3\rangle) - V(d-1,f-1,\langle s_1,s_2,s_3\rangle)}. \tag{A.12}$$

The equations for HP_2 and HP_3 follow the same form, but with $\langle s_1+1, s_2, s_3\rangle$ replaced with $\langle s_1, s_2+1, s_3\rangle$ and $\langle s_1, s_2, s_3+1\rangle$, respectively. While the hurdle probability is not a directly decision-relevant measure in this decision context, it was included in the daily recommendations as an intuitive illustration of the daily cost-benefit decision faced in each region. As flight successes are accrued in one region, the hurdle probability for that region increases to shift resources towards regions with fewer flight successes.

Bibliography

[1] Applequist, S., G.E. Gahrs, R.L. Pfeffer, X.F. Niu, 2002: Comparison of Methodologies for Probabilistic Quantitative Precipitation Forecasting. *Wea. Forecasting*, 17(4), 783-799.

[2] Arrow, K.J., H.B. Chenery, B.S. Minhas, and R.M. Solow, 1961: Capital-labor substitution and economic efficiency. *Rev. Econ. Stat.*, 225-250, doi:10.2307/1927286

[3] Barnett, W., 2003: The modern theory of consumer behavior: Ordinal or cardinal?. *Quart. J. Austrian Econ.*, 6(1), 41-65.

[4] Barth, M., W. Brune, C. Cantrell, S. Rutledge, 2012: *Deep Convective Clouds and Chemistry (DC3) Operations Plan.* [Available at http://www.eol.ucar.edu/projects/dc3/documents/DC3_Operations_Plan_4_Apr_2012.pdf].

[5] Bellman, R., 1957: *Dynamic Programming*. Princeton University Press, 339 pp.

[6] Berger, J.O., 1993: *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, 617 pp.

[7] Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78, 1-3.

[8] Dasgupta, S., C.H. Papadimitriou, U.V. Vazirani, 2006: *Algorithms*, McGraw-Hill Higher Education, Boston, 169.

[9] Epstein, E.S., 1969: Stochastic dynamic prediction. *Tellus A*, 21(6).

[10] Glahn, H.R., 1964: The use of decision theory in meteorology. *Mon. Wea. Rev.*, 92(9), 383-388.

[11] Glahn, H.R. and D.A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, 11(8), 1203-1211.

[12] Gleeson, T.A., 1960: A prediction and decision method for applied meteorology and climatology, based partly on the theory of games. *J. Meteor.*, 17(2), 116-121.

[13] Gneiting, T. and A.E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, 310(5746), 248-249.

[14] Gringorten, I.I., 1950: Forecasting by statistical inferences. *J. Meteor*, 7(6), 388-394.

[15] Hable, M., C. Meisenbach, and G. Winkler, 2002: Economically optimised power dispatch in local systems using evolutionary algorithms and dynamic programming. *Power System Management and Control, Fifth International Conference on (Conf. Publ. No. 488)*, 174-179, doi:10.1049/cp:20020030

[16] Hamill, T. M. and S.J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, 125(6), 1312-1327.

[17] Hanlon, C.J., J.B. Stefik, A.A. Small, J. Verlinde, G.S. Young, 2013: Statistical decision analysis for flight decision support: the SPartICus campaign. *J. Geophys. Res. Atmos.*, doi:10.1002/jgrd.50237.

[18] Hanlon, C.J., G.S. Young, J. Verlinde, A.A. Small, S. Bose, 2014a: Probabilistic forecasting for isolated thunderstorms using a genetic algorithm: the DC3 campaign. *J. Geophys. Res. Atmos.*, doi:10.1002/2013JD020195.

[19] Hanlon, C.J., A. A. Small, S. Bose, G. S. Young, and J. Verlinde, 2014b: Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign. *J. Geophys. Res. Atmos.*, doi:10.1002/2014JD021922.

[20] Haupt, R.L. and S.E. Haupt, 2004: *Practical Genetic Algorithms*, John Wiley, Hoboken, N.J., 272 pp.

[21] Katz, R.W. and M. Ehrendorfer, 2006: Bayesian approach to decision making using ensemble weather forecasts. *Wea. Forecasting*, 21(2), 220-231.

[22] Klein, W.H., B.M. Lewis, I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, 16(6), 672-682.

[23] Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.*, 35(9), 2739-2750.

[24] Leith, C.E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, 102(6), 409-418.

[25] Mace, J., E. Jensen, G. McFarquhar, J. Comstock, T. Ackerman, D. Mitchell, X. Liu, T. Garrett, 2009: *SPartICus: Small Particles in Cirrus Science and Operations Plan.*[Available at http://www.arm.gov/publications/programdocs/doe-sc-arm-10-003.pdf].

[26] Marsh, P.T., J.S. Kain, V. Lakshmanan, A.J. Clark, N.M. Hitchens, J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, 27(2), 531-538

[27] Marzban, C., S. Sandgathe, E. Kalnay, 2006: MOS, perfect prog, and reanalysis. *Mon. Wea. Rev.*, 134(2), 657-663.

[28] McFadden, D., 1963: Further results on CES production functions. *Rev. Econ. Stud.*, 30(2), 73-83, doi:10.2307/2295804

[29] Metropolis, N. and S. Ulam, 1949: The Monte Carlo Method. *J. Amer. Statist. Assoc.*, 44, 335-341, 10.2307/2280232

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

[31] Nilim, A., L. El Ghaoui, M. Hansen, V. Duong, 2001: Trajectory-based air traffic management (TB-ATM) under weather uncertainty. *Proc. 4th USA/EUROPE ATM R&D Seminar*, 64-72.

[32] Pesaran, M. H., and S. Skouras, 2002: Decision-based methods for forecast evaluation. *A companion to economic forecasting*, 241-267.

[33] Raftery, A.E., T. Gneiting, F. Balabdaoui, M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133(5), 1155-1174.

[34] Reap, R.M. and D.S. Foster, 1979: Automated 12–36 Hour Probability Forecasts of Thunderstorms and Severe Local Storms. *J. Appl. Meteor.*, 18, 1304-1315.

[35] Romine, G.S., C.S. Schwartz, C. Snyder, J.L. Anderson, M.L. Weisman, 2013: Model Bias in a Continuously Cycled Assimilation System and Its Influence on Convection-Permitting Forecasts. *Mon. Wea. Rev.*, 141, 1263Ű1284. doi:10.1175/MWR-D-12-00112.1

[36] Ross, S.M., 2010: *Introduction to Probability Models*, 10th Ed., 784 pp., Elsevier, Boston.

[37] Savage, L.J., 1951: The theory of statistical decision. *J. Amer. Statist. Assoc.*, 46(253), 55-67

[38] Sobash, R.A., J.S. Kain, D.R. Bright, A.R. Dean, M.C. Coniglio, S.J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, 26(5), 714-728.

[39] Small, A.A., J.B. Stefik, J. Verlinde, N.C. Johnson, 2011: The Cloud Hunter's Problem: An Automated Decision Algorithm to Improve the Productivity of Scientific Data Collection in Stochastic Environments. *Mon. Wea. Rev.*, 139, 2276-2289, doi:10.1175/2010MWR3576.1.

[40] Theis, S.E., A. Hense, U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorol. Appl.*, 12(3), 257-268.

[41] Thompson, J.C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. emphMon. Wea. Rev., 78(7), 113-124.

[42] Thompson, J.C., 1962: Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteor.*, 1(1), 13-17.

[43] Uzawa, H., 1962: Production Functions with Constant Elasticities of Substitution. *Rev. Econ. Stud.* 29(4): 291-299, doi:10.2307/2296305.

[44] Vislocky, R.L. and G.S. Young, 1989: The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting*, 4(2), 202-209.

[45] Vogelmann. A., G. McFarquhar, J. Ogren, D. Turner, J. Comstock, G. Feingold, C. Long, H. Jonsson, A. Bucholtz, D. Collins, G. Diskin, H. Gerber, P. Lawson, R. Woods, E. Andrews, H.-J. Yang, C. J. Chiu, D. Hartsock, J. Hubbe, C. Lo, A. Marshak, J. Monroe, S. McFarlane, B. Schmid, J. Tomlinson and T. Toto, 2012: RACORO Extended-Term, Aircraft Observations of Boundary Layer Clouds. *Bull. Amer. Meteor. Soc.*, 93, 861-878. doi:10.1175/BAMS-D-11-00189.1

[46] Weisman, M.L. and J.B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev.*, 110, 504-520.

[47] Weisman, M.L. and J.B. Klemp, 1986: Characteristics of isolated convective storms. *Mesoscale Meteorology and Forecasting*. P.S. Ray, Ed., Amer. Meteor. Soc., Boston, 353-354.

[48] Weisman, M.L., C. Davis, K.W. Manning, J.B. Klemp, 2008: Experiences with 0–36-h Explicit Convective Forecasts with the WRF-ARW Model. *Wea. Forecasting*, 23, 407-437, doi:10.1175/2007WAF2007005.1.

[49] Wilks, D.S. and D.W. Wolfe, 1998: Optimal use and economic value of weather forecasts for lettuce irrigation in a humid climate. *Agric. For. Meteorol.*, 89, 115-130.

[50] Williams, J.K., C.J. Kessinger, J. Abernethy, S. Ellis, 2008: Fuzzy logic applications. *Artificial Intelligence Methods in the Environmental Sciences*. A. Pasini, C. Marzban, S.E. Haupt, Eds., Springer, New York, NY, 347-377.

[51] Witten, I.H. and E. Frank, 2005: *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, CA, 145.

[52] Xie, S., et al., 2010: Clouds and more: ARM Climate Modeling Best Estimate Data. *Bull. Amer. Meteor. Soc.*, 91, 13-20.

[53] Young, G.S., S.D. Goldberger, J. Verlinde, C.J. Hanlon, 2015: Forecasting regional chance of occurrence through aggregation of MOS PoPs. *J. Operational Meteor.*, 3 (4), 3040.

Vita

Christopher Hanlon

Christopher Hanlon was born on 19 July 1989 to Charles and Nancy Hanlon of Drexel Hill, PA. He attended St. Andrew School in Drexel Hill from kindergarten through eighth grade and Monsignor Bonner High School in Drexel Hill. Chris enrolled at Penn State University in 2007 in the Schreyer Honors College. Chris graduated in 2011, majoring in meteorology and energy, business and finance and completing the honors thesis, "Verifying an Automated Decision Algorithm Used for Flight Decisions in the SPartICus Campaign". Chris continued his research in the Penn State meteorology department, beginning graduate school in 2011.

Beginning in July 2014, Chris works in Philadelphia for reinsurance intermediary JLT Re as a Catastrophe Modeling Consultant.

Publications

- Hanlon, C.J., J.B. Stefik, A.A. Small, J. Verlinde, G.S. Young, 2013: Statistical decision analysis for flight decision support: the SPartICus campaign. J. Geophys. Res. Atmos., doi:10.1002/jgrd.50237
- Hanlon, C.J., G.S. Young, J. Verlinde, A.A. Small, S. Bose, 2014: Probabilistic forecasting for isolated thunderstorms using a genetic algorithm: the DC3 campaign. J. Geophys. Res. Atmos., doi:10.1002/2013JD020195
- Hanlon, C.J., A.A. Small, S. Bose, G.S. Young, J. Verlinde, 2014: Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign. J. Geophys. Res. Atmos., doi:10.1002/2014JD021922

Conference presentations

- AMS 92nd Annual Meeting, New Orleans, LA, 24 January 2012: "Using Statistical Decision Analysis to Make Better Flight Decisions: The SPARTICUS Campaign"(C. Hanlon, J. Stefik, A. Small, J. Verlinde, G. Young)
- AMS 93rd Annual Meeting, Austin, TX, 9 January 2013: "Probabilistic forecasting for isolated thunderstorms using a genetic algorithm: the DC3 campaign" (C. Hanlon, G. Young, J. Verlinde, A. Small, S. Bose)
- AGU Fall Meeting, San Francisco, CA, 12 December 2013: "Automated decision algorithm applied to a field experiment with multiple research objectives: the DC3 campaign" (C. Hanlon, A. Small, S. Bose, G. Young, J. Verlinde)
- AMS 94th Annual Meeting, Atlanta, GA, 4 February 2014: "Algorithmic decision-making under weather uncertainty in atmospheric science field campaigns: a summary" (C. Hanlon, A. Small, J. Verlinde, G. Young)