UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

USING MACHINE LEARNING TO IMPROVE THE NSSL'S WARN-ON-FORECAST SYSTEM'S PREDICTION OF THUNDERSTORM LOCATION

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

CHAD WILEY
Norman, Oklahoma
2023

USING MACHINE LEARNING TO IMPROVE THE NSSL'S WARN-ON-FORECAST SYSTEM'S PREDICTION OF THUNDERSTORM LOCATION

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Corey Potvin, Chair

Dr. Amy McGovern, Co-Chair

Dr. Montgomery Flora

Dr. Cameron Homeyer

Acknowledgements

This thesis would not have been possible without the help and support of my advisors, friends and family, and colleagues. Without the support and sacrifices made by each person, I would not have been able to find the success I have found in my time at the University of Oklahoma and NSSL. I would like to first and foremost thank my advisors, Dr. Corey Potvin, Dr. Amy McGovern, and Dr. Montgomery Flora. Their guidance, mentorship, and expertise led to me reaching to goals I had set before myself. The bi-weekly meeting with Dr. Potvin and Dr. Flora challenged me to become a better scientist through discussion of findings, thought processes, and communication of my work, and I feel I have learned so much from that time spent meeting. Dr. Flora was also instrumental in the guidance of verification and explainability techniques in machine learning and proved crucial to the project. Dr. Amy McGoven provided guidance, advice, and new ideas to try with my project. Her support and guidance in our weekly meetings often re-centered me on the task, and I would often leave the meeting feeling revitalized and motivated to tackle the challenges the master's degree presented. I would also like to extend thank yous to Dr. Randy Chase and Tobias Schmidt, who both provided code and were patient teachers. Dr. Chase's extensive knowledge of deep learning models and supercomputing saved me time and enabled me to learn so much. Mr. Schmidt's patching code, debugging expertise, and friendship also saved me immense time and was always someone I could rely on for help when I was stuck. I also need to extend my deepest gratitude to my family and partner, Jess. Their love, support, and visits allowed me to disconnect from the challenges of the program and recharge. I am so blessed to have such an amazing support system. Lastly, I would be remiss if I failed to mention thanks to my two favorite sports teams, the Philadelphia Phillies and Eagles. Their long post-season runs provided me with a much-needed distraction and nightly break from my work. Their subsequent losses in the championships then also motivated me to focus more on my work afterward so that I could forget the heartbreaking losses both teams inflicted on me. Overall, the time each and everyone spent on me during my two years at OU lead me to grow into a better scientist, and I will be forever grateful.

This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758. Funding was provided by NOAA/Office of

Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

Table of Contents

List Of Tables

List Of Figures

Abstract

Deep learning (DL) models have become immensely popular in recent years, with many models creating accurate and high-skill predictions for a wide range of atmospheric phenomena. Using DL models for predicting convection and associated hazards has experienced some of the most substantial gains in skill. The National Severe Storms Laboratory (NSSL) has created the experimental Warn-On-Forecast System (WoFS) to increase warning lead times through probabilistic short-term forecasts of individual thunderstorms. Currently, the WoFS has a shortcoming of missing storms due largely to poorly initialized environments. To help mitigate this issue, we developed a U-Net deep learning model to predict locations of thunderstorms trained on WoFS model data consisting of environmental data, such as CAPE and CIN, and intra-storm variables, such as WoFS ensemble average, mean, and max composite reflectivity and updraft and downdraft velocities. To address the issue of poorly initialized environments and lagging data assimilation, the model also has access to Multi-Radar/Multi-Sensor System (MRMS) data valid at the WoFS initialization is also used as an input. To evaluate the skill of the DL-based guidance, different baseline methods were tested to ensure a substantial performance increase. Comparing the performances of the WoFS baseline and DL model on an independent testing dataset, we were able to increase the maximum critical success index from 0.17 to 0.27, along with increasing the reliability and discrimination of the predictions. Using MRMS composite reflectivity proved to be vital for the DL model's performance when predicting values $\geq$ 40 dBZ. Through this work, we demonstrate DL models are an effective and efficient solution to improving the skill of the WoFS forecast of convection with a 30-minute lead time.

Chapter 1

Introduction

Accurate and precise prediction of thunderstorm location is vital to the partners and stakeholders within the weather enterprise. Thunderstorms impact almost every aspect of the US economy. In 2011, thunderstorms and associated hazards cost $47 billion in economic losses (Sander et al. 2013). Through accurate predictions of thunderstorms, some of these economic impacts can be lessened or avoided. Whether NWS forecasters creating skillful forecasts and communicating them with local governments during potentially severe weather outbreaks, the Weather Prediction Center creating rainfall forecasts for nationwide communities, or aviation partners creating safe flight plans to avoid locations of intense turbulent air, predicting the location of thunderstorms is essential to fulfilling NOAA's future vision of creating resilient communities and economies.

Thunderstorm prediction heavily impacts the aviation sector. Aviation forecast discussions (AFDs) and thunderstorm forecasts from local NWS weather forecasting offices (WFOs) are used to avoid convection-induced turbulence and to uphold strict safety standards expected within the industry. Fast-forming convection can quickly and without notice impact the safety of an airplane. The incident on December 18, 2022, is an example of how quickly convection-induced turbulence can impact the safety of passengers on board a flight. According to the NTSB's preliminary report, Hawaiian Airlines Flight 35, an Airbus A330, was on the final approach reporting clear conditions visually and through the on-plane weather radar when, 'like a smoke plume,' a cloud shot up in front of the plane. Within seconds the plane encountered

severe turbulence, resulting in 6 major injuries and 19 minor injuries (National Transportation Safety Board). Multiple studies have looked into the impacts of weather on aviation accidents (Cornman and Carmichael 1993; Kaplan et al. 2005; Lane et al. 2012; Williams 2014) and have found that convection has a major role in turbulence-related aviation accidents. Up to 60 % of turbulence-related accidents were linked to convection (Cornman and Carmichael 1993), making it extremely important for forecasters and models to accurately communicate regions where thunderstorms are located or expected to form within a short time frame to mitigate the impacts of convection on operation teams on the ground and in the air.

Forecasting convection in the near future (within 60 minutes) is often referred to as nowcasting. Traditionally, nowcasting is performed through extrapolation of current observations to predict locations and intensities of storms. These statistical forecasting methods often lack the incorporation of numerical weather prediction (NWP) or physics, leading to rapid performance decay as lead times increase (Sun et al. 2014). By merging these nowcasting techniques and NWP, performance on longer times scales has increased (Browning 1997; Bowler et al. 2006). The NSSL has created the Warn-On-Forecast System (WoFS) to address the limited number of short-term probabilistic forecasting tools available to operational forecasters. The WoFS is an experimental, rapidly updating, high-resolution ensemble model system designed to provide probabilistic forecasts for individual thunderstorms and to increase the warning time for hazardous weather associated with convection (Stensrud et al. 2009, 2013). The WoFS is focused on filling the "watch-to-warning" gap and providing forecasters with probabilistic forecasting data on short temporal and spatial scales. WoFS forecasts have already been used to inform operational products used by the NOAA NWS Storm Prediction Center, Weather Prediction Center, and numerous NWS WFOs (Wilson et al. 2021). Feedback on the WoFS from NOAA's Hazardous

Weather Testbed Spring Forecasting Experiment and Aviation Weather Testbed have also been overwhelmingly positive (Wilson et al. 2021).

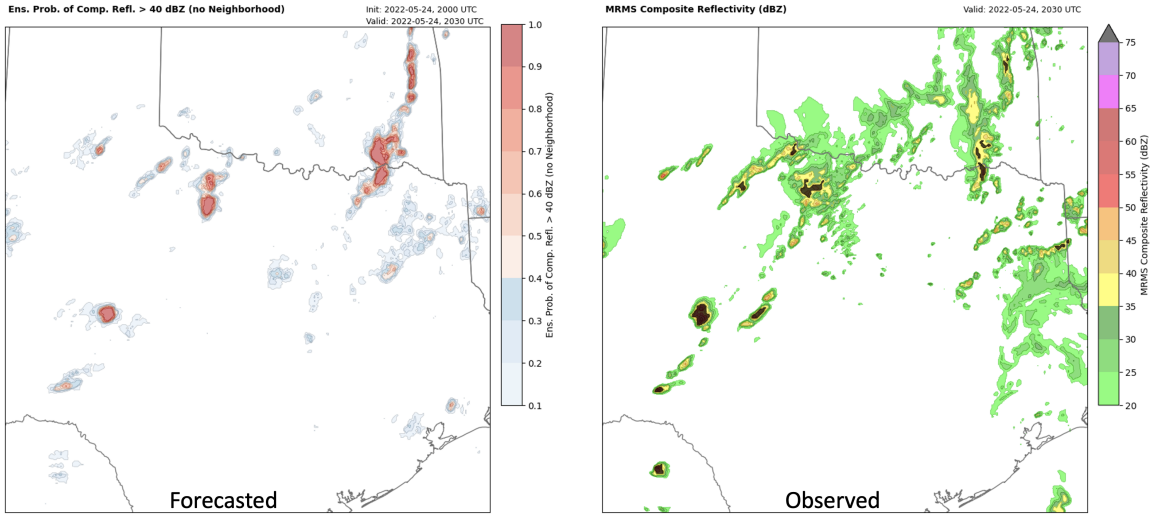

Figure 1.1: Comparison of the WoFS ensemble probability for composite reflectivity $\geq$ 40 dBZ, 30 minutes after initialization (denoted 'Forecasted') and actual MRMS composite reflectivity values (denoted 'Observed') at the forecast valid time. Black blobs are MRMS composite reflectivity values greater than 40 dBZ. The forecast is valid for 24 May 2020, at 2030 UTC, and was initialized at 2000 UTC on 24 May 2020.

Figure 1.1 is an example of output from the WoFS ensemble probability of composite reflectivity $\geq$ 40 dBZ. If a member's predicted reflectivity is $\geq$ 40 dBZ, the member is assigned a 1 and a 0 otherwise. An average across all ensemble members is computed to create a probability field. While the results from the WoFS have been promising, known issues have limited the performance of the system (Guerra et al. 2022). Phase errors, positive frequency bias, and poorly initialized environments are problems currently impacting the WoFS predictions. We believe utilizing a machine learning model to help calibrate the WoFS prediction is one way of addressing these problems.

Machine learning has been used both within the WoFS (Flora et al. 2021; Clark and Loken 2022) and outside of the WoFS (Lagerquist et al. 2020, 2021; McGovern et al. 2017; Cintineo et al. 2020; Gensini et al. 2021; Kotsuki et al. 2019) to increase prediction skill of convection and the associated hazards. The goal of the current project is to improve the WoFS ensemble probability of composite reflectivity ≥ 40 dBZ using a DL model. The WoFS outputs probabilities for reflectivity values ≥ 40 dBZ at each grid point, showing a strong overprediction bias. We hypothesized that using a DL model would lead to an increase in performance. This is due to previous success with utilizing DL models to increase skill for prediction of convection within the nowcasting time-scale (Lagerquist et al. 2021; Han et al. 2020; Li et al. 2023).

The objective of this thesis is to display how the use of a deep learning model can increase the skill of thunderstorm location prediction utilizing the WoFS model output and MRMS composite reflectivity. Also included in this thesis is an outline of the current literature pertaining to this project, which includes nowcasting methods, the WoFS model, and deep learning models for atmospheric prediction. A thorough overview of the data, which includes the steps of gathering, manipulating, and preprocessing data into multiple datasets in that a DL model could extract information. Preprocessing steps include creating a patching scheme, normalization of the data, and splitting the data into training, validation, and testing datasets. Due to the iterative nature of this project, multiple datasets and patching procedures were used to create multiple DL models. Each of these iterative steps improved upon the baseline and fixed issues observed in the previous DL model. Using the best-performing DL model for each iteration, different case studies will be shown to display the performance of the predictions and how they performed against the baseline. Finally, there will be a discussion on the explainability of the DL model and future work. The contributions of this thesis will bring together literature pertaining to the use of DL with convection, compare and contrast different approaches to achieving the goal of

this project, and show the best DL model and various case studies. These case studies will display the strengths and shortcomings of the models along with an analysis of the model's performance.

Chapter 2

Literature Review

2.1 Nowcasting Thunderstorms

Nowcasting techniques for convection have been in use operationally since the 1950s by extrapolating observations (Ligda 1953). Since then, nowcasting using different algorithms and methods has been devised to predict the weather in the short term. According to Wilson et al. (1998), there are two different techniques for extrapolation: steady-state assumption and following the trend of size and intensity. Early iterations of the steady-state assumption would examine two radar images at different times and use cross-correlation to determine the average storm motion (Hilst and Russo 1960; Noel and Fleisher 1960). This assumed no change in storm intensity or overall storm motion. The NSSL began to explore individual cell motion in the 1970s (Barclay and Wilk 1970; Wilk and Gray 1970; Zittel 1976). In this technique, cells were identified through radar images and computed into centroids. An algorithm was then applied to each centroid to determine the velocity at each cell. Dixon and Wiener (1993) developed the TITAN (Thunderstorm Identification, Tracking, Analysis, and Nowcasting) method in 1993. This system was able to match storm objects at different radar scans and forecast the future with some skill. This system was also novel because it could merge and split storm cells. Therefore, creating storm objects is an effective approach for short-term forecasting. In 1998, the SCIT (Storm Cell Identification and Tracking) algorithm was devised using WSR-88D radar information (Johnson et al. 1998). This algorithm identified 68% of storms with reflectivity values greater than 40 dBZ and 96% of storms with reflectivity values greater than 50 dBZ. The

incumbent system was only able to identify 24% and 41%, respectively. The SCIT was accurate as it tracked 90% of all storm cells. While these algorithms were important for determining storm location in the near future, they could not forecast intensity.

Tsonis and Austin (1981) attempted to use the TITAN system to extrapolate storms into the future to determine echo size and intensity for a 30-minute forecast. While the forecast was able to increase the Probability of Detection (POD), it also increased the False Alarm Ratio (FAR) (Wilson et al. 1998). Overall, traditional nowcasting algorithms used current observations to predict locations and intensities of storms. These statistical forecasting methods often lack the incorporation of numerical weather prediction (NWP) or physics, leading to a fast performance decay when lead times increase (Sun et al. 2014).

By merging NWP and nowcasting techniques, the strengths and weaknesses of both approaches should yield an overall improved forecast for all forecasting times (Browning 1997; Bowler et al. 2006). Historically, NWP has struggled with short-term forecasting, while nowcasting has been very skillful within 60-90 minutes. Figure 1 from Sun et al. (2014) displays the increase in skill using a blended method of both NWP and nowcasting. The Nimrod (Nowcasting and Initialization for Modeling Using Regional Observation Data System) system was one of the first attempts at creating a hybrid approach (Golding 1998). The system would use information from both observation and NWP model output for the different analyses, and the forecast used extrapolation of the NWP model's prediction to create a new forecast. Results from the Nimrod system beat both persistence forecasting and raw NWP predictions. More recently, convection-allowing models (CAMs) have been developed and utilized. These models have smaller grid spacing, allowing convective processes to be explicitly represented. These models have performed with greater accuracy than global models, which do not have the spacing which allows convection processes to be represented. While these models are useful in adding skill to thunderstorm forecasting, the models

often have a period of spin-up which result in the model's forecast struggling for nowcasting (Sun et al. 2014). This is due to interpolating the coarse resolution analysis into the finer resolutions seen in the CAM, often called a cold start. While uses of hybrid algorithms have improved the skill of forecasts, there are still problems with short-term forecasting and combining the NWP and nowcasting results in an optimal manner.

To improve on this problem, recent works have used DL models. Lagerquist et al. (2021); Han et al. (2020); Li et al. (2023) have all used deep learning models to nowcast convective storms. Lagerquist et al. (2021) used a U-Net trained on multispectral brightness-temperature images to predict convection out to 120 minutes. This approach outperformed persistence forecasting for lead times $\geq$ 60 minutes and provided a thunderstorm climatology closer to reality for all times. Han et al. (2020) utilized a Convolutional Neural Network (CNN) to incorporate 3-D Doppler radar for short-term forecasting of convective storms. In this case, multiple models were able to improve upon the performance of the baseline. Li et al. (2023) combined GOES-16 geostationary satellite infrared brightness temperature, lightning flashes from the geostationary lightning mapper, and vertically integrated liquid into a U-Net to predict convection and lightning out to 90 minutes. By incorporating the three variables into a DL model, the performance decay characteristic of nowcasting forecasts was delayed, and the forecast of lightning was improved. Overall, nowcasting has progressed from steady-state projections to incorporating observations and NWP predictions into deep learning models. The deep learning models can increase the skill within 60 minutes, which is when the nowcasting algorithms are strong, and past 60 minutes, when the NWP models are strong.

2.2 Warn-on-Forecast System Previous Works

The Warn-on-Forecast System is an experimental model system with the goal of creating rapidly-updating probabilistic ensemble analysis and forecasts on a convective scale (Stensrud et al. 2009). The motivation for this system was initially to increase warning times for tornadoes; then, the goal was expanded to all associated hazards and longer lead times. Currently, warnings associated with thunderstorms are only issued on visual representations or proxy signatures. Through the use of a rapid radar data assimilation system, forecasters could have an operational product to bridge the gap from "watch-to-warning" and increase lead times on severe hazards (Stensrud and Wandishin 2000; Stensrud et al. 2009).

Since the initial paper in 2009, many papers have been published on the WoFS and its advancements. Stensrud et al. (2013) described the progress and challenges of the WoFS; some of the progress included assimilating observations from multiple radars into a single analysis and creating convective-scale forecasts from a model ensemble. Through evaluations at the HWT, it was found that forecasters had greater confidence in issuing warnings when using a model system that used 3DVAR system for data assimilation. Similarly, forecasters felt more confident in warning operations with access to short-range ensemble forecasts. Stensrud et al. (2013) suggested that lead times would increase if forecasters had access to the different products the WoFS would aim to provide.

More recently, Guerra et al. (2022) examined the accuracy of the WoFS by storm age using object-based verification. The WoFS is a rapidly-updating ensemble model system designed for the convective scale, so the ability to assimilate storms quickly and accurately is paramount to the model's performance relative to other NWP systems. The results showed a sharp performance separation between established and newly formed storms. Storms that were an hour old leading into a new initalization of the model scored a probability of detection (POD) of 0.7-0.9. Storms that formed 2-3

hours after initialization of the model scored a POD of 0.3. Therefore, the WoFS is very good at predicting convection already assimilated into the model, while it struggles to predict convection not yet present at initalization.

Traditional machine learning (ML) methods have been successfully applied to the WoFS. Flora et al. (2021) used different traditional machine learning models to predict the overlap of WoFS 30-minute ensemble tracks with tornado, severe hail, or severe wind reports. In this paper, random forest, gradient-boosted trees, and logistics regression algorithms were all trained and tested on different WoFS-based features. Those features were categorized into WoFS intra-storm variables, environmental variables, and morphological attributes. The ML methods increased the prediction of all three hazards against the calibrated, surrogate severe baselines. The biggest increase was observed in severe wind prediction, where the Brier skill score was almost double that of the baseline prediction.

Clark and Loken (2022) also used an ML method with the WoFS to increase the skill of the WoFS forecast. In this paper, a random forest model utilized the WoFS model output of both intra-storm variables and environmental variables to predict severe weather probabilities at 0-3 hour lead times. The ML model created reliable probabilities and significantly outperformed the baseline prediction. Results from the model also displayed that intra-storm variables were far more important than the environmental variable in short-term severe weather prediction. They found that including the smoothed UH fields from each ensemble member improved predictions. Clark and Loken (2022) and Flora et al. (2021) displayed that using traditional machine learning methods on the WoFS output can lead to significant gains in performance for predicting severe weather and the associated hazards.

2.3 Deep Learning and the Atmosphere

Traditional machine learning and deep learning models have become increasingly popular in recent years. According to Figure 1 in Chase et al. (2022a), machine learning meteorology papers have been quickly increasing since 2015, and papers using deep learning methods have increased exponentially since 2019. Deep learning has been very successful in predicting convection and the associated hazards. Other than the papers described above (Lagerquist et al. 2021; Han et al. 2020; Li et al. 2023), many other papers have been published showing the performance of deep learning models predicting convection. In Lagerquist et al. (2020), a DL model was trained on radar images and proximity soundings from the Rapid Refresh model to predict tornadoes over the next hour. The performance of the DL model was comparable to that of an operational ML model (ProbSevere) for predicting severe weather. The largest increase in performance was reflected in EF2+ tornadoes.

DL models have also been useful in spatial analysis of severe hailstorms (Gagne et al. 2019). A CNN was trained on patches from the NCAR convection-allowing ensemble model output. These patches consisted of upper air data and thermodynamic variables. The model aimed to predict the probability of severe hail in the domain. This paper found that compared to traditional statistical approaches, the DL model could make predictions with more skill and sharper probabilistic predictions. The DL model also gave insight into storm structures, storm environments, and storm modes that produce severe hail. For severe hail to form, storms usually have strong lapse rates, directional wind shear, and seeding from graupel and small hail stones in weaker updrafts. Storms that produced large hail also were twice as likely to be a supercell thunderstorm as a QLCS thunderstorm.

On a larger spatial and time scale, DL models have also successfully predicted synoptic-scale fronts (Lagerquist et al. 2019; Justin et al. 2022). Using deep learning trained on different state variables at a variety of levels, the DL model was able to

predict where WPC forecasters would draw fronts. This method significantly outperformed the baseline of NWP frontal analysis by scoring a CSI score of 0.52 (Lagerquist et al. 2019). This project aimed to create a method of forming datasets and climatologies of fronts for research work. In the current iteration of the project, warm, cold, stationary, and occluded fronts are now plotted. On a 250 km neighborhood, the warm and cold front binary classification significantly outperformed the baseline and will serve as a first guess for forecasters to identify frontal boundaries more efficiently (Justin et al. 2022).

Overall, deep learning models have become very popular and successful in recent years (Chase et al. 2022b). These models have been able to successfully predict and outperform the baseline on a wide variety of atmospheric phenomena at many different spatial and temporal scales (Chase et al. 2022b; Lagerquist et al. 2021; Han et al. 2020; Li et al. 2023; Lagerquist et al. 2020; Justin et al. 2022; Lagerquist et al. 2019). Given previous successful ML applications and combining methods from similar works (Flora et al. 2021; Clark and Loken 2022; Lagerquist et al. 2021), we hypothesize that a DL model trained on the WoFS ensemble output and MRMS composite reflectivity will outperform the current WoFS forecast for reflectivity values exceeding 40 dBZ 30 minutes after initialization of the WoFS.

Chapter 3

Methods and Data

3.1 Warn-on-Forecast System Specifications

The WoFS is a regional, 3-km ensemble analysis and forecast system designed to produce rapidly updating probabilistic guidance for severe weather. The system comprises 18 forecast members (36 analysis) using the Weather and Research Forecast Model (WRF-ARW; Skamarock (2008)) as the dynamic core. Each member has a different physical parameterization provided in Skinner et al. (2018), Table 1. Initial and boundary conditions for the WoFS are provided by the experimental 3-km High-Resolution Rapid Refresh Ensemble (HRRRE; (Dowell et al. 2016)). The WoFS domain size for 2017-2019 was 750 $\times$ 750 km, but from 2019-current, it is 900 $\times$ 900 km. The WoFS domain is repositioned over the region where and when the greatest severe weather threat is anticipated. Forecasts are mostly initiated during the warm season, including during NOAA's Hazardous Weather Testbed Spring Forecasting Experiment (Gallo et al. 2017, 2022; Clark et al. 2022). Since 2021, the WoFS has been run more regularly and outside of the warm season when severe convection is expected. The WoFS assimilates radial velocity, radar reflectivity, Geostationary Operational Environmental Satellite (GOES-16) cloud water path, and, when available, Oklahoma Mesonet observations every 15 minutes with conventional observations assimilated hourly. Forecasts are initialized every 30 minutes with output every five

minutes out to three or six hours of lead time. The WoFS is expected to be incorporated into operations within the United Forecast System between 2025 and 2030. For more details about the WoFS, see Miller et al. (2022).[1]

3.2 U-Nets

U-Nets are a type of convolutional neural network (CNNs), initially created to segment medical images (Ronneberger et al. 2015). Since then, U-Nets have been widely used in many different applications with success across a variety of applications (John Saida and Ari 2022; Pan et al. 2020; Zhang et al. 2021).

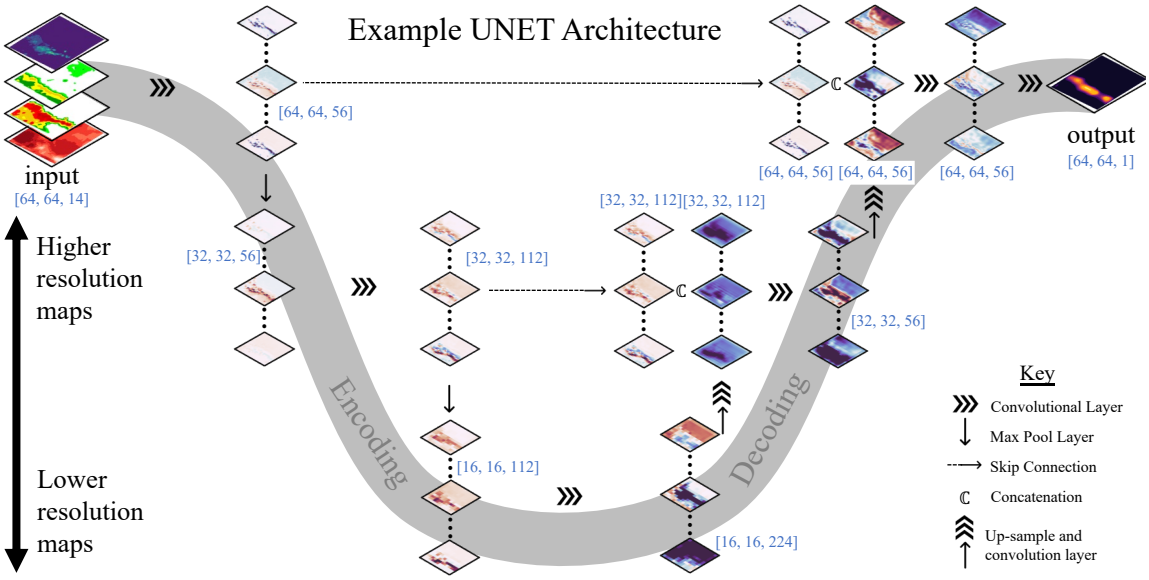

Figure 3.1: Example architecture of a U-Net that predicts reflectivity values $\geq$ 40 dBZ. A corresponding arrow indicates the different layers; the legend is displayed in the lower right corner. Dimensions for each layer are displayed in [latitude, longitude, channel]. The center images are convolved images. Only three are shown for space reasons. The exact architecture is displayed in table 3.3. Figure adapted, with permission, from Chase et al., 2022.

[1]These are the WoFS parameters valid for the dataset used herein; future specifications may be changed slightly.

Figure 3.1 displays an example U-Net architecture for predicting composite reflectivity $\geq$ 40 dBZ. U-Nets are a form of image-to-image model where they ingest one type of image and output other images. In this case, we are able to input composite reflectivity, CAPE and CIN, and vertical velocity. The U-Net outputs a prediction of reflectivity values $\geq$ 40 dBZ.

CNNs (and U-Nets) process images and identify different features within the images by utilizing convolutional layers (LeCun et al. 1989). These layers contain convolution kernels (also referred to as convolutional filters). Each convolutional layer is comprised of many convolutional kernels, with each kernel having the possibility to identify different features within the image. Each kernel is made up of an array of different learned weights that are applied to each of the input fields to create a convolved image. The different weights will allow for recognizing different features (e.g., straight lines, edges). Pooling layers are often interlaced in the convolutional layers. These layers reduce the resolution of the image and will allow subsequent convolutional layers to extract larger-scale features (colors, shapes, etc.). Therefore, utilizing a network with multiple convolutional and pooling layers can result in the extraction of complex features [e.g., fronts (Lagerquist et al. 2019), hail (Gagne et al. 2019)]. At the bottom of the convolutional network, the output from the previous layer is flattened into a one-dimensional vector and fed into an artificial neural network (ANN) where a prediction is computed. This prediction can be a regression or classification, but for this project's scope, only classification problems will be discussed. For calcification problems, a CNN outputs the probability that the goal feature is present in the image. An example of a desired feature could be whether there is a hook echo somewhere in a radar image. This can be a limitation for meteorologists since no spatial information is communicated through a single output prediction for the entire domain.

U-nets, unlike CNNs, are a form of autoencoder, where they take input at one resolution, encode it into a lower resolution, and then output back at the original

resolution. In a U-net (Figure 3.1), you can see the convolutional filters and pooling layers used to both downscale and upscale. The final output represents the probability of the target being present at each point in the original image; in this case, reflectivity values $\geq$ 40 dBZ. These probabilities are vital to understanding a model's confidence in predictions and provide spatial information to the predictions. Overall, by utilizing a U-Net model, we can use image-to-image translation to give the model different meteorological data to receive a map of probabilities in the same dimension as the input data for thunderstorms 30 minutes after the initialization of the WoFS. For a more detailed explanation and discussion of deep learning and U-Nets in meteorology, refer to Chase et al. (2023).

3.3 Dataset

The data primarily used to train, evaluate, and test the DL models was extracted from the WoFS ensemble model output valid 30 minutes after initialization. Variables from the WoFS output included both environmental and intra-storm variables shown in Table 3.1. This short list of variables were selected due to their connections to thunderstorm environments and to ensure the model would remain explainable by restricting dimensionality.

Due to the differences in ensemble members' variance and spread, environmental and intra-storm variables had different statistics computed. For intra-storm variables, ensemble average, 90th percentile, and ensemble maximum were computed. The 10th percentile and minimum values were used for downdraft due to the negative values indicating greater intensity. Since intra-storm variables are nearly zero in most of the domain they heavily skew spatial distributions. Therefore, we relied on higher percentile statistics to properly characterize these variables. Using the three different ensemble statistics for the intra-storm variables provides the model with more information about ensemble spread and confidence. For environmental

Variable Family	Variable Name
Intra-storm	Ensemble Maximum Composite Reflectivity
Intra-storm	Ensemble 90th Percentile Composite Reflectivity
Intra-storm	Ensemble Average Composite Reflectivity
Intra-storm	Ensemble Maximum Updraft Velocity
Intra-storm	Ensemble 90th Percentile Updraft Velocity
Intra-storm	Ensemble Average Updraft Velocity
Intra-storm	Ensemble Minimum Downdraft Velocity
Intra-storm	Ensemble 10th Percentile Downdraft Velocity
Intra-storm	Ensemble Average Downdraft Velocity
Environmental	Ensemble Average ML CAPE
Environmental	Ensemble Average SFC CAPE
Environmental	Ensemble Average ML CIN
Environmental	Ensemble Average SFC CIN
Other	MRMS Composite Reflectivity

Table 3.1: Data input to each of the DL models. Intra-storm and environmental variables are valid 30 minutes after the initialization of the WoFS model. MRMS composite reflectivity is valid at the time of the WoFS initialization.

variables, only ensemble averages were computed. Environmental variables are more Gaussian, and therefore their average values are more meaningful. Accompanying the WoFS model output, Multi-Radar/Multi-Sensor (MRMS) composite reflectivity (Smith et al. 2016) values were included at the initialization time of the WoFS model. Data used in the project were collected from 2017 - 2021, with a majority of the runs occurring during NOAA's HWT SFE.

Finally, the input data was scaled using min-max normalization (equation 3.1).

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

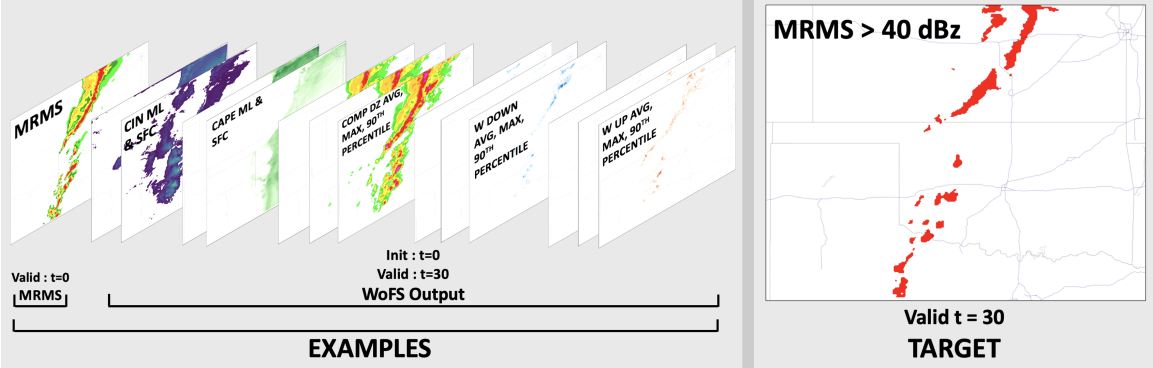

Figure 3.2: Data layout for the deep learning model. The examples are the input data into the DL model. MRMS composite reflectivity is valid at t=0 min. The WoFS model output is valid at t = 30 min. The targets are the prediction goal of the DL model. They are MRMS composite reflectivity at t = 30 min binarized on a 40 dBZ threshold.

Each variable is scaled by the minimum and maximum value determined from the entire dataset. Normalization is needed due to the variations in magnitudes between variables. For example, a DL model could struggle to find the optimal weights to apply to the variables since CAPE and CIN are on the magnitude of 10^3 while reflectivity values within thunderstorms are only on the magnitude of 10^1.

The target variable was MRMS composite reflectivity $\geq$ 40 dBZ 30 minutes after WoFS initialization. We chose a 40 dBZ threshold as it often separates weaker convection from moderately intense convection. The base rate for targets in the dataset is 0.006%. Figure 3.2 shows how the data is configured for each event and then fed into the DL model for training, validation, and testing.

3.4 Patching Scheme

Creating data patches out of the domain has multiple benefits compared to utilizing the WoFS domain for each event. Firstly, creating patches decreases the memory burden and increases learning efficiency. Secondly, breaking each time step in an

event into multiple patches allows for more samples to be created. This creates more training, validation, and testing data. Thirdly, input data when working with U-Nets should be in powers of two-size grid-point squares (32 x 32, 64 x 64, 128 x 128, etc.). This is due to the max pooling layers used in this model. Max pooling layers will halve the data every time the data passes through the layer. Multiple patching sizes were tested and the best performance was found at a patch size of 64 x 64. An algorithm was devised to aggregate the WoFS output and MRMS data were temporarily aligned based on initialization date and time. If all data was present for the given individual date and time, multiple non-overlapping patches were created with all relevant input WoFS data, input and target MRMS data, initialization date and time, and latitude and longitude. Overlapping patches were tried in the training dataset, but the performance was sub-optimal due to the reduction in mutual independence in each patch. Events for the dataset were defined as the total time the WoFS was running for a convective event. Often this was an eight-hour period (from 19 UTC to 03 UTC), where the WoFS initializes every 30 minutes.

3.4.1 Initial Patching Scheme

Patches were created from all WoFS cases from 2017-2021. Two patching schemes were developed since the overall WoFS domain size changed in 2019. From 2017-2018, the domain size was 250 x 250 grid points with a 3km grid spacing. To create the maximum number of patches possible, we added a 3-grid point border of zero padding to the domain to create a 256 x 256 grid point domain. Therefore each event time step had a 4x4 quilt of patches with the size of 64 x 64 grid points. For the cases from 2019-2021, the domain size was 300 x 300 grid points with a 3km grid spacing. In order to get the domain to split easily into a grid of 64 x 64 grid point patches, a 10 grid point border of zero padding was appended, making the domain 320 x 320 grid points. This allowed for a quilt of 5x5 patches at a size of 64 x 64 grid

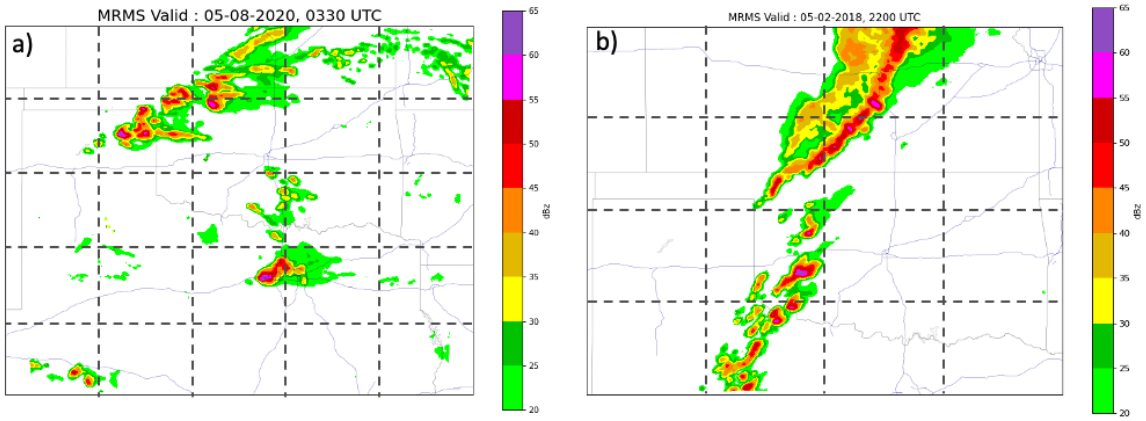

Figure 3.3: Example of the initial patching scheme. a) Patching method for events from 2019-2021 with zero padding of 10 grid points on the border. b) Patching method for events from 2017-2018 with a zero padding of 3 grid points on the border. The dashed lines represent the border of the patches.

points to be extracted from each event time step. Figure 3.3 displays how the initial patching scheme was executed. Through this scheme, a total of 41,132 patches were created from 2017-2021. The training, validation, and testing split was by time with 80%/10%/10% (32,906/4,113/4,113) partitions, respectively.

3.4.2 Current Patching Scheme

During experiments with the previous patching scheme (section 3.4.1), a few problems became evident, leading to the need for a revised patching scheme. The first problem was due to the way the patching was split between the training, validation, and testing datasets. In the initial scheme, all the patches were in one dataset and split based on the total size. This way could not ensure the independence of the training, validation, and testing datasets. This most likely resulted in one event being in both the training and validation datasets and another event being in the validation and testing datasets. The second problem arose with the zero padding on the domain border. While the cases from 2017-2018 with a smaller zero padding border likely minimally impacted

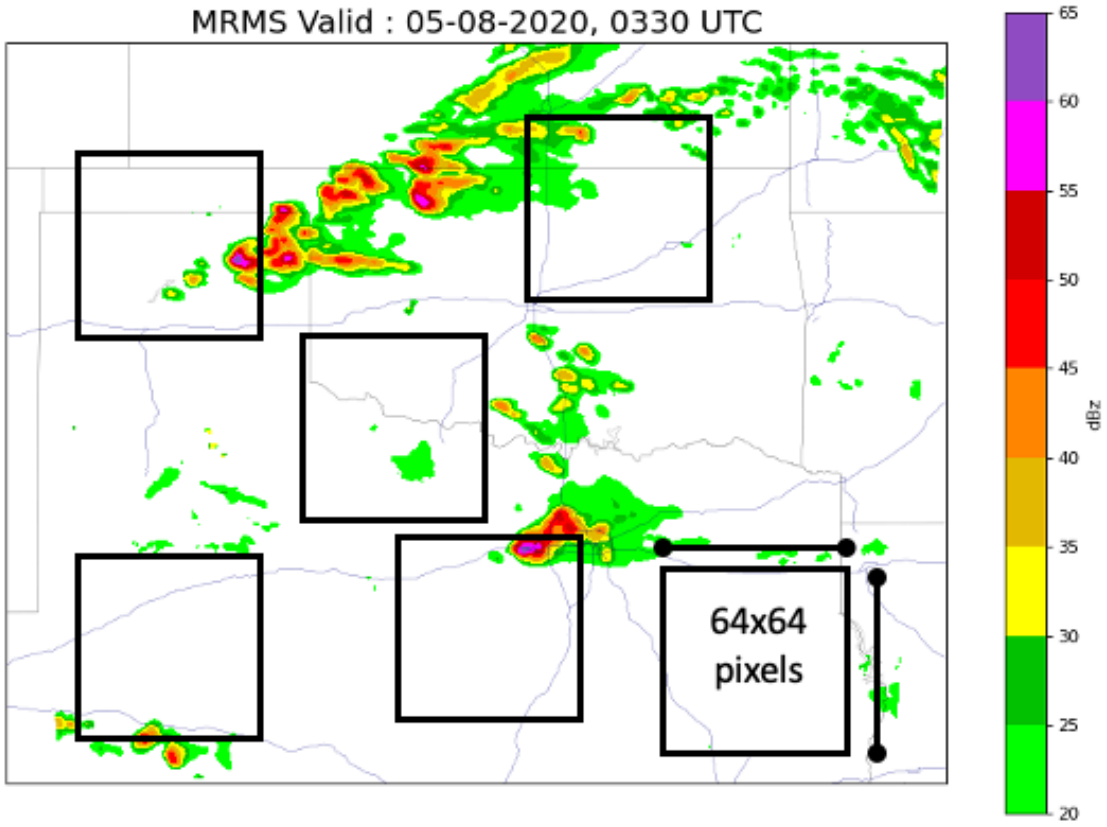

Figure 3.4: Example of the current patching scheme. Multiple patches were created out of the domain, with the location of the patches selected at random. All patches were 64 x 64 grid points.

the model, we believe that the larger zero padding border on the patches from 2019-2021 caused confusion in the model by adding in zero values in areas that could have been adjacent to strong reflectivity signals. For these two reasons, a new patching scheme was needed. This scheme would guarantee independence between the datasets and create a dataset more representative of reality.

In this patching scheme, each time step of each event had multiple patches created out of the domain. The patch location was randomly selected and did not allow any overlapping patches. Figure 3.4 displays an example of how the patches may be selected. Overall, this patching scheme produced 23,061 patches which were extracted from 2017 - 2021 events. The training dataset consisted of patches from 2017-2019, the

first half of 2020, and the first half of 2021. Therefore, validation and testing datasets were composed of the second half of 2020 and 2021, respectively. Patches from the same event were restricted to one of the datasets to ensure independence. This resulted in a training/validation/testing split of 73%/15%/12% (16,935/3,473/2,653).

3.5 Deep Learning Model Methods

For this study, we adopted a U-Net deep learning model (Section 3.2). Previous research has shown that the choice of loss equation can greatly impact the performance of the DL model (Ebert-Uphoff et al. 2021). This led us to test and evaluate multiple models with different loss equations. Both pixel- and spatial-based loss functions were tested. The Weighted Binary Cross-Entropy (WBC) loss function (equation 3.2) applies weights to the different classes in the binary cross-entropy equation. These weights are designed to penalize poor predictions of events more than poor predictions of non-events. When dealing with rare events, equally penalizing poor predictions of events and non-events results in a DL model skewed towards predicting low probabilities. The idea for the WBC loss function was derived from the weighted mean squared error function from Ebert-Uphoff et al. (2021):

$$Loss_{WBC} = -\frac{1}{N}\sum_{i=0}^{N} weight[Y_{i,Truth} log(Y_{i,Pred}) + (1 - Y_{i,Truth}) log(1 - Y_{i,Pred})] \quad (3.2)$$

where

$$weight = \begin{cases} 1 & \text{if } Y_{i,Truth} = False \\ w & \text{if } Y_{i,Truth} = True \end{cases} \quad (3.3)$$

In equation 3.2, the 'w' represents the different weights that were tested in the hyperparameter search. They ranged from 1.0 to 4.0. This loss function was able to produce the best models at the pixel evaluation level and was static for the remainder of the model search with regard to pixel-based models.

A spatial loss function was also tested and evaluated on the dataset described in section 3.4.1. This loss function was the Fractions Skill Score (FSS) (Roberts and Lean 2008). FSS evaluates a prediction on a neighborhood spatial size, in this case, predictions within 5 grid points (i.e., 15 km). The benefit of a spatial loss function is avoiding the double penalty effect. In a pixel-based loss function, if a prediction is one grid point from being a hit, the model is punished for two wrong predictions, one for the false alarm and one for the false negative. The drawback of the FSS loss function is that predictions are not rewarded for being correct at finer spatial scales. FSS loss functions can be very useful in problems where small phase errors can be tolerated. The equation for FSS and the derivation of the loss function is described in equation 3.4 (Roberts and Lean 2008).

$$FSS = 1 - \frac{FBS}{FBS_{worst}} \tag{3.4}$$

FBS is Fraction Brier Score and FBS_{worst} is valid for the worst possible forecast. FBS and FBS_{worst} are defined in equations 3.5 and 3.6. F_{pred} refers to the forecasted fraction, and F_{true} refers to the true fraction.

$$FBS = \frac{1}{N_{lat}} \frac{1}{N_{lon}} \sum_{lat}^{N_{lat}} \sum_{lon}^{N_{lon}} [F_{pred}(lat, lon) - F_{true}(lat, lon)]^2 \tag{3.5}$$

$$FBS_{worst} = \frac{1}{N_{lat}} \frac{1}{N_{lon}} [\sum_{lat}^{N_{lat}} \sum_{lon}^{N_{lon}} F_{pred}^2(lat, lon) + \sum_{lat}^{N_{lat}} \sum_{lon}^{N_{lon}} F_{true}^2(lat, lon)] \tag{3.6}$$

A loss function can then be made through the formula:

$$FSS_{loss} = 1 - FSS \tag{3.7}$$

A perfect loss score for the FSS loss function is 0. For more information regarding the FSS loss function, refer to Ebert-Uphoff et al. (2021) and Justin et al. (2022).

3.5.1 Initial Model Parameters

In the first iteration of this project, two separate models were trained, evaluated, and tested on the dataset from section 3.4.1. The first model was a pixel-based loss

function utilizing the WBC loss function, and the second model was a spatial-based loss function using the FSS loss function with a 15 km neighborhood.

U-Net Hyperparameters		
Loss Function	Weighted Binary Cross-Entropy	Fractional Skill Score
Loss Weights	4.0 to 1.0	n/a
Convolution Layers per Step	1	1
Convolution Kernel Size	7	5
Kernel Size	8	4
Activation Function	Leaky ReLU	ELU
U-Net Depth	3	2
Optimizer	adam	adam
Learning Rate	0.001	0.001
Batch Size	64	128
Neighborhood Size	n/a	15 km
Output Activation Function	Sigmoid	Sigmoid

Table 3.2: Best performing models for the pixel- and spatial-based loss functions. Both models were trained, evaluated, and tested on data from section 3.4.1.

Hundreds of models were trained, evaluated, and tested through a random hyperparameter search for both loss functions. Models were evaluated based on the Maximum Critical Success Index (maxCSI, described in Section 4.1) and loss score on the validation dataset. Models with the best performance scores and that produced stable loss diagrams were selected. The WBC model had a validation maxCSI of 0.1636 and a loss score of 0.2258 after 100 epochs. For the FSS model, the validation maxCSI was only around 0.0084, while the loss score was 0.022 after 100

epochs. This lower maxCSI score is understandable, given that CSI evaluates predictions on the pixel scale while the FSS model is evaluated on a 15km scale. For the post-processing results, the FSS model is evaluated using the FSS metric.

3.5.2 Current Model Parameters

U-Net Hyperparameters	
Loss Function	Weighted Binary Cross-Entropy
Loss Weights	2.0 to 1.0
Convolution Layers per Step	1
Convolution Kernel Size	5
Kernel Size	8
Activation Function	Leaky ReLU
U-Net Depth	4
Optimizer	rsmprop
Learning Rate	0.01
Batch Size	256
Output Activation Function	Sigmoid

Table 3.3: Hyperparameters selected after a gridded hyperparameters search over 100 separate models. After 100 epochs, validation maxCSI reached 0.20.

Once WBC was selected as the best loss function, a hyperparameter search was conducted wherein 100s of models were created, trained, and evaluated on the dataset using the new patching scheme (section 3.4.2). The best-performing model was defined as the model which achieved the highest maxCSI and the lowest loss score on the validation dataset. Table 3.3 displays the hyperparameters of the best-performing model. All hyperparameters were varied for this run, with the exception of the loss function and activation function. L2 regularization was also included in the models,

and a noticeable jump was observed across both validation and testing metrics when turned on (not shown). L2 regularization had a value of 0.001. The model displayed in Table 3.3 achieved a maxCSI of 0.20, while the validation loss was 0.16 after 100 epochs. This model outperformed the best-performing baseline significantly, which we believe justifies using a DL model to improve the WoFS composite reflectivity product.

Chapter 4

Results

4.1 Verification Metrics

Several verification metrics were used to evaluate the DL model and baseline predictions. Those metrics include the maxCSI, Probability of Detection (POD), and False Alarm Ratio (FAR). Each of these metrics can be used to understand a model's overall performance, discrimination, and skill. MaxCSI is the highest CSI score produced when calculating the score across many probability thresholds from 0-1. The formula for CSI is $TP/(TP+FN+FP)$, where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives. A maxCSI score of 1 indicates a perfect forecast.

POD and FAR are often shown together. POD is defined as $TP/(TP+FN)$. For POD, a perfect score is 1, meaning all forecasted positives lined up with observed positives. FAR is defined as $FP/(TP+FP)$. A perfect score for FAR is 0, meaning there were no false alarms. The success ratio (SR) can also describe the FAR; it is defined as 1 - FAR. Using POD and FAR in tandem will show the models's discrimination. Finding the right balance between maximizing POD and limiting FAR is important to creating a discriminative model. For example, POD can be inflated by lowering the probability threshold so that every hit case is correct, but this will result in a high FAR. Determining the correct balance between these two metrics varies on a case-to-case basis. In the case of tornadoes, a higher FAR may be tolerated due to the high impact of a missed tornado. In this project, finding the maxCSI determined

the best model, and POD and FAR were used to compare the performance of the WoFS baseline and the DL model.

The best model was determined through a performance diagram and a reliability diagram. In the performance diagrams herein, the bold 'X' denotes the maxCSI. The goal is to maximize this number and keep the 'X' close to the 1.0 bias line. A reliable model follows the dotted goal line.

Brier Score was used to evaluate the DL model's predictions. The formula for the BS is shown in equation 4.1:

$$BS = \frac{1}{n}\sum_{i=0}^{n}(P_{i,pred} - Y_{i,Truth})^2 \tag{4.1}$$

where

$$Y_i, Truth = \begin{cases} 0 & \text{if } Y_{i,Truth} = False \\ 1 & \text{if } Y_{i,Truth} = True \end{cases} \tag{4.2}$$

$P_{i,pred}$ represents the prediction probability of the pixel being $\geq$ 40 dBZ. BS ranges between 0 and 1, where 0 is a perfect score, and 1 is no skill. For the BS calculation, we excluded all points where the observation and predictions were zero to highlight correct forecast events better. [1].

4.2 WoFS Baseline

We used WoFS grid-scale ensemble probability of composite reflectivity exceeding 40 dBZ as our comparison baseline. Creating and testing multiple baselines was essential to ensure the DL model was substantially more skillful than any much simpler (and easier to develop) model. For a simple baseline, we computed the WoFS grid-scale ensemble probability of composite reflectivity exceeding a threshold (Schwartz and

[1]For more information about the verification metrics, see https://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts or https://glossary.ametsoc.org/wiki/Skill

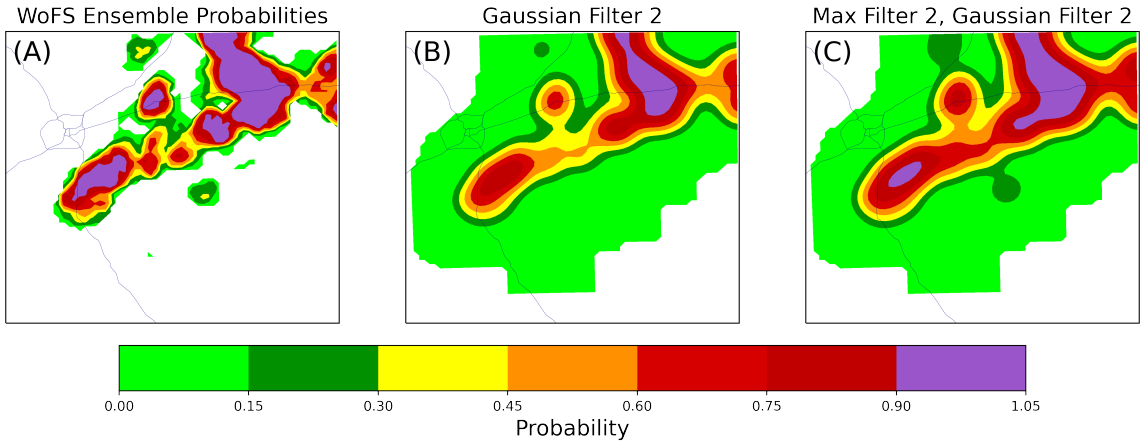

Figure 4.1: Examples of WoFS ensemble probability of composite reflectivity $\geq$ 40 dBZ with different post-processing applied.(a) grid-scale ensemble probability, (b) grid-scale ensemble probabilities with a two-grid point radius Gaussian filter, and (c) grid-scale ensemble probability with a two-grid point radius maximum filter and then smoothed with a two-point Gaussian filter.

Sobash 2017; Flora et al. 2021). To calibrate the probabilities, we adopted the simple framework of introducing maximum value and Gaussian filters, similar to other surrogate severe studies (Clark and Loken 2022; Loken et al. 2020; Sobash et al. 2016). Examples of the different baselines are shown in Fig. 4.1. The initial baseline (Figure 4.1a) is the grid-scale ensemble probability of composite reflectivity $\geq$ 40 dBZ. Ensemble probabilities with different filters (Gaussian and max filters) are shown in Figures 4.1b and 4.1c. A Gaussian filter with a grid point radius of two was applied to the probabilities (Figure 4.1b). By applying this filter to the probabilities, noise is reduced, and probabilities are spread out over a larger area. In the third approach (Figure 4.1c), before making the dataset binary, a max filter with a radius of 5 pixels was applied to the reflectivity values. The max filter applies an iterative five-by-five box wherein every grid point within the box is assigned to the maximum value within the box. Max filters are applied to reduce noise and spread the largest probability to the surrounding pixels. The performance of the baselines is shown below.

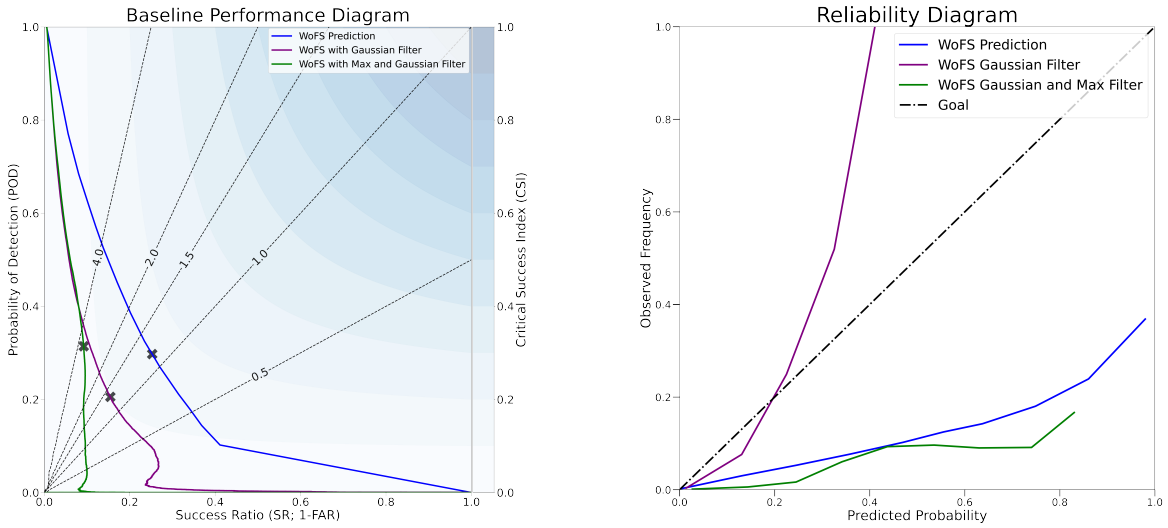

Figure 4.2: Performance diagram comparing each of the baselines. (left) The WoFS prediction (blue solid line), the WoFS prediction with a Gaussian Filter (purple), and the WoFS prediction with a Gaussian and Max Filter (green solid line). The X represents the maxCSI values for the corresponding model. The maxCSI was 0.16, 0.10, and 0.08 for the WoFS prediction, WoFS prediction with a Gaussian Filter, and WoFS prediction with a Gaussian and Max Filter, respectively. (right) Reliability diagram comparing each of the WoFS baselines. The WoFS prediction (blue solid line), the WoFS prediction with a Gaussian Filter (purple solid line), and the WoFS prediction with a Gaussian and Max Filter (green solid line). The black dashed line is the ideal line for a model's prediction.

The baselines were evaluated on the validation dataset using a reliability diagram, performance diagram, and maxCSI from the dataset with the current patching scheme (section 3.4.2). Figure 4.2 shows the performance and reliability diagrams for the three baselines. The WoFS predictions and the WoFS predictions with a max filter and Gaussian filter applied showed over-prediction biases. With just a Gaussian filter applied, the prediction showed an underprediction bias. The maxCSI scores for WoFS ensemble probabilities, WoFS prediction with a Gaussian filter, and WoFS prediction with a Gaussian and max filter applied were 0.16, 0.10, and 0.08, respectively. Based

on the performance of the three baselines, the WoFS ensemble prediction will be the baseline going forward for the iteration of models.

4.3 Initial Deep Learning Performance

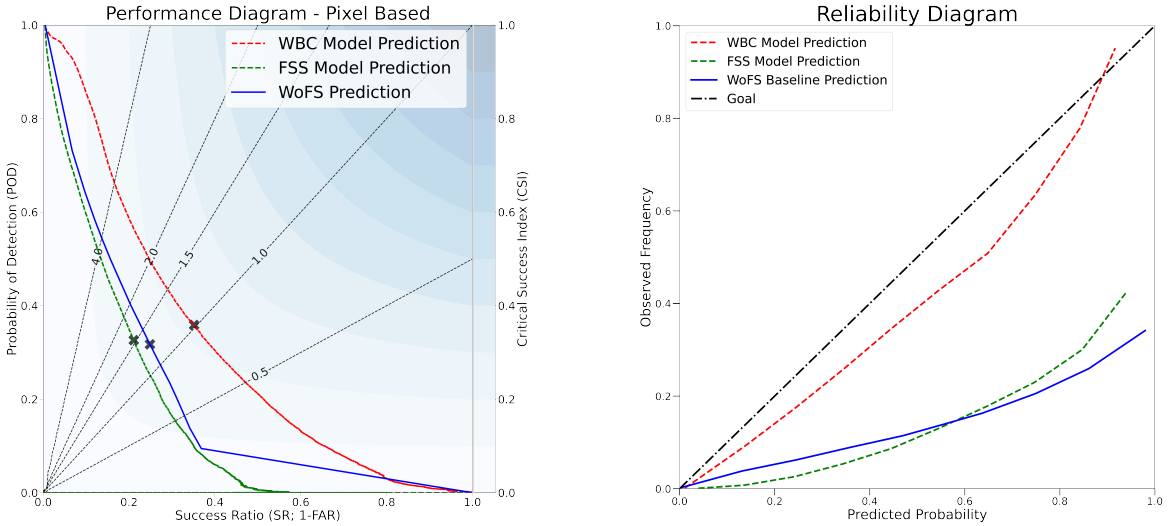

Figure 4.3: (left) Performance diagram comparing the WBC model (red dashed line), the FSS model (green dashed), and WoFS prediction (blue solid line). The X represents the maxCSI values for the corresponding model. The WBC model's maxCSI was 0.22; for the FSS model, the maxCSI was 0.15; and for the WoFS prediction, the maxCSI was 0.16. (right) Reliability diagram comparing the WBC model (red dashed line), the FSS model (green dashed), and WoFS prediction (blue solid line). The black dashed line is the ideal line for a model's prediction.

The DL models were trained and evaluated on the dataset from section 3.4.1. Due to the differing nature of the pixel- (WBC) and neighborhood-based (FSS) loss functions, two different verification metrics were performed. The first performance metric was evaluated at the pixel scale, and the second was evaluated on a 15km neighborhood. Figure 4.3 displays the pixel-based evaluation. In the performance diagram, the WBC model's predictions substantially outperformed the WoFS and

FSS model's predictions. The maxCSI scores were 0.22, 0.15, and 0.16 for the WBC model, FSS model, and WoFS model, respectively. This jump in performance shows that the WBC model has been able to increase discrimination and skill compared to the baseline. The bias score also shows the difference in performance. In the WoFS and FSS predictions, the bias scores are close to 1.5, while the WBC model has a bias score that is very close to 1.0. This is a notable improvement since the WoFS predictions have an over-prediction bias (Guerra et al. 2022). Turning to the reliability diagram, the reliability has also increased at the pixel-based evaluation; the WBC model prediction lies near the goal line for all predicted probabilities. The WoFS and FSS predictions continue to show the over-prediction bias observed in the performance diagram.

The overprediction bias in the FSS model was expected due to the properties of the FSS loss function. Using this loss function, the model attempts to predict on a 15km scale while the model is evaluated on a 3km scale. This motivated evaluation of the FSS model on the same scale it was trained on.

Figure 4.4 displays the performance and reliability diagrams with the labels increased to a 15km neighborhood. To create a 15km neighborhood, a max filter with a radius of 5 grid points was applied, then made binary on a 40 dBZ threshold. At this scale, a fair assessment of the FSS model can be made. In the performance diagram, the WBC model appears to be performing better than the WoFS and FSS prediction. The maxCSI score for the WBC model, FSS model, and WoFS model was 0.19, 0.15, and 0.15, respectively. The most notable change in the neighborhood evaluation versus the pixel-based evaluation is the frequency bias of the FSS model and WoFS model. In the FSS model, the bias score is 1.0 showing that the model frequency bias is correct when evaluated at the scale it was trained on. The WoFS bias is now below 1.0, showing an under-prediction bias. This was also expected since the WoFS is making predictions on the grid scale and is now being evaluated on a 15 km scale.

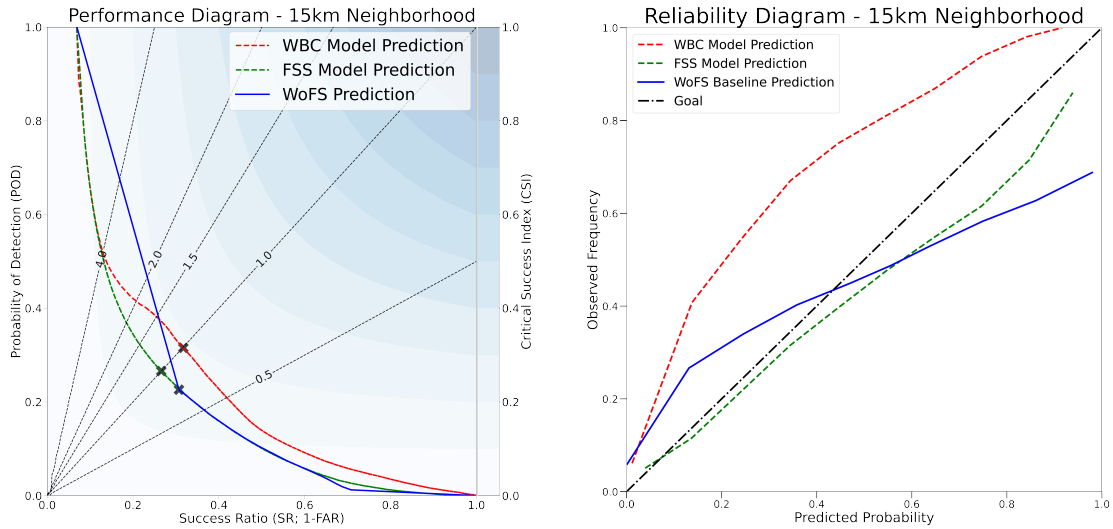

Figure 4.4: Performance (left) and reliability (right) diagrams comparing the WBC model (red dashed line), the FSS model (green dashed), and WoFS prediction (blue solid line). a) Performance diagram comparing the WBC model (red dashed line), the FSS model (green dashed), and WoFS prediction (blue solid line). The X represents the maxCSI values for the corresponding model. b)Reliability diagram comparing the WBC model (red dashed line), the FSS model (green dashed), and WoFS prediction (blue solid line). The black dashed line is the ideal line for a model's prediction.

The reliability diagram is also different when evaluated on the 15 km neighborhood. The most encouraging change in performance is the FSS model. The reliability of the FSS model is very much improved on this scale compared to the reliability diagram in figure 4.3. The FSS model is shown to be close to the goal line, while in figure 4.4, it showed a strong over-prediction bias. While we expected the WoFS model to show an under-prediction bias for all predicted probabilities, it showed an under-prediction bias in the lower probabilities and an over-prediction bias in the high probabilities. This gives insight into the over-predictive nature of the WoFS. Even at a larger scale, the WoFS predictions are still over-predicting at high probabilities. The WBC model's under-prediction bias is expected as it is trained on a 3km grid and is now being evaluated at a larger scale.

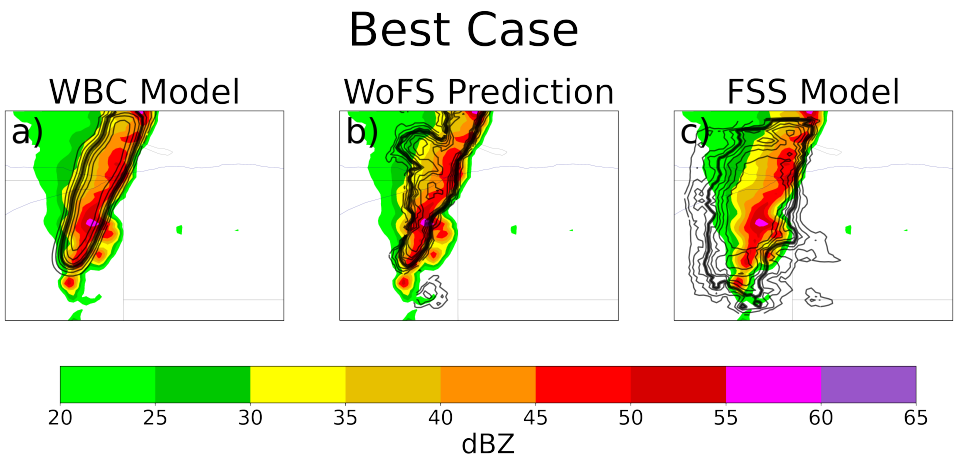

Figure 4.5: Case study displaying the best case wherein all three models create good predictions. a) WBC model's prediction. b) WoFS model's prediction. c) FSS model's prediction. In all three panels, the radar objects are the MRMS at t=30min. The black contours are the prediction probabilities in intervals of 10%. The bolded black line represents the 50% probabilities. All models are attempting to predict reflectivity values $\geq$ 40 dBZ.

4.3.1 Case Studies - Initial Results

The three case studies explored in this subsection were selected based on CSI and FSS scores. The first case study will show a case where all three models performed well. The second case will display a case where the WBC model's predictions outperform both the WoFS and FSS models' predictions. The last case exhibits a case where the FSS model's predictions outperform the WBC and WoFS models' predictions. In all cases, a 50% probability is the probability threshold for metrics that require a binary threshold.

4.3.1.1 Best Case

In Figure 4.5, there is an ongoing quasi-linear convective system (QLCS) with a strong linear area of reflectivity and a core of 55 dBZ near the center of the patch. The WBC model is able to handle this storm the best, while the FSS model performs

the worst. The WBC model was able to detect the main core of the storm and outline the edges in higher probabilities. This resulted in a CSI of 0.75 and an FSS of 0.93. The WoFS forecast performed well overall but struggled to find the exact center and edges of the storm. This resulted in slightly lower CSI and FSS scores of 0.72 and 0.85, respectively. Visually and quantitatively, the FSS model did not perform as well as the other two models. While the CSI and FSS scores were above the average for the testing dataset, the prediction appears to be lacking discrimination and skill. The prediction for the FSS model appears to increase both the probabilities and areas of the WoFS predictions. This trend will continue through the rest of the case studies and appears to be the main problem with the FSS model.

Overall, this case displays the discrimination and skill of the WBC model in identifying regions of strong reflectivity and encompassing the correct regions. As shown through both quantitative metrics and visualization of the prediction contours, the prediction of the WBC model outperformed that of the WoFS model. This case also showed the shortcomings of the FSS model. While the model was able to detect regions of strong reflectivity, it lost the smaller-scale features of the WoFS and WBC models' predictions.

4.3.1.2 Example of WBC Outperforming FSS

In Figure 4.6, there is a single discrete cell with a 55 dBZ core in the northwest corner of the domain. The WBC model was able to center its strongest prediction over the cell, while the other two models' predictions struggled to predict the center of the high reflectivity values. The WBC verification metrics were a CSI score of 0.63 and an FSS score of 0.68. For the WoFS model's prediction, it scored a CSI score of 0.58 and an FSS score of 0.70. The FSS model's prediction scored the worst, having CSI and FSS scores of 0.51 and 0.41, respectively.

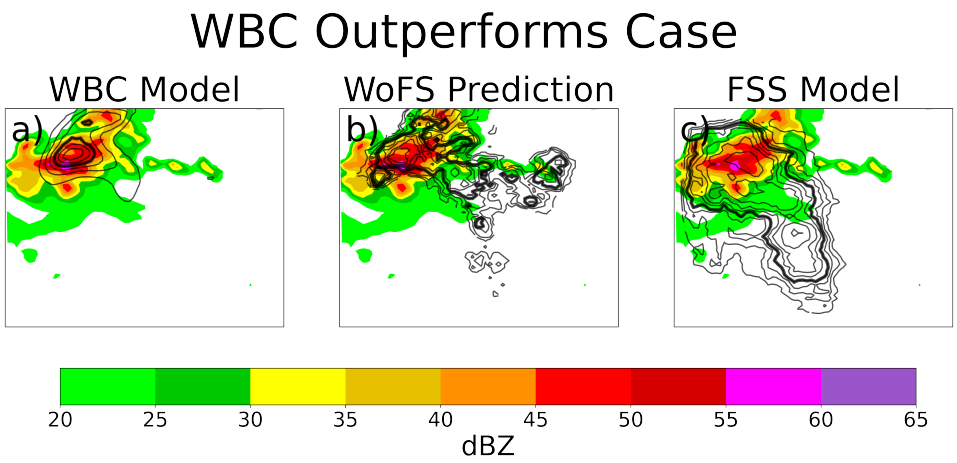

Figure 4.6: Case study displaying a case where the WBC model outperforms both the WoFS and FSS models' predictions. a) WBC model's prediction. b) WoFS model's prediction. c) FSS model's prediction. In all three panels, the radar objects are the MRMS at t=30min. The black contours are the prediction probabilities in intervals of 10%. The bolded black line represents the 50% probabilities. All models are attempting to predict reflectivity values $\geq$ 40 dBZ.

Though the FSS score of the WoFS prediction was slightly higher than the WBC model, visually, the WBC prediction appears to be more skillful. The WBC model identifies the region of strong reflectivity values and only places probabilities in that region. In contrast, the WoFS has high probabilities that are more spread out over the region. This supports the findings of the verification metrics and of Guerra et al. (2022), which showed the current WoFS predictions are over-predictive in nature. The reliability and performance diagrams in figure 4.3 showed that the WBC model was more reliable and had a frequency bias closer to 1.0. The WoFS performance and reliability diagrams showed the over-prediction bias. The FSS model's prediction is not very skillful (Figure 4.6c). The forecast placed a large area of high probabilities over regions of both high reflectivity values and no reflectivity values. This continues to show that the FSS model can not discriminate between areas of high reflectivity and no reflectivity. Overall, this case study showed that substantial prediction

improvements can be achieved through a DL model. In this case, the WBC model adjusts the forecast to overcome the over-prediction bias observed throughout the WoFS predictions.

4.3.1.3 Example of FSS Outperforming WBC

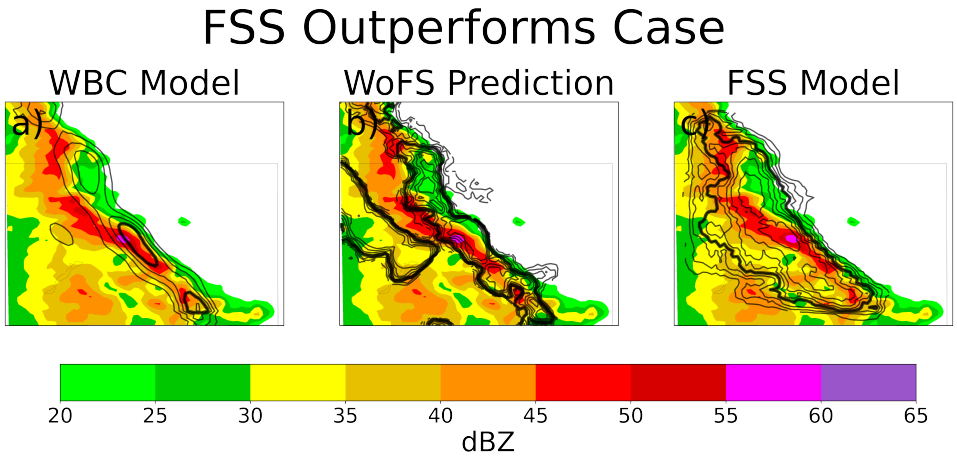

Figure 4.7: Case study displaying a case where the FSS model outperforms both the WoFS and WBC models' predictions. a) WBC model's prediction. b) WoFS model's prediction. c) FSS model's prediction. In all three panels, the radar objects are the MRMS at t=30min. The black contours are the prediction probabilities in intervals of 10%. The bolded black line represents the 50% probabilities. All models are attempting to predict reflectivity values $\geq$ 40 dBZ.

In this case study, there is a large ongoing QLCS with a strong leading edge to the storm with regions of higher reflectivity behind the line. In this case, the FSS model performed well, while the WBC model performed poorly. The WoFS baseline exhibits a leading area of high probabilities with a second, separate area of higher probability values. The WBC scored 0.48 and 0.18 for CSI and FSS, respectively. The WoFS model had a CSI score of 0.44 and an FSS score of 0.63. The FSS model performed best with a CSI score of 0.57 and an FSS score of 0.70.

This case is notable for two reasons. The first was how poorly the WBC model performed. While it did place higher probabilities over two regions of higher reflectivity values, it missed much of the QLCS leading edge and the high-reflectivity region in the southwest corner of the domain. This is very concerning for this model, being that it was a very strong signal, and the model should have been able to place high probabilities across much of the WoFS's domain. The FSS model's predictions did quite well in this case, wherein it captured a majority of the reflectivity values over 40 dBZ with high probabilities. But this case appears to be an outlier case. Being that FSS is a spatial loss function evaluated at 15km neighborhood, a larger patch domain is likely needed to create an accurate forecast. This thought is further supported by Justin et al. (2022). That paper utilized the FSS loss function over synoptic spatial scale fronts with great success. Therefore using this logic, the FSS performed so well in this case because there was a large region of strong reflectivity values it could analyze.

4.3.2 Iterative Goals for Next DL Models

Through the initial two DL models, there were two clear areas of findings that guided the next iteration of DL models. One finding was that the WBC loss function is the best loss function for this application due to its pixel-based nature and its increased weighting of positive cases. A neighborhood-based loss function is not appropriate due to the de-emphasis on small-scale features. The weighted aspect of the WBC models was important as well. Due to reflectivity values $\geq$ 40 dBZ being very rare– 0.006% of the total grid points– more heavily weighing the positive observed cases allowed the model to learn better what kind of patterns are conducive to reflectivity values $\geq$ 40 dBZ.

The initial experiments motivated tests of two changes to the data processing. The first was exploring how removing zero padding from the domain border would

impact the WBC model and its performance. Zero padding on the outside of the WoFS domain could lead to a skewing of performance, and it is important to explore how it impacts the models. The second was that data was not entirely independent, and a new patching and splitting scheme was required to ensure the evaluation of the models was correct. Overall, the first iteration of models led to promising results for proving that DL models can produce accurate and skillful predictions of intense reflectivity at a 30-minute lead time.

4.4 Current Deep Learning Performance

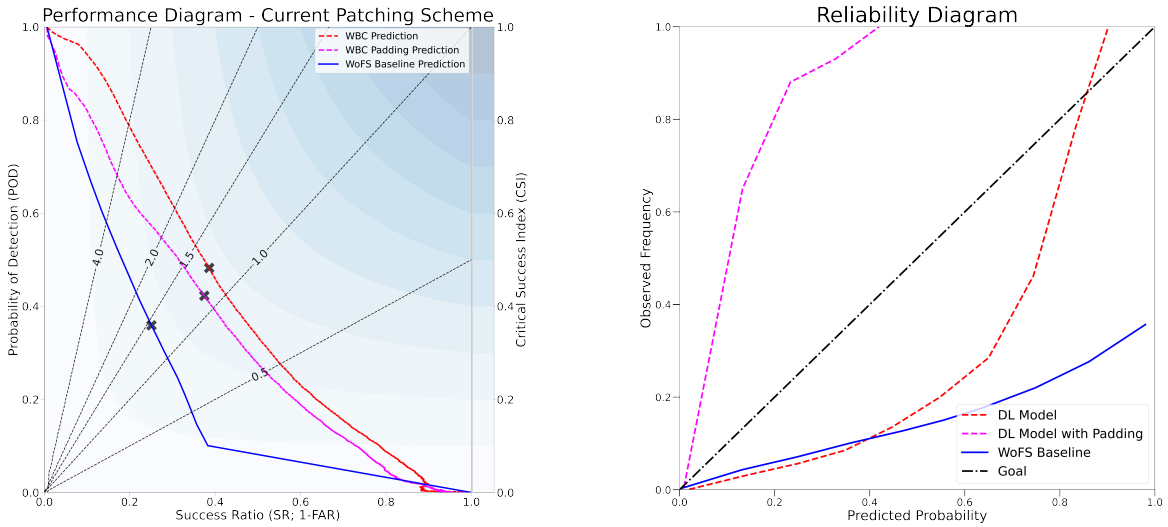

Figure 4.8: (left) Performance diagram comparing the new DL model (red dashed line), the previous DL model (magenta dashed line), and the WoFS baseline (blue solid line). The X represents the maxCSI values for the corresponding model. For the new DL model, the maxCSI was 0.27; for the old DL model, the maxCSI was 0.25; for the WoFS baseline, the maxCSI was 0.19. (right) Reliability diagram comparing the DL model (red dashed line), the previous DL model (magenta dashed line), and the WoFS baseline (blue solid line). The black dashed line is the ideal line for a model's prediction.

Results for the current DL model were produced from the testing dataset described in sections 3.4.2. The DL model achieved the goal of increasing the skill substantially when compared to the WoFS baseline. Figure 4.8 displays two separate performance metrics comparing the new DL model, the old DL model, and the WoFS baseline. Results from the performance diagram show a substantial increase in maxCSI and overall performance. MaxCSI values rose from 0.19 to 0.27 when comparing the WoFS baseline and the new DL model. Comparing the old DL model to the new DL model, the maxCSI increased from 0.25 to 0.27. The new DL model decreased the model's bias by lowering the maxCSI marker closer to the 1.0 bias line. Regardless of the probability threshold, the DL model outperforms the baseline in the performance diagram, especially when the probability threshold is high. The reliability was also much different between the old and new DL models. The old model is showing a very strong underprediction bias and is much less reliable than the new model. Therefore due to a higher maxCSI and better reliability, the new DL model will now be referred to as the DL model, and no more comparisons will be made between the old model and the new model. These results are further confirmed within the reliability diagram.

Isotonic regression (Niculescu-Mizil and Caruana 2005) was applied to the DL model to attempt to improve the reliability of the prediction. This method has been used in similar cases to increase the reliability of models with success (Burke et al. 2020; Flora et al. 2019). While the reliability of the DL model increased and closely matched the goal line in Figure 4.8, the model's discrimination was greatly impacted. The increase in reliability did not justify the loss of discrimination, and so the use of isotonic regression was abandoned.

4.4.1 Case Studies - Current Results

Multiple examples were extracted from the testing dataset to display the skill and discrimination of the DL model compared to the baseline. The three cases selected

were the best performance for the DL model, a median case, and a worst case. The best and worst cases were determined through the greatest difference in CSI between the DL model and baseline at each sample. The average case was determined through the median difference in CSI in the testing dataset, which was 0.032. Cases, where the difference was 0 were ignored. To measure the performance of the DL model and WoFS baseline in each of the cases, CSI and Brier Score (BS; Brier (1950)) were used. The probability threshold used to calculate the CSI score was determined by the probability threshold that resulted in the maxCSI score for each model. The thresholds were 0.575 and 0.475 for the DL model and WoFS baseline, respectively.

4.4.1.1 Best Performing Case: 0230 UTC 21 May 2021

Figure 4.9 displays a case from 0230 UTC on 21 May 2021 in South Dakota, where the DL model created a very good prediction. This case presents a quickly strengthening QLCS with two regions of high reflectivity. From the initialization of the model to the prediction time, the northern cell increased reflectivity values from 45 dBZ to 55 dBZ. In the southern cells, the reflectivity values increased from 40 dBZ to 50 dBZ. The DL model handled this rapid intensification of the storm much better than the WoFS baseline. This is shown in both the performance metrics and the locations of the prediction contours. The DL model predictions outperformed the WoFS baseline predictions by large margins in CSI (0.29) and BS (0.36). Rapid intensification of storms is a known problem for the WoFS (Guerra et al. 2022), and it is encouraging to see the DL model can improve prediction in an area the WoFS is currently lacking.

In addition to the large difference in performance metrics, the location of the prediction contours emphasizes the increase in discrimination and skill the DL offers in this case. Visual inspection of the DL model and WoFS baseline confirm the superior performance of the DL model in this case. The DL model was able to place the highest probabilities over the two regions of high MRMS reflectivity values. This

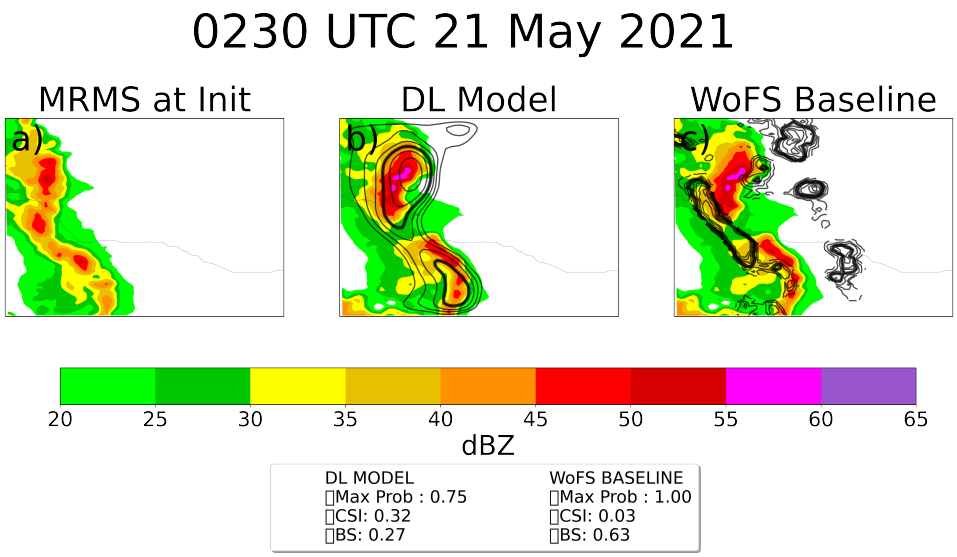

Figure 4.9: Case study from WoFS forecast initialized at 0230 UTC 21 May 2021 in South Dakota. a) Observed MRMS composite reflectivity at t=0 min, valid at 0230 UTC 21 May 2021. b) Black contour DL model predictions overlaid on observed MRMS composite reflectivity at t=30 min. c) Black contour WoFS baseline predictions overlaid on observed MRMS composite reflectivity at t =30 min. Any probability $\leq 10\%$ is masked out. The 50% contour line is in bold. Due to DL model only having information within a patch, contours do not extend out of the image like in the WoFS baseline predictions.

also included placing the highest probability (0.75) over the maximum reflectivity value (55 dBZ). In contrast, the WoFS baseline prediction was not able to handle this case well, wherein all the regions of higher probabilities are not over the regions with strong reflectivity values. The highest probabilities were placed behind the storms and completely missed the highest area of reflectivity values. The WoFS baseline also predicted 3 regions of $\geq$ 40 dBZ values which are ahead of the observed line of storms. These three areas, if interpolated out, line up with the three areas of strong reflectivity shown at the initialization of the model. This suggests that the WoFS prediction was expecting faster storm motions and decaying storm intensities.

Overall, this case is very encouraging for the DL model's performance. The DL model was able to accurately predict the location of the storms and where the greatest areas of intensification would be located. This is in vast contrast to the WoFS baseline prediction, which missed the location and strengthening of the storm and placed high probabilities in regions where no reflectivity was observed. This case shows the potential of the DL model to improve in the areas the WoFS currently is weak and generally increase the skill of the WoFS.

4.4.1.2 Average Case: 0300 UTC 19 May 2021

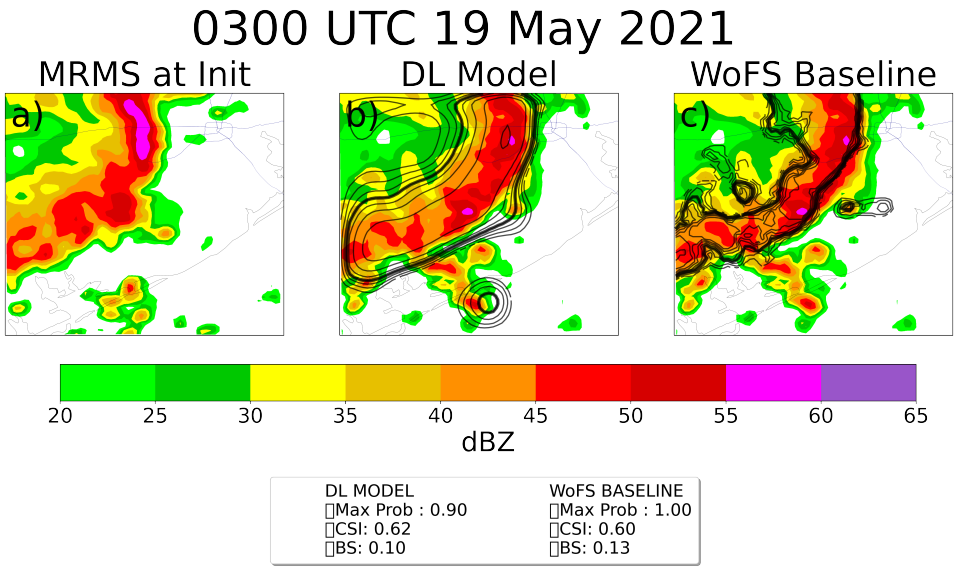

Figure 4.10: Case study from WoFS run at 0300 UTC 19 May 2021 from Southeastern Texas. a) Observed MRMS composite reflectivity at t=0 min, valid at 0300 UTC 19 May 2021. b) Black contour DL model predictions overlaid on observed MRMS composite reflectivity at t=30 min. c) Black contour WoFS baseline predictions overlaid on observed MRMS composite reflectivity at t =30 min. Any probability $\leq$ 10% is masked out. The 50% contour line is in bold. Due to DL model only having information within a patch, contours do not extend out of the image like in the WoFS baseline predictions.

Figure 4.10 is a case from 0300 UTC on 19 May 2021 with a strong ongoing QLCS over Southern Texas. Comparing the performance of the two models shows that both created good predictions with similar performance metrics. The DL model scored 0.62 and 0.10 on the CSI and BS, respectively. The WoFS baseline scored 0.60 and 0.13 on the CSI and BS, respectively. While the DL and WoFS Baseline models performed similarly on these metrics, two different aspects of this prediction are encouraging for the DL model. First, the highest DL probabilities are slightly better aligned with the highest MRMS reflectivity values than the BL probabilities, which lag the QLCS.

The second encouraging aspect of the DL model's prediction is at the bottom of the patch domain in figure 4.10. At this location, the DL model correctly predicts the cell east of the QLCS in the southern part of the domain, while the WoFS baseline prediction misses it. These results again suggest the DL model can occasionally predict rapid intensification when the WoFS fails to. The prediction of the cell also shows that the DL model is able to differentiate from the main area of high reflectivity values and is not just replicating the WoFS predictions. Overall, this case shows that even with similar performance metrics, the DL model will make subtle improvements leading to an overall more skillful prediction.

4.4.1.3 Worst Case: 0230 UTC 22 May 2021

Figure 4.11 displays a case from 0230 UTC 22 May 2021 over the Nebraska-South Dakota border. This case exhibits disorganized storm mode with two areas of high reflectivity. One region, in the northern portion of the domain, has slightly weakened, and one, in the southern portion of the domain, has strengthened. Most of the higher probability predictions in the DL model are centered over regions where no reflectivity values were observed. Based on the location of the storm at t=0 minutes and t= 30 minutes, it appears that the DL model did not handle the storm motion correctly. The DL model predicted the storm would move eastward, whereas the

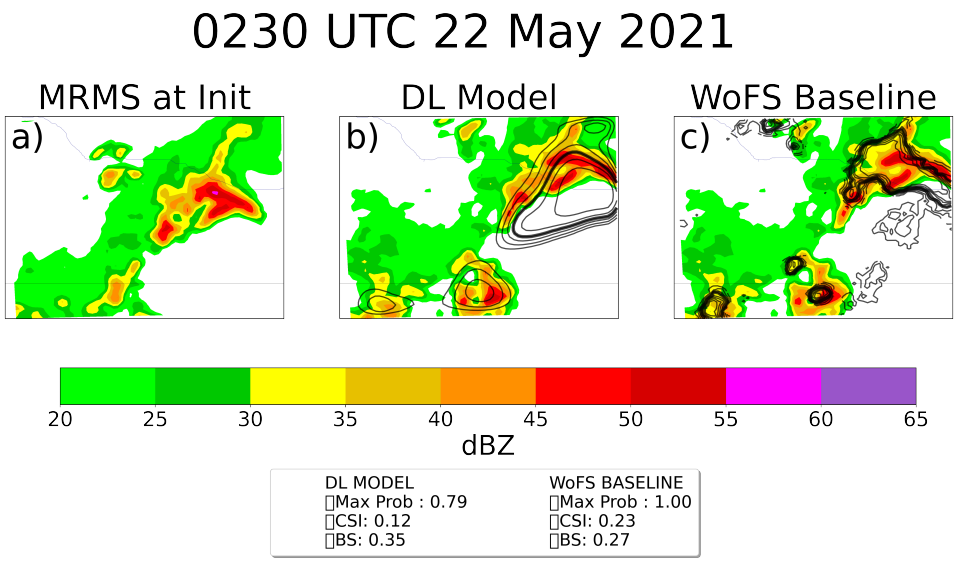

Figure 4.11: Case study from WoFS run at 0230 UTC 22 May 2021 from the Nebraska-South Dakota border. a) Observed MRMS composite reflectivity at t=0 min, valid at 0230 UTC 22 May 2021. b) Black contour DL model predictions overlaid on observed MRMS composite reflectivity at t=30 min. c) Black contour WoFS baseline predictions overlaid on observed MRMS composite reflectivity at t=30 min. Any probability $\leq$ 10% is masked out. The 50% contour line is in bold.

observed storm moved northeastward. In contrast, the WoFS baseline prediction did much better predicting the motion of the northern storm, causing a majority of the high probabilities to overlap with the strong observed reflectivity values. The evaluation metrics well reflected the difference in performance between the two models. The DL model scored a CSI of 0.12 and a BS of 0.35, whereas the WoFS baseline scored a CSI of 0.23 and a BS of 0.27.

This case shows a limitation of the DL model that should be fixed in future iterations of the model. Currently, the model only has access to the information in the patch's domain, while the WoFS baseline has information pertaining to the entire domain. This difference in information leads to the WoFS baseline outperforming the DL model when storms are located on the edges of the patch. In operations, this problem should go away as the DL model will have access to the full domain. We are

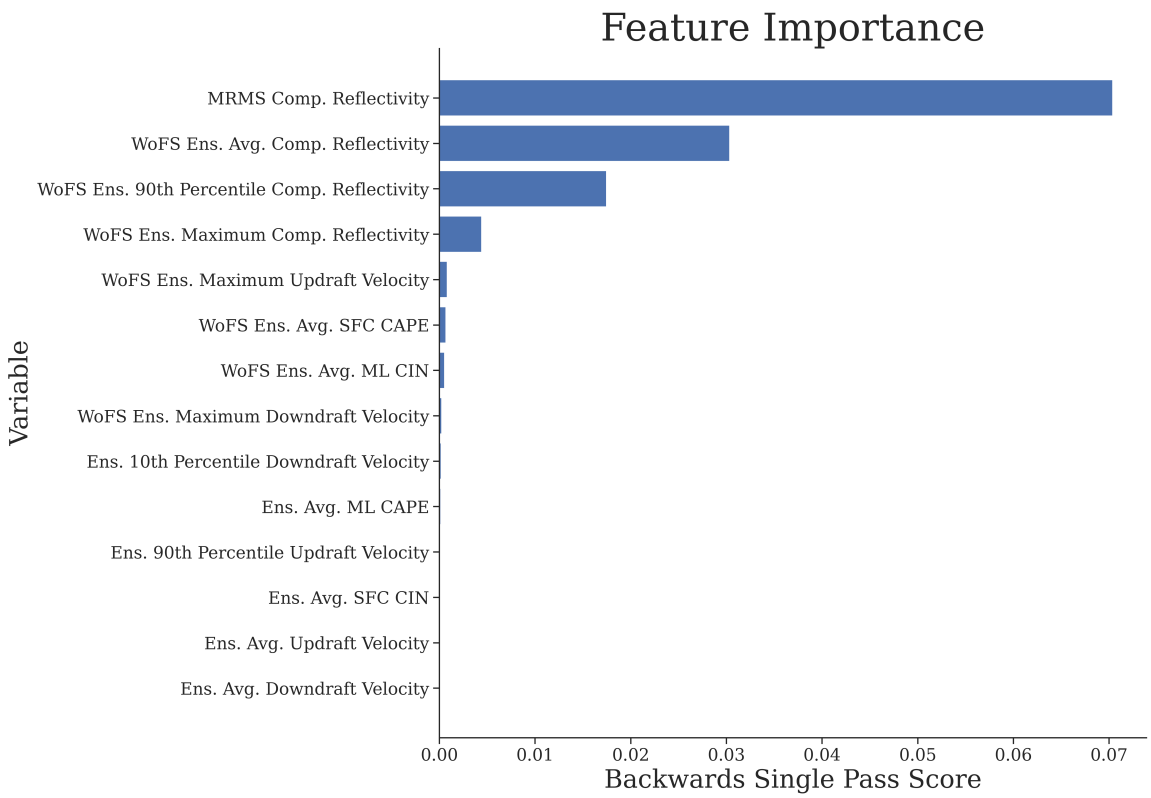

Figure 4.12: Feature importance plot with the most impactful variable at the top (MRMS Composite Reflectivity) and the least impactful variable at the bottom (Ensemble average Downdraft Velocity).

confident in this assessment due to how the DL model consistently outperforms the WoFS baseline when storms are located in the center of a patch's domain.

4.5 Explainability

Figure 4.12 shows the feature importance diagram for the DL model. To compute the feature importance we used backward permutation importance (McGovern et al. 2019). In this approach, a new instance of the model was used and evaluated without one of the input variables. If a variable was important to performance, the model would perform worse without that variable. Due to the limited number of variables, we were not concerned with correlations. The MRMS composite reflectivity at t =

0 min was the most important variable, while most of the environmental variables were unimportant (consistent with Flora et al. (2021); Clark and Loken (2022)). MRMS composite reflectivity is the most impactful by a wide margin, confirming our hypothesis that MRMS composite reflectivity at t= 0 minutes would help to overcome the phase and intensity errors in the WoFS initial conditions.

This outcome is reasonable given that the prediction lead time was so short (30 minutes). Unsurprising, the WoFS composite reflectivity fields were also quite impactful, specifically ensemble average and ensemble 90th percentile composite reflectivity. These were more impactful than the ensemble maximum for WoFS composite reflectivity. This was due to the DL model understanding the biases within each of the composite reflectivity fields. Other intra-storm variables, such as updraft and downdraft, were unimportant for the performance of the model. While these variables are vital for storm strength and longevity, they are less deterministic of observed composite reflectivity than the WoFS-predicted composite reflectivity variables.

Environmental variables had a negligible impact on the performance of the model. We hypothesize these variables are redundant for short forecasting times but could lead to greater impact as forecasting time increases. This is because if there is already an ongoing storm in the patch, it would be a safe assumption to assume that the environment is primed for convection in the short term. As the forecast time extends and storms have yet to form, the environmental variables should become more impactful for the DL model to predict whether or not convection can form and be sustained.

Chapter 5

Conclusion and Future Work

The goal of this project was to determine if a DL model would be able to substantially improve upon the current WoFS product for forecasting reflectivity values $\geq$ 40 dBZ 30 minutes after WoFS initialization. Through evaluation and comparison of the DL model and WoFS baseline performances on the entire testing dataset and individual patches, it is clear that the DL model was able to increase performance compared to the baseline substantially. The DL model was able to increase the MaxCSI from 0.17 to 0.27 while decreasing the frequency bias. Also, the DL model was able to improve the reliability of the prediction for predicted probabilities exceeding 0.30. Comparing individual test cases, the DL model was able to remedy some of the problems described in Guerra et al. (2022), where the WoFS predictions struggled with rapidly intensifying storms and phase errors. In multiple cases of both QLCS and discrete storms, the DL model was able to correctly shift the highest probabilities directly over the region of the strongest reflectivity when the WoFS prediction did not. One of the only limitations of the DL model came from the lack of skill when detecting storms along the edges of the patch domain. This problem should be resolved in future iterations of the model due to running on information from the entire domain.

Through multiple iterations of the DL models, the main findings were: 1) The weighted binary cross-entropy loss function was the best loss function for this application due to the rare and small-scale nature of reflectivity values exceeding 40 dBZ. 2) Including MRMS composite reflectivity data valid at initialization of the WoFS is very impactful in creating an improved forecast at a 30-minute lead time. 3) Creating

a randomly selected patching area creates a dataset more reflective of reality than adding padding to the outside of the WoFS domain.

Future plans for this project include decreasing the reliance on patches and utilizing the entire WoFS domain. This should resolve the issue of poor predictions by the DL model when storms are located on the edge of the patch's domain because the WoFS domain is often centered over the regions where strong convection is expected. It will also be expected that the DL model's performance should also increase, as seen in how the DL model performed when storms were located in the center of patches and limiting the instances of limited data on domain edges. Secondly, feature ablation should allow the model to run more quickly with little loss of skill. Based on the results from the explainability section, many of the environmental variables could be removed without substantially impacting the performance of the DL model.

A final and grandiose goal will be to increase the prediction timesteps to every five minutes out to an hour. This could be done by creating a 3-D, time-resolving U-Net where the output from the model would be multiple timesteps predictions at each grid point. This would be more memory intensive but, we hypothesize, would allow the model to understand the evolution of storms and their environments. Alternatively, the 2-D model could be retained, and the input fields sequenced in time. In this way, the output from the model would still predict out through an hour, but the model would not learn about storm evolution. Overall, both concepts have positives and negatives, and it is worth exploring both options.

In conclusion, the DL model was able to substantially outperform the baseline in predicting reflectivity values $\geq$ 40 dBZ 30 minutes after the initialization of the WoFS model. This was completed through the use of a U-Net deep learning model architecture with MRMS composite reflectivity at t=0min, WoFS model output intra-storm variables, and WoFS model output environmental variables as input variables. The DL model outperformed the WoFS baseline in all performance metrics and showed that the predictions were skillful in predicting thunderstorm locations. With ongoing

work to continue to improve this DL model's performance, there is evidence to be encouraged about the improvement of the WoFS forecast of thunderstorms through a post-processing DL model.

The experimental WoFS ensemble forecast data used in this study are not currently available in a publicly accessible repository. However, the data used to generate the results herein are available from the authors upon request. GitHub for all the code used in this project is found at: https://github.com/chadwiley14/wofs_40dbz.

Reference List

Barclay, P. A., and K. E. Wilk, 1970: *Severe thunderstorm radar echo motion and related weather events hazardous to aviation operations*. ESSA.

Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quart. J. Roy. Meteor. Soc.*, **132 (620)**, 2127–2155.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78 (1)**, 1–3.

Browning, K. A., 1997: Review lecture: Local weather forecasting. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, **371 (1745)**, 179–211.

Burke, A., N. Snook, D. J. Gagne, II, S. McCorkle, and A. McGovern, 2020: Calibration of machine Learning–Based probabilistic hail predictions for operational forecasting. *Weather Forecast.*, **35 (1)**, 149–168.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022a: A machine learning tutorial for operational meteorology. part i: Traditional machine learning. *Weather Forecast.*, **37 (8)**, 1509–1529.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2023: Machine learning tutorials for operational meteorology. *103rd AMS Annual Meeting*, AMS.

Chase, R. J., D. R. Harrison, G. Lackmann, and A. McGovern, 2022b: A machine learning tutorial for operational meteorology, part II: Neural networks and deep learning. `2211.00147`.

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 Forecast.*, **35 (6)**, 2567–2588.

Clark, A. J., and E. D. Loken, 2022: Machine Learning–Derived severe weather probabilities from a Warn-on-Forecast system. *Weather Forecast.*, **37 (10)**, 1721–1740.

Clark, A. J., and Coauthors, 2022: The second Real-Time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Am. Meteorol. Soc.*, **103 (4)**, E1114–E1116.

Cornman, L. B., and B. Carmichael, 1993: Varied research efforts are under way to find means of avoiding air turbulence. *ICAO J.*

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis, and Nowcasting—A radar-based methodology. *J. Atmos. Ocean. Technol.*, **10 (6)**, 785–797.

Dowell, D. C., and Coauthors, 2016: Development of a High-Resolution rapid refresh ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*.

Ebert-Uphoff, I., R. Lagerquist, K. Hilburn, Y. Lee, K. Haynes, J. Stock, C. Kumler, and J. Q. Stewart, 2021: CIRA guide to custom loss functions for neural networks in environmental sciences – version 1. 2106.09757.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate Storm-Scale probabilistic guidance of severe weather hazards in the Warn-on-Forecast system. *Mon. Weather Rev.*, **149 (5)**, 1535–1557.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-Based verification of Short-Term, Storm-Scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast system. *Weather Forecast.*, **34 (6)**, 1721–1739.

Gagne, D. J., II, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Weather Rev.*, **147 (8)**, 2827–2845.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous weather testbed spring forecasting experiment. *Weather Forecast.*, **32 (4)**, 1541–1568.

Gallo, B. T., and Coauthors, 2022: Exploring the watch-to-warning space: Experimental outlook performance during the 2019 spring forecasting experiment in NOAA's hazardous weather testbed. *Weather Forecast.*, **37 (5)**, 617–637.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the united states using ERA5 proximity soundings. *Weather Forecast.*, **36 (6)**, 2143–2160.

Golding, B. W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appl.*, **5 (1)**, 1–16.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of NSSL Warn-on-Forecast system accuracy by storm age using Object-Based verification. *Weather Forecast.*, **37 (11)**, 1973–1983.

Han, L., J. Sun, and W. Zhang, 2020: Convolutional neural network for convective storm nowcasting using 3-D doppler weather radar data. *IEEE Trans. Geosci. Remote Sens.*, **58 (2)**, 1487–1495.

Hilst, G. R., and J. A. Russo, 1960: *An objective extrapolation technique for semi-conservative fields with an application to radar patterns*. Travelers Insurance Companies.

John Saida, S., and S. Ari, 2022: MU-Net: Modified U-Net architecture for automatic ocean eddy detection. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1–5.

Johnson, J. T., P. L. MacKeen, A. Witt, E. De Wayne 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 Forecast.*, **13 (2)**, 263–276.

Justin, A. D., C. Willingham, A. McGovern, and J. T. Allen, 2022: Toward operational real-time identification of frontal boundaries using machine learning: A 3D model. *102nd American Meteorological Society Annual Meeting*, AMS.

Kaplan, M. L., A. W. Huffman, K. M. Lux, J. J. Charney, A. J. Riordan, and Y.-L. Lin, 2005: Characterizing the severe turbulence environments associated with commercial aviation accidents. part 1: A 44-case study synoptic observational analyses. *Meteorol. Atmos. Phys.*, **88 (3-4)**, 129–152.

Kotsuki, S., K. Kurosawa, S. Otsuka, K. Terasaki, and T. Miyoshi, 2019: Global precipitation forecasts by merging Extrapolation-Based nowcast and numerical weather prediction with locally optimized weights. *Weather Forecast.*, **34 (3)**, 701–714.

Lagerquist, R., A. McGovern, and D. J. Gagne, II, 2019: Deep learning for spatially explicit prediction of Synoptic-Scale fronts. *Weather Forecast.*, **34 (4)**, 1137–1160.

Lagerquist, R., A. McGovern, C. R. Homeyer, D. J. Gagne, II, and T. Smith, 2020: Deep learning on Three-Dimensional multiscale data for Next-Hour tornado prediction. *Mon. Weather Rev.*, **148 (7)**, 2837–2861.

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. *Mon. Weather Rev.*, **149 (12)**, 3897–3921.

Lane, T. P., R. D. Sharman, S. B. Trier, R. G. Fovell, and J. K. Williams, 2012: Recent advances in the understanding of Near-Cloud turbulence. *Bull. Am. Meteorol. Soc.*, **93 (4)**, 499–515.

LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, 1989: Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.*, **2**.

Li, Y., Y. Liu, R. Sun, F. Guo, X. Xu, and H. Xu, 2023: Convective storm VIL and lightning nowcasting using satellite and weather radar measurements based on Multi-Task learning models. *Adv. Atmos. Sci.*, **40 (5)**, 887–899.

Ligda, M. G. H., 1953: Horizontal motion of small precipitation areas as observed by radar. Ph.D. thesis.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic Next-Day severe weather forecasts from Convection-Allowing ensembles using random forests. *Weather Forecast.*, **35 (4)**, 1605–1631.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Am. Meteorol. Soc.*, **98 (10)**, 2073–2090.

McGovern, A., R. Lagerquist, D. J. Gagne, G. Eli Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.*, **100 (11)**, 2175–2199.

Miller, W. J. S., and Coauthors, 2022: Exploring the usefulness of downscaling free forecasts from the Warn-on-Forecast system. *Weather Forecast.*, **37 (2)**, 181–203.

National Transportation Safety Board, ????: Aviation investigation preliminary report. Tech. Rep. DCA23LA096, NTSB.

Niculescu-Mizil, A., and R. Caruana, 2005: Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*, Association for Computing Machinery, New York, NY, USA, 625–632, ICML '05.

Noel, T. M., and A. Fleisher, 1960: The linear predictability of weather radar signals. Tech. rep., Massachusetts Inst of Tech Cambridge.

Pan, G., Y. Zheng, S. Guo, and Y. Lv, 2020: Automatic sewer pipe defect semantic segmentation based on improved U-Net. *Autom. Constr.*, **119**, 103 383.

Roberts, N. M., and H. W. Lean, 2008: Scale-Selective verification of rainfall accumulations from High-Resolution forecasts of convective events. *Mon. Weather Rev.*, **136 (1)**, 78–97.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, 234–241.

Sander, J., J. F. Eichner, E. Faust, and M. Steuer, 2013: Rising variability in Thunderstorm-Related U.S. losses as a reflection of changes in Large-Scale thunderstorm forcing. *Weather, Climate, and Society*, **5 (4)**, 317–331.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from Convection-Allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Weather Rev.*, **145 (9)**, 3397–3418.

Skamarock, W. C., 2008: A description of theadvanced research WRF version 3. *Tech. Note*, 1–96.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Weather Forecast.*, **33 (5)**, 1225–1250.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Am. Meteorol. Soc.*, **97 (9)**, 1617–1630.

Sobash, R. A., G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016: Explicit forecasts of Low-Level rotation from Convection-Allowing models for Next-Day tornado prediction. *Weather Forecast.*, **31 (5)**, 1591–1614.

Stensrud, D. J., and M. S. Wandishin, 2000: The correspondence ratio in forecast evaluation. *Weather Forecast.*, **15 (5)**, 593–602.

Stensrud, D. J., and Coauthors, 2009: Convective-Scale Warn-on-Forecast system: A vision for 2020. *Bull. Am. Meteorol. Soc.*, **90 (10)**, 1487–1500.

Stensrud, D. J., and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16.

Sun, J., and Coauthors, 2014: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bull. Am. Meteorol. Soc.*, **95 (3)**, 409–426.

Tsonis, A. A., and G. L. Austin, 1981: An evaluation of extrapolation techniques for the short-term prediction of rain amounts. *Atmosphere-Ocean*, **19 (1)**, 54–65.

Wilk, K. E., and K. C. Gray, 1970: Processing and analysis techniques used with the NSSL weather radar system. *Conf. on Radar Meteorology, Tucson, AZ, Amer. Meteor ...*.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95 (1)**, 51–70.

Wilson, J. W., N. Andrew Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Am. Meteorol. Soc.*, **79 (10)**, 2079–2100.

Wilson, K. A., B. T. Gallo, P. Skinner, A. Clark, P. Heinselman, and J. J. Choate, 2021: Analysis of end user access of Warn-on-Forecast guidance products during an experimental forecasting task. *Weather, Climate, and Society*, **13 (4)**, 859–874.

Zhang, J., and Coauthors, 2021: LCU-Net: A novel low-cost U-Net for environmental microorganism image segmentation. *Pattern Recognit.*, **115**, 107 885.

Zittel, W. D., 1976: Computer applications and techniques for storm tracking and warning. *Conference on Radar Meteorology, 17 th, Seattle.*