UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

EXPERT-GUIDED MACHINE LEARNING FOR METEOROLOGICAL PREDICTIONS ACROSS SPATIO-TEMPORAL SCALES

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy

By

AMANDA BURKE
Norman, Oklahoma
2024

EXPERT-GUIDED MACHINE LEARNING FOR METEOROLOGICAL PREDICTIONS ACROSS SPATIO-TEMPORAL SCALES

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Amy McGovern (Chair)

Dr. Cameron Homeyer

Dr. Elinor Martin

Dr. Talayeh Razzaghi

Acknowledgments

I'm not the type of person to need a lot of words to convey what I feel. I want to thank Amy McGovern because without her I don't know where I would be and I certainly would not be in the position I'm in and have the success that I have had. She's been an amazing advisor and one that I am so grateful to have met. Next, big thank you to Cameron Homeyer for being so supportive during the last few years in my PhD. It's been a roller coaster but I can confidently say that you have been part of the bar keeping me in my seat.

Without the push and guidance from my Spirit Team I know I wouldn't be where I'm at in life. They don't allow me to stop fighting for myself and because of that, and my own Inner Strength, I've overcome so many challenges.

Last but very much not least a very large thank you to myself. Those nights where all you have is yourself and your inner voice is the enemy, you really learn a lot about yourself and the strength needed to pull through. Here's to the tower moments that led to this dissertation finally being finished, which I didn't think would have been possible 2 years ago.

Keep shining, Sunshine ∞

Table of Contents

List Of Tables

List Of Figures

Abstract

This dissertation emphasizes the contribution of expert knowledge in the development and assessment of machine learning (ML) models within the Earth sciences, specifically Meteorology. Despite the common focus on achieving high skill scores, conventional metrics may inadequately capture the nuanced patterns learned by these models. This dissertation underscores the importance of incorporating end-user feedback, demonstrating that with this feedback, tailored yet flexible ML models can effectively learn specific meteorological patterns while remaining applicable to broader contexts.

The first focus of ML development is in identification of above-anvil cirrus plumes (plumes). In satellite imagery, plumes serve as critical indicators of impending severe weather, often appearing 30 minutes before reported events. Their real-time identification is particularly valuable in radar-deficient regions, where they offer insights into the convective environment. However, manually labeling plumes is labor-intensive and requires specialized expertise. To streamline this process, I develop a deep learning (DL) model trained on expert-annotated data to create skillful pixel-level plume classifications using remote sensing data that is available globally. This approach was tested on combinations of spectral data across the contiguous United States, showing above-average object correspondence with human-derived labels.

Another focus of this dissertation is leveraging ML models for severe hail prediction on localized scales. Existing ML models have demonstrated proficiency across the United States during spring and summer but have struggled to capture the nuanced spatio-temporal dynamics of thunderstorm development in local contexts. Addressing this gap, I develop a novel localization technique that prioritizes storm object weighting without imposing substantial additional burdens on model developers. Results indicate that localized weighting of storm objects matches or outperforms existing ML

approaches, while improving the physical relevance of the top predictors in the trained ML model.

Lastly, leveraging the expansive archives of satellite data, this dissertation tackles the challenge of creating training sets that are representative of large scale Earth science datasets while maintaining efficiency. This work explores clustering approaches to extract regional nuances amidst a vast dataset of remote sensing data for a straightforward use case with an established baseline - land cover classification. Employing surface reflectance bands in a random forest (RF) model, I compare classification outcomes between training with randomly sampled datasets of varying size and datasets created using clustering. Using a clustering approach, a training sample was created that was 200% smaller than the largest sample studied, yet it achieved a 77% increase in F1 score. This suggests that clustering may offer an effective alternative (or addition) to increasing computing power when modeling "Big Data".

Chapter 1

Background

Artificial intelligence (AI) and other data science methods have pioneered evaluating vast datasets across diverse domains. In fact, over the past decade there has been a widespread embrace of AI techniques by forecasters and researchers, owing to their effectiveness in various applications such as post-processing calibration, decreasing cognitive load, and uncovering novel insights (McGovern et al., 2017). More recently, the expansion of computational resources and data availability (e.g., Haupt et al., 2018b,a) resulted in a spread of ML applications to different meteorological domains.

This dissertation covers the application of ML to three domains: hail prediction and land cover classification, which have years of development, and above anvil cirrus plume identification, which is relatively new to ML. The expansion of computational capabilities increases AI's potential applications in the Earth sciences and beyond, a topic further discussed in the representative sampling section (Chapter 4). However, experts, defined by their accredited scientific training and comprehensive knowledge of a particular field, remain essential for successful ML modeling strategies and deployment.

The concept of "success" will be explored in detail, particularly regarding severe hail prediction, in Chapter 3. Using these powerful ML models is only part of the challenge; another part involves developing solutions for tasks previously difficult or impossible to study. The progress in ML methodologies and capabilities in processing high-resolution data enable more detailed investigations into phenomena that influence

our climate, such as the transport of water vapor (WV) to the upper troposphere and lower stratosphere (UTLS) via Above-Anvil Cirrus Plumes (AACPs). This is especially important as the world deals with the increasing impacts of climate change and scientists seek to understand and address these issues. More detail on how ML methods and increasing data resolution have been crucial for successful model development, important for both climate impacts and protecting life and property, will be provided in Chapter 2 that discusses AACP identification.

1.1 Above Anvil Cirrus Plume Identification

In the realm of climate studies, accounting for stratospheric WV is critical as its a potent greenhouse gas, impacting both stratospheric cooling and surface warming. Even minor increases in stratospheric WV ($\leq 10\%$) can exert substantial influence on the Earth's radiation budget and climate dynamics (de F. Forster and Shine, 1999; Dessler and Sherwood, 2004; Solomon et al., 2010) as greenhouse effects can result in a positive feedback loop of surface warming. Furthermore, the enhancement of stratospheric WV may trigger the activation of organic chlorine compounds, leading to the depletion of stratospheric ozone (Anderson et al., 2012, 2017). These different aspects are important to understand as our climate warms and threatens livelihoods. AACPs contribute to stratospheric WV enhancement by at least an order of magnitude (Homeyer et al., 2017; O'Neill et al., 2021; Gordon and Homeyer, 2022) and were associated with the most significant WV increase in the stratosphere during NASA's SEAC4RS campaign over the Midwest United States. (Herman et al., 2017; Smith et al., 2017). Despite the clear importance of AACPs to stratospheric WV, and consequently the climate, there are still uncertainties about the frequency of WV transport to the UTLS via convection, particularly AACPs. One challenge in cataloging AACPs is that their warm

temperatures can cause cloud-top height retrieval algorithms to mistakenly identify them as tropospheric features, leading to a misrepresentation of plumes (Setvák et al., 2010).

With the advent of high spatio-temporal resolution data from visible (VIS) and infrared (IR) imagery, researchers have identified specific AACP characteristics. Plume temperatures can exceed those of the surrounding anvils by more than 20 K (Brunner et al., 2006). The warm anomalies within the plumes often contrast sharply with the much colder tropospheric anvil IR brightness temperatures (BTs), resulting in recognizable signatures such as "Enhanced-V" (EV, Brunner et al., 2006) or "cold ring" signatures (Setvák et al., 2010). Some plumes however are associated with the typical shadowing and smooth characteristics in VIS data but showing as cold (or colder) IR temperatures than the surrounding anvil. Based on the work in Murillo and Homeyer (2022), AACPs with warmer IR signatures (warm plumes) reside in the stratosphere and are associated with lower tropopause heights/warmer UTLS temperatures, whereas AACPs with colder IR plume characteristics (cold plumes) reside in the upper troposphere and are related to environments with higher tropopause heights/colder UTLS temperatures (Murillo and Homeyer, 2022). This distinction lends to the importance of identification of especially warm plumes for context with stratospheric water vapor increases. Additionally, the change in IR temperature values for a plume could prove challenging for identification when IR alone is input to a ML model.

Over the past 35 years, AACPs have been the subject of extensive study and have been identified as precursors to severe weather events, as observed in VIS and IR imagery (Fujita, 1982; McCann, 1983; Brunner et al., 2006; Setvák et al., 2010, 2013; Bedka et al., 2015; Homeyer et al., 2017; Kunz et al., 2017; Bedka et al., 2018; Liles et al., 2020; Mecikalski et al., 2021). Severe weather outbreaks often feature numerous long-lasting AACPs, with these phenomena commonly found above severe convection

worldwide. Bedka et al. (2018), analyzed over 4500 storms using combinations of radar, VIS/IR imagery, and lightning dataset to identify 405 storms that produced AACPs. The storms that produced AACPs were 14 times more likely to be severe compared to convection without AACPs, and resulted in 85% of the total 807 events of hail with diameters of 2+ inches (5+ cm) and EF-2+ tornadoes analyzed. Lending to the ability for AACP identification for severe weather prediction, AACPs appear on average 30 minutes before the first severe weather report produced by a storm. Where radar coverage is lacking, a purely satellite-based AACP detection product could offer valuable insight into storms likely to produce significant severe weather, aiding in warnings and protection of people and property. GOES satellites offer continuous IR and visible imagery, with observations collected at intervals as short as 10 minutes, and as frequently as every 30 seconds during field campaigns or particularly impactful weather events. Early detection of AACPs can provide warning lead time comparable to that of expert forecasters from the NOAA National Weather Service (Bedka et al., 2018),

Humans can manually identify AACPs, but the process is time-consuming and requires extensive training to accurately label them. Developers greatly desire an automated approach to reduce their workload. The GOES-R Aviation Algorithm Working Group (AWG) developed OT and EV/AACP detection algorithms using fixed criteria of temperature differences between OTs/AACPs and the surrounding anvil (Bedka et al., 2010; Bedka, 2011). However, this approach led to missed AACP detections. Moving from fixed thresholds to a to network-based deep learning (DL) approach optimized for spatial pattern recognition, the NASA LaRC explored DL approaches for OT and AACP detection using various satellite imagery sources, radar data, and human-based AACP identifications for training (Bedka et al., 2018; Cooney et al., 2024). Using convolutional neural networks (CNNs) and Unets trained with channels of 0.5 km VIS

reflectance and 2 km IR BT, the DL approach resulted in a validation intersection over union (IoU) of 0.3313. IoU relates how well a semantic segmentation model fits by dividing true positives ("hits") by the sum of the false positives ("false alarms"), true positives ("hits") , and false negatives ("misses") (Liles et al., 2020). An IoU of 1.0 is considered perfect, while values below 0.5 indicate that a predicted object is falsely identified.

Other approaches employing ML models on remote sensing data for convective prediction include the research conducted by the NOAA/CIMSS ProbSevere team that focused on identifying patterns atop convection indicative of severe weather (Cintineo et al., 2020). Utilizing approximately 64x64 km subsets of geostationary (GEO) imagery, storms were categorized as either "intense" or "non-intense", and a convolutional neural network (CNN) was applied to derive a "Probability of Intense Convection". While initial results were promising, this approach is less focused on storm cell-scale physical processes and operates without radar data. Apart from the ProbSevere team, Mecikalski et al. (2021) found that when using a random forest model to determine if 1-min satellite imagery is beneficial for severe weather warning detection, the presence of an AACP was in the top three most important variables in distinguishing severe convection from nonsevere convection. Kim et al. (2017) employed traditional machine learning models to identify overshooting tops, achieving a Probability of Detection (POD) of 77.06% and a False Alarm Rate (FAR) of 36.13%. Kanneganti (2020) used a CNN for overshooting top detection, resulting in a method with a POD of 79.31%, FAR of 90.94%, and critical success index of 0.088. Additionally, Lee et al. (2021) used satellite data input to a neural network to detect regions of convection. Cintineo et al. (2020) predicted convection using satellite and lightning data with a CNN. Finally, Wang et al. (2021) employed a random forest (RF, Breiman, 2001) algorithm to predict cloud-top height for tropical overshooting convection. Notably, the previously

mentioned algorithms do not evaluate data at the pixel scale and therefore cannot mirror human analysts.

In this study, we combine multi-sensor, multi-spectral data processed through state-of-the-art methods to offer a comprehensive understanding of the capabilities and limitations of DL in analyzing and detecting severe and tropopause-penetrating convective patterns within geostationary (GEO) imagery. The establishment of open-source methodologies for automated severe storm detection aims to support the Earth Science research community, as well as increase opportunity for transition to operations. This approach aims to identify AACP features at the individual pixel level, as this fine scale resolution is crucial for discerning the processes generating AACPs. The objective is to accurately map out the occurrence of these storms in terms of time and location to support NASA's research objectives and the severe weather, aviation weather, and climate research communities.

1.2 Severe Hail Prediction

Staying within the convective meteorology domain, we pivot to an area with several years of research and collaboration: severe hail prediction. According to Murphy (1993), a well-rounded forecast comprises three essential components: quality, or alignment with observation; consistency, reflecting how well a prediction aligns with forecaster judgment; and value, measuring its utility to end users. DL models can alleviate this spatio-temporal disparity through 2D convolutional methods, as applied to AACP identification, however conventional machine learning (ML) models typically prioritize quality by closely aligning with observations, sometimes neglecting consistency across spatial and temporal domains.

Despite the significant improvements demonstrated by AI/data science techniques in various high-impact weather domains, their operational integration is not straightforward. Forecasters must trust the forecasts generated by such techniques, as highlighted by experiences in projects like HWT/PHI (Karstens et al., 2018). Unlike previous years, multiple different ML models were submitted to the 2020 Hazardous Weather Testbed Spring Forecasting Experiment (HWT, Clark et al., 2021), underscoring the recent popularity of ML model development for predicting convective hazards (e.g., Burke et al., 2020; Loken et al., 2020; Sobash et al., 2020). The growth of novel dataset prediction with ML models underscores the importance of expert knowledge and empirical methods in developing not only quality ML models but also consistent predictions. One hazard that ML models have shown success at predicting, after expert tuning and input, is severe hail.

Without accounting for physical differences in data across time or space, Gagne et al. (2017) and Burke et al. (2020) demonstrated that an object-based RF method can produce quality forecasts across the contiguous United States (CONUS). Although skillful, under-representing the role of spatio-temporal variability of severe thunderstorm development (e.g., Kelly et al., 1985; Johns and Doswell, 1992; Shafer et al., 2010; Grams et al., 2012; Krocak and Brooks, 2018) on hail formation may result in models unable to capture important local environmental patterns (Smith et al., 2012; Allen et al., 2020). While emphasizing model proficiency is tempting, prior research suggests a preference among forecasters and other scientific users for physics-based models (McGovern et al., 2022). This makes sense as even if forecasts are accurate, optimal performance can come at the cost of learning non-physical relationships as most ML models do not generally model physical relationships between the inputs and observations (McGovern et al., 2019a).

In an effort to maintain optimal ML forecast quality *and* consistency, selecting physically-relevant ML training data for a given problem not only increases forecast interpretability without increasing model complexity but reduces computational load on model developers during the tuning stage. One method for selecting training data that considers thunderstorm development variability is through explicit bounding boxes in time and space (Hill et al., 2020). However, Burke (2019) found that strictly limiting the training dataset negatively impacts the RFs ability to learn useful prediction patterns. Hill et al. (2020) similarly reported decreased hail forecast performance in regions with low event frequencies. Additionally, hail forecasts calibrated on strict domain-dependent environmental parameters may be suboptimal outside a specific area and lack value (Brimelow et al., 2006; Jewell and Brimelow, 2009; Allen et al., 2020).

In this dissertation we introduce a procedure for statistically choosing weights applied to severe hail predictors, where storm examples in the relevant time (space) receive the highest weights. This method provides a flexible framework to "choose" which training data a RF model deems as important without increasing developer load. Thus, preserving the quality of the ML framework from Burke et al. (2020), while providing more information to forecasters and a potentially more consistent procedure for processing ML training data for optimal severe hail prediction. In fact, applying physics-based modeling techniques has shown promise at improving consistency while maintaining or even further improving quality (e.g., Willard et al., 2020; Beucler et al., 2021).

In addition, we demonstrate that the flexible method can be applied to multiple different domain problems with little oversight and provide ML models that learn environmental patterns consistent with previous literature detailing thunderstorm development variability in time and space, rather than the data with the highest sampling

frequency. As predictions derived from this ML method were submitted to the 2021 HWT, we were able to receive feedback on the algorithm. This provides a unique opportunity to incorporate different aspects of how trust plays a major role in ML predictions used by forecasters.

1.3 Representative Sampling of Global Data

The last part of this dissertation focuses on creating training datasets from extensive remote sensing data, aiming to capture regional and global patterns with a minimal number of samples. Abundant data resources, termed "Big Data", are the standard in today's technological landscape, particularly in remote sensing. For example, the Landsat satellite observes the globe every 16 days at a 30m spatial resolution, archiving data back to 1972; MODIS data are available at sub-1km resolutions twice daily, back to 2000; VIIRS, launched in 2011, is sending back daily global data at 375-5600m resolution every day. Although a significant volume of global data is at the disposal of the remote sensing community, *quickly* extracting *valuable* information from the data deluge is a challenge.

Traditional ML techniques, optimized for various data types (i.e. RGB images, audio, text, etc.), grapple with complexities in adapting to the nuances of remote sensing data (Vali et al., 2020). Differing from standard images input to ML models, different spectral signatures can indicate the same object while similar signatures can be attributed to different objects, because signatures can change depending on the materials, environmental conditions, seasonal variations, etc, within underlying images (Bao et al., 2013; Chi et al., 2016). The complexity of globally classifying images is heightened by the restricted range of spectral bands, leading to similarities in spectra

among different classes due to the overlap in signatures from distinct features (Zhou et al., 2020).

The effectiveness of utilizing RFs for remote sensing challenges relies heavily on the characteristics of the training data. In the image classification case, RF classifiers are sensitive to spatial autocorrelation of classes and the proportional representation of various classes within training samples (Dalponte et al., 2013; Millard and Richardson, 2015). Complications arise when separating remotely sensed data, as spectral band data are limited to broad wavelengths that can make distinguishing subtle changes in the Earth's surface challenging (Lu and Weng, 2007). Overlapping classes, and the challenge in class separation, can lead to a strength of the RF algorithm, bootstrapping to reduce the impact of outliers and mislabeled data, becoming a weakness. Where without balanced classes, a bootstrapped sample likely will contain examples solely from an individual class. This creates poor outcomes for classification accuracy with the minority class (Chen and Breiman, 2004). Consequently, when employing RFs with spectral data, the challenge lies in handling overlapping classes, with intentional sampling choices being essential for a successful result.

The importance of inter-class balanced training data, meaning all classes are evenly represented in a training dataset, is well documented. Implicit to training data sampling, intra-class variations are just as important to account for. This importance is seen in land cover products, where land classes can be separated into tree cover versus barren areas (Zhou et al., 2020). In fact, Zhu et al. (2016) state that a robust ML model that can classify at large (i.e., global) and small (i.e., regional) scales must capture the variation in each individual class. As mentioned previously, one strength of the RF method is bootstrapping multiple trees, however balanced classes are not the only pitfall that can weaken the ensemble-based method. Homogenous and heterogenous samples of each class are necessary to represent not only the different predictor classes,

but the intra-class variability (Stanimirova et al., 2023). Without this representation, the stability and robustness of random forest model output can decrease.

Stanimirova et al. (2023) highlight intra-class variation as a crucial element in representing individual regions within a global training dataset. By aggregating two decades of global Landsat data for land cover classification, the researchers crafted a high-quality dataset that captures homogeneous examples of land cover types through image analysis. They complemented this process by incorporating heterogeneous land cover signals, identified through unsupervised learning methods. The development of a diverse training dataset necessitated the collaboration of multiple image analysts and significant computational resources to ensure its representation at both regional and global levels. Streamlining the human and computational resources involved would enable more extensive exploration within the training dataset, shifting the emphasis away from dataset creation itself. Ramezan et al. (2019) observes the difficulty of selecting samples from high-resolution remote sensing maps, particularly within extensive regional datasets, a challenge further compounded on a global scale. An ideal solution would involve an unsupervised automated sampling approach that is efficient, reproducible, and capable of generating quality samples.

To assess the capability of an automated method to create a data sample representative of balanced inter-class and intra-class variability, the authors perform a straight-forward land cover classification task—land versus water. Large remotely sensed datasets, both temporally and spatially, offer a clear avenue for assessing the advantages and disadvantages of automating the sampling procedure, and the potential for achieving high-quality image classifications from a training sample representative of both regional and global data distributions.

For completeness' sake, we investigate multiple sampling strategies applied to a global archive of Moderate Resolution Imaging Spectroradiometer (MODIS) data. The main four strategies examined in this research are as follows:

- Simple random sampling:
  - 5 million examples
  - 800k examples, plus deliberate sampling
  - 27k examples
- Automated stratified random sampling
  - 27k examples

where an individual pixel of MODIS data is named an "example". We compare the above-mentioned sampling techniques against a long-standing water mask product, MOD44W (Carroll et al., 2009, 2016). However, this sampling method is not restricted to water masking, and can be applied with tree cover masking, and further beyond remote sensing applications.

1.4 Goals

This dissertation explores identifying AACPs, crucial severe weather indicators, on a pixel scale to emulate human analysts and paving the way for further research beyond the training dataset domain for plume identification. In addition, introduced is a framework for severe hail prediction that adjusts the importance of individual data points before input to an RF. Furthermore, this dissertation examines automatic training data selection from large datasets, focusing on remote sensing data to determine the minimal training sample required to represent vast datasets on global and regional

scales. Overall, the dissertation enhances the understanding of how ML can be applied across various scientific domains, addressing specific challenges in each domain.

Chapter 2

Real-Time Above Anvil Cirrus Plume Identification

In visible imagery, the intense updrafts associated with overshooting tops produce distinct textures and cast shadows on the surrounding anvil cloud, especially when the sun is near the horizon. Meanwhile, AACP temperatures can exceed those of the surrounding anvils by more than 20 K (Brunner et al., 2006). The warm anomalies within the plumes often contrast sharply with the much colder tropospheric anvil IR brightness temperatures (BTs), resulting in recognizable signatures such as "Enhanced-V" (EV, Brunner et al., 2006) or "cold ring" signatures (Setvák et al., 2010). In this study, we use state-of-the-art deep learning methods to offer a comprehensive understanding of the capabilities and limitations of deep learning in analyzing and detecting severe and tropopause-overshooting convective patterns within geostationary imagery

This dissertation integrates multiple satellite-derived datasets into a deep learning model, complemented by expertly annotated hand-drawn labels. The predictor data includes visible (VIS), infrared (IR), and the difference between multiple IR bands (Figure 2.1). Visible data (band 0.64 µm) aids in discerning shadowing effects from cirrus plumes and the bubbling associated with overshooting tops (OTs). IR data focuses on the water vapor (WV) absorption band (6.2 µm), capturing temperatures correlated with OTs and plumes, as brightness temperatures (BTs) are lowest at the coldest temperatures (lowest WV). The IR Difference variable, derived from the difference between the 6.2 µm and 10.8 µm IR bands, typically yields negative values.

However, under specific conditions like very cold cloud tops and temperature inversions, such as those supporting AACPs, the IR Difference variable can show positive values (Schmetz et al., 1997; Setvák et al., 2008). The hand labels consist of two classes, each pixel either within an AACP or outside. Experts hand labeled plumes for 12 days in 2019 including 30 April, 1 May, 5-8 May, 17-18 May, 20-21 May, 26-27 May. The days in 2019 are used for testing, with a separate day of hand drawn labels from 13 May 2020 used for validation.

2.1 Data Processing

Originally on a 2,000 by 2,000 grid of 1 km grid spacing, the input label data are sliced into 128 x 128 grids (determined to be the best based on initial hyperparameter tuning) with an overlap of 8 grid points. The overlap allows for stitching of tiles together to create a whole image at the end of the testing process. The slices must contain at least 10% pixels labeled as AACPs to be included in the training dataset. Each slice is then matched with the corresponding predictor data at that timestep and location. Multiple different percentages of pixels were tested but 10% provided not only a larger dataset for training the deep learning model, but also allowed for sliced labels

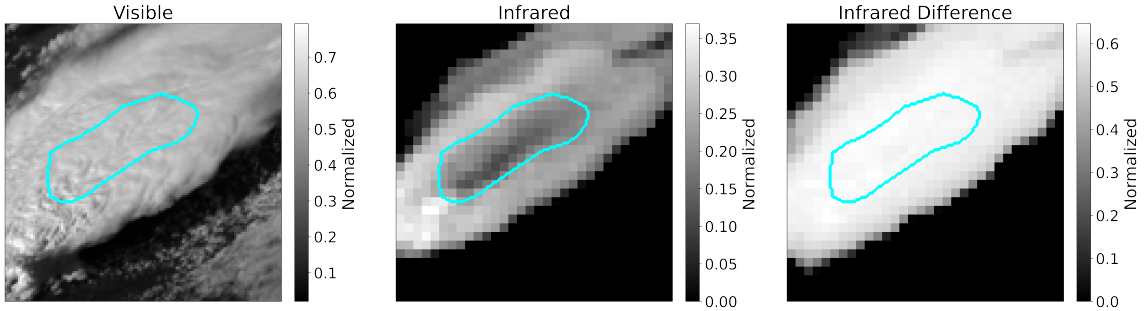

Figure 2.1: Gridded and normalized visible (0.64 µm, left), infrared (6.2 µm, center), and infrared difference (6.2-10.8 µm, right) data from 26 May 2019. Cyan contour is the expert label of a plume for this scene.

with large areas of the patches not containing plume labels. Training a model with more "negative" pixels provided a type of regularization that decreased probabilities in areas where a plume was not labeled, based off the validation dataset. The predictor data are normalized using the minimum and maximum of the data (each data point has the minimum value subtracted and subsequently divided by the range of the full dataset). This method was chosen because of the lack of outliers in the predictor data and the choice between standardization and normalization did not affect the validation results in substantial ways. The training minimum and maximum values are saved and applied to the validation and testing data, following standard ML practices to ensure the model generalizes well from training to validation/testing data. Further, data augmentation is applied to the training dataset consisting of 3407 slices of 128 x 128 grids input to a deep learning architecture. Augmentation is performed using the ImageDataGenerator class within keras, with a rotation range of 15 degrees for each image, a width shift range of 0.2 and a height shift range also of 0.2. These parameters were chosen given previous work on using the ImageDataGenerator class and ranges that would be appropriate for learning about plumes without creating images that are unrealistic.

2.2 Methods

The chosen machine learning model for this dataset is a Unet, given the ability of the Unet to learn spatial properties using convolved weighting parameters but also outputting predictions on the same scale as the input data due to the upsampling layers. More sophisticated Unet models can learn more complex spatial relationships using skip connections between the downsampling and upsampling layers. However, for this study, the regular Unet is employed as a first try. Hyperparameters that create the

most accurate model are chosen using a cross-validation approach, scored based off the correspondence ratio, or the intersection of two datasets, images in this case, divided by the union of both datasets (Stensrud and Wandishin, 2000). As the Unet outputs values between 0 and 1, a thresholding technique is used on the validation data. Values greater than a given threshold between 0 and 1 in steps of 0.05, are transformed to 1. The parameter IOU is determined based off the predicted validation labels converted to 1 and 0 using the thresholding and the best IOU and threshold parameter is saved.

One day out of the total twelve mentioned above is reserved for validation, with each day cycled through until all days are used to validate the model that trains using the other data. Although the data are close together in time (days apart) and space (all taken from West Texas), AACPs do not occur longer than a few hours at most and these events have little overlap, therefore there is little chance for autocorrelation and overlap in conditions in time for the data chosen as validation. However this does have implications to the robustness of the model to identify plumes in different regions and at different times of the year. Nevertheless, the model with the highest validation IOU is saved and compared togother models with different loss parameters and input predictors. This process, also called hyperparameter search, is accomplished over different combinations of hyperparameters to create the most skillful model.

To determine which combination of predictors is best for predicting AACPs, multiple models are created with varying predictors. One model contains all three input variables (VIS, IR, IR difference), another excludes VIS data but retains the IR variables, and so on for all the combinations of input predictors. Both the best and worst outputs (based on IOU score) are shown for certain models (for the sake of brevity). The best and worst cases are examined to indicate where the strengths of the modeling technique lie as well as potential challenges, both important information when using AI models in a real-world scenario. In this study, multiple Unets are compared

because of the different combination of input predictors and the resulting quality of the predictions. Therefore, each Unet will be slightly different because they are tuned for the given dataset. Details of each individual Unet are included in Appendix Table 5.1.

The best model with the various choices of predictors is examined with Class Activation Maps (CAMs, Zhou et al., 2016) to investigate what aspects of the input predictor data the deep learning models highlight, as part of the analysis of this method application in a new domain. For a given image, the CAM algorithm sums the weights learned from the DL model corresponding to the class that is selected. Meaning, the weights that correctly predict a class of 1, given that the class 1 is what the user is interested in, are the only weights applied throughout the network and then summed at the end of the process. This produces an average "heatmap" of where in an image the network is highlighting for its prediction. Typically, CAMs are evaluating an entire image with a single classification in the case of traditional CNNs, however Vinogradova et al. (2020) extended this procedure to be able to use 2D output from Unets by using a "region of interest" parameter. The region of interest in this study would be the areas that a DL model classifies as "plume" labels, and these data alone are applied to the CAM method and then regridded onto the total input predictor shape. While CAMs can be used after the network has been trained and implemented as an evaluation step, the algorithm is also useful in the debugging stage to determine if a DL model is relying upon physically relevant areas within an image, or spurious image noise that may or may not be relevant to a given problem domain.

2.3 Results

The ML plume classifications are exclusively examined in real-time and apply to that specific timestep, with the following time-step being one minute later according to the input remote sensing data. Classifications occur in western Texas. To evaluate the performance of the various Unet models based on their input data, both quantitative and qualitative assessments are examined. Quantitatively, boxplots of ML plume classifications indicate the IOU values for all the different 128 x 128 grid scenes available on 13 May 2020. Qualitatively, two scenes from specific ML models are analyzed visually to investigate further the reasoning behind an individual models classifications. These scenes feature the input predictor variables overlaid with expert truth labels and Unet model predictions, along with class activation maps indicating the basis of the Unet's predictions in each scene. Scene selection was based on the 128 x 128 grid with the highest and lowest IOU. This combined quantitative analysis, covering all available data for the test date, and qualitative investigation provides a more nuanced insight into each selected model, shedding light on their learning processes for potential future deployment or further exploration into model training.

2.3.1 Quantitative

Overall, any combination of variables on average outperforms the Unet models trained with a single variable (Fig. 2.2). The IR-only ML model produces the highest IOU score, yet the model displays the widest range of performance values spanning approximately 0.36 - 0.74 IOU with an average of 0.37. The Unet model trained with VIS-only data outputs a similar average IOU at 0.35 with a smaller range in values (0.32 - 0.6) but overall shows less skill than the IR-only model. The IRDiff-only Unet outperforms the other two single-variable models in average IOU at 0.4, however displays a smaller

range of output values from 0.37 - 0.45. This indicates that the IR-only and VIS-only models can outperform the IRDiff-only trained Unet on individual scenes, however the likelihood of outperformance is dependent on an individual case and not overall model performance.

Of the different Unets trained with a combination of input variables, the IR/IRDiff model outputs the lowest average IOU, 0.44, and skill scores ranging from 0.43 - 0.54. The VIS/IR model produces an average IOU of 0.55, with overall values ranging between 0.48 - 0.66. Falling sligthly behind the VIS/IR trained Unet, the VIS/IRDiff Unet showcases an IOU of 0.51, with a very small range of output values from 0.45 - 0.53. Finally, incorporating all three variables results in ML classifications with a higher average IOU value (0.47) than the IR/IRDiff model (0.44) but performs worse compared to the VIS/IR Unet (0.55) and VIS/IRDiff Unet (0.51). This suggests that introducing the IRDiff parameter does not offer substantial advantage to plume classification when compared to the combinations of VIS/IR. This is potentially due to the documented inconsistency in IRDiff providing distinct character within AACPs, where some events exhibit strong features while others lack noticeable distinctions.

2.3.2 Qualitative

For the qualitative visualisation, it's important to note that higher IR or IRDiff normalized values relate to colder temperatures, while lower values are indicative of warmer brightness temperatures. The darkest areas represent the coldest regions and are more likely to be linked with an OT.

Of the different input feature combinations to investigate, the VIS/IR ML model was selected to illustrate the qualitative performance of the Unet that produced the highest average IOU for the test case. Both the patch with the highest IOU and the

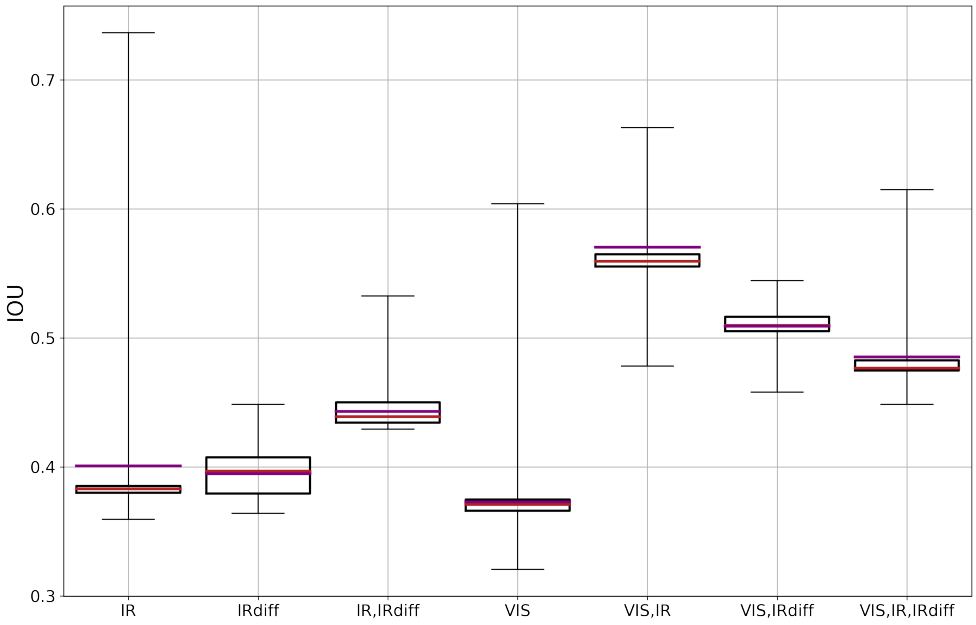

Figure 2.2: Boxplots showing the distribution of IOU values for all gridded data on 13 May 2020. Each boxplot represents the IOU value comparison between the expert labels and the Unets trained using the labeled input features. The rec line on each plot is the average IOU value, while the purple line is the median.

one with the lowest IOU are displayed for a more rounded interpretation of how the Unet processed specific input data. The patch with the highest IOU (Fig. 2.3a,b) reveals a large area of ML plume classification, larger compared to the expert labels. In this scene, the Unet particularly emphasizes AACP-related cloud structures in the visible data regions. Areas with minimal AACP-related cloud shadowing collocated with colder IR temperatures are also positively identified as plume regions.

The bottom row of Figure 2.3 illustrates the output of the segmented gradCAM (2.3 c,d). Regions highlighted in red are considered most crucial in predicting plumes, whereas those in blue are of lesser importance. The VIS/IR model focuses on colder IR temperature especially evident at the top of the scene. The highest importance covers the area nearest to the maximum in normalized IR values, indicating the likely presence of an OT. Another area with high importance, although not highlighted by the expert labels as being a plume, shows the ML model keying in on higher IR values towards the bottom of the scene. This may signify a plume extending beyond the gridded scene or could underscore the significance of AACP edges (and emergence of the anvil IR emission) to the prediction process. The combination of VIS and IR data to a Unet demonstrates how a ML model can highlight pertinent information for AACP classification, leveraging potential OT locations alongside VIS imagery indicating the presence of AACP-related cirrus clouds.

For the scene with the lowest IOU score produced by the VIS/IR Unet, expert labels delineate a diagonal plume identification from the left to the right of the image (Fig. 2.4 a,b). The ML model classifies most of the scene pixels as part of a plume. However, areas devoid of the plume classification include the upper-left region and a section in the middle of the image. Notably, for this patch of VIS/IR data, the majority of the area classified as a plume seem associated with cirrus clouds alongside gradients

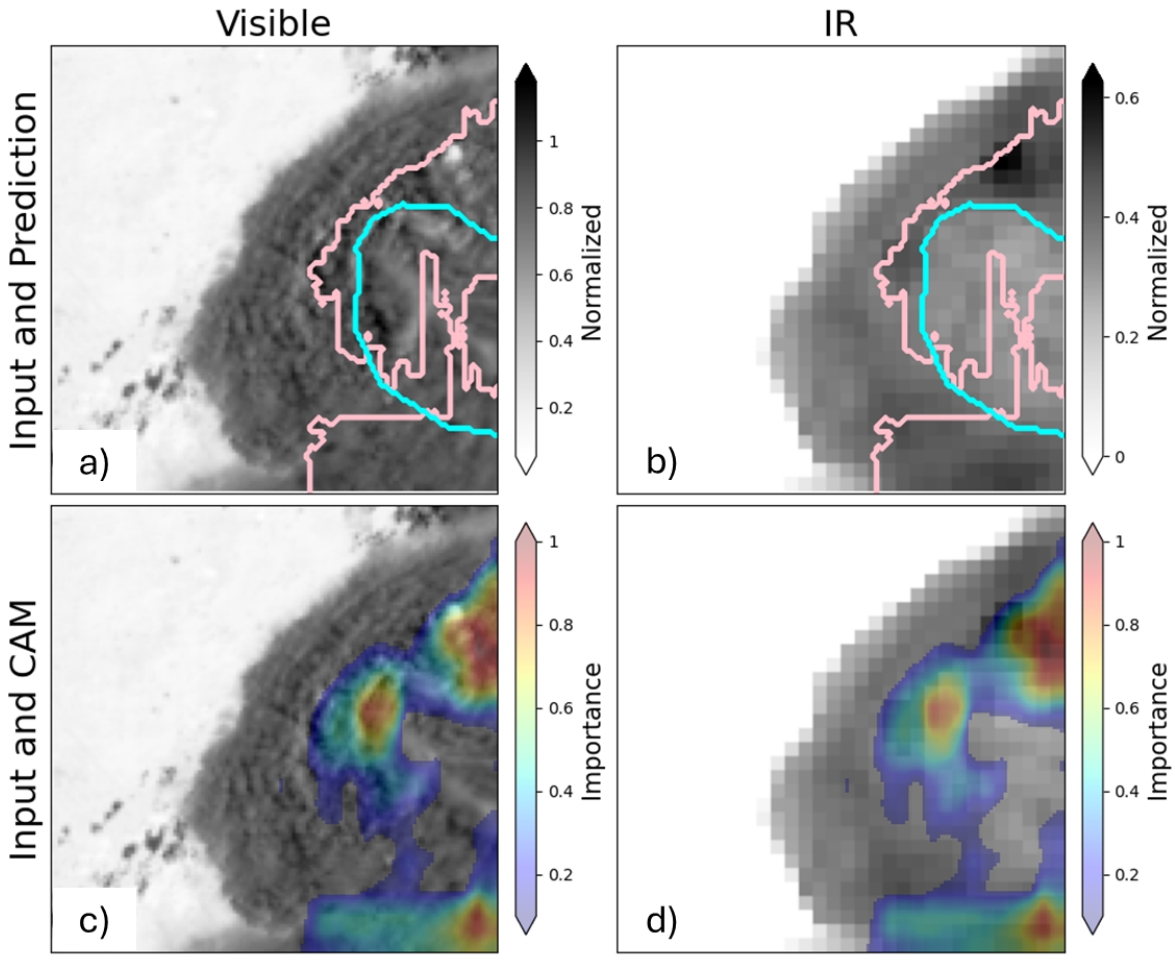

Figure 2.3: Individual features used for prediction with a Unet, overlaid with label (upper plots, cyan) and ML predictions (upper plots, pink). Also overlaid with the input features are class activation maps (lower plots), showing heat maps of importance. Individual case data is associated with the highest IOU score between the label and ML prediction.

of IR values compared to Figure 2.3. There isn't as pronounced a presence of cirrus clouds downstream of an overshooting top as in the scene with the lowest IOU score as compared to the highest IOU scene. This suggests that a neighboring OT may not be captured within the image patch. Given the reliance of the high-IOU prediction on this aspect, the absence of an OT within the scene (although likely present outside the patch) may have influenced this case.

When examining the importance values in the bottom row of Figure 2.4 it becomes apparent that the region of cirrus clouds closest to higher normalized IR values holds the highest importance, where the probability of upward motion and the presence of an OT is heightened (Fig. 2.4c,d,). Another region of notable importance occurs in the regions with high gradients in VIS shadowing, albeit with less prominent gradients in IR temperatures. Based on the assessment of the Unet's performance, this low-IOU scene suggests the necessity for larger scenes to effectively capture extensive areas when gradients lack distinctness. Moreover, employing a smaller sliding scale than 8 pixels could generate more patches for training, enabling the capture of additional segments of large-scale plume environments and ultimately refining plume classifications.

Next, the IR-only Unet is examined because of the large range in IOU values for the test date and overall greatest skill of the Unet models of this study. Reviewing the scene with the highest IOU, there is substantial areal agreement between the expert labels and ML classifications (Figure 2.5 a). In contrast to the previous figure, the segmented gradCAM output indicates the highest importance is in regions with the warmest IR temperatures. Since warm plumes are often characterized by their elevated temperatures resulting from gravity wave breaking and entrainment of downstream warmer temperatures compared to the rest of the anvil, this suggests that the Unet model has learned to associate warmer (lower) IR temperatures (values) with plumes.

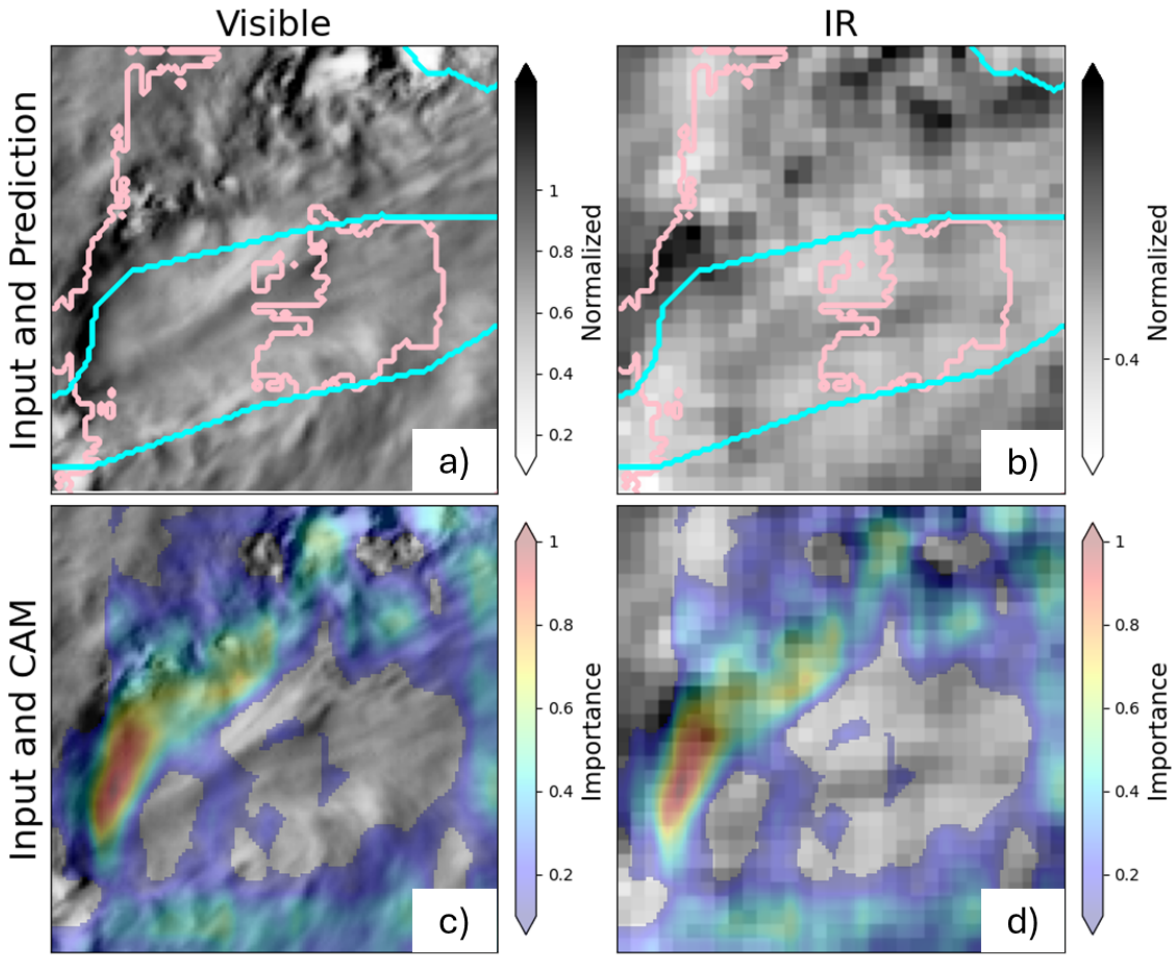

Figure 2.4: Individual features used for prediction with a Unet, overlaid with label (upper plots, cyan) and ML predictions (upper plots, pink). Also overlaid with the input features are class activation maps (lower plots), showing heat maps of importance. Individual case data is associated with the lowest IOU score between the label and ML prediction.

This differs from the VIS/IR model, which emphasized the proximity of cirrus clouds to the colder IR temperatures of the OT.

In the scene with the lowest IOU recorded for the IR-only ML model, the expert labels delineate a plume diagonally spanning Figure 2.6a, while the Unet identifies a smaller region to the right of the expert labels. Peak importances are nearly directly collocated with the ML plume classifications, in an area with relatively warmer IR temperatures. In contrast, the manually labeled AACP broadly encompasses lower IR temperatures, suggesting that this likely also represents a cold plume case. Unlike the VIS/IR model, the IR-only Unet appears to classify images based on the location of lower (warmer) IR values (temperatures), particularly those surrounded by large gradients of higher (colder) IR values (temperatures), as is the case for a warm spot with a cold ring or enhanced U/V signature. One explanation for the disparity in IR-only model outputs could be that the ML model more easily captures warm plumes, a signature that aligns with the pattern observed in the high-IOU scene. This is a potential bias within the expert labels that favor warm plumes, characterized by plume regions with distinctly warmer IR temperatures compared to the surrounding anvil. In contrast, cold plumes may exhibit IR temperatures equivalent to or colder than the surrounding anvil, like the scene with the low-IOU. Since the training dataset does not distinguish between cold and warm plumes, warm plumes may be more frequently labeled due to their more "distinct" IR signatures. It is also possible that these events predominantly involve warm-plume-only occurrences. Murillo and Homeyer (2022) undertook extensive efforts to identify a comparable number of cold and warm plumes across seasons for analysis. Often, one plume type dominates an event.

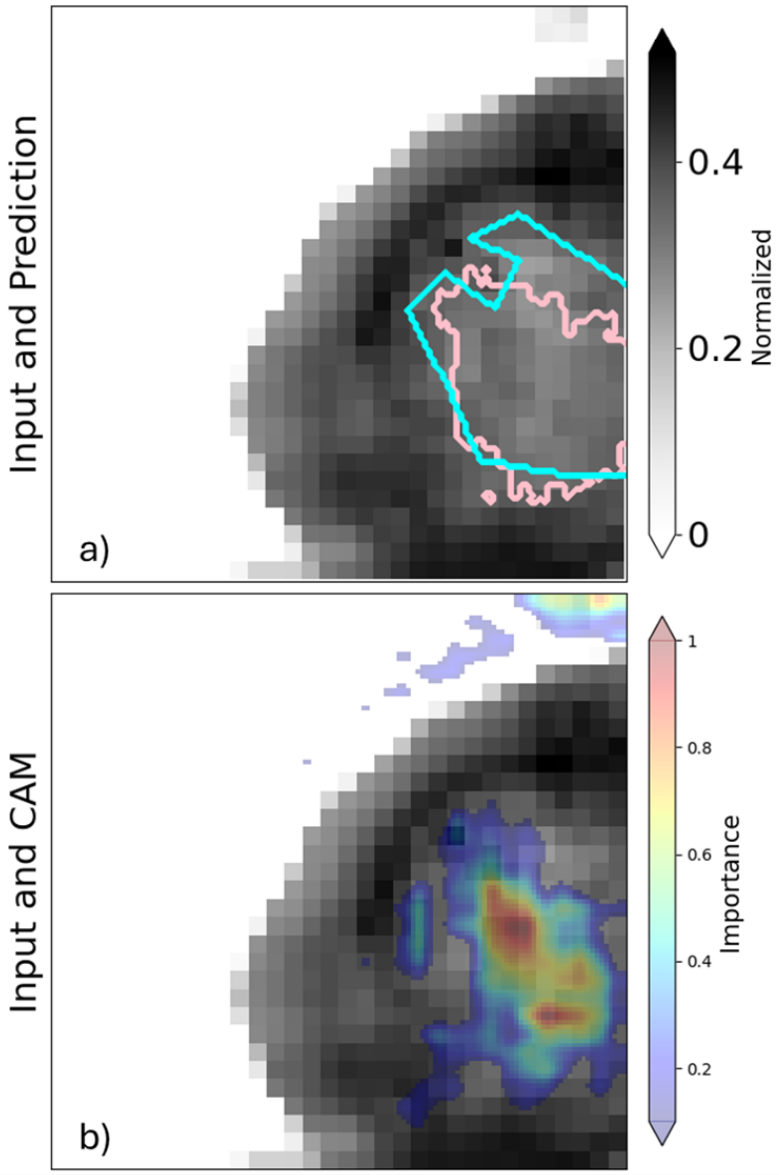

Figure 2.5: The same as Figure 2.3 with a Unet trained with only infrared data as an input feature.

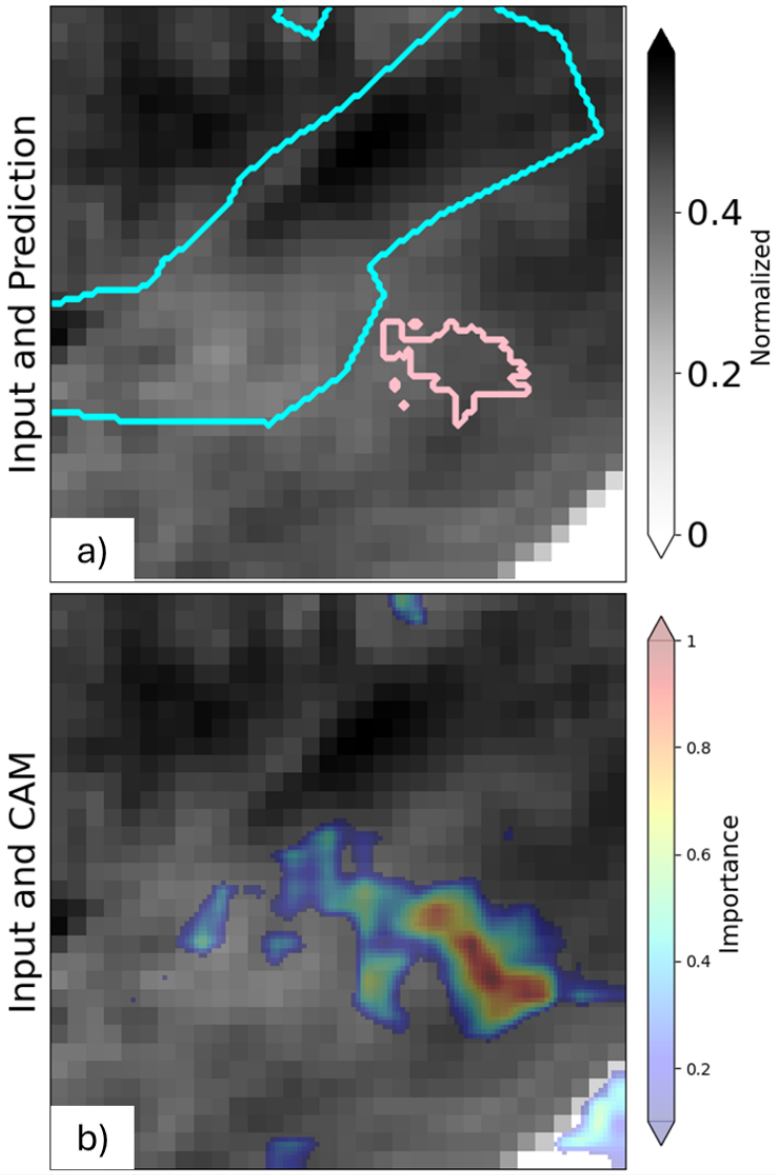

Figure 2.6: The same as Figure 2.4 with a Unet trained with only infrared data as an input feature.

2.4 Discussion

Based on the results above, one input variable alone struggles to train a Unet for consistently skillful plume classification with a limited dataset. For a single scene, the IR-only model performs the best on the test data (IOU of 0.74) while the VIS/IR model on average shows the greatest skill (IOU of 0.55). When warm plumes are present, the IR-only model shows great skill in focusing predictions to smaller areas than the VIS/IR model. A reason for the potentially higher performance ratings with the VIS/IR model is the combination of data that effectively leverages cirrus plume and overshooting top locations with warm plumes.

One caveat of this research centers around cold and warm plume presence, where the type of plume is a strong determining factor of classification success or failure. The lack of differentiation between cold and warm plumes in this study's processed data is an important limitation. A reason for the limitation between cold and warm plume analysis is due to the prevalence of warm plumes over cold plumes, especially highlighted by Murillo and Homeyer (2022). A repository of thousands of cold and warm plume cases could potentially help solve the issue of a ML model to distinguish in plume features however within this research there was a very limited training dataset, only $\sim$ 4000 training images and around 500 testing. One way to address this issue is adjusting how data are pre-processed, such as further augmenting the training scenes with larger ranges of a change in scene rotation. In addition, implementing a smaller window of slicing and potentially a larger area of gridding could enhance training; however, this approach restricts the number of available training scenes. The decision to use a 128 by 128 grid was based on its ability to capture the largest extent of a plume identifiable without excessively reducing the input dataset. Employing transfer

learning to establish a larger repository in a semi-supervised manner is another proposed option. A ML model can initially identify plumes, followed by expert validation of their accuracy, thus expanding the dataset. These annotated examples can then be used to train another model and update it with new labels.

Another caveat of this research pertains to potential errors in the labels generated by experts, some of which were undergraduate students trained by experts for a summer project. Human subjectivity means that one expert may label different regions as containing a plume, compared to another expert. To address this issue, mitigation strategies could involve incorporating a normalization factor to adjust predictions from binary values (1 or 0) to probabilities.

This approach would assign higher probabilities (closer to 1) to pixels closer to the labeled plume areas, gradually decreasing as distance increases but remaining non-zero. Empirical testing is essential to determine the optimal extent and shape of this probability decrease, taking into account factors such as storm motion. A probability decrease that considers storm motion would be particularly beneficial for training with a regression dataset rather than relying solely on binary classification. Further empirical testing is necessary to establish a probability threshold for converting plume probabilities back to binary classifications.

Chapter 3

Day-Ahead Severe Hail Prediction

In this study, the High-Resolution Ensemble Forecast System version 2 (HREFv2) serves as the dataset input for the machine learning models tasked with predicting severe hail occurrences across the CONUS. The HREFv2, a convection allowing model (CAM) ensemble, encompasses various microphysical schemes, four time-lagged members, planetary boundary layer schemes, grid spacing, and other parameters to enhance ensemble diversity (e.g., Jirak et al., 2018; Burke et al., 2020; Loken et al., 2017). Since 2021, 2 members have been omitted from the ML model's training and prediction, as they are set to be replaced in future HREF versions leading to only six of the eight HREFv2 members used for training. As each ensemble member data are trained separately, minimal changes were needed to adjust the forecast algorithm.

3.1 Data Processing

For the mapping of input feature variables essential for severe hail prediction, the ML models utilize the Maximum Expected Size of Hail (MESH, Witt et al., 1998), a product derived from NOAA/NSSL Multi-Radar Multi-Sensor radar data (Zhang et al., 2011; Smith et al., 2016). The dataset is partitioned into training, calibration, and testing sets to ensure model independence and robustness. Training data comprise HREFv2 and MESH records from April 1 to July 31, 2017, May 1 to August 31, 2018, and May 1 to August 31, 2019, totaling approximately 660,000 storm instances. The calibration

set spans May 1 to August 31, 2020, encompassing at least 226,000 storms. The testing set contains 173,000 storms occurring between May 1 and July 31, 2021. Notably, the 2017 training period differs due to data availability during the initial operational phase of HREFv2. Since adding weights doesn't affect data preprocessing, both unweighted and weighted ML models undergo training, calibration, and testing using the same datasets.

As in Burke et al. (2020), this study uses the HREFv2 and MESH datasets as input RFs for severe hail prediction. RFs are chosen for severe hail forecasts due to their speed in training and forecasting, computational efficiency, and cost-effectiveness when used in an operational setting. They excel in predicting rare events compared to linear models (Gagne et al., 2017; Herman and Schumacher, 2018), reduce model bias and variance (Breiman, 2001), and are easily interpretable (Herman and Schumacher, 2018), making them ideal for post-processing. Different pre-processing settings are implemented to address challenges noted in previous ML studies using this model. An object-based framework is applied to identify and track storm objects in both predictor and observational datasets over time and space. Storm objects are initially defined in regions where values exceed a user-defined threshold at a single time step. An enhanced watershed algorithm (Gagne et al., 2017) expands the object until the minimum threshold is met or the area exceeds 100 km^2. Objects smaller than this area threshold are excluded from the processed dataset. Storm objects within 240 km of each other between time steps are merged to form storm tracks. Predictor storm tracks are determined where Max Hourly Vertical Velocities (MAXUVV) exceed 8 ms^{-1}, but additional predictors are necessary for accurate hail formation prediction. Additional variables are extracted from the storm tracks, as detailed in Table 3.1. Statistical ML models require one-dimensional input data, so the standard deviation, skew, mean, max, and 10th and 90th percentiles of each predictor variable are extracted

from the storm tracks. This creates a dataset with multiple values for each feature predicting severe hail. Observations, or labels, are identified where MESH values exceed 12 mm (lowered from 19 mm in previous studies to prevent overfitting and the frequent prediction of severe hail). These changes were made to ensure the ML framework finds enough non-severe storm objects to avoid over-predicting severe hail.

Table 3.1: HREFv2 variables input to ML models (Burke et al., 2020). Geopotential height, U wind, and V wind features are extracted at 500, 700, and 850 hPa. Temperature and dew point temperature are additionally extracted at 1000 hPa. CAPE is convective available potential energy and CIN is convective inhibition.

Variable	**Level (s)**	**Type**	**Variable**	**Level (s)**	**Type**
Max Hourly Vertical Velocity	-	Storm	Geopotential Height	Multiple	Env
Storm Relative Helicity	1 and 3 km	Storm	U Wind	Multiple	Env
Max Hourly Downward Velocity	-	Storm	V Wind	Multiple	Env
Max Hourly Updraft Helicity	2-5 km	Storm	Max Hourly U Wind	-	Env
Precipitable Water	-	Env	Max Hourly V Wind	-	Env
Temperature	Multiple	Env	Surface Lifted Index	-	Env
Dew Point Temperature	Multiple	Env	CAPE	-	Env
			CIN	-	Env

In the final step, predictor storm tracks are matched with observational storm tracks based on the distance criteria specified by Gagne et al. (2017). If an observed MESH track is within a certain distance of a predictor storm track, it is classified as hail-producing (binary). These classifications are used as input for an RF classification model. Additionally, shape and scale parameters are extracted from the matched MESH storm tracks using a gamma distribution. These parameters serve as labels to train two separate RF regression models for predicting specific hail sizes, only for the storm tracks identified as hail-producing by the previous RF classification model. For

more detailed information on the pre-processing method, refer to Burke et al. (2020) and Gagne et al. (2017).

3.2 Unweighted Method

The unweighted ML framework for predicting severe hail from high-resolution data uses RF models, with an isotonic regression model for calibrating RF outputs. Each member of the HREFv2 is trained individually in this framework. Initially, predictor storm tracks and their binary hail-producing labels are inputs to a RF classification model to predict if a storm track will produce hail. If classified as hail-producing, the storm track is then input into two RF regression models. These regression models predict the shape and scale parameters of a gamma distribution to describe the possible distribution of hail sizes for the hail-producing track.

For each hail-producing storm track, MAXUVV pixel values within the storm track are compared to a percentile distribution derived from all MAXUVV values in the training dataset. The pixels in the storm track are then assigned the corresponding percentile value for hail size from the predicted gamma MESH distribution, where the distribution is based on the shape/scale parameters identified using the RF regressor. For instance, if a pixel with a MAXUVV value of 22 m/s is in the 45th percentile of the MAXUVV values in the training dataset, it will be matched with the 45th percentile gamma MESH value. This assigns predicted hail sizes to each grid point, allowing the ensemble to predict hail sizes for storm objects. Storm objects with hail sizes over 25 mm (severe hail) and 50 mm (significant severe hail) are assigned a binary label. The combination of ensemble members predicting severe or significant severe hail for a given pixel is converted into a probability. An isotonic regression model then calibrates these

neighborhood maximum ensemble probabilities, resulting in more skillful and reliable forecasts, with MESH values as the verification baseline.

3.3 Weighted Method

In the unweighted framework, each predictor storm track is assigned a uniform weight of 1. This means that the statistical values extracted from each storm track are equally likely to be classified as "important" by the RF model. By introducing weights, the likelihood that a specific storm track is deemed important increases, thereby influencing the decision tree splitting process more substantially.

Specifically, weights affect the likelihood that a split will occur through the "minimum impurity decrease" parameter. If the improvement in impurity (Equation 3.1) is less than or equal to 0, then the node will split.

$$\frac{N_{parent}}{N_{total}} * (\text{impurity}_{parent} - \frac{N_{right}}{N_{parent}} * \text{impurity}_{right} - \frac{N_{left}}{N_{parent}} * \text{impurity}_{left}) \quad (3.1)$$

Normally the N fields are example counts, however when weights are greater than 1 the counts become a weighted sum. With higher N values for the data points with greater weights, this decreases the improvement in impurity score without needing to add more examples, thus increasing the likelihood of a node splitting. Weighting the input data localizes the ML predictions, potentially producing a more useful forecast, while adding minimal computational expense since the RF hyperparameters and input data remain unchanged. Each storm track can be weighted based on timing information, distance from a specific location, or other priority criteria as needed.

In any of these weighting schemes, a ML model is trained to prioritize different predictor values, enabling it to focus on specific user-defined environments relevant to a particular problem domain. Though this approach is theoretically simple, deciding which data to prioritize can be difficult. Weights ≤ 1 are less likely to be considered important by the RFs. This study applies an exponential decay function to each input value. In Equation 3.2, x denotes the "distance" of a storm object from a reference point. If the reference is temporal (spatial), x could be the number of days (degrees) from a certain date (location).

$$f(x) = e^x \tag{3.2}$$

To increase flexibility, we introduce a parameter α, which controls the steepness of the exponential decay. As shown in Figure 3.1, an α of -0.32 results in the exponential weights dropping to 0 over a shorter distance in time (space). Conversely, an α of -0.1 maintains non-zero weights over a larger time frame (area). The decay function is multiplied by 5 (Equation 3.3) ensuring that prioritized data receive much higher weights compared to data outside the specified time period (region). This multiplication also requires the α/threshold value to approach 0 when the weights equal the natural log of one-fifth. The multiplication factor of 5 was arbitrarily chosen for this study as a means to test if this method can indeed impact training of a RF. Future work empirically validating this parameter could offer even more localized predictions using this method.

$$\text{Storm Weights} = 5e^{\alpha x} \tag{3.3}$$

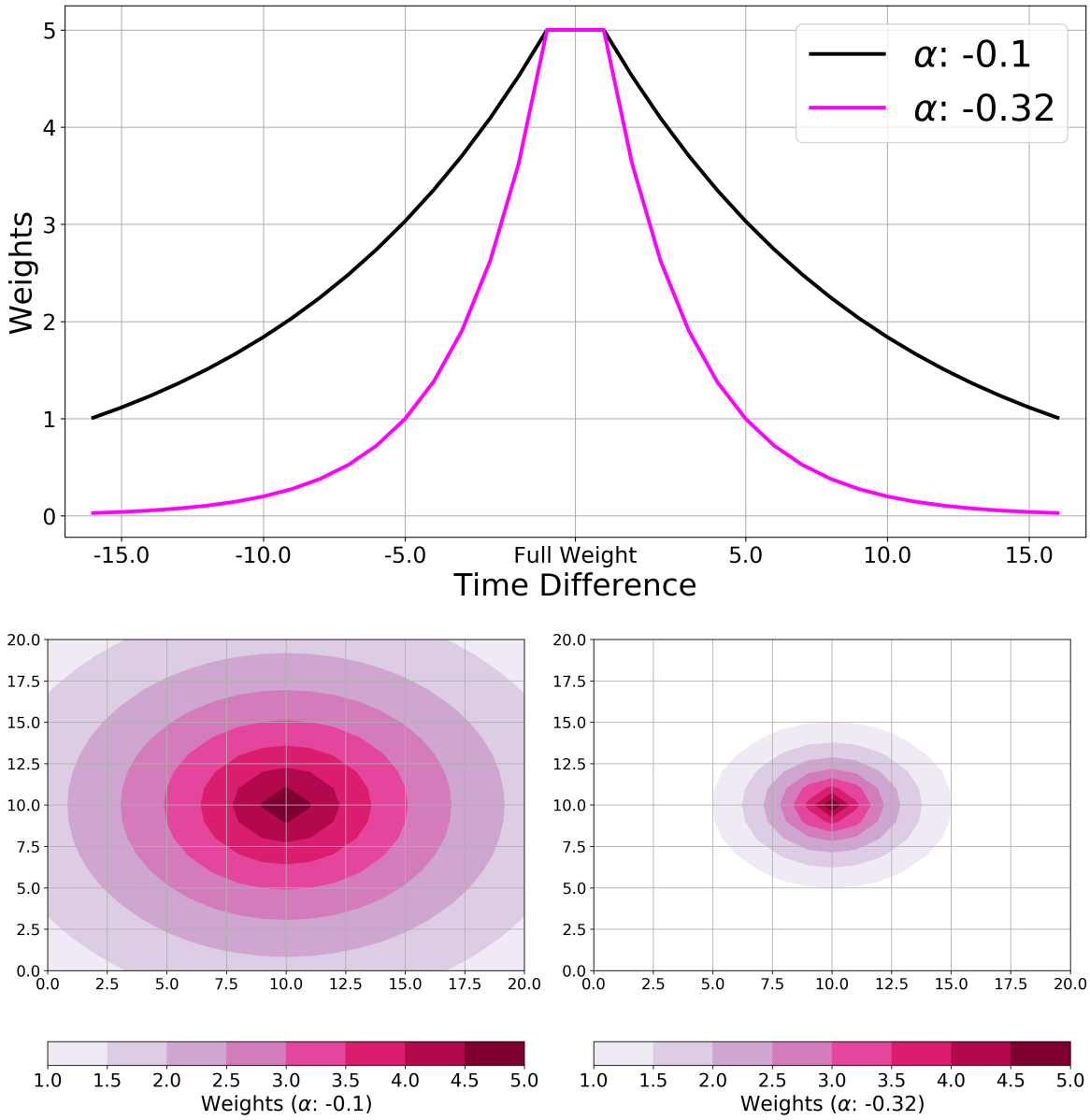

Figure 3.1: Schematic of weighting storm objects in time (upper) and space (unitless, lower) using various α parameters. Storm objects displaced from a given time period or center point are weighted higher if the α value is larger (less negative).

To calculate the α parameter (eq 3.4), we identify the point in time (or location) where the variable used to detect storm objects no longer correlates with itself. This is done by calculating the autocorrelation of the storm object variable over successive time periods (or degrees), such as beginning on 1 May and incrementing one day at a time. The number of days needed to achieve an autocorrelation of 0 is then used as the threshold factor for α and input to equation 3.3

$$\text{Threshold} = \text{Days/Distance when autocorrelation below 0}$$

$$\alpha = ln\frac{1}{5}(\text{Threshold})^{-1} \tag{3.4}$$

Following training, both the weighted and unweighted ML approaches predict whether storm objects in the test set are labeled as hail-producing. Storm tracks with hail designations are transformed from the RF-predicted gamma distribution into hail sizes, with the highest percentile storm object (MAXUVV) value linked to the highest MESH distribution value. Using the predicted hail sizes from each ensemble member, the algorithm generates 24-hour ensemble maximum size and neighborhood maximum ensemble probability forecasts.

3.4 Results

In this study, the data examples provided to the RFs are weighted by time, applying exponential decay to dates outside a single month, as well as spatially around a designated location. The ML predictions, weighted temporally and spatially, are subject to both subjective and objective comparisons across distinct datasets to analyze their performance across varied problem domains. Moreover, additional verification of hail forecasts is conducted against both types of ML predictions, including comparisons

with the SPC day 1 outlook at 1200 UTC and 2-5 km updraft helicity (UH) values >75 m^2s^2, which have been related to severe hail prediction (e.g., Sobash et al., 2016; Gagne et al., 2017). Temporally-weighted forecast evaluations are executed across the entire CONUS from 1 July to 31 July 2021. Spatially weighted forecast evaluations span from 1 May to 31 July 2021 across five states surrounding a reference point in west-central South Carolina (Figure 3.2). This location was selected for its proximity to the BMW US Manufacturing Plant in Greer, South Carolina. Car manufacturers need precise, high-quality hail forecasts to move their products efficiently despite the high costs involved. Improving forecast skill and reducing false alarms can save both time and money for these large manufacturers.

Each modeling scheme is examined using a case study and objective statistical measures pertinent to the aforementioned problem domain. Assessment metrics include reliability measures and a performance diagram, compared against MESH values. The reliability diagram is complemented by the Brier Skill Score (BSS, Brier, 1950), accompanied by a frequency diagram.

3.4.1 Case Study: 28 April 2021

On 28 April 2021, extensive hail struck parts of the southern plains, causing an estimated $ 2.4 billion in damages (Smith, 2010). Strong instability coupled with elevated deep-layer shear created a conducive environment for severe hail in Texas and Oklahoma. The day 1 SPC outlook at 1200 UTC indicated severe hail probabilities of up to 15 % over the southern plains, with a secondary maxima in the Great Lakes region (Fig. 3.3a). Similar to the SPC outlook, a proxy for updraft helicity (Fig. 3.3b) also predicted a swath of higher probabilities from southern Texas through Oklahoma and

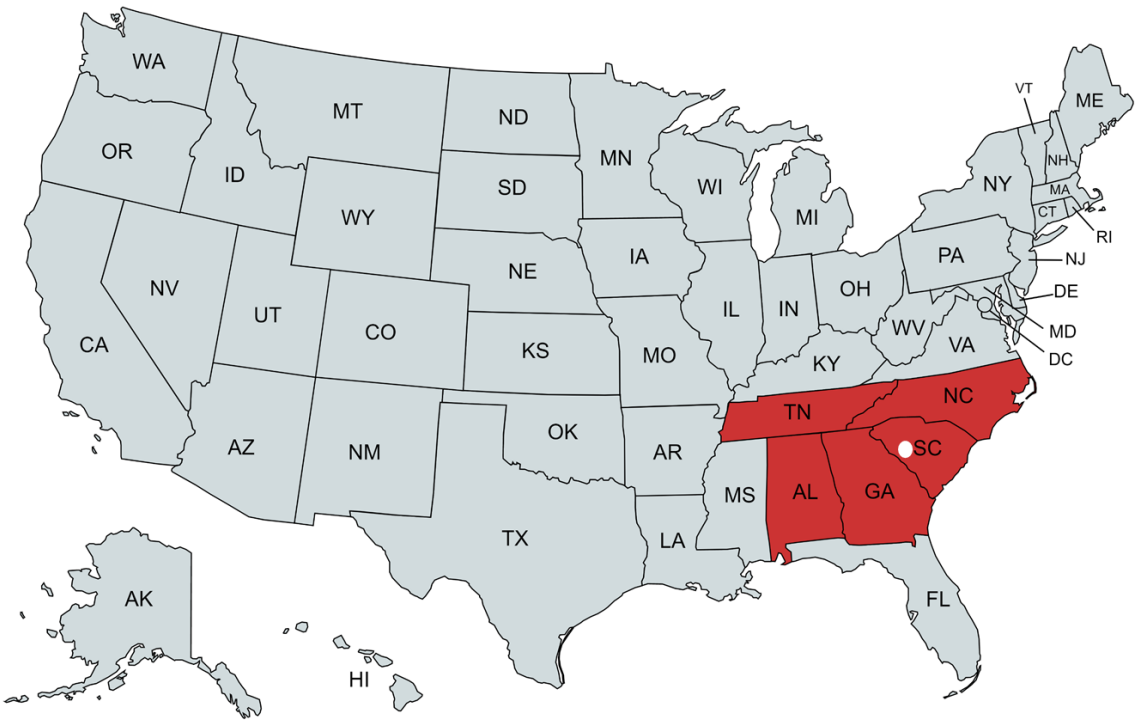

Figure 3.2: Map indicating the five states (highlighted in red) where the spatially weighted model is evaluated between 1 May and 31 July 2021. The white dot is the reference point of the weighted model, located approximately at the BMW US Manufacturing Plant in Greer, South Carolina.

northeastward. The updraft helicity proxy shows probabilities of severe hail surpassing 45 % across substantial portions of the southern plains, with a few localized areas reaching up to 60 %.

The temporally weighted and unweighted ML forecasts (Fig. 3.3c,d) demonstrate similar magnitudes and non-zero probability coverage compared to the SPC outlook. Both ML models indicate smaller areas with a 30 % probability, corresponding to the locations of highest hail threat. While the unweighted ML model effectively identifies regions with significant hail threat, thereby reducing false alarms from the SPC outlook and UH proxy, it fails to capture some hail reports in western Texas and Missouri. On the other hand, the temporally weighted ML model predicts similar probabilities but emphasizes the orientation of hail formation in Texas and reduces the area with 30% probabilities compared to the unweighted ML model. While the weighted ML model misses a few additional hail reports, it effectively pinpoints areas with the highest hail risk and reduces false alarms, particularly in regions where reports are clustered together. This contrasts with the model's less frequent highlighting of sporadic hail events. Generally, the ML models diminish false alarm regions from Missouri to New York compared to the SPC and UH forecasts, albeit at the expense of missed reports. The weighted ML model produces probability magnitudes akin to the unweighted model but highlights a smaller area of heightened hail threats, correctly predicting large hail occurrences from Texas into Oklahoma.

3.4.2 Case Study: 14 June 2021

On June 14, 2021, a quasi-linear convective system developed over the Carolinas, resulting in a few severe hail reports spanning from northern Georgia to western North Carolina. The SPC day 1 outlook at 1200 UTC (Fig. 3.4a) did not extend the 5%

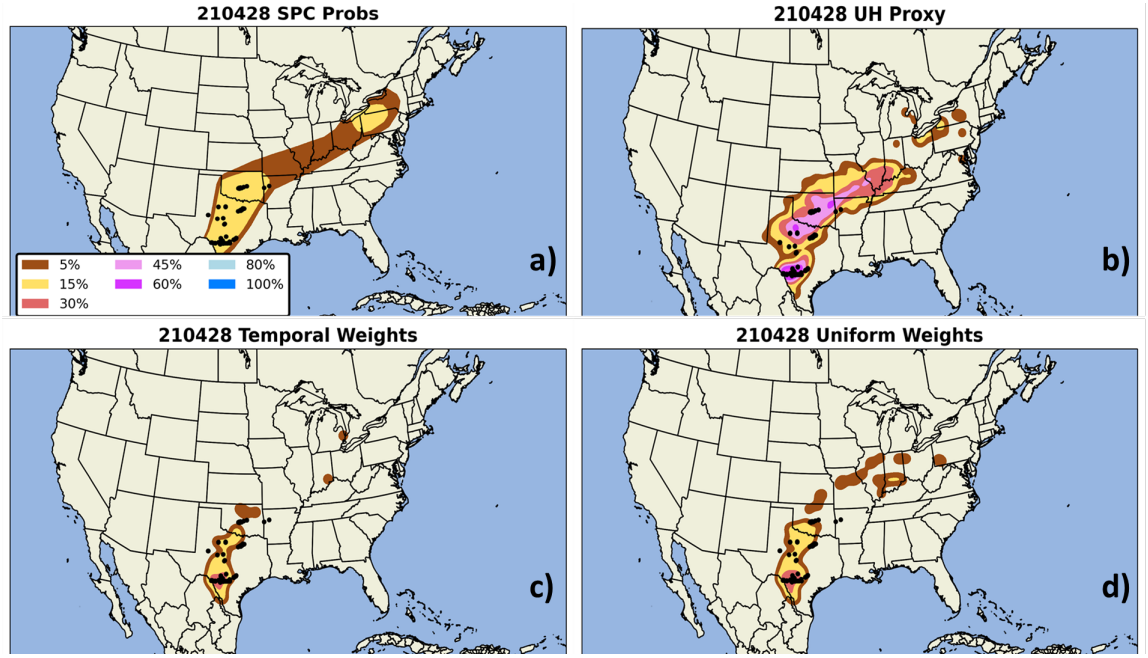

Figure 3.3: Severe hail case study on 28 April 2021 showing (a) the day 1 SPC outlook valid 1200 UTC, (b) updraft helicity (UH) proxy, (c) temporally weighted ML model trained to prioritize storm examples in May, and (d) unweighted ML model output. Black dots are severe storm reports.

probability of hail to the Carolinas due to a lack of significant large-scale forcing, as indicated in the SPC mesoscale discussion for the day. This particular day was selected for a case study evaluation due to the forecast's complexity, given the occurrence of missed storm reports outside of the SPC outlook. Even in an unfavorable environment, the UH proxy (Fig. 3.4b) yielded non-zero probabilities of hail over the Carolinas, extending across portions of the east coast. Since the UH proxy is derived from values in the HREFv2, this suggests that the ensemble did indicate forcing for convective storms, albeit to a limited extent.

The ML forecasts (Fig. 3.4c,d) use the forcing from the HREFv2 to forecast severe hail probabilities of up to 15% in the vicinity of the reported hail events near the Carolinas. The unweighted ML model outlines a broader area with 5% probabilities covering central North Carolina and western South Carolina, along with a smaller zone of heightened probabilities coinciding with regions of increased hail concentrations. In contrast, the spatially weighted ML model narrows down the area of 5% probabilities to only the locations of hail reports and centers the heightened probabilities directly over the reported incidents in central North Carolina. In essence, the ML models capture the subtle forcing and accurately predict non-zero probabilities of hail in areas where severe hail was observed. Moreover, the weighted ML model prioritizes regions with the highest hail threat, reducing the false alarm area in hail prediction while capturing most of the reports near the reference point in South Carolina. This is likely due to the enhanced emphasis on storms in the region within the RF training dataset, enabling the ML model to discern minute atmospheric patterns that other models may miss, resulting in hail formation.

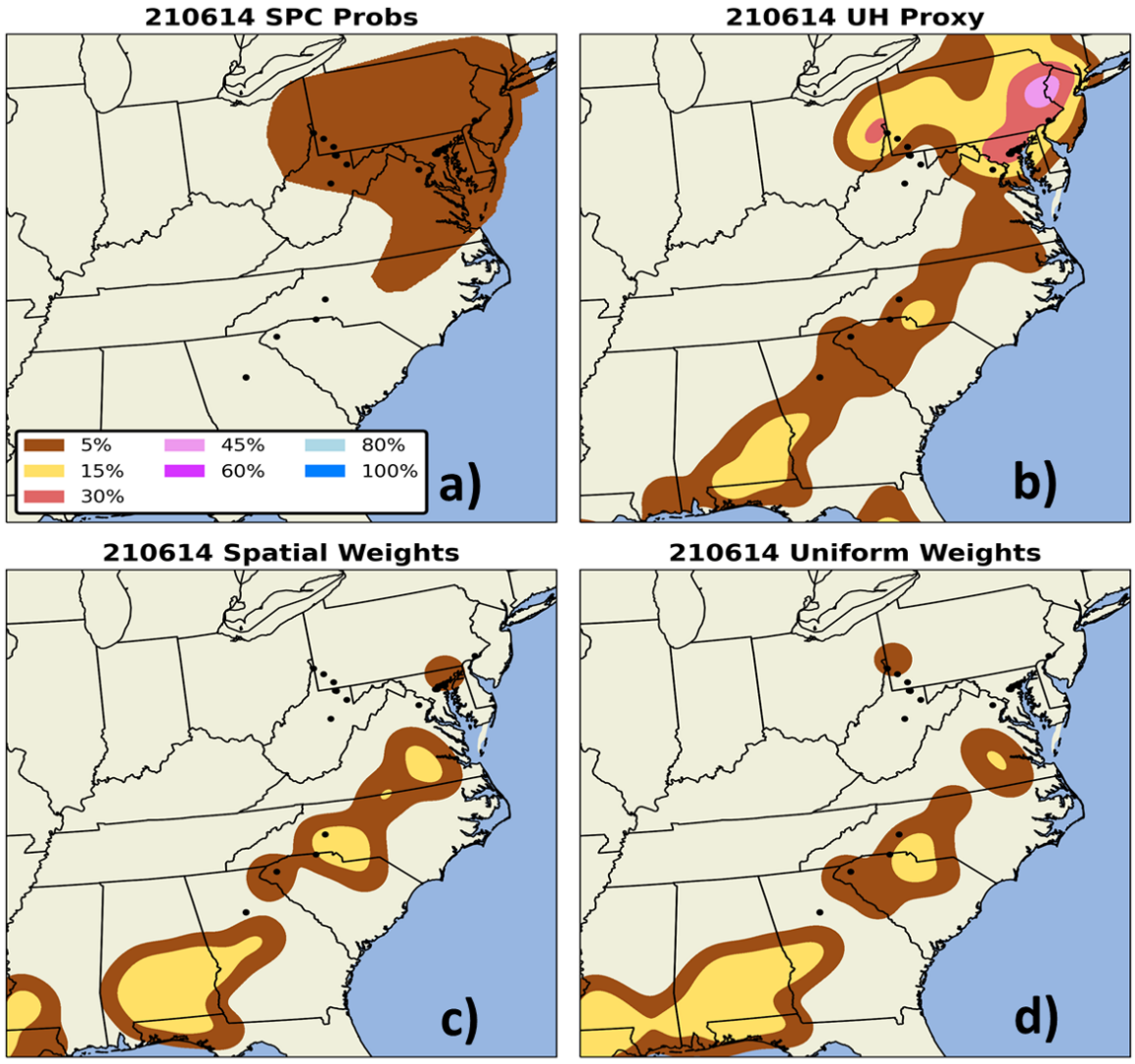

Figure 3.4: Severe hail case study on 14 June 2021, similar to Figure 3.3. The weighted ML model prioritizes storm examples spatially relevant to the reference point in Figure 3.2 instead of temporal weights.

3.4.3 Objective Evaluation: Temporal Weights

When using MESH values as observations, the temporally weighted ML model produces more reliable predictions within the 0 to 20% probability thresholds compared to both the SPC day 1 outlook and unweighted ML forecasts (Fig. 3.5a). However, the UH proxy forecasts demonstrate the highest reliability below 20%, yet tend to overestimate the most among all models above 20%. Furthermore, beyond the 20% probability threshold, the temporally and unweighted ML forecasts exhibit alternating levels of reliability, though they generally remain similar. Consistently, the day 1 SPC outlook underestimates the MESH observations across all available probability thresholds. A parallel trend is observed in the Brier Skill Score (BSS) analysis, where the SPC probabilities display less skill (BSS at -0.039) compared to both ML models. Additionally, the UH proxy exhibits a less skillful BSS (-0.002) compared to the temporally weighted model (0.019) and the uniformly weighted model (0.01). Through bootstrapped BSS analysis, it is revealed that although the difference in skill between the unweighted and weighted ML models is minor, it holds importance. This reinforces the idea that incorporating weights into a RF model maintains high-quality predictions. A detailed evaluation of the value and consistency of weighted severe hail ML models requires more research. However, previous studies indicate that forecasters are more likely to find models incorporating physically relevant information to be more consistent, with equal or better value (Willard et al., 2020; Beucler et al., 2021; McGovern et al., 2022).

Besides the ML forecasts surpassing the SPC outlook and UH proxy in terms of Brier Skill Score (BSS), the ML models generate forecasts with higher Critical Success Index (CSI) and Success Ratio values (Fig. 3.5b). While the UH proxy exhibits higher Probability of Detection (POD) values for each probability threshold, this is mitigated by low success ratio values. Both ML forecasts demonstrate a similar trend of POD and success ratio values, with the temporally weighted model producing higher POD

values and slightly higher success ratios for a given probability threshold, resulting in larger CSI values, particularly between 15% and 25%. Overall, the ML forecasts outperform the SPC probability outlook and UH proxy concerning MESH observations, with the temporally weighted model yielding slightly more skillful forecasts, likely due to reduced false alarms. While the statistical differences between the ML models are slight, they are more pronounced in qualitative assessments

3.4.4 Objective Evaluation: Spatial Weights

In order to focus the results on the region with the highest spatially weighted storms (Fig. 3.2), the ML forecasts, SPC probabilities, and UH proxy are assessed over a smaller domain than the entire CONUS. Below 15%, a consistent pattern emerges between the two weighted ML models, wherein the spatially weighted ML model yields more reliable forecasts compared to the unweighted model, while the UH proxy demonstrates the most reliable forecast (Fig. 3.6a). However, unlike the assessment of the temporally weighted model, the SPC probabilities produced over a smaller region near the southeast offer more reliable forecasts than both ML models below 15%. Above 15%, the ML models exhibit oscillation in terms of which forecast is more reliable, while the UH proxy consistently overestimates across all probabilities above 15%. Despite differences at lower probability thresholds, the Brier Skill Score (BSS) values of the various modeling outputs exhibit very similar patterns compared to the temporally weighted evaluation. The spatially weighted ML model achieves the highest skill score (0.025), closely followed by the unweighted model (0.024), then the UH proxy (-0.001), and finally the SPC probabilities (-0.026).

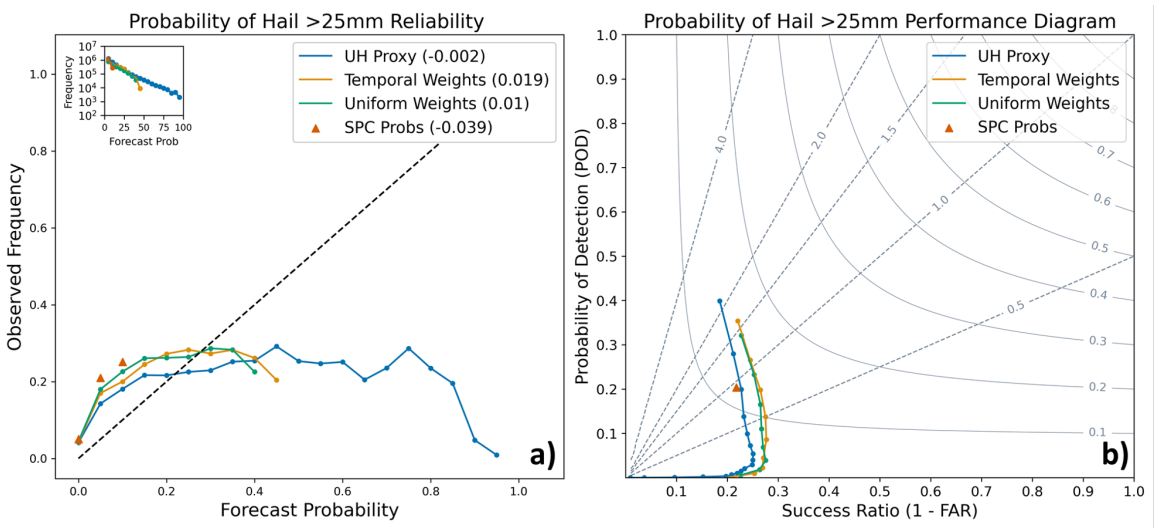

Figure 3.5: Quantitative verification of ML forecasts, updraft helicity (UH) proxy, and SPC outlooks using MESH as observations. Reliability (a) and performance diagrams (b) are calculated over the CONUS in July 2021. Reliability diagram includes the Brier Skill Score (BSS) in the legend, probabilities are labeled on the performance diagram.

The performance diagram (Fig. 3.6b) reveals a significant disparity between the two weighting schemes, with both ML models surpassing the UH proxy and SPC forecasts. However, beyond the 10% probability threshold, the spatially weighted model notably outperforms all other modeling types, including the unweighted model. Nevertheless, the UH proxy variable indicates higher Probability of Detection (POD) values for each threshold compared to the ML forecasts. However, similar to the assessment of the temporally weighted model, these values are counterbalanced by exceedingly low success ratio values. This likely indicates that the UH proxy generates larger areas of non-zero probabilities compared to both ML models, leading to a higher likelihood (evidenced by the maps) of false alarms compared to the ML forecasts. Nonetheless, smaller areas of non-zero probabilities result in more misses, penalizing the ML forecasts, albeit the spatially weighted ML model produces slightly higher POD values above 5% than the unweighted ML forecast. In general, the ML forecasts generate fewer false alarms than the UH proxy and SPC forecasts, with the spatially weighted

model enhancing the success ratio and reducing false alarms more effectively than the unweighted ML model.

3.4.5 Interpretation

For forecasters, trust hinges on two vital aspects: the model's performance and the ability to grasp its internal processes (McGovern et al., 2019a). In this research, we use multi-pass permutation variable importance to pinpoint the most critical predictor variables of a trained ML model (Lakshmanan et al., 2015), thereby enhancing our understanding of the learned patterns. Unlike single-pass variable importance, this method takes into account correlations between variables by permuting important variables while sequentially selecting other significant variables. For further details on this method, refer to McGovern et al. (2019b). In this study, the skill metric distinguishing the most important variables for a classification task is the area under the curve (AUC), indicating how effectively a classifier distinguishes binary classifications. Although regression RFs are part of the ML forecasting process, only the classification models are evaluated for brevity. Thus, the variables displayed were most important in classifying a storm as hail-producing. A total of 237 different predictor variables are assessed for each HREFv2 member from data spanning 1 May to 31 July, 2021. The number of predictors varies from the original 29 detailed in Section 3.1 because each variable is derived using statistical approximations (mean, max, min, etc.). The important variables of each member are averaged across the entire ensemble, resulting in a total ensemble variable importance for each trained ML model (Fig. 3.7).

The top five important variables input to the May weighted model include 700 hPa dewpoints, hourly maximum of upward vertical velocity from 100 to 1000 m above ground level (AGL), and the hourly maximum 10m AGL V-component of the wind

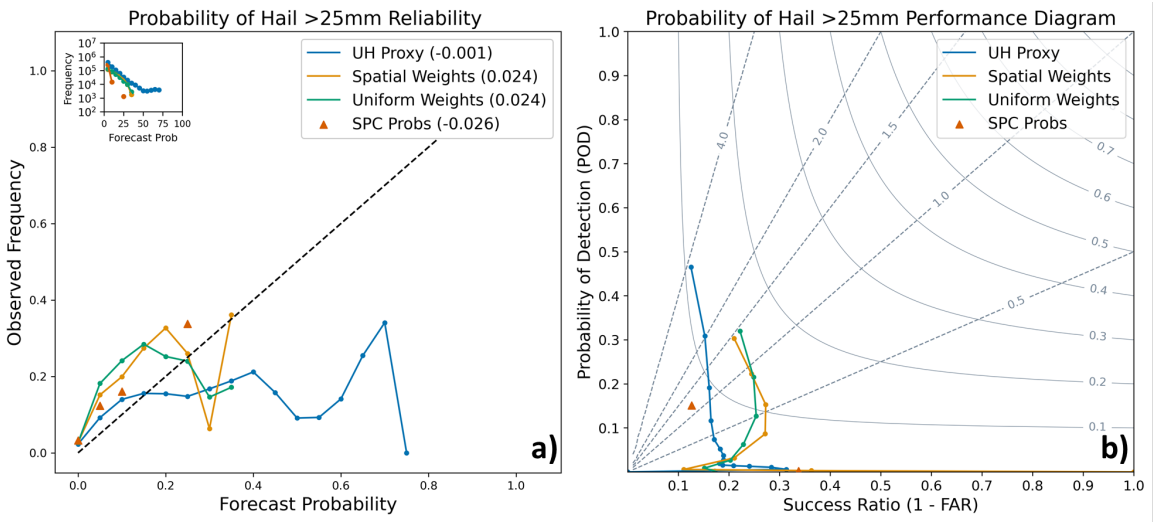

Figure 3.6: Similar to Figure 3.5 with spatial weights instead of temporal weights. Only data points within the highlighted red states in Figure 3.2 are verified. Reliability (a) and performance metrics (b) are calculated between 1 May to 31 July 2021.

(Fig. 3.7a). Notably, the May ML model is the only one that highlights dewpoints as important for classifying hail-producing storms. Alongside other significant variables like updraft velocities and surface north-south winds, this indicates that the ML models recognize the importance of moisture from the Gulf coast, as well as a minimum threshold required for hail storms. Conversely, the July temporally weighted ML model (Fig. 3.7b) emphasizes the hourly maximum wind's 10m AGL U-component, 0-1 km storm relative helicity (SRH), hourly maximum of downward vertical velocity from 100 to 1000m AGL, 700 hPa U-component of wind, and 0-3km SRH. Unlike the May ML model, the July model finds the east-west winds at various levels and storm helicity more crucial, indicating that shear is more significant for classifying hail storms in July compared to those in May. The variations in important variables between the two temporally weighted models suggest that the RFs have learned different patterns for hail classification. The July-trained RF focuses on vertical motions, while the May-trained RF highlights thermodynamic variables.

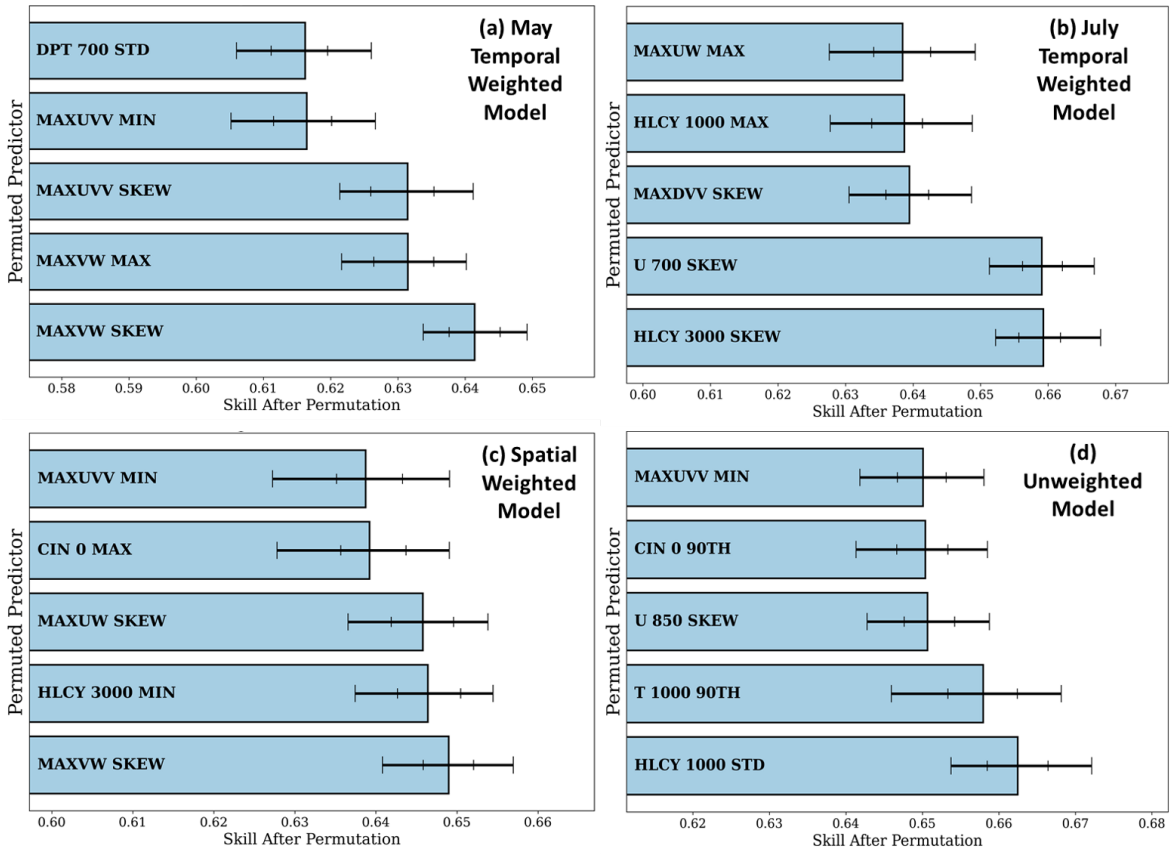

Figure 3.7: Multipass permutation variable importance results for the (a) May weighted, (b) July weighted, (c) spatially weighted, and (d) unweighted ML models using Area under the curve as the skill metric. The original (unpermuted) skill is 0.67 for each model. Each variable is bootstraped 100 times over a third of the 2021 data, with error bars indicating the 5th, 25th, 75th, and 95th percentiles of the bootstrap. The variable names include one of 29 input HREFv2 variables and the statistic applied to the storm objects found most important.

The spatially weighted model (Fig. 3.7c) highlights the most important variables as the hourly maximum of upward vertical velocity from 100 to 1000m above ground level (AGL), surface convective inhibition (CIN), 10m AGL U-component of the hourly maximum wind, 0-3 km storm relative helicity (SRH), and 10m AGL V-component of the hourly maximum wind. Unlike the temporally weighted models, the spatial model emphasizes the importance of both surface wind velocity components and mid-level shear. Notably, a minimum threshold value proves crucial for classifying hail storms in the South Carolina region (chosen for its proximity to a major car manufacturer, see Figure 3.2), similar to the May ML model. However, the inclusion of surface CIN highlights the importance of inhibiting factors to convection alongside convective variables themselves. On the other hand, in the unweighted model (Fig. 3.7d), the top variables include the hourly maximum of upward vertical velocity from 100 to 1000m AGL, surface CIN, 850 hPa U-component of wind, 1000 hPa temperatures, and 0-1 km SRH. While factors inhibiting convection are significant in both the unweighted and spatially weighted models, the unweighted model uniquely emphasizes low-level temperatures as crucial. While a minimum threshold of updraft velocities is important in classifying hail storms across all ML models, the unweighted model exhibits the most overlap in error bars among different variables. This suggests that, among all trained ML models, the top five variables in the unweighted model should be considered as a group rather than individual variables, each with varying degrees of importance.

Generally, the unweighted model accentuates multiple environmental variables pivotal for convection initiation, whereas the May ML model prioritizes moisture return. Conversely, the July model zeroes in on vertical motion, and the spatially weighted model highlights wind directions significant for classifying storms as hail-producing. Collectively, these models illuminate varying facets of the hail formation process contingent upon the weighting scheme employed. This indicates that the ML models

assimilate differing patterns contingent upon the weighting assigned to input examples.

Chapter 4

Representative Sampling of Global Geospatial Data

Large remotely sensed datasets, both temporally and spatially, are advantageous for assessing the capability of an automated method to create a data sample representative of balanced inter-class and intra-class variability. The data used for this task are the surface reflectance bands from MOD09GA/Q, with MOD09GA 500m data resampled to match MOD09GQ at 250m. At the pixel level, MOD09GA bands 3, 4, 5, 6, 7 and MOD09GQ bands 1, 2 were extracted with land or water labels based on MOD44W (Carroll et al., 2009, 2016) collection 6 classifications. Using these two datasets in conjunction it was possible to generate a training data set as large as desired (billions to trillions of examples if needed) and with evenly balanced classes. The daily availability of MOD09GA/Q and global availability of MODIS data in general made this an ideal dataset to derive a training sample that was both geographically and temporally diverse. From here, a "sample" is indicative of the examples from one of the four sampling strategies mentioned previously.

4.1 Data Processing

Using random sampling we generated a Python parquet file with over 5 billion class-balanced examples that was used as a base from which to select samples for each experiment. Specifically, a sample was created that was randomly subset from the overall

billion example dataset down to 1% of the data, resulting in a sample with 5 million random examples (referred to as sample 5mR). Sample 5mR, also class-balanced, comprises an even distribution of examples pulled from each highlighted tile in Figure 4.1. Another sample of the MODIS data contained $\sim$ 800 thousand class-balanced examples (referred to as sample 800kRD), most chosen randomly but with additional examples that were deliberately chosen by expert from regions (still within the highlighted tiles) that struggled with accurate water classifications in previous MOD44W versions. These specific tiles were chosen for their global geographic dispersion and the diverse array of land v water types (i.e., rivers, basins, small lakes, etc.)

Sample 5mR stands as the backbone of the clustering algorithm described in this study. The relatively smaller sample, rather than the full 5 billion example dataset, is used for the clustering analysis for more effective computing time. While clustering algorithms can handle high dimensional data, past 4 dimensions the results of the clustering analysis were difficult to investigate. Using expert knowledge and empirical trials, the initial predictors chosen were MODIS spectral bands 1, 2, 7. These bands verified as the most skillful prediction of land cover without losing accuracy, and most relevant to the land cover classification domain per expert knowledge. An additional variable, Normalized Difference Vegetation Index (NDVI), was also included due to the information added in areas of dense vegetation where the other reflectance bands struggled. Other MODIS bands were not included because of noisy data (bands 5,6), redundancy in the need for additional visible data (band 3), and the advantage band 1 has when dealing with sediment laden water. We tried other indices but found them, empirically, to not provide substantively more skill than the bands we were already using. Specifics of the clustering algorithm applied to these bands can be found in the next section.

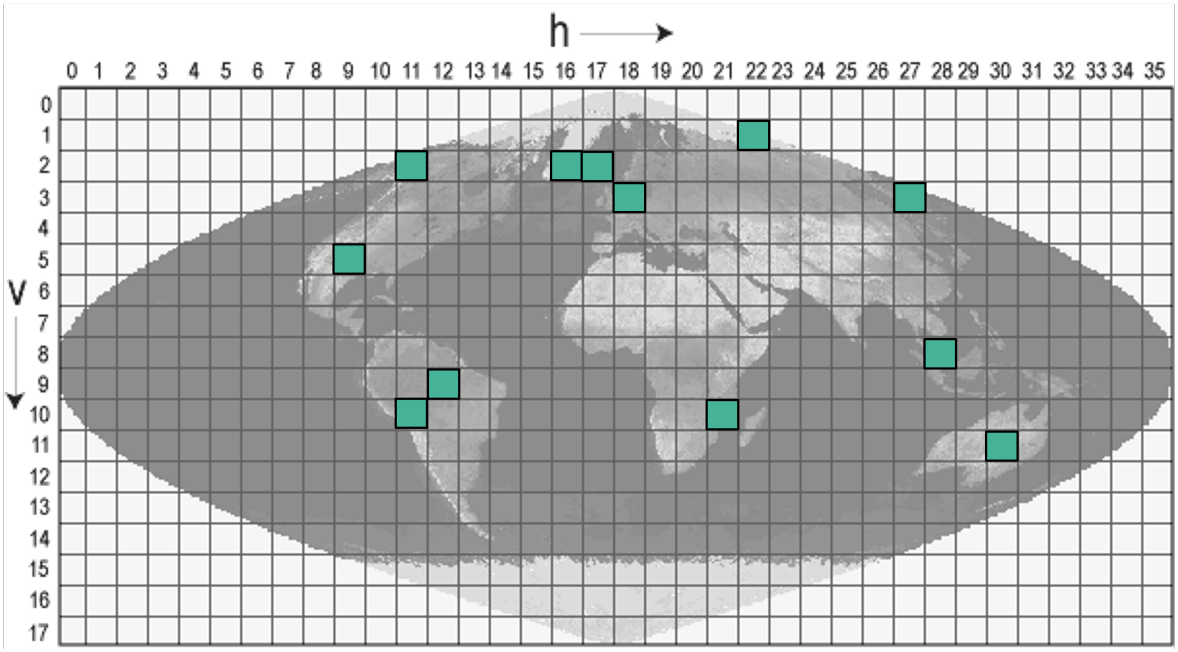

Figure 4.1: Tiles h09v05, h11v02, h11v10, h12v09, h18v03, h16v02, h17v02, h30v11, h28v08, h27v03, h21v10, h22v01 with data extracted from 2001, 2006 and 2019

4.2 Methods

Two initial RF models were individually trained using the 5mR and 800kRD samples, respectively. No data-processing went into the two samples and instead they were input to a RF classification model as is and validated using 5-fold cross validation. Five folds were chosen as they are a standard number for validation, with little difference occurring when tuning the model with other variations in folds. Twenty-five trials of tuning were accomplished, with the RF classifier having the highest validation f1 score of the total 25 trials, along with the hyperparameters, saved and used for testing on select tiles. This ML process of tuning and training is repeated for each individual RF classifier trained with the different samples in this study, resulting in 4 different RF classifiers. The hyperparameters of each RF classifier can be found in the appendix.

To examine ways of representing the variability and diversity of the 5mR sample, the authors investigated multiple different clustering techniques. The initial test went

the simplest route: apply a gridded approach, where examples within a certain range for each variable are grouped together. This approach was difficult to account for all different combinations for only two features, with the complexity not outweighing the potential benefits with more features. The next iteration of the clustering process involved pre-packaged algorithms.

Certain clustering methods were not efficient when applied to the size of the 5mR sample, where trying to cluster a 4-dimensional sample with over a million examples was not feasible with the Spectral (Jianbo and Malik, 2000) and Birch (Zhang et al., 1996) algorithms (Fig. 4.2a).The gaussian mixture model (Blei and Jordan, 2006) algorithm was able to handle the number of examples in the 5mR sample, however they did not produce meaningful clusters as debated by experts (Fig. 4.2b). Although Taşdemir et al. (2015) describe the k-means algorithm as having poor performance with remotely sensed data, due to highly spherical outputs (Xu and Wunsch, 2005; Gonçalves et al., 2008), in this work the authors discovered through empirical processes that k-means provided the best clusters with efficiency, regardless of sample size (Fig. 4.2c).

Although k-means clustering proved to be beneficial for the 5mR sample, one downside that accompanies the algorithm is the user defined number of clusters. Through empirical trials, 15 clusters were chosen to segment each label, land or water (Fig. 4.3) Clusters between 5 and 30 were explored, however visual inspection indicated that less than 15 clusters combined too much data while more than 15 clusters broke up the existing clusters without more information retained within the smaller clusters.

After expert deliberation, water labeled clusters with outlier values $> 10{,}000$ in the visible light spectrum, and the subsequent variables associated with such examples, were dropped. This caused the entire cluster with visible reflectances ranging from 4,000 to $> 10{,}000$ to be excluded from the study. Removing outliers manually with

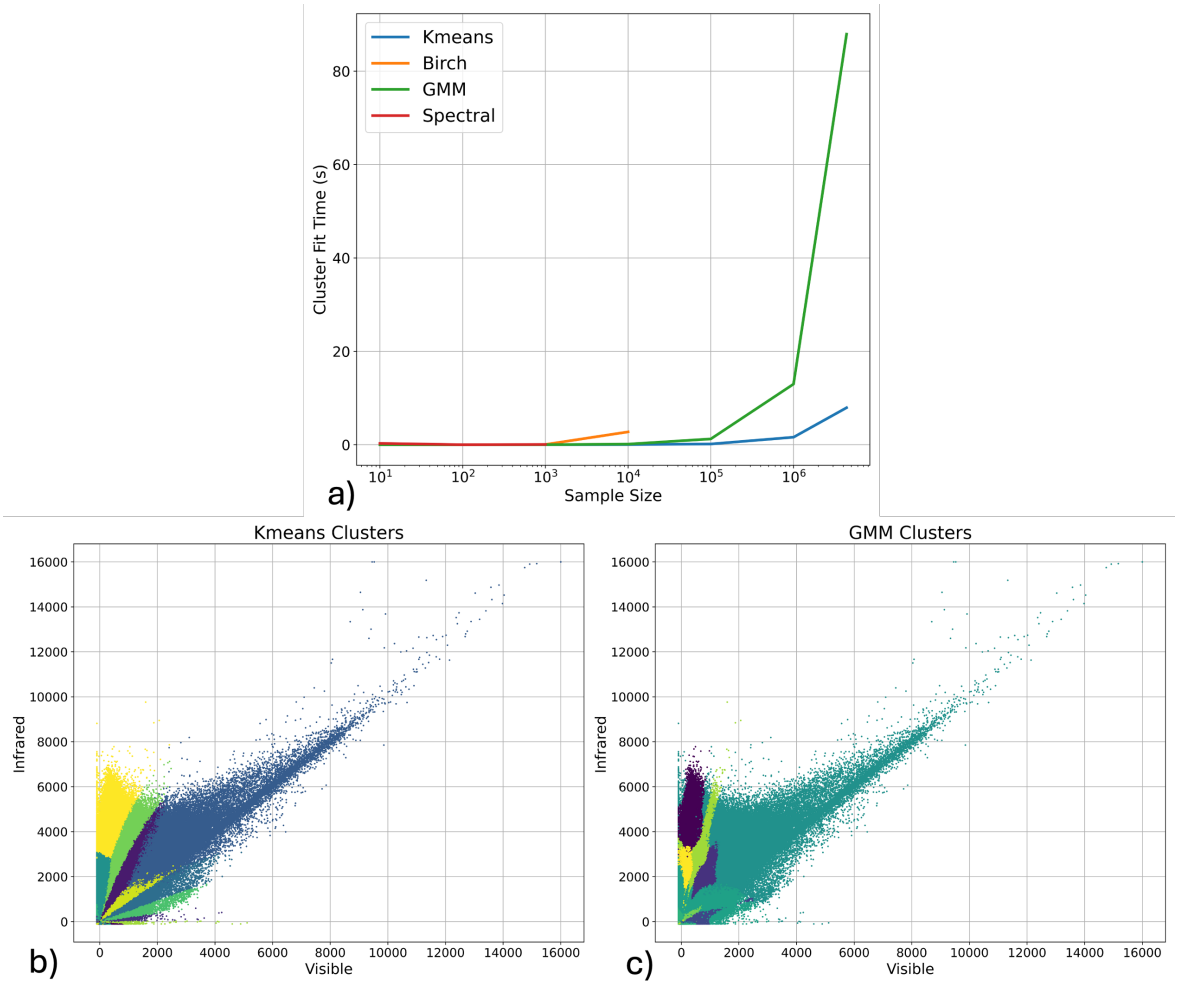

Figure 4.2: Sample size versus clustering time elapsed, each applied to the 5mR sample for 5 clusters (a). The two best clustering algorithms are applied to the whole 5mR sample with 15 cluster. Those include the K-means (b) clusters and Gaussian Mixture Model (c) clusters, where each color represents a different cluster group. All four MODIS bands are clustered, but for visual representation only bands 1 (visible) and 2 (infrared) are displayed.

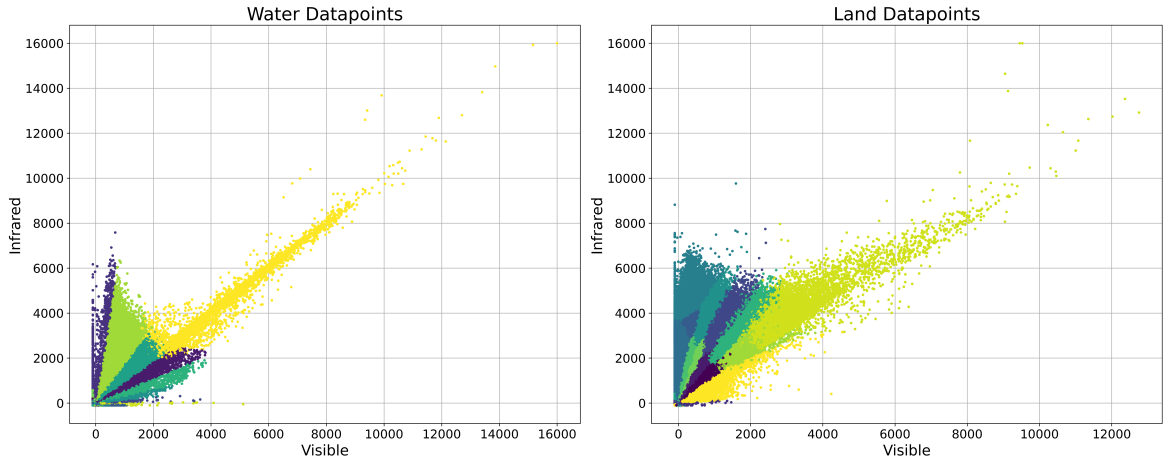

Figure 4.3: Kmeans clusters applied to land (right) and water (left) classes, separately. Different colors indicate different clusters.

the 10,000 threshold for the 5mR and 800kRD samples did not result in any changes in classification output, unsurprisingly, as the RF classifiers trained with more data are vastly more robust to outliers (Zhu et al., 2016). To maintain balance, one cluster from each land and water label are removed. Removing one of the land clusters that are well represented created a balanced sample, where empirical efforts indicate that keeping all land clusters slightly reduced the skill of a RF classifier for this problem domain.

Once clusters that represent the 5mR sample spectrum are established, the question became how to use the clustered data. Including all examples within each cluster would result in simply training another RF with the 5mR sample, but smaller and less robust to outliers and mislabeled data. Instead, a stratified random sampling approach proved the most appropriate for this data and problem domain. The smallest cluster size was used to determine the threshold for choosing random examples within each cluster. This created evenly balanced clusters and resulted in a much smaller training sample as the smaller clusters could range from around 900 to 1000 examples. In this way, each individual cluster of the labeled class was represented without higher weight towards

any individual cluster (Figure 4.4). The clustered data, now referred to as sample 27kC, contained a total land cover training set of around 27k examples. A second method was tested that maintained the proportional size of each cluster by randomly choosing a percentage of the total number of examples within each cluster. Further testing did not indicate any substantial difference in classification output with a change in the percentage threshold, indicating that intraclass balancing is just as important as interclass balancing for skillful classifications.

Finally, to determine if the advantage of the clustering method over simply subsetting the 5mR sample, a fourth sample was constructed with the same number of land and water examples as the 27kC sample. Differing from the clustering approach, the examples within the fourth sample were chosen from the 5mR sample at random. This last sample is hereafter referred to as 27kR. Outliers in the visible data for the water labels are included when selecting the random examples for the 27kR data, as empirical investigations indicated that removing the outliers decreases model skill. Both the 27kC and 27kR samples were input to separate RF classifiers and tuned/trained using the same method as the RFs with the 5mR and 800kRD samples as input. In total, 4 different random forest outputs are evaluated for the differences pre-processing makes on classification skill, specifically the representative-ness of each sample.

Table 4.1 includes the compute time for tuning and training the respective RFs, showing that the training with the 5mR and 800kRD samples took approximately one hour (5mR) and 20 minutes (800kRD), whereas the smaller samples took less than 5 minutes (27kC and 27kR). This large speed up showcases the efficiency of using smaller samples, however the need for representative samples is at the forefront in image classification when limiting data.

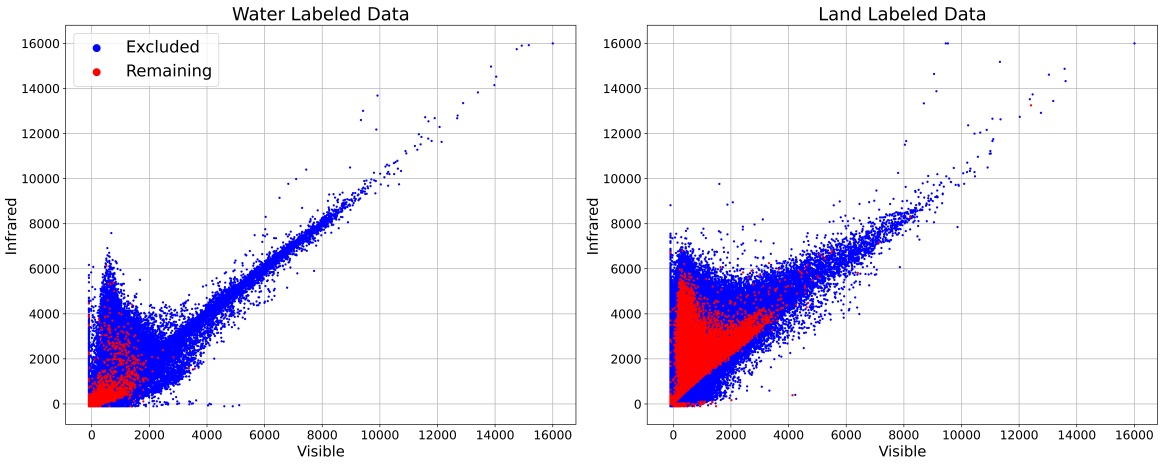

Figure 4.4: Kmeans clusters applied to land (right) and water (left) classes, separately. Different colors indicate different clusters.

Input Sample Name	Num of Examples	RF Tuning/Training Time
Master dataset	5,000,000,000	-
5mR	5,000,000	1:11 hour
800kRD	800,000	20 min
27kR	27,000	1:38 min
27kC	27,000	1:23 min

Table 4.1: Different samples explored in this study, with their respective size of examples and the time elapsed for tuning and training each individual RF. Accomplished on Explore HPC https://www.nccs.nasa.gov/systems/ADAPT run on top of 40 Intel Xeon Gold 6248 CPU @ 2.50GHz

4.3 Results

Exploring the various samples provided as input to the individual RF's reveal multiple differences. Figure 4.5 illustrates each sample, combining values from both water and land-labeled examples. Reflectance values from bands 1, 2, 7 span from -100 to $\sim$ 16,000, and NDVI values range from -30,000 to 30,000 (or -1 to 1 then multiplied by a factor of 30000). This analysis will primarily focus on the surface reflectance in bands 1, 2, 7 as NDVI is a combination of bands 1, 2 and for brevity the NDVI analysis will be excluded. In the 5mR and 800kRD samples, examples show nonzero frequencies for the majority range of values, whereas the smaller samples (27kR and 27kC) are associated with nonzero frequencies up to $\sim$ 8000. In the 27kC sample, this reflects part of the clustering process where outlier values $>$ 4,000 for water examples in the visible light spectrum are not kept. Although outliers are not explicitly dropped for the 27kR sample, the majority of values still lean towards lower reflectance.

Distinct disparities emerge when comparing the 800kRD sample with the other samples. The 800kRD sample not only comprises randomly selected examples from the 5mR sample but also includes additional examples chosen by experts. These expert-selected examples are particularly noticeable with the secondary peak appearing in bands 1 and 2 around 10,000 reflectance that is not seen in the other samples. Potential variations in the performance of RFs with different input samples may be attributed to each samples' data spectra, however in cases where distinct differences in data are not evident, the importance of very small variability in values becomes highlighted.

Finally, differences between the 27kR and 27kC samples are less noticeable until $\sim$ 6000, where the samples exhibit frequency differences ranging from hundreds to thousands of examples. Despite the seemingly modest numerical discrepancies, these

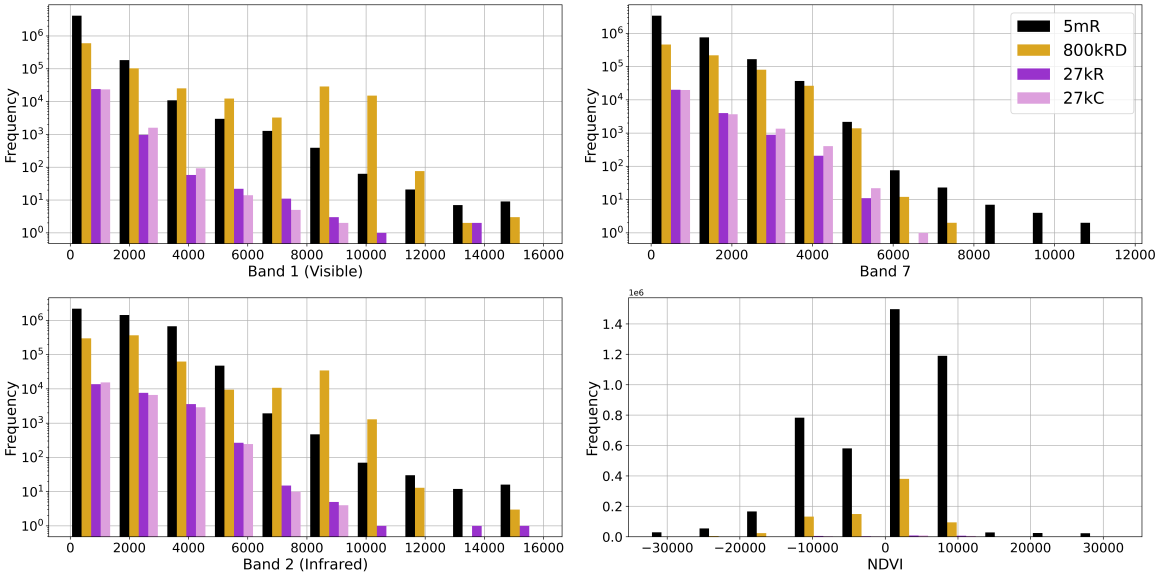

Figure 4.5: Frequency diagrams of the input predictors, Visible (upper left), Band 7 (upper right), Infrared (lower left), and NDVI (lower right). Different colored frequencies indicate the data input for each different RF classification model, with the samples 5mR (black), 800kRD (gold), 27kC (pink), 27kR (purple). Both labeled classes shown.

frequencies are important given 27kC and 27kR samples are limited in examples. Frequency differences are a mixed bag, some input predictors maintain higher frequencies in 27kC data at values > 6000, while others suggest the higher frequency of 27kR sample (Figure 4.6). One interesting difference between the two samples are their range of NDVI values, where the 27kR sample shows most examples logging around -10,000 to 10,000, and substantially dropping off in frequency outside of these values. Meanwhile, the 27kC sample shows higher frequencies between -10,000 – 10,000, but with much higher incidence of examples outside this zone compared to the 27kR sample. The NDVI examples are on the order of hundreds higher in frequency with the 27kC sample.

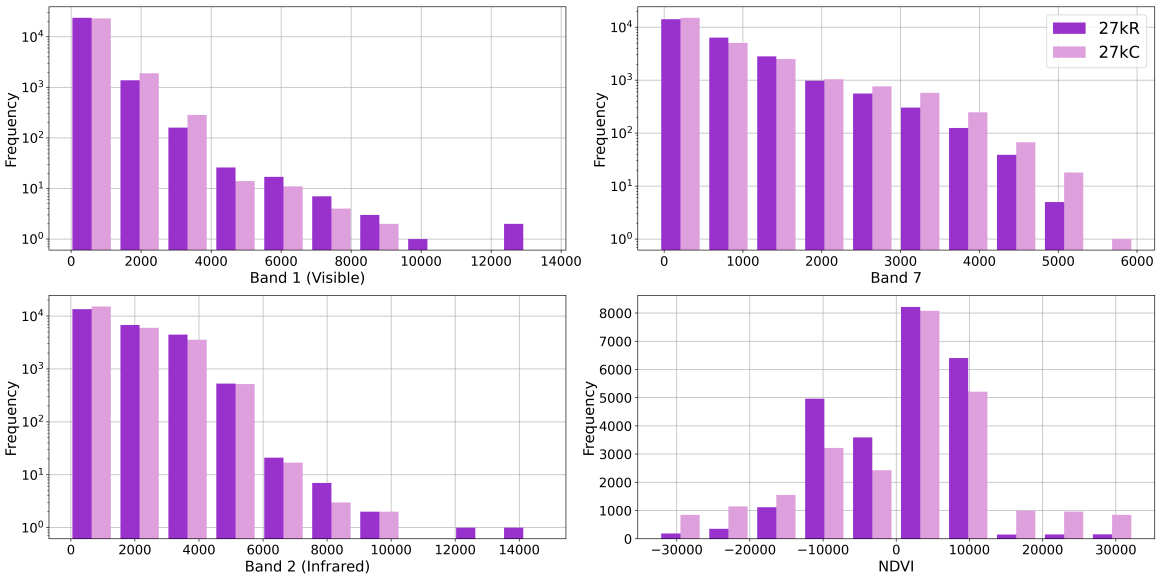

Figure 4.6: Same as Figure 4.5 but only showcasing the 27kR and 27kC samples.

4.3.1 Quantitative

Statistical measurements were accomplished for annual land cover classifications in tiles h12v09, h09v05, h22v01, and h21v10 (Fig. 4.7) for the years 2006 and 2019. These classifications are more likely to identify long-term changes in water sources compared to daily/weekly/monthly changes in rivers, lakes, etc. In addition although years/tile overlap for the data used in training and testing, the shear size of MODIS data removes worry of independence contamination. All of the RF models investigated in this study produce classifications with accuracies, Matthew Correlation Coefficient (MCC) values, and f1 scores at 1.0, the highest performance possible, indicating that the MODIS dataset in general is a good fit, as well as the data originally sampled in the 5mR sample, for this specific land cover classification.

Across the different tiles and years, the 27kC trained model outputs classifications with skill scores between 0.9 - 1.0 for accuracy, MCC, and f1 (Fig. 4.8). The next model with comparatively high skill is the 5mR trained model, with tile classifications

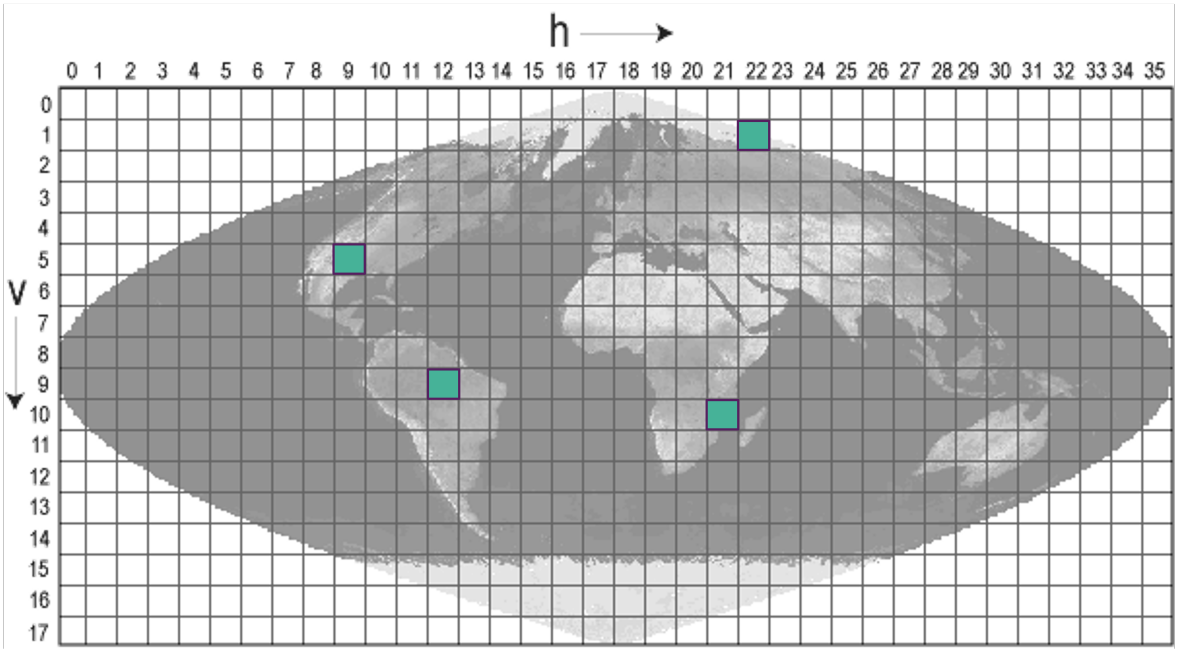

Figure 4.7: Tiles h22v01, h21v10, h12v09, and h0905 examined for years 2006 and 2019.

ranging from 0.95 - 1.0 in accuracy, 0.6 - 1.0 MCC values, and f1 scores from 0.55 - 1.0. The 27kR trained model produces classifications with accuracies in the range of 0.84 - 1.0, MCC values from 0.42 - 1.0, and f1 scores at 0.35 – 1.0. Finally, the 800kRD trained model outputs classifications with accuracy measurements ranging from 0.82 - 1.0, MCC values ranging from 0.4 – 1.0, and f1 scores in ranges of 0.3 - 1.0. Every model shows skill scores reaching 1.0 across the tested tiles and years, however unlike the other models, the RF trained with the 27kC sample demonstrates scores with the least variability, all very close to 1.0. The other models show greater variability in MCC and f1 values depending on the tile and year. Accuracy is similar across the different samples and may not be the best indicator in skill differences.

In deeper analysis of the classifications for each tile/year, it was discovered that in general the number of true negatives either balanced or far exceeded the number of true positives for each tile/year combination, inflating accuracy. Recall was very

high for each RF model classifications in the year/tile combinations, however precision was lower for the RFs trained with samples 800kRD, 5mR, and 27kR in both years for tiles h12v09 and h09v05. This lower precision is reflected in the lower F1 scores of the 800kRD, 5mR, and 27kR models, and is explored qualitatively in the next section. Lower MCC values are also attributed to the h12v09 and h09v05 tiles, however the classifications for the h12v09 tile are the least skillful for all the trained models besides the 27kC trained model. Further analysis of the h12v09 tile, and a comparison to a tile with higher skill scores, are detailed in the next sections.

4.3.2 Qualitative

As mentioned in the section covering the quantitative analysis, the classification output for the h12v09 tile (Fig. 4.9 a) in 2019 showed high recall with low precision, as well as lower MCC values. In a region that is potentially difficult to obtain cloud-free images (Oliveira et al., 2016), which affects the MODIS data selection algorithm, the RFs trained using the 5mR, 800kRD, and 27kR samples all show visually large false positive areas (Fig. 4.9 b). The f1 score analysis very clearly lines up with the visual inspection of the 2019 h12v09 tile, where the 800kRD model (yellow) outputs the lowest f1 score (lowest precision), followed by the 27kR (purple), and finally the 5mR (black) models, as compared to the MOD44W classifications (white). Diverging from the other three RFs, the 27kC model (Fig. 4.9 c) outputs water classifications nearly identical to the MOD44W output, with few pixel-wide false positives at the edges of spherical shaped water bodies as well as peppering of incorrect water classifications throughout the forested area of the Amazon.

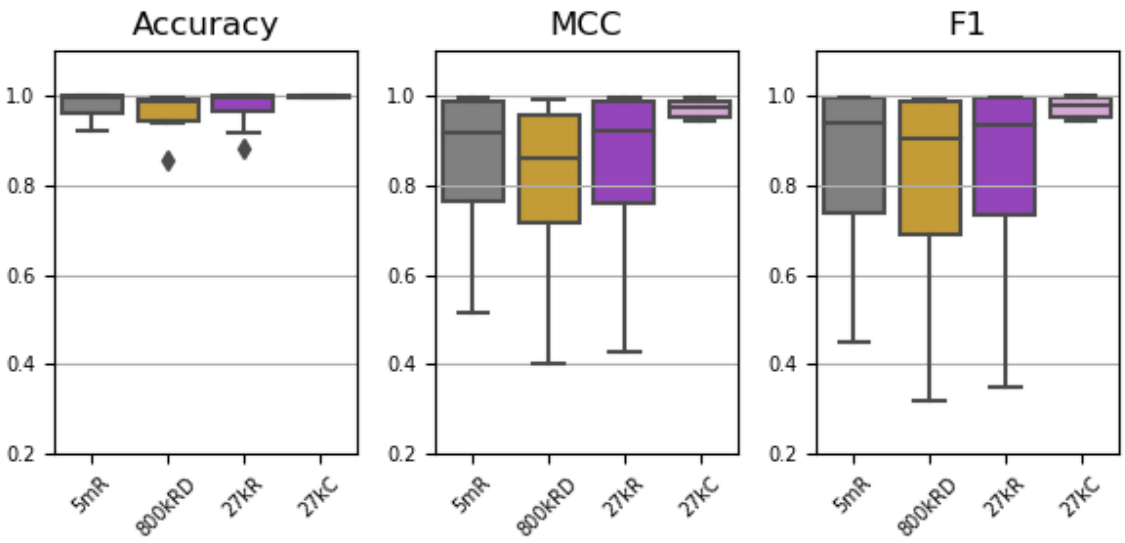

Figure 4.8: Box plots of Accuracy, Matthew Correlation Coefficient (MCC), and F1 score for RF trained with the different sampling strategies. Statistics computed over testing tiles in years 2006/2019. Each RF trained with a different sample is evaluated.

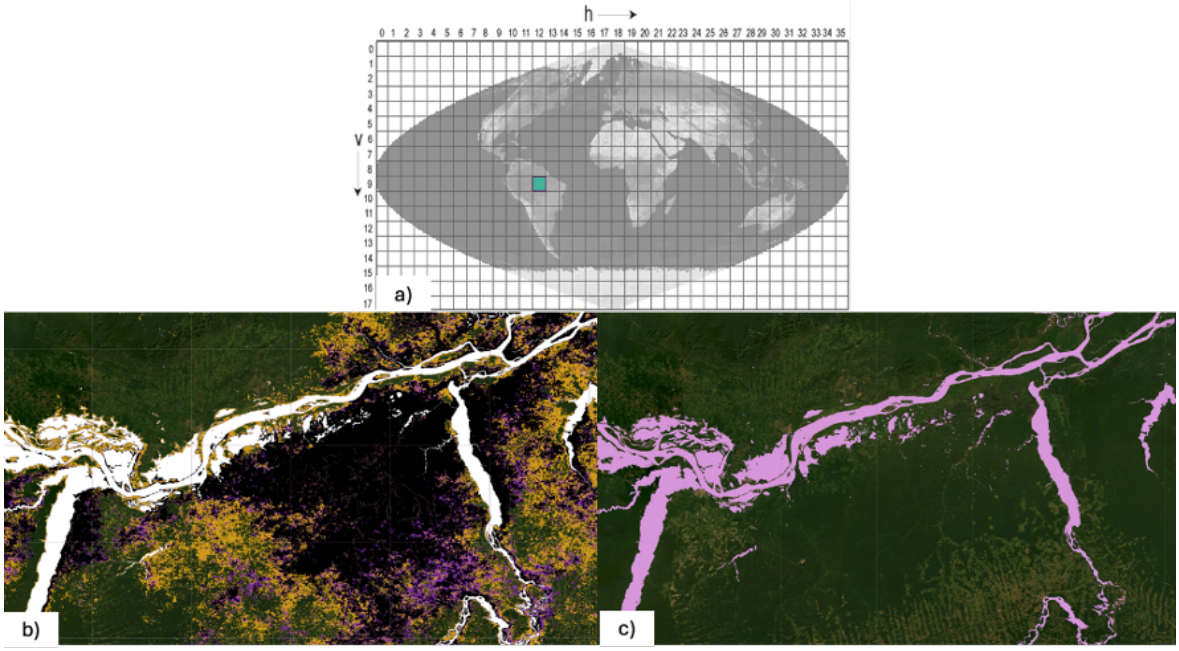

Figure 4.9: Tile h12v09 from 2019. The baseline target, MOD44W, is indicated in white. Also included are outputs from RF models trained with the 5mR (black), 800kRD (yellow), 27kR (purple), and 27kC (pink) samples.

Similar to the tile in the Eastern Amazon, 2019 classifications for tile h22v01 (Fig. 4.10a) show more prevalent 800kRD model (yellow, Fig. 4.10b) water outputs compared to the MOD44W output (white). The 27kC (pink, Fig. 4.10c) output is similar to the MOD44W classifications, although there are more false negatives in the arctic compared to the more southern tiles. The false negatives occur mostly on the edges of larger lakes and very small lakes not being accounted for. Differing from the southern tile in Eastern Amazonia, the 5mR and 27kR output show very few false positive water classifications, although of all the models the 27kC model best captures the water classifications that are part of the MOD44W product. The qualitative output lines up well with the quantitative analysis for all tiles and years.

4.4 Discussion

In the data sampling process, researchers specifically choose certain regions, dates, times, etc to implicitly select the most salient features and account for intra-class variability. This careful selection aims to capture the most salient features. Theoretically, with a diverse range of examples representing each class, a sufficiently large sample should cause machine learning algorithms to accurately capture both inter and intra-class variabilities within a much larger dataset (e.g., Halevy et al., 2009; Fassnacht et al., 2018). However, in this work, even though a larger class-balanced sample (5mR) was created by deliberately choosing geographically diverse tiles and data examples to encompass inter and intra-class variabilities between land and water classes, it struggles in certain regions such as h12v09 and h09v05.

A second sample was created by deliberately adding examples while reducing the overall sample size to reduce computational time, resulting in the 800kRD sample.

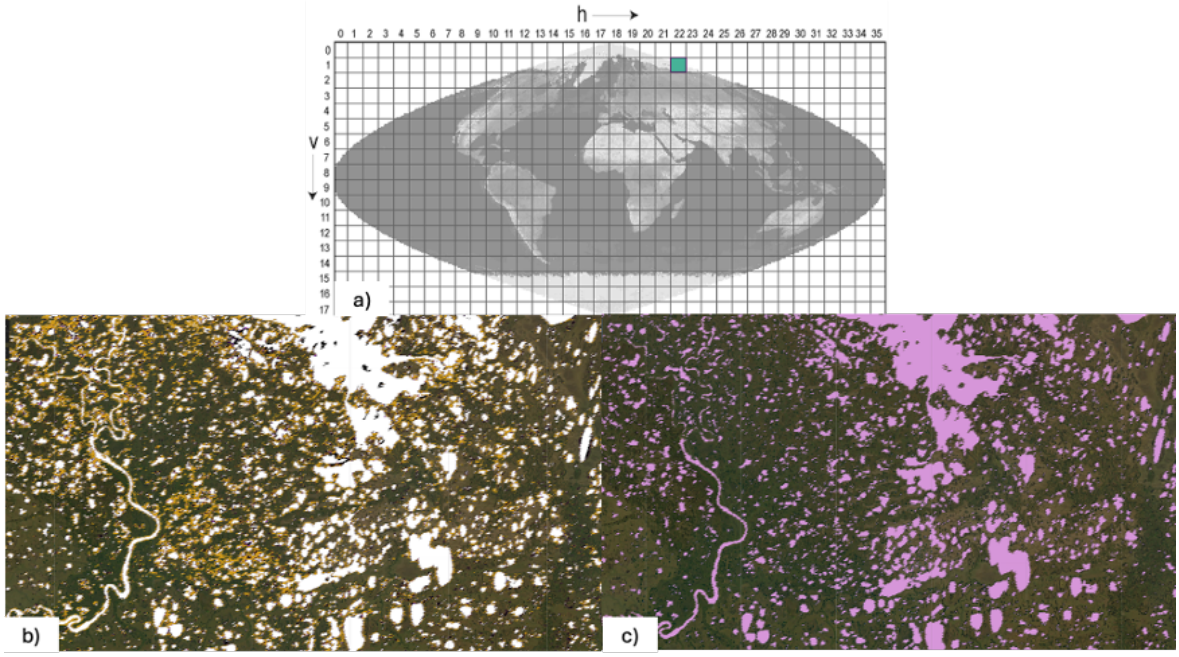

Figure 4.10: Same as Figure 4.9 but with tile h22v01

Despite these efforts, the ML output trained with this sample also exhibits poor performance in specific regions, notably h12v09 and h09v05, and tends to overclassify examples as water when they should be classified as land. Both the 5mR and 800kRD samples, crafted with a deep understanding of sampling representations, demonstrate that larger training sample size does not necessarily indicate improved performance of ML output.

It can be labor intensive to construct quality data samples that encompass both inter and intra-variabilities requiring researchers to devote extensive time and resources to generating their training data. The 800kRD sample was guided by domain expertise and still struggled to capture the full dimensions of the data. Clustering offers an automatable and efficient alternative, generating much smaller samples that still encapsulate the crucial features of Big Data. The authors investigated the reproducibility of this approach (not shown) by training multiple clustered samples using K-means, achieving comparable results to those presented in the study. As noted by

Chi et al. (2016), Big Data offers both challenges and important discoveries. Extracting valuable insights from massive datasets is a complex task, requiring substantial human and computational time investments. Given its speed and ability to produce condensed representations of large datasets, clustering presents a fundamental shift in the approach to handling Big Data, emphasizing exploration over time-consuming data sampling. This not only reduces burdens on developers and researchers, but also supports real-time global operational products like MOD44W, where rapid processing is essential.

Apart from the speed increase of training a ML model with a clustered sample, the results show that the effectiveness of such a model surpasses that of even a very large sample, which theoretically should cover the entire data spectrum. A crucial aspect of successful classification with clustered data involves the strategic elimination of entire clusters, underscoring the importance of both clustering itself and the empirical process of selecting which clusters to retain. In this study, the removal of a cluster associated with spectral signatures of ice had a notable impact on the accuracy of ML model classifications in Arctic regions, but otherwise provided increased global performance. Given this study's focus on land versus water classification, the examples representing ice could be substituted using a post-processing algorithm. Further research into preserving specific Arctic data presents an intriguing avenue for exploration, particularly in assessing whether additional clustering of the removed data or explicit inclusion of missed data, like the 800kRD sample approach, would yield benefits.

This study compares an operational land cover product with experimental machine learning outputs, specifically examining how representative sampling influences performance. The underlying distribution of this dataset is already well-understood, as evidenced by the baseline provided in the operational product. In scenarios where the underlying distribution is unknown or a baseline is absent, the clustering method

introduced in this work could empower researchers to engage in data exploration confidently, knowing that essential features are preserved through clustering. This approach enables the creation, training, and deployment of multiple clustered samples within a short timeframe, facilitating rapid exploration and offering quicker insights into potentially unfamiliar datasets.

Lastly, this approach prioritizes inputting data to a RF model, offering explainability inherent to the methodology compared to the black boxes of deep learning models. With fewer parameters to tune and the ease of describing how a random forest works, the ability to produce accurate land cover classifications resonates strongly with researchers familiar with this methodology, thereby overcoming a hurdle in utilizing and deploying machine learning methods (Maxwell et al., 2018).

Chapter 5

Summary and Conclusions

Numerous studies over the past few decades have proven that machine learning (ML) models can predict weather phenomena effectively (e.g., Krasnopolsky et al., 1995; Marzban and Stumpf, 1996; Elio et al., 1987; Campbell and Olson, 1987; Mcarthur et al., 1987; Gardner and Dorling, 1998; Hsieh and Tang, 1998). The research in this work moves beyond standard ML applications for weather prediction, exploring three different domains where ML is a tool for deeper exploration of meteorologically relevant data. This research underscores the vital role of expert knowledge in creating, evaluating, and deploying ML models in the earth sciences. Evaluating the success of an ML model solely by metrics such as accuracy or IOU creates boundaries to understanding the details underlying the patterns a ML model learns. The small, empirically driven details are, in the author's opinion, most crucial for meaningful model development and deployment. The end user is a key component throughout the ML model development process and should remain so, as illustrated by the knowledge gained from these three distinct projects using ML modeling with similar or related types of geophysical data across various spatial and temporal scales.

5.1 Above Anvil Cirrus Plume Identification

Detecting these plumes presents a challenge due to the time-intensive task of locating and accurately delineating their boundaries, resulting in a dataset that is limited

in scope. To mimic trained human analysts, a pixel-scale classification approach is essential. A binary classification of plumes across entire storm systems could prove beneficial in operational environments where the presence of a plume itself is more critical than its precise location. However, for studying plume extents and effects on stratospheric WV, an ideal dataset would include pixel-scale or high-resolution classifications. As noted by Ledesma Maldonado et al. (2022), when overshooting tops are observed, stratospheric WV advects downstream while remaining aloft of the anvil. Having fine-scale plume identification allows for tracking the occurrence, origins, and potential trajectories of local deposits of stratospheric WV.

Based on the pixel-scale approach for plume identification, this dissertation investigates training a Unet model using satellite data inputs to generate real-time plume classifications using a repository of approximately 4000 images. Out of various combinations tested, the Unet trained with VIS and IR data demonstrated the best performance. The study suggests that Unets excel in identifying warm plumes, particularly when features like OT and cirrus cloud structures are discernible. However, the absence of VIS data, especially during nighttime, limits applicability. This research highlights the ability of a shallow Unet, trained with minimal samples, to grasp the features of warm plumes while acknowledging the struggle in identifying cold plumes within this dataset. It proposes important insights for future dataset creation to enhance Unet performance in ML classification.

Apart from enhancing cold plume analysis, this method operates in real-time without considering previous storm movement. The incorporation of time, particularly in VIS data analysis, could be crucial. Currently, the classification is solely for real-time assessment, overlooking temporal dynamics. Animated imagery reveals the dynamic nature of plume analysis, capturing the "bubbling" and ejection of warm AACP clouds from an OT (Liles et al., 2020). The decision to exclude previous timesteps stems from

the slicing of gridded data approach rather than storm-centering. While slicing offers efficiency, it neglects storm motion, essential for training an ML model on AACP evolution. Introducing time would augment the training dataset, enabling recording of plumes at various stages, thereby facilitating the learning of incremental evolutionary steps and expanding the dataset.

Even with augmentation, the selected plume data are closely grouped in time (days apart) and space (all from West Texas), which could affect the model's capability to identify plumes across diverse regions and seasons. As pointed out by Hong et al. (2023), OTs (and subsequent plumes) are more frequent over the Intertropical Convergence Zone (ITCZ), central and southeastern North America, tropical and subtropical South America, southeastern and southern Asia, tropical and subtropical Africa, and northern middle to high latitudes. Seasonal variations are influenced by major climate systems such as the ITCZ and local monsoons. Therefore, plume characteristics and airflow patterns typical of the spring/summer in central United States may not adequately represent plume dynamics in tropical or monsoonal regions. To develop a robust dataset for training, further investigations into OTs and AACPs across diverse geographical regions are necessary. Nonetheless, employing a small-scale model initially to identify general plume areas and subsequently refining these data could enhance regional datasets to achieve broader global representation. Furthermore, a comprehensive repository of plumes could offer valuable insights into stratospheric water vapor levels and essential metrics for the climate science community.

5.2 Severe Hail Prediction

Next, is CONUS-scale, day-ahead severe hail prediction. Through this work, we introduce a flexible method for assigning weights to data that are physically or scientifically

relevant to earth science developers. The results reveal interesting differences between various weighting metrics and model evaluations. While there are qualitative and quantitative differences between the trained ML models, these differences are minimal and likely noticeable only to those who analyze them in detail, like the method's developer. A cursory glance would suggest little difference between the weighted and unweighted ML outputs.

Statistical testing was not performed in this dissertation, but it would be beneficial in the future to investigate similarities between the weighted and unweighted ML models. Nonetheless, though the physical relevance of the weighted models does not substantially increase their skill, they may be considered more trustworthy in specific situations. This raises interesting questions about the metrics that constitute a good AI prediction. The temporally-weighted ML models concentrate probabilities in regions of highest threat, as noted by participants of the 2020 Hazardous Weather Testbed Spring Forecasting Experiment. For instance, the 2020 HWT SFE summary document stated, "ML Burke tended to give sharper probabilities over smaller regions, which helped forecasters identify areas of greatest threat" (Clark et al., 2021), indicating end-user trust in these forecasts as initial guidance.

Reliability and performance diagrams show that the weighted ML forecasts reduce false alarms but maintain higher probabilities of detection (POD), similar to Brooks (01 Jun. 2004). While high POD values, like those in the UH proxy forecast, are desirable, reducing false alarms is essential for valuable forecasts (Murphy, 1993). Focusing on the highest hail threat increases skill in key regions but misses more storm reports. The UH parameter, although high in POD, has very low success ratio values, indicating that UH proxy outputs larger areas of non-zero probabilities compared to both ML models, leading to more false alarms than ML forecasts. Depending on the needs of individual forecasters or forecast offices, the different products, whether UH proxy or

an ML model, weighted or not, may prove more useful. Changing the regional domain for test case evaluations to either a larger or smaller area could affect these results, potentially showing more skill with spatial weights.

Permutation importance results show that each trained ML model highlights different features of severe hail production. This differentiation is crucial as it provides insights into the environments where each model performs best. For instance, the ML model trained on July storms emphasizes vertical air motions rather than thermodynamic variables as seen in the ML model trained for May storms. The unweighted ML model tends to favor the highest sample frequency, which might not be relevant to the entire CONUS throughout the year.

This method is highly flexible, capable of handling various functions and procedures for weighting input examples, making it essential for localizing ML models for specific problem domains. Implementing different weighting scenarios does not require new data, simplifying future localized ML model deployment. We plan to explore different weighting configurations for an optimal ML hail prediction model and apply this method to other regions with low-frequency hail events.

5.3 Representative Sampling of Global Data

Finally, the last domain involves global scale land cover classification using remotely sensed data. The prevalence of exceedingly large global datasets has become commonplace, accompanied by a rise in endeavors to refine their analytical methodologies. While ML modeling presents numerous advantages for efficiently processing vast datasets, skillful ML model classifications depend on training data. Prior research

(Halevy et al., 2009; Fassnacht et al., 2018, e.g.,) has indicated that by employing adequately large training data subsets, ML models can capture the comprehensive population spectra, encompassing the most crucial features. Nonetheless, this study found that introducing very large training samples to a RF classifier resulted in inadequate practical land cover classifications.

The primary objective of this investigation is to construct a highly representative training sample from a very large dataset. The task to examine how sampling affects classification skill when using ML models is simple land cover classification—differentiating land from water. By utilizing the k-means algorithm, the authors explored the potential of clustering large samples to maintain essential features while minimizing sample size. This approach facilitates the automated creation of ML training samples, thereby streamlining model tuning computational time and alleviating the burden on developers, allowing them to focus on data exploration rather than manual data sampling.

The utilization of a baseline global water mask, MOD44W, facilitated the comparison of clustering spectral data and other sampling methods with a widely recognized and extensive dataset. Multiple different random sample sizes are investigated as input to separate RFs to examine the effect of sampling on classification outputs. In general, the accuracy of each classification output were similar and indicated good performance as compared to the MOD44W mask, however practical applications and F1 scores describe a different story. Despite the similar test classification accuracy measurements observed across various models, a qualitative assessment of the ML classification outputs delineates a notably different scenario. RF classifiers trained using very large samples, alongside a random sample of similar size to the clustered sample data, show more false water classifications, particularly evident in tiles h12v09 and h09v05.

In addition to enhancements in classification performance and efficiency, this study emphasizes the efficacy of the original sampling approach. The original sampling strategy focuses on specific 12 tiles worldwide, a robust foundation for further analysis in global land cover classification. While operational water masks like MOD44W already exist, this research illuminates a methodology applicable across various problem domains, particularly advantageous in scenarios where the underlying distribution of a large dataset is uncertain. This straightforward approach proves remarkably effective, as demonstrated by the automatic generation of clusters that represent an extensive distribution of values with minimal effort besides applying the k-means algorithm. It underscores the significance of training data quality over sample size, provided data variability is adequately addressed. The clustering sample strategy not only outperformed other sampling methods both visually and objectively but also demonstrated the effectiveness of clustering in handling "Big Data". This method suggests that clustering could be the next frontier in 'Big Data' modeling, offering an alternative to merely increasing computing power. Moreover, clustering reduces training time, resulting in fewer computing hours and lower greenhouse gas emissions.

5.4 Conclusions

By moving beyond traditional machine learning (ML) applications for weather prediction, such as out of the box RF models, this dissertation explores three distinct areas: real-time plume identification, day-ahead severe hail prediction, and global land cover classification. Firstly, this work addresses the challenges of identifying above anvil cirrus plumes in real-time using satellite data, crucial for enhancing severe weather alerts in radar-deficient regions. This research highlights the effectiveness of a Unet model trained with VIS and IR data, emphasizing the need to incorporate temporal

dynamics for improved accuracy in plume classification. Secondly, this dissertation delves into CONUS-scale severe hail prediction, showcasing how weighted ML models concentrate probabilities in high-threat areas, reducing false alarms while maintaining high detection rates. This research underscores the flexibility of ML in adapting to different weighting scenarios without requiring new data, facilitating localized model deployment. Lastly, the dissertation tackles global land cover classification, demonstrating that clustering techniques with the k-means algorithm can optimize training sample efficiency. Underscored is the importance of data quality over quantity in ML applications, offering a robust methodology applicable across diverse environmental datasets.

Overall, this dissertation emphasizes the crucial role of expert knowledge in developing and evaluating ML models within the earth sciences, arguing that metrics such as accuracy or IOU fall short of fully capturing the intricate patterns learned by these models. This dissertation underscores the importance of incorporating end-user feedback throughout the ML model development process, demonstrating that tailored yet flexible ML models can effectively address specific meteorological challenges while remaining applicable to broader contexts.

5.5 Contributions

In this dissertation, I have made contributions to the meteorological field by spearheading the creation and application of a ML weighted methodology designed for localized severe hail prediction in varying temporal and spatial contexts. Additionally, I was part of a collaborative effort that implemented this advanced modeling approach during the 2020 HWT, yielding insightful feedback which further refined our methodologies.

Moreover, I devised a novel framework for deep learning AACP plume prediction, focusing on plumes while another NASA group emphasized OT identification. Adopting a sliding panel method over a storm-centered approach, my method improved computational efficiency but potentially reduced the available training data. I also provided not only a statistical evaluation of the DL AACP identification method, but also a two-dimensional XAI headmap of the DL outputs to understand the model's nuances, culminating in recommendations for future enhancements.

Separately, I analyzed a vast global repository of surface reflectance values and implemented an automated algorithm for data pre-processing, reducing a training dataset from roughly 5 million data points to 27 thousand, thereby substantially cutting computational time for training a ML model. This adaptable method is useful across multiple domains, not limited by the input dataset type, and is currently used in other projects at NASA GSFC including a meteorological group focusing on cloud dynamics. Overall, my work has led to the development of various methods for data handling in meteorology and Earth sciences, specifically tailored yet flexible. These pre-processing strategies for ML models are advantageous across numerous sectors and hold promise for applications beyond convective meteorology and remote sensing.

Appendix

Training Information					Validation
Input Features	Cases (number)	Epochs	Final IOU Loss	Duration (min)	Date
VIS	4081	29	0.6871	92	8 May
VIS, IR	4081	31	0.6338	81	8 May
VIS, IRDIFF	4081	50	0.6439	145	8 May
VIS,IR,IRDIFF	4081	39	0.6154	123	8 May
IR	3907	27	0.685	60	7 May
IRDIFF	3954	50	0.6991	71	27 May
IR, IRDIFF	3632	29	0.6569	31	25 May

Table 5.1: Information about each Unet trained to identify plumes, where each row indicates what features were used to train a Unet model. Training data are from 30 April, and 1,5,6,7,8,17,18,20,21,26,27 May 2019 excluding the validation date.

Reference List

Allen, J. T., I. M. Giammanco, M. R. Kumjian, H. Jurgen Punge, Q. Zhang, P. Groenemeijer, M. Kunz, and K. Ortega, 2020: Understanding hail in the earth system. *Reviews of Geophysics*, **58 (1)**, e2019RG000 665, doi:10.1029/2019RG000665.

Anderson, J. G., D. M. Wilmouth, J. B. Smith, and D. S. Sayres, 2012: Uv dosage levels in summer: Increased risk of ozone loss from convectively injected water vapor. *Science*, **337 (6096)**, 835–839, doi:10.1126/science.1222978, URL http://dx.doi.org/10.1126/science.1222978.

Anderson, J. G., and Coauthors, 2017: Stratospheric ozone over the united states in summer linked to observations of convection and temperature via chlorine and bromine catalysis. *Proceedings of the National Academy of Sciences*, **114 (25)**, doi:10.1073/pnas.1619318114, URL http://dx.doi.org/10.1073/pnas.1619318114.

Bao, J., M. Chi, and J. A. Benediktsson, 2013: Spectral derivative features for classification of hyperspectral remote sensing images: Experimental evaluation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **6 (2)**, 594–601, doi:10.1109/jstars.2013.2237758, URL http://dx.doi.org/10.1109/JSTARS.2013.2237758.

Bedka, K., J. Brunner, R. Dworak, W. Feltz, J. Otkin, and T. Greenwald, 2010: Objective satellite-based detection of overshooting tops using infrared window channel brightness temperature gradients. *Journal of Applied Meteorology and Climatology*, **49 (2)**, 181–202, doi:10.1175/2009jamc2286.1, URL http://dx.doi.org/10.1175/2009JAMC2286.1.

Bedka, K., E. M. Murillo, C. R. Homeyer, B. Scarino, and H. Mersiovsky, 2018: The above-anvil cirrus plume: An important severe weather indicator in visible and infrared satellite imagery. *Weather and Forecasting*, **33 (5)**, 1159–1181, doi:10.1175/waf-d-18-0040.1, URL http://dx.doi.org/10.1175/WAF-D-18-0040.1.

Bedka, K. M., 2011: Overshooting cloud top detections using msg seviri infrared brightness temperatures and their relationship to severe weather over europe. *Atmospheric Research*, **99 (2)**, 175–189, doi:10.1016/j.atmosres.2010.10.001, URL http://dx.doi.org/10.1016/j.atmosres.2010.10.001.

Bedka, K. M., C. Wang, R. Rogers, L. D. Carey, W. Feltz, and J. Kanak, 2015: Examining deep convective cloud evolution using total lightning, wsr-88d, and goes-14 super rapid scan datasets*. *Weather and Forecasting*, **30 (3)**, 571–590, doi:10.1175/waf-d-14-00062.1, URL http://dx.doi.org/10.1175/WAF-D-14-00062.1.

Beucler, T., and Coauthors, 2021: Climate-invariant machine learning. arXiv, URL https://arxiv.org/abs/2112.08440, doi:10.48550/ARXIV.2112.08440.

Blei, D. M., and M. I. Jordan, 2006: Variational inference for dirichlet process mixtures. *Bayesian Analysis*, **1 (1)**, doi:10.1214/06-ba104, URL http://dx.doi.org/10.1214/06-BA104.

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, doi:10.1023/A:1010933404324.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3, doi:10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial forecasts of maximum hail size using prognostic model soundings and hailcast. *Weather and Forecasting*, **21 (2)**, 206–219, doi:10.1175/WAF915.1.

Brooks, H. E., 01 Jun. 2004: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bulletin of the American Meteorological Society*, **85 (6)**, 837 – 844, doi:10.1175/BAMS-85-6-837, URL https://journals.ametsoc.org/view/journals/bams/85/6/bams-85-6-837.xml.

Brunner, J., S. Ackerman, A. Bachmeier, and R. Rabin, 2006: A quantitative analysis of the enhanced-v signature in relation to severe weather. *86th AMS Annual Meeting*.

Burke, A., 2019: Using Machine Learning Applications and HREFv2 to Enhance Hail Prediction for Operations. M.S. thesis, School of Meteorology, University of Oklahoma, [Available online at https://hdl.handle.net/11244/320425.].

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 and Forecasting*, **35 (1)**, 149–168, doi:10.1175/WAF-D-19-0105.1.

Campbell, S. D., and S. H. Olson, 1987: Recognizing low-altitude wind shear hazards from doppler weather radar: An artificial intelligence approach. *Journal of Atmospheric and Oceanic Technology*, **4 (1)**, 5–18, doi:10.1175/1520-0426(1987)004⟨0005:RLAWSH⟩2.0.CO;2.

Carroll, M. L., C. M. DiMiceli, J. R. G. Townshend, R. A. Sohlberg, A. I. Elders, S. Devadiga, A. M. Sayer, and R. C. Levy, 2016: Development of an operational land water mask for modis collection 6, and influence on downstream data products. *International Journal of Digital Earth*, **10 (2)**, 207–218, doi:10.1080/17538947.2016.1232756, URL http://dx.doi.org/10.1080/17538947.2016.1232756.

Carroll, M. L., J. R. Townshend, C. M. DiMiceli, P. Noojipady, and R. A. Sohlberg, 2009: A new global raster water mask at 250 m resolution. *International Journal of*

Digital Earth, **2 (4)**, 291–308, doi:10.1080/17538940902951401, URL http://dx.doi.org/10.1080/17538940902951401.

Chen, C., and L. Breiman, 2004: Using random forest to learn imbalanced data. *University of California, Berkeley*.

Chi, M., A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, 2016: Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, **104 (11)**, 2207–2219, doi:10.1109/jproc.2016.2598228, URL http://dx.doi.org/10.1109/JPROC.2016.2598228.

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

Clark, A. J., and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, **102**, E814 – E816, doi:10.1175/BAMS-D-20-0268.1, URL https://journals.ametsoc.org/view/journals/bams/102/4/BAMS-D-20-0268.1.xml.

Cooney, J. W., K. M. Bedka, C. A. Liles, and C. R. Homeyer, 2024: Open-source automated software detection of severe storm signatures using geostationary imagery. *Artif. Intell. Earth Syst.,*, submitted.

Dalponte, M., H. O. Orka, T. Gobakken, D. Gianelle, and E. Naesset, 2013: Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, **51 (5)**, 2632–2645, doi:10.1109/tgrs.2012.2216272, URL http://dx.doi.org/10.1109/TGRS.2012.2216272.

de F. Forster, P. M., and K. P. Shine, 1999: Stratospheric water vapour changes as a possible contributor to observed stratospheric cooling. *Geophysical Research Letters*, **26 (21)**, 3309–3312, doi:10.1029/1999gl010487, URL http://dx.doi.org/10.1029/1999GL010487.

Dessler, A. E., and S. C. Sherwood, 2004: Effect of convection on the summertime extratropical lower stratosphere. *Journal of Geophysical Research: Atmospheres*, **109 (D23)**, doi:10.1029/2004jd005209, URL http://dx.doi.org/10.1029/2004JD005209.

Elio, R., J. D. Haan, and G. S. Strong, 1987: Meteor: An artificial intelligence system for convective storm forecasting. *Journal of Atmospheric and Oceanic Technology*, **4 (1)**, 19–28, doi:10.1175/1520-0426(1987)004⟨0019:MAAISF⟩2.0.CO;2.

Fassnacht, F. E., H. Latifi, and F. Hartig, 2018: Using synthetic data to evaluate the benefits of large field plots for forest biomass estimation with lidar. *Remote Sensing of Environment*, **213**, 115–128, doi:10.1016/j.rse.2018.05.007, URL http://dx.doi.org/10.1016/j.rse.2018.05.007.

Fujita, T. T., 1982: Principle of stereoscopic height computations and their applications to stratospheric cirrus over severe thunderstorms. *Journal of the Meteorological Society of Japan. Ser. II*, **60 (1)**, 355–368, doi:10.2151/jmsj1965.60.1_355, URL http://dx.doi.org/10.2151/jmsj1965.60.1_355.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840, doi:10.1175/WAF-D-17-0010.1.

Gardner, M., and S. Dorling, 1998: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, **32 (14)**, 2627 – 2636, doi:https://doi.org/10.1016/S1352-2310(97)00447-0, URL http://www.sciencedirect.com/science/article/pii/S1352231097004470.

Gonçalves, M. L., M. L. A. Netto, J. A. F. Costa, and J. Zullo Júnior, 2008: An unsupervised method of classifying remotely sensed images using kohonen self-organizing maps and agglomerative hierarchical clustering methods. *International Journal of Remote Sensing*, **29 (11)**, 3171–3207, doi:10.1080/01431160701442146, URL http://dx.doi.org/10.1080/01431160701442146.

Gordon, A. E., and C. R. Homeyer, 2022: Sensitivities of cross-tropopause transport in midlatitude overshooting convection to the lower stratosphere environment. *Journal of Geophysical Research: Atmospheres*, **127 (13)**, doi:10.1029/2022jd036713, URL http://dx.doi.org/10.1029/2022JD036713.

Grams, J. S., R. L. Thompson, D. V. Snively, J. A. Prentice, G. M. Hodges, and L. J. Reames, 2012: A climatology and comparison of parameters for significant tornado events in the united states. *Weather and Forecasting*, **27 (1)**, 106–123, doi:10.1175/WAF-D-11-00008.1.

Halevy, A., P. Norvig, and F. Pereira, 2009: The unreasonable effectiveness of data. *IEEE Intelligent Systems*, **24 (2)**, 8–12, doi:10.1109/mis.2009.36, URL http://dx.doi.org/10.1109/MIS.2009.36.

Haupt, S. E., B. Kosović, S. W. McIntosh, F. Chen, K. Miller, M. Shepherd, M. Williams, and S. Drobot, 2018a: 100 years of progress in applied meteorology. part iii: Additional applications. *Meteorological Monographs*, **59**, 24.1–24.35, doi:10.1175/AMSMONOGRAPHS-D-18-0012.1.

Haupt, S. E., R. M. Rauber, B. Carmichael, J. C. Knievel, and J. L. Cogan, 2018b: 100 years of progress in applied meteorology. part i: Basic applications. *Meteorological Monographs*, **59**, 22.1–22.33, doi:10.1175/AMSMONOGRAPHS-D-18-0004.1.

Herman, G. R., and R. S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev*, **146**, 1571–1600, doi:10.1175/MWR-D-17-0250.1.

Herman, R. L., and Coauthors, 2017: Enhanced stratospheric water vapor over the summertime continental united states and the role of overshooting convection. *Atmospheric Chemistry and Physics*, **17 (9)**, 6113–6124, doi:10.5194/acp-17-6113-2017, URL http://dx.doi.org/10.5194/acp-17-6113-2017.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Monthly Weather Review*, **148 (5)**, 2135–2161, doi:10.1175/MWR-D-19-0344.1.

Homeyer, C. R., J. D. McAuliffe, and K. M. Bedka, 2017: On the development of above-anvil cirrus plumes in extratropical convection. *Journal of the Atmospheric Sciences*, **74 (5)**, 1617–1633, doi:10.1175/jas-d-16-0269.1, URL http://dx.doi.org/10.1175/JAS-D-16-0269.1.

Hong, Y., S. W. Nesbitt, R. J. Trapp, and L. Di Girolamo, 2023: Near-global distributions of overshooting tops derived from terra and aqua modis observations. *Atmospheric Measurement Techniques*, **16 (5)**, 1391–1406, doi:10.5194/amt-16-1391-2023, URL http://dx.doi.org/10.5194/amt-16-1391-2023.

Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, **79 (9)**, 1855–1870, doi:10.1175/1520-0477(1998)079⟨1855:ANNMTP⟩2.0.CO;2.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta Hail Growth Model Using Severe Hail Proximity Soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609, doi:10.1175/2009WAF2222230.1.

Jianbo, S., and J. Malik, 2000: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22 (8)**, 888–905, doi:10.1109/34.868688, URL http://dx.doi.org/10.1109/34.868688.

Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the Optimal Configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.].

Johns, R. H., and C. A. Doswell, 1992: Severe Local Storms Forecasting. *Wea. Forecasting*, **7**, 588–612, doi:10.1175/1520-0434(1992)007⟨0588:SLSF⟩2.0.CO;2.

Kanneganti, G. T., 2020: Detection of overshooting cloud tops with convolutional neural networks. Ph.D. thesis.

Karstens, C. D., and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Weather and Forecasting*, **33 (3)**, 715–737, doi: 10.1175/waf-d-17-0188.1, URL http://dx.doi.org/10.1175/WAF-D-17-0188.1.

Kelly, D. L., J. T. Schaefer, and C. A. Doswell, 1985: Climatology of nontornadic severe thunderstorm events in the united states. *Mon. Wea. Rev.*, **113**, 1997–2014, doi:10.1175/1520-0493(1985)113⟨1997:CONSTE⟩2.0.CO;2.

Kim, M., J. Im, H. Park, S. Park, M.-I. Lee, and M.-H. Ahn, 2017: Detection of tropical overshooting cloud tops using himawari-8 imagery. *Remote Sensing*, **9 (7)**, 685, doi:10.3390/rs9070685, URL http://dx.doi.org/10.3390/rs9070685.

Krasnopolsky, V. M., L. C. Breaker, and W. H. Gemmill, 1995: A neural network as a nonlinear transfer function model for retrieving surface wind speeds from the special sensor microwave imager. *Journal of Geophysical Research: Oceans*, **100 (C6)**, 11 033–11 045, doi:10.1029/95JC00857.

Krocak, M. J., and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the united states. *Weather and Forecasting*, **33 (1)**, 59–69, doi:10.1175/WAF-D-17-0123.1.

Kunz, M., U. Blahak, J. Handwerker, M. Schmidberger, H. J. Punge, S. Mohr, E. Fluck, and K. M. Bedka, 2017: The severe hailstorm in southwest germany on 28 july 2013: characteristics, impacts and meteorological conditions. *Quarterly Journal of the Royal Meteorological Society*, **144 (710)**, 231–250, doi:10.1002/qj.3197, URL http://dx.doi.org/10.1002/qj.3197.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol*, **32**, 1209–1223, doi:10.1175/JTECH-D-13-00205.1.

Ledesma Maldonado, K., C. Narotsky, O. Miyawaki, V. G. Anantharaj, and M. E. O'Neill, 2022: Understanding the Climate Impact of Above-Anvil Cirrus Plume (AACP) in Hydrating the Lower Stratosphere. *AGU Fall Meeting Abstracts*, Vol. 2022, A22D–1698.

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

Liles, C., K. Bedka, E. Xia, Y. Huang, R. Biswas, C. Dolan, A. H. Jafari, and T. Smith, 2020: Automated detection of the above anvil cirrus plume severe storm signature with deep learning. *19th Conference on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, American Meteorological Society.

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 and Forecasting*, **35**, 1605 – 1631, doi:10.1175/WAF-D-19-0258.1, URL https://journals.ametsoc.org/view/journals/wefo/35/4/wafD190258.xml.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble. *Wea. Forecasting*, **32**, 1403–1421, doi:10.1175/WAF-D-16-0200.1.

Lu, D., and Q. Weng, 2007: A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, **28 (5)**, 823–870, doi:10.1080/01431160600746456, URL http://dx.doi.org/10.1080/01431160600746456.

Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on doppler radar-derived attributes. *Journal of Applied Meteorology*, **35 (5)**, 617–626, doi:10.1175/1520-0450(1996)035⟨0617:ANNFTP⟩2.0.CO;2.

Maxwell, A. E., T. A. Warner, and F. Fang, 2018: Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, **39 (9)**, 2784–2817, doi:10.1080/01431161.2018.1433343, URL http://dx.doi.org/10.1080/01431161.2018.1433343.

Mcarthur, R. C., J. R. Davis, and D. Reynolds, 1987: Scenario-driven automatic pattern recognition in nowcasting. *Journal of Atmospheric and Oceanic Technology*, **4 (1)**, 29–35, doi:10.1175/1520-0426(1987)004⟨0029:SDAPRI⟩2.0.CO;2.

McCann, D. W., 1983: The enhanced-v: A satellite observable severe storm signature. *Monthly Weather Review*, **111 (4)**, 887–894, doi:10.1175/1520-0493(1983)111⟨0887:tevaso⟩2.0.co;2, URL http://dx.doi.org/10.1175/1520-0493(1983)111⟨0887:TEVASO⟩2.0.CO;2.

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. *Bulletin of the American Meteorological Society*, **98 (10)**, 2073–2090, doi:10.1175/bams-d-16-0123.1, URL http://dx.doi.org/10.1175/BAMS-D-16-0123.1.

McGovern, A., C. D. Karstens, T. Smith, and R. Lagerquist, 2019a: Quasi-operational testing of real-time storm-longevity prediction via machine learning. *Weather and*

Forecasting, **34 (5)**, 1437–1451, doi:10.1175/waf-d-18-0141.1, URL http://dx.doi.org/10.1175/WAF-D-18-0141.1.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019b: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **0 (0)**, doi:10.1175/BAMS-D-18-0195.1.

McGovern, A., and Coauthors, 2022: Nsf ai institute for research on trustworthy ai in weather, climate, and coastal oceanography (ai2es). *Bulletin of the American Meteorological Society*, **103 (7)**, E1658–E1668, doi:10.1175/bams-d-21-0020.1, URL http://dx.doi.org/10.1175/BAMS-D-21-0020.1.

Mecikalski, J. R., T. N. Sandmæl, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett, 2021: Random forest model to assess predictor importance and nowcast severe storms using high-resolution radar–goes satellite–lightning observations. *Monthly Weather Review*, doi:10.1175/mwr-d-19-0274.1, URL http://dx.doi.org/10.1175/MWR-D-19-0274.1.

Millard, K., and M. Richardson, 2015: On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, **7 (7)**, 8489–8515, doi:10.3390/rs70708489, URL http://dx.doi.org/10.3390/rs70708489.

Murillo, E. M., and C. R. Homeyer, 2022: What determines above-anvil cirrus plume infrared temperature? *Journal of the Atmospheric Sciences*, **79 (12)**, 3181–3194, doi:10.1175/jas-d-22-0080.1, URL http://dx.doi.org/10.1175/JAS-D-22-0080.1.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281 – 293, doi:10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/8/2/1520-0434_1993_008_0281_wiagfa_2_0_co_2.xml.

Oliveira, E. N., and Coauthors, 2016: Assessment of remotely sensed chlorophyll-a concentration in guanabara bay, brazil. *Journal of Applied Remote Sensing*, **10 (2)**, 026 003, doi:10.1117/1.jrs.10.026003, URL http://dx.doi.org/10.1117/1.JRS.10.026003.

O'Neill, M. E., L. Orf, G. M. Heymsfield, and K. Halbert, 2021: Hydraulic jump dynamics above supercell thunderstorms. *Science*, **373 (6560)**, 1248–1251, doi:10.1126/science.abh3857, URL http://dx.doi.org/10.1126/science.abh3857.

Ramezan, C. A., T. A. Warner, and A. E. Maxwell, 2019: Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, **11 (2)**, 185, doi:10.3390/rs11020185, URL http://dx.doi.org/10.3390/rs11020185.

Schmetz, J., S. Tjemkes, M. Gube, and L. van de Berg, 1997: Monitoring deep convection and convective overshooting with meteosat. *Advances in Space Research*, **19 (3)**, 433–441, doi:https://doi.org/10.1016/S0273-1177(97)00051-3, URL https://www.sciencedirect.com/science/article/pii/S0273117797000513, proceedings of the A0.1 Symposium of COSPAR Scientific Commission A.

Setvák, M., K. Bedka, D. T. Lindsey, A. Sokol, Z. Charvát, J. Šťástka, and P. K. Wang, 2013: A-train observations of deep convective storm tops. *Atmospheric Research*, **123**, 229–248, doi:10.1016/j.atmosres.2012.06.020, URL http://dx.doi.org/10.1016/j.atmosres.2012.06.020.

Setvák, M., D. T. Lindsey, R. M. Rabin, P. K. Wang, and A. Demeterová, 2008: Indication of water vapor transport into the lower stratosphere above midlatitude convective storms: Meteosat second generation satellite observations and radiative transfer model simulations. *Atmospheric Research*, **89 (1–2)**, 170–180, doi:10.1016/j.atmosres.2007.11.031, URL http://dx.doi.org/10.1016/j.atmosres.2007.11.031.

Setvák, M., and Coauthors, 2010: Satellite-observed cold-ring-shaped features atop deep convective clouds. *Atmospheric Research*, **97 (1–2)**, 80–96, doi:10.1016/j.atmosres.2010.03.009, URL http://dx.doi.org/10.1016/j.atmosres.2010.03.009.

Shafer, C. M., A. E. Mercer, L. M. Leslie, M. B. Richman, and C. A. Doswell, 2010: Evaluation of wrf model simulations of tornadic and nontornadic outbreaks occurring in the spring and fall. *Monthly Weather Review*, **138 (11)**, 4098–4119, doi:10.1175/2010MWR3269.1.

Smith, A. B., 2010: U.s. billion-dollar weather and climate disasters, 1980 - present (ncei accession 0209268). https://www.ncdc.noaa.gov/billions/events/US/1980-2021, doi:10.25921/stkw-7w73.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part i: Storm classification and climatology. *Weather and Forecasting*, **27 (5)**, 1114–1135, doi:10.1175/WAF-D-11-00115.1.

Smith, J. B., and Coauthors, 2017: A case study of convectively sourced water vapor observed in the overworld stratosphere over the united states. *Journal of Geophysical Research: Atmospheres*, **122 (17)**, 9529–9554, doi:10.1002/2017jd026831, URL http://dx.doi.org/10.1002/2017JD026831.

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

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from

a convection-allowing model. *Weather and Forecasting*, **35**, 1981 – 2000, doi:10.1175/WAF-D-20-0036.1, URL https://journals.ametsoc.org/view/journals/wefo/35/5/wafD200036.xml.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Weather and Forecasting*, **31 (1)**, 255 – 271, doi:10.1175/WAF-D-15-0138.1, URL https://journals.ametsoc.org/view/journals/wefo/31/1/waf-d-15-0138_1.xml.

Solomon, S., K. H. Rosenlof, R. W. Portmann, J. S. Daniel, S. M. Davis, T. J. Sanford, and G.-K. Plattner, 2010: Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science*, **327 (5970)**, 1219–1223, doi:10.1126/science.1182488, URL http://dx.doi.org/10.1126/science.1182488.

Stanimirova, R., and Coauthors, 2023: A global land cover training dataset from 1984 to 2020. *Scientific Data*, **10 (1)**, doi:10.1038/s41597-023-02798-5, URL http://dx.doi.org/10.1038/s41597-023-02798-5.

Stensrud, D. J., and M. S. Wandishin, 2000: The correspondence ratio in forecast evaluation. *Weather and Forecasting*, **15 (5)**, 593–602, doi:10.1175/1520-0434(2000)015⟨0593:tcrife⟩2.0.co;2, URL http://dx.doi.org/10.1175/1520-0434(2000)015⟨0593:TCRIFE⟩2.0.CO;2.

Taşdemir, K., B. Yalçin, and I. Yildirim, 2015: Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. *Pattern Recognition*, **48 (4)**, 1465–1477, doi:10.1016/j.patcog.2014.10.023, URL http://dx.doi.org/10.1016/j.patcog.2014.10.023.

Vali, A., S. Comai, and M. Matteucci, 2020: Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sensing*, **12 (15)**, 2495, doi:10.3390/rs12152495, URL http://dx.doi.org/10.3390/rs12152495.

Vinogradova, K., A. Dibrov, and E. W. Myers, 2020: Towards interpretable semantic segmentation via gradient-weighted class activation mapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, doi:10.1609/aaai.v34i10.7244.

Wang, G., H. Wang, Y. Zhuang, Q. Wu, S. Chen, and H. Kang, 2021: Tropical overshooting cloud-top height retrieval from himawari-8 imagery based on random forest model. *Atmosphere*, **12 (2)**, 173, doi:10.3390/atmos12020173, URL http://dx.doi.org/10.3390/atmos12020173.

Willard, J., X. Jia, S. Xu, M. Steinbach, and V. Kumar, 2020: Integrating scientific knowledge with machine learning for engineering and environmental systems. arXiv, URL https://arxiv.org/abs/2003.04919, doi:10.48550/ARXIV.2003.04919.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An Enhanced Hail Detection Algorithm for the WSR-88d. *Wea. Forecasting*, **13**, 286–303, doi:10.1175/1520-0434(1998)013⟨0286:AEHDAF⟩2.0.CO;2.

Xu, R., and D. Wunsch, 2005: Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, **16 (3)**, 645–678, doi:10.1109/tnn.2005.845141, URL http://dx.doi.org/10.1109/TNN.2005.845141.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. *Bull. Amer. Meteor.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.

Zhang, T., R. Ramakrishnan, and M. Livny, 1996: Birch: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, **25 (2)**, 103–114, doi:10.1145/235968.233324, URL http://dx.doi.org/10.1145/235968.233324.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, 2016: Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, doi:10.1109/cvpr.2016.319, URL http://dx.doi.org/10.1109/CVPR.2016.319.

Zhou, Q., H. Tollerud, C. Barber, K. Smith, and D. Zelenak, 2020: Training data selection for annual land cover classification for the land change monitoring, assessment, and projection (lcmap) initiative. *Remote Sensing*, **12 (4)**, 699, doi:10.3390/rs12040699, URL http://dx.doi.org/10.3390/rs12040699.

Zhu, Z., and Coauthors, 2016: Optimizing selection of training and auxiliary data for operational land cover classification for the lcmap initiative. *ISPRS Journal of Photogrammetry and Remote Sensing*, **122**, 206–221, doi:10.1016/j.isprsjprs.2016.11.004, URL http://dx.doi.org/10.1016/j.isprsjprs.2016.11.004.