Data-Driven Detection of En-route Convective Weather Avoidance and Development of a Weather Assessment Model

by

Rachel E. Price

S.B., Massachusetts Institute of Technology (2018)
S.M., Massachusetts Institute of Technology (2020)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN AERONAUTICS AND ASTRONAUTICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by:	Rachel E. Price Department of Aeronautics and Astronautics August 15, 2023
Certified by:	R. John Hansman T. Wilson Professor of Aeronautics and Astronautics, MIT Thesis Supervisor
Certified by:	Hamsa Balakrishnan William E. Leonhard (1940) Professor, MIT Thesis Committee Member
Certified by:	Tom Reynolds Group Leader, Air Traffic Control Systems, MIT Lincoln Laboratory Thesis Committee Member
Accepted by:	Jonathan P. How R. C. Maclaurin Professor in Aeronautics and Astronautics Chair, Graduate Program Committee

Data-Driven Detection of En-route Convective Weather Avoidance and Development of a Weather Assessment Model

by

Rachel Price

Submitted to the Department of Aeronautics and Astronautics
on August 15, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Aeronautics and Astronautics

Abstract

Convective weather has significant impacts on aviation operations around the world. It reduces available airspace capacity, and avoidance of hazardous convective systems increases the workload for both pilots and air traffic controllers. Optimizing airspace use in the presence of constantly evolving convective activity is an ongoing research effort. Today, convective weather avoidance models aid controllers in safely and efficiently managing their airspace. These models were built on past examples of aircraft-weather interactions drawn from archived weather observation data and air traffic data. However, the development of existing models was limited by the significant amount of work required to assess convective weather impacts on the final trajectory of each aircraft. This work explores data-driven approaches to weather avoidance modeling and introduces an image-based modeling paradigm.

At present, there is limited ability to examine historical aviation operations and automatically identify when aircraft deviated due to convective weather blocking the intended route. In particular, retrospectively assessing aircraft-weather interactions without the aid of human subject matter experts poses a difficult challenge. Each interaction is different, and simple heuristics have proven to be ineffective at classifying weather encounters. A key contribution of this work is the development of a deviation detection model, which takes in information about a flight's flown route and its intended route as well as local weather observation data and assigns a behavioral classification. Individual cases are classified as non-deviations, deviations for tactical weather avoidance, or deviations unlikely to be caused by the tactical weather situation. This classifier was developed using a supervised machine learning approach, with training data labeled by crowd-sourced subject matter experts, generally airline pilots. By leveraging transfer learning techniques, a small labeled dataset was sufficient to train the deviation detection classifier.

After the deviation detection classifier was developed, it was used to label a large set of aircraft-weather interactions. The goal of this step was to identify interactions containing useful information on weather avoidance decisions and techniques, which was previously a process carried out by human subject matter experts. This included

both non-deviations where the aircraft penetrated convective weather and tactical weather deviations where the aircraft maneuvered to avoid convective weather. The large set of non-deviations and tactical weather deviations was used to develop a deviation prediction model. For the deviation prediction model, only the intended route and local weather observations are considered. The model outputs a prediction as to whether the intended route will be flown or if there will be a deviation for convective weather.

The predictive model developed in this work outperforms existing convective weather avoidance models, and was able to do so based off of an entirely machine-labeled training dataset. This demonstrated the value of using a small dataset to effectively label a much larger dataset, as well as the image classification framework used for developing the deviation detection classifier. Reducing reliance on human data curation and labeling makes the creation of more specialized predictive models feasible and allows future research to leverage techniques requiring larger datasets than were previously possible.

Thesis Supervisor: R. John Hansman
Title: T. Wilson Professor of Aeronautics

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Acknowledgments

This work would not have been possible without the generous and dedicated backing of a great many people.

First and foremost, I want to thank my committee members, Prof. Hansman, Prof. Balakrishnan, and Dr. Reynolds. I am very grateful for your guidance, expertise, and patience throughout this process. Prof. Balakrishnan, I have greatly appreciated your encouragement and support over the past few years. Dr. Reynolds, thank you especially for coordinating the use of Lincoln Lab resources, as well as the feedback you provided me during the writing process. Finally, Prof. Hansman, I consider myself very fortunate to have been your advisee as an undergraduate, and even more fortunate that you became my graduate advisor. I have learned more about aviation from you than I ever thought possible.

I would also like to thank the thesis readers, Prof. Ravela and Dr. Wolfson. Your feedback and questions were especially helpful for bringing the thesis story together.

I would like to acknowledge the generous support of MIT Lincoln Laboratory, which provided data, computational resources, and technical expertise. Many members in Group 43 were instrumental in providing access to the weather and flight data. Likewise, the Lincoln Lab Supercomputer (LLSC) is an incredible asset, and I am grateful to the staff at the LLSC for providing such a high-quality resource. Finally, Mr. Mike Matthews contributed a great deal of technical expertise throughout the entire process, and I greatly appreciate his assistance.

This work relied on the contributions over 30 expert pilots, who participated in a study to assess aircraft behavior in the presence of convective weather. Additionally, Dr. Jensen and Mr. Hahn were indispensable in helping to recruit our study participants.

Finally, to my friends and family: thank you for your unwavering support for my entire educational journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Convective weather systems present a challenge both to pilots, who wish to minimize operational delay, and air traffic controllers, who are tasked with optimizing airspace usage. The impacts of convective weather on the National Airspace System (NAS) are diverse and have been repeatedly studied in efforts to achieve more efficient airborne operations. These impacts are seen in both terminal and en-route airspace. The FAA estimates that convective weather accounts for nearly 30% of delays in the NAS during the summer months [1], and roughly one-third of accidents occurring during Part 121 (commercial airline) operations from 2003-2007 were attributed to weather [2].

The FAA employs multiple mitigation strategies to enable efficient operations even during times of severe weather. For example, one of these strategies is the development of Severe Weather Avoidance Plans (SWAPs) in areas that are "particularly susceptible to severe weather" [3]. When aviation meteorologists in these regions forecast weather-imposed airspace constraints, a SWAP may be activated. These plans may involve the re-routing of traffic to less constrained areas and the use of pre-determined routes, such as plays from the FAA Playbook or published Coded Departure Routes (CDRs) [4]. These routes help aircraft avoid the hazard areas while lowering the burden of coordinating between air traffic controllers, aircrew, and airlines, allowing air traffic to continue to flow efficiently. However, a key limitation in many SWAPs is that the pre-determined routes may prescribe large deviations that

are not required to avoid the hazardous weather, unnecessarily increasing flight time and expense. Due to the financial and temporal costs associated with weather delays, research is ongoing into weather avoidance optimization.

This work is one of many efforts to reduce overall convective weather impacts in the NAS. By combining weather information with behavior modeling, it is possible to better understand what convective weather will be considered impenetrable to aircrew. This understanding could be used to improve real-time route management on the ground and decrease the amount of coordination required between aircrew and air traffic control when handling convective weather events.

The primary objective of this work was to develop a model that could predict convective weather avoidance for en-route aircraft. Given an aircraft's intended route of flight and local weather observation data, the model would predict if the aircraft would continue on its planned route or deviate to avoid convective weather. The model was developed using large air traffic and weather datasets, learning from tens of thousands of previous interactions between aircraft and convective weather. Aviation weather avoidance models do exist today, but these models were typically developed from relatively small datasets. Moreover, their development process often does not scale well for larger datasets. With these models, each case added to the dataset required a minimum level of human involvement, so time and labor costs scaled linearly with dataset size. This work takes a data-driven perspective, exploring methods for building a predictive model that can incorporate flight and weather data at a much larger scale.

1.1 Convective Weather

Convective weather, or the "vertical transport of heat and moisture in [an] unstable environment", typically impacts the NAS in the form of thunderstorms [5]. Both isolated thunderstorms and thunderstorm clusters called squall lines pose serious threats to aircraft. The primary concerns are hail and turbulence, though lightning and icing are also of concern [6]. Hail can cause structural damage to aircraft very quickly,

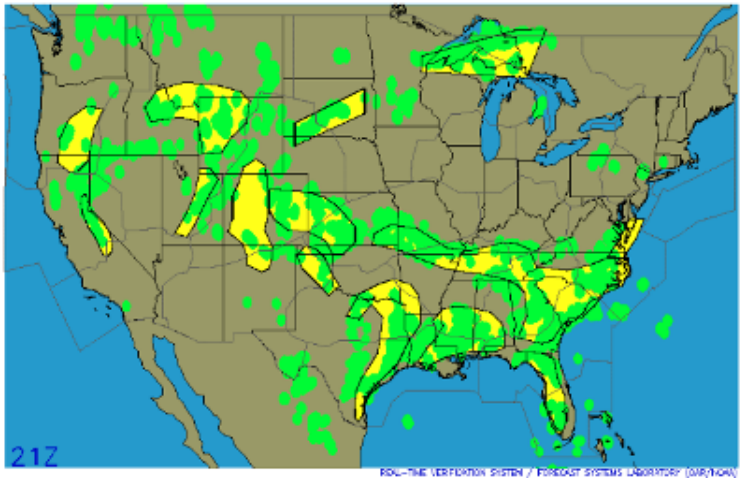


Figure 1-1: An illustration of airspace constraints from convective weather. The green areas represent areas of intense convection, plus a 20 nautical mile radius. Yellow areas are the Convective Constraint Areas from the Collaborative Convective Forecast Product (CCFP), and represent areas of reduced traffic flow due to forecast convective threats. Figure from [7].

even if the hailstones are only one-half inch in diameter. Turbulence can cause loss of control as well as structural damage to the aircraft.

In the air, thunderstorm clouds are the most easily visible indicators of hazardous weather. Similarly, weather radar detects areas of high air moisture content, such as clouds, to identify convective activity. Higher levels of air moisture correspond to more intense convective activity, and weather radar installations will observe higher levels of radar reflectivity for these area. However, it can be difficult to precisely locate aviation hazard areas. Both hail and severe turbulence may be present several miles away from the thunderstorm clouds. As a result, the FAA recommends staying at least 20 nautical miles (nmi) away from severe radar echoes [6]. However, as shown in Figure 1-1, the 20 nmi radius can severely constrain available airspace.

The process for avoiding convective weather hazards is a collaboration between aircrew and air traffic control. Ultimately, the person responsible for the safety of the aircraft is the pilot-in-command. However, air traffic controllers are responsible for the overall flow of traffic through the airspace. If an air traffic controller directs an aircraft on a route that the pilot-in-command deems to be unsafe, the aircrew will inform the controller that a deviation is required. The aircrew and the controller will work

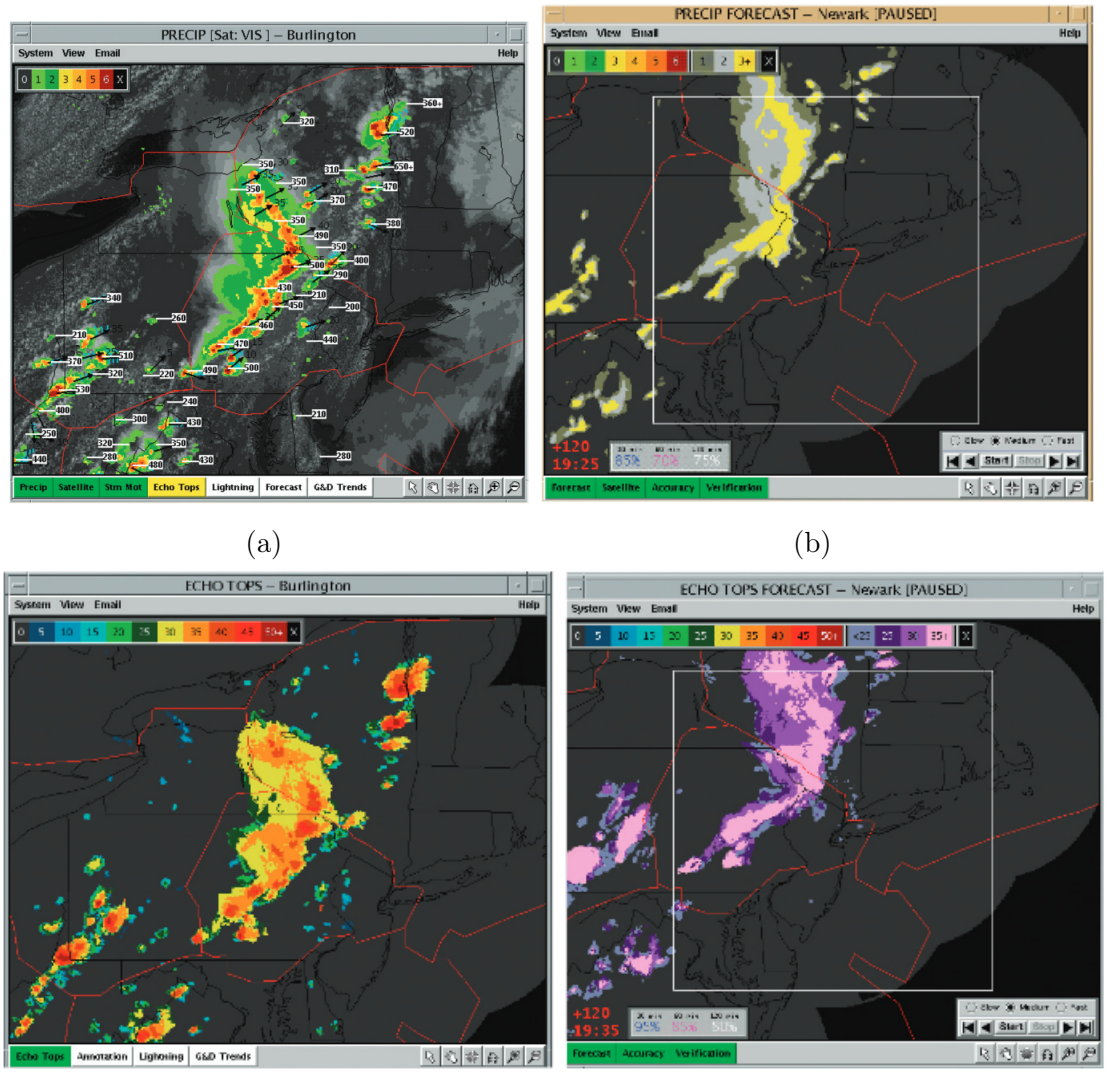
together to find a route that is acceptable to the aircrew and does not interfere with other traffic in the controller's sector. The additional burden of coordinating route changes with aircrew can reduce the number of aircraft a controller can safely handle, which in turn reduces the number of aircraft that can be transiting through a sector. Giving air traffic controllers better weather assessment tools can help reduce the number of deviations required as well as reduce the amount of coordination required for each deviation.

1.2 Weather and Behavior Modeling

1.2.1 Current State

Many tools have been developed to aid controllers in predicting what airspace will be usable both in real-time as well as in the future. Some of these tools, such as the Corridor Integrated Weather System (CIWS), integrate weather observations from multiple sources to provide updated forecasts and situational awareness. For example, CIWS combines ground-based radar data with satellite data and surface observations to produce multiple convective weather displays, including precipitation mosaics and echo top (storm height) mosaics, which inform controllers on the location and trends of convective weather systems in their airspace [8]. Figure 1-2 hows some examples of how this data can be shown to controllers on the ground. Other tools, like the Route Availability Planning Tool (RAPT), use weather forecasts to predict which of several departure routes will be flyable [9]. The development of the Convective Weather Avoidance Model (CWAM) in 2006 was an important milestone for convective weather avoidance tools. This model identified flights that had encountered a minimum level of convective weather and labeled them as deviated or non-deviated relative to their intended (filed) route. This model was one of the first to quantitatively correlate pilot behavior with different convective weather characteristics [10].

CWAM and its applications have been studied extensively over the past two decades [11]–[15]. However, as will be discussed in more detail in Chapter 2, CWAM



(a) (b)

(c) (d)

Figure 1-2: Some examples of different displays available using the Corridor Integrated Weather System (CIWS) data. Subplot (a) shows a Vertically Integrated Liquid (VIL) display; subplot (b) shows a VIL forecast display. Subplot (c) shows an Echo Tops display; subplot (d) shows an Echo Tops forecast display. Figures from [8].

comes with a few key limitations. At a high level, many of these limitations can be traced back to difficulty in accurately identifying examples of deviations and non-deviations in the data set. This process had not yet been automated reliably, which means that every labeled example trajectory must be verified by a human subject matter expert [10]. Building a deviation prediction model with a much larger dataset requires a new approach.

1.2.2 Research Potential

In the nearly twenty years since the inception of CWAM, computational abilities and techniques have advanced rapidly. In particular, great strides have been made in the areas that leverage large datasets, like deep learning and computer vision. As a result, there have been recent pushes to collect and store large quantities of data that might be useful in machine learning efforts. Today, very large air traffic and weather datasets exist and can be used in a data-driven approach for aviation weather avoidance modeling.

Fundamentally, the goal is to better understand interactions between aircraft and convective weather events. The key insight is that these interactions are already embedded in the existing datasets. Combining air traffic data and historical weather observations enables learning from tens of thousands of individual interactions that have already happened.

However, the size and depth of the available data also presents its own challenges. The interactions exist in the datasets, but sifting through the data to group interactions into useful behavioral categories is a non-trivial problem. To take full advantage of modern computational techniques, the data must be processed in a way that is efficient and minimizes the need for human input. Human involvement becomes infeasible as the size of the dataset grows and puts practical limitations on the amount of data used in machine learning solutions. The original CWAM work identified reliance on humans for data labeling as one of the key limitations of the approach. Since the case studies used in the work were labeled by experts, the final model was based on only about 500 cases, all from the same ATC supersector [10].

While the researchers developed an automated labeling algorithm, they found that it was too unreliable to be used unsupervised and suggested that further research be done into automatic labeling of cases.

Many machine learning paradigms exist, and the ability to find useful, interesting patterns in these large datasets will depend on both the machine learning approach and the framing of the data to the machine learning model. This is another challenge associated with the data-driven approach. Part of the contribution of this work is the exploration of machine learning techniques previously unused in the field of aviation convective weather research. In particular, this work examined the use of image classification methods for analyzing combinations of weather and flight data.

The applications of a more robust deviation prediction model extend beyond decision support aids. The results from CWAM have been used by other researchers to improve aviation path-planning techniques. For example, McNally et al. developed a dynamic weather re-routing model that could recognize when a en-route aircraft would encounter impassable weather based on its trajectory, then generate acceptable alternative trajectories [16]. The CWAM results informed the model's definition of impassable weather. In this case, CWAM is not being directly as a decision-support tool to inform controllers of impassable regions, but as an input to another model that is constantly checking for conflicts and presenting options for resolving conflicts when they are identified.

Another potential key application of a weather avoidance model is its use in trajectory planning and situational awareness for autonomous aircraft. For truly autonomous operations, aircraft will need to evaluate the surrounding weather for acceptable flight conditions and modify their trajectory as required with limited or no human supervision. An accurate model of where human pilots will choose not to fly could be valuable information for an autonomous vehicle.

1.3 Objectives

The objectives of this work are two-fold:

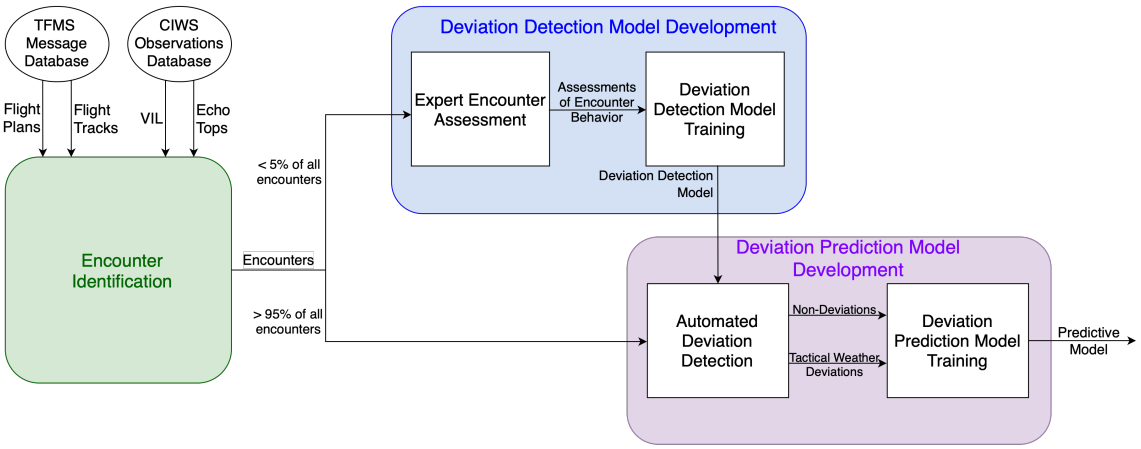


Figure 1-3: An overview of the approach taken in this work, which is broken into three distinct phases.

1. Develop a model to predict convective weather avoidance for en-route commercial aircraft

2. Develop an automated procedure for identifying enroute deviations due to convective weather

The primary objective is centered on better predicting when aircraft will deviate for convective weather. The focus is on en-route aircraft, as behavior in the terminal area is significantly different from behavior en-route. Likewise, the scope is limited to air transport aircraft (Part 121 operations). General aviation aircraft operate under very different conditions and their behavior is likely best modeled separately.

As a secondary objective, this work seeks to develop a deviation detection model that can automatically classify aircraft-weather interactions according to aircraft behavior. Such a model would enable learning from large air traffic and weather datasets.

1.4 Approach

This work can be broken into three distinct phases of development, as shown in Figure 1-3.

The first phase is encounter identification. The primary inputs to this phase are air traffic and weather data. Temporally and spatially relevant weather data is matched to each flight, and the planned route of flight is evaluated for minimum significant weather exposure. If the minimum weather exposure is met, it is considered a point of weather-aircraft interaction. The flight and weather data shortly before and after the point of interaction is captured as a "weather encounter." This process isolates the data that is most likely to be informative on aircraft behavior in and around convective weather events. Encounter identification is carried out on tens of thousands of flights across multiple days, yielding a large set of weather encounters that capture a range of different behaviors.

The next phase is deviation detection model development. In this phase, the goal is to develop an automated mechanism for classifying weather encounters according to their weather avoidance behaviors. This is a retrospective task using the flown route, planned route, and local weather conditions. In particular, the output of this stage is a classifier that could identify a) weather encounters that resulted in non-deviation and b) weather encounters where the aircraft ultimately deviated to avoid convective weather phenomena. First, a small set of weather encounters were selected, which would be used to train and evaluate the deviation detection. A number of subject matter experts (SMEs) were recruited to assess the weather encounters and determine the most appropriate behavior classification. The three behavior classifications were non-deviation, deviation like due to convective weather conditions, and deviation likely due to causes other than convective weather.Once the weather encounters had been assessed, they could be used in developing the deviation detection classifier, which was the output of this stage.

In the final stage, deviation prediction model development, the outputs from the previous two stages are combined. The large set of unlabeled weather encounters is labeled by the deviation detection classifier. Two key behaviors, non-deviation and tactical weather deviation, form the basis of the deviation prediction datasets. In this stage, the flown route data is masked for each example. Only the planned route and the weather conditions are known, simulating a real-time encounter where the

outcome is still unknown. The non-deviation cases and the tactical weather deviations are used to train and evaluate the deviation prediction model, which is the outcome of this stage.

A novel element of the deviation detection and deviation prediction model development was the incorporation of image classification techniques. Frequently, convective weather avoidance work downsamples weather data into a manageable number of spatial features, then correlates those spatial features with types of aircraft behavior. The aircraft behavior classifications are frequently based only off of flight data and do not incorporate weather data. As will be discussed in Chapters 5 and 6, both the deviation detection and the deviation prediction model were developed using standard image classification techniques with some task-specific adaptations.

1.5 Contributions of the Thesis

The primary contribution of this work is the development of a convective weather avoidance model that outperforms current state-of-the-art avoidance models. This model was trained using a dataset of over 30,000 weather encounters from across the continental United States. The data-driven framework developed in this work provides a way to leverage large air traffic and weather datasets. This was possible due to the development of the deviation detection model, which is another significant contribution from this work. The lack of a usable deviation detection classifier has been identified numerous times as a hindrance to developing more robust weather avoidance models.

This work also demonstrated the utility of framing tasks that involve flight and weather data as image classification problems. To the author's knowledge, the combination of weather and flight data being represented as an image directly to a model is a unique innovation.

1.6 Organization of the Thesis

The thesis is organized into eight chapters. Chapter 3 describes the data used in this thesis, its sources, and the processing done on the data. The process for identifying weather encounters from flight and weather data is laid out in Cahpter 4. Chapter 5 describes the development of the deviation detection model, the final model architecture, and the performance of the model. In Chapter 6, the deviation prediction model is developed, and its performance briefly analyzed. Chapter 7 provides further discussion on the performance of both classifiers and suggestions for further research. Finally, Chapter 8 summarizes the findings and contributions of this work.

Chapter 2

Literature Review

2.1 Convective Weather and Air Traffic Management

As previously stated, there has been extensive research into optimizing air traffic management in the presence of convective weather conditions. This research has produced operational tools for improving controller situational awareness and models that provide more specific decision support based on weather conditions. More recent data-driven attempts have demonstrated benefits for air traffic management in studies.

2.1.1 Improving Situational Awareness for Air Traffic Controllers

Modern real-time convective weather situational awareness for air traffic controllers can be traced back to the development of the Corridor Integrated Weather System (CIWS) in 2002. When CIWS was being developed, controllers did not have "accurate, timely information on locations of significant storms". Instead, two forecast products were available from the Aviation Weather Center: one-hour forecast contours via the National Convective Weather Forecast, or NCWF, and a set of 2, 4, and 6 hour forecasts via the Collaborative Convective Weather Forecast Product, or

CCFP. Notably, the NCWF did not provide forecasts for small convective cells. At the time, the FAA intended to field ground-based weather radar (NEXRAD) composite reflectivity displays for controllers, but there were concerns over how controllers might interpret different anomalies in the reflectivity products.

To address this shortcoming, CIWS consolidated aviation weather models and real-time observations, then provided this consolidated information to air traffic controllers in a digestible way. The resulting product is a tactical (0-2hrs) decision support tool for congested en-route airspace that is affected by convective weather. In their 2002 paper, Evans et al. identified three steps in improving forecasts: first, improve weather sensing; second, improve storm severity prediction algorithms; and third, establish a common situational awareness between key users (en route controllers, TRACON, airlines, etc.) [17]. Over the next several years, CIWS evolved, adding more products to its suite with the goal of "improv[ing] the accuracy and timeliness of the storm severity information and [providing] state-of-the-art, accurate, automated, high-resolution, animated three-dimensional forecasts of storms (including explicit detection of storm growth and decay)" [8]. Some of these products were shown previously in Figure 1-2. The CIWS products have since been used in a number of aviation weather products, including the Collaborative Storm Prediction for Aviation (CoSPA) [18], as well as the original and subsequent Convective Weather Avoidance Models (CWAM) [10], [14].

2.1.2 Decision Support Tools for Air Traffic Control

Route Availability Prediction Tool

The effects of convective weather delays on the National Airspace System (NAS) is demonstrated by the challenges of managing air traffic into and out of New York City-area airports. The NYC terminal area has an extremely high density, restricting options available for alternate routing in the presence of convective weather. To this end, multiple prototype decision support tools have been developed to aid air traffic controllers in optimizing departure flow from the New York City airports by

predicting the impact of convective weather on departure routes [19], [20]. The Route Availability Planning Tool (RAPT) first became operational in 2002. RAPT evaluates departure routes for convective weather impact up to 30 minutes in the future and assigns a color based on its assessment of the convective weather impact and airspace usage. Originally only incorporating the precipitation intensity contours, the blockage determination algorithm was updated to include the echo top height in 2004. Other changes were also made to reduce over-warning. DeLaura et al. found that further work was needed to reduce "over-tuning" to specific scenarios and allow for a more transparent route blockage algorithm that is more robust to small changes in the convective forecast [12].

Convective Weather Avoidance Model

One of the first attempts to correlate historical aircraft behavior and convective weather conditions was the work of DeLaura and Evans in 2006 [10]. This study analyzed hundreds of trajectories crossing through two Air Traffic Control super-sectors. Flights that encountered a minimum level of convective weather were considered to have had a "weather encounter." For each encounter, the route as filed and the route as flown were compared and the distance between the two routes was calculated for each point along the filed route. Encounters were then labeled according to whether or not they deviated from their filed routes. The labels were applied automatically, based on the flight's mean deviations distance from its filed trajectory, but were then verified by a subject matter expert. These weather encounters and their labels were then used to develop a quantitative model that could predict deviations based on weather condition inputs. The study characterized the convective weather using three different statistics: the vertically integrated liquid (VIL), which is a measure of precipitation intensity; the radar echo tops, which measure storm height; and cloud-to-ground lightning strikes, a measure of convective activity.

The study found that the most important factor for predicting weather-related deviations was the storm height relative to flight altitude. Precipitation measurements (in the form of VIL) could then further increase the predictive power of the model.

The application of this model was in the development of three-dimensional weather avoidance fields (WAFs), which gave the probability of a pilot deviation at each grid cell based on the input weather characteristics. An identified major limitation of the study was the necessity of expert labeling of deviation or non-deviation for the weather encounters. While there was an algorithm to automatically determine deviation or non-deviation, it was not reliable enough to be used without human verification. This limitation ultimately reduced the number of the available inputs to the deviation prediction model. The original Convective Weather Avoidance Model (CWAM) achieved roughly 75% accuracy in predicting deviations for its initial data set [10].

Post-CWAM Work

Subsequent work built on the original CWAM with a variety of different methods. Again relying on manual labeling methods, Chan et al. noted that in some circumstances, pilots flew through areas that CWAM had predicted to be impassable [21]. This indicated that factors not included in the original CWAM were necessary to consider when making airspace penetration assessments. Additionally, a 2014 study by Lin and Balakrishnan identified some factors that influenced whether aircraft would penetrate severe weather. These factors included, among others, operational factors, differences between pilots in their behavior, and the shape of the convective activity [22].

The authors of the original CWAM paper extended the study by adding more encounters and including more meteorological deviation predictors [12]. This study confirmed that the storm height (echo top height) relative to the flight altitude was the best predictor of pilot deviation. However, the study found that the "major challenge in convective weather avoidance modeling" was predictions where the flight altitude was close to the storm height. Once again, the authors identified the need for a reliable deviation detection algorithm, as this remained a major limitation to the number of cases considered in the deviation prediction model [12].

Another attempt to validate the two convective weather avoidance models, this

time via a comparison of the model outputs and observed aircraft-based data, was published in 2009 [11]. Again, the difficulty in identifying and validating pilot deviations around convective weather makes it difficult to validate both CWAM and other models that correlate weather characteristics and pilot deviation decisions. In order to get around this limitation, the authors sought a more direct comparison between weather avoidance field values and the experience of pilots in the cockpit (e.g. turbulence, airborne radar readings, visual sighting of convective activity). Four research flights were flown in and around areas of convective activity, and data collected from onboard instrumentation was compared to WAF values output by the two CWAMs. Turbulence was measured using an inertial navigation system (INS), and the aircraft also was equipped with a GPS, a satellite weather radar display, and an airborne weather radar. Over the course of the four flights, the study found that "regions of moderate turbulence in and around convective cores were characterized by high WAF deviation probabilities, ranging from 0.70 to 1.0 (70% to 100%)", with CWAM 2 outperforming CWAM 1. However, neither of the models was able to accurately differentiate between a downwind anvil and non-turbulent stratiform rain with a high echo top. Further, based on cockpit feedback, radar reflectivity is not sufficient to determine areas of visual meteorological conditions (VMC) and instrument meteorological conditions (IMC). This was the case even when using the on-board radar (which did tend to have good agreement with the ground-based radar). The study suggests that distinguishing between VMC and IMC may be a key input to a convective weather avoidance model. Finally, the authors indicated the need for more pilot input: "What is needed is a way to take data from studies that capture many details of the flight experience (ground-based radar, cockpit radar, visual cues, turbulence encountered, etc.) and package them in a way that can be widely distributed to pilots in an effort to gather data about the specific factors that influence their decision in convective weather. The presentation of the data must be sufficiently true to operational experience that pilots are able to make accurate judgments of their behavior in conditions similar to those presented" [11].

Matthews and DeLaura published another evaluation of CWAM performance in

2010 [23]. A retrospective analysis of the data set used to generate the quantitative avoidance model found that predictive ability varied by Air Route Traffic Control Center (ARTCC), the region in which the weather encounter occurred. The reason for these differences was not immediately obvious, though the authors speculated that such differences could be due to different predominant weather types in each region, air traffic control willingness to accommodate willingness, and whether or not acceptable alternatives exist. In addition to analyzing algorithm performance by region, the study evaluated CWAM performance for three different weather types and found a possible correlation between performance and weather type [13]. A 2012 paper by Campbell, Matthews, and DeLaura also investigated an adaptation of CWAM for terminal airspace [24]. The study found that the model was "able to *decisively* predict deviations and penetrations."

Despite the general accuracy and usefulness of the en-route convective weather avoidance model, the output weather avoidance fields can sometimes result in overly complex and unrealistic avoidance trajectories. For example, a flight that is encountering a string of isolated convective cells is more likely to simply avoid the area altogether, rather than weave around the cells. To this end, an algorithm was developed to use WAFs along with other storm characteristics to define a convective weather avoidance polygon (CWAP) [15]. Essentially, the CWAP attempts to take the convective “cores”, as defined by the WAF, and approximate the storm’s visual boundaries using the echo top edges. The assumption is that the visual boundaries represent areas that pilots will choose to avoid. This CWAP identification algorithm was tested on cases from two different days. The performance of the algorithm differed significantly by day. The authors attribute this difference to well-organized storms on one of the days, which allowed for less ambiguous storm boundary detection. On the other day analyzed, smaller-scale convective systems were present, and the CWAP algorithm was overly aggressive in grouping these systems into a single impassable region. The authors suggest that sharing of CWAP information between ATC and pilots could allow for pilots to choose from an assortment of suggested trajectories.This would mean that instead of ATC having to predict what trajectory would be accept-

able to a pilot, the pilot only needs to pick one acceptable option from a small set of trajectories, allowing for more efficient operations during convective weather.

Another application of CWAM was described by McNally et al. in a 2012 study. This study described an automated trajectory rerouting system that continuously evaluates flights en route for convective weather impacts on their planned trajectories, and suggests alternative routes for flights whose trajectory may require correction [16]. The Dynamic Weather Routing (DWR) system uses CIWS output to generate CWAM WAFs. These WAFs are inputs to the trajectory modeling system, which then attempts to find a rectified trajectory with minimum delay. The trajectory modeling system also takes traffic into account. Over 14 hours of data from the Fort Worth Center traffic, DWR resulted in an average 10 min time saving per flight over 171 flights [16].

2.1.3 Data-Driven Approaches

Another approach by Sheth et al. avoided the limitations associated with deviation labeling and prediction by not considering intent in their probabilistic avoidance model. This study correlated air traffic data and convective weather forecast data, deriving a measure of the probability of flight in the vicinity of forecasted convective weather [25]. This value could then be used to inform controllers of regions likely to be avoided by aircraft. This is similar to the approach taken by Kuhn. Kuhn also noted the difficulties associated with labeling a large number of trajectories, and instead estimated weather impact by counting the number of aircraft in grid cells with similar weather conditions, with the goal of understanding which cells would always or almost always be avoided [26].

In a 2009 paper, Michalek and Balakrishnan examined the relationship between route blockages and certain features of weather forecasts. Using forecast features that were strongly correlated with route blockage, a classifier was developed to predict open and closed routes, given a set of routes and weather forecast [27]. A follow-up 2012 paper by Pfeil and Balakrishnan extended this work to identify "robust routes" that are less sensitive to uncertainty in the weather forecast, as well as the features of a

route (as determined by the forecast) that might predict robustness [28].

Finally, in 2015 Arneson used a combination of supervised and unsupervised machine learning approaches to explore the use of certain FAA Playbook routes [29]. These pre-determined routes are published by FAA and can be used when convective weather conditions require that aircraft be re-routed. The intention of this work was to understand how weather conditions dictated which Playbook routes would be implemented. The study examined 2,208 hours of weather and flight data. Routes inbound to the NYC terminal airspace were clustered to determine which re-routes appeared in each hour under study. Then, weather features were calculated for each hour based on echo top data. For the twenty most commonly used re-routes, a model was trained to predict if the route would be used given a set of weather features. Initial results were mixed, with fifteen of the twenty models achieving true positive and true negative rates above 60%.

2.2 Image Classification

2.2.1 Convective Weather as an Image Classification Problem

As previously discussed, the predictive task and part of the deviation detection task were framed as image classification problems. Fundamentally, analyzing weather and traffic data together requires capturing spatial relationships between the weather flight and data. Typically, flight and weather data are processed separately, and important features are extracted from each dataset before being combined. The image-based approach allows simultaneous consideration of flight and weather data. Image classification models are designed to efficiently capture spatial relationships in the input data and learn from these patterns.

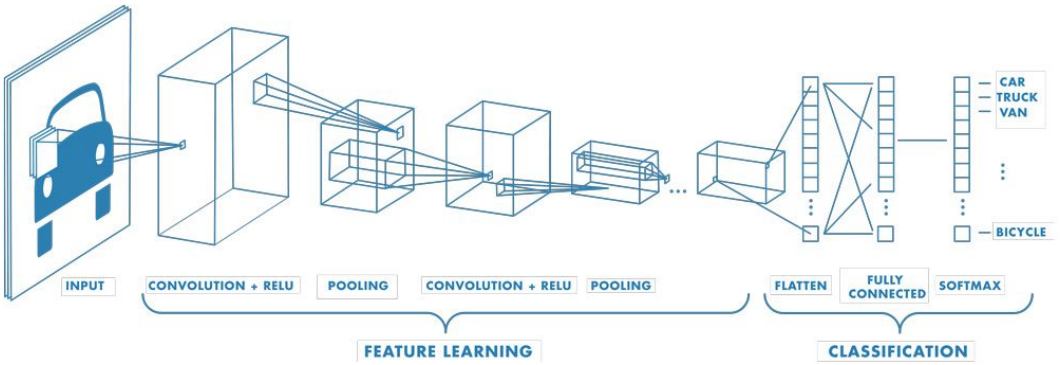


Figure 2-1: Typical convolutional neural net (CNN) architecture, with alternating convolutional layers (spatial filters) and pooling layers. Image from [31].

2.2.2 Convolutional Neural Nets and Transfer Learning

Convolutional Neural Networks

Image classification methods have been researched extensively in the past decade, and there is a wide body of literature to draw on for this work. In particular, convolutional neural nets, or CNNs, have been identified as especially useful for image processing tasks [30]. A typical CNN architecture consists of a series of alternating spatial filtering layers and pooling layers. Figure 2-1 illustrates this common architecture. Simplistically, early layers of the CNN look for small local patterns, later layers will look for patterns in those local patterns, and the final layers will consider all of the identified patterns to make a classification. The use of successive filters to identify key underlying relationships in the input image is known as "feature extraction".

Transfer Learning

Framing the deviation type classification problem as an image classification task also helps mitigate the challenges encountered when training a model on a small dataset. A technique known as transfer learning takes a model that has already been trained on a (usually large) dataset for a classification task, and adapts it for use with a different dataset and classification task. The idea is that the model has already been trained to extract useful features from the inputs, and those features can be used to make predictions on a different dataset with relatively slight modifications to the model. Transfer learning helps researchers with small datasets leverage state-of-the-art techniques in image classification. These techniques often use datasets with tens of millions of examples, with ImageNet being one of the most well-known examples [32]. Construction of these datasets is feasible because typical images do not require expert opinions, so image labels can be crowd-sourced across a very large population.

Transfer learning has seen a surge in popularity in the biomedical imaging community [33]. With biomedical imaging, image labels are very expensive to obtain because they require highly trained experts, so supervised learning approaches are typically working with relatively small datasets. The success of transfer learning in this domain demonstrates that a model trained on a very general image dataset can be successfully adapted for a much more specialized task. Similarly, assessing an encounter and inferring its cause of deviation requires a significant level of expertise, limiting the size of a labeled training dataset. However, transfer learning techniques had the potential to mitigate this limitation.

Chapter 3

Flight and Weather Data

3.1 Data Pre-Processing

3.1.1 Flight Pre-processing

The flight data in this study comes from the Lincoln Lab Traffic Flow Management System (TFMS) database. The FAA defines the Traffic Flow Management System as "a data exchange system for supporting the management and monitoring of national air traffic flow" [34]. The TFMS database receives and stores messages for flights in U.S. airspace. Examples of messages include position updates, flight plan messages, and flight cancellation messages [35]. For the purposes of this work, messages with flight plans, flight plan amendments, or a position update for the aircraft were considered.

The data stored in the Lincoln Lab TFMS database is compiled hourly, with each hour of each day having its own XML file. Therefore, messages for a single flight are spread out over multiple hours, and often multiple days, since flight plans will often be submitted the day prior to the flight. Each hourly file contained about 2 GB worth of messages stored as plain text in a standard XML formatting. Because of the size of the files and the lack of organization by flight, some pre-processing was required to parse the TFMS data into a usable format.

For each day, two blank CSV files were generated: a file for position updates and

a file for plan creation and amendments. Then, each hour of that day was processed. Each TFMS message in the hourly file has a "message type" tag that describes the kind of information contained in the message. Messages were screened according to their message type. The desired message types for this work were "track information", "flight plan information", and "flight plan amendment information" messages. If the message type was "track information", the message contents were extracted and written to that day's CSV file containing position updates. Similarly, if the message type was either "flight plan information" or "flight plan amendment information", the message was written to that day's CSV file with flight plans. The information included for each message type is shown in Table 3.1. The resulting CSV files for each day totaled about .75 GB combined, compared to >40 GB for a day's worth of TFMS messages in their original format.

The flight plan shown in Table 3.1 is a mix of waypoints, jet routes, and standardized arrival and departure routes. "KBOS" is the airport identifier for Boston Logan. "PATSS6" is a Standard Instrument Departure (SID) which defines a sequence of waypoints to traverse while exiting the BOS terminal airspace. "PATSS" is the final waypoint in the PATSS6 SID. "NELIE" is a waypoint with a defined latitude and longitude. "CMK" is a navigational aid, also with a defined latitude and longitude. "J75" is a jet route which consists of multiple waypoints in sequence; in this case only the portion of the jet route between "CMK" and "GVE" is part of the flight plan. "GVE" and "LYH" are also navigational aids. "CHSLY4" is a Standard Terminal Arrival Route (STAR); like a SID, the STAR defines a series of waypoints that will be traversed on approach to the Charlotte airport. Finally, "KCLT" is the airport identifier for the Charlotte airport.

To compile flight records for any given day, the CSV files for that day, the day before, and the day after were pulled. All unique flight reference numbers for that day's flight information file were identified. Any flight reference numbers that appeared in the flight information file for the day before were discarded. For each flight reference number, all messages associated with that number in the flight information files for the day in question or the day after were collected. The route files from the current

Track Information Message Data		**Flight Plan Message Data**	
Data Type	Example	Data Type	Example
Flight Ref. Number	123456789	Flight Ref. Number	123456789
Message Timestamp	1561953593	Message Timestamp	1561953593
Aircraft ID	JBU101	Aircraft ID	JBU101
Airline	JBU	Airline	JBU
Departure Airport	KBOS	Departure Airport	KBOS
Arrival Airport	KCLT	Arrival Airport	KCLT
			KBOS.PATSS6.
			PATSS..NELIE..
Latitude	40.48056	Plan	CMK.J75.GVE..
			LYH.CHSLY4.
			KCLT/0214
Longitude	-80.0875		
Speed (kts)	450		
Altitude (100s of ft)	310		

Table 3.1: Description and example of Traffic Flow Management System (TFMS) data used in this work. The left side gives an example of a TFMS track information example; the right side gives an example of a flight plan message.

day and the day before were checked for the flight reference number, and the routes associated with that number were also collected. This process associated each flight reference number with its available track and plan information.

Once the flight data had been compiled for each flight, the data was filtered to exclude cases not in the scope of this work. To ensure completeness of data for each of the flights, as well as availability of weather data, only flights that departed from and arrived at airports within the continental United States were considered. Only flights whose first and last observations were below 10,000 ft were considered, again to ensure completeness of flight data. Also, only air carrier operations (Part 121) were considered. This filtering was accomplished by checking the flight number listed in TFMS against a list of prefixes for U.S. air carriers.

Finally, flights were evaluated for route validity. The last route or route amendment before the time of the flight's departure was set as the flight route. Each element of the route was identified as a waypoint, NAVAID, airport, or airway using FAA databases. The latitude and longitude was recorded for each waypoint, NAVAID, or airport. Since airways are comprised of multiple waypoints in sequence, the appro-

priate start and end waypoints of the airway were identified in the filed route, and the latitudes and longitudes for waypoints in the relevant section were recorded. If any element of the route was not identifiable in the FAA databases, the flight was discarded.

3.1.2 Weather Pre-Processing

Weather Data Descriptions

The weather data was pulled from the Corridor Integrated Weather System (CIWS). First fielded in 2001, CIWS provides "accurate, automated, rapidly updating information on storm locations" that can be combined with the current flight radar data to provide decision support for air traffic controllers [17]. It uses multiple weather observation sources, such as satellite weather data and WSR-88D (NEXRAD) data to generate several weather products. These products include the Vertically Integrated Liquid (VIL) Mosaic, Storm Motion, Lightning, Echo Tops, and Growth/Decay Trends, among others.

For the sake of simplicity, this work primarily uses the VIL mosaic and Echo Tops mosaic data, as these two datasets summarize the intensity and 3-D coverage of convective weather events. Further work could be done to incorporate other factors like lightning or storm motion. In the original CWAM study, the most important predictive features were derived from VIL and echo tops data, though lightning was assessed as well [10].

Both the VIL mosaic and Echo Tops mosiac products are distributed in a gridded format. The projection used is Lambert Azimuthal Equal-Area projection, where each grid cell has the same area. The same projection was used to convert flight position data from latitude and longitude to x and y coordinates. This was accomplished with the PyProj and Cartopy python packages [36], [37].

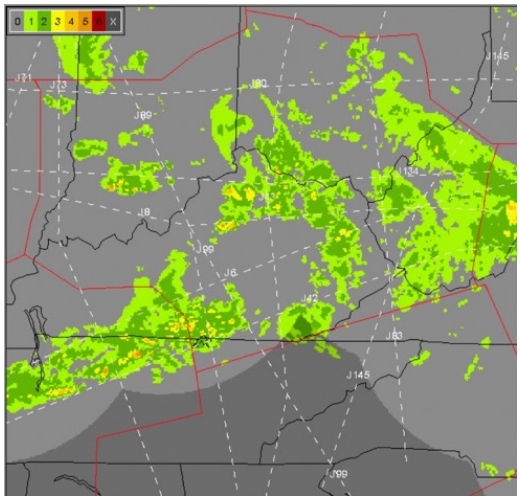


Figure 3-1: The CIWS VIL Mosaic product. [38]

Vertically Integrated Liquid

The Vertically Integrated Liquid is a measure of precipitation density. It is calculated by integrating the amount of water in a column of air. Radar installations scan through a range of elevations to capture reflectivity returns at multiple heights, creating a 3-dimensional reflectivity map of the surrounding space. There are several different methods for mapping this 3-dimensional data onto a 2-dimensional plane (for instance, a computer display showing the locations of flights). Some aviation weather products will display the maximum reflectivity from any beam elevation, while others will pre-select a scan elevation to display, often the base elevation [38], [39]. The VIL handles the mapping problem by integrating all available reflectivity data, giving a more complete weather picture that is less prone to error [38], [40]. An example of the CIWS VIL mosaic is shown in Figure 3-1. In CIWS products, VIL is colormapped to a six-level Video Integrator and Processor (VIP) color scale, as shown by the key in the upper-left corner of Figure 3-1.

Echo Tops

The Echo Top product estimates storm height. Generally, more severe convective weather will correspond to higher echo top heights. Echo top height is particularly important in an aviation context, as storms that are well below cruising altitude may

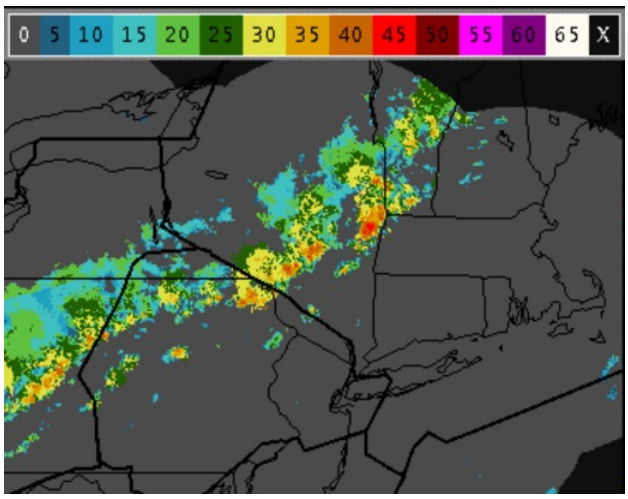


Figure 3-2: The Echo Tops CIWS product. [38]

not impact en-route aviation operations. In fact, a measure of the echo top height relative to the aircraft altitude was the most important predictor of deviation in the original CWAM work [10]. Figure 3-2 shows an example of this product in CIWS. Warmer colors are higher storm heights and cooler colors are lower storm heights. A potential limitation of this metric is that it will only register radar echoes at or above a threshold reflectivity, typically 18dBZ. Clouds may still be visible at altitudes above the estimated echo top [38].

3.2 Selection of Days To Analyze

Air traffic levels dropped significantly in 2020 due to the COVID-19 pandemic and have not returned to pre-pandemic levels as of December 2022 [41]. To maximize the number of flights per day, only days from 2019 and earlier were considered. To ensure broad coverage of convective weather phenomena, a mix of moderate and severe weather conditions was desired. The RAPT Evaluation Post-Event Analysis Tool (REPEAT) software was used to select days based on this criteria. The REPEAT software analyzes route weather blockage and air traffic flow in the New York City and Chicago terminal areas [42]. It also assigns an overall weather impact rating to each terminal area for each day, which can be little to no impact, moderate impact, or

Date (YYYY-MM-DD)	NYC Severity	Chicago Severity
2019-06-16	Moderate	Severe
2019-06-17	Moderate	Moderate
2019-07-03	Moderate	Severe
2019-07-17	Severe	Severe
2019-07-18	Severe	Severe

Table 3.2: The days that the aircraft-weather encounters were pulled from, and the corresponding weather impact rating in the New York City and Chicago terminal areas. The weather impact rating comes from the REPEAT software.

severe impact. These overall impact ratings were used to inform general convective weather conditions for each day. Ultimately, five days were chosen, all from 2019: June 6 and 7, July 3, and July 17 and 18. Table 3.2 shows the severity recorded in the REPEAT tool for each day for both terminal areas (New York City and Chicago).

Chapter 4

Encounter Identification

Together, flight databases and weather databases capture the interaction between millions of flights and the weather that those flights were exposed to. This information can be leveraged to gain insights into how aircrew make decisions about handling weather en-route. However, for a significant number of flights, a large portion of the time spent en-route will be spent in clear-air weather conditions. To better understand aircraft-weather interactions, the goal was to isolate short segments of flights with a nontrivial level of weather exposure for the planned route. In this work, the term "weather encounter" refers to a combined set of flight and weather data that records a relevant aircraft-weather interaction. This chapter will further define the criteria that are used for establishing weather encounters, and how those encounters are identified and constructed.

4.1 Encounter Criteria

4.1.1 Tactical Weather Encounter

The first consideration was that relevant weather encounters would involve decision-making on a tactical, or a short time/distance, scale. Ultimately, the goal is to model which weather situations are likely to be considered penetrable. It can be difficult to draw useful conclusions about a decision to penetrate or not penetrate when the

weather driving that decision is spatially or temporally separated from the decision itself.

For example, consider a hypothetical flight from Los Angeles (LAX) to New York City (JFK). This has filed for a common route, taking a northerly approach that passes just south of the Chicago area and approaches the New York City airspace from the north. Shortly after takeoff for this hypothetical flight, convective weather forecasts predict severe storms across the upper Midwest. In order to mitigate operational disruption downstream, this flight may reroute, taking a more southerly approach and avoiding the areas with more severe weather outlooks. This kind of rerouting can reduce the workload of air traffic controllers in the weather-burdened sectors. Deviating early may also result in a shorter distance flown than a flight that attempts to penetrate the forecasted storms but ultimately is forced to deviate around the storms.

The deviation of this hypothetical flight is a *strategic* deviation. The decision to deviate is an attempt to optimize downstream performance. It does not mean that the weather the flight would have encountered by staying on its filed route would not have been penetrable. The forecast may have overestimated the storm severity, or the locations of severe cells might not have blocked the filed route. Additionally, it is not known which cells would have been considered impenetrable. Some of the cells appearing on the filed route may have been penetrable, but it is impossible to know which cells these are, since a decision for these specific cells was never required. Since the goal is to understand and identify the decisions that were made about specific convective weather conditions, only *tactical* decisions were considered for this work.

The definition of tactical can vary quite a bit depending on application. Even within aviation weather modeling, different studies will define a tactical timescale differently. For instance, Wu considered the tactical time frame to be 0-2 hours when evaluating "the strategic aspect of pilot weather decision making in the tactical time frame" [43]. More explicitly, Latorella and Chamberlain sought to define the boundaries of strategic and tactical decision making in an aviation weather avoidance context [44]. In part of this study, seventeen general aviation (GA) pilots were asked

to evaluate eight different weather situations and estimate when they would switch from a strategic to a tactical response, in terms of minutes until weather contact. The reported times generally increased as the weather scenarios became more severe or covered a greater area. The median strategic/tactical boundary value was 17.5 minutes for the least severe scenario ("an isolated cell with yellow radar return along the route of return"), and 45 minutes for the most severe scenario ("an area of severe thunderstorms perpendicular to, and extending 40nm to either side of the route"). This indicates that the tactical response mode is likely to start earlier when conditions are more severe.

Notably, the study from Latorella and Chamberlain was conducted with general aviation pilots, so these subjects would not be anticipating the aid of onboard weather radar returns when making the distinction between a strategic and a tactical response. Also, the study was conducted in 2002, before ground-based weather observations in a visual format were commonly available in the cockpit of GA aircraft. Weather awareness in GA pilots was driven by knowledge of the forecast obtained on the ground before the flight, aural updates from automated weather reporting systems (ASOS, ATIS, etc.), and visual information outside the cockpit. Identifying the appropriate tactical range for this study required considering the information available in the cockpit to pilots operating a Part 121 flight in the modern era.

With respect to onboard radar equipment, Airbus recommends a 160 nmi range setting for the "long-term avoidance strategy" and a "80NM or below" range when deciding on "avoidance tactics" for weather [45]. Eighty nautical miles is about 150 km, which was set as the baseline for tactical decisions. However, initial investigations of weather encounters identified using the 150 km range revealed that some tactical decision points (the point of deviation) were being missed at a 150 km distance. The tactical range was increased to 200 km to better capture these points. With a typical cruise speed of 430 kts or 800 km/h, a 200 km distance is covered in 15 minutes.

VIL (kg/m^2)	Reflectivity (dBZ)	VIP Level
0.05	<18	0
.14	18	1
0.7	30	2
3.5	41	3
6.9	46	4
12.0	50	5
32.0	57	6

Table 4.1: A conversion between Vertically Integrated Liquid (VIL), Visual Integrator and Processor (VIP), and reflectivity values [39], [48].

4.1.2 Severity of Convective Weather

To define the term "weather encounter", a minimum level of convective weather needed to be established. In the original CWAM, the researchers determined that passing through "either VIL level 2[1] or greater or echo tops of 25 kft or greater for at least 2 minutes" qualified as a weather encounter [10]. These criteria were based on a 2002 study by Rhoda et al., which also analyzed aircraft encounters with convective weather observed to be VIP Level 2 or higher [46]. The VIP system consists of seven levels of precipitation intensity. A conversion between VIL and VIP values can be found in Table 4.1.

In the study by Rhoda et al., VIP Level 2 was chosen as the minimum qualifying weather because aircraft pass through VIP Level 1 weather with high frequency. When only considering a very limited number of encounters, it is reasonable to remove as many trivial encounters as possible from consideration. In the case of this work, the dataset size was less constrained, and it was desired that both weather deviations and non-deviations would be captured, so VIP Level 1 was set as the minimum precipitation intensity for consideration. Based on Table 4.1, VIP Level 1 is roughly equivalent to 0.14 kg/m^2, or a reflectivity of 18 dBZ. A reflectivity of 20dBZ is generally when falling precipitation starts [47]. This roughly corresponds to green or higher levels of reflectivity on NWS radar products.

[1]While the original published study cites a "VIL level 2" criterion, it is interpreted here to mean a "VIP level 2" criterion, as the VIL scale is continuous whereas the VIP scale has discrete levels. Additionally, the reference to the work done by Rhoda et al. further supports an interpretation of VIP level 2.

4.1.3 Initial Encounter Alignment

A flown route can only be deviated relative to its intended route, and so the intended route is an important baseline to establish. In this work, the planned route, defined as the last flight plan on file with the Traffic Flow Management System prior to departure, is used to model the intended route. However, if the flown route and the planned route are separated by a significant distance at the start of the encounter, the planned route will less accurately model the intended route over the course of the encounter. To remove these cases, a maximum distance between the flown and planned routes at the start of the encounter was established.

In the original CWAM study, DeLaura and Evans found the clear-air route widths to be 12 km and 24 km in the ZID and ZOB supersectors, respectively [10]. When calculating weather statistics, a 16 km neighborhood was used, which "approximates the clear-air route width". The 16 km distance was applied in this work as a distance filter for the weather encounters. If the planned route and the flown route were more than 16 km apart at the start of the encounter, the encounter was discarded.

4.1.4 Altitude Restrictions

The decision-making process differs greatly between en-route and terminal area convective weather encounters. In the terminal area, altitude and airspeed are constantly changing, whereas aircraft en-route are flying a much more stable altitude and airspeed. Previous studies have shown that the primary predictor of a weather deviation en-route is storm height relative to the aircraft altitude. Since aircraft are flying at lower altitudes in the terminal area, it is typically not possible to avoid a storm by flying over it, indicating that deviation behavior will significantly differ between en-route and terminal areas. Additionally, a higher density of aircraft in the terminal area can limit the ability to laterally deviate around convective cells. For arrivals, fuel constraints and the use of "corner posts" to manage traffic flow into the terminal airspace further complicate deviation decisions [49]. In research by Temme and Tienes, over half of the pilots interviewed (6 of 11) stated that their convective

weather avoidance decisions were based on the phase of flight [50]. Intuitively and empirically, terminal area decision-making differs significantly from en-route behavior, and it is reasonable to study them separately. For this reason, this work chose to focus on en-route weather encounters. To accomplish this, encounters were filtered for altitude, and only encounters that stayed at or above 10,000 ft were considered.

4.2 Encounter Identification Algorithm

4.2.1 Generating VIL Values

Weather encounters were identified based on convective weather exposure to the planned, not the flown, route. To illustrate the reason for using the planned route, consider a flight that deviates from its flight plan when it is 100 kilometers away from an isolated, intense convective cell that is blocking the intended route. The flight gives the cell a wide berth then returns to its planned route, and the resulting weather exposure for the flown route is minimal. If only the weather exposure on the flown route is considered, this case might not qualify as a weather encounter, even though the case is a relevant example for this work. If the weather exposure on the planned route is considered, this case can be captured as an instance of a convective weather deviation. Capturing successful weather deviations requires considering the weather that was avoided, which is identified based on information from the planned route.

The planned route is stored in TFMS as a series of waypoint or route names. Routes are expanded into lists of waypoints, and all waypoints in the planned route are looked up and converted to latitude and longitudes. Each (longitude, latitude) coordinate pair was mapped to an (x, y) coordinate pair using a Lambert Azimuthal Equal Area projection. This same projection is used in mapping the CIWS observations to a 1 km x 1 km grid. At this stage, the planned route typically consisted of a small number of coordinate pairs. To check for weather exposure along the planned route at a high frequency, this planned route was expanded by interpolating between

the original coordinate pairs at a constant distance. This distance was set so the number of coordinate pairs making up the planned route matched the number of coordinate pairs that made up the flown route. Finally, each coordinate pair in the planned route needed a timestamp to evaluate the weather exposure at that point. This was done by determining the average speed of the aircraft using the total distance of the flown route and its start and end timestamps. The planned route was assumed to start at the same time as the flown route, and its last timestamp was set using the average speed and the total distance of the planned route. After the first and last timestamps for the planned route were determined, the remaining positions' timestamps were evenly spaced between the start and end times. Once each position had a timestamp, the Vertically Integrated Liquid value for that (x, y, t) was retrieved from the weather data archives. Weather observations were taken at five minute intervals, so each timestamp was rounded to the nearest five minutes.

4.2.2 Bounding Encounters

After assigning VIL values, each point on the planned route was evaluated to determine if the VIL at that (x, y, t) met or exceeded 0.14 kg/m^2, which was the previously established VIL threshold. A point on the planned route that met or exceeded the VIL threshold was deemed a "point of encounter". After finding a point of encounter, the next step was to establish encounter bounds using the "tactical" distance of 200 km. The point of encounter was centered in the encounter bounds, and the encounter bounds encompassed 200 km north, south, east, and west of the point of encounter, yielding a 400 km by 400 km box. Establishing encounter bounds in this way guaranteed a consistent two-dimensional size and shape for every encounter. Figure 4-1 shows an identified point of encounter and its corresponding encounter bounds with respect to the entire flight. Figure 4-2 shows the same encounter once it has been cropped to the correct size, as well as the start of the encounter.

If the planned route crossed through a large cell or multiple discrete cells, many consecutive points on the planned route could potentially meet the VIL criteria for an encounter. To avoid treating all of these points as independent weather encoun-

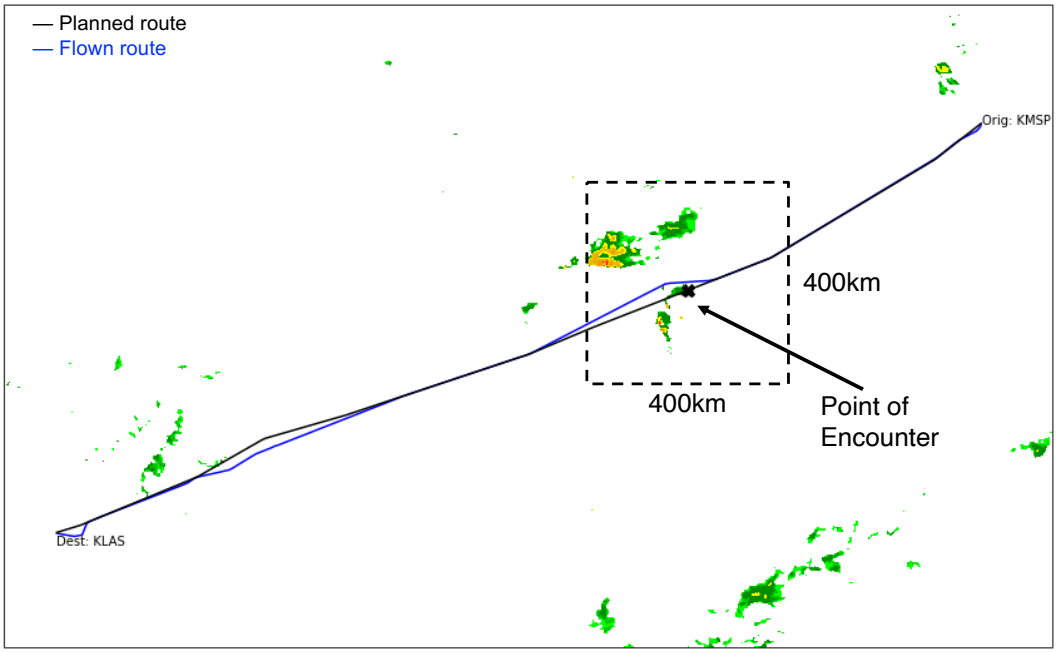


Figure 4-1: An example flight from Minneapolis-St. Paul to Las Vegas, with a weather encounter en-route. The weather shown is the VIL data associated with the timestamp of the point of encounter.

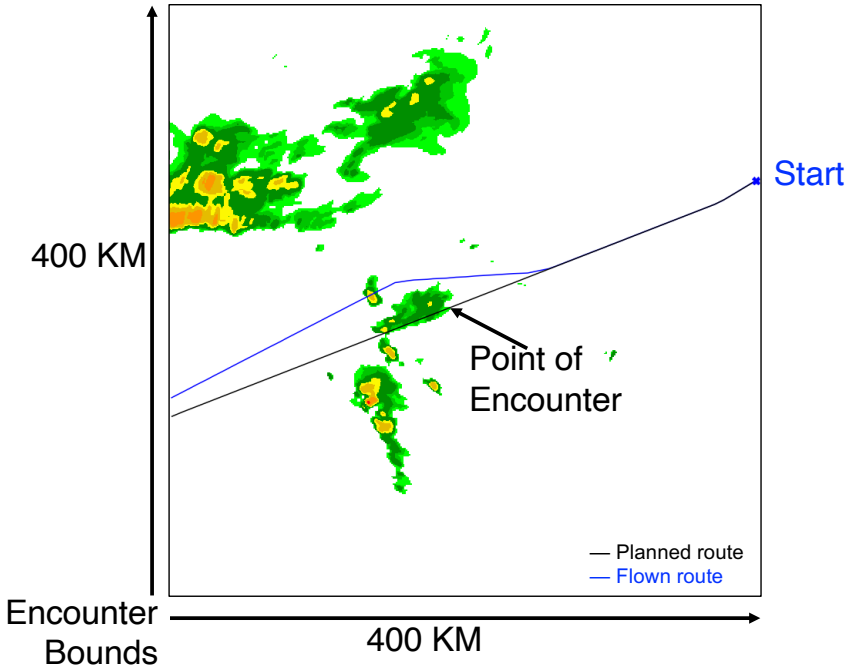


Figure 4-2: The encounter from Figure 4-1, cropped to the standard 400 km by 400 km window. The start of the encounter is labeled.

ters, minimum separation criteria were established. After a weather encounter was identified and validated, any planned route (x, y, t) points within its bounds were not evaluated for meeting encounter criteria. The encounter identification process resumed on the first downstream point outside the encounter's bounds. This method allowed flights to have multiple independent weather encounters.

4.2.3 Validating Encounters

After identifying an encounter, it was validated according to the criteria previously discussed in this chapter. Once the encounter bounds were determined, the distance between the flown route and the planned route at the start of the encounter was calculated. If this distance was greater than the 16 km clear-air route width, the encounter was discarded. The altitude of each point on the flown route within the encounter bounds was also checked. If any these points were below 10,000 ft, the encounter was discarded.

4.3 Results

After processing the five days of flights, 85,968 flights were identified that met the criteria outlined in Chapter 3. From these flights, 48,374 individual encounters were identified. Table 4.2 shows the breakdown of flights and encounters by day. Note that multiple encounters could be identified from the same flight.

Date	**Number of Flights**	**Number of Encounters**
2019-06-16	15357	8897
2019-06-17	18800	13045
2019-07-03	17546	10916
2019-07-17	17141	9100
2019-07-18	17124	6416
Total	**85968**	**48374**

Table 4.2: The number of flights and number of encounters recovered from each date in the study.

A small subset (<5%) of these encounters was used to train the deviation detection

model. After developing the deviation detection model, the remaining encounters were classified by the deviation detection model, then used in the development of the deviation prediction model.

Chapter 5

Deviation Detection Model Development

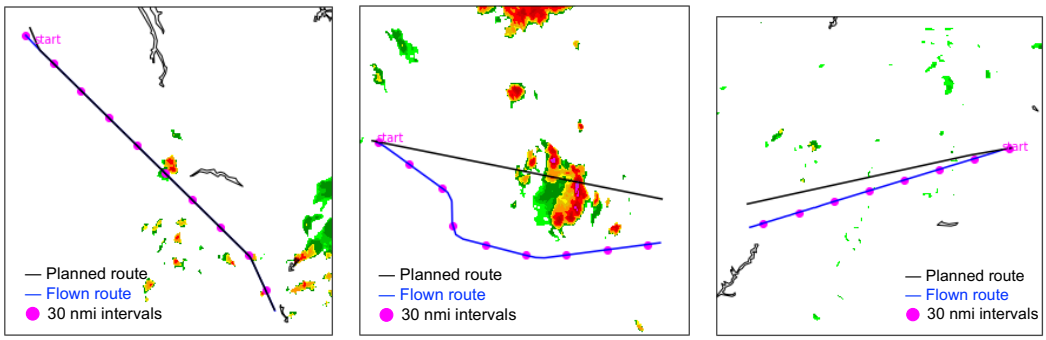
5.1 Overview

For this purposes of this work, three behavioral classes were established for encounters. Encounters could belong to one of three classes: non-deviations, deviations for tactical weather, and deviations for other reasons. Figure 5-1 shows an example of each encounter type. For the purposes of training a predictive model, the non-deviations and the tactical weather deviations are the most informative. The purpose of the deviation detection model is to identify those kinds of encounters for use in predictive model development.

The deviation detection model takes in an encounter, which is comprised of flight plan data, flight track data, and associated weather data. It outputs a behavior label, as shown in Figure 5-2. This chapter will detail the process used to develop the deviation detection model.

5.2 Approach

Figure 5-3 gives a high-level overview of the approach to the deviation detection model development as well as its integration into the work as a whole. A small sub-



(a) Non-deviation (b) Tactical WX Deviation (c) Other Deviation

Figure 5-1: Examples of the three kinds of encounters: non-deviations, tactical weather deviations, and other deviations.

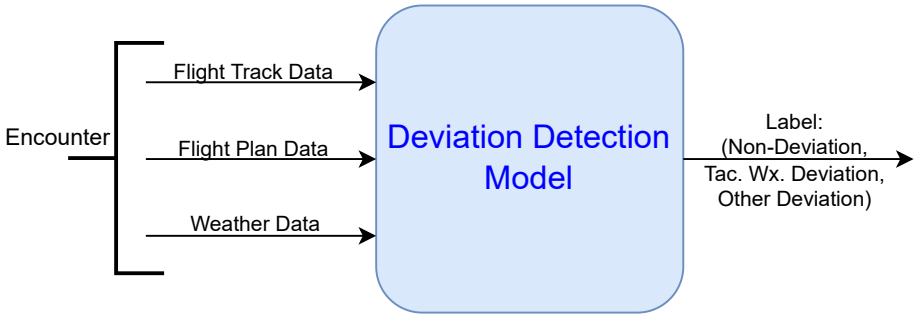


Figure 5-2: Inputs and outputs of the deviation detection model.

set of encounters, the deviation detection dataset, is selected from the large pool of encounters identified in the previous phase. These encounters are assessed by subject matter experts, who assign a behavior label. The encounters and their labels can then be used in a supervised learning approach for constructing a deviation detection model. The resulting model is used to perform automated deviation detection on the remainder of the encounters, identifying non-deviations and tactical weather deviations for use in the deviation prediction model development.

The determination of non-deviation vs deviation (either type) is independent of the weather situation, and can be made based only off the flight plan and the flight tracks. Once non-deviations are identified, the remaining encounters are sorted into tactical weather deviations and other deviations. This results in a two-step classification process, as shown in Figure 5-4. The first step determines if there is a deviation. If there is a deviation, the second step determines if it is likely that the tactical weather

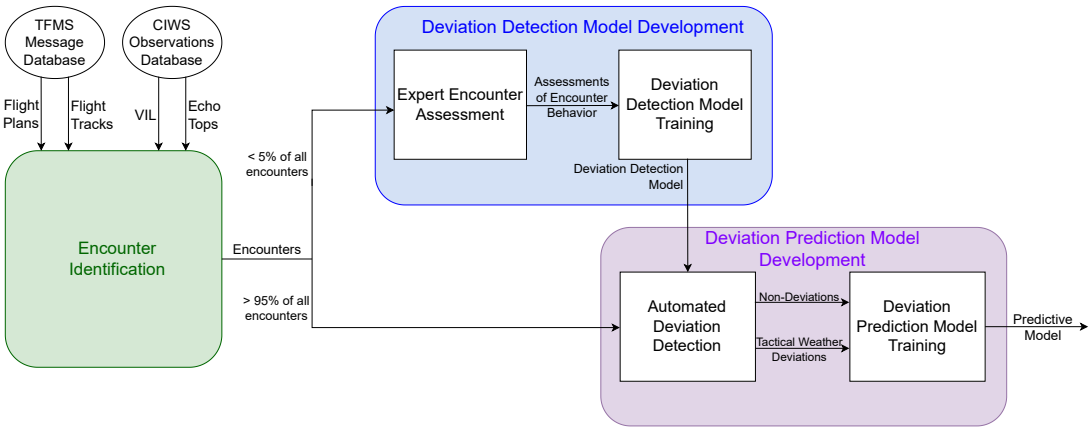


Figure 5-3: An overview of the deviation detection model development process and its integration into the work as a whole.

situation caused the deviation.

5.2.1 Non-Deviation Identifier

Previous work, namely the original CWAM research, attempted to identify deviations based on a "clear-air route width", which was the 90th percentile mean deviation distance on a clear day [10]. Encounters from the study days were classified as deviations if their mean deviation distance exceeded the clear-air route width. All of the studied encounters were reviewed by an expert after being classified to ensure the correct label was applied, which allowed the researchers to evaluate the performance of the deviation classifier. Ultimately, the researchers found that the classifier performance was too unreliable to be used on its own without expert review.

Rather than setting a threshold distance and then evaluating the performance of the threshold, this research used cases that were already labeled to construct a deviation identifier. Notably, this work also only considered weather encounters and did not use clear-air cases for establishing a deviation threshold. The mean deviation distance for each encounter was estimated by approximating the area between the flown and the planned routes, then dividing by the length of the planned route. After obtaining expert behavior classifications for each encounter in the deviation detection

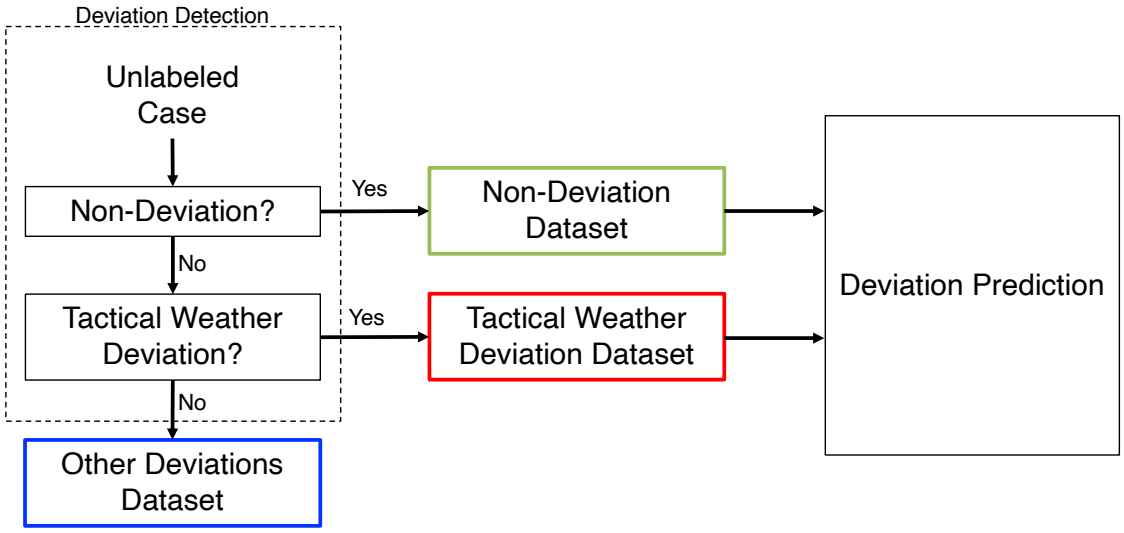


Figure 5-4: Conceptual flow of the deviation detection process, which is split into two distinct steps: non-deviation identification and deviation type classification.

dataset, mean deviation distances were calculated and then a threshold distance was set.

5.2.2 Deviation Type Classifier

A major challenge associated with building the deviation type classifier is handling the structure and dimensionality of the weather data. With an encounter size of 400 km by 400 km and an observation resolution of 1 km, each encounter has 160,000 unique VIL observations at a single point in time, one for each grid cell. Moreover, a great deal of information is encoded in the spatial relationships between the observations, the flight plan, and the flown route. If each VIL observation for each grid cell was to be treated as an independent value, it would obscure these spatial relationships. Additionally, with a number of observations on the order of $10x10^5$, an extremely large dataset would likely be required to learn the relevant patterns.

In previous work, spatial relationships are often captured by applying various two-dimensional filters to the weather data [10], [14], [29]. Typically, these filters are crafted by experts to summarize an aspect of the weather that the expert believes will be relevant to the work. For instance, the original CWAM work found that the

"L3PCT" feature was one of the best predictors of deviation [10]. The L3PCT value is calculated by determining the percentage of pixels in a given area showing a VIP[1] level of 3 or higher. If this value is calculated for a 120 km x 60 km array[2], for example, then those 7200 values have been reduced to a single feature that describes weather intensity. Using a number of these spatial filters allows feature-rich raw data to be represented with a relatively small number of features.

While there are certain advantages to this approach, especially the ability to incorporate domain-specific knowledge, it is not particularly compatible with a data-driven approach. There is limited additional benefit from greatly expanding the number of cases available for training if the number of weather features is held constant. Also, in the data-driven paradigm, it is desirable to make as few assumptions as possible about the structure of the data and which features will be most relevant.

This work uses an image-based approach as opposed to the hand-crafted features approach. Like image data, weather data and flight data are two-dimensional and spatially correlated. Also, important features of weather and flight data may appear in different locations for each encounter - an intense echo may be located one kilometer north of the flight track in one encounter, and one kilometer south of the flight track in another encounter. Similarly, one image of a dog may have the dog's face on the right side of the image, while another image of a dog may have the dog's face on the left side of the image. The same collection of features is present (the face) but a classifier only looking for a face on the right side of the image would not register the face in the second image. Because of the structural parallels between the encounter data and image data, it was hypothesized that treating the encounter data as image data might open new avenues for machine learning on aircraft-weather interactions.

The advantage of presenting the encounter as an image is that it allows the model to simultaneously consider the weather, flown track, and planned track without extensive pre-processing or feature extraction prior to model ingestion. With approaches like the one used for CWAM, the flight data is considered separately from the weather

[1]Video Integrated Processor, another scale for measuring weather intensity

[2]120 km x 60 km reflects a typical "neighborhood" that the original CWAM work would have applied this filter to.

data, and the weather data is used to construct features which are combined with the flight data. Image classification models such as convolutional neural networks are designed to efficiently capture spatial patterns in the training images. This is typically done through alternating layers of spatial filters and pooling. For more information on convolutional neural network architecture, please refer to Chapter 2.

Using an image-based approach also enabled the application of transfer learning techniques. With transfer learning, a model that has already been trained on a large dataset is modified for use with a different dataset. The premise behind transfer learning is that the model has already been trained to extract useful features from the inputs, and the features that the model has learned could be used for a different classification task. Transfer learning helps maximize the utility of small datasets. It has shown success in a number of specialized applications, such as biomedical imaging, where the expertise required to label a dataset makes dataset creation expensive [33].

To test the image-based approach, each encounter that would be classified by the deviation type classifier first needed to be represented as an image. In particular, to maximize transfer learning performance, the encounter image representations needed to be structurally similar to typical images used for training large-scale CNNs, i.e., the encounters would ideally be represented as a 3-dimensional array of RGB values. To accomplish this, the flight plan and flight track, plotted using different colors, were overlaid on a colormapped VIL image[3], and the combined image was converted to an RGB array before being fed into the deviation type classifier.

5.3 Expert Assessment of Weather Encounters

As described previously, a small subset of encounters identified in Chapter 4 was randomly selected for assessment by experts. The expert assessments were used to construct the non-deviation identifier and train the deviation type classifier. Four hundred encounters were randomly selected from each of the five days studied in this work, yielding 2000 encounters in total. The encounters were grouped into 40

[3]Information on the colormapping process can be found in Appendix A.

batches, with 50 encounters in each batch. Batches were distributed to subject matter experts, who applied one of four pre-set labels to each encounter. The labels were used in aggregate to construct a dataset for developing the deviation detection model.

5.3.1 Obtaining Aircraft Behavior Labels

A crowdsourcing approach was chosen for the encounter assessment in order to maximize quantity and quality of the encounter labels. This was due in part to a "scalable approach" mindset - having a single person or small group of people complete all of the labeling is inherently not a scalable effort, and therefore does not have the potential to leverage a very large dataset. Conversely, 20 participants could completely label the 2000 encounters in the initial dataset if each participant assessed 100 encounters. In addition, the inclusion of many different individuals helped to remove bias that might have existed with a smaller number of participants.

The participants needed to be able to look at a combination of weather and flight data and assess the relationship between weather data, flown routes, and planned routes. Ideally, participants would understand what different kinds of deviations looked like and infer the impact of local convective weather systems on the encounter outcome. For this reason, pilots were recruited to label cases. The minimum qualification to participate was an instrument rating, to ensure that the participant had some experience with weather-impacted operations and would be able to assess a weather situation from the information presented.

Because the encounters are limited to commercial flights (Part 121 operations), recruitment mostly focused on pilots who had operated Part 121 flights, or the equivalent for pilots based outside of the United States. This was done primarily through posting on popular online airline pilot forums, emailing aviation-focused email mailing lists, and using personal connections with commercial pilots. Some pilots with general aviation experience also participated, recruited with similar means. Of the 31 participants who returned encounters, 22 were ATP-certified (or its equivalent), 7 were general aviation pilots, and 2 did not specify their certification.

Due to the involvement of human subject participation, MIT COUHES (Com-

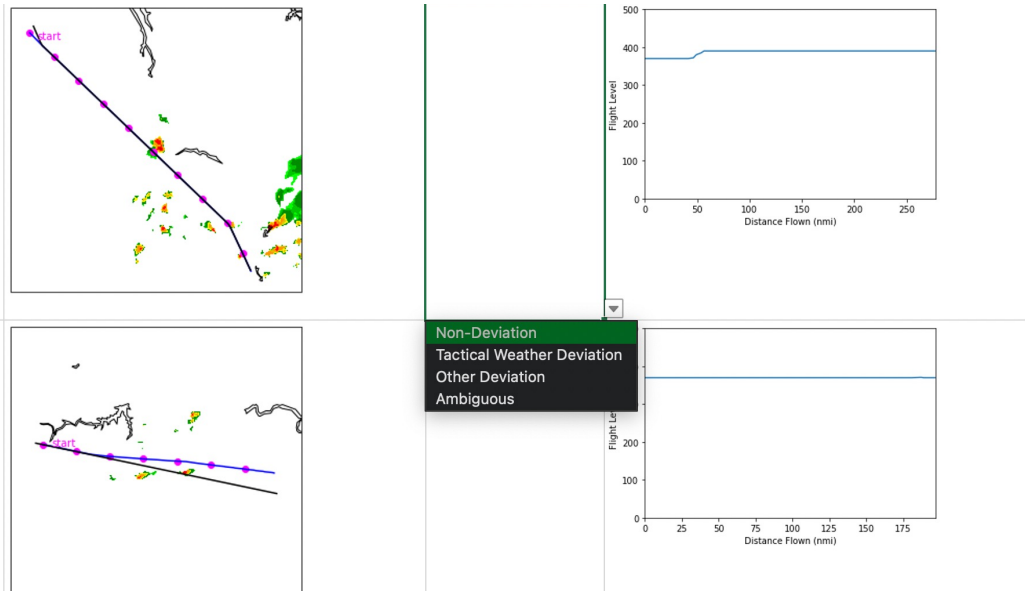


Figure 5-5: An example of the information provided for each encounter to the expert assessors.

mittee On the Use of Humans as Experimental Subjects) regulations applied. This study was classified as a "survey, questionnaire, or interview" and was approved for exemption from COUHES review as a result. Study information was provided to participants and informed consent was obtained before participants were sent any cases to label. The informed consent statement can be found in Appendix B.

Initially, each participant was sent 100 encounters for assessment, along with instructions. For each encounter, an image containing the flown route, the filed route, and the VIL data at the start of the encounter was provided, along with an altitude profile. Experts could apply one of four labels to each encounter: non-deviation, tactical weather deviation, other deviation, or ambiguous. Figure 5-5 shows an example of the information provided for each encounter. Experts were also provided with instructions, which can be found in Appendix C. Some participants were asked to label additional encounters if they expressed interest in further participation.

5.3.2 Results of Expert Assessment of Weather Encounters

Overall, 3050 encounters were returned with labels, with some encounters having at least two independently assigned labels. If an encounter had only two labels and the

Total Number of Labels	Number of Agreeing Labels	Label Confidence
1	1	Low
2	1	No label
2	2	Standard
3	1	No label
3	2	Medium
3	3	High

Table 5.1: Assignment of confidence level to encounter label based on number of total labels applied and number of agreeing labels.

labels did not agree, it was forwarded to a third expert for a tiebreaker label. Each encounter with a valid label was also assigned a confidence level for its label. Table 5.1 shows the relationship between number of labels and the confidence level in the label for any given encounter. In general, there is less confidence in the label when experts disagreed on the correct label. In order to collect cases with high-confidence labels, some batches of 50 encounters were forwarded to a third expert in their entirety.

Figure 5-6 shows the distribution of encounters by final label and label confidence. Note that the chart does not include encounters with no label, in cases where a tiebreaker label was not available. Non-deviations and tactical weather deviations occurred at about the same frequency and were more common than other deviations.

A notable result from the labeling process was a generally low but highly variable inter-rater agreement rate. Using the first two labels for each case, the average inter-rater agreement rate was about 65%. Agreement rates for batches of 50 cases varied between 36% and 88%. This result highlights the difficulty of this problem generally - even using expert humans as classifiers, it is a non-trivial task to draw clear boundaries between different behaviors and consistently apply broad labels. A great deal of information is encoded in the interaction between the weather and the flight, but extracting and decoding this information can be challenging. There are a number of reasons that aircraft might deviate, with shortcuts and traffic re-routing being some of the most common. In some cases, such as in Figure 5-7, the planned route does appear to approach the convective weather more closely than the flown route, but the flown route cuts the corner of the planned route. In this case, though the flown route

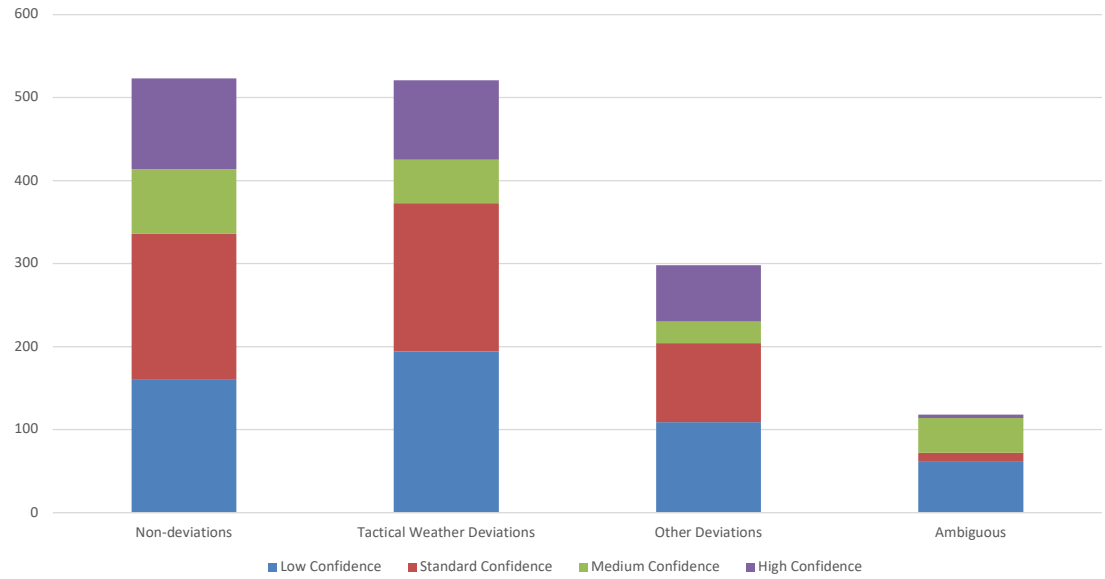


Figure 5-6: Distribution of encounters by final label and label confidence. Encounters with no final label are excluded.

did result in reduced weather exposure, the deviation may have been driven entirely by a desire to reduce the route length. Such shortcuts occur frequently, as it is not uncommon for controllers to "clear direct" to a future waypoint.

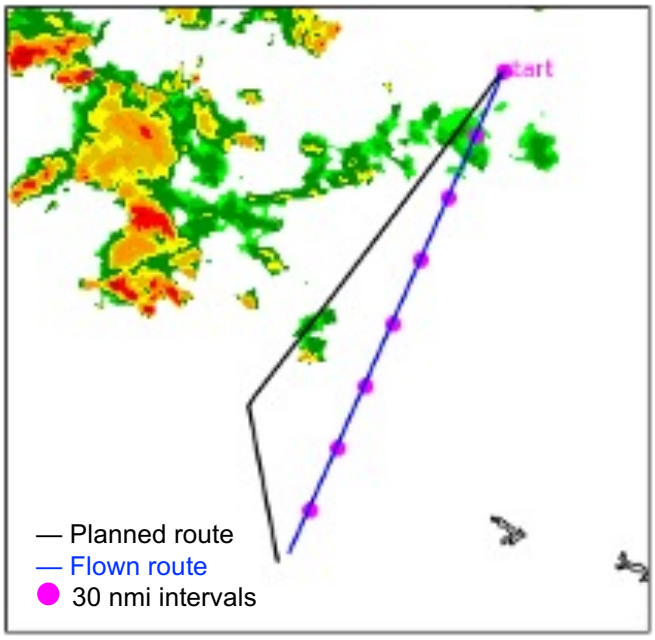


Figure 5-7: An encounter where the correct label may be unclear, since the flown route could be an attempt to reduce the weather exposure or could simply be a shortcut.

The low inter-rater agreement rate implies two things: first, that there is some noise in the ground truth labels being used to train the deviation type classifier; and second, that the difficulty associated with labeling these encounters can vary greatly between encounters, with some encounters having a clear categorization and others being more ambiguous. To reduce label noise in the training dataset, low confidence cases (encounters with a single label) were excluded from further analysis.

It can also be instructive to evaluate which disagreements were most likely to occur. Table 5.2 shows the number of times each pair of labels occurred together on the same case[4]. From the first two conflict pairings, it appears that disagreement over whether a deviation occurred and disagreement over what caused a deviation occurred at roughly the same frequency. For the other conflict pairings, the presence of the "ambiguous" label in many of these pairings makes it difficult to determine the root cause of the disagreement. It is notable that the overlap of "Non-Deviation" and "Tactical Weather Deviation" was relatively small. This indicates that there is little overlap in the behavior that typifies these cases, and that they are relatively distinguishable.

Label 1	**Label 2**	**Number of Occurrences**
Other Deviation	Tactical Weather Deviation	96
Non-Deviation	Other Deviation	94
Ambiguous	Other Deviation	87
Ambiguous	Tactical Weather Deviation	86
Ambiguous	Non-Deviation	26
Non-Deviation	Tactical Weather Deviation	18

Table 5.2: The most common two-way label conflicts.

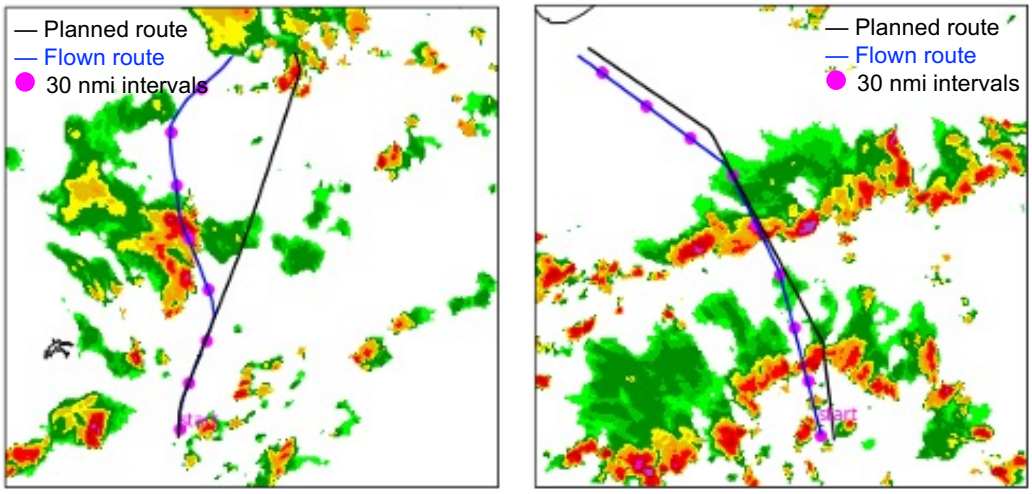
5.3.3 Expert Feedback on Assessment Process

The static nature of the encounter weather representation posed some challenges during the expert assessment phase. The VIL observations shown in each encounter image were the VIL observations at the start of the encounter, which was about 200

[4]Note that in rare instances, an encounter would receive three labels and still lack a final label, since the third label did not agree with either of the existing labels. These cases are not reflected in Table 5.2

km or about 15-20 minutes prior to the point of weather encounter identified along the planned route. However, in cases where there were high winds aloft speeds, it became more difficult to assess the relationship of the flight data relative to the weather data throughout the duration of the encounter. In some cases, aircraft appeared to increase their weather exposure by deviating. An example of this can be found in Figure 5-8a. A number of experts suggested that adding the direction of the winds aloft would be helpful. Knowing the direction of the weather and how the weather picture was likely to change would give the assessor a better idea of the cause of the deviation.

Another common challenge with labeling involved "hybrid" cases - cases where the aircraft appeared to penetrate weather, but then maneuvered around the weather (or vice versa). Generally, this can be thought of as any case where there plausibly may have been two weather-related decisions made. Figure 5-8b shows an example of such a case. Future work could consider developing a more robust set of labels to handle these encounters.



(a) An encounter that appears to increase weather exposure with its maneuvering.

(b) An encounter with mixed non-deviation and deviation behaviors.

Figure 5-8: Examples of encounters that posed special challenges to expert assessors.

5.4 Non-Deviation Identifier

5.4.1 Non-Deviation Identifier Development

After collecting the expert assessments, the next step was to create a non-deviation identifier based on the labeled encounters. Tactical weather deviations and other deviations were collapsed into a single category for this stage. In total, there were 830 encounters with a valid label (i.e., at least two independent labels as well as a consensus among the labels). The estimated mean deviation distance was calculated for each encounter, and the set of expert-labeled encounters was used to establish a deviation threshold value.

5.4.2 Non-Deviation Identifier Performance

The mean deviation distance was very effective in separating the non-deviations from the deviations. Figure 5-9 shows the Receiver Operating Characteristic (ROC) curve, which allows visual inspection of the false positive rate vs the true positive rate. In this case, non-deviations are considered to be the positive class. As the ROC curve shows, the non-deviation identifier is able to achieve a high true positive rate while maintaining a relatively low false positive rate, making it an excellent classifier. The area under the curve was calculated to be 0.991. A perfect classifier would have an area under the curve equal to 1, and an area of 0.991 indicates that the non-deviation identifier performs very well.

Finally, a threshold mean deviation distance needed to be chosen based on the true and false positive rates. Since the positive class is non-deviation, a false positive or false alarm would be a true deviation that has been mis-categorized as a non-deviation. In this specific case, a false alarm could be very detrimental - potentially, a weather encounter that had caused a deviation would be classified as a non-deviation for the predictive stage. Particularly for safety critical applications, it could be very harmful for the model to over-predict non-deviations. On the other hand, it is ideal to capture as many non-deviations as possible in this stage to provide a broad cross-section of

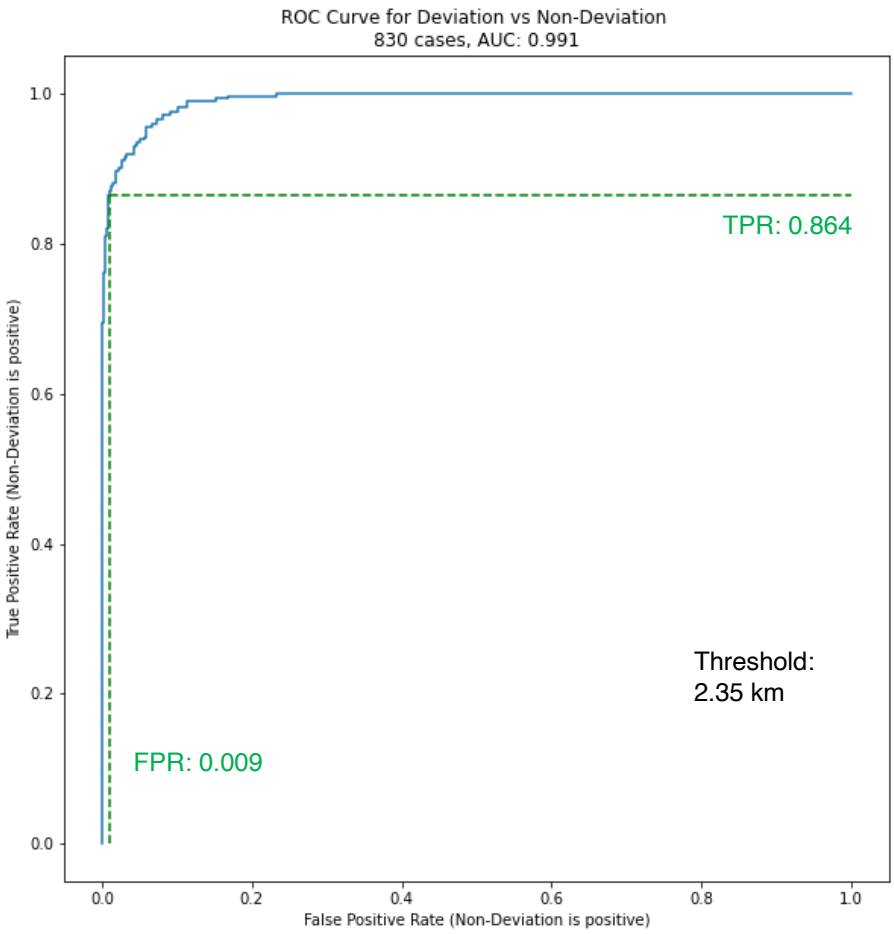


Figure 5-9: ROC curve for non-deviation identifier (non-deviation is positive class). The false positive rate and true positive rate associated with the 2.35 km threshold are shown.

non-deviation cases for downstream modeling efforts. However, given the high cost of a false alarm (a weather-related deviation that is misclassified as a non-deviation), the threshold was set to maintain a false-alarm rate of less than 1%. The resulting threshold had a false alarm rate of 0.9% and a probability of detection of 86%, as shown in Figure 5-9. This corresponded to a mean deviation distance of 2.35km, or 2350 meters.

Using this threshold, 258 non-deviations were identified in the set of 830 encounters, which left 572 deviations for use in training and validating the deviation type classifier. About 14% of the encounter that experts had assessed as non-deviations

exceeded the deviation threshold, and so were included in the 572 deviations. These cases were assigned a label of "other deviation" for the purposes of deviation type classification.

5.5 Deviation Type Classifier

5.5.1 Training and Validation Datasets

The 572 deviation encounters were split into training and validation datasets for the deviation type classifier. An 80/20 split was used, allocating 458 encounters into the training dataset and 114 in the validation dataset. However, the construction of the validation dataset was non-standard. As previously discussed, there existed multiple levels of "label confidence" for the dataset. Encounters with three labels which all agreed had a higher label confidence than encounters with three labels where one of the labels disagreed, since the existence of a disagreeing label indicated some uncertainty in the correct classification. The three kinds of encounters used in training and validating the deviation type classifier were high-confidence, medium-confidence, and standard-confidence encounters (see Table 5.1 for more information on how label confidence was established).

Since the standard confidence validation dataset is the most representative of the general distribution of cases, 50 of the 114 validation cases were allocated to this dataset, and randomly selected from the standard-confidence cases. The remaining 64 cases allocated to validation were divided evenly between the high-confidence and the medium-confidence datasets, and were randomly selected from the high-confidence and medium-confidence encounters, respectively. The remaining encounters in the standard-, medium-, and high-confidence subsets were grouped into a single training dataset. Once in the training dataset, encounters were used the same, regardless of label confidence.

The training dataset was relatively small, with 458 encounters. A common technique to mitigate small dataset size is to implement data augmentation. In the general

sense, data augmentation is simply the application of a transformation to a case or set of cases. These transformations are designed to reduce bias, amplify dataset size, or improve generalizability of the model. An example of bias that might be present in this work would be if convective weather patterns are dominated by movement from west to east. If all of the training cases implicitly assume that weather is moving from west to east, it will impact the model's ability to handle cases where weather is not following that pattern. Data augmentation helps to reduce this bias so the model is more generalizable, and also increases dataset size.

The augmentation strategy was to rotate each case 90, 180, and 270 degrees, as well as mirror cases horizontally and vertically. This helped to reduce any directionality bias present in the original dataset, and increased the training dataset size sixfold to 2748 cases. While the underlying data for an encounter and its augmented versions remained the same, a rotated RGB array is not identical to its unrotated sibling from the point of the view of the model, and the model will extract different features [51].

5.5.2 Deviation Type Classifier Architecture and Parameters

Architecture Considerations

The Python TensorFlow API was used to implement all of the deviation type classifier models. The TensorFlow API interfaces with the Keras machine learning library to build and train neural networks [52]. The library includes a number of state-of-the-art convolutional neural networks that can be used for transfer learning approaches. Early model testing indicated that transfer learning models using ResNet50 as the base model performed best, though the testing was not exhaustive and it is possible that another model architecture would have outperformed the ResNet50 model. Given the wide array of available architectures and hyperparameter choices, performing an exhaustive model evaluation and optimization was outside the scope of this work. The ResNet50 architecture is very popular for use in image classification tasks, including transfer learning tasks for biomedical image classification as well as trajectory classification [33], [53]–[55].

Hyperparameters

Generally, there are a number of hyperparameters that can be tuned to optimize neural network performance. For this work, the ability to optimize these values was limited by the small validation dataset sizes. Due to the large number of variables and the small validation datasets, it can be difficult to ensure that a change to the model structure or hyperparameters has a true effect on the model performance and gains in performance are not the result of random chance. In testing, most performance gains from tuning hyperparameters were modest and the best models were typically close on performance metrics. Still, some general considerations for the most salient hyperparameters are discussed here. Many of these parameters relate to how the model performs weight updates, known as back-propagation. During back-propagation, the model will update its weights a small amount in the direction that would minimize loss [56].

- Epochs: During each epoch, each training example passes through the neural network once. More epochs means more weight updates for the neural net (for a given dataset and batch size). Initial epochs will typically see a rapid improvement in model performance; if nothing else changes, the model will eventually reach a point where additional epochs do not improve performance. This is known as convergence. In testing, models generally reached convergence within 10 epochs.

- Optimizer: TensorFlow provides a number of options for the optimizer used for backpropagation. One of the most commonly used optimizers is Adam, which combines desirable properties from other optimizers into a single method [56], [57]. Adam was used in the construction of the models described here.

- Learning rate: Larger learning rates mean each weight update can be more significant (larger delta from original weight to new weight). This generally means the model reaches convergence more quickly, but can also be unstable or oscillate. Lower learning rates converge more slowly. The default learning rate

for the Adam optimizer is 0.001, but models tended to reach convergence very quickly, so the learning rate used in the models described here was 0.0005.

- Batch size: The batch size determines how many cases' losses will contribute to the backpropagation calculation. Larger batch sizes mean that the gradient is calculated from a larger number of cases. The default batch size for TensorFlow is 32. Modifying the batch size did not change performance in a consistent manner, so the batch size was left at 32.

Dropout

Because the training dataset is small, special care must be taken to avoid overfitting. In addition to the steps already identified that combat overfitting, a dropout layer was also introduced. Dropout layers will hide outputs from randomly selected neurons (the "dropout" of these neurons' outputs) during training. This forces the model to be robust and not become overly reliant on a few key neurons, making the model more generalizable. Dropout tends to increase validation performance relative to training performance, since validation data has the advantage of being run on the full model, while training data is using the model with dropout.

Early Stopping

It is difficult to correctly estimate the optimal number of epochs to train a model for before stopping. To address this, TensorFlow allows the setting of an early stopping condition. At the end of each epoch, certain performance metrics are calculated, such as the accuracy of the model evaluated on the training dataset. The model will stop training when a specified performance metric stagnates, indicating model convergence, then restores the best-performing model weights. Again, this helps to ensure the model is trained to convergence without overfitting the model.

For this work, the validation loss was evaluated each epoch for early stopping. Unlike accuracy, which only considers if each example is labeled correctly or incorrectly, loss values give a measurement of how close the model's prediction is to being

correct. For each case, the model output is a value between 0 and 1. For a positive case (label is 1), the loss will be higher the lower the model output is, because the output is further from the truth value. Binary cross-entropy loss was used in the training of this model, which is a standard loss metric. Validation loss is evaluated instead of training loss because training loss will continue to improve (decrease) when the model begins to overfit.

For the deviation type classifier, early stopping was implemented with a minimum delta of 0.005 and a patience of three epochs. This means that if the validation loss had not decreased by at least 0.005 within the previous three epochs, the model would stop training and return its best weights up until that point.

Dual-Input Model

One challenge encountered in the expert labeling process was handling the temporal aspect of the encounters. Since encounters are represented by a static image, the weather data shown in the image is based on a single timestamp. As many experts noted, lack of information on weather evolution could make some cases more difficult to assess. In order to address this issue, a double-input deviation type classifier was developed, with dual inputs for each case. For this model, the input for each encounter was made up of two images: the standard encounter image and was complemented by another image that showed the weather at approximately the halfway point of the encounter. This gave the neural net more information about weather movement with respect to the planned and the flown routes.

In order to use a dual-image input, the general transfer learning strategy needed to be slightly modified. The pre-trained neural net was duplicated, so each type of input image trained its own base neural net. The classification layer then used features from both images to produce a label.

Final Architecture

The final hyperparameters chosen for the model are shown in Table 5.3. The ResNet50 model was used as the base architecture for the transfer learning model. Pre-processing

Base Model	Learning Rate	Optimizer	Batch Size	Dropout
ResNet50	5e-4	Adam	32	0.5

Table 5.3: The parameters used to train the final deviation type classifier.

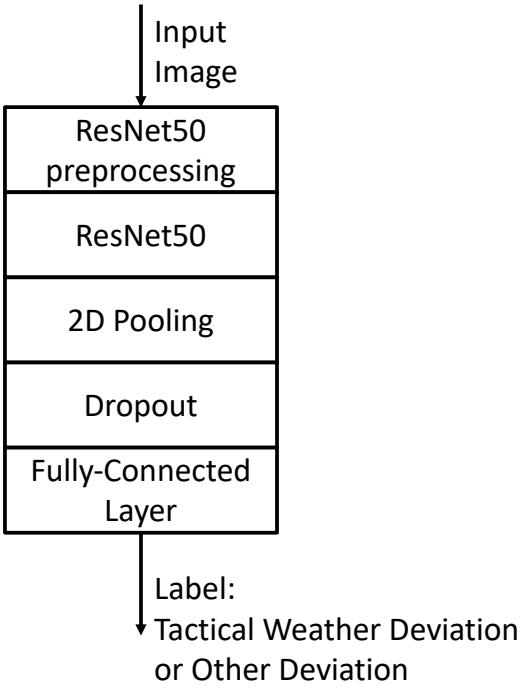


Figure 5-10: Transfer learning architecture of the deviation type classifer.

layers specific to ResNet50 were added before the ResNet50 layers. The ResNet50 layers output into a pooling layer, which connects to a dropout layer followed by a fully-connected layer. The weights in the fully-connected layer are the weights that are trained during each epoch. An illustration of this architecture can be seen in Figure 5-10.

5.5.3 Deviation Type Classifier Performance

Single-Input vs Dual-Input Model

It was hypothesized that neural net performance could be improved by considering weather data at more than one point in time for each encounter. However, this hypothesis was not supported by the results, as shown in Table 5.4.

The single input model trained for eight epochs before the early stopping condition

was met and the best model weights, from the fifth epoch, were restored. The double input model trained for only four epochs before meeting the early stopping condition. In this case, the best weights were from the first epoch. The double input model performs notably worse on the medium-confidence validation dataset, with a much higher loss.

A conclusion that might be drawn from this comparison is that while more information may be helpful for an expert, doubling the amount of information for each input runs the risk of making it more difficult to extract the most important features for a limited dataset. With a larger training dataset, the dual model could have perhaps realized the hypothesized performance gains.

Dataset	**Metric**	**Single-Input**	**Dual-Input**
Standard Confidence	Loss	0.5056	0.4989
	Accuracy	72%	74%
High Confidence	Loss	0.4552	0.5046
	Accuracy	81.25%	75%
Medium Confidence	Loss	0.5281	0.703
	Accuracy	71.88%	65.62%

Table 5.4: Comparison of single-input and dual-input model for the deviation type classifier.

Fine-Tuning

Fine-tuning is sometimes used in transfer learning approaches to further improve performance after the model has reached convergence using the frozen pre-trained feature extraction layers. A small number of layers at the end of the pre-trained model are "unfrozen"; that is, their weights are also updated during the training process, and the model is trained for a few more epochs at a small learning rate. This allows for a closer representation of the target dataset.

Fine-tuning can induce overfitting, where the model begins to "memorize" the training dataset in a way that is not generalizable to other cases. This is evident when the accuracy and loss performance improve for the training dataset, but not for the validation dataset.

Two fine-tuning approaches were tried: one that unfroze the last five layers, and one that unfroze the last 15 layers. In both cases, the learning rate was one-tenth of the original learning rate, and the single input model was used. In the model that unfroze the last five layers, validation loss failed to improve after the first epoch, though training loss did continue to improve. This indicates that even with a very small number of unfrozen layers, overfitting happens very quickly. This phenomenon was also observed with the model with a larger number of fine-tuned layers. The model with the larger number of fine-tuned layers appears to have overfit more quickly, which is an expected result. Compared to the model that was not fine-tuned, the model that was fine-tuned on five layers did show some modest performance gains, so fine-tuning was kept for the final model. The results for the single-input model with and without fine-tuning are shown in Table 5.5. Since the single-input model with fine-tuning performed the best, it was selected as the deviation type classifier for this work.

Dataset	**Metric**	**No Fine-Tuning**	**With Fine-Tuning**
Standard Confidence	Loss	0.5056	0.4881
	Accuracy	72%	74%
High Confidence	Loss	0.4552	0.4441
	Accuracy	81.25%	81.25%
Medium Confidence	Loss	0.5281	0.4951
	Accuracy	71.88%	75%

Table 5.5: Performance of the deviation type classifier on the different validation datasets, both with and without fine-tuning.

5.5.4 Examples and Discussion

A sample of the deviation type classifier results can be seen in Figure 5-11, which shows six randomly selected deviation encounters that have been labeled by the deviation type classifier. Two of the encounters were classified as tactical weather deviations and the remainder were classified as other deviations. At a glance, these results seem generally reasonable. The upper middle tactical weather deviation looks like a classic weather deviation, with an approximately 30° deviation from the filed flight plan followed by a rapid return to the planned route once clear of the most intense

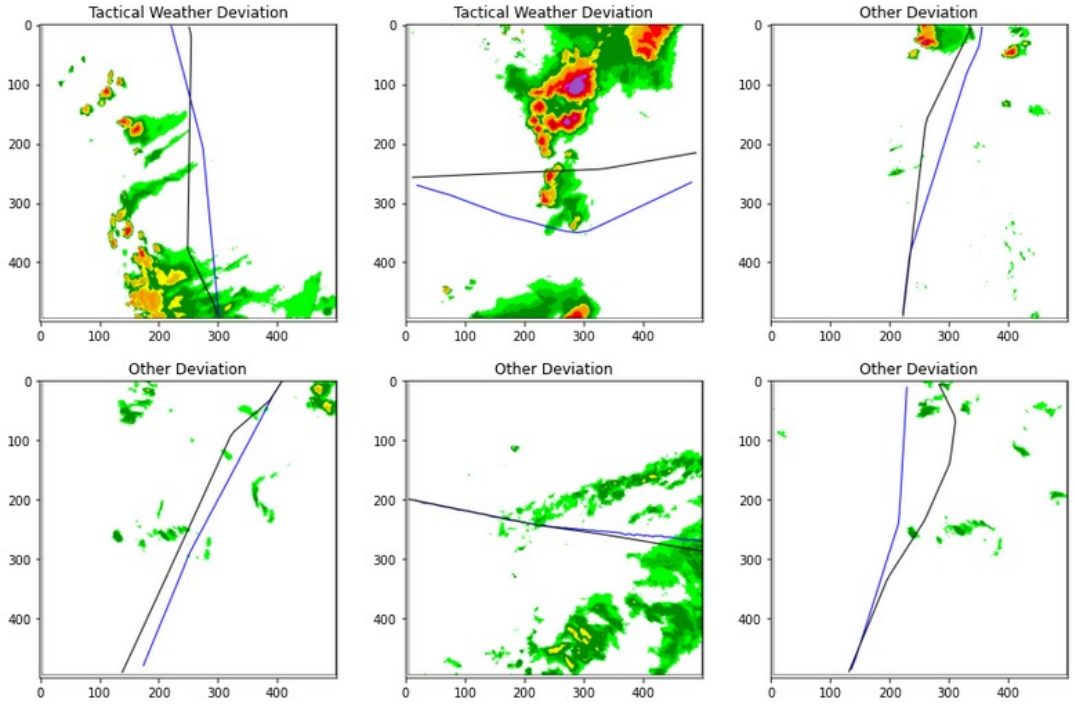


Figure 5-11: Examples of encounters that were processed and labeled by the deviation type classifier, along with the predicted class.

convective cells. Many of the cases classified as other deviations show the flown route taking a shorter path than the planned route, with overall weather exposure being generally mild for both routes.

This is generally the desired behavior for the deviation type classifier. Because the minimum level of convective weather for an encounter is relatively low at 14 kg/m^2 (VIP Level 1), it is expected that there will be a number of deviations where the local convective activity was incidental. The goal of the deviation type classifier is to identify these cases so they can be excluded from the dataset used to train the deviation prediction model. Looking at the cases in Figure 5-11, the model is successful in identifying the clear tactical weather deviation, and the maneuvering in the four cases classified as other deviations could reasonably be due to causes other than weather. The upper left case is a little more ambiguous, and does not appear to fit neatly in either category. Overall, this sample of cases indicates the deviation detection model is performing reasonably well.

When evaluating the deviation detection model, it is important to consider the

performance of the model relative to the inter-rater reliability. The deviation detection model is about 74% accurate on the standard confidence dataset, while the overall inter-rater agreement rate was about 65%. With a low inter-rater agreement rate, it is expected that the performance of the deviation detection model will be somewhat limited.

It was hypothesized that some cases might be more difficult to classify than others, both for experts and for the deviation type classifier. This was tested by evaluating the deviation type classifier performance on the different validation datasets (high, medium, and standard confidence). It was expected that the classifier would perform better on the high-confidence validation validation dataset, since these cases had the highest inter-rater agreement, indicating less ambiguity associated with the encounter behavior. As shown in Table 5.5, the deviation type classifier did perform better on the high-confidence dataset than on the medium or standard confidence datasets.

5.6 Deviation Detection Model Development Summary

Figure 5-12 summarizes the deviation detection model developed in this chapter. The final deviation detection model is a two-step classifier that takes in flight plan data, flight track data, and weather data. Non-deviations are identified based only on flight plan and flight track data. Encounters with deviations are passed to a deviation type classifier, which separates cases by likely cause of deviation. With this model in place, the set of encounters generated in Chapter 4 can be assigned behavior labels for use in a supervised deviation prediction learning task.

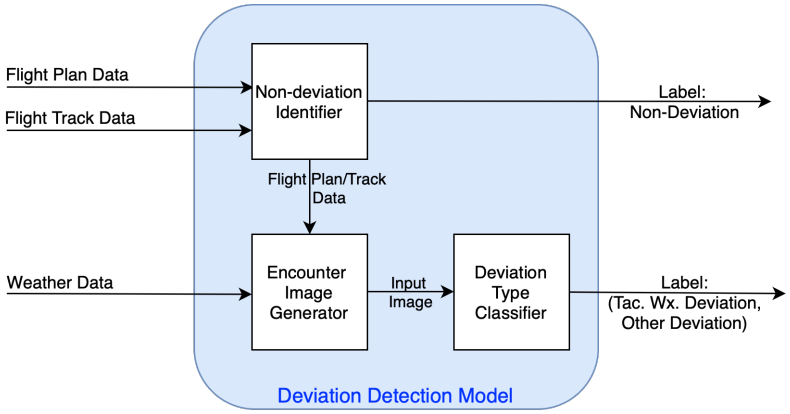


Figure 5-12: Flow of the deviation detection model developed in this chapter.

Chapter 6

Deviation Prediction Model Development

6.1 Overview

In this stage, a model to predict convective weather avoidance is trained, fulfilling the overall objective of this work. Figure 6-1 illustrates the problem formulation for the predictive model. The goal is to take in a current aircraft position, the aircraft's intended route, and the local convective weather situation as quantified by the Vertically Integrated Liquid (VIL) and Echo Tops (ET). Given these inputs, the output of this model is a prediction on whether the aircraft will deviate from its intended route.

The prior stages of this work, the encounter identification and the development of the deviation detection model, were crucial to the development of the deviation prediction model. Figure 6-2 shows the interaction between the different phases of this work. In particular, these previous stages provided the ability to isolate examples where aircraft encountered a nontrivial level of convective weather and either a) did not deviate from their intended path or b) did deviate from their intended path, and the deviation was likely due to the local convective weather situation. These examples of non-deviations and tactical weather deviations were saved for use in training the deviation prediction model.

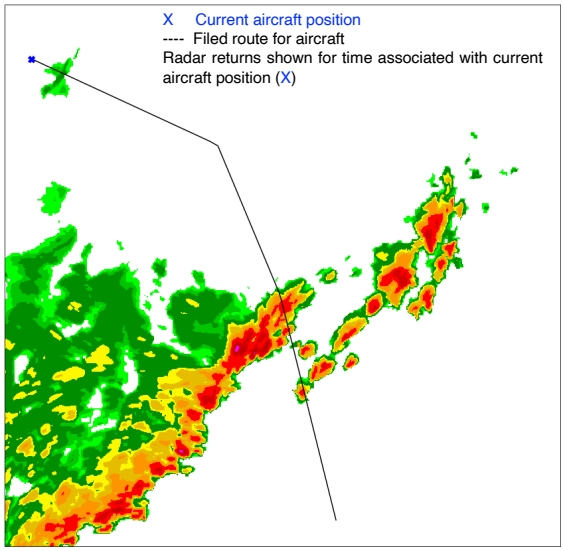


Figure 6-1: An illustration of the problem formulation for the deviation prediction model development stage.

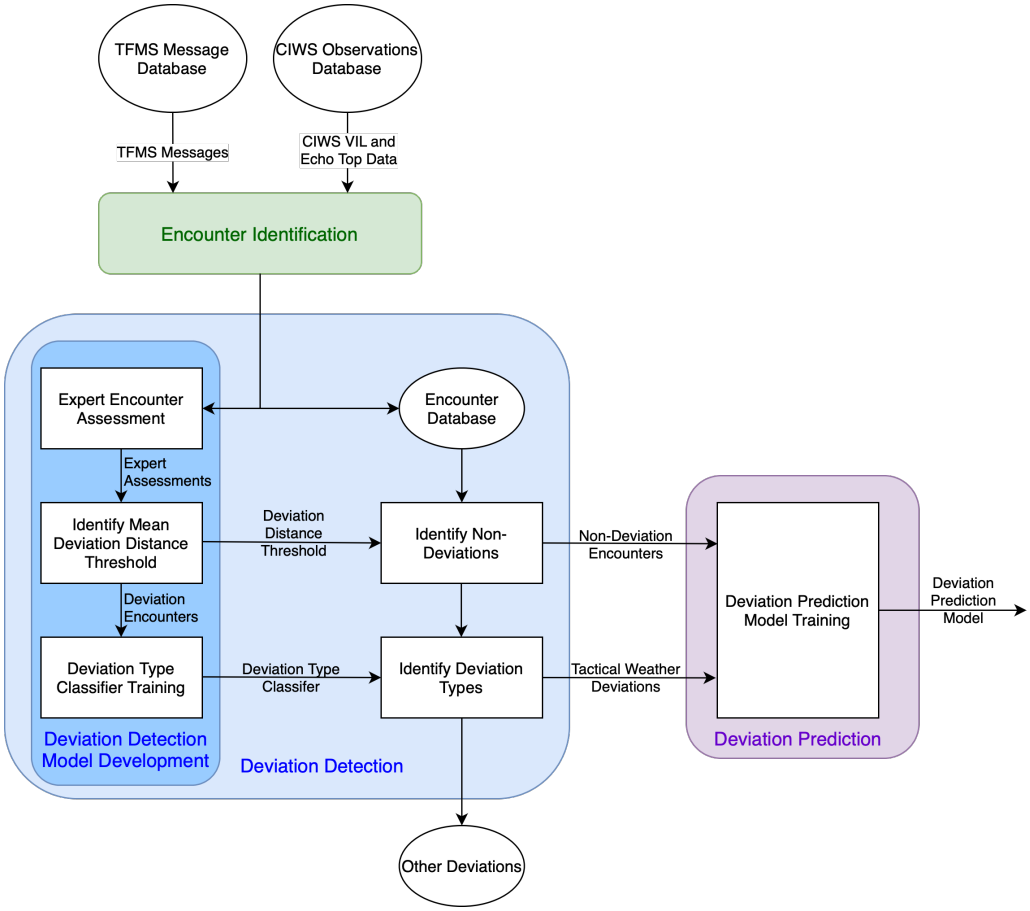


Figure 6-2: The interaction between the deviation prediction stage and different components of the earlier "encounter identification" and the "deviation detection" stages.

6.2 Dataset Construction

The deviation prediction stage started with a large set of weather encounters. As discussed in Chapter 4, the term "weather encounter" refers to a combined set of flight and weather data that records a relevant aircraft-weather interaction. Encounters were identified based on the observed aircraft trajectory, the last flight plan filed in TFMS prior to departure, and local weather observation data (relative to the flight). Weather and flight data from five days in 2019 was processed to identify weather encounters. Running these encounters through the deviation detection classifier output a behavioral label for each encounter. The classifier labeled an encounter as belonging to one of three types of behavior: a non-deviation, a deviation that was likely due to the tactical weather situation, or a deviation that was likely due to causes other than the local weather situation. The non-deviations and tactical weather deviations were be used to train the predictive model.

As discussed in Chapter 5, sorting encounters into behavior categories was a two-step process. First, for each encounter the mean deviation distance (MDD) was calculated by finding the area between the flown route and the planned route, then dividing by the length of the planned route. An MDD threshold was established based on hand-labeled cases to differentiate between deviations and non-deviations. Encounters with an MDD below this threshold were saved as non-deviations.

For encounters with an MDD at or above the non-deviation threshold, the appropriate encounter images were generated and run through the deviation type classifier. The classifier distinguished between deviations likely caused by the tactical convective weather situation and deviations likely caused by factors other than the tactical convective weather situation.

Then, predictive model input images were generated for each of the non-deviation and tactical weather deviation encounters. The non-deviation and tactical weather deviation datasets were then divided into training, test, and validation datasets. This overall process is summarized in Figure 6-3.

Nearly 50,000 unlabeled encounters were sorted into non-deviations, tactical weather

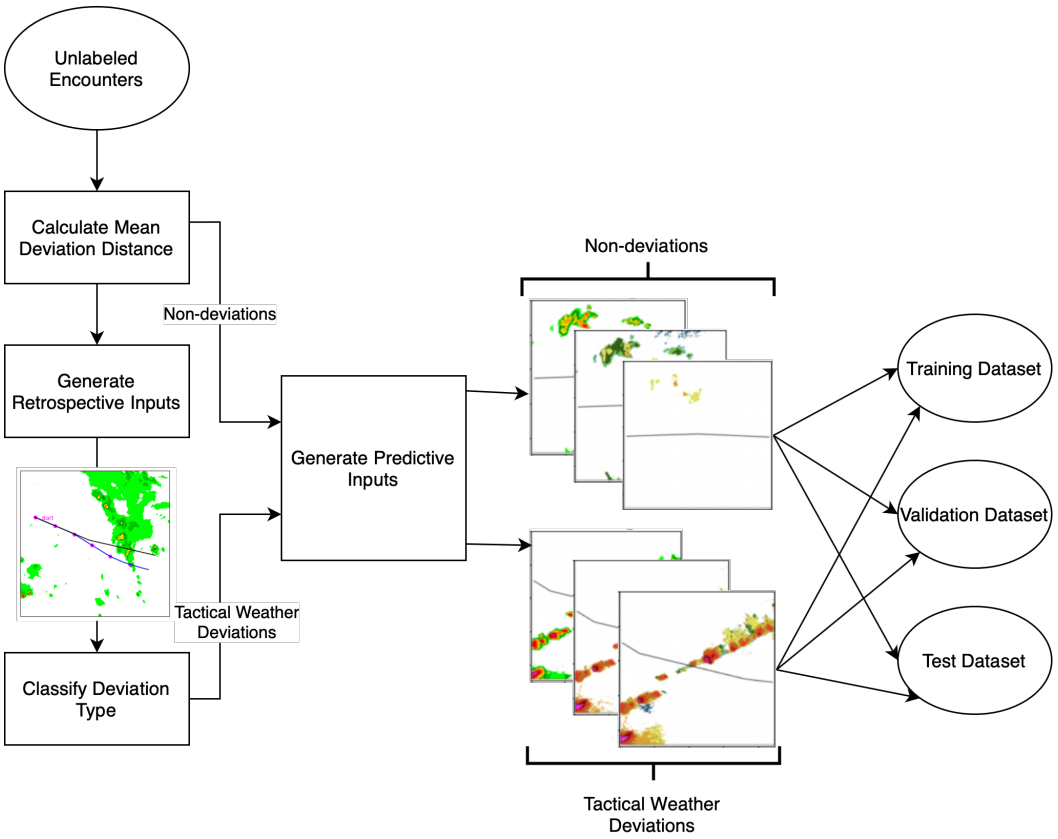


Figure 6-3: The data pipeline that transforms unlabeled encounters into the datasets needed to build and evaluate the predictive model.

deviations, and other deviations. Figure 6-4 shows the breakdown of these 50,000 cases into their respective groups.

As with the deviation type classifier, an image-based approach was used for the deviation prediction model. This decision informed the construction of the inputs for the deviation prediction model. The key data to include for each case was the planned route and the local convective weather observations. As with the deviation detection stage, the route and weather data are spatially correlated. However, the deviation detection stage only considered the VIL and the flown/planned routes, which are all 2-D datasets. The echo tops dataset, included in the deviation prediction stage, added a third dimension to the data. While the echo tops dataset itself is two-dimensional, with a single value for each grid square, the vertical distance between the storm height and the aircraft was the most predictive feature in the original CWAM model. For deviation prediction, it was important to represent not just the lateral relationship

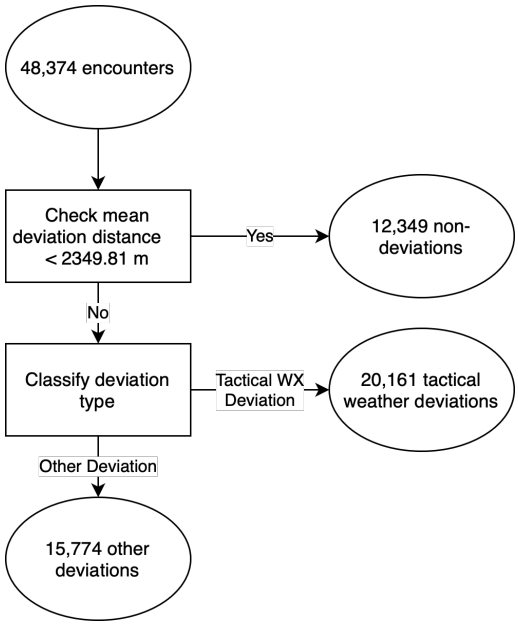


Figure 6-4: The quantitative breakdown of the encounters into each of the three classes of behavior: non-deviation, tactical weather deviation, or other deviation.

between the flight and the local storms, but also the vertical relationship.

Ultimately, the final data representation took the form of three two-dimensional images that collectively formed the input for a single encounter. The first image was very similar to the image generated for the deviation detection stage for each encounter, but did not include the flown route alongside the planned route, since the goal was to make a prediction and not a retrospective classification. The other two images contained spatial representations of the echo tops data along with the planned route. One of these images showed all of the echo tops data within the encounter bounds, colored according to the standard CIWS ET color scale, along with the planned route. The other image was similarly structured but masked all of the echo tops data below the altitude of the aircraft. In other words, grid cells would not show convective activity if the aircraft would overfly the storm at its current altitude. Using two images to represent the echo top activity allowed both the absolute and relative echo top data to be considered while maintaining the two-dimensional image structure required for use in a pre-trained convolutional neural net. Figure 6-5 shows the three-image input for two different encounters. The top set of images

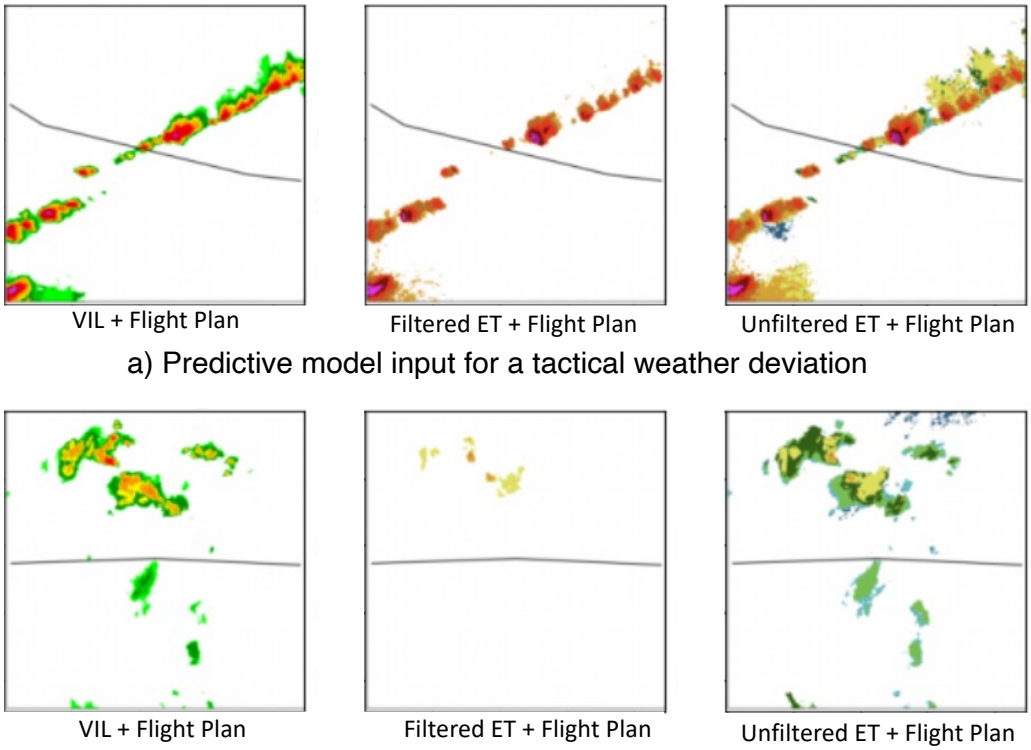


Figure 6-5: Example inputs for the deviation prediction model. The top row of images is for a tactical weather deviation, and the bottom row is for a non-deviation.

is from a tactical weather deviation, and the bottom set of images is from a non-deviation. Comparing the "filtered ET" image for the two encounters helps illustrate the predictive value of this data representation. Much of the convective activity for the non-deviation appears to drop out when the aircraft altitude is considered.

To allow for parallel computing and reduced file size, encounters were processed and stored by date. After all of the non-deviations and tactical weather deviations for each day were identified and transformed into predictive model input images, there were 10 total datasets, a non-deviation and a tactical weather deviation dataset for each of the five days examined. The final step for data processing was to use these datasets to generate training, test, and validation datasets. The training dataset was the dataset used to train neural net weights. The validation dataset was used to tune neural net hyperparameters, such as number of epochs of training. The test dataset was used as a final indication of the model's performance on a "raw" dataset, since it

is possible that the final model selected had been overfit to the training and validation datasets. An 80/10/10 split was used for training, test, and validation.

6.3 Classifier Development

6.3.1 Classifier Options

In keeping with the image-based approach, a convolutional neural network (CNN) was chosen as the base architecture for the deviation prediction model. Within the broad class of convolutional neural nets, a wide variety of structures could be used. Since this training dataset included a reasonably large number of examples, one option would be to design a custom CNN and train the model from scratch. Alternatively, a transfer learning solution could be implemented in which a fully trained CNN could be adapted for the deviation prediction task by training a small number of final layers on the deviation prediction dataset while keeping feature extraction layers static. Finally, there was a hybrid approach in which a pre-designed neural net architecture would be implemented but all layers were trained from scratch on the deviation prediction data. All three options were tried and compared before selecting a single approach to develop further.

Custom Model

Generally speaking, the standard building block of a convolutional neural net is a convolutional layer followed by a max pooling layer. These blocks are stacked in succession, with a classification head at the end. For the custom CNN, four blocks of convolutional layer + max pooling layer were used. Even this small CNN had almost 450,000 trainable parameters. With a relatively small number of examples compared to the number of trainable parameters, more layers were unlikely to improve performance, simply due to the limited amount of training information available. Additionally, increasing the number of layers presents new challenges. For example, the vanishing/exploding gradient problem is a well-known hurdle in deep learning

research. There are solutions to this problem, but in general, more layers will not necessarily improve predictive performance without complementary adjustments that allow the model to take advantage of the additional layers [58].

Transfer Learning Models

For strategies that involved use of a pre-designed model, multiple models were evaluated. Again, the TensorFlow Python package was used to manage the data pipeline, including implementation and training of CNNs. This package comes with a number of ready-to-use CNNs which can be initialized with pre-trained weights or trained from scratch. The CNNs included in TensorFlow are some of the best-performing image recognition neural nets available. The CNNs tested included DenseNet121, ResNet50, and MobileNet. Due to the large dataset size, all models were implemented, trained, and validated on the Lincoln Lab Supercomputer (LLSC) [59]. To accommodate the triple-input structure of the encounter inputs, three CNNs were used in parallel, each one taking in one of the three input images for each case. A single layer combined the outputs from these independent CNNs to make a prediction. Figure 6-6 shows the general structure of this approach.

6.3.2 Classifier Selection

Multiple neural net architectures were implemented, trained, and then evaluated. The most promising architecture was selected for further optimization. Implementation and training required selection of certain hyperparameters, such as batch size and learning rate. Since there are many variables that must be set when training any given CNN, it was not feasible to try every combination of variable and architecture. Instead, reasonable or default values were chosen for these hyperparameters, and were generally kept the same between models when suitable to do so. The basis for comparison between models was primarily the binary cross-entropy loss, and average time for each training epoch was also recorded. The results of this comparison are shown in Table 6.1.

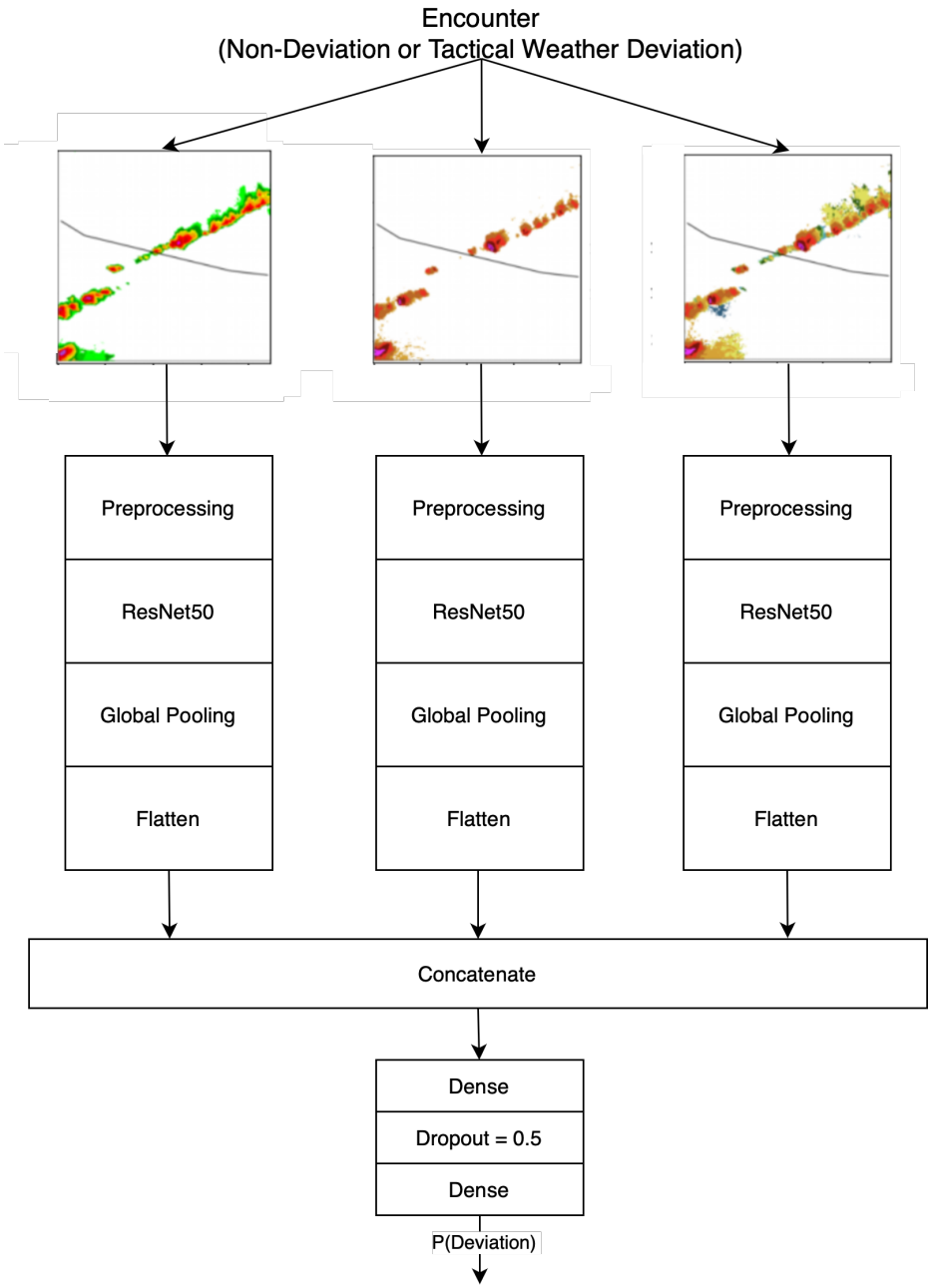


Figure 6-6: The triple-input architecture used for the predictive model. Three instances of a base model are instantiated in parallel, along with separate pre-processing and output layers. Outputs are flattened and concatenated into a single vector, which is fed into a set of final output layers.

Base Model	Pre-Trained	Learning Rate	Best Validation Loss	Avg Time per Epoch (s)
ResNet50	Yes	5e-5	0.4092	428
DenseNet121	Yes	5e-5	0.4457	436
MobileNet_v2	Yes	5e-5	0.4221	172
Custom model	No	1e-3	0.4683	142
Custom model	No	5e-5	0.4714	151
ResNet50	No	1e-3	0.4754	1258
ResNet50	No	5e-5	0.4766	1222

Table 6.1: A comparison of different model architectures that were implemented and evaluated for use in a predictive model. The first three models were transfer learning models and were trained on a very large set of generic images first. The other models were initialized with empty (random) weights.

All of the models in Table 6.1 were trained using a batch size of 32 and a dropout rate of 0.5. Learning rates are as shown. Larger learning rates were initially used for the models trained from scratch, since these models do not have a trained baseline to work from and learning rates that are too slow may cause the models to get stuck in a local minimum. When these models performed poorly, a lower learning rate was also trialed, which did not yield better performance. All of the models converged relatively quickly, within 15 epochs at the most. The validation loss shown in Table 6.2 is the best validation loss, using a binary cross-entropy function, within the first 15 epochs. As with the deviation type classifier development, validation loss was used as the basis for comparison to help avoid model overfitting.

In general, the models trained from scratch underperformed the models with pre-trained weights. While significantly larger than the initial hand-labeled dataset, the machine-labeled dataset used to train these models still contained fewer than 50,000 examples. Pre-trained models are typically trained on datasets containing over well one million examples [58]. Thus, it is not surprising that the dataset used in this work is not large enough to take full advantage of the additional complexity of the pre-trained deep CNNs. Additionally, there are computational benefits associated with using pre-trained models. Models are much quicker to train when only a classification head or a few final layers are being trained, and not thousands or tens of thousands of individual weights.

Model ID	Learning Rate	Epochs	Validation Loss	Area Under ROC Curve	Avg Time per Epoch (s)
A	5e-4	7	0.3525	0.894	441.0
B	1e-4	9	0.3522	0.895	442.2
C	5e-5	5	0.3532	0.891	474.5
B(fine)	5e-5	10	0.3546	0.902	476.8

Table 6.2: A comparison of predictive model performance with different learning rates. Model B was selected for optimization with fine-tuning, yielding slight improvements in the area under ROC curve.

Given the advantages of transfer learning solutions in both performance and computation, a pre-trained model was chosen for further development. There were only minor differences in performance between the different pre-trained CNNs used. While ResNet50 did slightly outperform the other two CNNs in terms of validation loss, MobileNet performed almost as well with significantly less time per epoch. The training time was not a limiting factor for this work, so the best-performing architecture (ResNet50) was selected as the base model for the deviation prediction model. However, if future work examines the effects of a larger training dataset, it may be useful to use MobileNet as the base model to allow for faster training and iteration on models.

6.3.3 Final Model

Once ResNet50 had been selected as the overall model architecture, model hyperparameters were tuned. Learning rates varied between 0.0005 and 0.00005, as shown in Table 6.2. As with the deviation type classifier, early stopping was used to reduce overfitting. Once the validation loss was no longer improving, the model would cease training and revert to the weights yielding the best performance (in this case, the lowest validation loss)[1]. Finally, the best-performing model was fine-tuned by unfreezing the last five layers of the base model, reducing the learning rate to $\frac{1}{10}$ of the original learning rate, and training for a further 10 epochs.

[1]See Chapter 5 for more information on early stopping techniques.

6.4 Classifier Results

6.4.1 Overall Results

The results of the final classifier are visualized in Figure 6-7 with a receiver operating characteristic (ROC) curve, which shows the relationship between the false positive rate and the true positive rate. For the deviation prediction classifier, the positive class is tactical weather deviation. This ROC curve was generated using the test dataset, which had been previously unused in the model selection and tuning process. This is a standard practice in machine learning model development and evaluation. If the ROC curve is significantly worse for the test dataset vs the validation dataset, the model hyperparameters may have been overtuned, and the model may not generalize well. In this case, the area under the ROC curve (AUC) for the test dataset is 0.915, which slightly outperforms the AUC of 0.902 for the model on the validation dataset.

Figure 6-7 also shows where select deviation prediction thresholds fall on the ROC curve. For each example that is run through the predictive CNN, the output is a number between 0 and 1. This output value may be thought of as the model's confidence that the example belongs to the positive class. Generally, any example with an output value of 0.5 or higher is predicted to be a deviation case, and any example with an output value below 0.5 is a non-deviation case. However, by adjusting the threshold at which a case is predicted to deviate, the probability of detection and the false alarm rate can be adjusted.

6.4.2 Performance on Expert-Labeled Dataset

The deviation prediction model was trained and evaluated entirely on encounters labeled via the deviation detection model described in Chapter 5. To further verify the model performance, an expert-labeled dataset was generated using the encounters and expert assessments from the deviation detection model development. The results of this comparison are shown in Figure 6-8. There are some slight differences in the shape of the two ROC curves. This can be partially attributed to the different

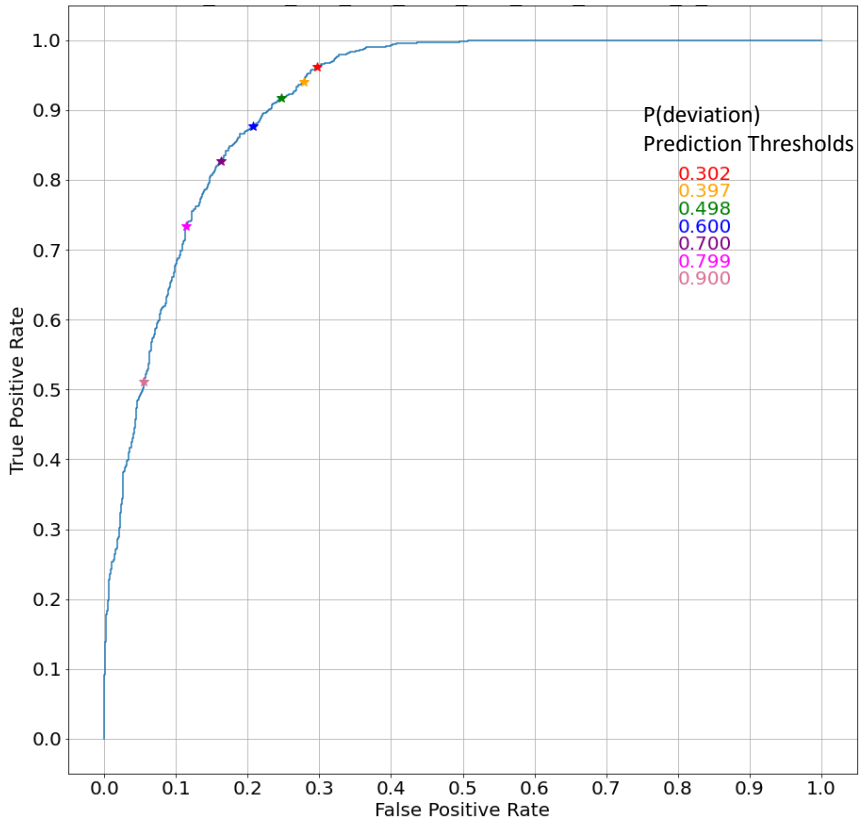


Figure 6-7: ROC curve showing performance of final predictive model on the test dataset. Different thresholds for classifying inputs as tactical weather deviations are shown on the curve. The area under the curve is 0.915.

.

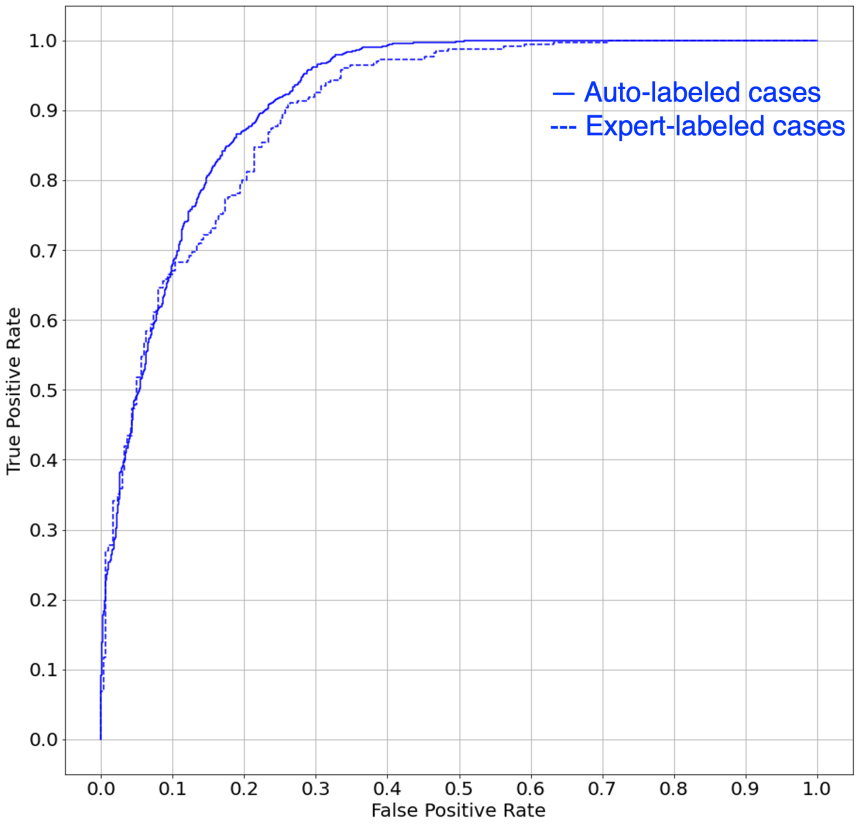


Figure 6-8: Comparison of deviation prediction model performance on the automatically labeled dataset and on the expertly labeled dataset.

dataset sizes. The expert-labeled dataset has only 633 cases, while the machine-labeled test dataset has 3260, so the expert-labeled ROC will be comparatively less smooth. The area under the ROC curve for the expert-labeled dataset is 0.899, compared to the AUC of 0.915 for the machine-labeled test dataset, indicating largely similar performance.

6.4.3 Comparsion to CWAM

Finally, the deviation prediction classifier performance was compared to the original (2006) Convective Weather Avoidance Model performance. Figure 6-9 shows the relative performance of CWAM and the classifier developed in this work. The researchers generated a number of different models, but only Model #9 is shown in the comparison, because it was the model the researchers chose to further analyze and discuss in the published work [10]. For the same false alarm rate, the new model increases

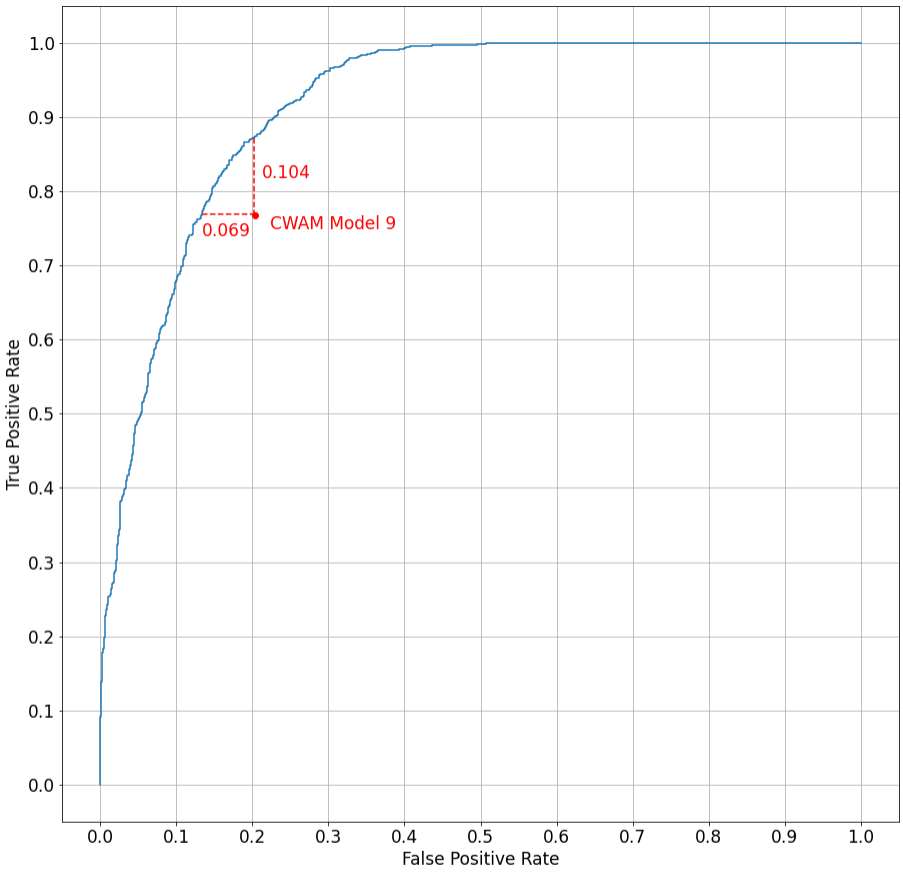


Figure 6-9: Comparison of the deviation prediction model to CWAM (2006) performance.

the probability of detection by more than 10 percentage points, as shown in Figure 6-9. For the same probability of detection, the false alarm rate decreases by almost 7 percentage points.

6.5 Examples

Figures 6-10 through 6-13 show some example inputs that were fed to the deviation prediction classifier. A non-deviation encounter and a tactical weather deviation encounter were selected from the sets of encounters with a high (>0.7) and a low (<0.3) deviation prediction classifier output value. Again, the output value will directly scale with the classifier's confidence that the encounter belongs to the positive

(tactical weather deviation) class.

6.5.1 Cases with High Deviation Prediction Classifier Output Values

Figure 6-10 shows the input images of a true tactical weather deviation and Figure 6-11 shows the input images of a true non-deviation. The output value for both of these inputs was 0.81, meaning that the model had a relatively high confidence that both cases would result in a deviation. Visually, the convective weather in each encounter is generally similar, with some scattered intense echoes visible in the reflectivity (VIL) image and some cells with echo tops at or above the altitude of the aircraft and in close range of the planned route. The visual similarity is unsurprising, given that the classifier gave the same output value to the cases. These two cases help illustrate that different pilots will have different responses to similar weather conditions, and their decisions may be impacted by factors other than weather, such as traffic conditions. Because these cases have the same classifier output value, 0.81, any decision threshold would misclassify one of the two cases. However, the classifier is behaving as expected in the sense that it predicts the same outcome for similar combinations of weather and flight data.

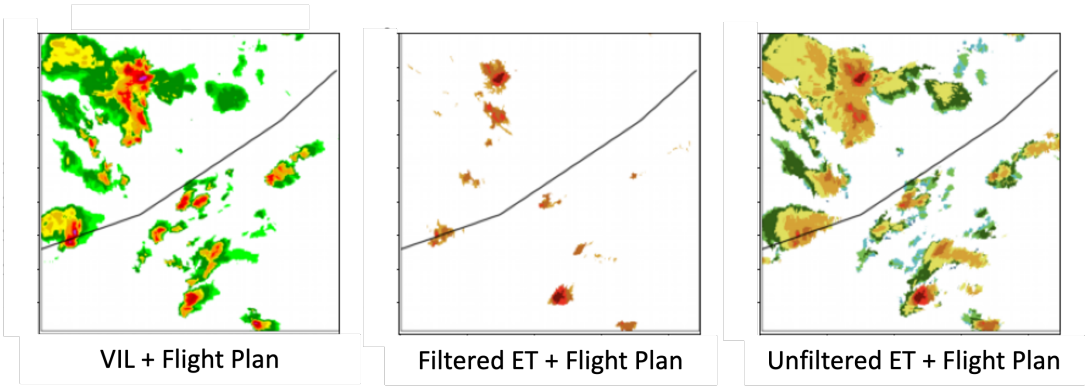


Figure 6-10: Input images for an encounter with a ground truth label of tactical weather deviation. The classifier output was 0.81, indicating relatively high model confidence that the input images belong to a tactical weather deviation.

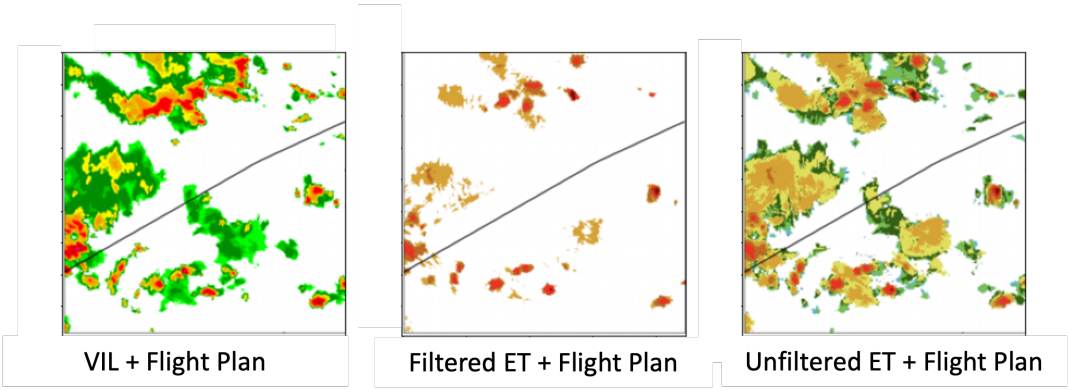


Figure 6-11: Input images for an encounter with a ground truth label of non-deviation. The classifier output was 0.81, indicating relatively high model confidence that the input images belong to a tactical weather deviation.

6.5.2 Cases with Low Deviation Prediction Classifier Output Values

Figure 6-12 shows the input images of a true tactical weather deviation with a classifier output value of 28. Figure 6-13 shows the input images of a true non-deviation with a classifier output value of 18. In both cases, there is very little convective activity at or above the aircraft altitude, and there are relatively few convective cells with high reflectivity values. Again, the fact that the responses to these two cases varied indicates that more information would likely be needed to correctly predict their outcomes.

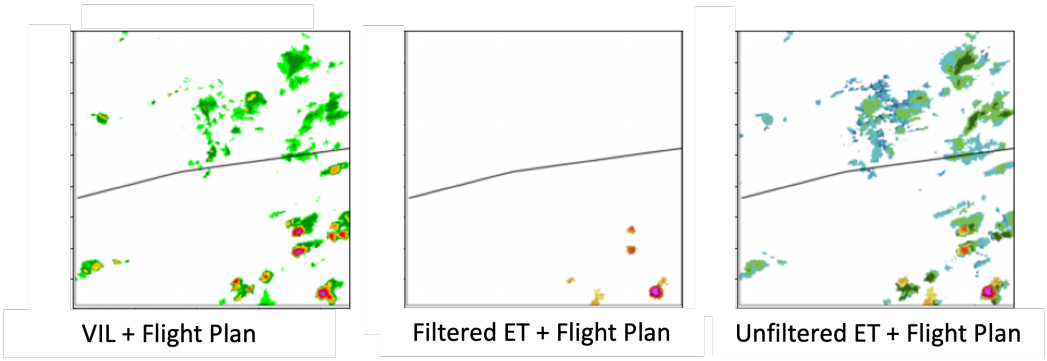


Figure 6-12: Input images for an encounter with a ground truth label of tactical weather deviation. The classifier output was 0.28, indicating relatively low model confidence that the input images belong to a tactical weather deviation.

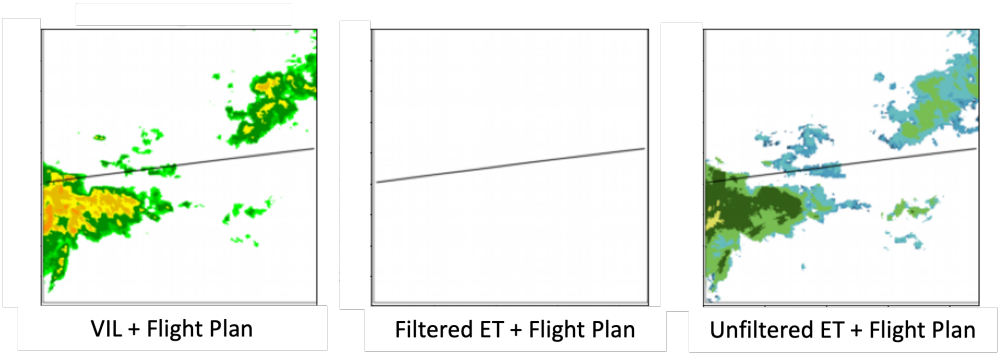


Figure 6-13: Input images for an encounter with a ground truth label of tactical weather deviation. The classifier output was 0.18, indicating relatively low model confidence that the input images belong to a tactical weather deviation.

Chapter 7

Discussion

With the data pipeline in place, transforming raw traffic and weather data into a predictive model, this chapter takes a step back to consider the models that have been developed. The performance of the deviation detection and deviation prediction models have been discussed briefly, but a more detailed analysis is warranted. Additionally, it is important to consider the limitations of the model development, and how these models can be improved in the future. Finally some operational considerations are discussed, as well as some applications outside of the originally intended context of air traffic control.

7.1 Comparison with Previous Aviation Weather Avoidance Research

The original Convective Weather Avoidance Model created in 2006 (and expanded upon in 2008) has served as the foundation for many aviation weather tools. It also informed this work in a number of ways. The original CWAM study first identified the need for a deviation detection algorithm. The factors identified as most predictive of deviation served as a starting point when deciding what weather features to include in our modeling efforts. In particular, CWAM researchers found that the difference in altitude between the echo tops and the aircraft cruising altitude was the most

significant feature for predicting a deviation. This information directly influenced the way echo top data was represented to the deviation prediction model. CWAM and its successors have set the standard for convective weather avoidance research, and as such make a good benchmark for comparison.

7.1.1 Deviation Detection Model

The focus of the 2006 CWAM work was not on developing a robust deviation detection model, but the work did develop a model, albeit one that was considered insufficient at the time. The lack of an adequately performing model was a considered a significant limitation by the researchers. Conversely, this work does focus on developing a deviation detection model for use on large-scale datasets. There are also some key differences between the problem formulation for the CWAM work and this work. The CWAM work did not make an attempt to distinguish between deviations for weather and deviations for other reasons. Instead, the work only sought to distinguish between deviations and non-deviations.

The CWAM deviation detection model development process considered five days of air traffic in the Indianapolis and Cleveland Air Route Traffic Control Centers (ARTCCs). Briefly, the methodology consisted of 1) establishing a clear-air route width for each supersector, 2) calculating the mean deviation distance for each encounter, and 3) using the mean deviation distance to determine if the encounter was outside the clear-air route width (deviated). Ultimately, encounters from the Cleveland ARTCC were not used in the predictive model development. The general deviation behavior in the Cleveland ARTCC differed significantly from deviation behavior in the Indianapolis ARTCC, and there were too few deviations in the Cleveland ARTCC.

For the Indianapolis ARTCC that was kept, the probability of non-deviation detection across all days was 87.9%. The false alarm rate for non-deviations was 10%. In comparison, the non-deviation identifier for this work has a probability of non-deviation detection of 86.4%, and a false alarm rate of 0.9%. In other words, for almost the same probability of detection, the false alarm rate was an order of magni-

tude lower.

At least in terms of deviation vs non-deviation, the model developed in this work offers some clear benefits. It is not immediately clear why the model performance is significantly improved, given that they are essentially using the same measurement (mean deviation distance). Notably, the threshold for a deviation in this work is about 2.5 km, as opposed to the 12 km deviation threshold used for Indianapolis ARTCC encounters in the CWAM work. It may be that today, aircraft typically fly closer to their filed routes with the widespread availability of precise navigational procedures.

Additionally, the definition of "non-deviation" is slightly different between the two approaches. In this work, because deviations not for weather were expected to be filtered out at a later step, almost any sort of maneuvering would cause the encounter to exceed the deviation threshold, regardless of weather conditions. In the CWAM work, the deviation threshold was set based on the clear-air route width. Though shortcuts and re-routes were excluded, other kinds of non-weather-related maneuvering would have been factored into the clear-air route width, increasing the deviation threshold value. This works very well if encounters that result in a weather-related deviation have a significantly larger mean deviation distance than encounters with a deviation unrelated to weather. However, if the maneuvering is happening on roughly the same scale for both weather and non-weather deviations, it will be difficult to distinguish between the two.

7.1.2 Deviation Prediction Model

As discussed in Chapter 6, the deviation prediction model developed in this work outperforms the original CWAM model. However, the CWAM work was revisited in 2008, with more cases and a new set of weather features to evaluate [12]. Flight data was collected from the Washington, D.C. ARTCC in addition to the original Indianapolis and Cleveland ARTCCs. New predictive models were generated to find the most effective combinations of weather features as well as the performance impact of including more features in the model.

The researchers found that the overall performance increase was very modest with

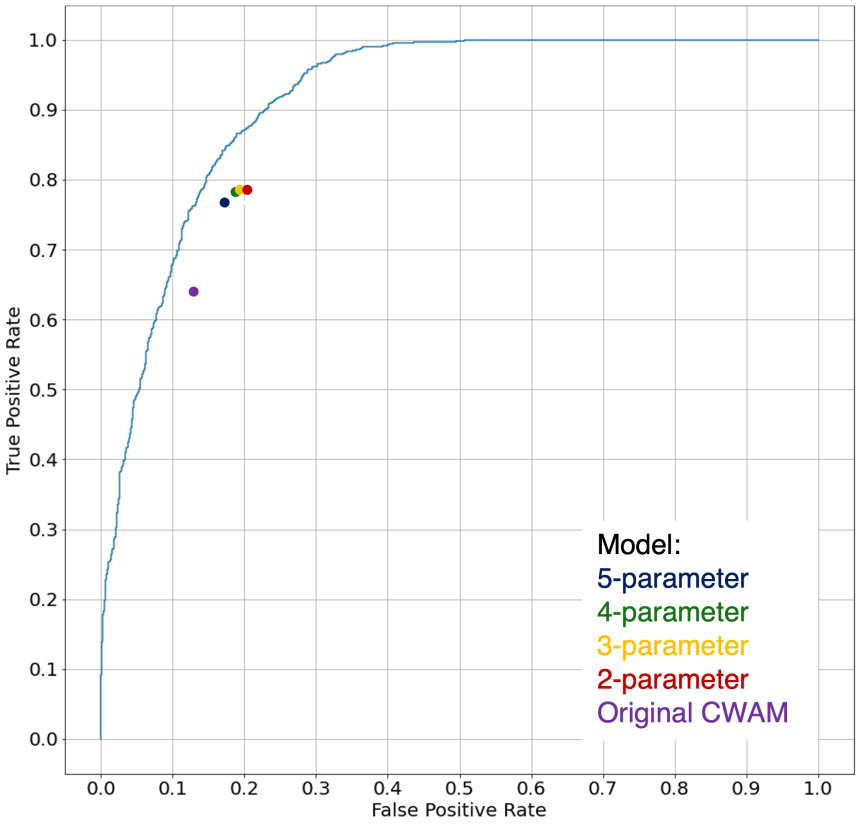


Figure 7-1: Comparison of deviation prediction model to CWAM (2008) models.

the new weather features and larger dataset size. Moreover, increasing the number of features included in the model did not yield significant performance benefits. A summary of the performance of these models, as well as the performance of the original CWAM model on the new dataset, is compared to the deviation prediction model developed in this work in Figure 7-1.

It should be noted that the relatively constant overall performance hides significant differences in the models' performance on different ARTCCs. Additionally, some models had very large spreads in accuracy for the different classes. For example, when evaluated only on the Washington, D.C. encounters, the three-feature model had about a 15% prediction error for deviations, but over a 30% error for predicting non-deviations.

In addition to outperforming the other models, the advantage of the model pre-

sented in this work is that it is a more flexible baseline from which to build other models. This work only considered VIL and Echo Tops, but it is relatively simple to add historical weather or forecast data (e.g., trends or storm motion). If the image-based approach is used, no new features would need to be crafted to incorporate the new data, making it easier to expand the model.

7.2 Limitations

7.2.1 Static Weather Information

Using an image-based paradigm for deviation detection and prediction meant that the weather represented in each example image was only the weather at a single instant in time, while the flown aircraft track could represent over thirty minutes' worth of flight data. For both the deviation detection and the deviation prediction problems, the timestamp of the weather data in the encounter images was the time at the beginning of the encounter. In the deviation detection problem, the intent was to give the experts the picture of the weather as it was when the aircraft entered the tactical decision-making range. For the deviation prediction problem, the image is simulating a real-time prediction where future weather data is unavailable. However, in both problems, the lack of weather trend data may have limited model performance.

A common remark from the subject matter experts who labeled the initial set of cases was that there was no information on weather movement in the example images. The actual point of weather exposure was typically about fifteen minutes after the timestamp on the weather data used to generate each example. In some cases, this could have made it more difficult to assign the correct behavioral label and overall may have contributed to low inter-rater reliability.

While the deviation prediction task did not allow for use of "future" weather information, i.e., the weather data beyond the start of the weather encounter, additional past weather information may have been useful for improving model performance. Generally, aircrew will consider movement of convective cells when deciding to devi-

ate or stay on the route. Including weather trend information would better replicate the information available to controllers and aircrew.

7.2.2 Inter-rater Reliability

Low inter-rater reliability was also a significant limitation in this work. Many labeled cases were unused because the low overall inter-rater reliability meant that there was low confidence in single labeled cases. This reduced the number of cases usable for the deviation detection model development from about ?? to 830. There were also a number of double-labeled cases with no consensus. Overall, it was difficult to capitalize fully on the expert assessments because of the low agreement.

Additionally, the low inter-rater reliability presents challenges for evaluating the the deviation type classifier. The deviation type classifier performance is quantified on the encounters that have been labeled by experts, but it is known that there is some disagreement as to the ground truth label for some of these encounters. In other words, some cases may not have been correctly assessed by the experts, and this impacts the ability to evaluate overall performance.

This uncertainty carries over, to an extent, to the deviation prediction model training and evaluation, since these datasets were labeled by the deviation detection classifier. However, the impact is mitigated by the fact that non-deviations are identified via a separate process, which has a much higher level of accuracy. The main effect of inaccuracies in the deviation type classifier would be the presence of true "other deviation" encounters in the tactical weather deviation dataset. This could cause the deviation prediction model to overpredict tactical weather deviations, since weather conditions that do not necessarily merit deviation are being fed to the deviation prediction model as weather deviations.

7.3 Future Work

Suggestions for improvement are focused on improving deviation type classifier and deviation prediction model performance. The non-deviation identification step al-

ready performs very well and further improvements would likely be marginal. Improving performance of the deviation type classifier would help reduce some of the uncertainty in quantifying predictive model performance, and could potentially improve the performance of the predictive model by providing more consistent training datasets. In addition to addressing some of the weather data limitations discussed previously, there are avenues for tweaking deviation detection performance. Improving deviation detection performance would be the ultimate goal in either case.

7.3.1 Improving Inter-Rater Reliability

Low inter-rater reliability did pose a significant limitation, and would be a good place to start for improving the framework. The intent would be to improve the deviation type classifier via providing more consistent behavioral classifications. A better deviation type classifier would potentially improve deviation prediction by cutting down on the number of cases not requiring deviation that have a tactical weather deviation label.

Providing more weather information to the expert assessors, particularly information about weather trends and movement, might resolve some of the ambiguity for tougher cases when using experts to assess and provide better ground-truth labels. As discussed above, one of the challenges of using a single image is that the dynamic component of the weather is hard to capture. Having better knowledge of how the weather situation evolved would likely reduce some of the ambiguity and increase inter-rater reliability. This could take the form of providing multiple images or providing storm motion or growth trends products.

Additionally, using ATC recordings in conjunction with the images could help produce a dataset with very high inter-rater agreement. The timestamp and location of the aircraft at the beginning of an encounter is already known; if ATC recordings are archived in a way that allows indexing by center and timestamp then the relevant chunk of the recording could be identified. However, it is not practical to have experts listen to ten minutes of ATC communication for each example, much of which would not be relevant to the case at hand. One mitigation for this might be limiting use of

ATC recordings to cases where experts disagree, or where experts indicate a need for more information. Alternatively, speech-to-text could be employed to translate the recording into a transcript. Speech-to-text for air traffic control recordings can be particularly challenging due to distorted speech over radio, and until recently was not considered mature enough for large-scale use [60]. However, some modern work has shown a good deal of promise in reliable ATC transcription [61]. Of particular note is the ability to recognize call signs. If it is possible to identify snippets of relevant communication for a flight, then the additional information provided to experts could be limited to only pertinent messages, reducing the labeling task burden.

7.3.2 Expanding Model Input Data

For both the deviation type and deviation prediction classifiers, providing additional information to the model could potentially yield performance improvements. This includes weather information as well as other data, such as data on air traffic in the vicinity.

Deviation Detection

This work did perform a cursory evaluation on the inclusion of additional weather information for the deviation type classifier. The deviation detection modeling phase included testing of a dual-input model. In the dual-input model, a case consisted of two example images: one with the weather shown as it was at the beginning of the encounter (the standard image), and one with the weather shown as it was in the middle of the encounter (roughly around the point of weather encounter). This was in contrast to the single-input model, which had weather information from only the beginning of the encounter. It was hypothesized that the additional weather information in the dual-input model would improve performance for the deviation detection classifier as compared to the single-input model. However, the single-input model very slightly outperformed the dual-input model. The extra weather information did not confer a benefit.

It is possible that there was potential for a benefit, but it was unable to be realized. The additional weather information would be most likely to change the outcome for ambiguous cases where experts disagreed. However, if the experts disagreed and no third expert-applied label was available for the case, or if the third label did not match either of the first two labels, the case could not be used in the training or validation process. With fewer of these ambiguous cases in the dataset, the advantages of additional weather information might have been too slight to detect. The small dataset size may also have precluded the model from finding useful patterns in the additional data.

There is also a potential for using the ATC communications for large-scale deviation identification and not just as an adjunct for human labeling. Intentional deviations from the intended route will always be acknowledged in ATC communications. Using transcripts and natural language processing, after a deviation is identified, communications around the point of deviation could give evidence for the cause of the deviation.

Deviation Prediction

There are also many opportunities for improvements to the deviation prediction process. For this work, a minimalist approach was taken, and a limited number of weather features were considered, namely the VIL and Echo Tops observations from CIWS. As with the deviation type classifer, other weather data may be useful for improving predictive power. This includes other kinds of weather observations as well as dynamic VIL and Echo Tops data. The potential usefulness of winds aloft data and/or trend data is supported by studies that examine pilot decision-making for convective weather. For example, Temme and Tienes found that pilots named wind speed and wind direction as significant factors when considering a deviation [50]. Simply including weather observation history in the model inputs may yield improvement over the current inputs. Weather history in the form of a pilot report (PIREP) may also be informative, particularly in corridors with high traffic where multiple flights are likely to be in trail on the same route of flight within a short period of time.

Additionally, research has shown that there are factors other than weather that might influence weather penetration or deviation. Temme and Tienes found that six of the eleven pilots interviewed named phase of flight as a consideration when deciding a weather avoidance strategy [50]. Other research on terminal area weather avoidance has found that non-weather predictive features include the time spent in the terminal area, the total range of the flight, and if the aircraft is flying a route that other aircraft have flown recently [22]. While the study that identified these factors was focused on terminal area weather penetration, it is not unreasonable to infer that external factors may also influence deviation decisions in en-route airspace. Traffic considerations may also a useful inclusion to future models, particularly because some pilots will request a deviation for very minor storm presence if traffic is light and the controller can accommodate such a request[1].

7.3.3 Model Architecture Improvements

Finally, some improvements made be made to the model architecture itself. In this work, transfer learning was used to mitigate performance deficits caused by a relatively small training dataset. Ordinarily, with a small dataset, we would be restricted to shallow neural networks. Because the neural network has already been trained, we can take advantage of much deeper architectures. However, recent research suggests that ensemble models may help overcome limitations associated with small datasets [62]–[64]. This is a particularly relevant problem for biomedical image classification, where obtaining correct labels can be a very expensive task since the labels are generally only applied by highly trained physicians. Liz et al. and Islam and Zhang both found that ensemble classifiers composed of shallow CNNs and trained on a small dataset outperformed deep learning/transfer learning solutions [63], [64]. Phung and Rhee similarly demonstrated high performance of an ensemble of simple CNNs with a small dataset, though there was no direct comparison to deeper models [62].

The model architecture could also be modified to be more robust to noisy labels. As previously discussed, the deviation type classifier does not produce 100% accurate

[1]Discussion with one of the labeling SMEs on overall weather avoidance strategy

labels. Therefore, the datasets used for training and evaluating the deviation prediction classifier are noisy. This is not a unique problem, and various methods have been devised to make classifiers more robust to training data with noisy labels [65]–[68].

Another way to take advantage of off-the-shelf resources would be to investigate predictive model performance when the base predictive model has been trained on a weather-specific dataset. Hadad et al. tested transfer learning solutions for an MRI image classification task on two different models: one was a deep neural net trained on a very large but general dataset, and the other was a shallower neural net trained on an X-ray image classification task with a small dataset. The shallower neural net proved to be superior to the deeper neural net when transfer learning performance on the MRI images was evaluated [69]. The authors hypothesized that the MRI and X-ray images were more similar in structure, which meant the feature extraction layers trained on X-ray images were more applicable and useful for the MRI image classification task. Similarly, using neural nets that have been trained on weather radar data may yield performance benefits for both the deviation prediction and deviation type classification tasks.

One of the key challenges for using this approach would be obtaining a labeled dataset large enough to train a model from scratch. Fortunately, other researchers have recognized the potential value in a large, annotated, structured weather dataset that is available to weather researchers and can be used as a baseline dataset for machine learning techniques. The Storm Event Imagery (SEVIR) dataset, available as of 2020, contains weather data from over 10,000 weather events, including VIL data [70]. Many of the weather events in the dataset have been mapped to storm descriptions in the NOAA Storm Events Database, which can help fulfill the need for a large, labeled dataset. Images from the SEVIR VIL dataset, along with accompanying storm descriptions, could be used to train a shallow neural network on a storm classification task. The trained neural network would then be used as the basis for transfer learning on the weather avoidance prediction task. There is an advantage to this approach over simply generating an extremely large weather encounter dataset and training a predictive model from scratch. Creating a baseline, generalized weather-focused neu-

ral network would retain the advantages of using transfer learning to quickly generate customized prediction models on smaller datasets.

7.3.4 Other Potential Avenues

Increasing Dataset Size

With the deviation detection model, the deviation type classifier was trained on a fairly small dataset. Only 830 cases were available in total, and 258 of those were identified as non-deviations. The 80/20 split for the remaining 572 cases left only 458 cases in the training dataset. Many cases were unusable, because they had either only a single label or two disagreeing labels with no tiebreaker. Realistically, any increases in dataset size requiring intensive human expert involvement will be modest, but even an increase to 1000 or 2000 cases could yield performance increases, particularly if steps are taken to increase the dataset quality as well.

Increasing the dataset size would also allow for better comparison between the different validation datasets. Some cases will remain harder to distinguish (unless an absolute truth source like ATC communication is included) so the different confidence validation datasets would likely remain. Quantifying these differences and why they exist could further help improve the automated deviation detection. To reduce overall labeling burden, only deviations would need to be considered, with non-deviations being first filtered out.

With the deviation prediction model, increasing dataset size is fairly easy thanks to the deviation detection model. The focus of this work was not necessarily to optimize the deviation prediction model but rather provide a framework for developing and improving the predictive model, so the additional resources to increase dataset size were not expended. However, with appropriate improvements to the deviation detection model, it would be reasonable to process more cases for the deviation prediction step. More cases might make it desirable to train a custom neural network, though this CNN would likely still be shallower than something like ResNet50. It would be interesting to baseline the performance of the deviation prediction model

with increasing dataset size to determine where diminishing returns on larger dataset sizes are happening.

Customized Prediction Models

One of the primary benefits of the off-the-shelf methodology described in this work is the ability to quickly generate a customized predictive model. For example, the CWAM follow-on study in 2008 found significant variations in predictive performance when models were evaluated in different ATC super-sectors. To address these gaps in performance, a specific predictive model could be built for each super-sector. Using the encounter generation process outlined in Chapter 4 and including some geographic filters would yield a set of unlabeled encounters that are representative of behavior within a specific super-sector. These encounters would then be labeled by the deviation detection classifier, and the non-deviation and tactical weather deviation cases would form the training and evaluation datasets for the predictive model. Using the same off-the-shelf CNN transfer learning methods, the regional weather avoidance prediction model is trained. The same process could be applied to specific weather conditions as well. When aircraft altitudes are very close to echo top heights, the existing predictive models struggle [10], [12]. Again, a large set of cases that meet this criteria could be identified, labeled, and used to generate a model that is tailored for the more difficult weather situations.

7.4 Roadmap to Operations

Ultimately, the vision for this work is deployment as a real-time decision aid for air traffic controllers. The predictive model approach aligns neatly with the FAA's move to trajectory-based operations (TBO). In a TBO framework, air traffic management centers around 4-D flight trajectories (latitude, longitude, altitude, and time) to plan and optimize air traffic flow [71], [72]. With real-time weather information and the flight trajectories already stored and updated in the TBO system, all the data needed to run the current state of the flight through the predictive weather avoidance

model is provided. Essentially, there would be continuous online testing for route acceptability with respect to the current weather conditions. Preemptively informing the controller that a tactical weather deviation might be necessary could reduce the amount of back-and-forth communication generated by an unexpected deviation request. Additionally, if a controller is devising alternate routes due to severe weather conditions, they would be able to use the classifier to predict if the alternate route will be acceptable.

It would be worth exploring some of the model improvements discussed above before bringing this work into an operational context. In rough order of implementation complexity, from simplest to most complex, those suggestions include increasing the deviation prediction dataset size, including more weather and non-weather features in the model, devising custom models for specific geographical areas or specific weather circumstances, doing transfer learning off of a neural net originally trained for weather-related tasks, and implementing robust methods for handling noisy label data. It would also be useful to study the optimal ratio of true positives to false positives in an operational context. This ratio would be application-specific, and it would be helpful to have feedback from aircrew and air traffic controllers regarding the relative costs of missed alerts and false alarms.

Additionally, improving consistency between SME labels as well as performance of the deviation type classifier would be desirable prior to operational deployment. The SME labels are the ground truth for all of the modeling efforts, both retrospective and predictive. The current low inter-rater agreement rate indicates that either there is not information present in some cases to make a decision, or the behavioral labels are not jointly but independently covering the sample space. More SME labels per case may provide more insights on which cases are hardest to categorize and why those ambiguities arise. Some possible factors leading to behavior label ambiguity are discussed in Chapter 4. Model performance varied depending on how confident we were in the labels for that particular dataset. As expected, the deviation type classifier performed best on the "high confidence" (3/3 SME label agreement) dataset. Further stratifying the dataset may yield additional insight on why some cases are

more difficult than others, and how the more difficult cases should be handled.

7.5 Implications for Autonomy

While the primary focus of this effort was on air traffic control needs and improvements to the ability to optimize air traffic flow in the NAS, there are also some implications for the development of autonomous air vehicles. As the potential for beyond-visual-line-of-sight (BVLOS) operations expands, the onboard situational assessment capabilities of autonomous vehicles must also increase. Learning from previous weather interactions is particularly attractive when discussing higher-order processes like situation awareness and assessment. Aircraft interactions with weather are complex and a rule-based behavioral model would struggle to capture much of the decision-making process that humans go through. However, learning patterns from previous aircraft-weather interactions may allow the autonomous vehicle to imitate human decision-making. In that sense, much of what has been discussed in this work could be applied in an autonomous vehicle context.

However, there are two major breaks between the air traffic control context that this work is geared for and the autonomous operations context. The first is simply that the current predictive modeling approach asks a narrow question and gets a simple answer – "yes, the aircraft will deviate", or "no, the aircraft will stay on its planned route." For an autonomous vehicle, a more holistic evaluation of the weather would be required. It would also be desirable to have a more instructive model that could more readily be used for trajectory generation and modification. The current predictive model assumes that there is an operator with a route in mind, and only gives feedback on that route. The second major break is that this work very explicitly focused on Part 121 airline operations. The current state-of-the-art for autonomous air vehicles is mostly focused on much smaller aircraft, and the certification challenges associated with an autonomous airliner are significant. In the event that such certification was desired and realistic, it would be better to have more explainable results than this work currently does. The use of CNNs allowed leveraging state-of-the-art techniques

when it comes to image classification tasks, but this approach did not value the explainability of the CNN predictions. It is more of a "black box" approach, which may be fine in a decision aid but is problematic when the model is a responsible decision-making entity. Still, certain elements of this work, like the encounter generation method, may prove to be useful in building datasets for autonomous vehicle research.

Chapter 8

Conclusion

8.1 Summary of Thesis Objectives and Approach

This work sought to better predict tactical weather deviations for en-route aircraft. In particular, a major aim of the work was to explore potential avenues for leveraging existing traffic and weather datasets to better understand interactions between aircraft and convective weather. Previous convective weather avoidance modeling approaches have often relied on relatively constrained datasets with limited ability to scale the methodology to large datasets.

There were three major phases to this work. First, an algorithm for automatically identifying potentially relevant aircraft-weather interactions was developed. This algorithm was used to generate a large set of "weather encounters," short segments of flights that were exposed to a minimum level of convective weather. These weather encounters could then be used in downstream modeling efforts. Next, a deviation detection model was built based on experts' assessments of a small set of weather encounters. The resulting model was used on the large set of weather encounters, assigning each encounter to a behavioral class. Finally, encounters in the non-deviation and tactical weather deviation behavioral classes were used to train and validate a deviation prediction model.

8.2 Summary of Thesis Results and Analysis

Using the approach described above, a deviation prediction model that outperforms current state-of-the-art predictive weather avoidance models was developed. The deviation prediction model was trained entirely on machine-labeled encounter data, but the trained model performs similarly on machine-labeled data and expert-labeled data. Besides substantiating the deviation prediction model, the performance on expert-labeled data helps demonstrate the efficacy of the deviation detection model. While not perfect, the deviation detection model performance was sufficient for generating a usable machine-labeled dataset.

Evaluation of the expert-labeled encounters revealed that non-deviations and deviations were distinguishable with a high degree of confidence based only on flight data. Notably, the threshold for declaring an encounter a deviation based on mean deviation distance was set much lower for this work than similar thresholds from previous work. This work used a threshold of about 2.5 km, compared to the CWAM deviation thresholds of 12 km and 24 km [10].

Determining if deviations were intentional weather avoidance maneuvers involved significantly more ambiguity and was a difficult task, both for experts and for the deviation type classifier. In general, inter-rater reliability for behavior classifications was low. The high label variability illustrates the difficulty of conducting a post-event analysis with only weather and traffic data. It is possible that the low inter-rater reliability would have been partially alleviated by providing the experts with more weather data for each encounter. Even so, assessing the encounter behavior is a complex endeavor which requires a nuanced understanding of the factors involved, and correct classification of some encounters will remain uncertain. Interestingly, the degree of expert uncertainty in the encounter classification was reflected in the deviation type classifier performance. Cases with a lower inter-rater agreement rate tended to also be more difficult for the deviation type classifier, with a degraded performance on these cases as compared to cases with a high inter-rater agreement rate.

8.3 Major Contributions

The overall objective - using large air traffic and weather datasets to achieve better prediction of en-route tactical weather deviation - was accomplished. Furthermore, the approach outlined in this work is flexible and can be readily expanded upon. The deviation prediction model developed for this work is relatively simple, considering only two kinds of weather observations. However, the deviation prediction model is a basis on which more complex models can be built. For instance, weather forecast data could be easily included as an input into the model with a straightforward change to the model architecture.

Additionally, the development of the deviation detection model enables further research in aviation weather. The classifier can be directly used on encounter images to generate behavioral labels, facilitating the creation of large datasets for supervised machine learning techniques. Indirectly, the methods employed to build the deviation detection model may be of use in future work. The deviation detection model assumes a number of properties for input images and can only provide labels according to the classes used in this work. The constraints on model inputs and outputs may not be appropriate for some aviation weather research problems, but the existence of a viable model construction process opens up new possibilities for solving these problems.

This work also demonstrated the utility of framing aircraft-weather behavior classification as an image recognition task. While recent work has found considerable success with robust and scalable techniques such as clustering for air traffic data, incorporating weather data in a meaningful way has proven to be a difficult challenge. Much of the current literature modeling aircraft-weather interaction relies on hand-crafting spatial filters to extract weather features for use in model development. This approach helps to capture important spatial relationships between flight and weather data, which is essential to predicting deviations. It also reduces the dataset dimensionality, potentially reducing the size of the dataset required to learn predictive patterns in the data. Despite its benefits, this approach is not necessarily well-suited to a large-scale machine learning task. It requires that experts estimate which weather

features will be most useful, and the downsampling of the weather data reduces the potential benefits of increasing dataset size.

The image-based approach offers a few key advantages. It inherently encodes the spatial relationships between the aircraft and weather data, and requires fewer *a priori* assumptions about encounter behavior. Furthermore, image classification is a heavily researched field that has seen recent rapid advancement, and this work benefited greatly from the application of mature, state-of-the-art image classification techniques. In particular, the adaptation of models trained on extremely large generic image datasets helped to maximize the usefulness of the much smaller expert-labeled encounter dataset.

Finally, the 1700+ encounters with expert behavior assessments collectively form a valuable resource for advancing aviation weather research. It is costly and time-consuming to collect expert opinions on aircraft-weather interactions. Based on current literature, the expert-labeled dataset used in this work appears to be among the largest of its kind. The dataset has been made available on the Harvard Dataverse repository with the goal of enabling other researchers to continue work in this field [73].

Appendix A

Colormapping

Visualizing the VIL and echo top observations required adoption of a colormapping scheme. The standard CIWS Echo Top colormap (see Figure 3-2) was used for the echo tops data in this work. To approximate the same reflectivity coloring scale used by the National Weather Service, the VIL was colormapped using a scale that approximated the NWS's 5 dBZ color bands. It was desirable to approximate the NWS reflectivity coloring because the expert participants would be most accustomed to weather information presented in that format.

This was done using the VIL/VIP/reflectivity equivalencies found in Table 1 and estimating the rough VIL equivalents for each of the reflectivity bounds. Figure A-1 shows the relationship between VIL and reflectivity, with points colored according to the NWS reflectivity coloring. Values below 20 dBZ were masked to avoid cluttering the radar weather pictures, and because these values are typically not shown in en-route weather radar displays.

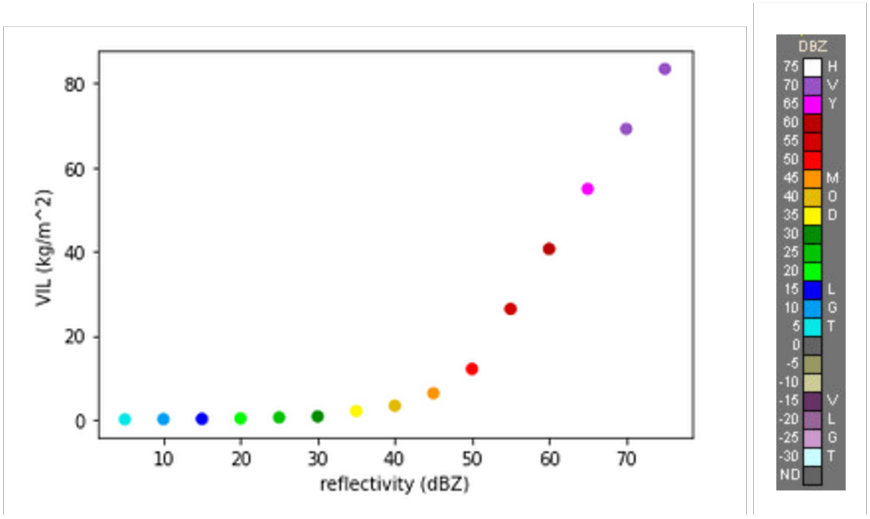


Figure A-1: Interpolation of estimated reflectivity equivalents for VIL values, and the NWS reflectivity colormap on the right.

Appendix B

Informed Consent

The informed consent statement sent to potential participants reads as follows:

Thank you for your interest in our study! More information about the study is listed below. If, after reviewing the following information, you would like to participate, please reply with a statement that you have read and understood the following information.

Participation: We expect that participation will take about 15-30 minutes. It is done entirely remotely and asynchronously (we will not need to schedule any meetings). Participation consists of categorizing images of previous aircraft weather encounters as non-deviations, deviations due to weather, or deviations for other reasons.

Purpose: The end goal of this work is the development of better predictive algorithms for en-route convective weather avoidance. Development of these models relies on labeled historical aircraft weather encounters. This study is being conducted to develop an algorithm for recognizing those behaviors in previous weather encounters.

Methods: Current weather avoidance models are based on small sets of labeled weather encounters. We seek to improve these models by leveraging machine learning models to generate larger labeled datasets. At this time, we are planning to use an artificial neural network to process our labeled data. The expert-labeled encounters collected during this study will be used to train models to label significantly larger sets of encounters. This larger set of labeled encounters will be used to build more robust weather avoidance models.

Benefits: Successful development of a predictive convective weather avoidance model may result in improved decision support tools for air traffic controllers and other personnel involved with the routing of aircraft. These tools may help to reduce flight plan deviations due to convective weather as well as delays associated with enroute convective weather.

Privacy: The personal information that is being collected are the email addresses, the types of aircraft certified for, and the number of flying hours for each labeler. Each participant will be assigned a randomized alphanumeric code to maintain anonymity during data analysis. These assignments are kept on a separate hard drive from the data analysis. The labels themselves will be used only in aggregate. We will not release or publish individual data for email addresses, type certificates, or flying hours, or any other personal information that is provided to us.

Participation in this study in voluntary, and there are no adverse consequences for choosing not to participate. If you choose to participate, you may decline to answer any or all questions. You may withdraw from participation at any time, with no adverse consequences.

If you have any other questions about this study, please reply to this email or contact wx_avoidance@MIT.edu.

Appendix C

Labeling Instructions

The following documents are the instructions sent to the experts who participated in assessing the encounters.

MIT Study of Pilot Weather Deviations

As part of an MIT study on the impact of convective weather on flight rerouting we are studying historical cases where flights may have deviated from their original flight plans. We are training machine learning algorithms to recognize convective weather conditions which would cause pilots to deviate around the weather.

In order to develop a training set of cases we are asking experienced pilots to evaluate flight segments which may (or may not) have deviated due to the tactical weather situation.

You will be asked to evaluate a series of enroute flight segments, each about 200nmi long, where you will be presented with:

1 – The filed flight plan route (in black);

2 - The route flown (in blue, with magenta dots every 30 nmi and the start point identified);

3 – A reconstruction of the weather radar image at the time of the start of the segment;

4 – The vertical profile of the flight.

You will be asked to determine if the there was a deviation or not. If there was a deviation you will be asked to assess if it was due to the tactical weather situation or for some other reason such as ATC instructions or direct routing. Some cases may be very difficult to assess, in which case an “ambiguous” label can be applied.

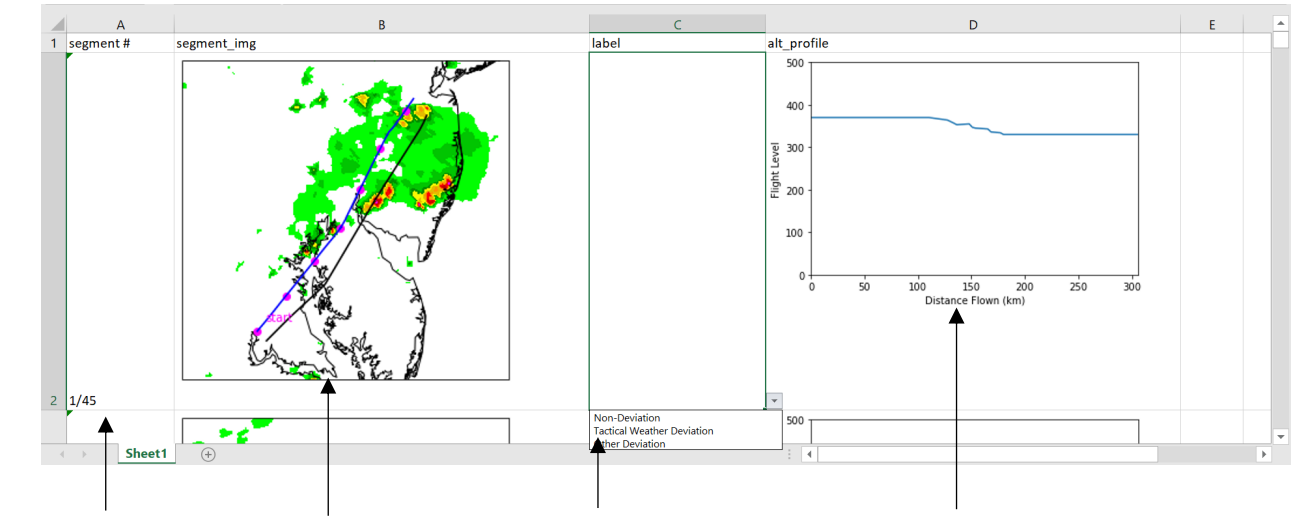
All flights were Part 121 operations in air transport jet aircraft.

A set of about 50 flight segments will be presented to you in an Excel spreadsheet. A detailed example of this spreadsheet can be found on the following page.

Please label each flight as (Non-Deviation, Tactical Weather Deviation, Other Deviation or Ambiguous) in the pull-down menu in Column C.

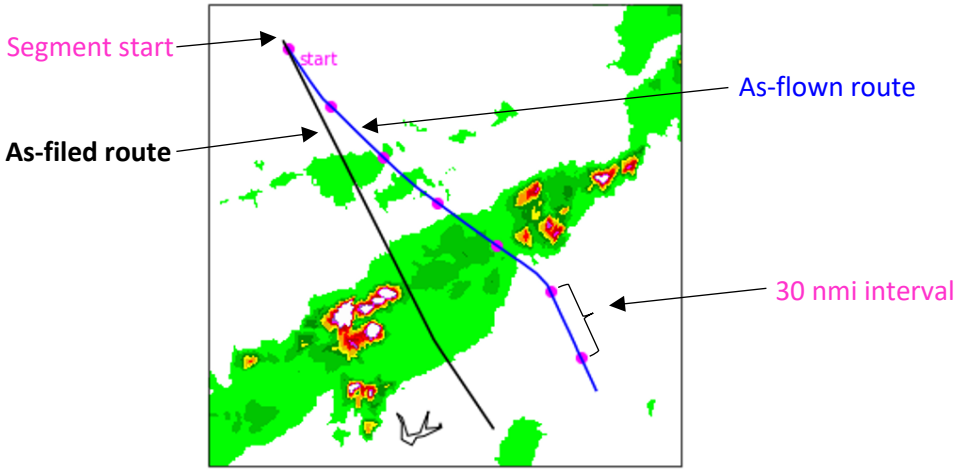
After you complete your evaluation please save the spreadsheet without changing the title and send it back to wx_avoidance@mit.edu.

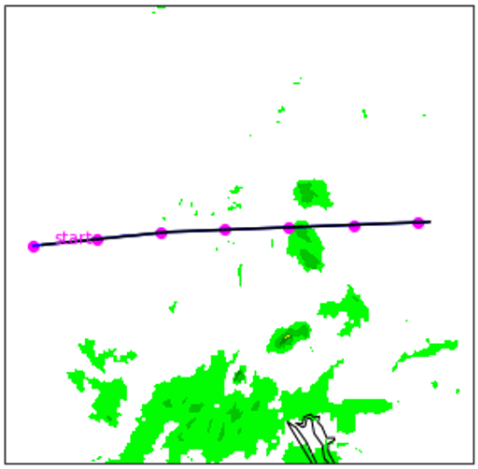
Thanks for your help on this study.

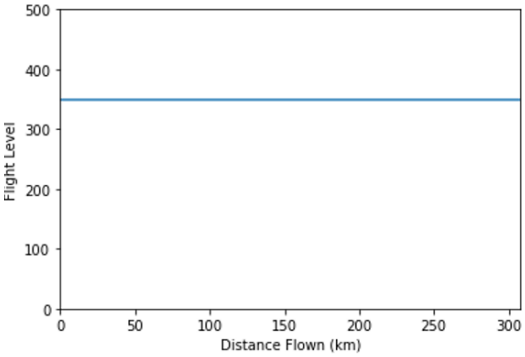


Each segment can be labeled as one of three categories: a non-deviation, a deviation for tactical weather, or a deviation for other reasons. In cases where none of these categories are appropriate, an "ambiguous" label can instead be applied. These labels can be selected from the drop-down menu in the column in between the segment image and the altitude profile. Some examples of each of the three categories can be found on the following page (no example is given for the ambiguous designation).

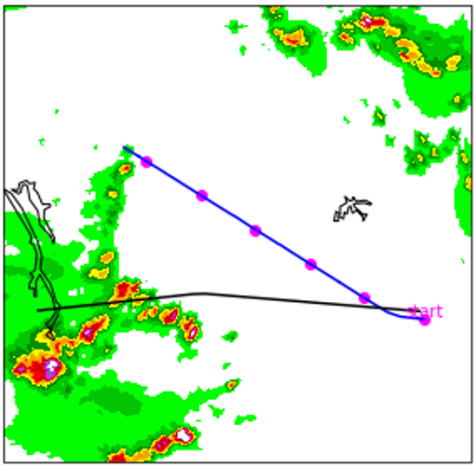
In each segment image, the as-flown route is shown in blue, and the **as-filed** (planned) route is shown in **black**. The start of the segment is indicated by the magenta 'start', and the magenta points along the flown route are spaced 30 nmi apart.

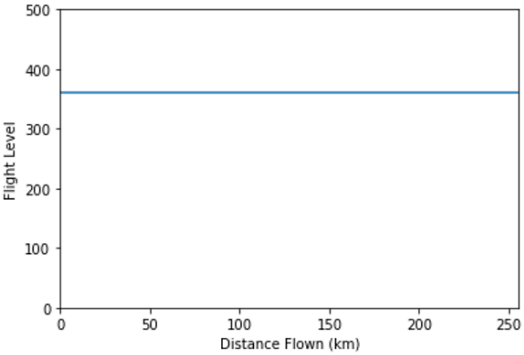




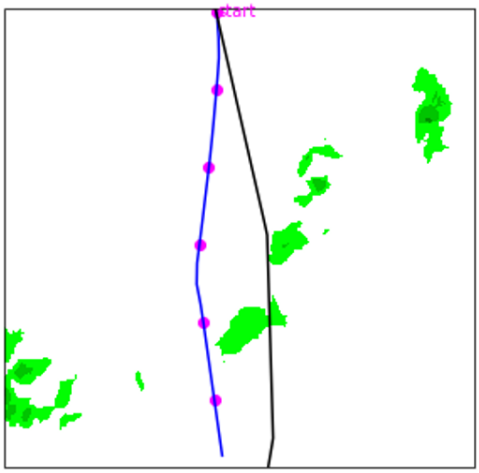


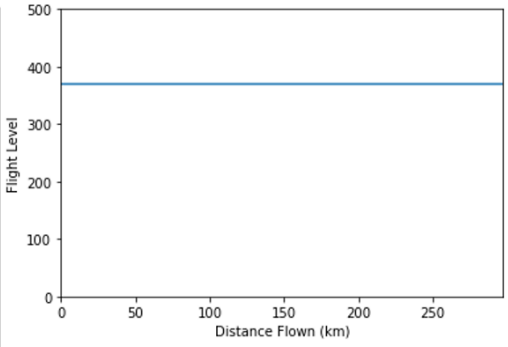
This segment image does not show a deviation, since the flown route tracks the filed route very closely. Its label would be Non-Deviation.





This is a case in which the aircraft deviated in order to avoid the convective weather shown on the filed route. Its label would be Tactical Weather Deviation.





This segment image shows a deviation from the filed route. This deviation may or may not be an attempt to avoid the weather conditions shown in the encounter. If the deviation is most likely due to factors other than the weather conditions shown in the image, its label would be Other Deviation.

Bibliography

[1] "FAQ: Weather Delay," Federal Aviation Administration. (), [Online]. Available: https://www.faa.gov/nextgen/programs/weather/faq.

[2] "Weather-Related Aviation Accident Study 2003-2007," Federal Aviation Administration, Feb. 2, 2010. [Online]. Available: https://www.asias.faa.gov/i/studies/2003-2007weatherrelatedaviationaccidentstudy.pdf.

[3] "FAA Joint Order 7210.3, Chapter 18, Section 16. Severe Weather Avoidance Plans (SWAP)," Federal Aviation Administration. (Apr. 20, 2023).

[4] K. Sheth, A. Clymer, A. Morando, and F.-T. Shih, "Analysis of multiple flight common route for traffic flow management," in *16th AIAA Aviation Technology, Integration, and Operations Conference*, Washington, D.C.: American Institute of Aeronautics and Astronautics, Jun. 13, 2016, ISBN: 978-1-62410-440-4. DOI: 10.2514/6.2016-4207. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.2016-4207 (visited on 07/06/2023).

[5] A. Kenitzer, D. Gebru, and J. Leffler, *Convective Weather and How It Could Affect Your Flight*. [Online]. Available: https://www.faa.gov/podcasts/the-air-up-there/convective-weather-and-how-it-could-affect-your-flight.

[6] *Pilot's Handbook of Aeronautical Knowledge*. Federal Aviation Administration, 2023.

[7] J. Luppens Mahoney, S. Seseske, J. Hart, M. Kay, and B. Brown, "Defining Observation Fields for Verification of Spatial Forecasts of Convection: Part 2,"

American Meteorological Society, 2004. [Online]. Available: https://esrl.noaa.gov/fiqas/publications/articles/ARAM04-mahoney-final.pdf.

[8] J. E. Evans and E. R. Ducot, "Corridor Integrated Weather System," *Lincoln Laboratory Journal*, vol. 16, no. 1, pp. 59–80, 2006.

[9] M. Robinson, R. DeLaura, and N. Underhill, "The Route Availability Planning Tool (RAPT): Evaluation of Departure Management Decision Support in New York during the 2008 Convective Weather Season," presented at the Eighth USA/Europe Air Traffic Management Research and Development Seminar, Napa, CA, USA, 2009.

[10] R. DeLaura and J. Evans, "An Exploratory Study of Modeling Enroute Pilot Convective Storm Flight Deviation Behavior," presented at the 12th Conference on Aviation Range and Aerospace Meteorology, Atlanta, GA, USA: American Meteorological Society, 2006, p. 15. [Online]. Available: https://ams.confex.com/ams/pdfpapers/103755.pdf.

[11] R. DeLaura, B. Crowe, R. Ferris, J. F. Love, and W. N. Chan, "Comparing Convective Weather Avoidance Models and Aicraft-Based Data," p. 17, 2009.

[12] R. DeLaura, M. Robinson, M. Pawlak, and J. Evans, "Modeling Convective Weather Avoidance in Enroute Airspace," presented at the 88th Annual Meeting of the American Meterological Society, New Orleans, LA, USA: American Meteorological Society, 2008, p. 12. [Online]. Available: https://ams.confex.com/ams/pdfpapers/132903.pdf.

[13] M. Matthews and R. DeLaura, "Assessment and Interpretation of En Route Weather Avoidance Fields from the Convective Weather Avoidance Model," in *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, Fort Worth, TX, USA: American Institute of Aeronautics and Astronautics, Sep. 13, 2010, ISBN: 978-1-62410-159-5. DOI: 10.2514/6.2010-9160. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2010-9160 (visited on 06/24/2020).

[14] R. DeLaura, M. Robinson, R. Todd, and K. MacKenzie, "Evaluation of Weather Impact Models in Departure Managament Decision Support: Operational Performance of the Route Availability Planning Tool (RAPT) Prototype," presented at the 88th Annual Meeting of the American Meteorological Society, New Orleans, LA, USA: American Meteorological Society, Jan. 2008.

[15] M. Rubnich and R. DeLaura, "An Algorithm to Identify Robust Convective Weather Avoidance Polygons in En Route Airspace," in *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, Fort Worth, TX, USA: American Institute of Aeronautics and Astronautics, Sep. 13, 2010, ISBN: 978-1-62410-159-5. DOI: 10.2514/6.2010-9164. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2010-9164 (visited on 06/23/2020).

[16] D. McNally, K. Sheth, C. Gong, *et al.*, "Dynamic Weather Routes: A Weather Avoidance System for Near-Term Trajectory-Based Operations," in *28th Congress of the International Council of the Aeronautical Sciences*, Brisbane, Australia: International Council of the Aeronautical Sciences, Sep. 2012, p. 18. [Online]. Available: http://www.icas.org/ICAS_ARCHIVE/ICAS2012/PAPERS/366.PDF.

[17] J. E. Evans, K. Carusone, B. Crowe, D. Meyer, M. M. Wolfson, and D. Klingle-Wilson, "The Corridor Integrated Weather System (CIWS)," in *Proceedings of the 10th Conference on Aviation Range and Aerospace Meteorology*, Portland, OR, USA: American Meteorological Society, 2002, p. 6.

[18] W. Dupree, D. Morse, M. Chan, *et al.*, "The 2008 CoSPA Forecast Demonstration," in *Aviation, Range and Aerospace Meteorology Special Symposium on Weather - Air Traffic*, Phoenix, AZ, USA: American Meteorological Society, Jan. 2009, p. 19.

[19] S. S. Allan, S. Gaddy, and J. Evans, "Delay Causality and Reduction at the New York city Airports using Terminal Weather Information Systems," Lincoln Laboratory, Lexington, MA, USA, Project Report ATC-291, Feb. 16, 2001, p. 59.

[20] R. DeLaura and S. Allan, "Route Selection Decision Support in Convective Weather: A Case Study of the Effects of Weather and Operational Assumptions on Departure Throughput," presented at the 5th Eurocon/FAA ATM R&D Seminar, Budapest, Hungary, Jun. 2003, p. 10.

[21] W. Chan, M. Refai, and R. DeLaura, "An Approach to Verify a Model for Translating Convective Weather Information to Air Traffic Management Impact," in *7th AIAA ATIO Conf, 2nd CEIAT Int'l Conf on Innov and Integr in Aero Sciences,17th LTA Systems Tech Conf; Followed by 2nd TEOS Forum*, Belfast, Northern Ireland: American Institute of Aeronautics and Astronautics, Sep. 18, 2007, ISBN: 978-1-62410-014-7. DOI: 10.2514/6.2007-7761. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2007-7761 (visited on 04/28/2022).

[22] Y.-H. Lin and H. Balakrishnan, "Prediction of Terminal-Area Weather Penetration on the Basis of Operational Factors," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2400, no. 1, pp. 45–53, Jan. 2014, ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2400-06. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2400-06 (visited on 04/29/2022).

[23] M. Matthews and R. DeLaura, "Assessment and Interpretation of En Route Weather Avoidance Fields from the Convective Weather Avoidance Model," in *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, Fort Worth, Texas: American Institute of Aeronautics and Astronautics, Sep. 2010, ISBN: 978-1-62410-159-5. DOI: 10.2514/6.2010-9160. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2010-9160 (visited on 06/24/2020).

[24] S. Campbell, M. Matthews, and R. Delaura, "Evaluation of the Convective Weather Avoidance Model for Arrival Traffic," in *12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Indianapolis, Indiana:

American Institute of Aeronautics and Astronautics, Sep. 17, 2012, ISBN: 978-1-60086-930-3. DOI: 10.2514/6.2012-5500. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.2012-5500 (visited on 04/29/2022).

[25] K. Sheth, T. Amis, S. Gutierrez-Nolasco, B. Sridhar, and D. Mulfinger, "Development of a Probabilistic Convective Weather Forecast Threshold Parameter for Flight-Routing Decisions," *Weather and Forecasting*, vol. 28, no. 5, pp. 1175–1187, Oct. 1, 2013, ISSN: 0882-8156, 1520-0434. DOI: 10.1175/WAF-D-12-00052.1. [Online]. Available: https://journals.ametsoc.org/doi/10.1175/WAF-D-12-00052.1 (visited on 04/29/2022).

[26] K. Kuhn, "Analysis of Thunderstorm Effects on Aggregated Aircraft Trajectories," *Journal of Aerospace Computing, Information, and Communication*, vol. 5, no. 4, pp. 108–119, Apr. 2008, ISSN: 1542-9423. DOI: 10.2514/1.34830. [Online]. Available: https://arc.aiaa.org/doi/10.2514/1.34830 (visited on 04/23/2022).

[27] D. Michalek and H. Balakrishnan, "Identification of Robust Routes using Convective Weather Forecasts," p. 11, 2009.

[28] D. M. Pfeil and H. Balakrishnan, "Identification of Robust Terminal-Area Routes in Convective Weather," *Transportation Science*, vol. 46, no. 1, pp. 56–73, Feb. 2012, ISSN: 0041-1655, 1526-5447. DOI: 10.1287/trsc.1110.0372. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/trsc.1110.0372 (visited on 04/29/2022).

[29] H. Arneson, "Initial Analysis of and Predictive Model Development for Weather Reroute Advisory Use," in *15th AIAA Aviation Technology, Integration, and Operations Conference*, Dallas, TX: American Institute of Aeronautics and Astronautics, Jun. 22, 2015, ISBN: 978-1-62410-369-8. DOI: 10.2514/6.2015-3395. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.2015-3395 (visited on 07/05/2023).

[30] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: A survey," *Optical Engineering*, vol. 58, no. 4, pp. 040 901–040 901, 2019.

[31] "What is a convolutional neural network?" MathWorks. (), [Online]. Available: https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html?s_tid=srchtitle_site_search_1_convolutional%20neural%20network.

[32] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. arXiv: 1409.0575. [Online]. Available: http://arxiv.org/abs/1409.0575.

[33] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[34] F. A. Administration, *Traffic Flow Management System (TFMS)*. [Online]. Available: https://aspm.faa.gov/aspmhelp/index/Traffic_Flow_Management_System_(TFMS).html.

[35] J. A. V. N. T. S. Center, *SWIM flight data publication service (SFDPS) data consumer reference manual*.

[36] A. D. Snow, J. Whitaker, M. Cochran, *et al.*, *Pyproj4/pyproj: 3.6.0 release*, Jun. 2023. DOI: 10.5281/ZENODO.2592232. [Online]. Available: https://zenodo.org/record/2592232.

[37] Met Office, *Cartopy: A cartographic python library with a matplotlib interface*, Exeter, Devon, 2010 - 2015. [Online]. Available: https://scitools.org.uk/cartopy.

[38] D. Klingle-Wilson and J. Evans, "Description of the Corridor Integrated Weather System (CIWS) Weather Products," Lincoln Laboratory, Lexington, MA, USA, Project Report ATC-317, Aug. 1, 2005, p. 120.

[39] M. Robinson, J. E. Evans, and B. A. Crowe, "En Route Weather Depiction Benefits of the NEXRAD Vertically Integrated Liquid Water Product Utilized by the Corridor Integrated Weather System," presented at the 10th Conference on Aviation, Range, and Aerospace Meteorology, Portland, OR, USA: American Meteorological Society, 2002, p. 4. [Online]. Available: https://archive.ll.mit.edu/mission/aviation/publications/publication-files/ms-papers/Robinson_2002_ARAM_MS-15345_WW-10470.pdf.

[40] D. R. Greene and R. A. Clark, "Vertically Integrated Liquid Water—A New Analysis Tool," *Monthly Weather Review*, vol. 100, no. 7, pp. 548–552, Jul. 1972, ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1972)100<0548:VILWNA>2.3.CO;2. [Online]. Available: http://journals.ametsoc.org/doi/10.1175/1520-0493(1972)100%3C0548:VILWNA%3E2.3.CO;2 (visited on 12/28/2022).

[41] *Air traffic by the numbers*, Apr. 2023. [Online]. Available: https://www.faa.gov/air_traffic/by_the_numbers/media/Air_Traffic_by_the_Numbers_2023.pdf.

[42] N. Underhill, R. DeLaura, and M. Robinson, "Severe weather avoidance program performance metrics for new york departure operations," in *14th Conference on Aviation, Range and Aerospace Meteorology*, 2010.

[43] S.-C. Wu, C. Gooding, A. Shelley, C. Duong, and W. Johnson, "Pilot convective weather decision making in en route airspace," in *28th Congress of the International Council of the Aeronautical Sciences*, Brisbane, Australia: International Council of the Aeronautical Sciences, 2012, p. 8.

[44] K. A. Latorella and J. P. Chamberlain, "Tactical *vs.* Strategic Behavior: General Aviation Piloting in Convective Weather Scenarios," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 1, pp. 101–105, Sep. 2002. DOI: 10.1177/154193120204600121. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193120204600121 (visited on 06/23/2020).

[45] D. Marconnet, C. Norden, and L. Vidal, *Optimum Use of Weather Radar*, Jul. 2016.

[46] D. A. Rhoda, E. A. Kocab, and M. L. Pawlak, "AIRCRAFT ENCOUNTERS WITH THUNDERSTORMS IN ENROUTE VS. TERMINAL AIRSPACE ABOVE MEMPHIS, TENNESSEE," in *10th Conference on Aviation, Range, and Aerospace Meterology*, Portland, OR, USA: American Meteorological Society, May 2002, p. 4.

[47] "Radar images: Reflectivity," National Oceanic and Atmospheric Administration. (), [Online]. Available: `https://www.noaa.gov/jetstream/reflectivity`.

[48] S. Troxel and C. D. Engholm, "Vertical reflectivity profiles: Averaged storm structures and applications to fan beam radar weather detection in the us," in *16th Conference on Severe Local Storms, American Meteorological Society, Kananaskis Park, Alta, Canada*, Citeseer, 1990.

[49] M. P. Matthews and R. DeLaura, "Modeling Convective Weather Avoidance of Arrivals in the Terminal Airspace," presented at the 2nd Aviation, Range, and Aerospace Meterology Special Symposium on Weather-Air Traffic Management Integration, American Meteorological Society, Jan. 2011. [Online]. Available: `https://www.ll.mit.edu/r-d/publications/modeling-convective-weather-avoidance-arrivals-terminal-airspace`.

[50] M.-M. Temme and C. Tienes, "Factors for Pilot's Decision Making Process to Avoid Severe Weather during Enroute and Approach," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, London, UK: IEEE, Sep. 2018, pp. 1–10. DOI: `10.1109/DASC.2018.8569357`. [Online]. Available: `https://ieeexplore.ieee.org/abstract/document/8569357`.

[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: `http://www.deeplearningbook.org`.

[52] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[53] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia computer science*, vol. 132, pp. 377–384, 2018.

[54] Y. Yang, L. Zhang, M. Du, *et al.*, "A comparative analysis of eleven neural networks architectures for small datasets of lung images of covid-19 patients toward improved clinical decisions," *Computers in Biology and Medicine*, vol. 139, p. 104 887, 2021.

[55] I. Kontopoulos, A. Makris, and K. Tserpes, "A Deep Learning Streaming Methodology for Trajectory Classification," *ISPRS International Journal of Geo-Information*, vol. 10, no. 4, p. 250, Apr. 8, 2021, ISSN: 2220-9964. DOI: 10.3390/ijgi10040250. [Online]. Available: https://www.mdpi.com/2220-9964/10/4/250 (visited on 07/19/2022).

[56] *6.036 lecture notes*, Dec. 12, 2019.

[57] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[58] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." arXiv: 1512.03385 [cs]. (Dec. 10, 2015), [Online]. Available: http://arxiv.org/abs/1512.03385 (visited on 07/06/2023), preprint.

[59] A. Reuther, J. Kepner, C. Byun, *et al.*, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*, IEEE, 2018, pp. 1–6.

[60] V. N. Nguyen and H. Holone, "Possibilities, Challenges And The State Of The Art Of Automatic Speech Recognition In Air Traffic Control," Aug. 3, 2015. DOI: 10.5281/ZENODO.1108428. [Online]. Available: https://zenodo.org/record/1108428 (visited on 07/06/2023).

[61] S. Badrinath and H. Balakrishnan, "Automatic Speech Recognition for Air Traffic Control Communications," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2676, no. 1, pp. 798–810, Jan. 2022, ISSN:

0361-1981, 2169-4052. DOI: 10.1177/03611981211036359. [Online]. Available: http://journals.sagepub.com/doi/10.1177/03611981211036359 (visited on 07/06/2023).

[62] Phung and Rhee, "A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets," *Applied Sciences*, vol. 9, no. 21, p. 4500, Oct. 23, 2019, ISSN: 2076-3417. DOI: 10.3390/app9214500. [Online]. Available: https://www.mdpi.com/2076-3417/9/21/4500 (visited on 06/19/2023).

[63] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, and D. Camacho, "Ensembles of Convolutional Neural Networks models for pediatric pneumonia diagnosis," *Future Generation Computer Systems*, vol. 122, pp. 220–233, Sep. 2021, ISSN: 0167739X. DOI: 10.1016/j.future.2021.04.007. arXiv: 2010.02007 [eess]. [Online]. Available: http://arxiv.org/abs/2010.02007 (visited on 07/06/2023).

[64] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informatics*, vol. 5, no. 2, p. 2, Dec. 2018, ISSN: 2198-4018, 2198-4026. DOI: 10.1186/s40708-018-0080-3. [Online]. Available: https://braininformatics.springeropen.com/articles/10.1186/s40708-018-0080-3 (visited on 06/19/2023).

[65] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust Learning at Noisy Labeled Medical Images: Applied to Skin Lesion Classification," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy: IEEE, Apr. 2019, pp. 1280–1283, ISBN: 978-1-5386-3641-1. DOI: 10.1109/ISBI.2019.8759203. [Online]. Available: https://ieeexplore.ieee.org/document/8759203/ (visited on 09/22/2022).

[66] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101 759, Oct. 2020, ISSN: 13618415. DOI:

10.1016/j.media.2020.101759. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841520301237 (visited on 09/22/2022).

[67] Y. Jiang, Y. Liu, D. Liu, and H. Song, "Applying Machine Learning to Aviation Big Data for Flight Delay Prediction," in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Calgary, AB, Canada: IEEE, Aug. 2020, pp. 665–672, ISBN: 978-1-72816-609-4. DOI: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114. [Online]. Available: https://ieeexplore.ieee.org/document/9251206/ (visited on 04/29/2022).

[68] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Swedent, 2018.

[69] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit, "Classification of breast lesions using cross-modal deep learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, Australia: IEEE, Apr. 2017, pp. 109–112, ISBN: 978-1-5090-1172-8. DOI: 10.1109/ISBI.2017.7950480. [Online]. Available: http://ieeexplore.ieee.org/document/7950480/ (visited on 07/19/2022).

[70] M. S. Veillette, S. Samsi, and C. J. Mattioli, "SEVIR : A Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology," presented at the 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020.

[71] "Trajectory based operations and verification and validation," Federal Aviation Administration. (), [Online]. Available: https://www.faa.gov/sites/faa.gov/files/2022-02/11TrajectoryBasedOperations-PamelaWhitley.pdf.

[72] "Trajectory-based operations (TBO)," Federal Aviation Administration. (), [Online]. Available: https://www.faa.gov/air_traffic/technology/tbo.

[73] R. Price, *Convective Weather Encounters with Behavior Assessments*, 2023. DOI: 10.7910/DVN/YDUAL4. [Online]. Available: https://doi.org/10.7910/DVN/YDUAL4.