DEEP LEARNING FOR THUNDERSTORM NOWCASTING USING SATELLITE AND RADAR OBSERVATIONS

by

Stephanie M. Ortland

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Atmospheric and Oceanic Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 10/07/2024

The dissertation is approved by the following members of the Final Oral Committee:
Michael J. Pavolonis, Physical Scientist, NOAA/NESDIS
Grant W. Petty, Professor, Atmospheric and Oceanic Sciences
Angela K. Rowe, Assistant Professor, Atmospheric and Oceanic Sciences
Tristan S. L'Ecuyer, Professor, Atmospheric and Oceanic Sciences
Ángel F. Adames-Corraliza, Assistant Professor, Atmospheric and Oceanic Sciences
Dimitris Papailiopoulos, Associate Professor, Electrical and Computer Engineering

To Alexander Ortland
For his encouragement and unending supply of hugs.

ACKNOWLEDGMENTS

The common phrase "standing on the shoulders of giants" is interwoven throughout historical and contemporary texts for good reason. Each generation is able to pursue a deeper understanding of the Earth and its surroundings by building on research and ideas from previous generations. This dissertation would not be possible without the scientific advancements made by generations of scientists, known and unknown, since humankind's humble beginnings. I am especially grateful to all the women in science who have made it possible for me, a woman, to pursue a career in physics and atmospheric science with minimal hinderances associated with historical gender discrimination in the natural sciences.

Along these same lines of thinking, there are many people I would like to thank personally for their support during my pursuit of a Ph.D. in Atmospheric and Oceanic Sciences at the University of Wisconsin-Madison. First and foremost, I thank my research advisor, Mike Pavolonis, for his mentorship. I find it difficult to put my gratitude for Mike's support into words. Besides listening to my ideas, both rational and outlandish, Mike has taught me to strive for excellence while not letting the pursuit of perfection hinder advancement in doing so. Without a doubt, I would not be the research scientist I am today without him. Second, I thank my academic advisor and career confidant, Grant Petty, for his willingness to discuss obscure science questions and my goals for my career post-graduation. Grant is the kind of person you meet once in a lifetime and cherish endlessly; I am forever grateful to know him and be a beneficiary of his tutelage. For their support throughout the doctoral process, I also thank the members of my Ph.D. committee I have yet to mention: Angela Rowe, Tristan L'Ecuyer, Ángel Adames-Corraliza, and Dimitris Papailiopoulos. Their feedback has positively influenced my research and encouraged me to keep meteorological science questions at the heart of my work rather than technical details.

Next, I would like to thank my colleagues, fellow graduate students, and friends in the broader Atmospheric, Oceanic, and Space Science community at the University of Wisconsin-Madison. My esteemed colleague John Cintineo (also a collabo-

rator on my first publication), deserves my thanks for letting me adapt some of the code he developed for a previous project to my Ph.D. work and for his willingness to answer coding questions on short notice. Also, I have been lucky to share offices with some brilliant minds and compassionate listeners throughout the pursuit of my doctorate. Cameron Bertossa, Hamish Prince, and Brianne Andersen are three such people, and I am grateful for their support. Likewise, thank you to Lena Heuscher-Stewart, Anthony Bernal Ayala, Leila Gabrys, and Alicia Hoffman whose friendships bring me joy.

Outside of science, I stand on the shoulders of family members who have encouraged me to pursue my dreams and have provided me the financial and emotional support necessary to do so. My academic pursuits would not be possible without my wonderful parents, Todd and Sheri Bradshaw. Throughout my life they have celebrated my successes and boosted my spirits when I have failed along with my sisters, Megan and Rachel. It is a privilege and a blessing to be a part of this family. My husband, Alex Ortland, is well deserving of the dedication of this dissertation because he has always believed in my abilities and supported my career decisions even when those decisions have required us to relocate. Also, for instilling in me an appreciation of education and spoiling me with their affections, I thank my grandparents: Dale Bradshaw, Sharon Bradshaw, Sylvester (Bud) Louis II, and Nancy Louis. Although three of them were unable to see this achievement, their influence remains one of my greatest blessings.

Last but not least, I would like to acknowledge this dissertation would not be possible without the funding provided by the National Oceanic and Atmospheric Administration (NOAA) grant NA20NES4320003. With this funding, portions of this dissertation have been published in Ortland et al. (2023) ©American Meteorological Society. Additional publications with materials from this dissertation are forthcoming, pending reviews by peer-reviewed journals (Ortland and Pavolonis (2024) ©American Meteorological Society). The materials in the publication(s) are used with permission throughout this dissertation. The scientific results and conclusions, as well as any views or opinions expressed herein, are my own and do not necessarily reflect those of NOAA or the Department of Commerce.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The current scientific understanding of convection from a geostationary satellite and ground-based radar perspective is used to inform the development and evaluation of a deep learning model (called ThunderCast) for convective storm (thunderstorm) nowcasting in the continental United States. The model is trained with four satellite channels in the visible, shortwave infrared, and longwave infrared bands from the Geostationary Environmental Operational Satellite-16 (GOES-16) Advanced Baseline Imager (ABI) as predictors. The target is Multi-Radar Multi-Sensor (MRMS) radar reflectivity at the -10°C isotherm in the atmosphere, where any reflectivity values $\geqslant$ 30 dBZ are positive for thunderstorm activity. ThunderCast has high accuracy, recall, and specificity but has low precision, indicating the model is prone to false alarms. To determine the extent of true and false positive predictions with and without lightning activity in ThunderCast, an object tracking software is used to identify and track predicted storms. Over half of the true positive tracks were not associated with lightning and a radar reflectivity threshold cleanly separating tracks with and without lightning was not found, demonstrating a limitation of a purely radar-based definition of convection in thunderstorm nowcasting models. ABI observations for the tracks indicate electrified and non-electrified convective storms can appear similar in satellite imagery. Four case studies, each containing at least one instance of a non-electrified storm with red-green-blue (RGB) false color image signatures consistent with cloud-top glaciation, demonstrate non-electrified convective storms with similar visual appearances as thunderstorms are not limited by latitude and can form in an assortment of meteorological environments.

1 INTRODUCTION

Artificial intelligence is a broad term encompassing computer programs designed to automate intellectual tasks usually performed by humans. Machine learning refers to particular cases of artificial intelligence where, when presented with many examples of inputs and the corresponding target (the “truth" or the desired outcome known for each example) for a given task, a machine learns the rules to map input parameters to the desired output(s) (Chollet, 2018; Stevens et al., 2020). The appropriate structure of a machine learning model varies depending on the task. Models requiring many successive layers, where non-linear data transformations (associated with patterns in the input dataset) occur, are called deep learning models. The number of layers are referred to as the depth of the model, hence many layers categorize the model as deep (Chollet, 2018).

Deep learning has rapidly gained popularity in the atmospheric and oceanic sciences because of its ability to automatically identify patterns, often associated with physical phenomena by forecasters, in datasets often viewed visually (e.g., imagery comprised of bands from space-borne satellites). Some of the recent applications include, but are not limited to, tropical cyclone intensity estimates (Wimmers et al., 2019; Griffin et al., 2022), synoptic-scale front prediction (Lagerquist et al., 2019), short term tornado detection (Lagerquist et al., 2020), satellite-driven convective intensity (Cintineo et al., 2020), nowcasting radar echoes (Cuomo and Chandrasekar, 2021; Ravuri et al., 2021), radar estimations for numerical weather prediction models (Hilburn, 2023), and lightning prediction (Zhou et al., 2020, 2022; Cintineo et al., 2022). Many deep learning applications in atmospheric science operate in the 0-60 min “nowcasting" timeframe and are designed for short-term forecasting of weather phenomena. Nowcasting weather models, and their subsequent improvement through innovations in artificial intelligence techniques, are critical for protecting life and property because they support forecasters’ decision making processes by synthesizing vast quantities of environmental data into actionable insights. This statement is reinforced by the 115th United States Congress’ passing of the Weather Research and Forecasting Act (the Weather Act) in 2017 and its

reauthorization in 2019. The Weather Act prioritized the National Oceanic and Atmospheric Administration's (NOAA) and subsequent organizations' improvement of weather data, modeling, computing, forecasting, and issuance of weather advisories, watches, and warnings.

One forecasting area where deep learning is hypothesized to improve the lead time forecasters are able to obtain prior to the onset of hazards is thunderstorm nowcasting. Thunderstorms have many hazards associated with them including hail, strong winds, lightning, and flooding from heavy precipitation. Any additional time forecasters can provide to the public to prepare for these hazards is critical for protecting life and property. Geostationary satellite imagery is commonly used by forecasters to identify the beginning stages of thunderstorm growth. Because deep learning is well suited for computer vision tasks, and, since satellite data is often interpreted visually (e.g. through red-green-blue images), it is hypothesized that deep learning methods can be used to provide additional lead time to thunderstorm hazards like lightning.

Some models already use deep learning (in the form of a U-Net convolutional neural network) to predict hazards from thunderstorms with satellite imagery (as inputs) and lightning observations (as the target), like the LightningCast model presented in Cintineo et al. (2022). However, ground-based radar signatures associated with convective initiation of thunderstorms (see Chapter 2 for more details) occur earlier in a thunderstorm's life cycle than lightning. Targeting radar-based signatures of convective initiation instead of lightning events in a deep learning model could provide more lead time to hazards (like lightning) than those targeting observed lightning events. Hence, to provide additional lead time to hazards, this dissertation develops a model using a similar deep learning architecture to Cintineo et al. (2022) (a U-Net), geostationary satellite imagery, and ground-based radar observations to predict the occurrence of midlatitude convection associated with thunderstorms in the next hour.

In addition to supporting forecasting efforts, artificial intelligence models can provide valuable scientific insight. Deep learning models are often referred to as "black boxes" because the computational rules learned to map the inputs to

the desired output(s) are not straightforward. Methods for understanding these computational rules are an active area of research and remain exceptionally difficult to digest. However, each model is only as good as the data supplied for the model inputs and target (a process described further in Chapter 2). If the data used are based on the scientific understanding of the phenomena the model is making predictions for, then the performance of the model can indicate areas where the current scientific understanding is either adequate or inadequate. Based on this, the scientific understanding can be improved upon and model improvements can be made.

This dissertation focuses on using deep learning and the existing scientific understanding of convection from geostationary satellite and ground-based radar observations to develop and evaluate a thunderstorm nowcasting model (ThunderCast) for the contiguous United States. In doing so, the following research questions are explored:

1. How well can the occurrence of convection in the next hour be predicted with a deep learning model developed from the current scientific understanding of thunderstorm signatures in satellite imagery and ground-based radar observations?

2. What environmental conditions and/or observational factors (associated with the selection of ThunderCast's inputs and target) impact the reliability and applicability of ThunderCast for thunderstorm prediction?

3. What scientific insights can be gained, what limitations are exposed, and what potential avenues for model improvement can be identified from evaluating ThunderCast case studies?

Each of these questions are further motivated and discussed individually in the subsequent Chapters in this dissertation. Chapter 2 focuses on developing the deep learning model (called ThunderCast) for detecting convection associated with thunderstorms in the contiguous United States and exploring the initial capabilities of the model. A method to evaluate cases of predicted thunderstorm events using

object tracking is presented in Chapter 3. Chapter 3 also uses statistics and characteristics associated with the tracked predicted storms to highlight the limitations of using satellite imagery and radar for prediction of initiating thunderstorms. Lastly, Chapter 4 takes a closer look at some of the limiting cases found in Chapter 3. All together, this dissertation demonstrates how a deep learning model can provide valuable insight into the physical processes associated with thunderstorm development by determining the limitations of a model built from the current scientific understanding of thunderstorm characteristics from a ground-based radar and geostationary satellite perspective.

2 THE DEVELOPMENT AND INITIAL CAPABILITIES OF THUNDERCAST

2.1 Background

Field studies in the twentieth century documented the relationship between convection and precipitation radar echoes, forming the basis for radar-based definitions of convective storms (synonymously referred to as thunderstorms in this paper). An extensive field campaign called the National Hail Research Experiment aimed to increase the understanding of the microphysics and cloud dynamics of severe convective storms by analyzing thunderstorms located in northeastern Colorado, southeastern Wyoming, and/or southwestern Nebraska (Morgan and Squires, 1982; Fankhauser and Wade, 1982). As a part of this campaign, Dye and Martner (1982) and Dye et al. (1982) recorded the environmental conditions, reflectivity structure from ground-based radar, and microphysical characteristics of a thunderstorm on 25 July 1976 with moderate to heavy precipitation and hail. The Dye and Martner (1982) case study has acted as the source of “legacy-based" definitions of convective initiation from a ground-based radar perspective.

According to the non-inductive charging mechanism for thunderstorm electrification, the occurrence of lightning in a thunderstorm depends on collisions between of ice and graupel particles in developing cumulus clouds (Saunders, 1993). During passes through the developing thunderstorm stage of the Dye and Martner (1982) thunderstorm the University of Wyoming Queen Air, N10UW, and a sailplane observed the areas exhibiting the first radar echoes contained ice and graupel with concentrations consistent with radar echoes between 30 dBZ and 40 dBZ between 0°C and −5°C. Dye and Martner (1982) also asserted −10°C or colder cloud temperatures are necessary for adequate concentrations of nuclei for significant precipitation in thunderstorms. Thus, in the past few decades, radar signatures (thresholds) of convective initiation associated with thunderstorms and radar reflectivity measurements used for thunderstorm tracking tend to range

between 30 dBZ and 40 dBZ at various temperature levels including $-10°C$.

For example, during a field campaign designed to determine the importance of radar-observed boundary layer convergence lines in initiating convective storms over the Colorado plains, Wilson and Schreiber (1986) defined a storm as a reflectivity echo of $\geqslant 30$ dBZ. This same reflectivity was used as a starting threshold for the Storm Cell Identification and Tracking (SCIT) algorithm (Johnson et al., 1998). Alternatively, Roberts and Rutledge (2003) used a radar reflectivity threshold of 35 dBZ to distinguish between weakly precipitating storms (< 35 dBZ) and vigorous convective storms (> 35 dBZ). The 35 dBZ threshold was also used in the verification process for the Auto-Nowcast System detailed in Mueller et al. (2003) and studies citing a "legacy-based" radar definition of convective storms or initiation of convective storms (Mecikalski and Bedka, 2006; Mecikalski et al., 2010a,b, 2015; Walker et al., 2012). Although 40 dBZ, the highest end of the reflectivity range observed in Dye and Martner (1982), was used as a proxy for thunderstorm initiation in Sieglaff et al. (2011), it has not been used often in radar-based definitions of convective initiation of thunderstorms, because it has been shown to correlate with lightning instead of convective initiation (Zipser and Lutz, 1994; Gremillion and Orville, 1999; Elsenheimer and Gravelle, 2019). To obtain greater lead time to thunderstorm hazards, including lightning, ground-based radar signatures are targeted in this research study rather than lightning (see Chapter 1).

Based on the above discussion, a threshold of either 30 dBZ or 35 dBZ would be appropriate to reference as a radar signature of convective initiation of a thunderstorm. To maximize the amount of lead time obtained with the Thunderstorm Nowcasting Tool (ThunderCast), the research presented herein defines convective initiation leading to thunderstorm formation as the first occurrence of a radar reflectivity echo of $\geqslant 30$ dBZ at the $-10°C$ isotherm level in the atmosphere. This definition is consistent with the 25 July 1976 case study in the National Hail Research Experiment (the source of the "legacy" definition used in modeling convective initiation in the last few decades) and will be used in the methodology presented in Section 2.2.

As evident from the radar-based definition of thunderstorm initiation, it is clear

radar is a powerful tool for observing and diagnosing convection associated with thunderstorms. However, the network of radar instruments in the contiguous United States (CONUS), although extensive, does not provide full coverage of the land area. This is especially prevalent in less populated regions in the western United States, where beam blocking from mountainous terrain is commonplace, and the nearest radars may be out of range. Additionally, ground-based precipitation radars (S-band) do not provide information on storm development, prior to the first echoes detected from precipitation. Satellites can supplement this gap in observations and provide greater forecast lead time. The Geostationary Operational Environmental Satellite (GOES) program provides satellite data from the Advanced Baseline Imager (ABI), summarized in Schmit et al. (2017), with high temporal and spatial resolution for the entire CONUS, including areas without ground-based radar coverage.

ABI spectral bands with 0.64 μm (channel 2), 1.6 μm (channel 5), 10.3 μm (channel 13), and 12.3 μm (channel 15) wavelengths and/or combinations of these bands are sensitive to pertinent features in thunderstorms, such as cloud type, overshooting tops, cloud particle size, cloud-top glaciation, and cloud-top height (Pavolonis et al., 2005; Elsenheimer and Gravelle, 2019). These four ABI bands are commonly referred to as the red, snow/ice, clean longwave window, and dirty longwave window bands, respectively (Schmit et al., 2017). With passive visible and infrared satellite data, various geostationary satellite–based convection nowcasting tools have been developed. As an example, Roberts and Rutledge (2003), using GOES-8 (0.62 μm and 10.7 μm wavelengths) and Weather Surveillance Radar-1988 Doppler (WSR-88D) ground-based data, found the rates of cooling of cloud-top brightness temperatures were important for discriminating between weakly precipitating storms and vigorous convective storms. Their work increased the lead time and accuracy of convective storm forecasts produced by the Auto-Nowcast System originally developed in Mueller et al. (2003)).

Both ground-based radar and geostationary satellite data contribute valuable information for detecting thunderstorms throughout their various life stages. Because of this, a radar-based definition of thunderstorm initiation was used in studies such

as Mecikalski and Bedka (2006), Mecikalski et al. (2010a,b), and Walker et al. (2012) to evaluate satellite cloud-top signatures for developing convective storms. The underlying purpose of these studies was to enhance convective storm forecasting. Further improvements to convection nowcasting models came soon after with the adoption of machine learning techniques (Mecikalski et al., 2015; Lagerquist et al., 2021; Bradshaw, 2021).

Recently, Lagerquist et al. (2021) applied deep learning to nowcast convection in Taiwan with Himawari-8 satellite data using U-Net convolutional neural network architectures, originally designed for classification of biomedical imagery in Ronneberger et al. (2015). U-Nets have become popular for semantic segmentation tasks in atmospheric sciences where pixel-by-pixel predictions are valued (as opposed to categorizing whole images). However, in order to apply the results of Lagerquist et al. (2021) to the United States, adjustments are necessary to account for differing meteorological conditions and satellite coverage. With this in mind, there are indications that deep learning models, such as the vanilla U-Net in Lagerquist et al. (2021), can be applied to the CONUS with adjustments to the deep learning model organization (i.e. size of the images used and hyperparameter changes) as well as to the selected model inputs and target.

Instead of using Himawari-8 satellite data as inputs (Lagerquist et al., 2021), geostationary satellite data from the GOES-16 satellite can be used as inputs because the data has full coverage of the contiguous United States and the satellite bands can be related to physical cloud processes as suggested by Lee et al. (2021) and Bradshaw (2021). Lee et al. (2021) identified present time convection using GOES-16 data (at wavelengths of 0.65 μm and 11.2 μm) and an echo-classification algorithm (detailed in Zhang et al. (2016)), indicating machine learning models can learn physical properties of clouds from GOES-16 high-spatial and high-temporal resolution data. This was also demonstrated in Bradshaw (2021) where a U-Net was implemented for a small portion of the CONUS to predict daytime convection for the next hour.

In this chapter (Chapter 2) a U-Net, similar to the one implemented for convective storm nowcasting in Lagerquist et al. (2021), is adapted to nowcast thunderstorms in the CONUS, using a 60 minute prediction window, based on complex

spatial and spectral patterns in GOES-16 ABI imagery. The model, referred to as ThunderCast, is trained using NEXRAD radar data. If successfully demonstrated in a testbed environment, ThunderCast could be used to support operational applications by providing lead time to thunderstorm initiation (see Chapter 1 for more information regarding lead time importance) prior to radar thunderstorm signatures in all CONUS regions, including those without radar coverage. ThunderCast can also be used to gain insight into the relationship between radar and satellite products and physical processes.

2.2 Methodology

Model Structure

To train a deep learning model, a set of inputs and a corresponding target (also referred to as the "truth" or desired values) are presented to a deep learning architecture through many iterations to optimize the resemblance of the model output to the given target. This process is depicted in Fig. 2.1 and described in detail in Chollet (2018). The model architecture includes successive layers to transform the data into meaningful representations, which are characterized by a set of weights to determine the relative contribution of each layer to the final output. Weight updating occurs after a subset of the data, called a batch, is processed. A cycle through all the batches (the entire training dataset) forms an epoch. The model trains for many epochs until the difference between the model output and inputs, called the loss as measured by the loss or cost function, is minimized (Chollet, 2018; Stevens et al., 2020). Once trained, the model goes through validation to tune selected model hyperparameters, such as the batch size and the learning rate. After training and validation are complete, a set of inputs can be passed through the model to obtain a set of predictions without using the target. As shown in Fig. 2.1, cross entropy for a binary model [as defined in Cintineo et al. (2022)] and Adam (Kingma and Ba, 2017) were chosen as this application's loss function and optimizer, respectively. The final model hyperparameters for this study are

displayed in Table 2.1 and in the code repository provided in the data availability statement at the end of this paper. The hyperparameters were chosen based on computing restraints (e.g., number of graphical processing units (GPUs)), trial and error (e.g., early stopping patience), and the scientific understanding of the given task as discussed throughout this paper (e.g., number of inputs and class weights).

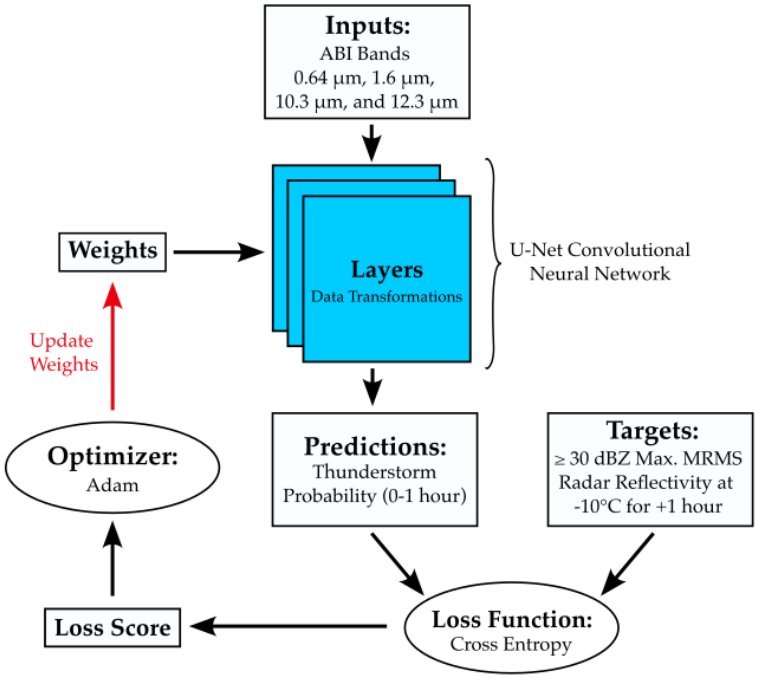

Figure 2.1: Diagram of model training for ThunderCast. The cycle is repeated until the model loss is minimized.

ThunderCast's layers are arranged in a U-Net convolutional neural network, as shown in Fig. 2.2. The initial layers (Fig. 2.2 upper left) are the inputs to the model, with subsequent layers representing data filters, which encode or decode features of the input data (Chollet, 2018). The number of layers, or depth of the U-Net, is consistent with the U-Nets described in Cintineo et al. (2022) and (Ronneberger et al., 2015) and was not varied as a part of the hyperparameter

Table 2.1: Select hyperparameters configured for the training process.

Hyperparameter	Value(s)
Number of Inputs	4
Number of Classes	2
Class Weights	[0.05, 0.95]
Number of GPUs	4
Batch Size	32
Learning Rate	10^{-3}
Early Stopping Patience	4
Reduce Learning Rate on Plateau Patience	1 epoch
Reduce Learning Rate on Plateau Cooldown	2 epochs
Reduce Learning Rate on Plateau Factor	0.1
Trainable Parameters	1.9 Million

tuning process. Three types of data transformations are used in the U-Net: two-dimensional convolutions, max pooling, and upsampling. Convolutions use the scalar product of the model weights (kernel) and the inputs to extract translationally invariant spatial hierarchies of features (Chollet, 2018; Stevens et al., 2020). During most of the convolutions, the data are split into 3×3 overlapping windows with a stride of one and padding with zeros to ensure the output has the same dimension as the inputs. These are accompanied by a rectified linear unit (ReLU) activation function, allowing for nonlinearity (Maas et al., 2013). However, the last convolution uses a 1×1 window and a softmax activation function to obtain one gridded segmentation map of probability scores with values between zero and one at 1-km resolution (320 km $\times$ 320 km, matching the target dataset's resolution) as output. The convolutions are important for isolating and learning local data patterns, and max pooling allows for learning at multiple spatial scales, which is for learning patterns regarding convection. Max pooling aggressively downsamples the data by taking the maximum of the series with 2×2 windows and a stride of two (downsamples by a factor of two). The size of the layers is halved during this procedure, so after max pooling, a window will view data from a larger area. Thus,

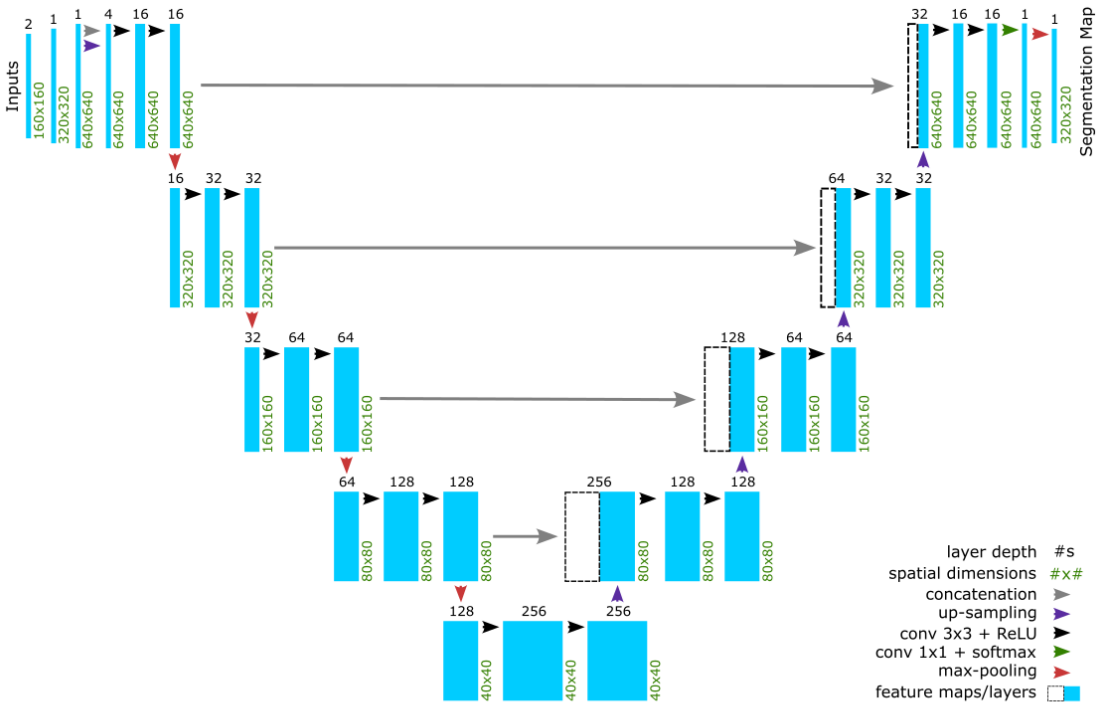

Figure 2.2: Depiction of ThunderCast's U-Net convolutional neural network model architecture. The figure is adapted for this application from Ronneberger et al. (2015).

initially, the U-Net learns small-scale data patterns, and it learns large-scale patterns after max-pooling transformations. At the bottom of the "U" in the U-Net, shown in Fig. 2.2, the image has a low resolution. Upsampling acts to return the image to a high resolution in order to obtain a pixel-by-pixel result. In particular, upsampling is performed with a combination of a three-dimensional transposed convolution and a concatenation (skip connection). Concatenations retain overall prediction details while helping to converge on a loss value during training (Ronneberger et al., 2015).

Data

GOES-16 (also currently designated GOES-East) ABI spectral bands with wavelengths of 0.64 μm (channel 2 reflectance at 0.5 km resolution), 1.6 μm (channel 5 reflectance at 1 km resolution), 10.3 μm (channel 13 brightness temperature at 2 km resolution), and 12.3 μm (channel 15 brightness temperature at 2 km resolution) compose the model input for ThunderCast. These ABI spectral bands were selected because they consist of multispectral imagery that is commonly utilized by forecasters for diagnosing trends and patterns in cumuliform clouds, including cloud-top glaciation, cloud-top temperature, and morphology (Elsenheimer and Gravelle, 2019). All four channels are included in the predictors regardless of solar illumination. However, in the absence of sunlight, the visible and shortwave infrared channels are near zero and are not expected to contribute to predictions. As shown in Fig. 2.2, the channels with resolutions of 1 or 2 km are upsampled to 0.5-km resolution at the beginning of the U-Net architecture to preserve the fine resolution of the visible band. Additionally, each input is normalized by subtracting the mean and dividing by the standard deviation of each respective spectral band in the training dataset.

To convey prediction uncertainty, the desired output of the model is interpreted as a 1-km grid of 60-min thunderstorm probabilities. To train the model to produce an optimal prediction of this nature, the target represents whether a thunderstorm has occurred between the input time and an hour afterward and is obtained with a grid of maximum Multi-Radar Multi-Sensor [MRMS; described further in Zhang et al. (2016) and Smith et al. (2016)] radar reflectivity at the -10°C isotherm in the atmosphere. The grid is binarized such that any pixel with maximum reflectivity greater than 30 dBZ during the hour after the input time is considered positive for thunderstorm occurrence. To maximize the lead time obtained with the model, 30 dBZ is selected as the radar threshold because it corresponds to the earliest thunderstorm radar signature.

Preprocessing is necessary to address data, computing, and machine learning limitations. Because large data files require more computing memory than is

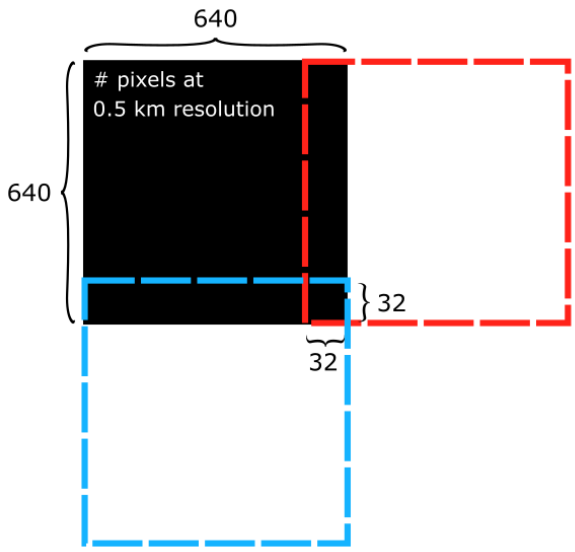

Figure 2.3: Example of a sample corner patch (black) with overlapping adjacent patches (blue and red). The number of pixels are labeled in 0.5 km resolution (black) and the patches are not drawn to scale.

typically available on current GPU servers for model training, it is not feasible to supply data from full-domain, high-resolution GOES-16 scans to the model. To account for this sort of limitation, Liu et al. (2018) implemented a patch-wise sampling technique, and a similar approach is used here. The GOES-16 scans are automatically split into small 640 $\times$ 640 pixel patches at 0.5-km resolution (simply referred to as patches or data patches in this paper). To avoid a loss of contextual information from splitting apart storms along patch borders, patches overlap adjacent patches by 32 pixels (0.5-km resolution) on each of their sides, as shown in Fig. 2.3. Additional errors in training or an unreasonable model output caused by data artifacts is avoided by rejecting patches containing at least one of the following: any "not a number" (NaN) values, MRMS data at -10°C with more than 1% of pixels with no coverage (designated by -999 values within the MRMS radar reflectivity at the -10° isotherm dataset), or any other fill values.

At any one time in the CONUS, the number of negative pixels (without thunder-

Table 2.2: The number of 640 $\times$ 640 pixel (0.5 km resolution) data patches in the training, validation, and testing datasets are provided. The years used for each dataset are indicated in parentheses.

Dataset	Number of Patches
Training (2019)	131,042
Validation (2020)	65,497
Testing (2021)	64,184

storm activity) far outnumbers the number of positive pixels (with thunderstorm activity). A sample is considered a positive patch when more than 1% of pixels are positive. Even positive patches generally have a class imbalance, as there are many negative pixels in a given scene. If this goes unchecked, a model can be highly accurate by never making a positive prediction. To mitigate this, a majority of negative patches are not used. Within a randomized list of negative patches, every 100th case is selected for use in the model to ensure the model has some exposure to a variety of midlatitude weather phenomena. The average percentage of positive pixels in the patches is used to set the class weights prior to training (the class weights are shown in Table 2.1).

Developing a deep learning model can be broken into three stages, each with its own separate dataset: training, validation, and testing. The training dataset is used to learn the weights necessary for an optimized model; the validation dataset is used for evaluating a model's skill during training for hyperparameter tuning; and the testing dataset is used to determine the model's after-training statistics, such as accuracy, precision, and recall. To ensure the samples between datasets are separated temporally, the training, validation, and testing datasets were taken from the years 2019, 2020, and 2021, respectively. The total data patches in each dataset are provided in Table 2.2. Within the datasets, patches are collected from all months in a year, as shown in Fig. 3.4. Additionally, to ensure ThunderCast has exposure to many types of terrain and environments, patches can originate from all climate regions in the CONUS as well as from areas with radar coverage near

the coasts, referred to as outside CONUS (OCONUS). The CONUS climate regions are shown in Fig. 2.5, where a patch is considered a part of the region where the center pixel is located. The numbers of patches in each climate region for training, validation, and testing are shown in Fig. 3.5. Variations in the number of patches per month and per climate region can be accounted for due to the temporal and spatial variability of thunderstorms in the United States. The monthly and regional distributions (Figs. 3.4 and 3.5) indicate that all three datasets (training, validation, and testing) well represent the thunderstorm climatology for 2019–21.

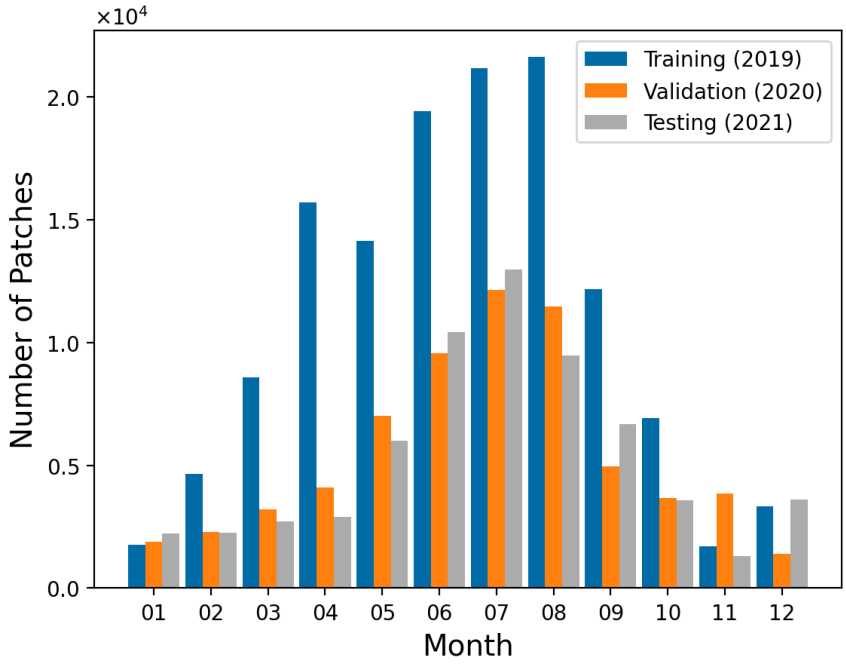

Figure 2.4: Temporal distribution of data patches comprising the training, validation, and testing datasets.

During initial model validation, predictions were biased toward west-to-east storm motion, which is the predominant atmospheric steering pattern in the CONUS. For example, for a known eastward-moving storm, predictions were elongated eastward of existing radar reflectivity at −10°C, suggesting in the next hour there could be storm development toward the east of the existing reflectivity, or there could be

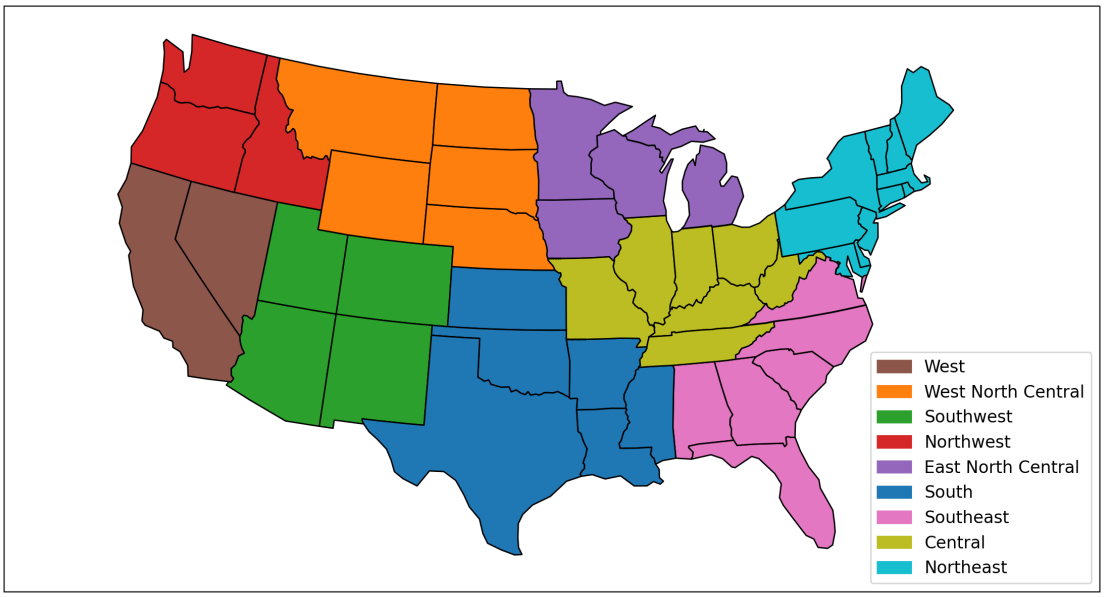

Figure 2.5: Map of the continental United States with state colors corresponding to the state's respective climate region in accordance with Karl and Koss (1984).

eastward motion of the existing storm system. The eastward prediction elongation continued to occur when the storm was flipped over the vertical axis (becoming a westward-moving storm in a time series of patches). To address this bias, data augmentation is utilized. When batches of data patches are imported for training, each batch has a 25% chance the patches will be flipped over the vertical axis, followed by a 25% chance of being flipped over the patches' horizontal axis. This results in some patches being flipped horizontally, vertically, or both horizontally and vertically during model training.

Evaluation Statistics

Traditional machine learning and forecasting statistics are used to evaluate model performance in Section 2.3. Although ThunderCast's outputs are probabilistic predictions for the next hour, the predictions are binarized for statistical evaluation. Predictions greater than or equal to a probability threshold (e.g., greater than or

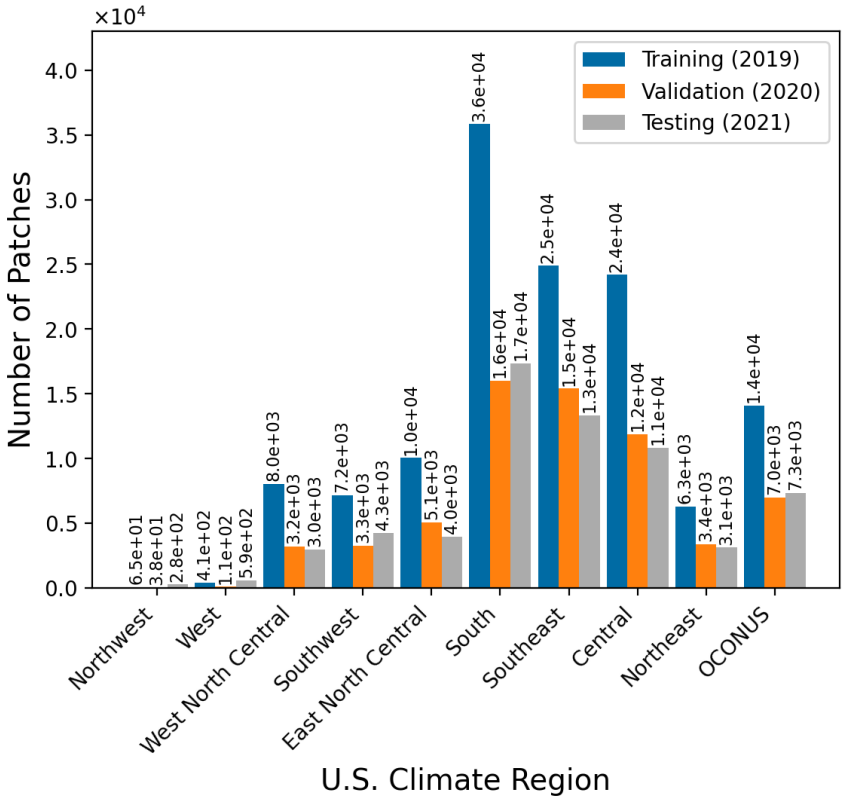

Figure 2.6: Spatial distribution of data patches comprising the training, validation, and testing datasets sorted by climate regions in the CONUS and nearby the United States coasts with radar coverage (OCONUS). The number of patches are indicated by the height of the bars as well as the values above each bar ("$e+0X$" indicates "$\times 10^{X}$").

equal to 50% probability) can either be a "yes" thunderstorm (positive) or a "no" thunderstorm (negative) case. A prediction can then be placed in one of four categories: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). True positives or negatives occur when both the prediction and observation are positive or negative, respectively, false positives are categorized when the prediction is positive while the observation is negative, and false negatives are categorized when the prediction is negative while the observation is positive. Since ThunderCast makes pixel-by-pixel predictions, the total pixels in each category

are counted for the entire training, validation, or testing set [represented as sums in Eqs. 2.1-2.6] and used to compute accuracy, precision, recall, specificity, false alarm ratio, critical success index, and frequency bias [represented in Eqs. 2.1–2.7]. With potential values in the range of $[0, 1]$, 1 is an ideal value for accuracy, precision, recall, specificity, and critical success index and 0 is ideal for the false alarm ratio. Frequency bias ranges over $[0, \infty)$ and is ideally close to 1:

$$\text{accuracy} = \frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{TP} + \sum \text{FP} + \sum \text{FN} + \sum \text{TN}} \tag{2.1}$$

$$\text{precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \tag{2.2}$$

$$\text{recall} = \text{Probability of Detection (POD)} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \tag{2.3}$$

$$\text{specificity} = \frac{\sum \text{TN}}{\sum \text{TN} + \sum \text{FP}} \tag{2.4}$$

$$\text{false Alarm Ratio (FAR)} = \frac{\sum \text{FP}}{\sum \text{TP} + \sum \text{FP}} \tag{2.5}$$

$$\text{critical Success Index (CSI)} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN} + \sum \text{FP}} \tag{2.6}$$

$$\text{frequency bias} = \frac{\text{POD}}{1 - \text{FAR}} \tag{2.7}$$

For visual interpretation of the ThunderCast model's performance, an attribute diagram [developed in Hsu and Murphy (1986)] and a performance diagram [developed in Roebber (2009)] are provided in Section 2.3. Forecasts are considered reliable or calibrated when the relative frequency of occurrence of events is equal to the probability forecast for those events. Thus, plotting the conditional event frequency with respect to the forecast probability gives a reliability curve. The attributes diagram displays the reliability curve, with reference lines for skill levels (Hsu and Murphy, 1986). The performance diagram combines the recall or

probability of detection [Eq. 2.3], the success ratio [the false alarm ratio in Eq. 2.5 subtracted from 1], and the critical success index [Eq. 2.6] into one diagram. A perfect performance diagram case results when the area under a curve with data points representing many probability thresholds is 1.

Another statistical metric called the fraction skill score (FSS) is used to determine how ThunderCast's forecast skill varies with spatial scale (Roberts and Lean, 2008). For every grid point in a binarized grid of predictions (of size N_x by N_y in the Cartesian grid), the average number of positive grid points within a centered square (window) of length n is computed. The result forms a density field $P_{(n)i,j}$, where i and j are indices for the x–y Cartesian grid. A similar grid, $T_{(n)i,j}$, is also calculated for the target. This process can be described as taking the average pooling of the binarized predictions and the target. The density fields $P_{(n)i,j}$ and $T_{(n)i,j}$ are then used to calculate the mean-square error (MSE) for the target and prediction densities [$MSE_{(n)}$; Eq. 2.8] and for a reference case [$MSE_{(n)ref}$; Eq.2.9]. These quantities form the basis for the fraction skill score shown in Eq. 2.10). The definitions are

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[T_{(n)i,j} - P_{(n)i,j} \right]^2, \tag{2.8}$$

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} T^2_{(n)i,j} + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P^2_{(n)i,j} \right], \text{ and} \tag{2.9}$$

$$\text{fraction skill score (FSS)} = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}} \tag{2.10}$$

In addition to statistical evaluations, ThunderCast is applied to three distinct thunderstorm growth environments: a mesoscale convective vortex, sea breezes, and a southwestern monsoon. Each case study includes visual representations of the satellite spectral band (inputs), radar reflectivity at the -10°C isotherm (target), and ThunderCast predictions. Forecast lead times from ThunderCast's predictions to the occurrence of 30 dBZ at -10°C are presented.

2.3 Results

Model Performance

Evaluating the model on data the model has not seen before with the testing dataset provides context on how the model is performing. The pixel-to-pixel statistics shown in Table 2.3 indicate the model is performing well for accuracy, recall, and specificity but has a low value for precision. Based on the equations for these metrics [Eqs. 2.1-2.4], ThunderCast tends to have a high number of false positives, resulting in low precision. However, a pixel-to-pixel evaluation does not take into account any slight offsets between the predictions and the target. The fraction skill score [2.10)] provides insight into whether predictions are skillful within a spatial range. Fraction skill scores for probability thresholds ranging from 10% to 90% for varied spatial window sizes are shown in Fig. 2.7. In Fig. 2.7, the fraction skill score slightly increases for all probability thresholds as the window size increases, although minimal improvements occur after the windows reach a length (n) between 13 and 17 pixels (equivalent to 13–17 km, since the spatial resolution is 1 km). This indicates the skill of the model improves when evaluated within a spatial range of the predictions.

If the scope of the target values resulting in true positives is broadened, such that any positive target value within a 15 km $\times$ 15 km window centered on the location of a prediction would result in a true-positive value, then the precision [Eq. 2.3] improves, as shown in Table 2.3. The broadened target in the testing dataset is referred to as the buffered target or buffered dataset in this paper. Although other values could be chosen for the buffer window, the fraction skill score improves minimally with windows greater than 13–17 km, and 15 km was found to represent the model well in a trial-and-error attributes diagram analysis. The increased precision achieved with the buffered dataset indicates ThunderCast's predictions may be useful within a buffered area surrounding a prediction. However, there is a tradeoff in doing this. In Table 2.3, the improvements in precision are concurrent with decreases in recall. The recall equation [Eq. 2.3] is the same as that for

Table 2.3: Model performance for the full, daytime-only, and nighttime-only testing datasets, where probabilities greater than or equal to 20% are considered to be positive for thunderstorm activity. Additional columns include values computed using an alternate test set with the same data patches but the target values have been adjusted such that all pixels within a 15 km $\times$ 15 km window centered on a pixel containing a maximum reflectivity of 30 dBZ or greater at -10°C for the next hour are also considered positive for thunderstorm occurrence (called a buffered dataset here).

	Full Set		Day Only		Night Only	
Buffer	None	15 km $\times$ 15 km	None	15 km $\times$ 15 km	None	15 km $\times$ 15 km
Accuracy	0.796	0.854	0.805	0.865	0.783	0.839
Precision	0.156	0.585	0.158	0.634	0.154	0.527
Recall	0.927	0.747	0.928	0.737	0.927	0.761
Specificity	0.790	0.878	0.800	0.896	0.777	0.855

precision, except recall has the sum of false negatives in the denominator instead of false positives. Thus, a decrease in recall with the buffered dataset indicates there will be more false negatives than with the original nonbuffered testing dataset.

Using the buffered target in the testing dataset, an attributes diagram and a performance diagram are provided in Figs. 2.8 and 2.9, respectively. In Fig. 2.8, the 1:1 dashed gray line represents an ideal model. For example, in an ideal model, predictions of a 40% probability of thunderstorms in the next hour would result in a thunderstorm (true positive) 40% of the time. The attributes diagram in Fig. 2.8 indicates predictions between approximately 20% and 35% result in a thunderstorm more often than anticipated, while forecast probabilities greater than approximately 38% result in a thunderstorm less often than in the ideal model case. However, almost all forecast probabilities demonstrate skill because they are above the diagonal no-skill line shown. In Fig. 2.9, the critical success index [Eq. 2.6], a measure often used for forecast performance evaluation, falls between 0.4 and 0.5 for all thunderstorm probabilities, with the highest critical success indices occurring between 20% and 40%. Probabilities less than 50% are slightly

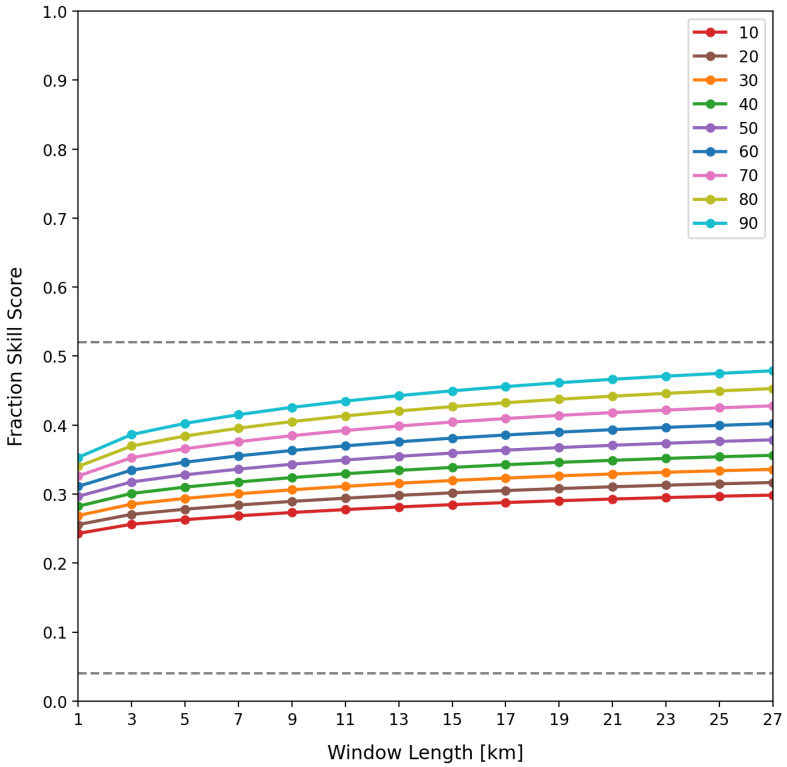

Figure 2.7: Fraction skill score diagram based on the method presented in Roberts and Lean (2008). The fraction skill scores [Eq. 2.10] are calculated for probability thresholds ranging from 10% to 90% for various window lengths. The window lengths [n in Eq. 2.10] are given in kilometers but can be referred to as pixels or grid spaces since the spatial resolution is 1 km. The probability thresholds (%) are colored according to the legend in the upper-right-hand corner of the diagram. The lower dashed gray line represents the fraction of observed points exceeding 30 dBZ at -10°C over the domain and is called the random fraction skill score. The upper dashed gray line is the uniform fraction skill score and marks halfway between the random and perfect skill scores.

overforecast (frequency bias is greater than 1), and probabilities greater than 50% are underforecast (frequency bias is less than 1). The 20% probability value, although overforecast, has one of the highest values for the critical success index and is skillful according to the attributes diagram, so probabilities greater than

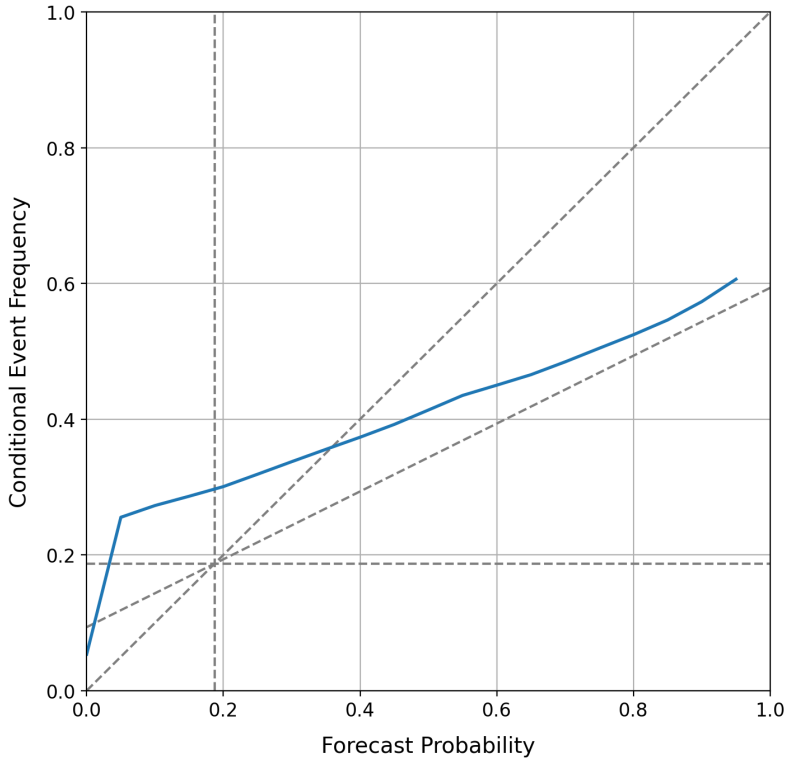

Figure 2.8: Attributes diagram for the ThunderCast model constructed with the method presented in Hsu and Murphy (1986). The blue line represents the conditional event frequency (true positives per total positive predictions) for given forecast probabilities for the testing dataset. The dashed gray lines are reference lines for determining model resolution. The vertical and horizontal dashed gray lines are the no-resolution lines equal to the testing dataset's overall relative frequency of thunderstorm occurrence. The 1:1 dashed gray line (upper diagonal line) represents a perfect reliability or forecast calibration. The lower dashed diagonal gray line is the no-skill line, where anything below this line is considered to have no skill (Brier skill score of 0).

or equal to 20% are considered positive for thunderstorm activity in the statistical calculations in Eqs. 2.1–2.4, as presented in Table 2.3 and Fig. 2.10.

For further statistical evaluation, the testing dataset is broken into subsets to analyze performance across the diurnal (day and night) and monthly cycles and

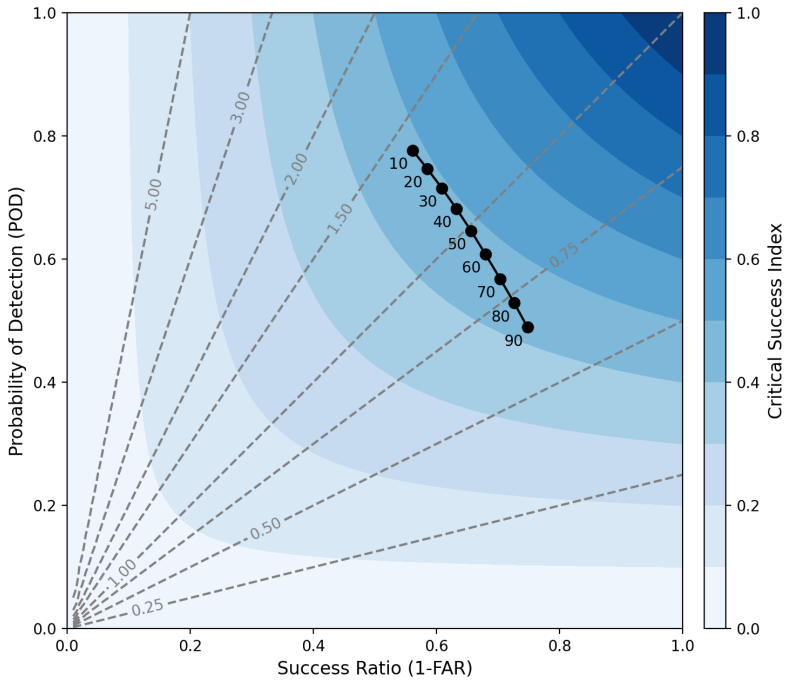

Figure 2.9: Performance diagram based on Roebber (2009) for the ThunderCast model on the testing dataset. The y axis displays the probability of detection [Eq. 2.3], and the success ratio [i.e., 1−FAR, or 1−2.5] is on the x axis. The background colors denote the critical success index [Eq. 2.6], and the dashed contours represent the frequency bias. Each black data point is labeled with the corresponding probability threshold (%) used for calculating the indicated quantities.

spatially across the CONUS and nearby OCONUS regions. In Table 2.3, the pixel by-pixel accuracy, precision, recall, and specificity for the day and night subsets are consistent with the full testing dataset results. With the target buffer, the differences in model precision between the day and night subsets are more pronounced, indicating the model may be more precise within a 15 km $\times$ 15 km centered window during the day than during the night. Fig. 2.10 demonstrates the accuracy, precision, recall, specificity, and critical success index's spatial and monthly variability. Some regions (e.g., the northwest) do not have data points for each month in the testing dataset (2021), because data patches were not present in those regions during the

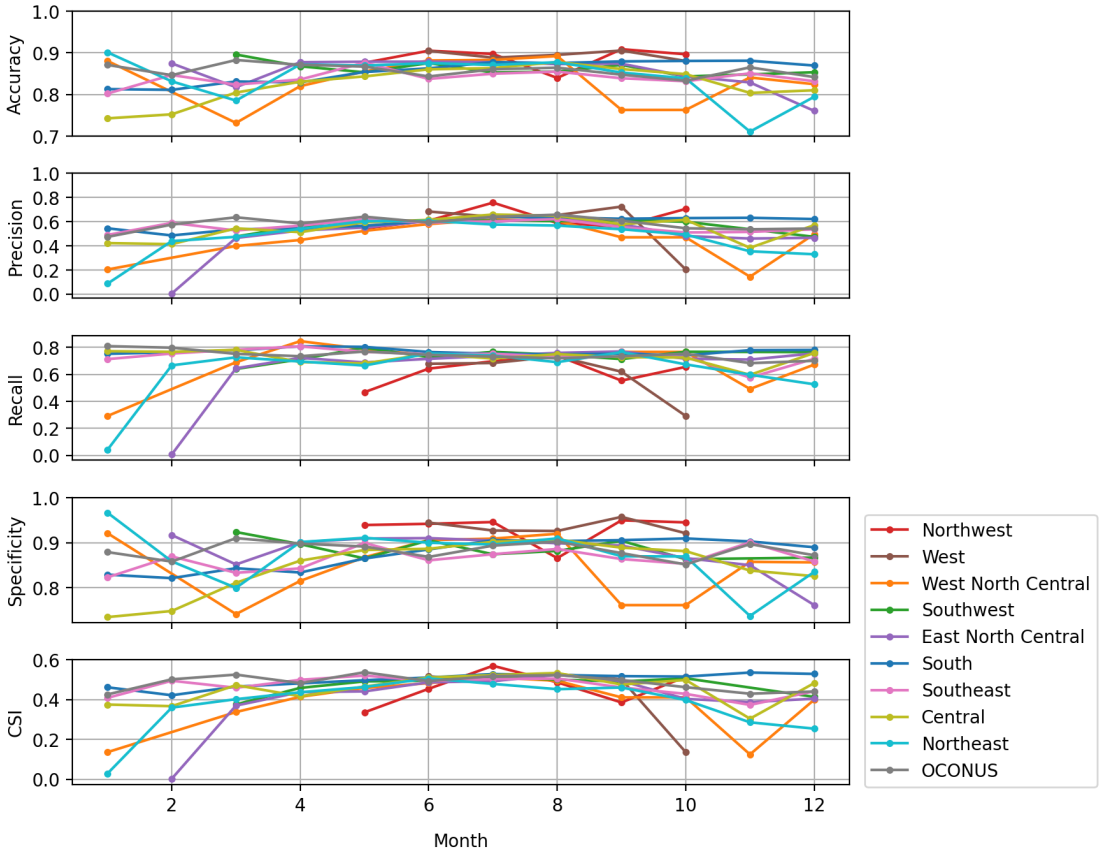

Figure 2.10: Accuracy, precision, recall, specificity, and critical success index [Eqs. 2.1-2.4 and 2.6] for ThunderCast's predictions on the testing dataset, with target values buffered by a 15 km $\times$ 15 km centered window. Probabilities greater than or equal to 20% are considered to be positive for thunderstorm activity in the calculations for each statistic. The testing dataset is sorted by region, and each data point represents the corresponding statistic value for only the testing data patches in the month given. For spatial reference, the climate regions are distinguishable by their colors, matching Fig. 2.5.

respective months. Across CONUS and OCONUS regions, the model performs similarly during the main convective months (midlatitude summer), with higher variability during the winter months. Although statistical insight is valuable, it is also important to determine in which cases ThunderCast is performing well and not

as well. A detailed evaluation of this is left for future work, but a sampling of case studies is presented in the remainder of this Chapter to demonstrate ThunderCast's initial prediction capabilities.

Case Study 1: Mesoscale Convective Vortex

Mesoscale convective systems are organized cloud structures containing convective (cumulonimbus) and stratiform (nimbostratus) clouds, with a mesoscale cirriform cloud shield in the topmost cloud layers. In the stratiform region of a mesoscale convective system, a cyclonic vortex can form as a result of pre-existing cyclonic absolute vorticity and heating gradients. This vortex, often referred to as a mesoscale convective vortex, can trigger new convection within a mesoscale convective system (Houze Jr., 2014). On 25 August 2021, a mesoscale convective vortex over eastern Iowa resulted in thunderstorm development in western Illinois. Fig. 2.11 shows a daytime cloud-phase distinction false-color red–green–blue (RGB) image [following the methods presented in Elsenheimer and Gravelle (2019)] of one such storm. The image depicts the beginning of the mature stage of the thunderstorm. The cold, bubbling, overshooting top is visible in the bright orange/red part of the cumulonimbus cloud, as is the beginning of the thunderstorm's anvil. The underlying green clouds are glaciated cumulus. An animation provided in the online supplemental material for this paper shows the initiation and development of the thunderstorm through the mature stage. An animation provided in the Supplementary Materials in Ortland et al. (2023) shows the initiation and development of the thunderstorm through the mature stage.

Fig. 2.12 shows ThunderCast's predictions for the storm's initiation along with ABI 0.64 μm mm reflectance and the associated MRMS radar at -10°C. ThunderCast's first predictions of 40% or greater occur at 18:31 UTC, when low-level cumulus clouds are present. The first signature of convective initiation from radar occurs at 18:56 UTC, when the clouds have become glaciated, 25 min after ThunderCast's initial predictions. To provide further context for this event, Fig. 2.13 displays the storm's later development between 19:31 and 20:46 UTC with processed data

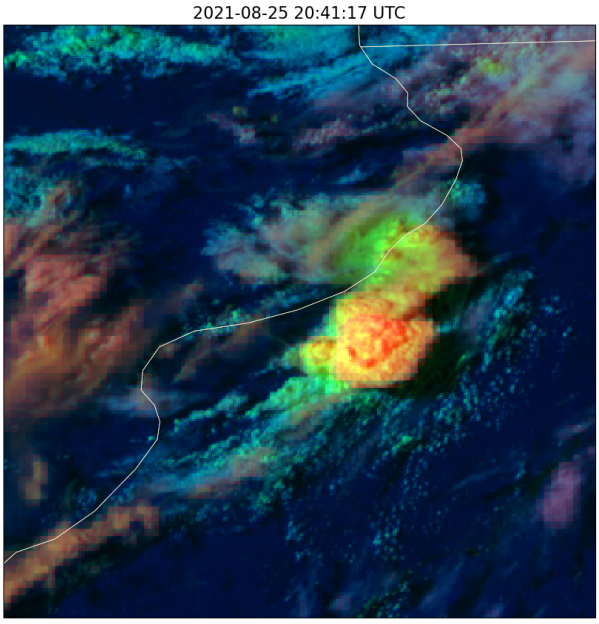

Figure 2.11: A daytime cloud-phase distinction false color red-green-blue (RGB) image for August 25, 2021 at 20:41 UTC centered at 41.44° latitude and −90.56° longitude. Following Elsenheimer and Gravelle (2019), red is ABI 10.3 µm (channel 13) brightness temperatures, green is ABI 0.64 µm (channel 2) reflectance, and blue is ABI 1.6 µm (channel 5) reflectance. The light tan line shows state borders.

from the GOES-16 Geostationary Lightning Mapper (GLM) (Rudlosky et al., 2019; Goodman et al., 2013). According to Fig. 2.13, the storm becomes electrically active at 19:46 UTC (50 min after 30 dBZ at −10°C) prior to anvil formation and has moved northeast of its convective initiation location in Fig. 2.12. ThunderCast's predictions in Fig. 2.12 at 18:56 UTC are elongated toward the northeast of the active convection (30 dBZ at −10°C) into nonclouded areas. Thus, ThunderCast's next-hour thunderstorm predictions demonstrate ThunderCast may be anticipating something about storm motion through learned data patterns.

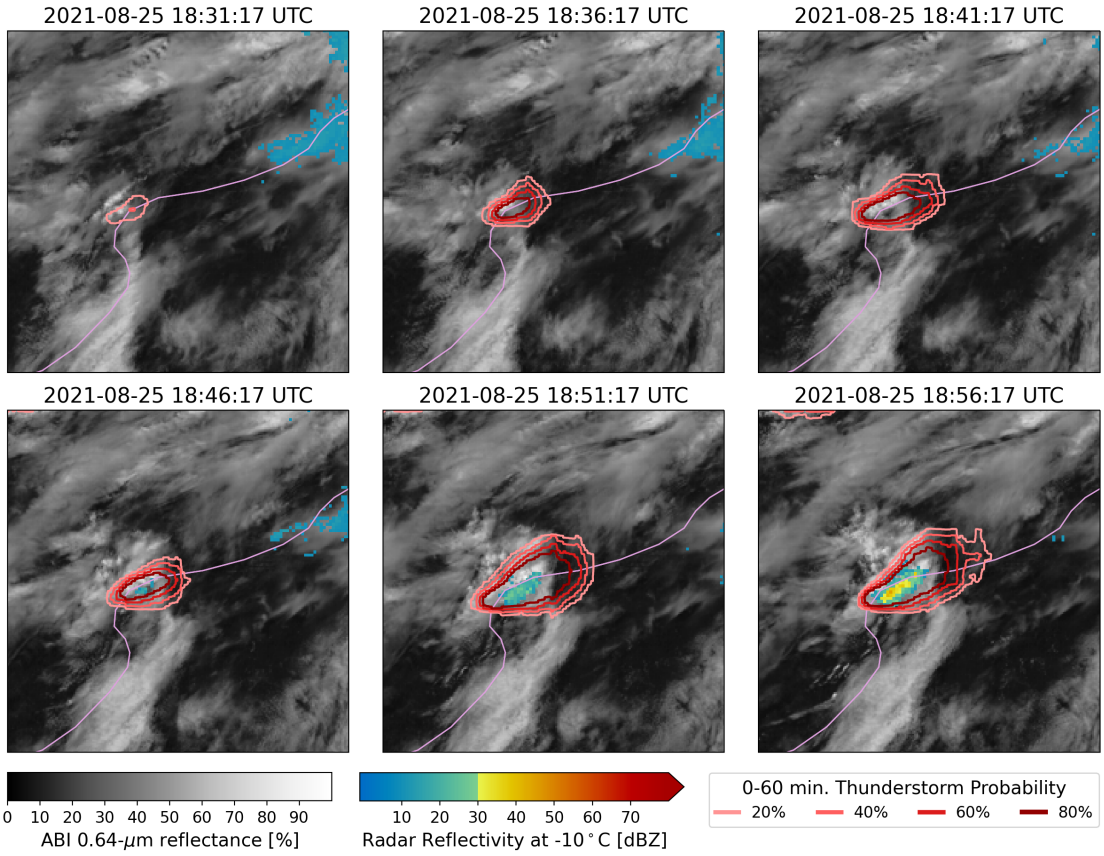

Figure 2.12: A paneled image time series from August 25, 2021 between 18:31 and 18:56 UTC centered at 41.41° latitude and −90.85° longitude. Each panel contains a background of ABI 0.64 μm reflectance layered with radar reflectivity at −10°C and ThunderCast's thunderstorm probabilities displayed as contours. The thunderstorm probabilities are valid for up to 1 h from the time listed above each panel. The radar reflectivity colorbar is adapted from Helmus and Collis (2016) and state borders are indicated by the light purple line for spatial context. Each image is 112 km × 112 km.

Case Study 2: Sea Breeze Convection

Differential heating between the land and the sea is a forcing mechanism for new convection. High-pressure (cooler) sea air is driven toward the low-pressure (warmer) air over the land, forcing the warmer air to rise. In satellite imagery, this

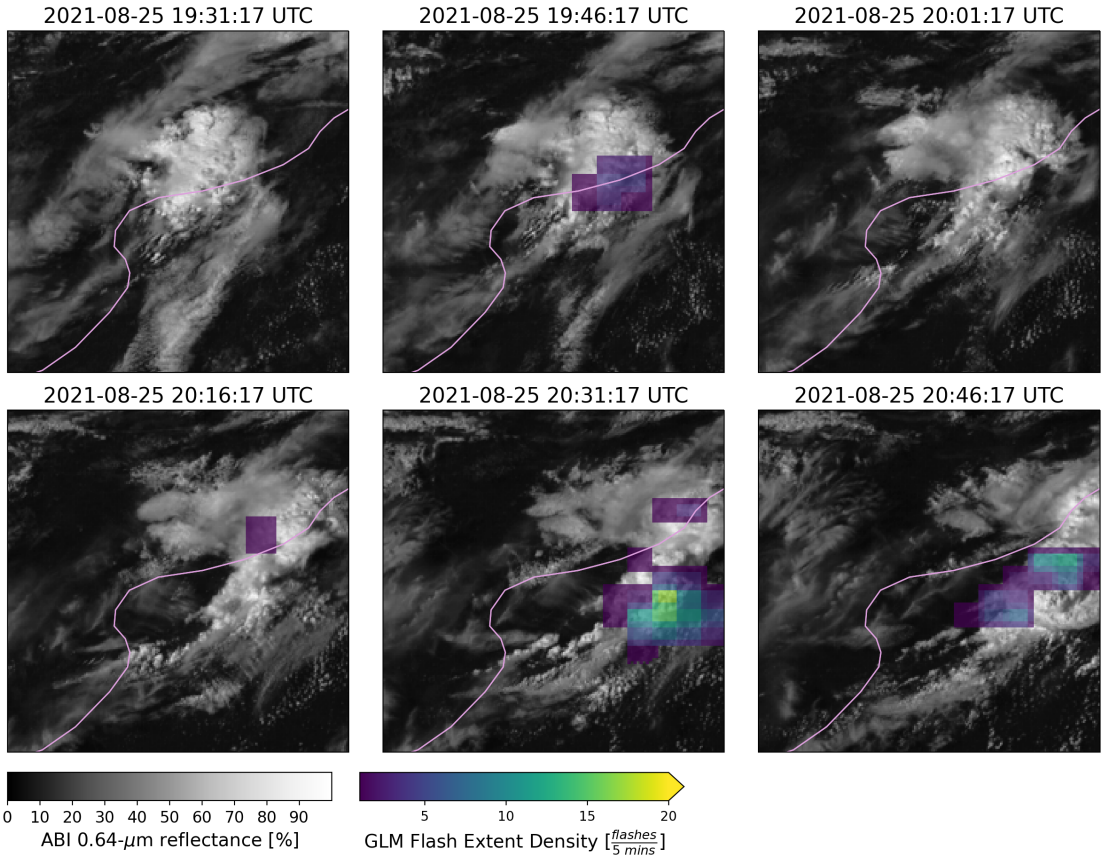

Figure 2.13: A paneled image time series between 19:31 and 20:46 UTC August 25, 2021 centered at 41.41° latitude and −90.85° longitude. Each panel contains ABI 0.64 µm reflectance, GOES GLM flash extent density in flashes per 5 min (Bruning et al., 2019), and light purple state borders. Each image is 112 km × 112 km.

type of convection can be identified along coasts with landward-moving surface winds, cumulus development along the land–sea boundary, and activity in the afternoon when the temperature difference between the land and the sea is greatest (Scofield and Purdom, 1986). Oftentimes, sea-breeze convection can develop rapidly, making it difficult for forecasters to issue timely advisories and warnings. For this case study, the 45th Weather Squadron, U.S. Space Force, provided 159 cases between July and August 2022 for which they were unable to achieve the

desired 30 min of lead time prior to observed thunderstorm hazards for regions of interest in and around Cape Canaveral, Florida. Of these cases, 88 (55%) were associated with sea-breeze activity, one of which is shown in Fig. 2.14.

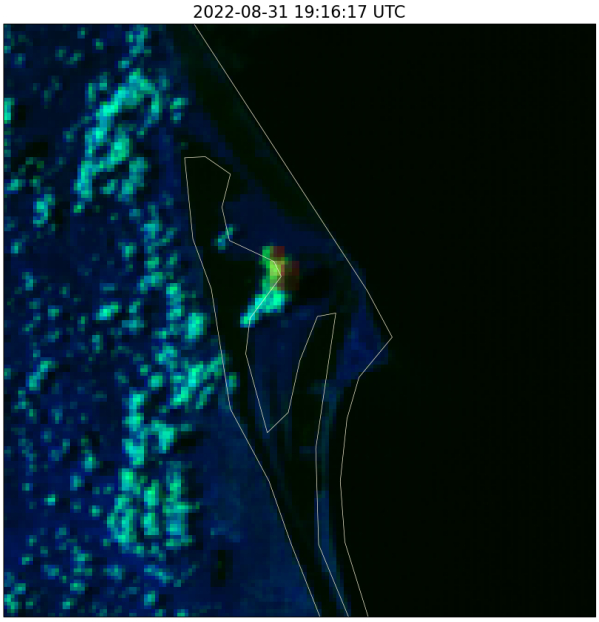

Figure 2.14: Daytime cloud-phase distinction false-color RGB image at 19:16 UTC 31 Aug 2022 centered at 28.52° latitude and −80.6° longitude. Following Elsenheimer and Gravelle (2019), red is ABI 10.3 μm (channel 13) brightness temperatures, green is ABI 0.64 μm (channel 2) reflectance, and blue is ABI 1.6 μm (channel 5) reflectance. The image is 80 km × 80 km, and the light-tan line shows state borders.

In the sea-breeze example shown in Fig. 2.14, scattered low-level cumulus clouds line the coastline, with an area of towering cumulus-containing ice particles, signified by the green and orange colors present in the RGB image, near the center of the figure. An RGB animation of the life cycle of this short-lived storm is provided in the online supplemental material in Ortland et al. (2023). As shown in Fig. 2.15, ThunderCast is able to identify which cumulus, among the cumulus field, are likely to initiate a thunderstorm. ThunderCast's predictions (20% probabilities or greater) first appear at 19:06:17 UTC, 5 min prior to cloud-top glaciation and MRMS radar reflectivities greater than or equal to 30 dBZ at −10°C. The 45th Weather

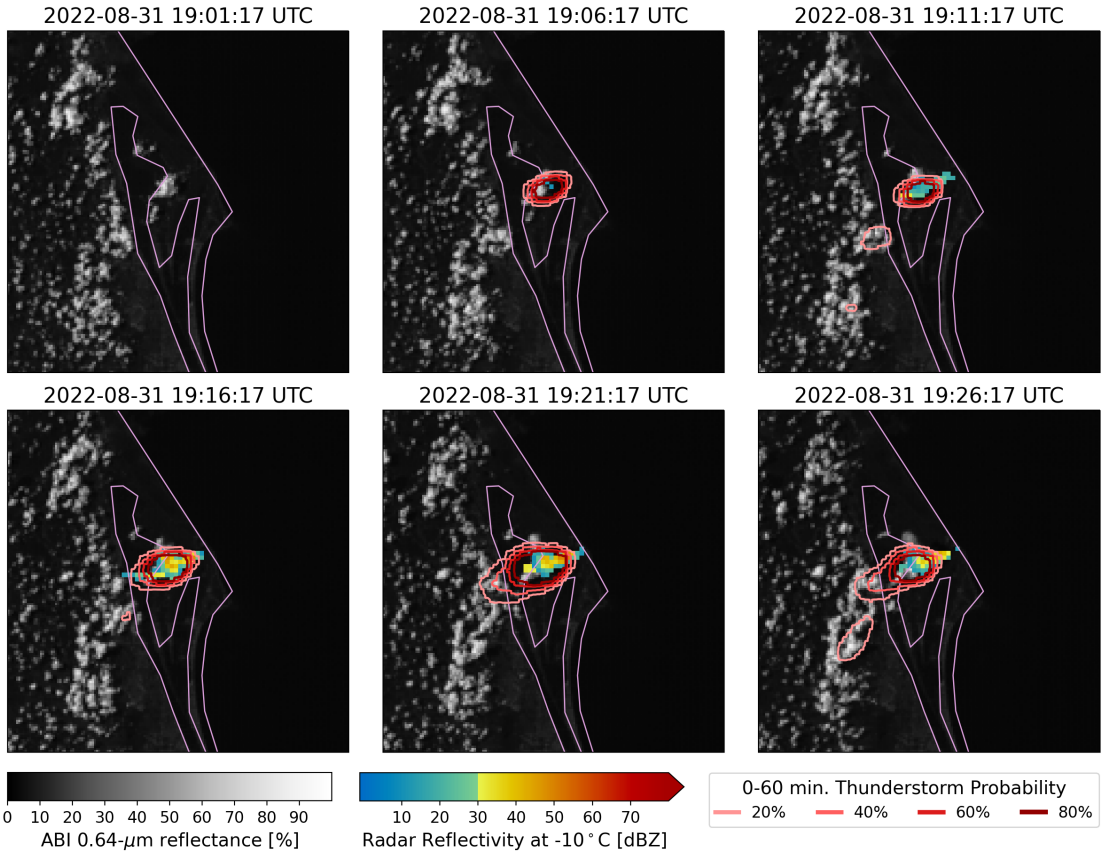

Figure 2.15: Paneled image time series between 19:01 and 19:26 UTC 31 Aug 2022 centered at 29.52° latitude and 280.65° longitude. Each panel contains a background of ABI 0.64 μm reflectance layered with radar reflectivity at −10°C and ThunderCast's thunderstorm probabilities displayed as contours. The thunderstorm probabilities are valid for up to 1 h from the time listed above each panel. The radar reflectivity color bar is adapted from Helmus and Collis (2016), and coastal borders are indicated by the light purple line for spatial context. Each image is 80 km × 80 km.

Squadron, U.S. Space Force, issued a lightning hazard watch at 19:23 UTC and recorded observed lightning at 19:26 UTC. Thus, ThunderCast's 30 dBZ predictions occur 20 min before the storm becomes electrically active in this case.

To assess lead time for sea-breeze cases, one randomly selected sample dur-

ing each of the 22 sea-breeze days highlighted by the 45th Weather Squadron is collected for statistical analysis. Only one sample was selected each day to avoid overrepresentation of a single storm due to the close proximity of the sampled regions near Cape Canaveral (i.e., one storm may span multiple regions and have multiple data points). On average, ThunderCast's first instance of a 20% probability prediction occurred 18.4 ± 5.8 minutes (95% confidence interval) prior to the first instance of 30 dBZ at -10°C within a 15 km $\times$ 15 km window centered on the prediction; 30 dBZ at -10°C occurred, on average, 46.6 ± 16.3 minutes (95% confidence interval)(95% confidence interval) before observed lightning. Although more work is needed to test the model in varying thunderstorm environments, this demonstrates the potential ThunderCast may have for providing forecasters advance notice of clouds with the potential to become a thunderstorm in the next hour, perhaps aiding in decision support.

Case Study 3: Southwestern Monsoonal Convection

On 27 August 2022, scattered non-severe thunderstorms were present in the western United States across Utah, Arizona, New Mexico, and Colorado. Fig. 2.16 shows some of these storms in northeastern Arizona and northwestern New Mexico, the area of interest for this case study. Many cumulus at various development stages are visible in Fig. 2.16, including some thick, high-level cumulus with ice particles (yellow/orange colors) exhibiting overshooting tops with anvils as well as some lower-level water-based cumulus and towering glaciated cumulus. An animation following the development of these clouds is provided in the Supplementary Materials in Ortland et al. (2023). Fig. 2.17 shows ThunderCast's predictions as these cumulus cloud clusters develop.

ThunderCast's prediction lead time varies depending on the storm cluster of interest. However, the initial predictions tend to appear when the clouds are low-level cumulus, and 30 dBZ at -10°C coincides with glaciation and colder cloud tops. In the 17:26 UTC panel in Fig. 2.17, 20% and 40% probability contours are present in the bottom left. This area reaches 30 dBZ at -10°C at 18:01 UTC, achieving 35 min

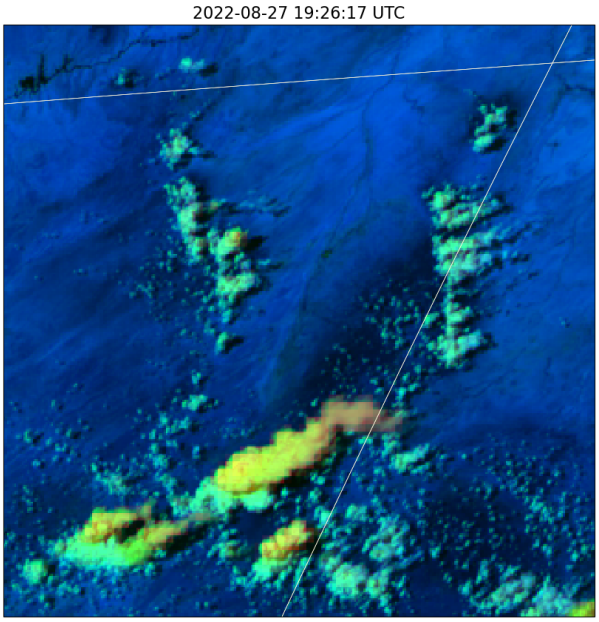

Figure 2.16: Daytime cloud-phase distinction false-color RGB image at 19:26 UTC 27 Aug 2022 centered at 36.13° latitude and −109.56° longitude. Following Elsenheimer and Gravelle (2019), red is ABI 10.3 μm (channel 13) brightness temperatures, green is ABI 0.64 μm (channel 2) reflectance, and blue is ABI 1.6 μm (channel 5) reflectance. The image is 160 km × 160 km, and the light-tan line shows state borders.

of lead time. Lightning is observed from the GLM at 1831 UTC, 30 min after 30 dBZ at −10°C. In the 18:11 UTC panel of Fig. 2.17, an area containing a 60% probability contour is present in the upper left. This area's first predictions of 20% or greater appear at 18:01 UTC, 30 dBZ at −10°C is recorded at 19:26 UTC, and lightning is observed from the GLM at 20:01 UTC. Thus, ThunderCast achieves 25 min of lead time to 30 dBZ at −10°C, which occurs 35 min before observed lightning. A lead time is not available for the storm in the upper right, spanning the Arizona and New Mexico border, because this area does not have MRMS radar coverage. However, ThunderCast's ability to make predictions in a no-coverage radar zone could be useful for forecasting thunderstorms in similar regions across the CONUS.

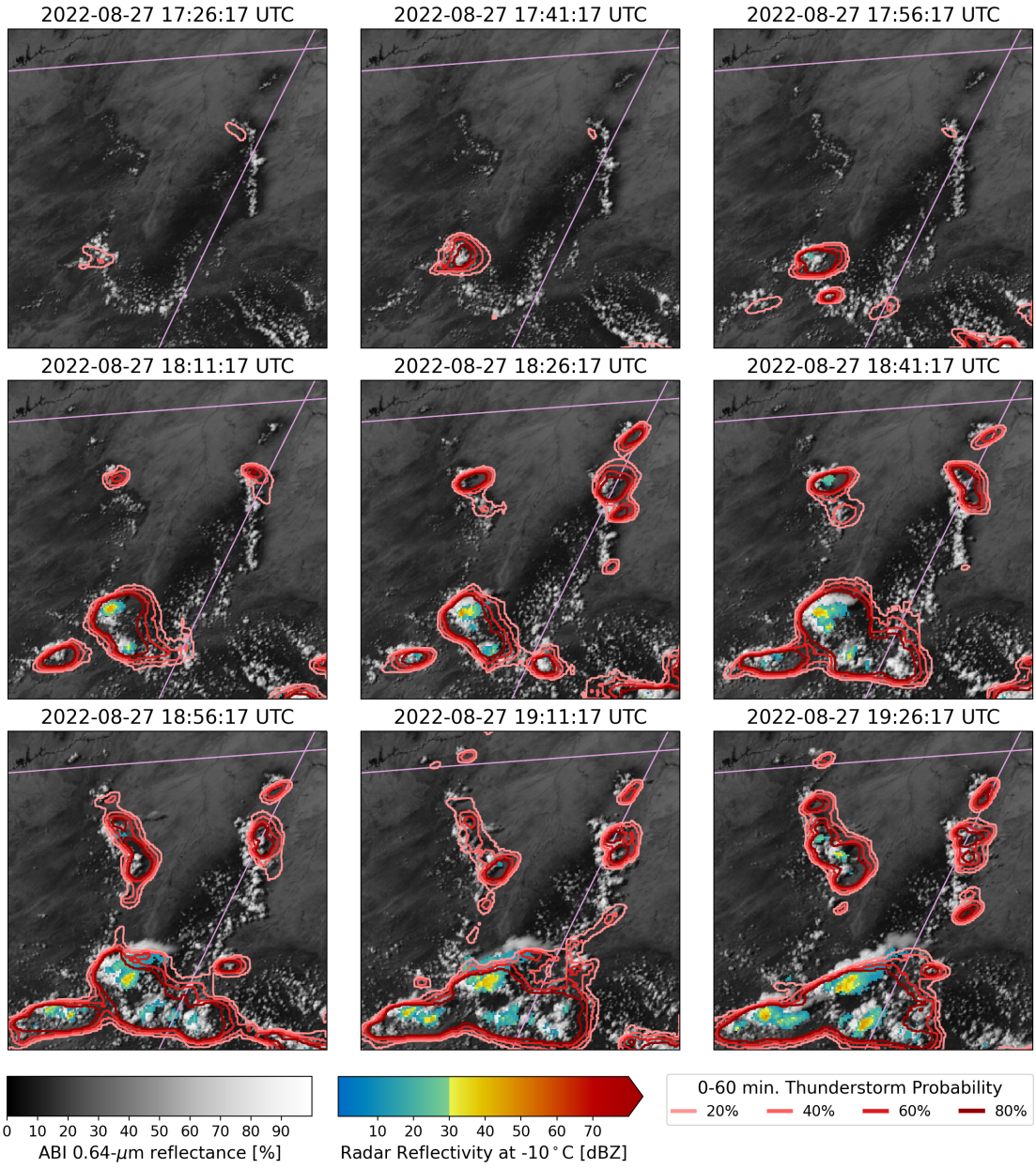

Figure 2.17: Sequence of images depicting the evolution of ThunderCast probabilities for weak thunderstorms in Arizona and New Mexico between 17:26 and 19:26 UTC 27 Aug 2022. Centered at 36.13° latitude and −109.56° longitude, all images contain a black-and-white background of ABI 0.64 µm reflectance, the available radar reflectivity at −10°C, and contours of ThunderCast probabilities. The probabilities are valid for up to 1 h from the time listed above each panel, and the reflectances are not corrected for parallax. The radar reflectivity color bar is adapted from Helmus and Collis (2016). Each image is 160 km × 160 km, and state borders are indicated by the light purple lines.

2.4 Summary

The Thunderstorm Nowcasting Tool, or ThunderCast, was developed for predicting thunderstorm occurrence for both developing and existing storms in the next 0–60 min in the conterminous United States. ThunderCast is a deep learning model built with a U-Net convolutional neural network for semantic segmentation. The model was trained with four predictors: 0.64-μm reflectance (channel 2; red band), 1.6-μm reflectance (channel 5; snow/ice band), 10.3-μm brightness temperature (channel 13; clean longwave window band), and 12.3-μm brightness temperature (channel 15; dirty longwave window band) from the GOES-16 Advanced Baseline Imager. All four channels are included in the predictors regardless of solar illumination; however, in the absence of sunlight, the visible and shortwave infrared channels are not expected to contribute to predictions due to near-zero values. The target dataset was the maximum Multi-Radar Multi-Sensor radar reflectivity at -10°C in the next hour, where anything greater than or equal to 30 dBZ was considered positive for thunderstorm occurrence. To address computational limitations and potential biases in the datasets, the model was trained on 320 km $\times$ 320 km data patches, where patches could be flipped vertically or horizontally during data augmentation, and the model weights were heavily in favor of the positive class due to the negative pixels being far more common than the positive.

On a pixel-by-pixel basis, ThunderCast performs well statistically in terms of accuracy, recall, and specificity, but it has a low value for precision resulting from a high amount of false positive predictions. The precision improves when evaluating the model within a buffered target region, where any occurrence of 30 dBZ at -10°C within a 15 km $\times$ 15 km window centered on the prediction is considered positive for thunderstorm activity. Considering the buffered target, ThunderCast was found to make skillful predictions with critical success indices ranging between 0.4 and 0.5 for all thunderstorm probabilities, with the highest values occurring between 20% and 40%. These statistics are consistent across CONUS climate regions, with some increases in variability during the Northern Hemisphere's winter months. Similarly, ThunderCast performs fairly consistently during the day and the night

but tends to have higher precision during the day.

ThunderCast was applied to three case studies to demonstrate the model's initial prediction capabilities. Each case study represented convection from varying convective environments across the CONUS, including convection associated with a mesoscale convective vortex, sea breezes, and a southwestern monsoon. At least 5 min and up to 35 min of prediction lead time to 30 dBZ at -10°C was achieved, and the 30 dBZ at -10°C radar threshold occurred 30–50 min prior to observed lightning. In addition to lead time, ThunderCast demonstrated it is able to make predictions in no-coverage radar zones and anticipate storm motion from a single timestamp of four GOES-16 ABI satellite bands without additional ambient environmental context. Although more work is needed to test the model in varying thunderstorm environments, these initial case studies demonstrated ThunderCast is capable of providing forecasters advance notice on the location of thunderstorms in the next hour in the CONUS.

Further analysis is presented in the next chapter (Chapter 3) to determine if ThunderCast's identification of areas with the potential for convective initiation can be used as a proxy for thunderstorm hazards like lightning. In theory, all thunderstorms will have lightning, because it is needed to produce the characteristic thunder for which the storms are named. However, there may be instances where 30-dBZ radar reflectivity at -10°C is observed without lightning or where 30 dBZ at -10°C is not observed when lightning is produced. To investigate how often these cases occur, chapter 3 uses an object-based method to collect instances of ThunderCast's predicted thunderstorms along with the corresponding lightning, satellite, and radar observations.

3 AN OBJECT-BASED EVALUATION OF OUTPUT FROM THUNDERCAST

3.1 Background

Semantic segmentation, a pixel-by-pixel labeling technique for computer vision tasks, was applied to near-term probabilistic forecasting (nowcasting) of thunderstorms in the contiguous United States in Chapter 2. The deep learning model, called the Thunderstorm Nowcasting Tool (ThunderCast), can be used day or night and in any season to predict the occurrence of convection associated with thunderstorms in the next 0-60 minutes. As described in Section 2.2, ThunderCast was built with a U-Net convolutional neural network architecture. Four satellite channels in the visible, shortwave infrared, and longwave infrared bands from the *Geostationary Operational Environmental Satellite-16* (*GOES-16*) Advanced Baseline Imager (ABI) served as the predictors. The training target, chosen based on a radar definition of convective initiation of thunderstorms, was the next-hour maximum Multi-Radar Multi-Sensor (MRMS) radar reflectivity at the -10°C isotherm in the atmosphere. Any next-hour maximum reflectivity at $-10^{\circ}\text{C} \geqslant 30$ dBZ was considered a thunderstorm.

Although performing well during the case studies presented in Section 2.3, there are situations where ThunderCast's predictions are not ideal. The statistical analysis in Section 2.3 showed ThunderCast has low precision (the number of correct predictions divided by the total predictions), indicating it is prone to false alarms. This is not surprising because other models involving the prediction of convective initiation have a similar tendency (McGovern et al., 2023). The second version of the Satellite Convective Analysis and Tracking (SATCASTv2) convective initiation algorithm reported an average false alarm ratio for all study domains of 55% in Walker et al. (2012). Mecikalski et al. (2015) reported lower, but still present, false alarm percentages (between 22 and 36%) for their random forest machine learning approach with incorporation of data from numerical weather

prediction models. More closely related to ThunderCast because of the similar U-Net architecture and incorporation of all thunderstorm events instead of only convective initiation, is Lagerquist et al. (2021)'s models for thunderstorm prediction in Taiwan with *Himawari-8* satellite data. The false alarm rates ranged between approximately 40 and 60% for the best probability thresholds (Lagerquist et al., 2021). Likewise, ThunderCast's false alarm ratio for the 20% probability threshold falls in this range at approximately 45%. Although the false alarm rates between models are not directly comparable due to differences in methodologies and study areas, all the mentioned models are prone to false positive predictions, which affects the reliability of the models' nowcasts.

Additional influences on ThunderCast's reliability and applicability to thunderstorm nowcasting relate to the radar-based definition of convective initiation used for classifying the target dataset when training the deep learning model. As discussed in Section 2.1, radar thresholds for convective initiation of thunderstorms originate from field campaigns such as the National Hail Research Experiment in northeastern Colorado, southwestern Wyoming, and/or southwestern Nebraska (Morgan and Squires, 1982; Fankhauser and Wade, 1982; Dye and Martner, 1982). Microphysical observations during the organizational stage of the storms observed in the experiment corresponded to reflectivity values between approximately 30 and 40 dBZ (Dye and Martner, 1982). To maximize ThunderCast's prediction lead time, a radar reflectivity threshold of 30 dBZ at $-10°$C was chosen for convective initiation of thunderstorms, which agrees with the threshold used in Wilson and Schreiber (1986). However, defining thunderstorm initiation from a radar definition is imperfect. Dye and Martner (1982)'s variability in the correspondence of radar reflectivities to the microphysical observations in developing storms attests to this. The imperfections associated with the use of a radar reflectivity threshold have resulted in other techniques for convective storm identification like Lee et al. (2017)'s echo-class algorithm approach, which was favored by Lagerquist et al. (2021). With regards to ThunderCast, there could be cases where $\geqslant 30$ dBZ reflectivities do not result in an electrified storm, and lightning is a necessary ingredient for a convective storm to be classified as a thunderstorm. Alternatively, there could

be false positive cases where areas with less than 30 dBZ correspond to lightning activity. Furthermore, the regionality of radar reflectivity thresholds for convective initiation could potentially cause erroneous nowcasts when applying one threshold across a vast spatial area containing many types of terrain (e.g. mountains, coasts, and plains).

To assess the reliability and applicability of ThunderCast for thunderstorm prediction in the United States, additional analysis is needed to determine the extent of false positive and true positive predictions with and without lightning activity in ThunderCast. However, existing statistical analyses of deep learning models focus on pixel-by-pixel instead of case-by-case evaluations. This is problematic because it does not allow for the collection of lead times for storm events, and it is not aligned with a forecasting standpoint. Forecasters tend to look for regions of predictions instead of individual pixels through time when evaluating large areas for potential severe weather activity. To address this, case studies are often included in convective machine learning applications (Mecikalski et al., 2015; Han et al., 2019; Lee et al., 2021; Lagerquist et al., 2021; Cintineo et al., 2022) including ThunderCast. A large-scale case-by-case evaluation expanding on the cases presented in Section 2.3 could provide the desired information for the evaluation of false positive and true positive cases in ThunderCast.

However, the copious case studies needed to assess the model across the entirety of the contiguous United States (CONUS) in a statistically meaningful way are not easily collected due to the need for human evaluation of each case. To overcome this, the present paper applies object-tracking software to track regions of predictions in ThunderCast. The tracked predicted storm objects are then classified into true positive or false positive with or without lightning based on observational data and used for large-scale statistical analysis and interpretation of predicted storm events. The resulting method can be applied to other machine learning applications for forecasting in atmospheric science and support improvements for convective nowcasting models through the evaluation of potential physical or spatial influences on erroneous nowcasts.

3.2 Methodology

Identifying and Tracking Predicted Thunderstorm Objects

To identify and track predicted thunderstorm objects in ThunderCast, tobac, an established tool for the Tracking and Object-Based Analysis of Clouds described in Heikenfeld et al. (2019) and Sokolowsky et al. (2023), is applied to ThunderCast's deep learning model output. Although developed for cloud tracking, the open-software Python package is input variable agnostic because it does not rely on specific input variables or a specific grid. Thus, ThunderCast's model predictions can be input into tobac to obtain trackable prediction objects. The process for doing so is described in the steps below, and tobac's Python package parameter configuration for ThunderCast predictions is summarized in Table 3.1.

Table 3.1: Specifications for tobac's Python package.

Parameter	Value(s)
Input Feature Detection Threshold(s)	[0.1, 0.9]
Target	Maximum
Position Threshold	Weighted Difference
Erosion Threshold	4
Sigma Threshold	2
Minimum Number of Pixels	3
Minimum Distance Between Features	15000 m
Segmentation Method	Watershed
Linking Method	Random
Maximum Tracking Velocity	100 $\frac{m}{s}$
Minimum Tracking Timestamps	6

Detect Features

In tobac, features are points representing contiguous areas of values above or below a threshold or series of thresholds (Heikenfeld et al., 2019). The thresholds can be

based on minima or maxima in a dataset depending on the use case. For this case, the output from ThunderCast is more likely to result in thunderstorm activity with higher values, so features are detected based on maxima in the input dataset. The location of the feature point within a large area of predictions meeting a threshold is determined with weighted differencing. Following the approach presented in Heikenfeld et al. (2019), the thresholds are determined by analyzing the distribution of one day of predictions with a histogram as shown in Figure 3.1. The date chosen (2022-08-03) represents a variety of convection events during the active convection season in the United States. Due to the two distinct bins with the highest number of predictions in Figure 3.1, the thresholds chosen are probabilities of 10% and 90%. Assigning multiple thresholds allows larger areas of probabilities with multiple maxima to be broken into pieces. This is beneficial for identifying predicted storms located close together spatially (e.g. predictions for a multicellular thunderstorm cluster).

To aid in the feature detection process, ThunderCast's input values are filtered with a Gaussian filter and binary erosion. The Gaussian filter smooths ThunderCast's predictions and binary erosion reduces the size of the detected objects (areas surrounding a feature) in all directions. The filters are controlled by the sigma threshold and the erosion threshold integers, shown in Table 3.1, for Gaussian smoothing and binary erosion, respectively. Furthermore, to reduce spurious objects, there must be at least three pixels greater than one or both of the probability thresholds for a feature to be identifiable and the features must be at least 15 km from other features. This distance is the equivalent of 15 pixels because ThunderCast's output is in 1-km resolution (see Section 2.2).

Obtain Objects

After identifying features, a watershed segmentation technique is used within tobac to associate areas with each identified feature (Heikenfeld et al., 2019). These areas are referred to as objects throughout the present paper. Inputs into the watershed algorithm are, as the name implies, treated similarly to a topographic

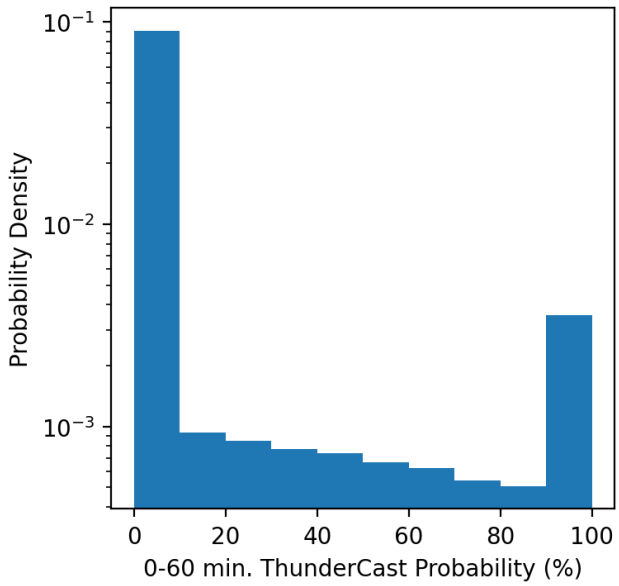

Figure 3.1: A histogram of all the pixel-by-pixel ThunderCast predictions for one day (2022-08-03). The y-axis is logarithmic.

map in a geological context where the map is separated into drainage basins called watersheds (Meyer, 1994; Soille and Ansoult, 1990; Heikenfeld et al., 2019). The watershed segmentation technique in tobac is built with scikit-image, an image processing library in Python (van der Walt et al., 2014; Heikenfeld et al., 2019). An example of ThunderCast prediction objects with their associated features is provided in Figure 3.2.

Track Features

To organize the features and associated objects into tracks of predicted ThunderCast storms through time, the features are linked using trackpy (Allan et al., 2023) within tobac. In particular, the tracker looks for a feature within a search radius in consecutive timestamps with the random method. In the random method, the feature point is the center of the search radius and the distance from the feature to the edge of the search radius is determined from the specified tracking velocity in

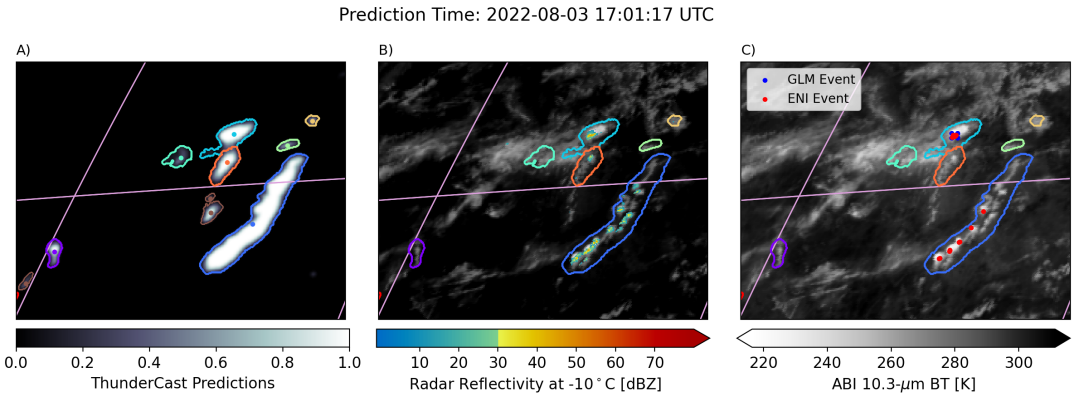

Figure 3.2: An example of the features and objects collected for 2022-08-03 17:01:17 UTC with corresponding observational data for times closest to the prediction time in each dataset. The latitude and longitude for the upper left-hand corners are 39.4° and −111.8°, respectively. All images contain state borders around the four corners region of the U.S. in light purple. Panel A contains the output from ThunderCast with all identified features as colored points and the features' corresponding objects in matching colors. Any brown objects were initially detected with tobac but were removed during post-processing. Both Panel B and C display the objects kept for analysis and the parallax-corrected ABI 10.3 μm brightness temperatures. Panel B shows radar reflectivity at −10°C on top of the ABI. The radar reflectivity color bar is adapted from Helmus and Collis (2016). Finally, Panel C contains points for lightning events from both the GLM (blue) and ENI (red).

Table 3.1 (Heikenfeld et al., 2019). The velocity chosen is larger than typical thunderstorm velocities because the location of the feature from weighted differencing can shift farther than a storm would travel naturally, especially for larger objects. Additionally, if multiple features are detected in the search area, the feature with the lowest sum of squared distance between features is connected as a part of the track. To ensure the feature tracks are persistent in time, tracks are only recorded if the track duration is 25 minutes or longer (6 timestamps when timestamps are 5 minutes apart as specified in Table 3.1). More information about tracking with tobac can be found in Heikenfeld et al. (2019).

Identify Connected Tracks

Thunderstorm environments can be complex; there are cases where individual storms merge to create a larger storm system (mergers) or thunderstorms can split into separate storms (splits). Since ThunderCast predicts regions of thunderstorm activity, there are similarly complex mergers and splits in the predicted storm tracks. To identify these complex predicted storm cases, tobac's merge/split submodule is implemented. The submodule uses a minimum euclidean spanning tree along with Kruskal's algorithm and parent/child nomenclature to assign parent track identification numbers (IDs) to all tracked features (Kruskal, 1956; Sokolowsky et al., 2023). The amount of tracks identified with the same parent ID are referred to as the number of children for the parent track ID. A more detailed description of the merge/split submodule in tobac can be found in Sokolowsky et al. (2023). Identifying complex predicted storms is useful for post-processing the ThunderCast object-based dataset.

Post-Processing

After the tracked predicted storms are obtained and connected feature tracks are identified, the predicted storm dataset undergoes post-processing to ensure the dataset avoids superfluous feature tracks and focuses on simple cases containing the initiation stage of development. First, any feature point not associated with a predicted storm track is disregarded to avoid short-lived predicted storms. Next, predicted storms that may be incomplete spatially due to the close proximity to the border of the study domain (CONUS) are ignored as well as those that may be incomplete temporally due to their existence at the first and/or last timestamps during a given day. Furthermore, in initial testing some predicted storm tracks were associated with existing convective systems rather than new convection. To avoid these cases contaminating calculations of lead time or inflating true or false positive counts by counting one storm multiple times, any track with an initial object containing greater than 2000 pixels is removed. Lastly, complex predicted storm cases with merging or splitting are considered out of the scope of this work,

so any tracks with greater than one child are ignored.

Analysis Techniques

To evaluate the extent of false positive and true positive cases with and without lightning in ThunderCast, a variety of data is collected for the predicted storm tracks (along with their associated features and objects) generated with tobac for 7 randomly selected days from each month in the 2022 convective season in the United States (April to September). For diurnal, seasonal, and spatial analysis, each track's time, date, and location are recorded. The location for each track is simplified to the CONUS climate region of the track's initial feature. The climate regions are west, west north central, southwest, northwest, east north central, south, southeast, central, northeast, and outside of the CONUS (OCONUS), matching those specified in Section 2.2 and Karl and Koss (1984). For simplicity, only tracks originating over land and tracks in areas with MRMS radar coverage are used during the analysis in this paper.

Determining if a case is true positive or false positive requires comparison to the target data used in ThunderCast, so radar reflectivity at -10°C is obtained from the Multi-Radar Multi-Sensor [MRMS; described further in Zhang et al. (2016) and Smith et al. (2016)] dataset. If reflectivity values $\geqslant$ 30 dBZ are present in a storm track, the track is considered a true positive case. All other cases are false positive. To avoid extraneous false positives from storm tracks without MRMS radar coverage, any objects (and associated tracks) containing data points without radar coverage are disregarded. The time difference between the object's first prediction greater than or equal to a specified probability threshold and the first occurrence of $\geqslant$ 30 dBZ reflectivity is recorded as the lead time to 30 dBZ at that probability threshold.

Lightning data observed from cloud-top via the Geostationary Lightning Mapper [GLM; Bruning et al. (2019)] and from Earth Networks, Inc. [ENI; Earth Networks (2024)] ground-based sensors (total lightning dataset) are used to determine which true positive and false positive cases achieve electrification. Both GLM

and ENI are used to ensure all possible lightning observances are captured as either dataset could miss instances collected by the other (e.g. satellites are limited to observing flashes visible from the cloud tops and ground-based sensors may miss in-cloud lightning). Because the format of the GLM and ENI lightning data for this analysis are points, a spatial polygon is created with SciPy's convex hull algorithm (Virtanen et al., 2020) from the outer points of the predicted storm objects within each track. If any lightning occurs within a timespan of $\pm$ 2.5 min of an object in the predicted storm track and the lightning's latitude and longitude are within the associated polygon, the predicted storm is considered electrified. Similar to the lead time to 30 dBZ, the time difference between the object's first prediction greater than or equal to a specified probability threshold and first observed lightning in the storm track is recorded as the lead time to lightning at that probability threshold.

Additionally, observational ABI data from the GOES-16 satellite are collected for the predicted storm tracks to gather information about cloud-top properties. To ensure ABI multispectral cloud properties are captured by the appropriate predicted storm objects, the ABI data are parallax corrected using a constant cloud-top height of 8 km. The ABI spectral bands are evaluated for statistical patterns with respect to the true positive and false positive classifications to determine if any of these quantities could help reduce false positive or non-electrified cases. A work-flow chart describing the order of information collection in each predicted storm track is shown in Figure 3.3. For each storm track, the coldest cloud-top brightness temperature prior to electrification is collected by identifying the pixel with the minimum 10.3 μm (channel 13) brightness temperature. At the same location and time the 1.6 μm (channel 5) reflectance is recorded. These bands are sensitive to pertinent cloud-top features of thunderstorms, such as cloud type, cloud-top glaciation, cloud particle size, cloud-top height, and overshooting tops (Pavolonis et al., 2005; Pavolonis, 2010; Elsenheimer and Gravelle, 2019). Because classification into true positives and false positives with and without lightning happens toward the initiation part of each storm track, ABI data are collected until lightning is observed from both ENI and GLM or the end of the predicted storm track is reached.

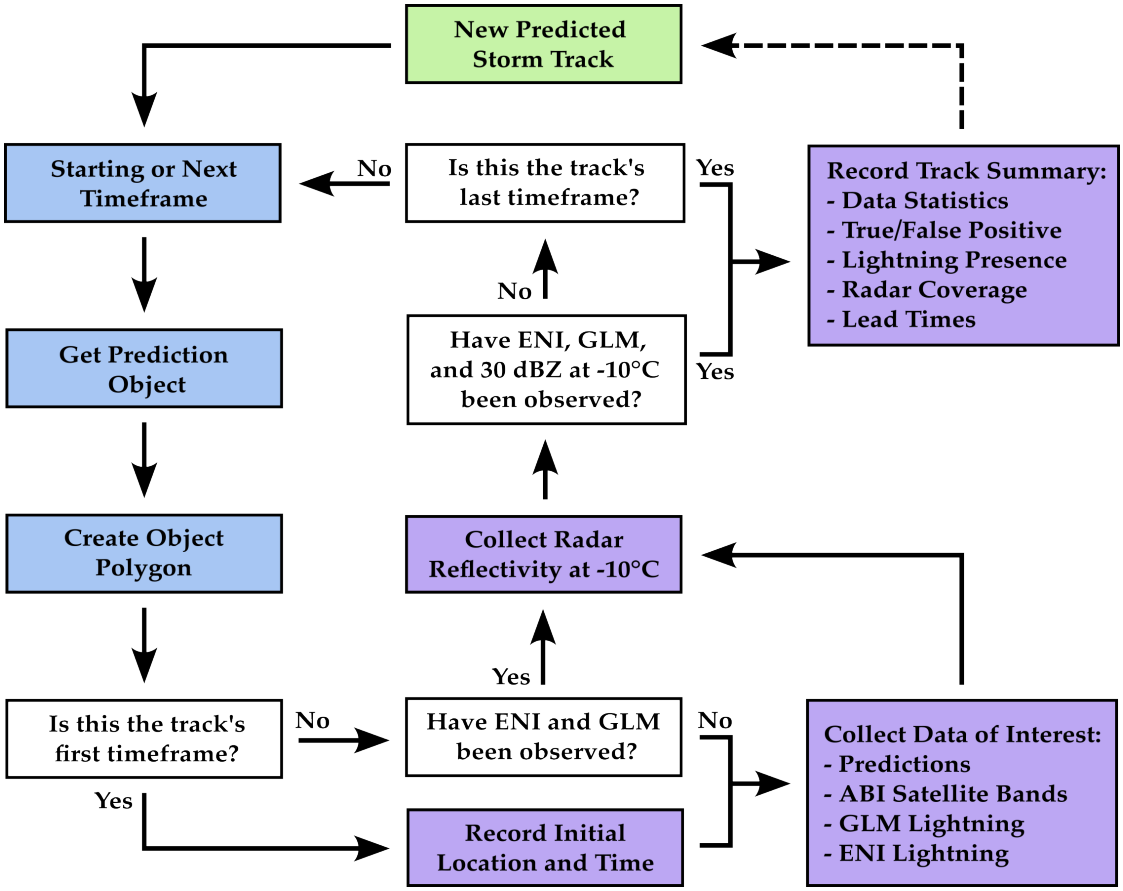

Figure 3.3: The workflow for collecting observational and model data for each predicted storm track.

Anticipated Track Limitations

Due to the decisions and specifications necessary to obtain objects and tracks for ThunderCast's predictions as detailed above and specified in Table 3.1, there are some inherent sources of errors within the dataset. For example, although a maximum tracking velocity of 100 ms^{-1} may be appropriate for the majority of cases, it could result in a new, distinct object nearby an older object from the previous timeframe being recorded as one continuous track instead of a new track. To quantify the extent and potential impact the errors may have within the analysis,

100 randomly sampled predicted storm tracks from the predicted storm dataset (discussed in Section 2.2) were evaluated manually. Each track was labeled in a “yes" or “no" style where tracks were recorded as having a noticeable error during tracking or not. Of the 100 tracks evaluated, 9 of them were identified as having an error from tracking. This indicates less than 10% of predicted storm tracks are anticipated to have errors from the tracking specifications chosen for this task. Since fewer than 10% of the analyzed predicted storms are expected to have an imperfect tracking history, the dataset is suitable for assessing the accuracy and lead time of ThunderCast predictions.

Additionally, the tracks were separately labeled in a “yes" or “no" fashion for noticeable errors in child counts from tobac’s merge/split algorithm. Although any tracks with more than one feature (more than one child) were filtered out, there are cases that appear to have merges or splits that were incorrectly counted as only having one feature/child by tobac. This was identified in 33 of the 100 tracks, so some of the merge/split cases unintentionally remain in this dataset.

3.3 Results

Each predicted storm track identified with tobac is sorted into true positive ($\geqslant$ 30 dBZ) and false positive ($<$ 30 dBZ) categories with and without lightning based on the data collected for the track using the flow-chart in Fig. 3.3. The total counts in each category are shown in Table 3.2. To the nearest tenth of a percent, true positives compose 67.8% of tracks. Because precision is the total true positives divided by the total predictions made (optimal closest to 100%), this value is synonymous to precision. The precision from the pixel-by-pixel evaluation in Section 2.3 was recorded as 15.6% and the precision buffered by a 15 km x 15 km window was 58.5%. The precision from the object-based evaluation of ThunderCast is closest to the buffered precision. This is expected because the object-based method and the buffered method in Chapter 2 are both informed by areas outside of an individual pixel. However, the object-based prediction is larger than the buffered precision, which can be accounted for due to methodology differences. The buffered method

is based on a fixed distance instead of areas of related pixels and could be inflated by reflectivities $\geqslant$ 30 dBZ from unrelated prediction clusters within the buffered window or deflated if a storm larger than the buffered window contains reflectivities $\geqslant$ 30 dBZ outside of the area specified. Additionally, it is important to keep in mind that the precisions from Section 2.3 incorporate all possible occurrences of convection while the object-based method presented in this paper filters the dataset to focus on new convection. Taking these conditions into account, the statistics collected with the object-based method presented in this paper are thought to better represent initiating predicted storms than the buffered approach in Chapter 2 because it uses a watershed method to determine related pixels instead of assuming a distance.

Table 3.2: The number of tracks identified from ThunderCast predictions with tobac for true positive (TP) and false positive (FP) predicted storms. The table is divided into two sections. Each section can be separately summed to equal the total predicted storm tracks. The bottom section splits the TPs and FPs into subcategories according to electrification status. All percentages are rounded to the nearest tenth of a percent.

Category	Number of Predicted Storm Tracks	Percent of Total Tracks (Total = 24770)
All TPs	16790	67.8%
All FPs	7980	32.2%
TPs without lightning	9964	40.2%
TPs with lightning	6826	27.6%
FPs without lightning	7940	32.1%
FPs with lightning	40	0.2%

The second section of Table 3.2 differentiates between electrified and non-electrified predicted storm tracks by identifying whether lightning occurred in the track's lifetime. True positive cases without lightning activity consisted of 40.2% of all predictions. Just considering the true positives, 59.3% of all true positive predictions do not result in electrified convection. This indicates the radar-based def-

inition of convection initiation is not well correlated with lightning initiation. Thus, ThunderCast is more skilled at predicting convective precipitation than lightning initiation.

The sorted predicted storm tracks are further sorted temporally and spatially in Figs. 3.4 and 3.5, respectively. Consistent with the percentages presented in Table 3.2, Fig. 3.4 shows most of the months used for the convective season (April-September in 2022) contain more true positive tracks than false positive tracks and all months contain more true positive tracks without lightning than those with lightning. April's false positives are an exception to this generalization because April contains more false positives than true positives. Characterized by the transition from spring to summer, April typically includes midlatitude cyclones that generate a variety of cloud types and precipitation modes. More variability in the type of convection present could contribute to the larger false positive to true positive ratio in April.

Fig. 3.5 shows all climate regions contain more true positive tracks than false positive tracks, which is consistent with Table 3.2. Also, all of the climate regions contain more true positives without lightning than with lightning except for the southwest. During the 2022 convective season, the southwest was impacted by the North American Monsoon System from mid-June through September. The monsoon transports moisture from the Pacific Ocean, the Gulf of Mexico, and the Gulf of California to normally dry areas, fueling thunderstorm development. In 2022, moisture levels were at a record high during the monsoon season and steering winds were generally weak until early September. It is not clear from this analysis why predictions for convection primarily resulting from the southwestern monsoon produce lightning more often than predictions in other regions of the United States. Carlaw et al. (2017) found the greatest frequencies of lightning in southwestern Arizona were concentrated in the regions with the highest terrain suggesting terrain-driven mesoscale circulations have an important role in the development of afternoon thunderstorms in the area. Future work comparing the predicted storm track characteristics in southwestern Arizona with the results from Carlaw et al. (2017) could help inform an improved understanding of observable

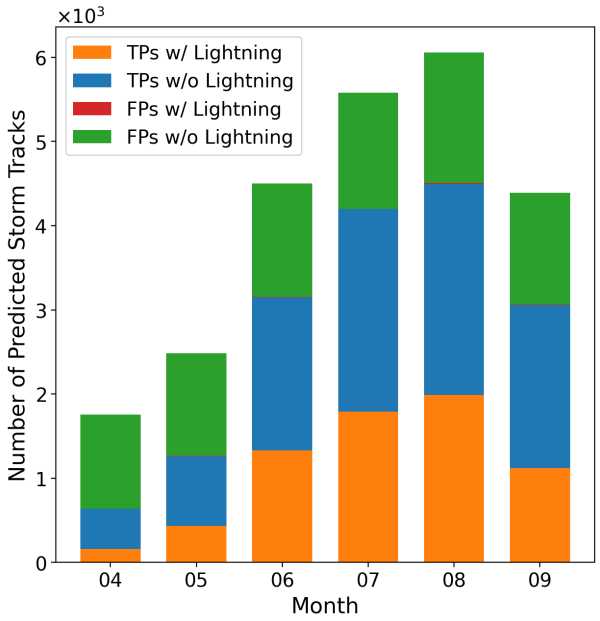

Figure 3.4: Temporal distribution of ThunderCast's predicted storm tracks. The tracks in each month are labeled true positive (TP) and false positive (FP) with (w/) and without (w/o) lightning.

convection processes necessary for electrification in the southwest. Comparison of such results to evaluations for other U.S. climate regions could also be useful to understand how the electrification processes in the southwest differ from the other climate regions, which could help inform improvements to thunderstorm prediction models.

During data collection for each predicted storm track, the maximum MRMS radar reflectivity at $-10°C$ is recorded. The maximum is taken from reflectivities between the start of the track and lightning initiation (both GLM and ENI are observed) or the end of the track for non-electrified cases. Fig. 3.6 presents the storm tracks' maximum radar reflectivity distribution in a histogram. Although the number of true positives without lightning decreases with higher reflectivity values, predicted storm tracks without lightning can reach high reflectivity values ($\geqslant 40$ dBZ). A distinct reflectivity threshold separating the tracks with and without

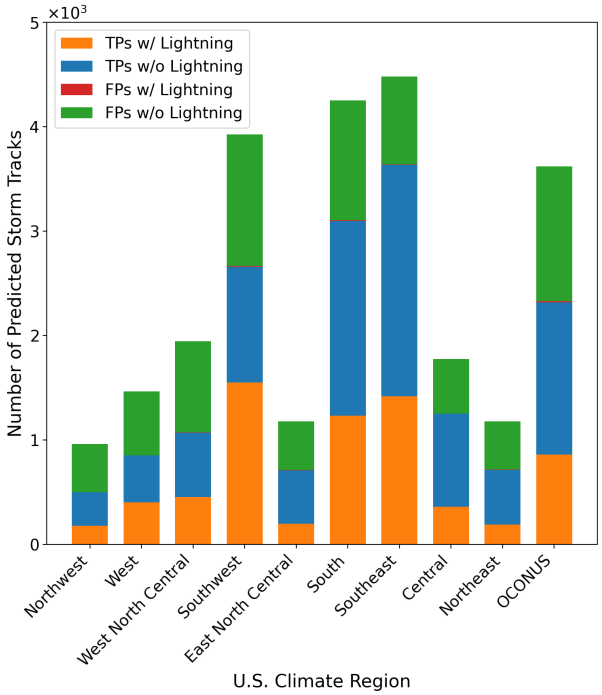

Figure 3.5: Spatial distribution of ThunderCast's predicted storm tracks. The tracks in each U.S. climate region are labeled true positive (TP) and false positive (FP) with (w/) and without (w/o) lightning.

lightning is not clearly discerned, demonstrating the limitations of using a radar reflectivity threshold as a definition of a thunderstorm. In the false positive cases, the lowest reflectivities recorded can contain stratiform cloud structures rather than convective structures. Also, ThunderCast sometimes predicts convection in areas associated with physical features like rivers or lakes due to their high contrast in the satellite bands used as predictors. ThunderCast's identification of these features could be improved by adding more samples of clear-sky cases over these physical features to the deep learning training dataset. Although shifting the deep learning model's target reflectivity to a higher value could filter out more stratiform convection cases, it will not completely filter out cases without lightning. Thus, the use of a "legacy-based" radar reflectivity threshold at the $-10°$C isotherm in the atmosphere is not adequate to define a thunderstorm when used solely. Additionally,

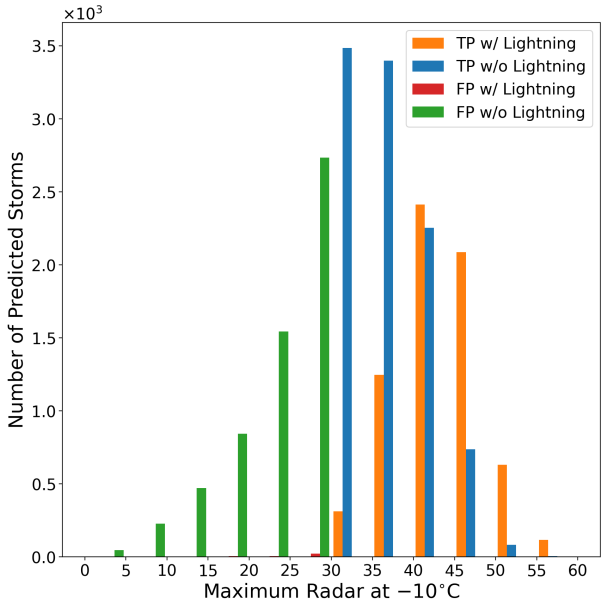

Figure 3.6: MRMS maximum radar at −10°C for ThunderCast's predicted storm tracks. The maximum radar reflectivity is measured up to whichever comes first: both GLM and ENI are observed or the end of the storm track. The tracks are labeled true positive (TP) and false positive (FP) with (w/) and without (w/o) lightning.

this indicates the microphysical observations and their corresponding reflectivity values in the organizing stage of storms from midlatitude field campaigns such as Dye and Martner (1982) do not necessarily represent midlatitude thunderstorms broadly because storms can form in a variety of environments and their growth is influenced by a variety of factors such as atmospheric moisture content, wind shear, and entrainment.

Rather than using a maximum radar reflectivity at an atmospheric temperature threshold to characterize thunderstorms, some studies have focused on evaluating the importance of areas or volumes of radar reflectivity for lightning intensity (Carey et al., 2019; Liu et al., 2012). Supporting this, a research study detailed in Liu et al. (2012) found a strong correlation between the area of 35 dBZ at −10°C to lightning flash rate for land-based storms. All of the precipitating storms in Liu et al.

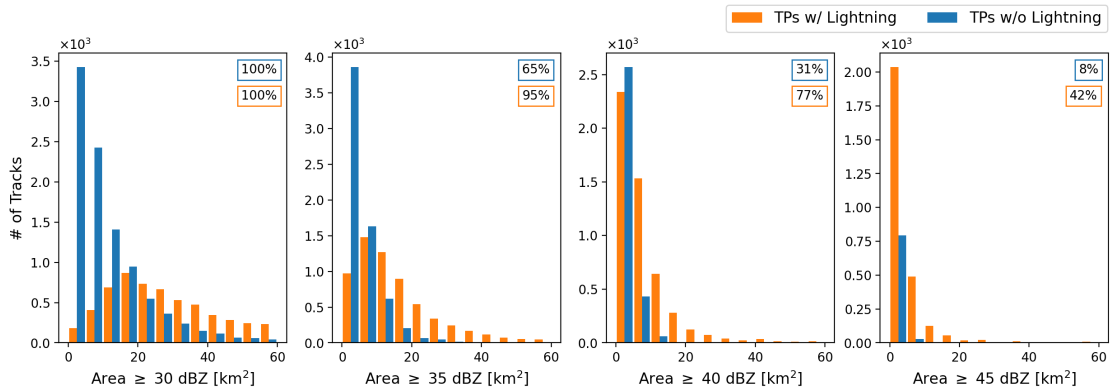

Figure 3.7: Predicted storm tracks' maximum area of pixels greater than or equal to an MRMS maximum radar at −10°C threshold during the true positive tracks' initiation stage. To obtain the maximum, areas are recorded until both ENI and GLM lightning are observed. The areas are given in km^2 because the dataset has 1-km resolution, and all areas included are non-zero. True positives (TPs) with (w/) and without (w/o) lightning are included in orange and blue, respectively. The percentages in the upper right hand corner of the plots indicate the percent of true positives with or without lightning with areas greater than or equal to the given thresholds. The colors of the boxes around the percents match the corresponding datasets shown in the histograms. Any predicted storms with areas greater than the values shown on the x-axis are omitted for ease of viewing, but they are included in the percentages provided.

(2012) contained at least one lightning flash, so the reflectivity area correlations found strictly apply to lightning producing storms. To get a better idea whether areas of radar reflectivity can be used to differentiate between electrified and non-electrified storms in a thunderstorm prediction model, Fig. 3.7 provides histograms of the true positive predicted storm tracks by area greater than or equal to four reflectivity thresholds: 30 dBZ, 35 dBZ, 40 dBZ, and 45 dBZ. It is important to note that not all true positive tracks contain reflectivities $\geqslant$ 35 dBZ, so the percent of the total datasets included in the histograms decreases with higher reflectivity. Consistent with Fig. 3.6, a higher percentage of the true positives with lightning reach reflectivities $\geqslant$ 35 dBZ, 40 dBZ, and 45 dBZ than those without lightning. The true positives without lightning are concentrated at lower areas $\geqslant$ 30 dBZ than

those with lightning. However, there is still a significant amount of true positives with lightning at the same areas as those without lightning. This is similar for areas $\geqslant$ 35 dBZ, 40 dBZ, and 45 dBZ, but the areas with the concentrations of each dataset are more aligned.

For further evaluation, the area histograms are broken up by U.S. climate region in Figs. 3.8 and 3.9. The areas with the highest concentrations of true positives with and without lightning are similar to Fig. 3.7 for all climate regions, indicating areas of radar reflectivity are unable to differentiate between true positive predicted storm tracks with and without lightning. One important difference between the regions is the percentages of true positives with and without lightning containing areas $\geqslant$ 35 dBZ, 40 dBZ, and 45 dBZ. Regions typically associated with higher humidities during the convective season have more predicted storm tracks with larger reflectivities than those without. For example, in the southeast climate region, 77% (99%) of the true positive predicted storm tracks without (with) lightning reach reflectivities $\geqslant$ 35 dBZ, 43% (89%) reach $\geqslant$ 40 dBZ, and 12% (53%) reach $\geqslant$ 45 dBZ. In the northwest only 44% (93%) of the true positive predicted storm tracks without (with) lightning reach reflectivities $\geqslant$ 35 dBZ, 13% (69%) reach $\geqslant$ 40 dBZ, and 2% (27%) reach $\geqslant$ 45 dBZ. This indicates the maximum radar reflectivity achievable by both true positives with and without lightning depend on the location, and corresponding environmental conditions, of convective initiation.

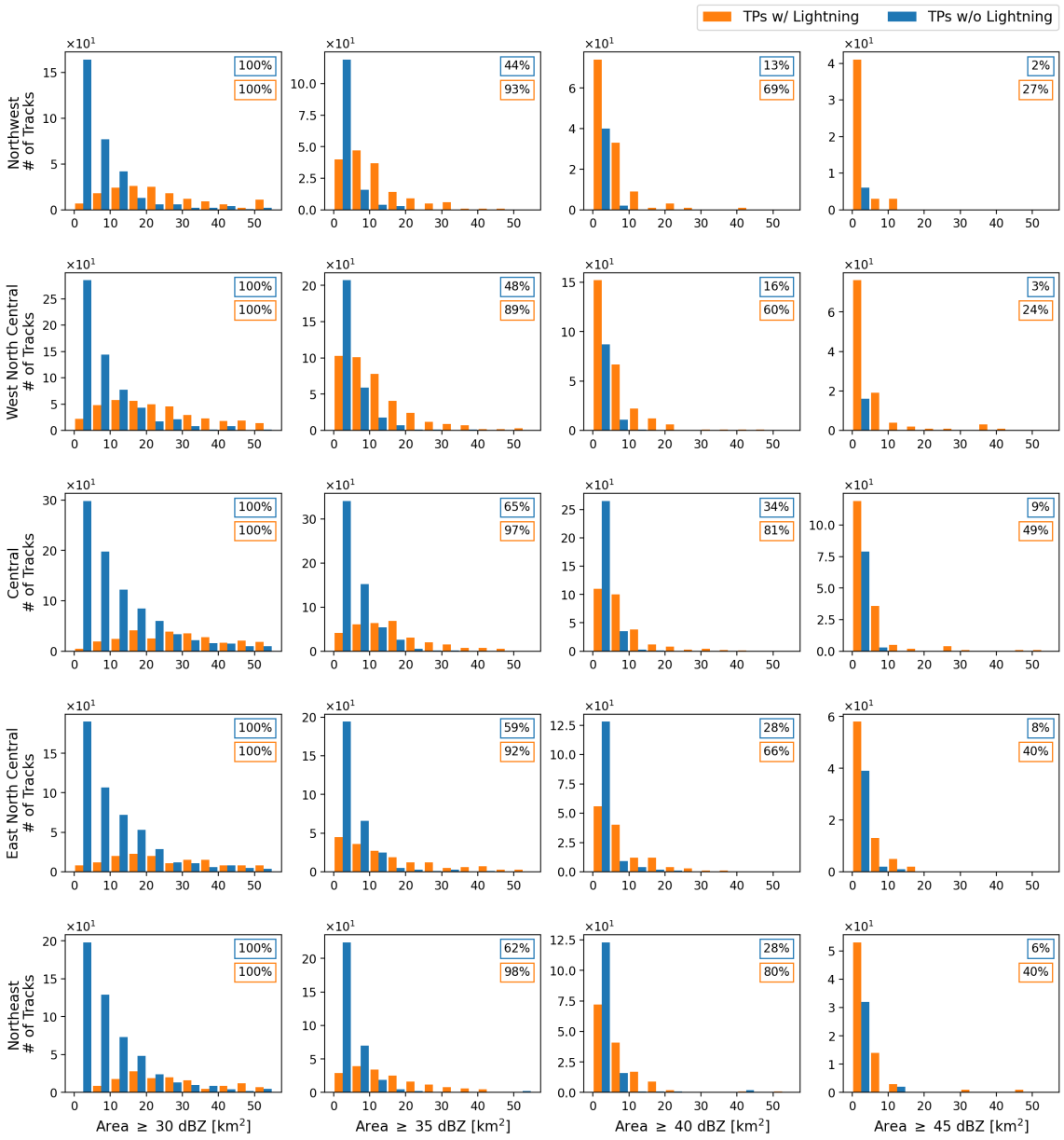

Figure 3.8: The maximum MRMS radar reflectivity area at −10°C sorted into northwest, west north central, central, east north central, and northeast U.S. climate regions. The plots are set-up similarly to Fig. 3.7 except each row contains predicted storm track datasets for the U.S. climate region indicated on the left-hand side of the row.

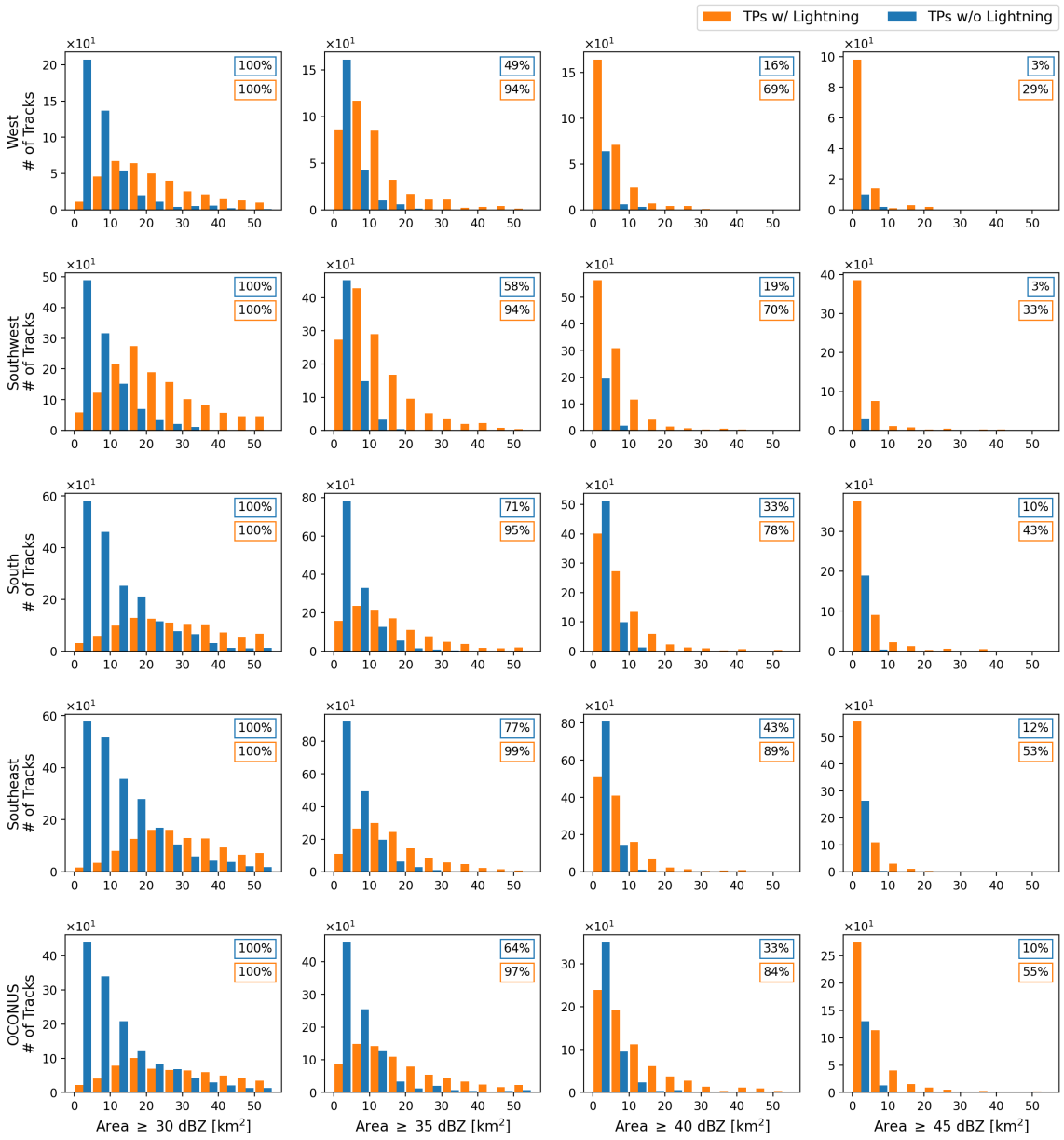

Figure 3.9: The maximum MRMS radar reflectivity area at −10°C sorted into west, southwest, south, southeast, and outside of the continental United States (OCONUS) U.S. climate regions. The plots are set-up similarly to Fig. 3.7 except each row contains predicted storm track datasets for the U.S. climate region indicated on the left-hand side of the row.

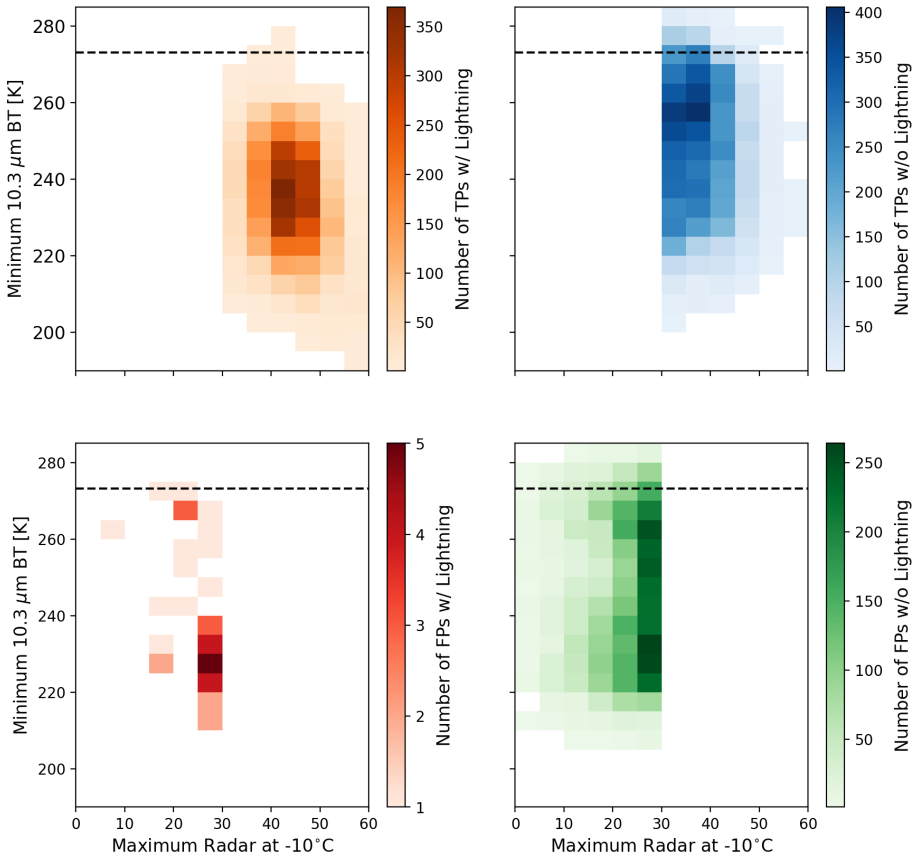

Figure 3.10: Two-dimensional histogram of the minimum 10.3 µm brightness temperature and the MRMS maximum radar reflectivity at −10°C for ThunderCast's predicted storm tracks. The maximum radar reflectivities and the brightness temperatures are measured up to whichever comes first: both GLM and ENI are observed or the end of the storm track. The tracks are labeled true positive (TP) and false positive (FP) with (w/) and without (w/o) lightning. The horizontal dashed line marks the freezing temperature of water.

The ABI spectral bands used as predictors in ThunderCast's deep learning architecture were selected because they form multispectral imagery commonly utilized by forecasters for diagnosing trends and patterns in cumuliform clouds such as cloud-top glaciation, cloud-top temperature, and cloud morphology. In order for the deep learning model to differentiate between electrified and non-electrified

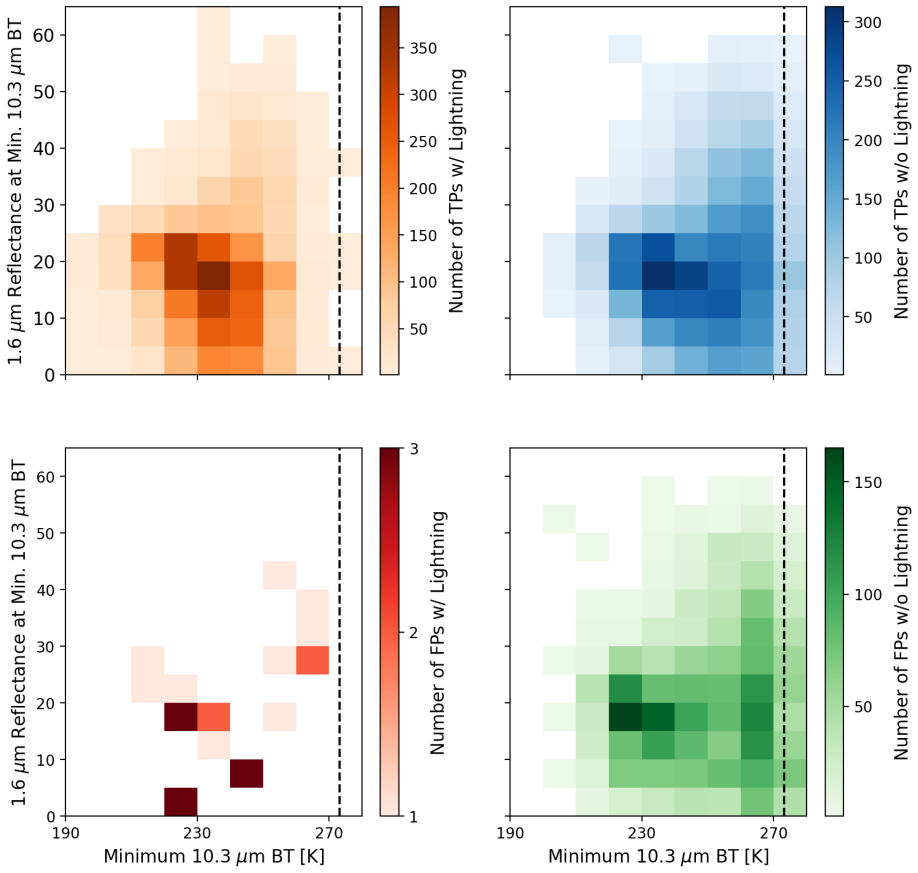

Figure 3.11: Two-dimensional histogram of the minimum 10.3 μm brightness temperature and the 1.6 μm reflectance at the minimum brightness temperature for ThunderCast's predicted storm tracks. The values are measured up to whichever comes first: both GLM and ENI are observed or the end of the storm track. The tracks are labeled true positive (TP) and false positive (FP) with (w/) and without (w/o) lightning. Because the 1.6 μm satellite band is only available during the day, the storm tracks are daytime only ($\leqslant 85°$ solar zenith angle). The vertical dashed line marks the freezing temperature of water.

storms, patterns in the predictors must be distinguishable and relatable to the target dataset (radar reflectivity at $-10°$C). To identify relationships between storm tracks' radar reflectivities and the ABI bands, Figs. 3.10 and 3.11 present two-dimensional histograms of maximum radar reflectivity at $-10°$C with cloud-top brightness tem-

perature from the 10.3 μm band and the daytime only 1.6 μm reflectance (snow/ice band), respectively. In Fig. 3.10, most of the predicted storm tracks have brightness temperatures below freezing indicating most cases, regardless of the presence of lightning or radar reflectivity, achieve cloud-top glaciation and/or contain supercooled water at cloud-top. The bins with the highest concentrations (darker colors) in the histograms shown in Fig. 3.10 generally cover similar regimes between true positive and false positives with and without lightning. Although the brightness temperature mean of the true positives with lightning appears colder than the other categories, there are still high concentrations of predicted storm tracks at the same temperatures. The predicted storm tracks' 1.6 μm reflectances at their minimum 10.3 μm brightness temperatures are shown in Fig. 3.11. Most of the predicted storms have low 1.6 μm reflectance values. Snow and ice surfaces are strongly absorbing at 1.6 μm, so these values indicate the predicted storms likely contain ice at cloud-top, which is considered to be necessary for thunderstorm development. The predicted storms warmer than freezing in Figs. 3.10 and 3.11 tend to be small cumuliform structures discernible with higher resolution bands such as the 0.64 μm band. The relatively low resolution of the 10.3 μm band can cause a small feature's cloud-top to blend with the surrounding warmer area, making the 10.3 μm brightness temperatures appear larger than the actual cloud-top temperature. The similarities between true positive and false positive categories with and without lightning indicate the predicted storm tracks likely appear similar visually in multispectral imagery, especially since the values shown in Figs. 3.10 and 3.11 indicate the presence of glaciation and ice at cloud-top. Considering this, additional predictors from additional data sources may be necessary to reduce the number of ThunderCast's true positives without lightning.

One of the benefits of an object-based evaluation technique is it allows for a bulk analysis of lead times for predicted storm events. Fig. 3.12 and Fig. 3.13 show ThunderCast's median and mean lead times, respectively, to 30 dBZ, GLM lightning, and ENI lightning calculated at various probability thresholds with this object-based technique. Because small probabilities often appear earlier than larger probabilities, they have the highest lead time values in Figs. 3.12 and 3.13. Section

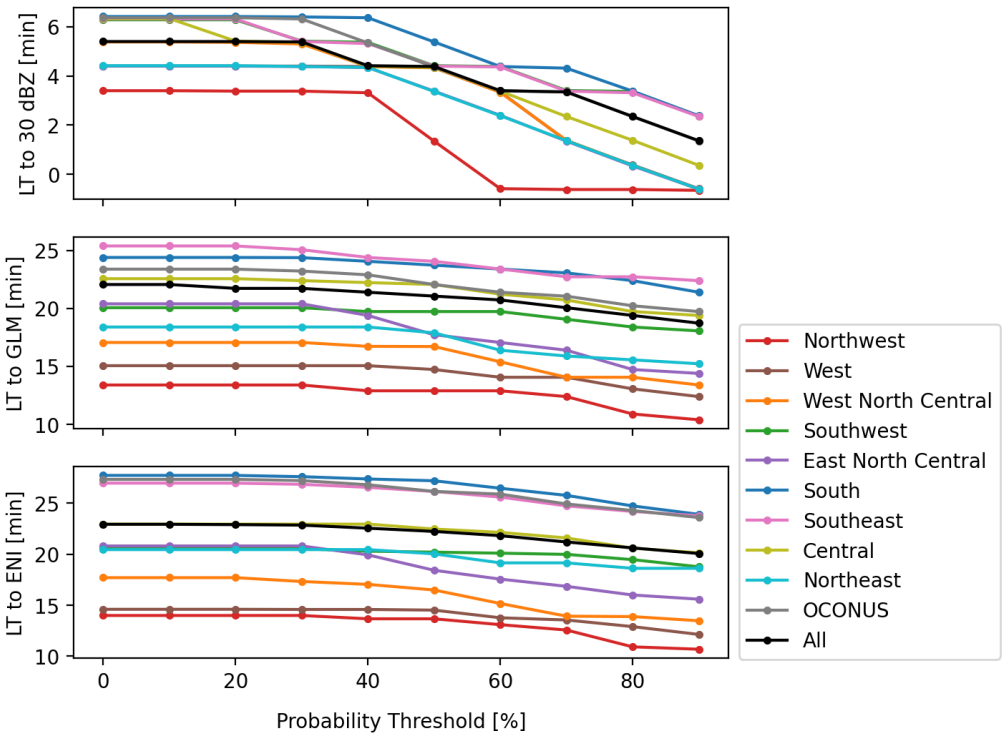

Figure 3.12: Median ThunderCast lead time (LT) to the first observation of a radar reflectivity echo $\geqslant$ 30 dBZ at -10°C, lightning from the Geostationary Lightning Mapper (GLM), and lightning from the total lightning product from Earth Networks, Inc. (ENI). The lead times are calculated from the first occurrence of a probability $\geqslant$ to the value [%] indicated except for the 0% probability threshold where lead time is calculated from the first occurrence of any probability $> 0\%$. The black line incorporates all predicted storm tracks in the dataset, while the other colored lines break the dataset into climate regions.

2.3 indicated ThunderCast's highest critical success index occurred at approximately 20% probability. The median lead times at 20% for all climate regions (black line in Fig. 3.12) are 5.4 min to 30 dBZ, 21.7 min to GLM lightning, and 22.9 min to ENI lightning. The mean lead times at 20% for all climate regions (black line in Fig. 3.13) are 11.8 min to 30 dBZ, 30.9 min to GLM lightning, and 31.7 min to ENI lightning. The southeast, south, and OCONUS regions consistently have some of the longest

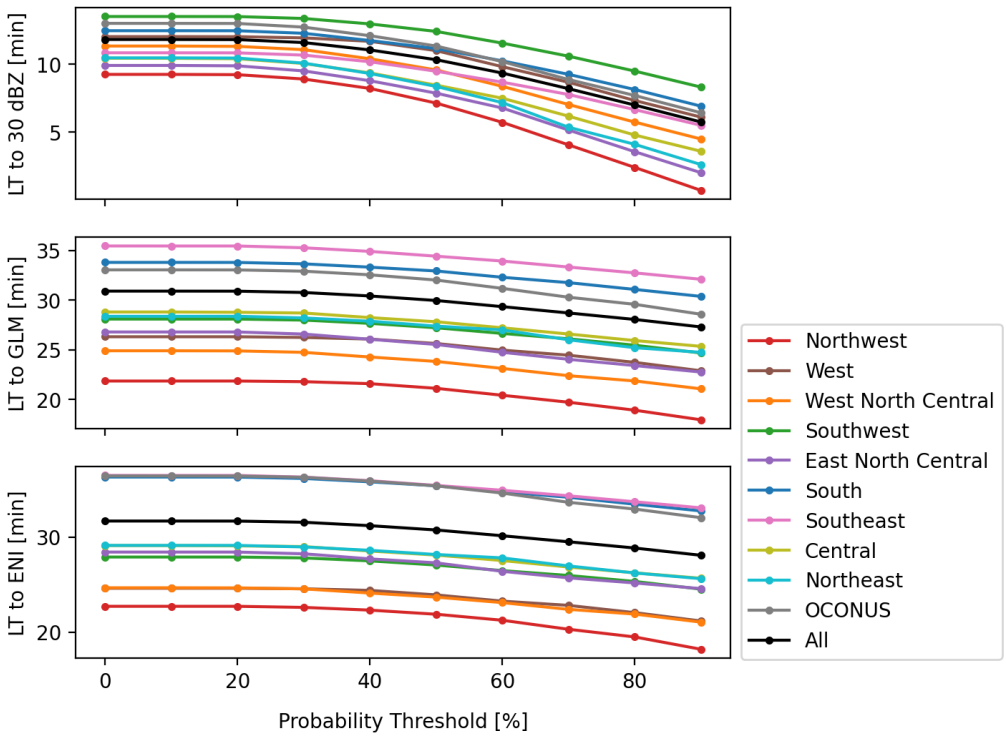

Figure 3.13: Mean ThunderCast lead time (LT) to the first observation of a radar reflectivity echo $\geqslant$ 30 dBZ at -10°C, lightning from the Geostationary Lightning Mapper (GLM), and lightning from the total lightning product from Earth Networks, Inc. (ENI). The lead times are calculated from the first occurrence of a probability $\geqslant$ to the value [%] indicated except for the 0% probability threshold where lead time is calculated from the first occurrence of any probability $> 0\%$. The black line incorporates all predicted storm tracks in the dataset, while the other colored lines break the dataset into climate regions.

lead times while the northwest has the shortest. A variety of factors can contribute to lead time differences including sampling during deep learning model training, the variation in atmospheric conditions leading to and sustaining convection in different regions, and the presence of terrain-based forcing mechanisms.

Because 30 dBZ is considered a threshold of convective initiation for ThunderCast, the lead time to 30 dBZ can be compared to those obtained from other

convective initiation models. When comparing lead times, it is important to keep in mind the method for obtaining the values could influence their reliability. Many machine learning models do not present bulk statistical analyses of lead time, and, instead, use case studies to demonstrate the model's lead time capabilities. Case studies are useful for model interpretation, but it is not clear how well their values generalize to a broader range of situations, hence, this paper presents an object-based evaluation technique for large-scale analyses of lead time. One convective initiation model with a large-scale lead time analysis applicable to the CONUS is SATCASTv2, an object-based 0-2-h nowcasting algorithm (Walker et al., 2012). Walker et al. (2012) reports the following lead times to convective initiation for SATCASTv2: 23.8 min. in Melbourne, FL, 32.7 min in Memphis, TN, 32.9 min in the central United States, and 27.4 min in the northeastern United States. These values are all larger than ThunderCast's average lead times to 30 dBZ at 20% probability in Fig. 3.13, which range from 9.2-13.5 min across the United States. The greater lead times observed in SATCASTv2 could indicate incorporating thunderstorm objects into the deep learning target dataset may result in larger lead times. Also, Mecikalski et al. (2015)'s random forest method improved on SATCASTv2 along with their incorporation of parameters from numerical weather prediction models. Perhaps numerical weather model parameters or other environmental observations could improve the performance of ThunderCast.

There are other machine learning models designed specifically to predict lightning occurrence instead of convective initiation in the CONUS such as ProbSevere LightningCast (Cintineo et al., 2022). Similarly to ThunderCast, LightningCast uses a U-Net architecture and the same ABI predictors, but it uses GLM lightning observations as the target instead of radar reflectivity at -10°C. One of the reasons ThunderCast used radar instead of lightning observations as a target was to increase lead time to thunderstorm hazards like lightning by using a predictor that occurs prior to lightning observations. This idea is consistent with the findings in Fig. 3.6, Table 3.2, and Fig. 3.12. Lightning typically occurs after 30 dBZ as shown in Fig. 3.6 and Table 3.2 where less than 1% of the predicted storm tracks with lightning corresponded to reflectivities $<$ 30 dBZ. Also, Cintineo et al. (2022)

reported LightningCast's median lead time as 17.5 min at LightningCast's 30 and 40% probabilities while ThunderCast's lead time to GLM is about 21.7 min at these probabilities. ThunderCast achieves approximately 4 min of additional lead time to GLM lightning than LightningCast when applied to electrified storm cases. Although models using a radar reflectivity threshold target can achieve greater lead times to lightning hazards than models with lightning observations as a target, it is important to keep in mind the limitations associated with such a model like the tendency to make predictions for non-electrified convection.

3.4 Summary

ThunderCast is a deep learning model designed to predict the occurrence of convection associated with thunderstorms in the next 0-60 minutes from a radar reflectivity perspective. To determine the extent of true positive and false positive predictions with and without lightning in ThunderCast, an established tool for the Tracking and Object-Based Analysis of Clouds (tobac) was applied to ThunderCast's probabilistic output for 7 randomly selected days in each month of 2022's convective season (April to September). Tobac's flexible software framework detected features in ThunderCast, used a watershed technique to obtain regions of related probabilities surrounding the features (objects), tracked the features through time, and identified connected tracks with a merge/split algorithm. A series of post-processing steps simplified the resulting dataset to avoid non-persistent predictions, incomplete tracks, predictions associated with existing convection, and complex merge/split cases. Each track's date, time, and location were recorded along with Multi-Radar Multi-Sensor (MRMS) radar reflectivity at -10°C, Geostationary Lightning Mapper (GLM) lightning, Earth Networks, Inc. (ENI) total lightning, Advanced Baseline Imager (ABI) brightness temperature at 10.3 μm, and ABI reflectance observations at 1.6 μm.

Each predicted storm track was sorted into true positive and false positive categories with and without lightning based on the data collected for each track. Any predicted storm track with a MRMS radar reflectivity at $-10^{\circ}\text{C} \geqslant 30$ dBZ was

categorized as a true positive due to the definition of convection used in ThunderCast's development. Out of the 24,770 land-based tracks with radar coverage, 67.8% were true positive and 59.3% of the true positive cases were not associated with GLM or ENI lightning. Because less than half of the predicted storms achieved electrification, ThunderCast is unable to differentiate between convective structures with and without lightning activity. Although the amount of true positives without lightning decreases with higher reflectivity values at $-10°C$, they can reach high reflectivity values ($\geqslant 40$ dBZ). This indicates a radar-based definition of convection alone is inadequate for isolating convection associated with thunderstorms. Additionally, because the "legacy-based" radar reflectivity definition of a thunderstorm came from microphysical and radar observations during the organizing stage of thunderstorms from early field campaigns in the midlatitudes, case studies from field campaigns, like those in Dye and Martner (1982), do not necessarily represent midlatitude thunderstorms broadly. The maximum reflectivity achievable by a predicted storm depends on the location of and the environmental conditions present for convective initiation.

ABI 10.3 μm brightness temperatures were compared to the maximum radar reflectivity at $-10°C$ and the corresponding 1.6 μm reflectance for each predicted storm track. Most of the predicted storm tracks, regardless of their true positive or false positive with or without lightning categorization, had cloud-top brightness temperatures less than freezing and, for the daytime cases, low corresponding 1.6 μm reflectances. These observations are consistent with the presence of ice at cloud-top and cloud-top glaciation, so the predicted storm tracks likely appear similar visually in multispectral imagery. Since the predictors appear similar for the cases the model is supposed to distinguish between, additional predictors from additional data sources may be necessary for ThunderCast to reduce the number of false positives and true positives without lightning.

Existing statistical analyses of deep learning models focus on pixel-by-pixel instead of case-by-case evaluations, which make automated, large-scale collection of lead times for storm events difficult. Lead times are useful to understand the benefits of a model for forecasting meteorological phenomena. The object-based

approach presented in this paper allows for a bulk analysis of lead-time for predicted thunderstorm events and can be applied to other deep learning models. ThunderCast's median lead times at 20% probability for all CONUS climate regions were 5.4 min to 30 dBZ, 21.7 min to GLM lightning, and 22.9 min to ENI lightning. The mean lead times at 20% probability for all CONUS climate regions were 11.8 min to 30 dBZ, 30.9 min to GLM lightning, and 31.7 min to ENI lightning. The lead times to 30 dBZ were shorter than Walker et al. (2012)'s model for convective initiation nowcasting, but the lead time to lightning was longer than LightningCast (Cintineo et al., 2022), a prominent deep learning lightning nowcasting model with a similar U-Net architecture and the same ABI predictors as ThunderCast. Using a radar-based target for convection associated with thunderstorms can achieve greater lead times than a lightning target. However, not all predictions from a radar-based model will produce lightning.

Further analysis is needed to determine how additional, scientifically informed predictors and/or a different target dataset can improve thunderstorm predictions by reducing false positive predictions or true positive predictions without lightning. An alternative target dataset to evaluate could be based on the echo-class algorithm used by Lee et al. (2021). For the predictors, perhaps numerical weather prediction model parameters, atmospheric moisture at various levels in the atmosphere, or information pertaining to cloud entrainment properties could provide the additional environmental context needed to improve thunderstorm predictions. Additional evaluation of predicted storm tracks' temporal and spatial characteristics could inform an improved scientific understanding of midlatitude convection and electrification, which may better inform what predictors or target would be optimal.

4 THE LIMITATIONS OF USING SATELLITES TO IDENTIFY CONVECTIVE INITIATION

4.1 Background

The Geostationary Operational Environmental Satellites (GOES; 16, 17, or 18) Advanced Baseline Imager (ABI) Day Cloud Phase Distinction (DCPD) red-green-blue (RGB) false color composite images, detailed in Elsenheimer and Gravelle (2019), are commonly utilized by forecasters in the United States for diagnosing trends and patterns in cumuliform clouds such as cloud-top glaciation, cloud-top temperature, and cloud morphology. The red component of the imagery uses the inverse of the 10.3 μm ABI band, known as the clean longwave infrared window band, to identify surface and cloud-top brightness temperatures. A small (large) red component in the imagery indicates warm (cold) brightness temperatures, characteristic of low (high) clouds and warm (cold) surface features. The green component uses the 0.64 μm visible ABI band to distinguish between highly reflective surfaces (large green component) like clouds and snow with lower reflectance (small green component) features such as water and vegetation. Particle phase is distinguishable by the blue component comprised of the 1.6 μm near-infrared ABI band, commonly known as the snow/ice band. Because ice clouds absorb more radiation at 1.6 μm than liquid water clouds, large (small) blue components in the RGB indicate highly (poorly) reflective liquid (ice) clouds (Elsenheimer and Gravelle, 2019; Connell et al., n.d.).

At the beginning of a thunderstorm's lifecycle, the cumulus cloud appears light blue or cyan in the DCPD RGB. As the cumulus cloud grows vertically into agitated cumulus and towering cumulus, the top of the cloud changes from liquid to mixed phase (with supercooled water and ice) to predominately ice crystals, a process known as cloud-top glaciation (Houze Jr., 2014). In the DCPD RGB, this change in phase is represented by the color of the cloud changing to green-yellow because the blue component decreases as the ice increases at the cloud-top (Elsenheimer

and Gravelle, 2019). This stage of a storm's development is important for lightning formation because the non-inductive theory for charge separation in thunderstorms relies on the presence of ice, graupel, and supercooled water in the cloud, as well as the particles' collisions in turbulent air and their separation by the cloud's updraft (Baker et al., 1987; Saunders, 1993; Baker and Dash, 1994; Dash et al., 2001; Saunders et al., 2006). As the storm continues to grow vertically to the upper levels of the troposphere and transitions to a cumulonimbus cloud, the cloud-top will begin to appear orange in the DCPD RGB and overshooting tops can be clearly discerned.

As the DCPD's name implies, the color transitions in the DCPD during thunderstorm development are only applicable during the day. At night, an alternative false color RGB image called the Nighttime Microphysics (NtMicro) RGB can be used to identify cloud types in the mid and upper atmosphere (Connell et al., n.d.). The red, green, and blue components of this RGB are comprised of the difference between ABI 12.3 μm and 10.3 μm bands, the difference between ABI 10.3 μm and 3.9 μm bands, and the 10.3 μm band, respectively. Large (small) contributions of the red component in the NtMicro RGB indicate thick (thin) clouds because the 12.3 μm and 10.3 μm band difference physically relates to optical depth. Because the difference between the 10.3 μm and 3.9 μm bands relates to particle phase and size, small (large) contributions of green component correspond to ice (small water) particles or surface features. Large (small) blue contributions from the 10.3 μm band indicate warm (cold) surface features. Thick, developing cumulus clouds appear dark red in the NtMicro RGB and can contain some yellow when the clouds are very cold. The visual color transition of cumuliform clouds in the thunderstorm lifecycle contribute to both the DCPD and NtMicro RGBs' usefulness in forecasting thunderstorms and their associated hazards.

Because the convective lifecycle is well captured visually by satellite spectral bands and satellite data is routinely available in areas where data from ground-based precipitation detecting radars are unavailable or are of reduced quality due to beam blocking from mountainous terrain, convective storm (thunderstorm) nowcasting models use spectral bands to identify and predict convection. For example, varied combinations of satellite channels from the visible, shortwave infrared,

and/or longwave infrared bands, including those with wavelengths matching the DCPD RGB, were used as predictors in Walker et al. (2012), Mecikalski et al. (2015), Lagerquist et al. (2021), Ortland et al. (2023), and Fan et al. (2024). In addition to relying on satellite data, all of the mentioned models were trained and/or validated with Multi-Radar Multi-Sensor (MRMS) data, where MRMS radar reflectivities $\geqslant$ a threshold (30 dBZ or 35 dBZ) were considered positive for thunderstorm activity. However, models involving the prediction of convective initiation from a satellite and radar perspective tend to have high false alarm ratios (McGovern et al., 2023). For the models mentioned, the false alarm ratios ranged from 22% to 60%.

To determine the extent of false alarms in Ortland et al. (2023)'s model, called the Thunderstorm Nowcasting Tool (ThunderCast), Chapter 3 applied object-tracking software to track regions of ThunderCast predictions. The predicted storm tracks were then labeled as either true positive ($\geqslant$ 30 dBZ at -10°C per ThunderCast's definition of a thunderstorm in Ortland et al. (2023)) or false positive ($<$ 30 dBZ at -10°C). The tracks were further classified by the presence (or lack thereof) of lightning from either the Geostationary Lightning Mapper (GLM) or Earth Networks, Inc.'s (ENI's) total lightning product. Over half of the true positive tracks identified in Chapter 3 were not associated with GLM or ENI lightning, and a radar reflectivity threshold cleanly separating tracks with and without lightning was not found. This indicates a purely radar-based definition of convection associated with thunderstorms is inherently flawed because large radar signatures are not always indicative of thunderstorm activity, and false alarm rates in such models may be higher than originally estimated.

Additionally, Chapter 3 compared the maximum radar reflectivities at -10°C for each predicted storm tracks to the tracks' ABI 10.3 μm brightness temperatures and 1.6 μm reflectances. Regardless of their true positive or false positive with or without lightning classifications, most of the predicted storm tracks had minimum 10.3 μm brightness temperatures below freezing and low corresponding 1.6 μm reflectances in the daytime cases. These results are consistent with the presence of ice at cloud-top, and indicate the predicted storm tracks likely appear similar visually in multispectral imagery like the DCPD RGB. In order for computer vision

artificial intelligence methods, like those presented in Mecikalski et al. (2015), Lagerquist et al. (2021), Ortland et al. (2023), and Fan et al. (2024), to differentiate between electrified and non-electrified storms, patterns in the predictors must be distinguishable. The similarities between the electrified and non-electrified storms from the ABI perspective during convective initiation could make it difficult to identify unique patterns at cloud-top and could contribute to high false alarms.

Following the results found in Chapter 3, this chapter presents four case studies where ABI cloud-top signatures in DCPD RGBs (or NtMicro RGB imagery in cases at night) appear consistent with thunderstorm development during convective initiation but do not end up producing lightning. The cases were chosen to demonstrate breadth in meteorological conditions, formation environments, and locations during the midlatitude convective season. For each case, the meteorological conditions, according to National Weather Service forecast discussions, leading up to and during the convective initiation event are discussed. Additionally, the satellite imagery, MRMS radar reflectivity at $-10°C$, lightning observations (ENI and GLM when available), and model predictions from ThunderCast are provided.

4.2 Case Studies

May 15, 2022: Severe Weather Day in Oklahoma

Conditions were favorable for thunderstorm development in northeastern Oklahoma (OK) and northwestern Arkansas during the afternoon and evening hours on May 15, 2022. According to National Weather Service Tulsa Office (2022), the convective available potential energy (CAPE) reached extreme values (4000-5000 J kg^{-1}) and convective inhibition (CIN or sometimes referred to as the strength of the cap) was weakening. Developing flow aloft contributed to deep level wind shear (50-60 kts), sufficient for sustaining supercells (National Weather Service Tulsa Office, 2022). Near surface relative humidity ranged between 62% and 66% between 21:00 UTC and 23:00 UTC at a nearby Mesonet site in Pryor, OK (Oklahoma Climatological Survey, 2022). These conditions resulted in rapid thunderstorm

development and many severe hail and wind reports were recorded (National Weather Service, 2022a). The time series of images in Figs. 4.1 and 4.2 follow the development of two storm cells (near the center of the images) in this primed thunderstorm environment to the east of Tulsa, OK. The first column in Figs. 4.1 and 4.2 display the Day Cloud Phase Distinction (DCPD) red-green-blue (RGB) images, discussed previously in Section 4.1, for eight time steps in the late afternoon.

In Fig. 4.1, the DCPD RGB shows glaciating cumulus clouds (light green in color) concentrated around the two developing cumulus towers. The towering cumulus clouds appear oranger than the surrounding cumuli, indicating vertical growth, cold cloud-tops, and the presence of ice. The bright orange colored cloud-tops are shown distinctly for both the left-most and right-most storms in the DCPD RGBs in Fig. 4.1. Fig. 4.2 focuses on the right-most storm from Fig. 4.1 during it's most active initiation stages, after the left-most storm has dissipated. Both towering cumulus clouds start growing at similar times in environments exhibiting key meteorological ingredients for thunderstorm development (high CAPE, low CIN, and large wind shear). However, the left-most tower quickly starts to dissipate at 22:11 UTC (about 50 min after glaciating) before producing lightning while the tower on the right continues to grow into a mature thunderstorm with a distinct overshooting top and observed lightning (at about 23:01 UTC, shown in Fig. 4.2). Both towering cumulus structures were identified by ThunderCast as having a high probability of developing into a thunderstorm and both contained radar reflectivities $\geqslant 30$ dBZ at -10°C, supporting Ortland and Pavolonis (2024)'s finding that artificial intelligence models based on satellite and radar observations are unable to differentiate between electrified and non-electrified storms. Furthermore, key meteorological quantities for general areas used in forecasting like CAPE, CIN, and wind shear are not sufficient, in some cases, to determine which cumulus towers will initiate into thunderstorms because both cumulus towers in this case study formed in similar environments. As a result, this case study demonstrates the complex nature of thunderstorm development. More work is needed to determine the causes of premature storm death and what datasets can be used to identify those cases with current or more complex satellite imagery pattern recognition techniques.

Such work is critical for understanding thunderstorm physical processes and what datasets can be used in convective initiation models to reduce false alarms.

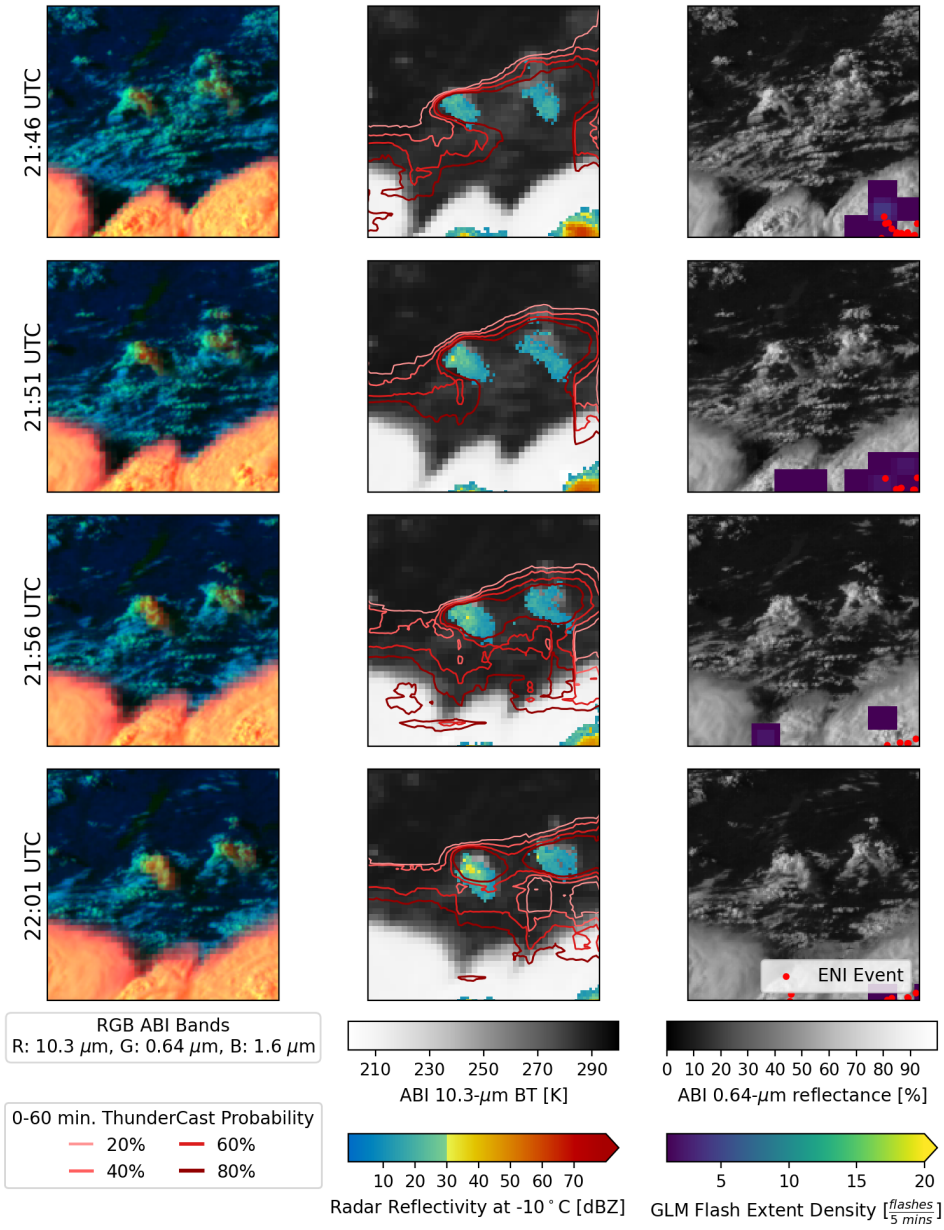

Figure 4.1: A time series of GOES-16 ABI images with corresponding radar, lightning, and model (ThunderCast) data for the times closest to the ABI time (written to the left of each row) collected from 21:46 to 22:01 UTC on 2022-05-15. The first column contains GOES-16 DCPD false color RGB images. The second column contains output from ThunderCast, MRMS radar reflectivity at $-10°$C, and GOES-16 ABI 10.3 µm brightness temperatures. The last column contains ENI total lightning events (points), GOES-16 GLM flash extent density, and GOES-16 ABI 0.64 µm reflectances. Each image is 80 km $\times$ 80 km and the latitude and longitude coordinates of the lower lefthand corners are $35.7°$ and $-96.1°$, respectively.

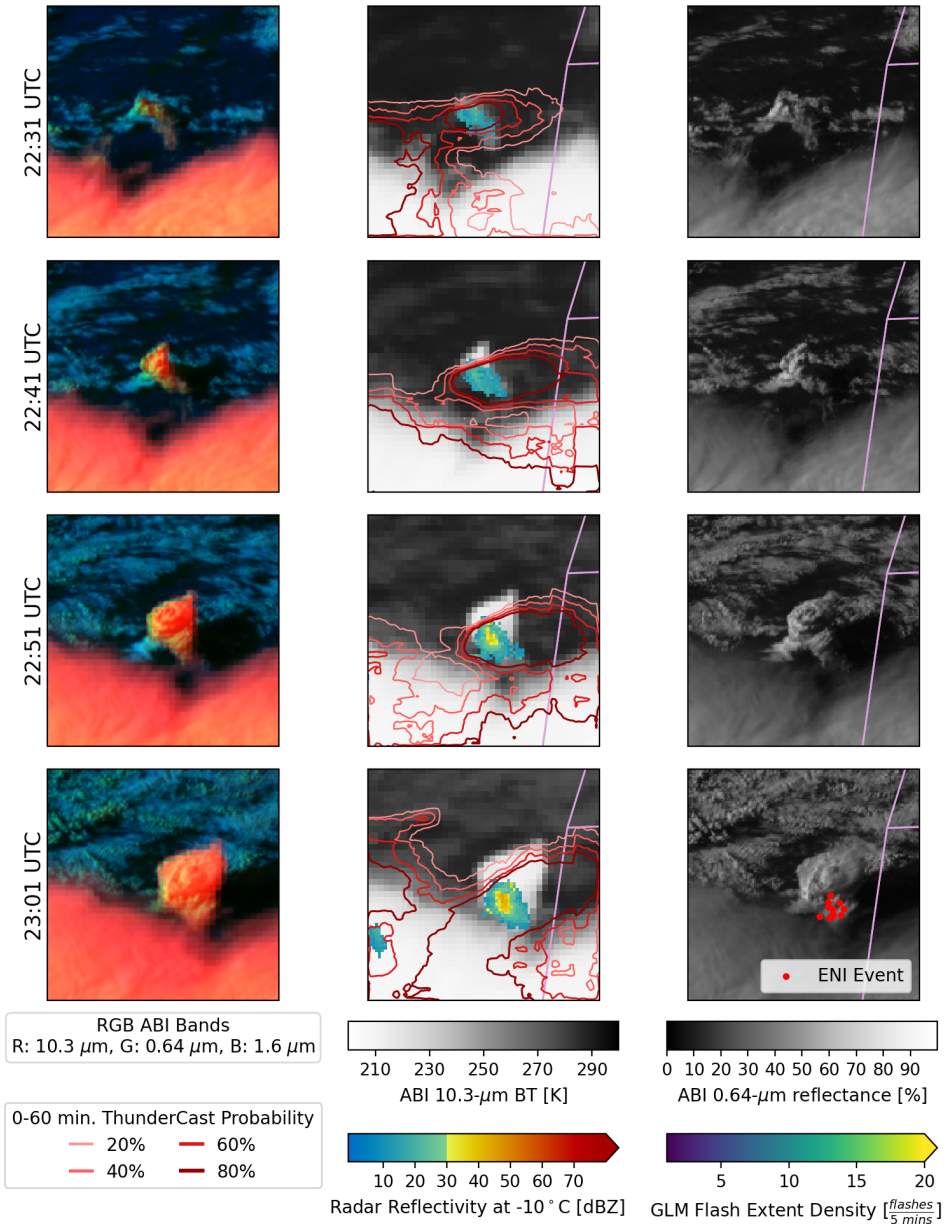

Figure 4.2: A time series of GOES-16 ABI images with corresponding MRMS radar data, model (ThunderCast) predictions, and GLM and ENI lightning data for the times closest to the ABI time (written to the left of each row) collected from 21:51 to 22:01 UTC on 2022-05-15. The plots are set-up in the same manor as Fig. 4.1 except the latitude and longitude coordinates of the lower lefthand corners are 35.7° and −95.6°, respectively.

July 24, 2022: Sea Breeze Convection along Texas' Gulf Coast

Differential heating between the Gulf of Mexico and the Texas coast can fuel new convection as cooler sea air is driven inland, forcing the warmer air over the land to rise (Scofield and Purdom, 1986). The resulting sea breeze convection can rapidly develop into thunderstorms. On July 24, 2022, the mid-morning wind surface conditions, shown in Fig. 4.6, were favorable for sea breeze convection development since the wind direction along the coast indicated landward air motion. Local National Weather Service forecasters described a potential for showers and thunderstorm development along the sea breeze during the morning and early afternoon (National Weather Service Houston/Galveston Office, 2022). Consistent with the forecast, thunderstorms did develop along the Texas coast, such as the case shown in Fig. 4.4. However, some of the cumuli initiating along Texas' Gulf Coast, with similar characteristics to the initiating stage of sea breeze thunderstorms nearby, did not produce lightning. Fig. 4.5 shows one such case for four time steps starting at 15:31 UTC (10:31am CDT).

In Figs. 4.4 and 4.5, similar to the Oklahoma case study, the green and yellow colors visible in the DCPD RGB images indicate the presence of ice at cloud-top, the radar reflectivities at -10°C reach at least 30 dBZ, and the vertical cumulus towers are maintained for a similar amount of time (approximately 45 min from glaciating to dissipating). However, the meteorological conditions in the coastal environment where this case study takes places is much different that of Oklahoma's in the first case study. This demonstrates patterns associated with convective storms in satellite imagery and radar can fail to distinguish between electrified and non-electrified cases in varied meteorological environments.

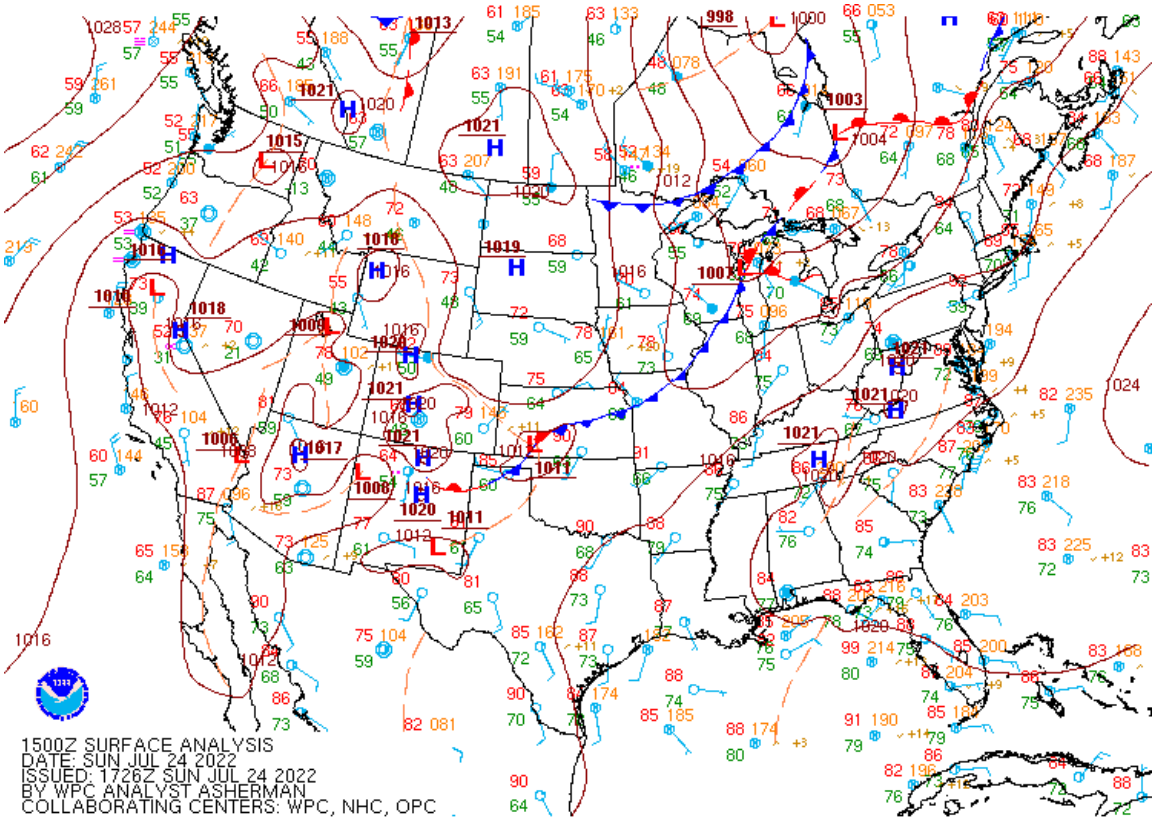

Figure 4.3: Meteorological surface conditions for Sunday, July 24, 2022 at 15:00 UTC. This image was produced by the National Oceanic and Atmospheric Administration (NOAA) and is available to the public (National Weather Service, 2022b). Each standard weather station plot in the image contains information for wind direction, wind strength, sky cover, temperature, dew point, pressure, and pressure trend. Surface pressure also provided by the dark red contours, and frontal boundaries are marked in red and blue.

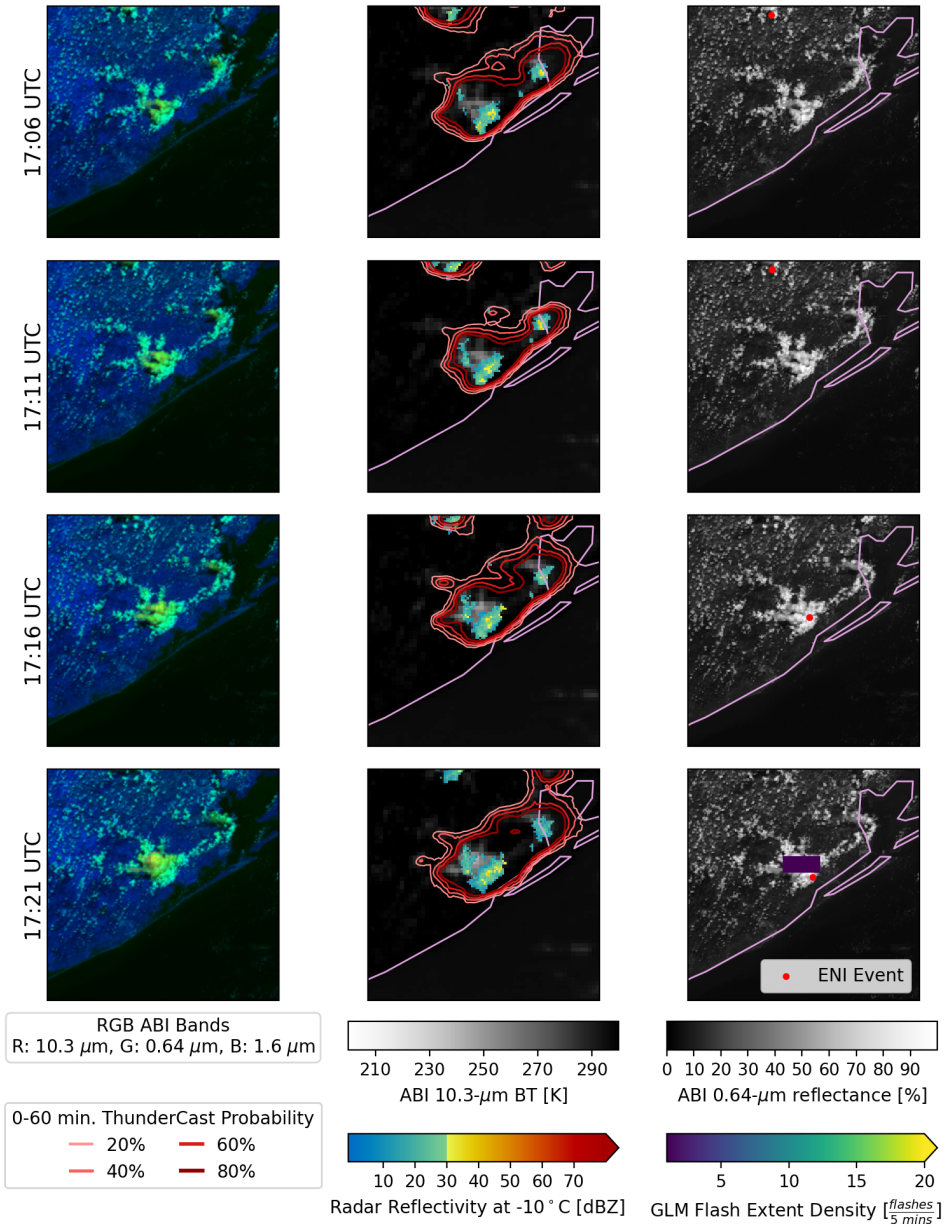

Figure 4.4: A time series of GOES-16 ABI images with corresponding MRMS radar data, model (ThunderCast) predictions, and GLM and ENI lightning data for the times closest to the ABI time (written to the left of each row) collected from 17:06 to 17:21 UTC on 2022-07-24. The plots are set-up in the same manor as Fig. 4.1. In the second and third columns, light purple lines display the Texas coastline. Each image is 112 km $\times$ 112 km and the latitude and longitude coordinates of the lower lefthand corners are 28.6° and $-96.0°$, respectively.

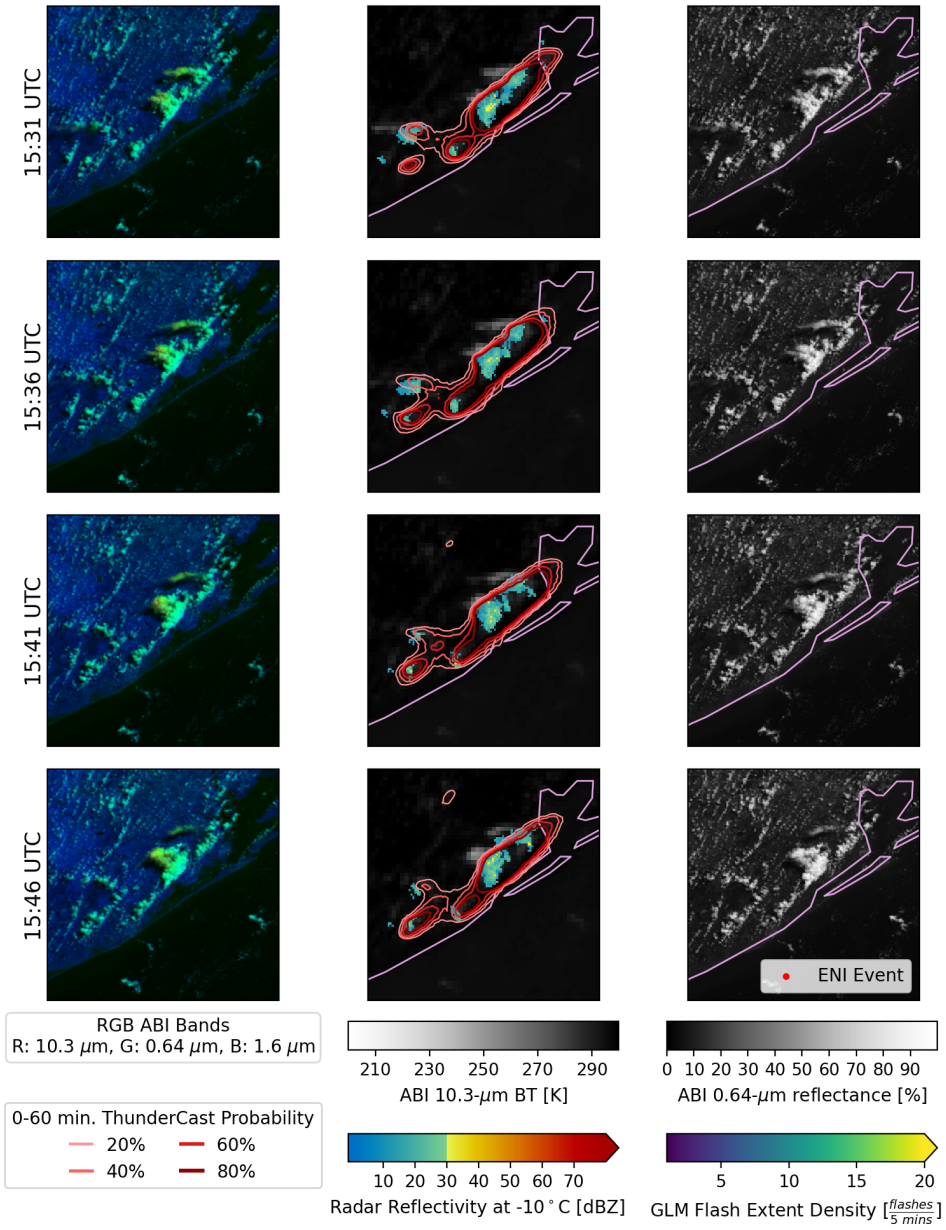

Figure 4.5: A time series of GOES-16 ABI images with corresponding MRMS radar data, model (ThunderCast) predictions, and GLM and ENI lightning data for the times closest to the ABI time (written to the left of each row) collected from 15:31 to 15:46 UTC on 2022-07-24. The plots are set-up in the same manor as Fig. 4.1 with the same dimensions and coordinates as Fig. 4.4. In the second and third columns, light purple lines display the Texas coastline.

September 7, 2022: Sustained Convection in Tennessee with Tropical Characteristics

Convective storms are influenced by the heating and circulation characteristics of their region of origin because growth and transport within a cloud depend on the cloud's interactions with the environment. In the tropics, towering, sustained cumuli are commonplace and do not always result in lightning production. This has been attributed, in part, to differences in entrainment rates between electrified and non-electrified storms (Singh and O'Gorman, 2013). According to Singh and O'Gorman (2013), large entrainment rates can reduce buoyancy, especially in warmer atmospheres, and decrease the strength of storm updrafts. Strong updrafts are important for charge separation in the non-inductive charging theory for lightning formation (Saunders, 1993), so towering cumulus cloud structures with high entrainment rates and, consequently, weaker updrafts could contribute to the frequency of non-electrified convective storms. Although high entrainment rates and weaker updrafts have largely been attributed to tropical convection, it is possible for them to be present in the midlatitudes as demonstrated in Tennessee (TN) on September 7, 2022.

On this date, moderate instability (CAPE values between 1000 J kg^{-1} and 2000 J kg^{-1}), weak wind shear, and high precipitable water (about 2 inches) were present in the mid-south region of the U.S throughout the afternoon and evening. CIN values indicated a weak cap that strengthened throughout the evening (National Weather Service Memphis Office, 2022). The surface temperature (82°F) and dew point temperature (72°F) at 18:00 UTC from the westernmost weather station in TN, shown in Fig. 4.6's surface plot, indicated high relative humidity (about 71%). Relative humidities remained high throughout the afternoon and evening.

In the early afternoon, scattered cumulus clouds were visible over western TN. The cumuli grew vertically and aggregated into an organized line traveling to the southwest along TN's western state border. At approximately 20:01 UTC the cloud-tops appeared to change from cyan to green in the DCPD RGB, indicating cloud-top glaciation. In Fig. 4.7, the DCPD RGBs in the first column display some orange

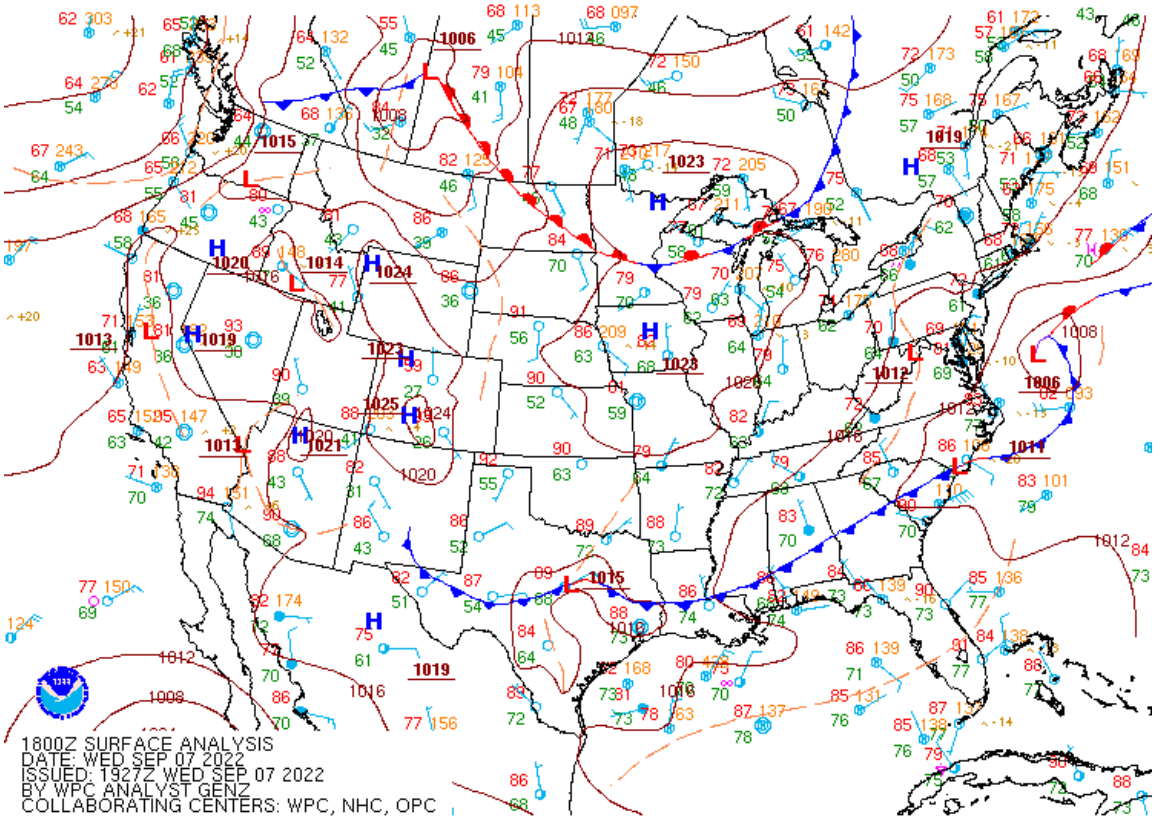

Figure 4.6: Meteorological surface conditions for September 7, 2022 at 18:00 UTC. This image was produced by the National Oceanic and Atmospheric Administration (NOAA) and is available to the public (National Weather Service, 2022b). Each standard weather station plot in the image contains information for wind direction, wind strength, sky cover, temperature, dew point, pressure, and pressure trend. Surface pressure also provided by the dark red contours and frontal boundaries are marked in red and blue.

pixels at cloud-top, suggesting continued vertical growth. The orange pixels are only present for the last two time steps in Fig. 4.7. Afterwards, the cumuli sustain their glaciated coloring for at least 2 more hours before dissipating. Although only showing reflectivities around 30 dBZ at $-10°C$ in Fig. 4.7, high reflectivity values (up to 51.0 dBZ) were observed at other time steps not shown.

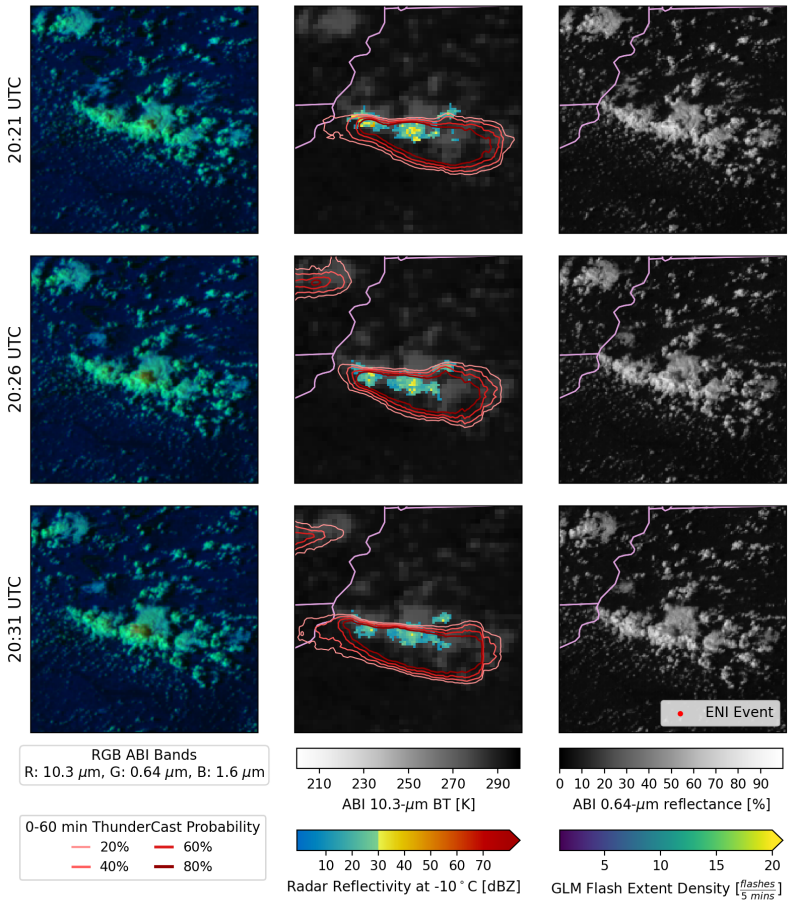

Figure 4.7: A time series of GOES-16 ABI images with corresponding MRMS radar data, model (ThunderCast) predictions, and GLM and ENI lightning data for the times closest to the ABI time (written to the left of each row) collected from 20:21 to 20:31 UTC on 2022-09-07. The plots are set-up in the same manor as Fig. 4.1. In the second and third columns, light purple state lines show TN's western border. Each image is 160 km $\times$ 160 km and the latitude and longitude coordinates of the lower lefthand corners are 34.7° and −90.7°, respectively.

This case study's non-electrified, long-lasting, and organized cumulus structure, formed in an environment with high humidity, is reminiscent of the towering tropical cumulus structures discussed earlier. It is speculated that high entrainment rates, weak updrafts, or the strengthening cap could have contributed to the lack of

lightning in this case. The existence of a tropical-like convective case in the midlatitudes indicates these structures are not confined to the tropics and could complicate the reduction of false alarms in thunderstorm nowcasting models. Additionally, more work is needed to determine how the frequency of such structures could change with the changing climate and the corresponding Hadley cell circulation expansion (Lionello et al., 2024).

July 27, 2023: Convective Activity in Alaska

The cases presented thus far have all been selected from central or southern U.S. regions. However, non-electrified towering cumuli, with similar appearances to thunderstorms in satellite imagery, have been observed at higher latitudes including the upper midwestern U.S. states and Alaska. In this final case study from Alaska on July 27, 2023, imagery from two non-electrified convective storms are compared to nearby thunderstorms. On this day, an upper level ridge in the area caused well above average temperatures and scattered thunderstorms across the Central and Eastern Interiors (National Weather Service Fairbanks Office, 2023). Multiple time steps for each towering cumulus case for this day are presented in Figs. 4.8, 4.9, and 4.10.

In Fig. 4.8, ThunderCast identifies an area with cold 10.3 μm brightness temperatures in the lower right-hand corner of the image as having a high probability of producing a thunderstorm in the next hour. The corresponding NtMicro RGB displays the potential thunderstorm as dark red with some lighter colors in the middle. The large red contribution in the RGB indicates the cloud is thick and the lighter yellow colors indicate it is very cold, both characteristics of towering cumulus or mature cumulonimbus. There are no ENI lightning events observed. However, in the upper left-hand corner similarly colored areas in the NtMicro RGB with ThunderCast predictions have observed lightning.

Additionally, in Fig. 4.9, the developing cumulus just left of center exhibits colors (bright oranges and greens) in the RGB consistent with the early development stages of thunderstorm growth. The indicated developing cumulus in Fig. 4.9 does

not produce lightning, but the dissipating cumulonimbus to its right did contain lightning about 40 min earlier. The initiation stage of the dissipating cumulonimbus in Fig. 4.9 is shown in Fig. 4.10 and contains similar colors in the RGBs. The cloud-tops are more saturated with red in Fig. 4.10 than Fig. 4.9, but this can be accounted for by the close temporal proximity to sunrise. Altogether, Figs. 4.8-4.10 demonstrate non-electrified convective storms can appear similar to electrified storms in satellite imagery at high latitudes.

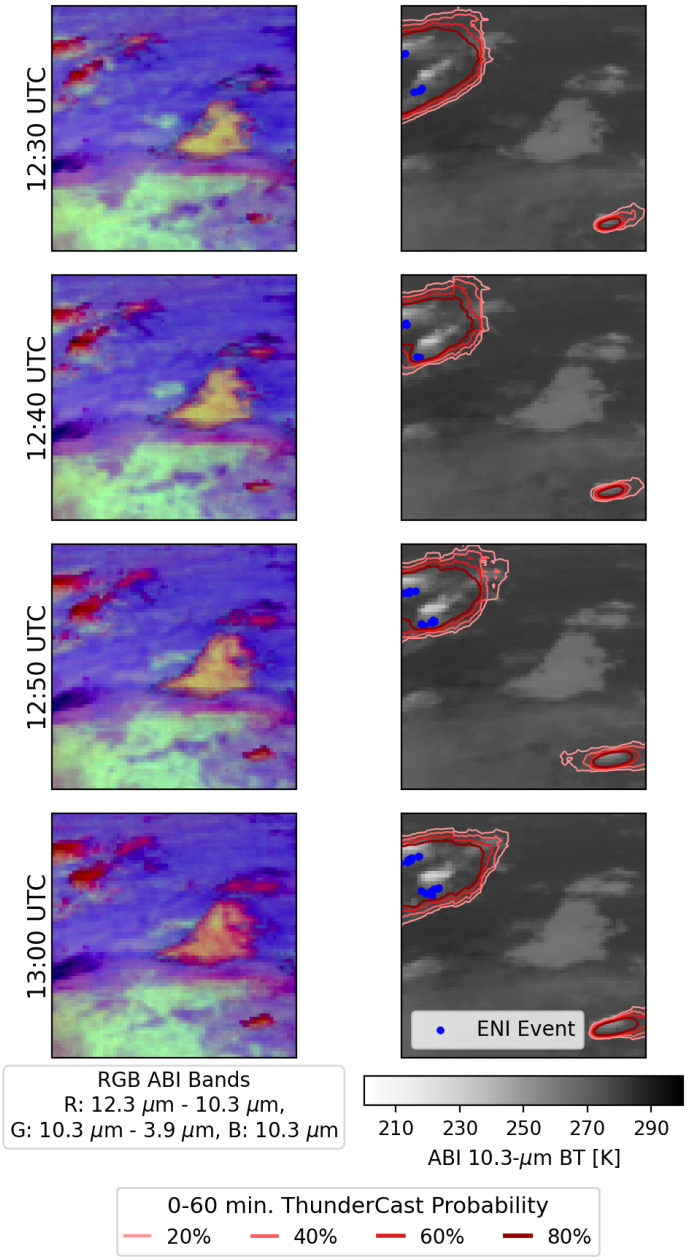

Figure 4.8: A time series of GOES-18 ABI images with corresponding lightning and model (ThunderCast) data for the times closest to the ABI time (written to the left of each row) collected from 12:30 to 13:00 UTC on 2023-07-27. The first column contains GOES-18 NtMicro false color RGB images. The second column contains output from ThunderCast, GOES-18 ABI 10.3 μm brightness temperatures, and ENI total lightning (blue points). Each image is 144 pixels × 144 pixels at 1-km resolution in a geostationary projection and the latitude and longitude coordinates of the lower lefthand corners are 62.3° and −151.6°, respectively.

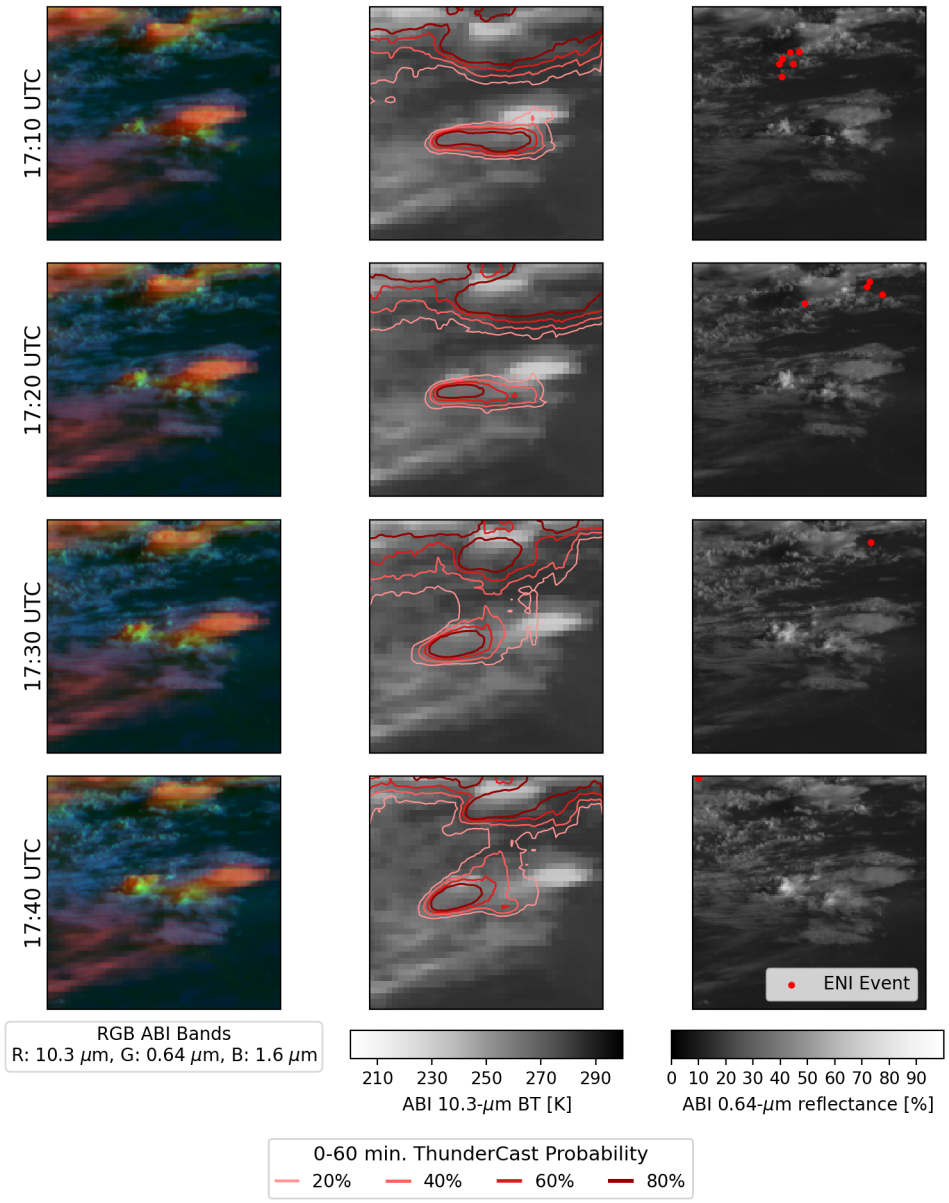

Figure 4.9: A time series of GOES-18 ABI images with corresponding lightning and model (ThunderCast) data for the times closest to the ABI time (written to the left of each row) collected from 17:10 to 17:40 UTC on 2023-07-27. The first column contains GOES-18 DCPD false color RGB images. The second column contains output from ThunderCast and GOES-18 ABI 10.3 μm brightness temperatures. The last column contains ENI total lightning events (red points) and GOES-18 ABI 0.64 μm reflectances. Each image is 96 pixels × 96 pixels at 1-km resolution in a geostationary projection and the latitude and longitude coordinates of the lower lefthand corners are 61.4° and −148.1°, respectively.

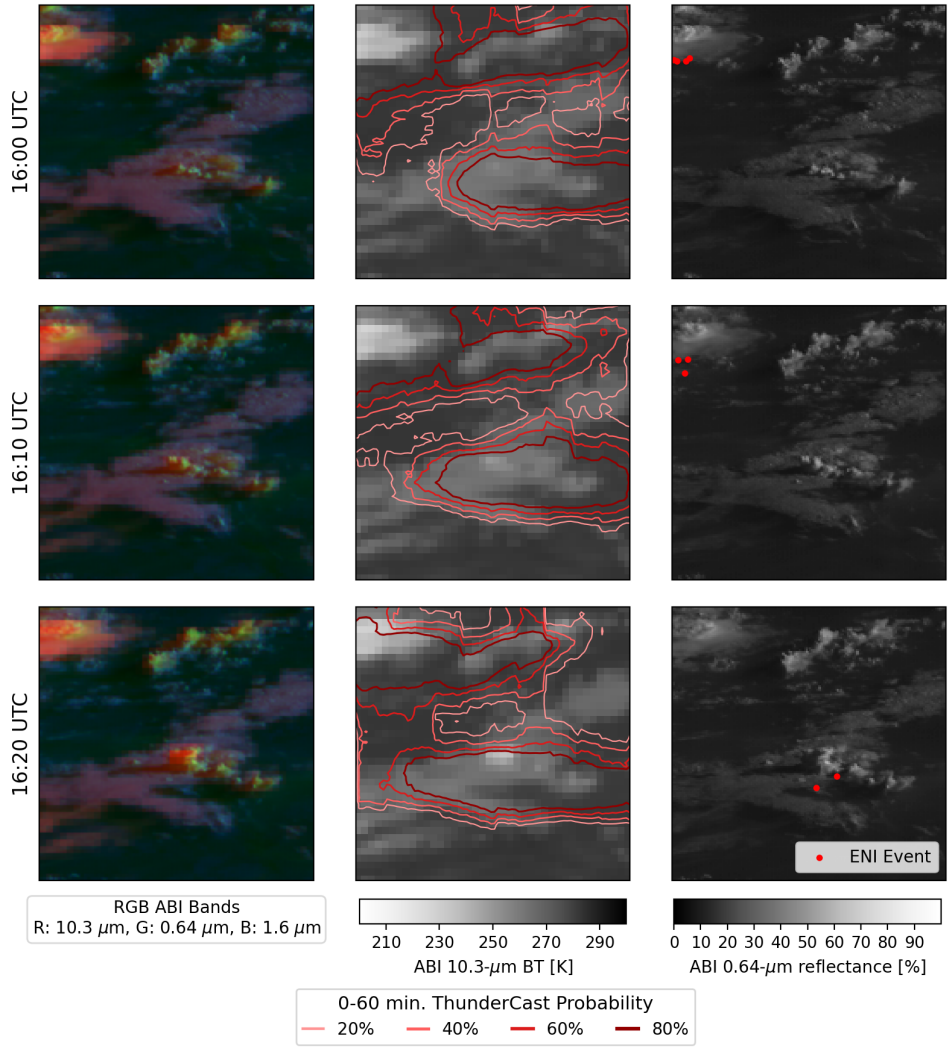

Figure 4.10: A time series of GOES-18 ABI images with corresponding lightning and model (ThunderCast) data for the times closest to the ABI time (written to the left of each row) collected from 16:00 to 16:20 UTC on 2023-07-27. The plots are set-up in the same manor and with the same dimensions as Fig. 4.9

4.3 Summary

False color red-green-blue (RGB) composite imagery, such as the Day Cloud Phase Distinction (DCPD) or Nighttime Microphysics (NtMicro) RGBs, are important for visually identifying convective lifecycle stages because the satellite spectral bands comprising the imagery are physically relatable to cloud microphysics and morphology. Because of their applicability to convective initiation and existing, manual thunderstorm identification forecasting methods, satellite spectral bands, including those used in the DCPD and NtMicro RGBs, have been incorporated along with radar data into artificial intelligence models for short-term forecasting of thunderstorms. However, recent work from Chapter 3 suggested developing convective storms with lightning (thunderstorms) often appear similar in satellite imagery and contain similar radar signatures to those that do not produce lightning. Four case studies containing non-electrified convective storms with similar appearances, in either the DCPD or NtMicro RGB images, to thunderstorms were presented in this chapter, verifying the results from Chapter 3.

In the presented case studies, non-electrified convective storms with similar appearances to thunderstorms were not limited by latitude or known environmental conditions. These non-electrified storms were located at a variety of latitudes, ranging from approximately 28° (Texas' Gulf Coast) to 63° (Alaska). Also, the storms formed in an assortment of meteorological environments including a coastal environment with an active sea breeze (Texas on July 24, 2022), a high humidity case with similar properties to tropical environments (Tennessee on September 7, 2022), and an environment exhibiting key meteorological ingredients for severe weather activity like high convective available potential energy, low convective inhibition, and large wind shear (Oklahoma on May 15, 2022). In some of these environments, other convective storms were able to successfully produce lightning, demonstrating the complexity of thunderstorm formation processes. Additionally, characteristics of the Tennessee case were similar to convection observed in the tropics, demonstrating tropical and midlatitude convection may not be as distinctive as the latitudinal labels imply.

Further analysis is needed to determine why some convective storms die prematurely while others continue to grow and if any remote sensing observations can be used to distinguish between initiating non-electrified and electrified storms. In doing so, potential avenues for convective initiation model improvement can be explored. The frequency of observing convection with tropical properties in the midlatitudes remains a topic of future work as well as how such frequencies could change with a warming climate.

5 CONCLUSION

5.1 Summary of Research Questions

Deep learning, a form of artificial intelligence, has been increasingly used for short-term forecasting (nowcasting) of weather phenomena because of its ability to synthesize vast quantities of environmental data into actionable insights. In this dissertation, the current scientific understanding of convection from a satellite and radar perspective was used to inform the development and evaluation of a deep learning model for convective storm (thunderstorm) nowcasting in the continental United States. In doing so, three primary research questions were explored. In this section, each research question is reiterated and the key findings, as they relate to the research questions, are summarized.

1. How well can the occurrence of convection in the next hour be predicted with a deep learning model developed from the current scientific understanding of thunderstorm signatures in satellite imagery and ground-based radar observations?

The Thunderstorm Nowcasting Tool (ThunderCast) was developed for predicting thunderstorm occurrence in the next 0-60 min using a U-Net convolutional neural network for semantic segmentation. The model was trained with four satellite spectral bands from the Geostationary Environmental Operational Satellite-16 (GOES-16) Advanced Baseline Imager (ABI) as predictors: 0.64-μm reflectance (channel 2; red band), 1.6-μm reflectance (channel 5; snow/ice band), 10.3-μm (channel 13; clean longwave window band) brightness temperature, and 12.3-μm (channel 15; dirty longwave window band) brightness temperature. These particular bands were selected because they are commonly utilized by forecasters for diagnosing trends and patterns in cumuliform clouds including cloud-top glaciation, cloud-top temperature, and cloud morphology. The target dataset, used for training and model validation, was the maximum Multi-Radar Multi-Sensor (MRMS) radar reflectivity at -10°C in the next hour, where anything $\geqslant 30$ dBZ was

considered positive for thunderstorm occurrence. This target dataset was chosen because it represents the earliest radar signature of convective initiation in previous research studies and it agrees well with radar signatures observed in storms during field campaigns.

Statistical analysis of ThunderCast revealed the model performs consistently across the contiguous united states (CONUS) and does well in terms of accuracy, recall, and specificity. However, the model is prone to false alarms, resulting in low precision. Accentuating this issue, the precision values were found to be higher during the day than at night, indicating false alarms are further elevated at night. This demonstrates a deep learning model, built with the current scientific understanding of thunderstorm signatures in satellite imagery and radar observations, can be highly accurate when predicting thunderstorm occurrence in the next hour but exhibits a tendency to over predict storms. Despite the false alarms, ThunderCast effectively identifies regions of interest for convective development and can be used to inform situational awareness.

2. What environmental conditions and/or observational factors (associated with the selection of ThunderCast's inputs and target) impact the reliability and applicability of ThunderCast for thunderstorm prediction?

To gather environmental and observational data for ThunderCast's predicted storms, an established tool for the Tracking and Object-Based Analysis of Clouds (tobac) was applied to ThunderCast's probabilistic output. In doing so, regions of ThunderCast predictions were identified and tracked for 7 randomly selected days in each month of the convective season in 2022. For simplicity, the dataset was filtered to focus on new convection without complex merge or split scenarios. MRMS radar reflectivity at -10°C, lightning observations from the Geostationary Lightning Mapper (GLM) and Earth Networks Inc.'s (ENI) total lightning product, and satellite observations from the 1.6 μm and 10.3 μm spectral bands were collected for each predicted storm track. Based on the observations collected for each track, the tracks were sorted into true positive (MRMS radar $\geqslant$ 30 dBZ at -10°C per

ThunderCast's radar definition of convective initiation) and false positive (< 30 dBZ at $-10°$C) categories with and without lightning.

Out of the 24,770 land-based predicted storm tracks identified, 67.8% were true positive and 59.3% of the true positive cases were not associated with GLM or ENI lightning observations. The large percentage of new convection without lightning affects the reliability and applicability of ThunderCast for thunderstorm prediction because the predictions can not be relied upon to correspond to thunderstorms even when they reach MRMS reflectivities $\geqslant 30$ dBZ. Although the number of true positive cases without lightning decreased with higher maximum reflectivity at $-10°$C, there were still non-electrified tracks with high reflectivity values ($\geqslant 40$ dBZ at $-10°$C). This indicates definitions of convective initiation that rely on thresholds in radar observations are not adequate for isolating convection associated with thunderstorms. Additionally, most of the predicted storm tracks, regardless of true or false positive with or without lightning categorization had cloud-top brightness temperatures less than freezing and low 1.6 μm reflectances, consistent with the presence of ice at cloud-top and cloud-top glaciation. Because true and false positive tracks with and without lightning contain these signatures, the tracks likely appear similar in satellite imagery, which could contribute to high false alarm rates. Thus, both cloud-top conditions and radar observations can appear similar for non-electrified and electrified convective storms, impacting ThunderCast's reliability and applicability for thunderstorm prediction.

3. What scientific insights can be gained, what limitations are exposed, and what potential avenues for model improvement can be identified from evaluating ThunderCast case studies?

In the exploration of the second research question, non-electrified and electrified convective storms were found to likely exhibit similar cloud-top properties from a satellite perspective. For further investigation, four case studies were presented where ABI cloud-top signatures in the Day Cloud Phase Distinction (DCPD) and Nighttime Microphysics (NtMicro) RGBs appeared consistent with thunderstorm development during convective initiation but did not go on to produce lightning.

The case studies demonstrated non-electrified convective storms can occur at a wide range of latitudes (ranging from 28° and 63° in the case studies) and under a variety of meteorological conditions (e.g., sea breeze, high humidity, and in a primed severe weather environment). In some of the case studies' environments, other convective storms were able to successfully produced lightning. This demonstrates known, generalizable precursors for thunderstorm initiation from remote sensing and ground-based observations are not able to provide enough information to determine which developing cumuli will grow into mature, electrified cumulonimbi. This limitation impacts the ability of artificial intelligence models to make precise predictions, and more work is needed to determine how to address this knowledge and observation gap.

In addressing the three primary research questions in this dissertation, a deep learning model, built from the current scientific understanding for thunderstorm characteristics from a radar and satellite perspective, provided valuable insight into how well our observations and scientific understanding can be used for thunderstorm prediction. The high false alarm rates in the model and the model's inability to distinguish between electrified and non-electrified storms reveals the model's limitations. Future research will be needed to address these limitations in terms of both improvements to the scientific understanding of the thunderstorm lifecycle and thunderstorm prediction models.

5.2 Future Work

There are many avenues for future work stemming from the research described in this dissertation. One option is to test different types of deep learning models to determine which architecture is best for thunderstorm nowcasting. ThunderCast is a deep learning model built with a U-Net convolutional neural network. However, there are many other types of architectures that could, potentially, perform better than a U-Net. Some possibilities include using transformers (Lin et al., 2022) or adapting attention models for attention-based semantic segmentation (Guo et al., 2022). Also, ThunderCast did not use time as a parameter in the model.

Instead, it used 1-hour maximum radar reflectivity values from ground-based radars as the target dataset to provide some context of convective activity in the next hour. Other model types could incorporate time into the analysis such as U-Time (Perslev et al., 2019) or satellite image time series (SITS) models (Pelletier et al., 2019). Although this dissertation did not focus on testing different model types, future research studies could investigate which models perform optimally for thunderstorm prediction.

In Chapters 2 and 3, statistical analysis of ThunderCast's output revealed ThunderCast is prone to false alarms. In Chapter 3, some of the false alarms corresponded to regions of predictions where clouds were not present, and these cases tended to occur near high-contrast physical features in the satellite imagery like lakes and rivers. To reduce false alarms, more clear-sky training cases near high-contrast physical features could be incorporated into ThunderCast's training dataset to ensure the model has adequate exposure to such features. Additionally, in Chapter 3, using a ground-based radar threshold of convective initiation did not completely filter out stratiform convection from the dataset, which resulted in thunderstorm predictions for non-cumuliform cloud types. Future work can investigate alternative methods to identify convection associated with thunderstorms in the target dataset, which could reduced false alarm rates. Another motivation for using an alternative method for identifying convection from thunderstorms is the majority of ThunderCast's predicted thunderstorm cases, with a ground-based radar signature meeting the defined threshold (MRMS $\geqslant$ 30 dBZ at -10°C), did not have any associated GLM or ENI lightning. This indicates the ground-based radar threshold was not able to distinguish between electrified and non-electrified storms, so additional research is needed to determine if an alternative methodology could eliminate non-electrified storms from the training dataset.

Furthermore, the non-electrified and electrified predicted storms were found to have similar 10.3 μm brightness temperatures and 1.6 μm reflectances in Chapter 3. They also appeared similar during initiation in the satellite imagery presented in the case studies in Chapter 4. In order for a deep learning model to distinguish between non-electrified and electrified storms, patterns need to be discernible in the satellite

data. This indicates, if forecasters desire a model identifying convection only associated with electrified storms, additional predictors may need to be incorporated into future deep learning models for thunderstorm nowcasting in order to avoid identification of non-electrified cases. More work is needed to determine which predictors would be appropriate, but some potential options include additional satellite bands (6.2 μm, water vapor for example), dual-polarization radar data, radar data from satellites, or numerical weather prediction models.

Another avenue for future work is to implement more spatial and temporal comparisons of predicted storm characteristics. For example, in Chapter 3, all of the U.S. climate regions had more cases with MRMS reflectivity values $\geqslant 30$ dBZ at -10°C without lightning than with lightning except for the southwest. Carlaw et al. (2017) recently found the greatest frequencies of lightning in southwestern Arizona were concentrated around the highest terrain. The location of ThunderCast's predicted storms with and without observed lightning could be compared to the findings in Carlaw et al. (2017). Additionally, new ThunderCast models could be trained for specific regions in the United States (e.g., a model trained just for the southwest). The resulting model statistics could be compared to the original ThunderCast model to see if a smaller regional model could have improved performance over the larger national model in the same area.

As detailed in this, final section of this dissertation, there are many avenues for future work. More research is needed to deepen the current scientific understanding of thunderstorm lifecycles, the relationship between observations and thunderstorms, and how observations can be best synthesized for forecasting thunderstorms and their associated hazards.

REFERENCES

Allan, D. B., T. Caswell, N. C. Keim, C. van der Wel, and R. W. Verweij, 2023: soft-matter/trackpy: Trackpy v0.6.1. URL https://doi.org/10.5281/zenodo.1213240.

Baker, B., M. B. Baker, E. R. Jayaratne, J. Latham, and C. P. R. Saunders, 1987: The influence of diffusional growth rates on the charge transfer accompanying rebounding collisions between ice crystals and soft hailstones. *Quarterly Journal of the Royal Meteorological Society*, **113 (478)**, 1193–1215, https://doi.org/https://doi.org/10.1002/qj.49711347807, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711347807, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711347807.

Baker, M. B., and J. G. Dash, 1994: Mechanism of charge transfer between colliding ice particles in thunderstorms. *Journal of Geophysical Research: Atmospheres*, **99 (D5)**, 10 621–10 626, https://doi.org/https://doi.org/10.1029/93JD01633, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93JD01633, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/93JD01633.

Bradshaw, S. M., 2021: A deep learning model for nowcasting midlatitude convective storms. M.S. thesis, Atmospheric and Oceanic Sciences, University of Wisconsin-Madison.

Bruning, E. C., and Coauthors, 2019: Meteorological imagery for the geostationary lightning mapper. *Journal of Geophysical Research: Atmospheres*, **124 (24)**, 14 285–14 309, https://doi.org/https://doi.org/10.1029/2019JD030874, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD030874.

Carey, L. D., E. V. Schultz, C. J. Schultz, W. Deierling, W. A. Petersen, A. L. Bain, and K. E. Pickering, 2019: An evaluation of relationships between radar-inferred kinematic and microphysical parameters and lightning flash rates in alabama storms. *Atmosphere*, **10 (12)**, https://doi.org/10.3390/atmos10120796, URL https://www.mdpi.com/2073-4433/10/12/796.

Carlaw, L. B., A. E. Cohen, and J. W. Rogers, 2017: Synoptic and mesoscale environment of convection during the north american monsoon across central and southern arizona. *Weather and Forecasting*, **32** (**2**), 361 – 375, https://doi.org/10.1175/WAF-D-15-0098.1, URL https://journals.ametsoc.org/view/journals/wefo/32/2/waf-d-15-0098_1.xml.

Chollet, F., 2018: *Deep Learning with Python*. First edition ed., Manning Publications Co., Shelter Island, NY.

Cintineo, J. L., M. J. Pavolonis, and J. M. Sieglaff, 2022: Probsevere lightningcast: A deep-learning model for satellite-based lightning nowcasting. *Weather and Forecasting*, **37** (**7**), 1239 – 1257, https://doi.org/10.1175/WAF-D-22-0019.1, URL https://journals.ametsoc.org/view/journals/wefo/37/7/WAF-D-22-0019.1.xml.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020: A deep-learning model for automated detection of intense midlatitude convection using geostationary satellite images. *Weather and Forecasting*, **35** (**6**), 2567 – 2588, https://doi.org/10.1175/WAF-D-20-0028.1, URL https://journals.ametsoc.org/view/journals/wefo/35/6/waf-d-20-0028.1.xml.

Connell, B., E. Dagg, M. Bowlan, and K. Fuell, n.d.: Daynightcloudmicrocombo quick guide. Tech. rep., National Oceanic and Atmospheric Administration. URL https://www.star.nesdis.noaa.gov/goes/documents/ABIQuickGuide_DayNightCloudMicroCombo.pdf.

Cuomo, J., and V. Chandrasekar, 2021: Use of deep learning for weather radar nowcasting. *Journal of Atmospheric and Oceanic Technology*, **38** (**9**), 1641 – 1656, https://doi.org/10.1175/JTECH-D-21-0012.1, URL https://journals.ametsoc.org/view/journals/atot/38/9/JTECH-D-21-0012.1.xml.

Dash, J. G., B. L. Mason, and J. S. Wettlaufer, 2001: Theory of charge and mass transfer in ice-ice collisions. *Journal of Geophysical Research: Atmospheres*, **106** (**D17**), 20 395–20 402, https://doi.org/https://doi.org/10.1029/2001JD900109,

URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD900109, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001JD900109.

Dye, J. E., L. J. M. andBrooks E. Martner, and Z. Levin, 1982: The 25 july 1976 case study: environmental conditions, reflectivity structure, and evolution. *Hailstorms of the Central High Plains*, C. A. Knight, and P. Squires, Eds., Vol. 2, Colorado Associated University Press, 197–209.

Dye, J. E., and B. E. Martner, 1982: The 25 july 1976 case study: microphyscial observations. *Hailstorms of the Central High Plains*, C. A. Knight, and P. Squires, Eds., Vol. 2, Colorado Associated University Press, 211–228.

Earth Networks, 2024: Lightning data. URL https://www.earthnetworks.com/product/lightning-data/.

Elsenheimer, C. B., and C. M. Gravelle, 2019: Introducing lightning threat messaging using the goes-16 day cloud phase distinction rgb composite. *Weather and Forecasting*, **34 (5)**, 1587 – 1600, https://doi.org/10.1175/WAF-D-19-0049.1, URL https://journals.ametsoc.org/view/journals/wefo/34/5/waf-d-19-0049_1.xml.

Fan, D., S. J. Greybush, E. E. Clothiaux, and D. J. Gagne, 2024: Physically explainable deep learning for convective initiation nowcasting using goes-16 satellite observations. *Artificial Intelligence for the Earth Systems*, https://doi.org/10.1175/AIES-D-23-0098.1, URL https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0098.1/AIES-D-23-0098.1.xml.

Fankhauser, J. C., and C. Wade, 1982: The environment of the storms. *Hailstorms of the Central High Plains*, C. A. Knight, and P. Squires, Eds., Vol. 1, Colorado Associated University Press, 5–33.

Goodman, S. J., and Coauthors, 2013: The goes-r geostationary lightning mapper (glm). *Atmospheric Research*, **125-126**, 34–49, https://doi.org/https://doi.org/10.1016/j.atmosres.2013.01.006, URL https://www.sciencedirect.com/science/article/pii/S0169809513000434.

Gremillion, M. S., and R. E. Orville, 1999: Thunderstorm characteristics of cloud-to-ground lightning at the kennedy space center, florida: A study of lightning initiation signatures as indicated by the wsr-88d. *Weather and Forecasting*, **14 (5)**, 640 – 649, https://doi.org/10.1175/1520-0434(1999)014<0640:TCOCTG>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/14/5/1520-0434_1999_014_0640_tcoctg_2_0_co_2.xml.

Griffin, S. M., A. Wimmers, and C. S. Velden, 2022: Predicting rapid intensification in north atlantic and eastern north pacific tropical cyclones using a convolutional neural network. *Weather and Forecasting*, **37 (8)**, 1333 – 1355, https://doi.org/10.1175/WAF-D-21-0194.1, URL https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0194.1.xml.

Guo, M.-H., C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, 2022: Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, **35**, 1140–1156.

Han, D., J. Lee, J. Im, S. Sim, S. Lee, and H. Han, 2019: A novel framework of detecting convective initiation combining automated sampling, machine learning, and repeated model tuning from geostationary satellite data. *Remote Sensing*, **11 (12)**, https://doi.org/10.3390/rs11121454, URL https://www.mdpi.com/2072-4292/11/12/1454.

Heikenfeld, M., P. J. Marinescu, M. Christensen, D. Watson-Parris, F. Senf, S. C. van den Heever, and P. Stier, 2019: tobac 1.2: towards a flexible framework for tracking and analysis of clouds in diverse datasets. *Geoscientific Model Development*, **12 (11)**, 4551–4570, https://doi.org/10.5194/gmd-12-4551-2019, URL https://gmd.copernicus.org/articles/12/4551/2019/.

Helmus, J. J., and S. M. Collis, 2016: The python arm radar toolkit (py-art), a library for working with weather radar data in the python programming language. *Journal of Open Research Software*, **4**.

Hilburn, K., 2023: Gremlin: Goes radar estimation via machine learning to inform nwp. Ph.D. thesis, Colorado State University.

Houze Jr., R., 2014: *Cloud Dynamics*. 2nd ed., Academic Press.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293, https://doi.org/https://doi.org/10.1016/0169-2070(86)90048-8, URL https://www.sciencedirect.com/science/article/pii/0169207086900488.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. W. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The storm cell identification and tracking algorithm: An enhanced wsr-88d algorithm. *Weather and Forecasting*, **13 (2)**, 263 – 276, https://doi.org/10.1175/1520-0434(1998)013<0263:TSCIAT>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/13/2/1520-0434_1998_013_0263_tsciat_2_0_co_2.xml.

Karl, T. R., and W. J. Koss, 1984: Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. *Historical Climatology Series 4-3*, National Climatic Data Center, Asheville, NC, 1–38.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, https://doi.org/10.48550/ARXIV.1412.6980.

Kruskal, J. B., 1956: On the shortest spanning subtree of a graph and the traveling salesman problem. URL https://api.semanticscholar.org/CorpusID:120068278.

Lagerquist, R., A. McGovern, C. R. Homeyer, D. J. G. II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, **148 (7)**, 2837 – 2861, https://doi.org/10.1175/MWR-D-19-0372.1, URL https://journals.ametsoc.org/view/journals/mwre/148/7/mwrD190372.xml.

Lagerquist, R., A. McGovern, and D. J. G. II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, **34 (4)**, 1137 –

1160, https://doi.org/10.1175/WAF-D-18-0183.1, URL https://journals.ametsoc.org/view/journals/wefo/34/4/waf-d-18-0183_1.xml.

Lagerquist, R., J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using deep learning to nowcast the spatial coverage of convection from himawari-8 satellite data. *Monthly Weather Review*, **149 (12)**, 3897 – 3921, https://doi.org/10.1175/MWR-D-21-0096.1, URL https://journals.ametsoc.org/view/journals/mwre/149/12/MWR-D-21-0096.1.xml.

Lee, S., H. Han, J. Im, E. Jang, and M.-I. Lee, 2017: Detection of deterministic and probabilistic convection initiation using himawari-8 advanced himawari imager data. *Atmospheric Measurement Techniques*, **10 (5)**, 1859–1874, https://doi.org/10.5194/amt-10-1859-2017, URL https://amt.copernicus.org/articles/10/1859/2017/.

Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff, 2021: Applying machine learning methods to detect convection using geostationary operational environmental satellite-16 (goes-16) advanced baseline imager (abi) data. *Atmospheric Measurement Techniques*, **14 (4)**, 2699–2716, https://doi.org/10.5194/amt-14-2699-2021, URL https://amt.copernicus.org/articles/14/2699/2021/.

Lin, T., Y. Wang, X. Liu, and X. Qiu, 2022: A survey of transformers. *AI Open*, **3**, 111–132, https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001, URL https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Lionello, P., R. D'Agostino, D. Ferreira, H. Nguyen, and M. S. Singh, 2024: The hadley circulation in a changing climate. *Annals of the New York Academy of Sciences*, **1534 (1)**, 69–93, https://doi.org/https://doi.org/10.1111/nyas.15114, URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15114, https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.15114.

Liu, C., D. J. Cecil, E. J. Zipser, K. Kronfeld, and R. Robertson, 2012: Relationships between lightning flash rates and radar reflectivity vertical structures in thunderstorms over the tropics and subtropics. *Journal of Geophysical Research: Atmospheres*, **117 (D6)**, https://doi.org/https://doi.org/10.1029/2011JD017123,

URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD017123, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JD017123.

Liu, Y., Q. Ren, J. Geng, M. Ding, and J. Li, 2018: Efficient patch-wise semantic segmentation for large-scale remote sensing images. *Sensors (Basel, Switzerland)*, **18**, https://doi.org/10.3390/s18103232.

Maas, A. L., A. Y. Hannun, A. Y. Ng, and Coauthors, 2013: Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, Atlanta, Georgia, USA, Vol. 30, 3.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artificial Intelligence for the Earth Systems*, **2 (3)**, e220 077, https://doi.org/https://doi.org/10.1175/AIES-D-22-0077.1, URL https://journals.ametsoc.org/view/journals/aies/2/3/AIES-D-22-0077.1.xml.

Mecikalski, J. R., and K. M. Bedka, 2006: Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime goes imagery. *Monthly Weather Review*, **134 (1)**, 49 – 78, https://doi.org/10.1175/MWR3062.1, URL https://journals.ametsoc.org/view/journals/mwre/134/1/mwr3062.1.xml.

Mecikalski, J. R., W. M. MacKenzie, M. Koenig, and S. Muller, 2010a: Cloud-top properties of growing cumulus prior to convective initiation as measured by meteosat second generation. part i: Infrared fields. *Journal of Applied Meteorology and Climatology*, **49 (3)**, 521 – 534, https://doi.org/10.1175/2009JAMC2344.1, URL https://journals.ametsoc.org/view/journals/apme/49/3/2009jamc2344.1.xml.

Mecikalski, J. R., W. M. MacKenzie, M. König, and S. Muller, 2010b: Cloud-top properties of growing cumulus prior to convective initiation as measured by meteosat second generation. part ii: Use of visible reflectance. *Journal of Applied Meteorology and Climatology*, **49 (12)**, 2544 – 2558, https://doi.org/10.1175/2010JAMC2480.1, URL https://journals.ametsoc.org/view/journals/apme/49/12/2010jamc2480.1.xml.

Mecikalski, J. R., J. K. Williams, C. P. Jewett, D. Ahijevych, A. LeRoy, and J. R. Walker, 2015: Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, **54 (5)**, 1039 – 1059, https://doi.org/10.1175/JAMC-D-14-0129.1, URL https://journals.ametsoc.org/view/journals/apme/54/5/jamc-d-14-0129.1.xml.

Meyer, F., 1994: Topographic distance and watershed lines. *Signal Processing*, **38 (1)**, 113–125, https://doi.org/https://doi.org/10.1016/0165-1684(94)90060-4, URL https://www.sciencedirect.com/science/article/pii/0165168494900604, mathematical Morphology and its Applications to Signal Processing.

Morgan, G. M., and P. Squires, 1982: Introduction. *Hailstorms of the Central High Plains*, C. A. Knight, and P. Squires, Eds., Vol. 1, Colorado Associated University Press, 1–4.

Mueller, C., T. Saxen, R. Roberts, J. Wilson, T. Betancourt, S. Dettling, N. Oien, and J. Yee, 2003: Ncar auto-nowcast system. *Weather and Forecasting*, **18 (4)**, 545 – 561, https://doi.org/10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/18/4/1520-0434_2003_018_0545_nas_2_0_co_2.xml.

National Weather Service, 2022a: Storm prediction center product and report archive. URL https://www.spc.noaa.gov/archive/, webpage.

National Weather Service, 2022b: Weather prediction center's surface analysis archive. URL https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php, webpage.

National Weather Service Fairbanks Office, 2023: Area forecast discussion. Tech. Rep. FXAK69 PAFG 271054, Iowa Environmental Mesonet. URL https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=AFDAFG&e=202307271054.

National Weather Service Houston/Galveston Office, 2022: Area forecast discussion. Tech. Rep. FXUS64 KHGX 241125, Iowa Environmental Mesonet. URL https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=AFDHGX&e=202207241125.

National Weather Service Memphis Office, 2022: Area forecast discussion. Tech. Rep. FXUS64 KMEG 071738, Iowa Environmental Mesonet. URL https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=AFDMEG&e=202209071738.

National Weather Service Tulsa Office, 2022: Area forecast discussion. Tech. Rep. FXUS64 KTSA 151852, Iowa Environmental Mesonet. URL https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=AFDTSA&e=202205151852.

Oklahoma Climatological Survey, 2022: Oklahoma's weather network mesonet. Mesonet data files, University of Oklahoma and Oklahoma State University. URL https://www.mesonet.org/past-data/mesonet-resources/mesonet-data-files?ref=1778.

Ortland, S. M., and M. J. Pavolonis, 2024: An object-based evaluation of output from a deep learning model for thunderstorm nowcasting, in review.

Ortland, S. M., M. J. Pavolonis, and J. L. Cintineo, 2023: The development and initial capabilities of thundercast, a deep learning model for thunderstorm nowcasting in the united states. *Artificial Intelligence for the Earth Systems*, **2** (**4**), e230 044, https://doi.org/10.1175/AIES-D-23-0044.1, URL https://journals.ametsoc.org/view/journals/aies/2/4/AIES-D-23-0044.1.xml.

Pavolonis, M. J., 2010: Advances in extracting cloud composition information from spaceborne infrared radiances—a robust alternative to brightness temperatures. part i: Theory. *Journal of Applied Meteorology and Climatology*, **49** (**9**), 1992 – 2012, https://doi.org/10.1175/2010JAMC2433.1, URL https://journals.ametsoc.org/view/journals/apme/49/9/2010jamc2433.1.xml.

Pavolonis, M. J., A. K. Heidinger, and T. Uttal, 2005: Daytime global cloud typing from avhrr and viirs: Algorithm description, validation, and comparisons. *Journal*

of Applied Meteorology, **44** (**6**), 804 – 826, https://doi.org/10.1175/JAM2236.1, URL https://journals.ametsoc.org/view/journals/apme/44/6/jam2236.1.xml.

Pelletier, C., G. I. Webb, and F. Petitjean, 2019: Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, **11** (**5**), https://doi.org/10.3390/rs11050523, URL https://www.mdpi.com/2072-4292/11/5/523.

Perslev, M., M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, 2019: U-time: A fully convolutional network for time series segmentation applied to sleep staging. URL https://arxiv.org/abs/1910.11162, 1910.11162.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597** (**7878**), 672–677, https://doi.org/10.1038/s41586-021-03854-z, URL https://doi.org/10.1038/s41586-021-03854-z.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136** (**1**), 78 – 97, https://doi.org/https://doi.org/10.1175/2007MWR2123.1, URL https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml.

Roberts, R. D., and S. Rutledge, 2003: Nowcasting storm initiation and growth using goes-8 and wsr-88d data. *Weather and Forecasting*, **18** (**4**), 562 – 584, https://doi.org/10.1175/1520-0434(2003)018<0562:NSIAGU>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/18/4/1520-0434_2003_018_0562_nsiagu_2_0_co_2.xml.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24** (**2**), 601 – 608, https://doi.org/https://doi.org/10.1175/2008WAF2222159.1, URL https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *CoRR*, **abs/1505.04597**, URL http://arxiv.org/abs/1505.04597, 1505.04597.

Rudlosky, S. D., S. J. Goodman, K. S. Virts, and E. C. Bruning, 2019: Initial geostationary lightning mapper observations. *Geophysical Research Letters*, **46**, 1097–1104.

Saunders, C. P. R., 1993: A review of thunderstorm electrification processes. *Journal of Applied Meteorology and Climatology*, **32 (4)**, 642 – 655, https://doi.org/10.1175/1520-0450(1993)032<0642:AROTEP>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/apme/32/4/1520-0450_1993_032_0642_arotep_2_0_co_2.xml.

Saunders, C. P. R., H. Bax-norman, C. Emersic, E. E. Avila, and N. E. Castellano, 2006: Laboratory studies of the effect of cloud conditions on graupel/crystal charge transfer in thunderstorm electrification. *Quarterly Journal of the Royal Meteorological Society*, **132 (621)**, 2653–2673, https://doi.org/https://doi.org/10.1256/qj.05.218, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.218, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.218.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the abi on the goes-r series. *Bulletin of the American Meteorological Society*, **98 (4)**, 681 – 698, https://doi.org/10.1175/BAMS-D-15-00230.1, URL https://journals.ametsoc.org/view/journals/bams/98/4/bams-d-15-00230.1.xml.

Scofield, R. A., and J. F. W. Purdom, 1986: The use of satellite imagery data for mesoscale analyses and forecasting applications. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., American Meteorological Society, 118–150.

Sieglaff, J. M., L. M. Cronce, W. F. Feltz, K. M. Bedka, M. J. Pavolonis, and A. K. Heidinger, 2011: Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *Journal of Applied Meteorology and Climatology*, **50 (1)**, 110 – 126, https://doi.org/10.1175/2010JAMC2496.1, URL https://journals.ametsoc.org/view/journals/apme/50/1/2010jamc2496.1.xml.

Singh, M. S., and P. A. O'Gorman, 2013: Influence of entrainment on the thermal stratification in simulations of radiative-convective equilibrium. *Geophysical Research Letters*, **40 (16)**, 4398–4403, https://doi.org/https://doi.org/10.1002/grl.

50796, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50796, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/grl.50796.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97 (9)**, 1617 – 1630, https://doi.org/https://doi.org/10.1175/BAMS-D-14-00173.1, URL https://journals.ametsoc.org/view/journals/bams/97/9/bams-d-14-00173.1.xml.

Soille, P. J., and M. M. Ansoult, 1990: Automated basin delineation from digital elevation models using mathematical morphology. *Signal Processing*, **20 (2)**, 171–182, https://doi.org/https://doi.org/10.1016/0165-1684(90)90127-K, URL https://www.sciencedirect.com/science/article/pii/016516849090127K.

Sokolowsky, G. A., and Coauthors, 2023: *tobac* v1.5: Introducing fast 3d tracking, splits and mergers, and other enhancements for identifying and analysing meteorological phenomena. *EGUsphere*, **2023**, 1–37, https://doi.org/10.5194/egusphere-2023-1722, URL https://egusphere.copernicus.org/preprints/2023/egusphere-2023-1722/.

Stevens, E., L. Antiga, and T. Viehmann, 2020: *Deep Learning with PyTorch*. Manning Publications Co., Shelter Island, NY.

van der Walt, S., and Coauthors, 2014: scikit-image: image processing in Python. *PeerJ*, **2**, e453, https://doi.org/10.7717/peerj.453, URL https://doi.org/10.7717/peerj.453.

Virtanen, P., and Coauthors, 2020: Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, **17**, 261–272, URL https://doi.org/10.1038/s41592-019-0686-2.

Walker, J. R., W. M. MacKenzie, J. R. Mecikalski, and C. P. Jewett, 2012: An enhanced geostationary satellite–based convective initiation algorithm for 0–2-h nowcasting with object tracking. *Journal of Applied Meteorology and Climatology*,

51 (**11**), 1931 – 1949, https://doi.org/10.1175/JAMC-D-11-0246.1, URL https://journals.ametsoc.org/view/journals/apme/51/11/jamc-d-11-0246.1.xml.

Wilson, J. W., and W. E. Schreiber, 1986: Initiation of convective storms at radar-observed boundary-layer convergence lines. *Monthly Weather Review*, **114** (**12**), 2516 – 2536, https://doi.org/10.1175/1520-0493(1986)114<2516:IOCSAR>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/114/12/1520-0493_1986_114_2516_iocsar_2_0_co_2.xml.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, **147** (**6**), 2261 – 2282, https://doi.org/10.1175/MWR-D-18-0391.1, URL https://journals.ametsoc.org/view/journals/mwre/147/6/mwr-d-18-0391.1.xml.

Zhang, J., and Coauthors, 2016: Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97** (**4**), 621 – 638, https://doi.org/10.1175/BAMS-D-14-00174.1, URL https://journals.ametsoc.org/view/journals/bams/97/4/bams-d-14-00174.1.xml.

Zhou, K., Y. Zheng, W. Dong, and T. Wang, 2020: A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of Atmospheric and Oceanic Technology*, **37** (**5**), 927 – 942, https://doi.org/10.1175/JTECH-D-19-0146.1, URL https://journals.ametsoc.org/view/journals/atot/37/5/jtech-d-19-0146.1.xml.

Zhou, X., Y.-a. Geng, H. Yu, Q. Li, L. Xu, W. Yao, D. Zheng, and Y. Zhang, 2022: Lightnet+: A dual-source lightning forecasting network with bi-direction spatiotemporal transformation. *Applied Intelligence*, **52** (**10**), 11 147–11 159, https://doi.org/10.1007/s10489-021-03089-5, URL https://doi.org/10.1007/s10489-021-03089-5.

Zipser, E. J., and K. R. Lutz, 1994: The vertical profile of radar reflectivity of convective cells: A strong indicator of storm intensity and lightning probability? *Monthly Weather Review*, **122 (8)**, 1751 – 1759, https://doi.org/10.1175/1520-0493(1994)122<1751:TVPORR>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/122/8/1520-0493_1994_122_1751_tvporr_2_0_co_2.xml.