Application of Random Forest-based supervised ensemble learning method for hail nowcasting in the Midwestern United States

Zhicheng (Chris) Jing
Department of Atmospheric and Oceanic Sciences
McGill University, Montreal
December 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Timely and accurate nowcasting of hail events has significant societal and economic benefits yet remains a challenge due to the limitation of existing numerical weather prediction models in resolving mesoscale convective weather and the complex interactions of factors relevant to hailstorms. The evolution of hail in thunderstorms is a complex process that depends on several interacting factors, including atmospheric instability, updrafts strength, supercooled water content, and dry air in the troposphere. We investigated the possibility and skill of a machine learning method called Random Forests (RF) to nowcast hailstorms. This study applies the RF algorithm to nowcast hail in the Midwestern United States. Hail reports from NOAA's Storm Event Database between May and August from 1999 to 2003 are used to screen out 240 severe hail days in the region. The hail nowcasting model is constructed with 20 relevant parameters and indices derived from the MYRORSS remote sensing data and ERA5 hourly reanalysis data during the same period. The model predicts the probability of hail occurrence with the maximum estimated hail diameter classified into one of the four hail-size categories (< 1 mm, 1–5 mm, 5–20 mm, ≥ 20 mm) for two forecast lead times of 15 minutes and 60 minutes. The Random Forest model identified several hail predictors of significance and demonstrated skills in forecasting hail threat areas in the next 0–1 hour with high accuracy and low error rates.

Résumé

La prévision rapide et précise des événements de grêle présente des avantages sociétaux et économiques importants, mais reste un défi en raison des limites des modèles numériques de prévision météorologique existants à simuler des phénomènes météorologiques convectifs à méso-échelle et des interactions complexes des facteurs pertinents pour les tempêtes de grêle. L'évolution de la grêle au sein des orages est un processus complexe qui dépend de plusieurs facteurs interagissant, notamment l'instabilité atmosphérique, l'intensité des courants ascendants, la quantité d'eau surfondue et l'air sec dans la troposphère. Nous avons étudié la possibilité et les capacités d'une méthode d'apprentissage automatique appelée Forêts aléatoires (Random Forests, RF) pour la prévision immédiate des tempêtes de grêle. Cette étude applique l'algorithme RF à la prévision des chutes de grêle dans le Midwest des États-Unis. Les rapports de grêle de la base de données des événements de tempête de la NOAA entre mai et août de 1999 à 2003 sont utilisés pour identifier 240 jours de grêle sévère dans la région. Le modèle de prévision de grêle est construit avec 20 paramètres et indices pertinents dérivés des données de télédétection MYRORSS et des données de réanalyse horaire ERA5 pendant la même période. Le modèle prédit la probabilité d'occurrence de grêle avec le diamètre maximal estimé de grêle classé dans l'une des quatre catégories de taille de grêle (< 1 mm, $1 – 5$ mm, $5 – 20$ mm, ≥ 20 mm) pour deux délais de prévision de 15 minutes et 60 minutes. Le modèle de Forêts aléatoires a identifié plusieurs prédicteurs de grêle significatifs et a démontré des compétences dans la prévision des zones de menace de grêle dans les 0 à 1 heure suivantes avec une grande précision et de faibles taux d'erreur.

Acknowledgements

This project has been sponsored by the Government of Canada and provided through the Canada Graduate Scholarship-Master's (CGS M) NSERC Award and the Natural Science and Engineering Research Council of Canada Discovery grant program. We also acknowledge the support of the Lorne Trottier Science Accelerator Fellowship of McGill University and the Peter Zwack Award of the Canadian Meteorological and Oceanographic Society. The remote sensing data used in the study can be obtained from the MYRORSS project website (https://osf.io/9gzp2/). The reanalysis data can be obtained from the ECMWF website (https://cds.climate.copernicus.eu/#!/home). The historical hail reports data can be obtained from NOAA's Storm Event Database (https://www.ncdc.noaa.gov/stormevents/).

First, I would like to thank my graduate supervisor Dr. Frédéric Fabry, for accepting me to pursue a Master's degree in Atmospheric and Oceanic Sciences at McGill University under his professional mentorship and guidance. He provided extensive recommendations and insights into the design, implementation, and post-analysis of this research project. In addition, he offered timely and considerate feedback to me on the writing of this thesis. I would also like to thank my colleague Dustin Fraser for providing critical comments on my department student seminar and poster for the CMOS 2023 Congress. Special thanks to Dr. Djordje Romanic for being the examiners of this thesis. Finally, I would like to thank my mom and dad for their financial and spiritual support throughout my graduate studies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AGL - Above ground level

AI - Artificial Intelligence

AMS - American Meteorological Society

ARL - Above radar level

BRN - Bulk Richardson Number

BWER - Bounded weak echo region

CAPE - Convective available potential energy

CART - Classification and Regression Tree

CIN - Convective inhibition

CL - Classic supercells

CMAX - Column maximum radar reflectivity

CNN - Convolutional Neural Networks

CONUS - Contiguous United States

CSI - Critical Success Index

CV - Cross-validation

DT - Decision Tree

ECMWF - European Centre for Medium-Range Weather Forecasts

ERA5 - Fifth generation ECMWF atmospheric reanalysis of the global climate

ETOP18 - Echo top heights for 18 dBZ

ETS - Equitable Threat Score

F - False Alarm Rate

FAR - False Alarm Ratio

GEFS/R - Second Generation Ensemble Forecast System Reforecast

HDA - Hail Detection Algorithm

HP - High-precipitation storms

HSS - Heidke Skill Score

IBC - Insurance Bureau of Canada

KI - K Index

LLAS - Low-level azimuthal shear

LLAS_QI - Low-level azimuthal shear quality index

LLSH - Low-level specific humidity

LLVS - Low-level vertical shear

LP - Low-precipitation storms

MCC - Mesoscale convective complex

MCS - Mesoscale convective system

MDI - Mean decrease in impurity

MESH - Maximum estimated size of hail

ML - Machine Learning

MLAS - Mid-level azimuthal shear

MLAS_QI - Mid-level azimuthal shear quality index

MLSH - Mid-level specific humidity

MLVS - Mid-level vertical shear

MRMS - Multi-Radar Multi-Sensor

MSC - Meteorological Service of Canada

MYRORSS - Multi-Year Reanalysis of Remotely Sensed Storms

NCEI - National Centers for Environmental Information

NEXRAD - Next-Generation Weather Radar

NOAA - National Oceanic and Atmospheric Administration

NSSL - National Severe Storms Laboratory

NWP - Numerical Weather Prediction

NWS - National Weather Service

PC - Proportion Correct

POD - Probability of Detection

RALA - Reflectivity at the lowest altitude

RAM20C - Reflectivity at isotherm –20°C

RF - Random Forest

RV500MB - Relative vorticity at 500 mb pressure level

S06 - 0–6 km vertical bulk wind shear

SHI - Severe Hail Index

SHIP - Significant Hail Parameter

SPC - Storm Prediction Center

SRH - Storm-relative helicity

SRRF - Spatiotemporal Relational Random Forest

SSI - Surface solar irradiance

SVR - Severe Thunderstorm Warning

SWEAT - Severe Weather Threat Index

TBSS - Three-body scattering signature

TT - Total Totals Index

VIL - Vertically integrated liquid

WB0 - Wet-bulb-zero level

WEA - Wireless Emergency Alert

WFO - Weather Forecast Office(s)

WSR - Weather Surveillance Radar(s)

ZDL - Zero-degree level

1. Introduction and Literature Review

1.1 Hail and Hailstorms

1.1.1 Hail Fundamentals

Hail is a solid precipitation phenomenon consisting of spherical, conical, or irregular lumps of ice known as hailstones. A unit diameter of at least 5 mm is the size convention for hail. Hail is a common by-product of severe thunderstorms, but their occurrences are not guaranteed. According to the American Meteorological Society (AMS, 2012), the term hailstorm is usually reserved for strong convective storms that produce ground-reaching hailstones of significant amount or size. As hailstones fall out of a moving hailstorm, they create a damage path on the ground known as a hail swath. Data from the National Severe Storms Laboratory (NSSL) suggests these paths can range in size from a few acres to 16 km wide and 160 km long. On average, they are just about 1.6 km wide and a few kilometers long. The size distribution of hailstone hitting the ground can be approximated using a shifted gamma distribution (Wong et al., 1988). According to Stull (2017), most hailstones fall within the diameter range of 5 to 15 mm (pea-sized to mothball-sized), while approximately 25% of the stones exceed a diameter of 15 mm. Hailstones with a diameter ≥ 70 mm (baseball-sized) are very rare and capable of smashing windshields and causing enormous damage to properties.

1.1.2 Thunderstorm Fundamentals

The evolution of hail can be better understood as part of the life cycle of a deep moist convective storm, commonly referred to as thunderstorms. Thunderstorms develop in sufficiently moist and convectively unstable air masses and can produce heavy rain, lightning, and hail (Rogers and Yau, 1996). The strength of convection is determined by the amount of instability, while the type of thunderstorm that forms after instability is released is determined by vertical wind shear (Weisman and Klemp, 1986). There are three broad categories of thunderstorms: single-cell (air-mass) development in weak wind shear, multi-cell development in moderate shear, and supercell development in strong shear (Dudhia, 1997). Notably, supercells are known for their propensity to produce larger hail than the other two storm types (Kumjian et al., 2021).

In thunderstorms, the "cell" serves as the fundamental unit. It represents a dynamic entity marked by a concentrated area of robust vertical air movement, detectable by weather radar due

to the associated volume of intense precipitation (Browning, 1977). Byers and Braham (1949) discovered that thunderstorms consist of one or more units of convective circulation, comprising an updraft zone and a region of compensating downward motion. These convective cells exhibit similar structure and behavior across most storms. Browning (1962) then classified cells as either ordinary cells or supercells, an important distinction for hail growth due to their difference in the driver and dynamics of the updraft.

Single-cell convection is comprised of a single ordinary cell. The updraft of an ordinary cell is driven by buoyancy. Beyers and Braham (1949) identified three stages in the evolution of an ordinary cell by the predominant direction and magnitude of the vertical air motion: the towering cumulus stage (with updraft alone), the mature stage (with updraft and downdraft together), and the dissipating stage (with downdraft alone). The life cycle of an ordinary cell begins as the updraft-producing thermal causes the cloud to grow in the towering cumulus stage. With continued upward motion, the water vapor in the air rapidly condenses into liquid. The mature stage commences with the production of precipitation particles heavy enough to fall through the updraft. The hydrometeor loading from the falling precipitation and its subsequent evaporative cooling induces a downdraft that reduces updraft buoyancy. The downdraft eventually cuts off the updraft from its source as the cell enters its dissipating stage. The leading edge of the downdraft upon reaching the ground level marks the gust front. Severe weather such as heavy wind gusts, hail, and intense rainfall usually occur near this transition. As the updraft decays with the elimination of the source of rainfall, the downdraft also weakens and eventually dies out, reducing the cumulonimbus cloud into an orphan anvil cloud residue composed entirely of ice crystals. Single-cell thunderstorms have a short duration of less than an hour with only weak to moderate updrafts, thus rarely producing severe hail on the ground (Markowski and Richardson, 2010).

In contrast to single-cell convection, multi-cell (or multicellular) convection is characterized by the repeated development of new ordinary cells along the upwind part of the gust front, where the downdraft of the mature cell meets the environmental wind, lifting air parcels and triggering new convection. The formation of multi-cell thunderstorms requires a moderate degree of vertical wind shear, which offsets the updraft from the downdraft (Stull, 2017). Despite the individual cells lasting for only 30-60 minutes, the continual initiation of new cells allows for the survival of a larger-scale convective system. This prolonged multicellular convection can result

in hours of severe weather, generating extensive areas of damaging straight-line winds and golf-ball-sized hail (Markowski and Richardson, 2010). In multi-cell hailstorms, hailfalls are typically scattered and discontinuous, either organized within a larger hail swath for structured multi-cells or distributed irregularly (North et al., 2015).

A supercell thunderstorm is characterized by the presence of a persistent, deep mesocyclone within the updraft. A mesocyclone is a region of rotation (vertical vortex) typically 3–8 km wide and extends over at least half of the depth of the updraft (Markowski and Richardson, 2010). Strong wind shear in the bottom 6 km of the troposphere creates a rotation in the horizontal airflow, which is then tilted vertically by the updraft within the storm. This tilting process causes the rotation to extend through a significant depth of the storm, forming the mesocyclone. The enhanced updraft lifts and sustains large amounts of moisture and energy, promoting the development and growth of some of the most severe weather phenomena such as tornadoes (Stull, 2017). Supercells are classified into three categories: low-precipitation (LP) storms, medium-precipitation or classic (CL) supercells, and high-precipitation (HP) storms (Stull, 2017). As the name suggests, LP supercells have a relatively low amount of precipitation. They are typically associated with dry environments, have a longer lifespan, and produce more severe weather than their high precipitation counterparts, especially large hails (Bluestein and Woodall, 1990). HP supercells tend to produce much more intense rainfall but are shorter-lived because they often become rain-wrapped and lead to storm dissipation. Although much less common than single-cell and multi-cell storms, supercells are responsible for almost all hail having a diameter of 5 cm or larger and nearly all violent tornadoes (Markowski and Richardson, 2010).

A thunderstorm system is often composed of several convective cells in various stages of development, making it difficult to identify any individual cell. Complex of thunderstorms organized on a scale larger than individual storm units are known as mesoscale convective systems (MCS). Examples of MCS include squall line, bow echo, and mesoscale convective complex (MCC). On average, about 8–17% of hail in the United States is associated with MCSs (Wang et al., 2023). Storm-relevant challenges in hail nowcasting include rapid evolvement in storm organization, significant variability in a storm complex over short distances, and the interaction of hail with other precipitation. Certain thunderstorm types possess distinct characteristics that make them more prone to hail formation. Accurate identification of storm type(s) associated with a given severe weather system can aid in hail prediction.

1.1.3 Hail Development

Hailstones form through the accretion and freezing of supercooled water droplets onto ice particles in cumulonimbus clouds (Rogers and Yau, 1996). This process involves three stages: hail embryo formation, growth, and melting. Embryos are formed as cloud droplets coalesce, freeze, and aggregate into clear or graupel ice particles (Figure 1.2). "Coalescence" refers to the collision and merging of cloud droplets to form raindrops while "aggregation" refers to the clumping of snow crystals to form snowflakes. A freezing raindrop can grow into a hailstone through the collection of supercooled water droplets via “accretion”. Sometimes, the accretion of supercooled cloud droplets in a low-density deposit (riming) produces graupels, a hail embryo during the development of hailstones. Some embryos are carried away by the updraft, preventing further growth. Successful embryos that initiate hail growth occur in updraft regions where they avoid expulsion. As they rise, hailstones collect supercooled droplets, freeze upon contact, and grow in the updraft. Studies have shown altitude with an air temperature interval between −10°C and −30°C favours hail growth (Miller, 1988). Most of the growth occurs as the hailstones suspend in the updraft, drifting horizontally across a narrow altitude range with temperatures ranging from –15 to –20°C (Stull, 2017).

The hailstones' layered structure arises from varying liquid water content and air temperature. Growth is prominent beneath the hailstone due to the accretion and freezing of water droplets. Embryos may lead to hailstone layers of various densities. Larger hailstones tumble more and have non-spherical shapes. The growth rate of hailstones increases with their size. Terminal velocity also rises with size, potentially exceeding 50 m/s (Knight and Knight, 2001). Upon falling through warmer air, hailstones melt, with melting influenced by their descent trajectory. Small hailstones melt before hitting the ground, but larger ones may reach the ground as hail with diameters ≥ 5 mm. Despite large hail being highly destructive, the water mass in hail on the ground is generally only 2 to 3% of the mass of rain from the same thunderstorm (Stull, 2017).

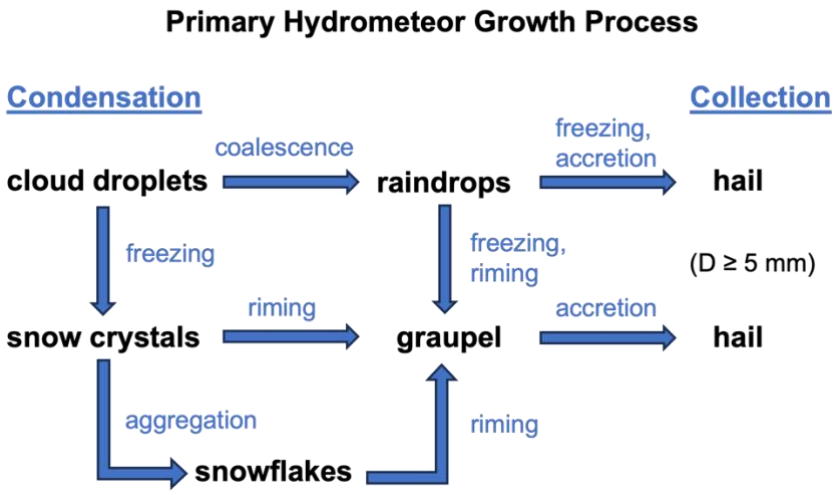


Figure 1.1. A rough classification of the primary hydrometeor growth process. Adapted from Knight and Knight, (2001).

1.1.4 Hailstorm Climatology and Statistics

Understanding the preferred locations of hailfall requires consideration of favourable environments for hail formation. Warm air with high humidity at low levels and cold temperatures aloft create the instability necessary for thunderstorm updrafts that can support large hail. Significant changes in the horizontal wind speed or direction (shear) with height also contribute to updraft velocity by influencing the vertical pressure gradient (North et al., 2015). These conditions are common within flat terrain continental interiors east of high mountain ranges at mid-latitudes during the late spring to summer months. With the interaction between the predominant westerlies aloft and warm, moist wind from the oceans near the ground, larger hail generally occurs east of the Rocky Mountains in North America, the Andes in South America, and the Himalayas in Asia (North et al., 2015). Although thunderstorms are more frequent in tropical regions than in the midlatitudes, they are less likely to produce large hail due to weaker updrafts with lower atmospheric instability and significant melting of hailstones thanks to a higher freezing level. Hail is also infrequent at high latitudes due to the generally cold and dry conditions not conducive to thunderstorm formation (Allen, 2019).

Within the contiguous United States (CONUS), hailstorms are most frequent in the central and southern Great Plains, as depicted in Figure 1.2. A study on hail climatology using radar

products from 1995 to 2017 by Allen et al. (2019) has shown the mean annual maximum hail size to be the largest (above 2.5-inch or 63.5 mm) for a region centered in central western Kansas, extends into southern Nebraska, west-central Oklahoma, and north-central Texas. This region also overlaps well with the area that receives the most annual severe hail days (≥ 7 days per year) from a study by Cintineo (2012) using radar-derived data from 2007 to 2010. Data analysis has also shown an increase in hail days (≥ +3 days) per decade for the same region mentioned above (Biryukov et al., 2021). Seasonally speaking for CONUS, hail activity is quiet in the winter months. It starts to gain ground in March in the southern U.S. before gradually moving the center of influence north over time into the central Great Plains. The hail activity coverage extends to the furthest north in August before retreating to the south into late fall (Cintineo, 2012). Research conducted by Johns and Hart (1998) revealed that nearly all tornadic storms in the central U.S. are accompanied by large hail (diameter ≥ 19 mm). However, along the Gulf and Atlantic coasts and in the southeastern states, some severe tornadic storms occur without being accompanied by large hail.

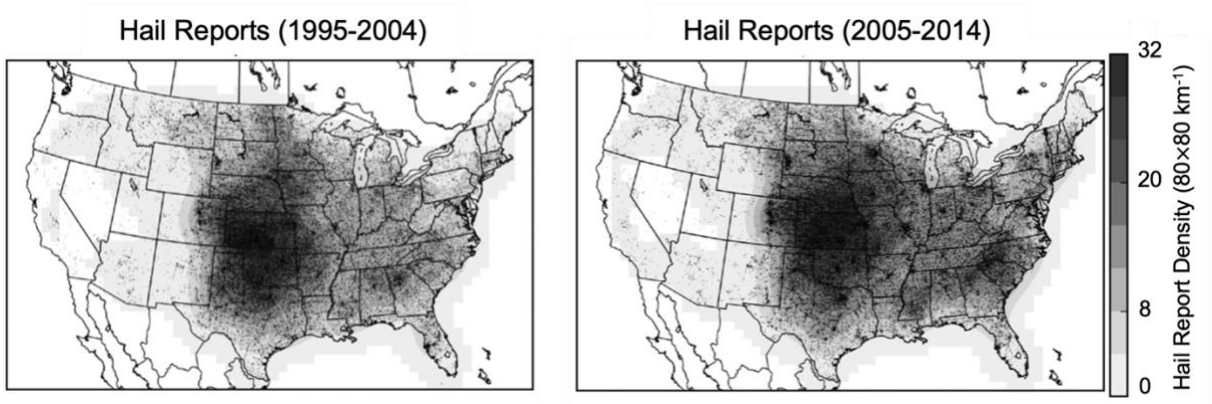


Figure 1.2. Mean annual U.S. Gaussian kernel-smoothed sub-severe (≥ 0.75-inch or 19 mm) hail report density for two-decade intervals for 1995–2014. Image adapted from Allen et al. (2019).

When the severity of the hail (size) is not considered, the region with the most hail reported in the U.S. is the area where Colorado, Nebraska, and Wyoming meet, known as "Hail Alley" (Marshall, 2000). Cheyenne, Wyoming is the most hail-prone U.S. city with an average of eight to nine days of hail per year (Doesken, 1994). Despite the high hail incidence in this high mountainous region, most of the hailstones are very small (North et al., 2015). This feature can be attributed to the ground being exceptionally close to the cloud base, making pea-sized ice

particles (normal forms of convective precipitation) that ordinarily melt when descending to the ground to stay frozen. (Knight and Knight, 2001).

In Canada, the most significant hailfalls are on the leeward side of the Rockies of central and southern Alberta, with the region between Calgary and Edmonton known as Canada's "Hailstorm Alley" (Watson, 2008). Hail activities are also common in central British Columbia and the southern prairies of Saskatchewan and Manitoba (Stull, 2017). The largest hailstone ever recorded in Canada was discovered in August 2022 near Red Deer, Alberta measuring 293 g with a maximum diameter of 123 mm (Renaud, 2022).

1.1.5 Hail Hazards and Mitigation

Hail is a significant natural hazard as large falling hailstones can cause severe damage to crops, tree foliage, vehicles, aircraft, and buildings (Stull, 2017). Strong winds can make the hailstones travel near-horizontally as they fall, causing extensive damage to building windows and sidings. In the U.S., annual average property losses have surpassed 10 billion dollars, and hail events with substantial impacts on major cities often result in losses of 1 billion U.S. dollars (Gunturi and Tippett, 2017). Risks associated with hail events are of particular concern to the insurance industry. Hailstorms can result in catastrophic losses for insurance companies, especially when they occur in regions with a high concentration of insured properties and automobiles (Allen, 2020). For example, the Calgary Hailstorm of June 2020 damaged 77,000 homes and caused an estimated $1.4 billion in insurable damages (Porter, 2022). Much of that money was paid for roof repairs. According to the Insurance Bureau of Canada (IBC), it remains the fourth costliest natural disaster in the country's history (IBC, 2023).

Agricultural crop typically experiences more significant impacts with higher hailfall density and directly depend on the length and width of the hail swath (Changnon, 1977; Changnon, 1999; Sánchez et al., 1996). Crop damage from hail in the U.S. averaged 119 million dollars per year from 2014 to 2018 (Rudden, 2022). While receiving less attention than property losses, the Department of Agriculture maintains an extended record of agricultural loss events (Changnon and Changnon, 2000). Hail also poses a significant threat to aviation. Hailstones hitting the plane can easily smash the windshield and block the pilot's view from the cockpit. In extreme cases, they can severely damage the aircraft radome and engines (Hayduk, 1973). Pilots are usually instructed to avoid flying their aircraft into or near hail-producing cumulonimbus clouds.

1.2 Hail Forecasting

1.2.1 Environmental Conditions Favouring Large Hailstones

Predicting hail size is challenging due to the complexity of hail formation and our limited understanding of associated processes (Allen et al., 2020). Determining relationships between hail size and the environment is difficult because of the small sample sizes of reliable hail observations and regional variations in necessary parameters (e.g., Brimelow et al., 2006; Edwards and Thompson, 1998). Nonetheless, there exists a consensus among meteorologists on favourable environmental conditions for large hail occurrence.

Forecasting the potential for large hail is directly linked to predicting the updraft strength in thunderstorms since heavier hailstones can be sustained aloft against their terminal fall velocities only in stronger updrafts (Stull, 2017). To generate such significant vertical velocities against multiple opposing forces, strong positive buoyancy is necessary. That is, the atmosphere must be convectively unstable to lift an air parcel from the lower level into the upper level of the atmosphere. Convective available potential energy (CAPE) is a frequently used parameter in forecasting updraft strength, as it is the amount of work that the upward buoyancy force would perform on a given rising air parcel. Numerical simulations by Lin and Kumjian (2022) show that hail size does not increase monotonically with CAPE as conventional wisdom suggests but instead maximizes for an intermediate range of "optimal" CAPE values. Despite its representativeness, CAPE and its negative counterpart, CIN (Convective Inhibition), are not always good predictors of updraft strength (Markowski and Richardson, 2010).

A strong vertical shear of the horizontal wind is another necessary condition for large hail. Since it takes about 40 to 60 minutes to create hailstones reaching the surface, large hail can only be produced by long-lived thunderstorms (e.g., supercells) with relatively steady, organized updrafts (Stull, 2017). Vertical shear tilts the updraft and shifts the precipitation-induced downdraft horizontally. This horizontal displacement is crucial for establishing a thermally direct circulation with positive buoyant forcing in the updraft and negative forcing in the downdraft (Borland et al., 1977). The 0–6 km vertical bulk wind shear (S06) and the 0–3 km storm-relative helicity (SRH) are both relevant parameters in measuring vertical shear. Punge et al. (2017) discovered that over an 11-year period in France and Germany, both S06 and SRH are critical factors for large hail (≥ 5 cm) when combined with long-duration storms and extended storm

tracks. Another observation study by Tang et al. (2019) found that an increase in days with instability and vertical wind shear was necessary for the increasing trend in days with favorable large hail environments in the central United States.

For large hail to form, there must be an adequate amount of supercooled water available to grow millimeter-sized embryos to centimeter-sized hailstones within the relatively short time they spend in the updraft growth zone. A high specific humidity (expressed as mass of water vapor per unit mass of air in g/kg) at mid-levels is a critical factor that influences the availability of water vapor for condensation and the subsequent formation of supercooled water droplets. Based on empirical studies of hail-producing storms, the specific humidity in the updraft should be greater than approximately 6 g/kg (Borland et al., 1977). However, as a hailstone descends toward the ground, lower specific humidity in the low troposphere is favoured for large hail. This is because, with a decrease in environmental moisture, there is less water vapor available to condense on the hailstone's surface, while sublimation cools the hailstones and slows their melting. Consequently, the rate of melting may be slower as the release of latent heat is reduced. Melting does not occur until the hailstone descends below the wet-bulb-zero level (WB0), the altitude at which the wet-bulb temperature is 0°C (Markowski and Richardson, 2010). The WB0 is always closer to the surface than the freezing level unless the troposphere is already saturated, in which case, they would be equal. The lower the height of the WB0, the larger a hailstone can retain when reaching the ground.

While we can identify and predict conditions favourable for hail embryo formation and growth with reasonable accuracy, we are unable to determine the critical set or sets of sufficient conditions. Even if all the conditions addressed above are satisfied, hail is not guaranteed. Although CAPE, vertical wind shear, and wet-bulb-zero height have limited individual utility for hail size forecasting, considering them together offers the best strategy for predicting large hail. A combination of a large CAPE-shear product and a low wet-bulb-zero height is especially favourable for the occurrence of the largest hail (Markowski and Richardson, 2010). Researchers are working on developing a unified forecast parameter that consolidates several factors favourable for hail prediction, with the Significant Hail Parameter (SHIP) being one such example. Developed by the U.S. National Weather Service (NWS), SHIP is a composite index proportional to the product of five height-specific parameters: CAPE, vertical shear, water vapor

mixing ratio, environmental lapse rate, and temperature. Typically ranges from 0 to 4, values of SHIP > 1 indicate a favourable environment for significant hail (i.e., hail diameters ≥ 5 cm).

1.2.2 Hail-relevant Environmental Indices

The Bulk Richardson Number (BRN) is a reliable indicator of convective storm classification within given environments. This nondimensional ratio assesses the relationship between buoyant energy (such as CAPE) and the vertical wind shear of the horizontal wind. Both are pivotal factors in determining storm development, evolution, and organization.

$$\mathrm{BRN} = \frac{\mathrm{CAPE}}{0.5\ (\Delta \mathrm{U})^2} \tag{1}$$

Here, ΔU is the wind speed difference between the density weighted 0-6 km mean wind and the lowest 500 m mean wind. BRN values below 10 suggest the shear may be too strong given the weak buoyancy to develop sustained convective updrafts unless given sufficient forcing. For BRN values over 50, multicellular thunderstorm development is likely as multiple non-steady updrafts develop due to the high buoyancy but weaker wind shear. Only when BRN is between 10 and 45 does the environment support supercell development. Since supercells are responsible for producing most of the large hail (Kumjian et al., 2021), BRN is a hail predictor candidate that correlates the storm cell type with the expected hail size.

The K Index (KI) is a measure of thunderstorm potential by evaluating the vertical temperature lapse rate and the quantity and vertical reach of low-level moisture within the atmosphere. Developed by George et al. (2014), it is determined by the following equation:

$$\mathrm{KI} = (\mathrm{T}_{850\mathrm{mb}} - \mathrm{T}_{500\mathrm{mb}}) + \mathrm{T}_{\mathrm{d}_{850\mathrm{mb}}} - (\mathrm{T}_{700\mathrm{mb}} - \mathrm{T}_{\mathrm{d}_{700\mathrm{mb}}}) \tag{2}$$

where T is the temperature and $\mathrm{T_d}$ is the dewpoint (in °C) at the pressure level indicated. A higher temperature and dewpoint at the lower troposphere (i.e., 850 mb), combined with a lower temperature at the mid-level (i.e., 500 mb) and minimal dewpoint depression in between (i.e., 700 mb), result in a higher K Index. The higher the KI value, the greater the likelihood of heavy rain. A KI value ≥ 40 has the best potential for very heavy rain. Since low-precipitation storms tend to produce larger hail (Bluestein and Woodall, 1990), the K Index is a hail predictor candidate that correlates precipitation (in terms of rainfall) intensity with the expected hail size.

The Severe Weather Threat Index (SWEAT) is a stability index developed by the U.S. Air Force to help distinguish severe thunderstorms from non-severe thunderstorms. SWEAT incorporates instability, wind shear, and wind speeds as follows:

$$\mathrm{SWEAT} = 12\,\mathrm{T_{d_{850mb}}} + 20(\mathrm{TT} - 49) + 2\mathrm{U_{850mb}} + \mathrm{U_{500mb}} + 125(\mathrm{S} + 0.2) \tag{3}$$

where $\mathrm{T_d}$ is the dewpoint (in °C) and U is the wind speed (in knots). S is the sine of the angle between the 500 mb and 850 mb wind directions (the shear term). TT represents the Total Totals Index, another stability index that accounts for the lapse rate and moisture content:

$$\mathrm{TT} = (\mathrm{T_{850mb}} - \mathrm{T_{500mb}}) + (\mathrm{T_{d_{850mb}}} - \mathrm{T_{500\,mb}}) \tag{4}$$

A higher TT value implies a higher probability of deep convection and contributes positively to SWEAT. A SWEAT value > 300 means severe thunderstorms are possible, while a value > 400 means thunderstorms with tornadoes are possible. As large hail correlates well with tornadoes (North et al., 2015), SWEAT is a hail predictor candidate for correlating storm severity with expected hail size.

In summary, simple parameters such as vertical wind shear, specific humidity, wet-bulb zero height, CAPE, and CIN as well as composite stability indices such as BRN, KI, and SWEAT are all considered relevant environmental hail predictors. These are therefore good candidates as input to a hail forecasting algorithm. Comprehensive yet sophisticated parameters such as SHIP that combine many factors favorable for hail are also being developed to improve hail prediction.

1.2.3 Radar-based Hail Detection and Nowcasting

Nowcasting refers to weather forecasting with local detail, by any method, over a period from the present to six hours ahead (World Meteorological Organization, 2019). It usually includes a detailed description of the present weather conditions from the latest observation. Early detection of situations conducive to hail aloft or indications of hail presence is crucial in hail nowcasting. Therefore, detection and nowcasting are highly integrated in hail prediction.

Weather radar stands as the most potent remote sensing tool for hail threat detection and nowcasting (Allen et al., 2020). Most weather radars operate by emitting bursts or pulses of electromagnetic waves that travel through the atmosphere and scatter off cloud droplets and precipitation particles. Post-analysis of the backscattered radiation received by the radar provides

details about scatterers, including their location, motion parallel to the wave path, size, shape, orientation, and physical composition (Fabry, 2015).

The most familiar radar-derived quantity is the radar reflectivity factor, hereafter "reflectivity" (Z_H), which provides an estimate of the precipitation size and concentration. Reflectivity is proportional to the particle's equivalent spherical diameter to the 6th power, but this holds only for small particles relative to the radar wavelength (S, C, and X bands), and large hail does not meet this condition. Thus, there is no direct relationship between hail size and observed Z_H (Allen et al., 2020). However, because even the heaviest rain rarely produces Z_H above 50 dBZ, any Z_H value on the order of 60 to 70 dBZ is a reliable indicator of the presence of hail (Stull, 2017). In addition, the presence of large hail can sometimes be inferred from a hail spike in reflectivity data, also known as a flare echo or three-body scattering signature (TBSS). This signature is the outcome of the radar transmission being scattered by airborne hailstones toward the ground, then back to the hailstones, and ultimately back to the radar (Markowski and Richardson, 2010).

Reflectivity features linked to severe storms with large hailstones have been identified by many researchers. Waldvogel et al. (1979) made a significant breakthrough in hail detection research by proposing the difference between the height of 45 dBZ contours and the freezing level as an indicator for hail detection. Lemon (1980) argued that reflectivity of at least 50 dBZ at a height of 8 km above ground level (AGL) is a threshold for hail risk. It assumes that a storm with a significant accumulation of hydrometeors (mainly supercooled droplets) located high above the freezing level is likely to produce large hailstones. In addition, the presence of high low-elevation reflectivity with a threshold of 60 dBZ usually indicates a risk of hail (Battan and Bohren, 1986).

Weather radars with dual polarization (Dual-Pol) can transmit and receive pulses in both horizontal and vertical polarizations, thus enabling measurements of the horizontal and vertical dimensions of the target. Big raindrops exhibit consistent flattening, resulting in a significantly stronger horizontally polarized radar echo when compared to the vertically polarized echo. On the other hand, hail shows a nearly equal radar echo strength at both polarizations due to small hail's lack of flattening and larger hail tumbling, distributing its elongations more evenly in space. Therefore, the difference, known as differential reflectivity (Z_{DR}), serves as a hail signal

when the radar echo is intense (North et al., 2015). The combination of high Z_H and low Z_{DR} has aided the detection of large hail (Allen et al., 2020). Other features that could offer skill in diagnosing hail size include the bounded weak echo region (BWER, Marwitz, 1972) in Z_H or Z_{DR} columns (Kumjian et al., 2014).

Most modern weather radars are dual-pol Doppler radars that can detect and interprets the Doppler effect in terms of the radial velocity of a target (AMS, 2012). Doppler radars can measure the Z_H-weighted mean velocity component projected onto the beam propagation path, called "radial velocity" (Allen et al., 2020). If the radial velocity field exhibits a rotational pattern, it may imply the storm is organized as a supercell with mesocyclone circulation, statistically more likely to produce hail (Stull, 2017). Some indicators of storm severity in the radial velocity field have demonstrated correlations with hail size, such as storm-top divergence magnitude (Witt and Nelson, 1991) and maximum rotation velocity (Witt et al., 2018).

The Next-Generation Weather Radar (NEXRAD) system is a network of 160 high-resolution S-band Doppler weather surveillance radars (WSR-88D) operated by the NWS. It contains numerous algorithms that collect radar-based data as input to produce meteorological and hydrological analysis products (Crum and Alberty, 1993). The needs of various interest groups for hail information led to the development of the operational hail detection algorithm (HDA) for the WSR-88D. For each detected storm cell, the HDA produces information on the probability of hail (of any size), probability of severe hail, and maximum expected hail size (Witt et al., 1998).

1.2.4 Radar-derived Parameters for Hail Nowcasting

Vertically integrated liquid (VIL) has long been used as a warning index for hail (Knight and Knight, 2001). It is an estimate of the total mass of liquid water content in a column obtained from vertical integration of radar reflectivity with a dimension of mass per unit area (AMS, 2012). Since VIL varies greatly based on airmass characteristics, dividing the VIL by the echo top height would "normalize" the VIL and produce a quotient (defined as VIL density) independent of airmass characteristics:

$$\text{VIL density} = \frac{\text{VIL}}{\text{echo top}}. \tag{5}$$

Studies by Amburn and Wolf (1997) using an echo top of 18 dBZ showed a substantial increase in large hail reports (≥ 19 mm) as VIL density increased above 3.5 g m^{-3}.

To determine the presence of severe hail, HDA adopted a reflectivity-to-hail relation instead of a reflectivity-to-liquid-water relation as VIL does (Witt et al., 1998). The algorithm transforms the reflectivity data into flux values of hail kinetic energy $\dot{E}$ (Federer and Waldvogel, 1978):

$$\dot{E} = 5 \times 10^{-6} \times 10^{0.084Z} W(Z), \tag{6}$$

where

$$W(Z) = \begin{cases} 0 & \text{for } Z \leq Z_L \\ \frac{Z - Z_L}{Z_U - Z_L} & \text{for } Z_L < Z < Z_U \\ 1 & \text{for } Z \geq Z_U \end{cases}. \tag{7}$$

Here Z is in dBZ, $\dot{E}$ is in Joules per square meter per second, and the weighting function $W(Z)$ can be used to define a transition zone between rain and hail with a default threshold set to Z_L = 40 dBZ and Z_U = 50 dBZ. Since hail growth exclusively takes place at temperatures < 0°C and severe hail growth is mainly concentrated around temperatures near –20°C or colder (Browning, 1977), the subsequent temperature-based weighting function is used:

$$W_T(H) = \begin{cases} 0 & \text{for } H \leq H_0 \\ \frac{H - H_0}{H_{-20} - H_0} & \text{for } H_0 < H < H_{-20} \\ 1 & \text{for } H \geq H_{-20} \end{cases}, \tag{8}$$

where H is the height above radar level (ARL), H_0 is the height ARL of the environmental melting level, and H_{-20} is the height ARL of the –20°C environmental temperature. Here, H_0 and H_{-20} are determined from a nearby sounding or numerical model of upper-air data. By incorporating $\dot{E}$ with the above two weighting functions, one can obtain a radar-derived parameter known as the severe hail index (SHI) defined as

$$\text{SHI} = 0.1 \int_{H_0}^{H_T} W_T(H) \dot{E} \, dH, \tag{9}$$

where H_T is the height of the top of the storm cell. The units of SHI are Joules per meter per second.

The SHI is also used to estimate the maximum estimated size of hail (MESH), which is probably the most difficult and challenging aspect of HDA. Witt et al. (1998) developed a model

such that around 75% of the hail size observations would be less than the corresponding predictions. This led to the following relation:

$$\mathrm{MESH} = 2.54(\mathrm{SHI})^{0.5}, \tag{10}$$

with MESH in millimetres. MESH is useful for assessing the 2-D distribution of hail according to analysis from the NWS. Both MESH and VIL exhibit comparable skills in distinguishing hail-size categories. Nevertheless, MESH holds an advantage as it exclusively pertains to convective echoes, unlike VIL, which can be observed for any echo (Ortega, 2018). Recent examinations have also indicated that recalibrating MESH thresholds using a more extensive collection of hailstorm samples than the original study by Witt et al. (1998) can lead to enhanced overall proficiency in detecting severe hail (Murillo and Homeyer, 2019).

In summary, simple reflectivity-based parameters such as VIL and echo top height, as well as HDA model products such as POSH and MESH are among the most important hail predictors developed and used in operational hail nowcasting.

1.2.5 Hail Threat Warning Systems

The NWS does not issue "Hail" warnings but incorporates hail threats under "Severe Thunderstorm" warnings. This approach is based on the understanding that hail is one of the primary hazards associated with severe thunderstorms. A Severe Thunderstorm Warning (SVR) is issued when a thunderstorm is occurring or imminent in the warning area with winds of at least 58 mph (50 knots or ~93 km/h) and/or hail of at least 1 inch (quarter-sized, ~2.5 cm) in diameter. A thunderstorm with winds equal to or greater than 40 mph (35 knots or ~64 km/h) and/or hail of at least ½ inch is defined as approaching severe. To further improve communication, NWS also includes the damage threat categories (Table 1.1) for all SVRs issued. The highest of the categories will be invoked from either a qualifying wind or hail value or both. Wireless emergency alert (WEA) messages will be activated on mobile devices when an SVR with a "Destructive" tag is issued or updated.

Damage Threat Category	Wind	Hail Diameter	WEA Activated?
Base (default)	58 mph (~93 km/h)	1.00-inch (quarter)	No
Considerate	70 mph (~113 km/h)	1.75-inch (golf ball)	No
Destructive	80 mph (~129 km/h)	2.75-inch (baseball)	Yes

Table 1.1. NWS thunderstorm damage categories and corresponding expectations.

Some NWS weather forecast offices (WFO) in hail-prone regions also issue a "Severe Hail Hazard Map" that depicts the local threat of severe hail within 12 miles (~19.3 km) of a location. It is based on the likelihood that severe hail (≥ 1-inch) will occur combined with the anticipated size (diameter) of the biggest hailstones. The Storm Prediction Center (SPC) under the NWS also maintains an open database of hail reports (with location, time, and size description) from trained local weather spotters and professionals that update in real-time.

The Meteorological Service of Canada (MSC) also incorporates hail threats under the Severe Thunderstorm Warnings. The warning is issued when there is evidence based on radar, satellite images, or from a reliable spotter that any of the following conditions is imminent or occurring:

- wind gusts of at least 90 km/h
- hail of at least 2 cm in diameter
- heavy rainfall, as per rainfall criteria (≥ 15 mm, 25 mm, or 50 mm per hour depending on the geographic region)

Thus, the MSC has slightly lower warning threshold criteria than the NWS. The 2 cm minimum hail diameter is nearly identical to the original ¾ inch (penny-sized, ~19 mm) threshold set by the NWS before it revised to the current 1-inch (~25 mm) definition for severe hail in 2010. The MSC includes a rough qualitative note on the maximum expected hail size in its warning statement but provides no further spatial and temporal description of the areas under the hail threat. The fact that neither the NWS nor MSC releases a "Hail Warning" highlights the lack of information provided to the public on hail risk levels compared to other thunderstorm hazards such as tornadoes, strong winds, or flash floods.

1.2.6 Other Methods in Hail Forecasting

Alternative methods for hail forecasting have employed one-dimensional coupled hail and cloud models, like HAILCAST, to estimate potential hail size based on atmospheric profiles (e.g., Brimelow et al., 2006; Jewell and Brimelow, 2009). Others use explicit microphysics to predict graupel or hail size from simulated updrafts (e.g., Gagne et al., 2019; Labriola et al., 2019). While not yet widely implemented operationally, both approaches seem to offer

reasonable hail size predictions (e.g., Gagne et al., 2019; Labriola et al., 2019). An instance is the integration of HAILCAST into high-resolution WRF model output using simulated storms (Adams-Selin and Ziegler, 2016). Further assessments and operational trials are essential to ensure these estimates provide accurate guidance, as simulated hail sizes have demonstrated significant sensitivity to the underlying weather model.

A new approach for analyzing parameters beneficial for hail involves machine learning techniques. These strategies have addressed the challenge by accommodating a broader range of variables, encompassing model-simulated updrafts, derived microphysical portrayals of hail, environmental factors, and observed radar information (e.g., Czernecki et al., 2019; Gagne et al., 2019; McGovern et al., 2017). These endeavors aim to offer operational insight for hail prediction and expand the dimensionality of forecast parameters. More details on machine learning and its application in hail forecasting will be presented in the next subsection.

1.3 Random Forest as a Tool in Hail Nowcasting

1.3.1 Supervised Machine Learning

Coined by IBM scientist Arthur Samuel (1959), Machine Learning (ML) gives computers the ability to "learn" without being explicitly programmed. It is a specialized field within artificial intelligence (AI) that involves leveraging data and algorithms to replicate the process of human learning, gradually enhancing accuracy in making predictions and decisions (Zhou, 2021). Without reliance on pre-established equations as models, ML algorithms use computational techniques to acquire insights autonomously from data. As the volume of learning data grows, these algorithms adaptively refine their performance.

Supervised learning is a subtype of machine learning that involves training algorithms using labeled (i.e., tagged with expected output) input datasets to accurately determine outcomes (IBM, 2023). The labeled data acts as a supervisor teaching the algorithm to predict outputs correctly. As the input data is fed into the model, the algorithm aims to establish a mapping function between input and output variables by adjusting its weights through cross-validation until an appropriate fit is achieved. In contrast, unsupervised learning involves training algorithms on unlabeled data to detect patterns independently without predefined output labels. This distinction showcases the supervised approach's focus on precise classification and prediction with predefined outcomes, making it applicable in processes such as image classification, risk

assessment, and anomaly detection (IBM, 2023). In essence, these processes are all critical to modern weather prediction, especially in severe weather nowcasting.

1.3.2 Decision Tree

A decision tree (DT) is a supervised learning approach used to categorize or make predictions based on the answer to a set of true-or-false questions. It predicts the value of a target variable by learning decision rules inferred from the data features and can be seen as a piecewise constant approximation (Breiman et al., 2017). As depicted in Figure 1.3, decision trees start with an ultimate question that forms the root node. From there, a series of questions that make up the decision nodes in the tree can be asked, forming binary branches that split the input data. Observations that fit the criteria will follow the "True" branch, and those that don't will follow the alternate "False" path. When a decision node does not further split into additional sub-nodes, it is denoted as a leaf node and represents a possible final decision outcome. The deeper the tree, the more complex the decision rules and the fitter the model (Hastie et al., 2009).

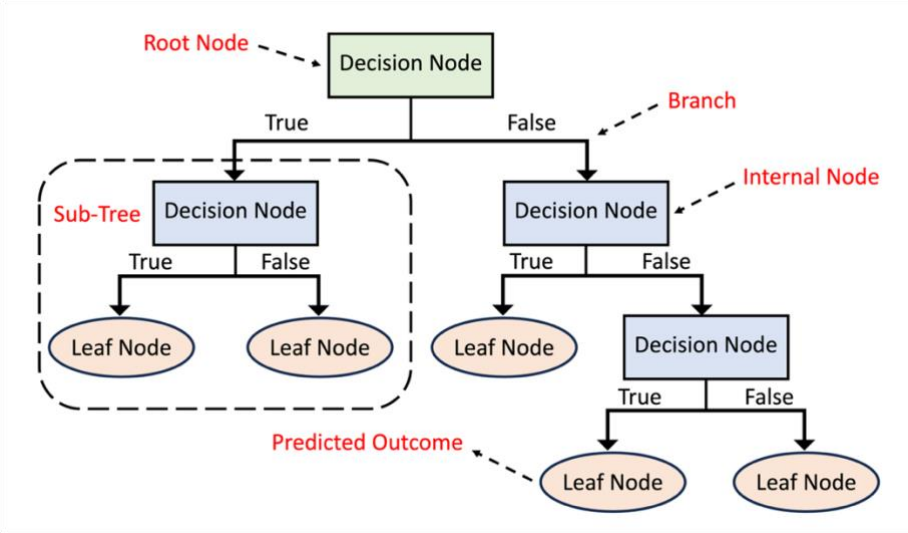


Figure 1.3. Schematic of a simple binary decision tree with five leaf nodes (original work).

Breiman et al. (1984) developed CART (Classification and Regression Tree), a decision tree-based predictive algorithm. It recursively splits the dataset into subsets based on the values of input features, ultimately creating a binary tree-like structure of decisions leading to final outcomes. It is versatile and can handle both categorical and numerical features in multi-class and multi-output tasks. The CART algorithm works in the following steps:

1. Tree Construction: Starts with the whole dataset as the root node and selects the best feature to split the data. The splitting criterion is typically chosen to maximize the separation of classes in classification tasks or to minimize the variance in regression tasks.
2. Recursive Splitting: Repeats the process of selecting the best feature and splitting the data into child nodes for each decision node until a stopping condition is met. This condition could be maximum tree depth, minimum number of samples per node, or other user-defined criteria.
3. Leaf Node Assignment: The growth of the tree terminates at leaf nodes. In classification tasks, the majority class within a lead node determines the predicted class label. In regression tasks, the predicted value can be the mean or median of the target values within the leaf node.

Decision trees are simple to understand and interpret when visualized. They use a white box model, with the decision process to any outcome easily explained by Boolean logic. However, decision trees tend to overfit if they are allowed to grow too deep. Pruning is often applied to remove unnecessary branches and improve model generalization. The CART algorithm has been influential in the development of various ensemble methods, including Random Forests and Gradient Boosting.

1.3.3 Random Forest

Decision trees can exhibit bias and instability due to minor fluctuations in the data, potentially leading to entirely different tree structures. To address this concern, multiple independent decision trees can be constructed to form a random forest (RF) or random decision forest. Like most ensemble methods, random forest enhances the generalization and robustness of a single estimator (e.g., decision tree) by combining predictions from multiple base estimators constructed with a given learning algorithm (Ho, 1995). In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., bootstrap sample) from the training set. In addition, the best split at each decision node is usually found from a random subset of input features. These two sources of randomness help the random forest to achieve a reduced variance by combining diverse trees and aggregating their predictions to identify the most likely outcome (Scikit-Learn, 2023).

Breiman (1996) introduced the Random Subspace Method as part of his research on improving the accuracy and robustness of classification algorithms. This method involves selecting a random subset of features from the original feature set and training a classifier on this subset. By combining the idea of bagging (bootstrap aggregation) with the Random Subspace Method, Breiman developed the RF algorithm (2001) as we know it today. The typical RF algorithm has following main steps:

1. Bootstrapping: Multiple random subsets of the original pre-labeled dataset are created by randomly selecting samples with replacement. Each subset is used to train an individual decision tree.
2. Random Feature Selection: For each decision tree in the ensemble, a subset of features is randomly selected. This procedure introduces diversity among trees and prevents any single feature from dominating the decision-making process.
3. Building Decision Trees: Each decision tree is constructed using the bootstrapped data subset (from Step 1) and the randomly selected features subset (from Step 2). The trees are built using a criteria-based recursive process (e.g., CART algorithm).
4. Majority Voting or Averaging: With all decision trees in the RF ensemble now completely built, each tree produces its own prediction, either a class label in classification or a numerical value in regression. For classification tasks, the class with the majority vote among the trees is chosen as the final prediction. For regression tasks, the predictions from all trees are averaged to produce the final prediction.

Note that in the last step above, the final prediction is deterministic. However, it is also possible to generate a probabilistic prediction by analyzing the frequency of each class from the aggregation of predictions by all trees. Figure 1.4 illustrates the schematic of a random forest with a probabilistic classification prediction. Compared to standalone decision trees, the random forest algorithm gained popularity due to its effectiveness in reducing overfitting and improving predictive accuracy. It also provides a measure of feature importance, aiding in understanding the significant attributes in the dataset. In summary, random forest combines the strengths of decision trees and ensemble methods to create a powerful and versatile algorithm for supervised learning tasks.

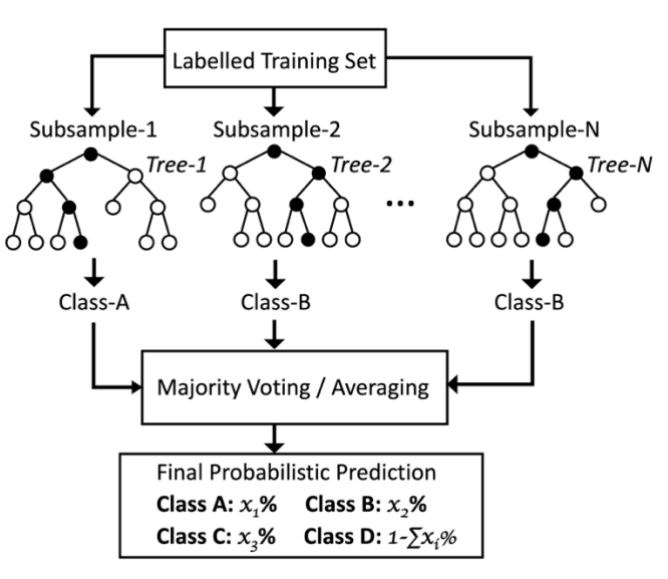


Figure 1.4. Sample schematic of a four-class random forest classifier with N decision trees (original work).

1.3.4 Application of Random Forest in Hail Nowcasting

In contrast to physics-based weather forecasting models that formulate prognostic and diagnostic equations from primitive equations, a set of assumptions, and initial and boundary conditions, ML models recognize patterns within a dataset and subsequently establish a connection between these patterns and an outcome, optimizing an error metric and producing calibrated predictions (Gagne et al., 2017). Compared to other ML methods, random forests stand out because they can handle predictors without normalization, be more computationally efficient, and require little tuning (Loken et al., 2022).

McGovern et al. (2010) were among the earliest to apply random forest to severe weather forecasting with the development of the Spatiotemporal Relational Random Forest (SRRF). Their research demonstrated that the SRRF algorithm is a strong predictor with variable importance estimation that could aid predictions of turbulence, tornadoes, and droughts. In the domain of hail prediction, studies have shown RF classifiers to outperform deep learning-based methods such as convolutional neural networks (CNN, Pulukool et al., 2020) and other ML classifiers, such as support vector machine (Gensini et al., 2021) with lower false alarms and higher precision scores.

Gagne et al. (2015) were the first to develop an RF-based approach to day-ahead expected hail size prediction by identifying potential hailstorms in storm-scale numerical weather prediction (NWP) models and matched them with observed hailstorms. They selected 18 input variables to test on three ML methods (random forest, gradient boosting regression trees, and linear regression) and found that RF produced the subjectively best forecast. However, none of the models can predict hail above 60 mm in diameter, as there was very little training data at these sizes. In a following study, Gagne et al. (2017) used an RF classifier to forecast the probability of hail occurring and another RF regressor to forecast the MESH size distribution of each hailstorm. For 24-hour hail outlooks, the RF methods identified hail threat areas while minimizing false alarms with better performance than HAILCAST and other storm-surrogate methods. Building on the work by Gagne et al. (2017), a study that compared two methods (individual members vs. ensemble mean predictors) to create RFs for next-day severe weather prediction was carried out by Loken, Clark, and McGovern (2022). Their findings suggest that the most important predictors are storm variables (e.g., 2–5 km updraft helicity), followed by index and environmental variables.

RF classifier algorithms have also been applied to hail nowcasting on the mesoscale. Using 72 physical quantities and convection parameters calculated from the ERA-Interim reanalysis data, Yao et al. (2020) built an RF classifier for hail prediction in the 0–6 hour forecast range for a local region in China. Some of the most significant predictors include CAPE, the height of the −20°C/ −10°C/0°C isotherm, and several storm composite indices. The RF model achieved a probability of detection (POD) of 90.1% and a false alarm ratio (FAR) of 9.7%, and the predicted potential area of hail event is in good agreement with the actual area of hailfall. In another study, 15 radar products from a single C-band Doppler radar were used to construct an RF classifier for hail disastrous weather nowcasting in 15 minutes to 1 hour (Huang et al., 2019). By classifying the hail weather into four scenarios (with or without strong wind and/or short-time heavy precipitation), the RF model obtained a mean POD of 74.8% and a mean FAR of 24.4%.

The concept of data fusion is gaining recognition in severe weather forecasting with ML methods. The incorporation of input data from multiple sources with the ML approach increases the diversity among predictors and makes it possible to compare contributions by predictors from different origins. By coupling radar reflectivity, lightning occurrence, and convective parameters derived from the ERA5 reanalysis data as inputs, Czernecki et al. (2019) developed an RF

classifier for large hail (> 2 cm) prediction in Poland. Among the 35 hail predictors used in the study, the sole radar-derived variable, column maximum radar reflectivity (CMAX), ranked the highest in feature importance. The RF classifier captured over half of all large hail events (POD = 50.7%) despite a high proportion of false alarms (FAR = 86.7%). Another example of the data fusion approach involved evaluating geostationary satellite and lightning information with ground-based radar to identify severe (with hail > 2.5 cm) versus non-severe storms by Mecikalski et al. (2021). By analyzing 49 atmospheric variables related to May, daytime Great Plains convective storms over two years, this study concluded that the radar-derived Z_{DR} minimum where Z_H > 45 dBZ was the most important predictor. Compared to the experiment by Czernecki et al. (2019), this RF model achieved better performance scores with a POD of 72% and a FAR of 17%.

RF-based probabilistic predictions of severe weather across CONUS have been compared to operational forecast products. Using 15 atmospheric variables collected over nine years of the historical forecast from the Second Generation Ensemble Forecast System Reforecast (GEFS/R) ensemble of the National Oceanic and Atmospheric Administration (NOAA), Hill et al. (2020) trained RF models separately for hail, tornado and wind prediction on Day 1 and collectively for Day 2–3 analogous to SPC's convective outlook. Their study revealed that RF outlooks produce probabilistic forecasts that slightly underperform SPC outlooks on Day 1 but significantly outperform them on Day 2 and Day 3. Further investigation by McCloskey et al. (2021) showed that the RF-based hail prediction is skillful in reducing false alarms by forecasting one outlook category lower than that of the SPC on Day 1. RFs could improve the SPC outlooks by calibrating hail probabilities based on the strength of simulated storms.

RF has demonstrated substantial skill in hail forecasting compared to traditional physics-based numerical models, even some alternative ML methods. It can capture the complex interactions between various meteorological parameters with an ensemble of independent decision trees. With carefully selected input predictors from multiple sources, appropriate algorithm design, and a well-crafted model training and validation process, RF models can be tailored to meet specific hail forecast needs. Some of the most pressing challenges in RF-based hail forecasting currently include:

1. Data Preparation: Accurate hail prediction requires a set of key physical parameters from various sources (radar, satellite, reanalysis), which might not always be available.
2. Limited Hail Observations: Hailstorms are relatively rare events leading to imbalanced datasets. This can affect model performance and lead to biased predictions.
3. Generalization: Ensuring that RF models generalize well to different geographic locations and diverse meteorological conditions remains a challenge. Overfitting specific datasets can hinder the model's performance on new data.
4. Lead Time Specificity: Tailoring RF models to specific lead times can be complex, as the importance of variables changes with different forecast horizons.
5. Model Interpretability: While RF provides feature importance rankings, interpreting the exact physical processes behind these rankings can be challenging, limiting our ability to improve physical understanding.

In conclusion, RF has shown promise in improving hail prediction accuracy and understanding the associated severe convective weather. However, challenges related to data availability, generalization, and model interpretation persist. Addressing these challenges will likely require interdisciplinary collaboration between meteorologists and machine learning experts.

1.4 Research Objectives and Contribution of Authors

1.4.1 Statement of the Research Problem and Objectives

The main objective of this research is to improve the existing hail threat warning system by developing a random forest classifier hail nowcasting model that takes a set of predictors from remote sensing data and environmental variables as inputs and produces a probabilistic prediction of the maximum estimated hail size range for the hail threat area. The study will focus on severe hail activities in the months of May to August during the second half of the day (between noon and midnight) in a rectangular domain that approximately overlaps with the Midwestern United States geographic region. The study will first investigate remote sensing (i.e., radar-derived) predictors-only, followed by environmental (i.e., reanalysis-derived) predictors-only, before coupling the most important predictors from the two sources to develop an integrated RF-based hail nowcasting model. The RF classifier will be trained using four years of

data with 5-fold cross-validation and tested with one year of data. The final product is a map of probabilistic prediction on the maximum estimated hail diameter in four size categories: < 1 mm, 1–5 mm, 5–20 mm, and ≥ 20 mm, available at two lead times: 15 minutes and 60 minutes. Performance analysis will determine the most important hail predictors used and the forecast skill of the RF classifier.

1.4.2 Contribution of Authors

All chapters of this thesis are written by graduate student Zhicheng Jing in the Summer of 2023. The thesis is a result of the original work by the graduate student. I, Zhicheng Jing, prepared the source dataset, developed the Python program for the random forest model, carried out training and testing of the model, and analyzed the results from the experiment. Co-author Professor Frédéric Fabry provided guidance on the design of the research, supervised the project, and edited the thesis.

2. Methodology

2.1 Source Data Collection and Preprocessing

2.1.1 Spatiotemporal Scope of the Study

This study focuses on hail prediction within a domain that overlaps with the Midwestern United States. As depicted in Figure 2.1, it is bounded by 105°W to the west, 85°W to the east, 49°N to the north, and 39°N to the south. With an approximate area of 1.78M km^2, this region lies between the lee side of the Rockies and the Great Lakes region, with analogous continental climates and mostly flat terrain. The northern edge of the domain borders Canada, making it viable for future studies in a Canadian context.

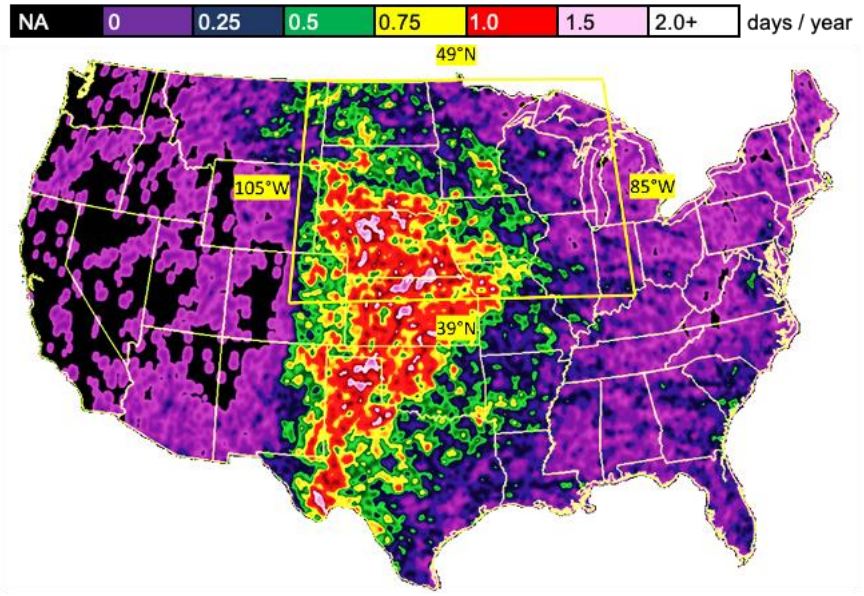


Figure 2.1. Annual hail days per year for CONUS during 2007–2010 with the yellow grid specifying the spatial domain of this study. Colored map adapted from Cintineo, (2012).

According to SPC hail reports from 1955 to 2015, over 80% of hail events in this region occurred during the second half of each day in the late spring and summer months. Taking this seasonal and diurnal hail frequency distribution into consideration, we chose to limit the temporal scale of this study to the 12 hours between local noon and midnight (12:00 to 00:00 Central Standard Time) over the four months of May, June, July, and August. This temporal scope reduces the amount of data to be analyzed while capturing the majority of hail reports.

2.1.2 Hail Reports

The Storm Events Database is a comprehensive and publicly accessible repository maintained by the National Centers for Environmental Information (NCEI), a part of the NOAA. This database contains detailed records of significant weather occurrences, such as tornadoes, thunderstorm winds, and hail from severe convective storms. As part of the storm data, reports of hail that is ¾ of an inch or larger in diameter are available from 1996 onward. Each hail record includes the beginning and end time of hail to the nearest 5 minutes with a single point location in latitude and longitude to the nearest minute resolution.

In this study, we defined a severe hail day as any day with at least 10 severe hail reports (≥ 19 mm) from the Storm Event Database in the spatiotemporal domain specified. A data filter was constructed to screen out severe hail days over the five years from 1999 to 2003. As depicted in Figure 2.2, there are a total of 240 severe hail days, an average of 48 days per year or 12 days per month. The number of severe hail days usually peaks in July, with about half of the days in the month witnessing significant hail activities somewhere within this region.

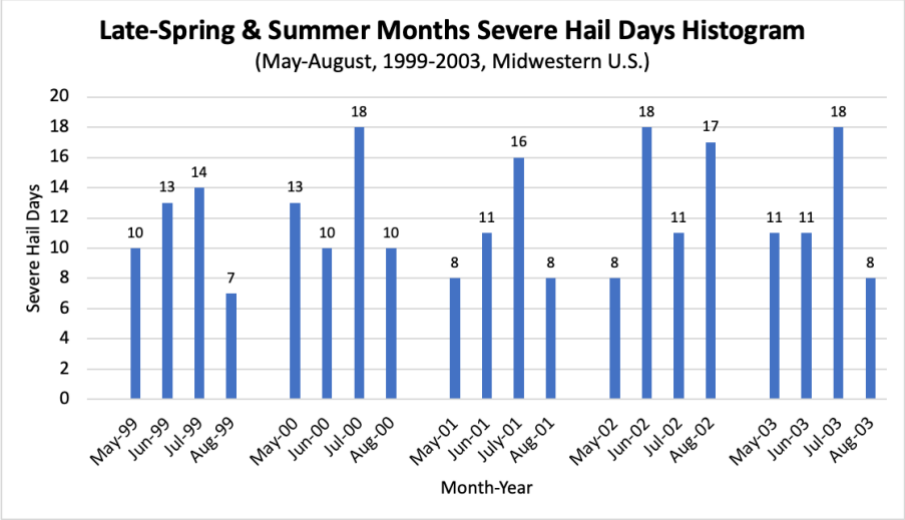


Figure 2.2. Histogram of monthly severe hail days for the spatiotemporal domain specified in this study. Data analyzed from the Storm Event Database of NCEI.

It is worth noting that a severe hail day does not indicate widespread hail occurrence across the entire domain. While some regions may experience severe thunderstorms capable of producing large hail, other areas may have fair weather throughout the day. Therefore, there

usually exists a wide range of atmospheric conditions at any given time that may or may not produce hail. This element is crucial since a hail nowcasting model should be trained on a comprehensive dataset that contains both hail and hail-free cases.

2.1.3 Remote Sensing Data

The Multi-Year Reanalysis of Remotely Sensed Storms (MYRORSS) initiative employs the Multi-Radar Multi-Sensor (MRMS) framework to integrate radar data and nearby environmental information, generating a comprehensive three-dimensional radar representation across CONUS (Williams et al., 2022). At the time of this study, MYRORSS has completed the reprocessing of the WSR-88D radar archive spanning the years 1998 to 2011. All products have a horizontal resolution of 0.01 × 0.01 latitude/longitude degree, except the azimuthal shear products at 0.005 × 0.005 latitude/longitude degree. Temporal resolution is approximately 5 minutes. With improved data quality and accuracy, the MYRORSS dataset leads in many aspects over other remote sensing data sources for meteorological research. In this study, we used a selection of 2D Products from the MYRORSS dataset to compute the 9 remote sensing parameters (see Table 2.1) for the hail nowcasting model.

Group	Abbreviation	Description	Unit	Variants
Composite	MESH	Maximum Estimated Size of Hail	mm	20-km, 40-km
Reflectivity	ETOP18	Echo top heights for 18 dBZ	km	20-km, 40-km
	RALA	Reflectivity at the lowest altitude	dBZ	20-km, 40-km
	RAM20C	Reflectivity at isotherm –20°C	dBZ	20-km, 40-km
	VIL	Vertically integrated liquid	kg m^{-2}	20-km, 40-km
Dynamic	LLAS	Low-level azimuthal shear (0–3 km AGL)	s^{-1}	20-km, 40-km
	MLAS	Mid-level azimuthal shear (3–6 km AGL)	s^{-1}	20-km, 40-km
	LLAS_QI	LLAS quality index	unity	20-km, 40-km
	MLAS_QI	MLAS quality index	unity	20-km, 40-km

Table 2.1. List of the 9 remote sensing parameters used in this study (all derived from the 2D Products of the MYRORSS dataset).

MESH was chosen as the only readily available composite index from the MYRORSS dataset. As introduced in Section 1.2.4, MESH is an ideal reference parameter for hail size prediction. The four parameters ETOP18, RALA, RAM20C, and VIL provide insights into the vertical structure of the thunderstorm. Higher ETOP18 indicates a greater vertical extent of the storm updraft, which is conducive to the development of large hailstones (Amburn and Wolf, 1997). Since the altitude range around −20°C favours hail growth (Miller, 1988), RAM20C values above 60 dBZ could indicate hail aloft, while RALA values that high suggest potential hail near the ground level. VIL was included given its role in indicating the presence of supercooled water required for hailstone growth. Lastly, the azimuthal shear at the low-level and mid-level were picked as dynamic factors. Azimuthal shear is a key indicator of storm rotation. Rotating updrafts (i.e., mesocyclones) are often associated with supercells, the most favourable type of thunderstorms for producing large hail.

In the context of designing a short-lead forecasting system, the above parameters were all computed in two variants by taking the average of all data points within a 20 km × 20 km (40 km × 40 km) grid with its center displaced to the west by 10 km (20 km) from the location of interest. The argument behind this approach is the following:

- The westward displacement: Since storm systems travel as they evolve, the core of a storm that impacts one place would be located some distance away from it (upstream) before it arrives. Since most thunderstorms in this region travel eastward, a shift to the west would better account for the location of the storm when the forecast is made.
- The square-grid averaging method: To account for the scale of the storm system and uncertainties in its development, information about the volume centered around the storm can be more valuable than a single column inside the storm.
- Two grid sizes: Two versions of grid size were used to account for uncertainties in the horizontal dimension and the relative speed of motion of the storm cell for a forecast lead time of up to 1 hour.

For each azimuthal shear parameter, a quality index was also computed. The quality index tells the proportion of non-zero data points used to compute the average shear value for each grid (0 indicates none, and 1 indicates all). Since azimuthal shear is almost always zero except in

supercells, the quality index can reveal more about the scale and intensity of the overall storm system than an averaged shear value can.

In total, 18 candidate hail predictors (9 parameters × 2 variants) were derived from the MYRORSS remote sensing data. Depending on the relative performance of each parameter in Phase 1 of the experiment, they may be subject to change or elimination in Phase 2 to improve the accuracy and efficiency of the RF model.

2.1.4 Environmental Data

In addition to the storm-relevant remote sensing data, we considered beneficial to have information about the general environmental condition before and at the time of the storm. The ERA5 product is the most up-to-date, comprehensive, and high-resolution atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis involves using a combination of observations from various sources with a numerical weather model to create a detailed and consistent representation of past atmospheric conditions. The ERA5 data are available in single levels and pressure levels with a spatial resolution of 0.25°, an interval of 1 hour, and a vertical resolution of 25 hPa from 1000 hPa to 100 hPa. The ERA5 and its predecessor ERA-Interim have been applied in investigating hail-producing thunderstorm environments in many studies (e.g., Czernecki et al., 2019; Yao et al., 2020; Pulukool et al., 2020).

In this study, a total of 12 physical quantities and convection parameters closely related to hail were initially selected and calculated as the prediction factors of the hail nowcasting RF model. A summary of these hail predictors is presented in Table 2.2.

Group	Abbreviation	Description	Unit
Composite	BRN	Bulk Richardson Number	unity
	SWEAT	Severe Weather Threat Index	unity
	KI	K Index	unity
Thermodynamic	CAPE	Convective available potential energy	$J\ kg^{-1}$
	CIN	Convective inhibition	$J\ kg^{-1}$
	SSI	Surface solar irradiance (averaged over 1 hour)	$W\ m^{-2}$
Dynamic	LLVS	Low-level vertical shear (0–3 km AGL)	s^{-1}

	MLVS	Mid-level vertical shear (3–6 km AGL)	s^{-1}
	RV500MB	Relative vorticity at 500 mb pressure level	s^{-1}
Humidity	LLSH	Low-level specific humidity (0–3 km AGL)	kg kg^{-1}
	MLSH	Mid-level specific humidity (3–6 km AGL)	kg kg^{-1}
	ZDL	Zero-degree level (AGL)	km

Table 2.2. List of the 12 environmental parameters used in this study (all derived from the ERA5 dataset).

CAPE and CIN are essential components in understanding the atmospheric conditions conducive to thunderstorm formation. While high CAPE values indicate the potential for strong updrafts and instability, CIN represents the initial resistance to convection. SSI indicates the solar radiation per unit area reaching the ground surface, which depends on the solar angle. It is included to investigate the potential seasonal and diurnal influence on convective destabilization and hail. LLVS and MLVS account for bulk vertical shear, while LLSH and MLSH account for moisture content (in the 0–6 km AGL). Both are relevant to ingredients to hail formation. As described in Section 1.2.2, both WB0 and SRH are highly relevant to forecast large hail. However, due to the difficulties in computing WB0 and SRH directly from the ERA5 data, we replaced them respectively with ZDL and RV500MB as substitute hail predictors in this study.

All three composite indices selected were introduced in Section 1.2.2 as candidate hail predictors with distinct characteristics. K Index is a readily available parameter in the ERA5 dataset that measures the potential for a thunderstorm to develop, calculated from the temperature and dew point in the lower part of the atmosphere. BRN is computed from CAPE, LLVS, and MLVS. As a ratio of buoyancy to vertical shear, high values of BRN favour sustained supercells, which are associated with most hail. The SWEAT index is computed by incorporating several dynamic and thermodynamic factors into one to assess severe weather potential, which is associated with tornadoes and large hail. There exist more well-known indices that can be computed using ERA5 data but were not added to the list as the three indices selected already provide decent representation of different aspects of thunderstorm potential.

Since the ERA5 reanalysis data have a spatiotemporal resolution coarser than that of the MYRORSS remote sensing data, linear interpolation on the gridded environmental parameters

was carried out to match its resolution with the radar parameters. The domain of study was divided into a network of 0.01° (latitude) by 0.01° (longitude) grid cells, with 30 hail predictors computed at a 5-minute time interval for each cell. The grid points data are used directly as input to the RF hail nowcasting model. The prediction covers 0–1 hr.

According to the ECMWF, the uncertainty estimation for ERA5 comes from the Ensemble of Data Assimilations system, which addresses some uncertainties of the model and data assimilation system, but not everything. The uncertainty estimates mostly account for random errors and not for systematic ones. Although it is the latest reanalysis product, ERA5 does not resolve the boundary layer at a desirable resolution for mesoscale severe weather analysis.

2.1.5 Training, Validation and Test Data Sets

In supervised machine learning, it is common practice to split the available learning data into training and test sets. The training set is used to train the model by exposing it to a large portion of labeled data, allowing it to learn patterns and relationships. The test set serves as an independent dataset to assess the model's performance after training. Empirical studies show that the best results are obtained if one uses 70–80% of the data for training, and the remaining 20–30% of the data for testing (Gholamy, 2018). Since exactly 20% of the five-year hail data come from the final year in the period of study, we carried out an 80/20 split using data from the first four years (192 days from 1999 to 2002) for training and the remaining year (48 days in 2003) for testing. In addition, we constructed a validation set by splitting the training set into five equal folds for cross-validation. More details on the purpose and generation of the validation set can be found in Section 2.2.3. In short, it provides an evaluation of the model fit on the training data while tuning the model's hyperparameters. Process regarding how input grid points data is screened and integrated to build the training set and test set are presented in Section 2.2.4, as this context is highly integrated with the ML model design principle.

2.2 Experiment Design

2.2.1 Random Forest Algorithm

As introduced in Section 1.3.3, the RF algorithm is a decision-tree-based ensemble learning method. Compared to a single decision tree, RF is less prone to overfitting and has better generalization to new, unseen data. Compared to deep learning, RF offers better model

interpretability and requires less computational resources and time for training. Also, RF provides a measure of feature importance, indicating which features contribute the most to the prediction. In this study, the hail nowcasting model is created using the random forest classifiers algorithm from the Python module Scikit-Learn (version 1.3.0, Pedregosa et al., 2011).

The steps to implement the RF classifier algorithm in Scikit-Learn are as follows:

1. Assuming in total there are N samples in the original training set, N_{tree} sample subsets are generated with random sampling with replacement (i.e., bootstrapping). Each bootstrapped subset is used to train an individual decision tree (N_{tree} trees in total).
2. Assuming there are M features (aka. predictors, attributes), M_{try} (usually a fraction of M) features are selected with random sampling at every node of every tree.
3. Each individual decision tree is constructed using its corresponding subset of training data (from Step 1) and a subset of features (from Step 2). An optimal feature is chosen from M_{try} features as the splitting factor at each node using a pre-specified criterion.
4. The RF classifier aggregates the predictions from each of the N_{tree} classification trees to make the final prediction through a majority vote.

The above procedure is a highly simplified breakdown of the RF classifier algorithm used (i.e., sklearn.ensemble.RandomForestClassifier). It was built on the foundation of the original work by Breiman (2001), which is beyond the scope of this thesis. The specific details of the implementation, such as optimization techniques, data structures, and parallelization strategies, have been developed and tailored in Scikit-Learn (2023) to ensure efficiency and usability within the library's framework.

2.2.2 Hyperparameter Tuning

In the RF classifier algorithm, N_{tree} and M_{try} are just two of the many customizable hyper-parameters. Hyper-parameter tuning is a crucial step when working with any ML algorithms to optimize their performance, enhance generalization, and prevent overfitting. Hyperparameters are settings defined before training a model and cannot be learned from the data. These hyperparameters determine the characteristics of the trees and of the forest that will be used in the RF algorithm. Tuning these hyperparameters helps find the best configuration for the specific problem and dataset.

Table 2.3 outlines the 10 hyperparameters investigated for tuning during the construction of the RF model. We narrowed down the tuning range for each hyperparameter to three candidate values. Here, the middle value was initially decided based on empirical knowledge in RF, then another two values were chosen to be a certain level lower or higher than that initial estimate. Take N_{tree} as an example, the number of trees in the random forest (n_estimators) needs to be sufficiently large to stabilize the error rate. Boehmke and Greenwell (2019) suggest starting with 10 times the number of features (M) while Ellis (2021) recommends value ranging from 50 to 400 trees. As a compromise, we considered three values, N_{tree} = 100, 10*n_features, or 400 in this study. Next, consider M_{try} as another example: typically, default values for the number of features to consider when looking for the best split (max_features) are $M_{try} = \frac{M}{3}$ for regression or $M_{try} = \sqrt{M}$ for classification (Boehmke and Greenwell, 2019). Since we are working with an RF classifier, we took $M_{try} = \sqrt{M}$ - 1, $\sqrt{M}$, or $\sqrt{M}$ + 1 rounded to the nearest integer as the three candidates for tuning. The optimal tuning range for all other hyperparameters in Table 2.3 was determined following the same principle. The only exception is the tree node split quality criterion, which has exactly three supported criteria pre-determined in Scikit-Learn. Criterion "gini" is for Gini impurity and both "log_loss" and "entropy" are for the Shannon information gain (Scikit Learn, 2023). The hyperparameters not listed in Table 2.3 are mostly Boolean parameters irrelevant to the construction of the trees and were set to default values throughout the training of the RF classifier.

Abbreviation	Description	Tuning Value Candidates
n_estimators (N_{tree})	Number of trees in the RF	100, 10*M, 400
max_samples	Size proportion of the training dataset for bootstrap sampling	0.10, 0.15, 0.20
max_depth	Maximum depth of the tree	10, 15, 20
criterion	Tree-node split quality criterion	"gini", "entropy", "log_loss"
max_features (M_{try})	Number of features to consider when looking for the best split	$\sqrt{M}$ - 1, $\sqrt{M}$ or $\sqrt{M}$ + 1
min_samples_split	Minimum proportion of samples required to split an internal node	0.10, 0.15, 0.20

min_samples_leaf	Minimum proportion of samples required to be a leaf node	0.05, 0.075, 0.10
max_leaf_nodes	Maximum number of leaf nodes	20, 25, 30
min_impurity_decrease	Minimum decrease in impurity to split an internal node	0.01, 0.025, 0.05
ccp_alpha	Complexity parameter used for Minimal Cost-Complexity Pruning	0.01, 0.015, 0.02

Table 2.3. List of the 10 hyperparameters and their tuning values investigated in this study.

While assessing various configurations in hyperparameters for estimators, there remains a risk of overfitting on the test set due to parameter adjustments leading to optimal performance. Consequently, knowledge about the test set might influence the model, rendering evaluation metrics less indicative of generalization performance. To address this, a separate "validation set" is needed between the training set and the test set for unbiased hyperparameter tuning. However, partitioning the available data into three sets comes at the expense of reducing the size of the other two sets.

To address this problem, there is a procedure known as cross-validation (CV). In the basic approach called k-fold CV, the test set is still held out for final evaluation, but the training set is split into k smaller subsets or folds. The model is trained k times, each time using k–1 folds for training and the remaining fold for validation. The performance measure of hyperparameters from k-fold CV is the average of loop-computed values. Common choices for k are 5 or 10, with larger values of k providing more reliable performance estimates but requiring more computational resources. In this study, we determined the optimal hyperparameters by applying 5-fold cross-validation, as depicted in Figure 2.3.

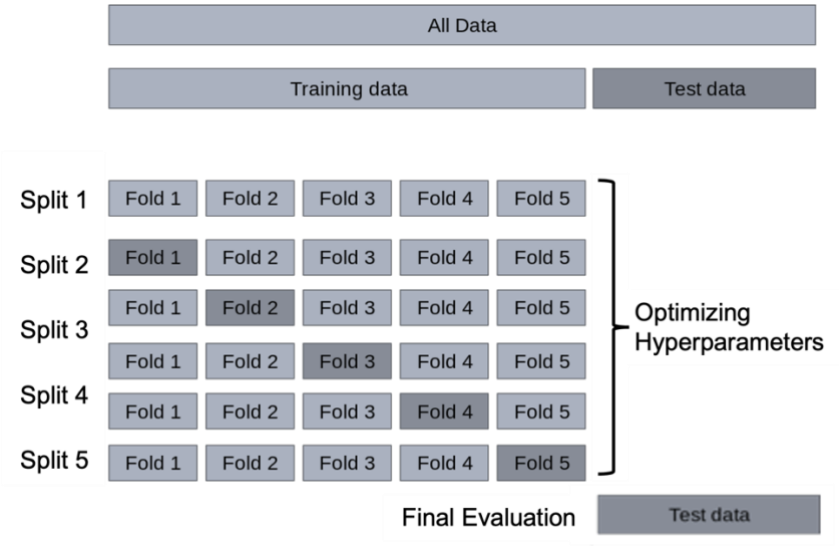


Figure 2.3. Data splitting demonstration for a 5-fold cross-validation. Image adapted from Scikit-Learn developers (2023).

Since the possible combinations of hyperparameters are on the order of 3^{10}, it is extremely inefficient and unrealistic to perform an exhaustive search for the optimal hyperparameters. Scikit-Learn offers an alternative method called RandomizedSearchCV, which implements a randomized search over all parameter settings. A total of 1000 search iterations were made for each of the 5 splits (k = 5), generating a total of 5000 combinations to estimate near-optimal hyperparameters.

2.2.3 Hail Nowcasting Model Setup

The core of the hail nowcasting model developed in this study was implemented through the RF classifier algorithm described. Pre-processing of the raw data was made to derive the 18 radar hail predictors and 12 reanalysis hail predictors for all samples within the determined spatiotemporal domain over 240 severe hail days. For the training set, each sample was labeled with the actual hail size (0 if no hail) at the lead time for learning purposes. For the test set, the hail size prediction made for each sample was compared with the actual hail size for verification purposes. Figure 2.4 is a simplified flow chart of this hail nowcasting model whose different elements were presented earlier.

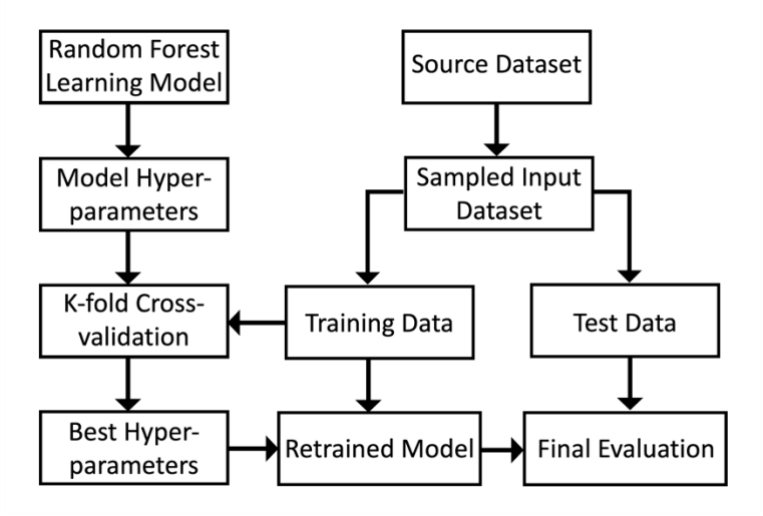


Figure 2.4. Design flow chart of the random forest-based supervised learning model suite for hail nowcasting.

At the beginning of this study, we planned to use hail reports from the Storm Event Database for labeling training samples and verifying test samples. However, there are two problems with this approach. First, manual hail reports have coverage bias toward populated areas and roads during the daytime. Second, negative hail reports (i.e., no-hail situations) do not exist in the database. To avoid the spatiotemporal bias and class imbalance that exist with the manual hail reports dataset, we opted for the MESH value at the time of forecast verification for this role. The MESH parameter is always available for all grid cells (spatiotemporally unbiased) and can have zero values (no-hail situation). Although MESH is not the actual observed hail size but an estimated one, it is the least problematic candidate for hail verification. The output of the hail nowcasting model is thus the predicted MESH for each grid cell in the next 0–1 hr. An alternative approach would be accepting an imbalanced set of hail reports by resampling appropriately to minimize the bias, which could be more reliable than verifying the hail prediction results against a non-perfect parameter like MESH. Nonetheless, this study opted for forecasted MESH as the validation parameter instead of hail reports statistics.

Multiple RF models were developed in three consecutive stages of the experiment, with variable importance and performance metrics calculated separately:

Stage 1: Based only on the MYRORSS remote sensing parameters (18 hail predictors), the binary-class deterministic RF classifiers predict whether hail above a certain size will occur. Nine independent models were developed for three hail diameter thresholds (≥ 1 mm, ≥ 5 mm, ≥ 20 mm) at three lead times (15-min, 30-min, 60-min).

Stage 2: Based only on the ERA5 reanalysis parameters (12 hail predictors), the binary-class deterministic RF classifiers predict whether hail above a certain size will occur. Nine additional independent models were developed for three hail diameter thresholds (≥ 1 mm, ≥ 5 mm, ≥ 20 mm) at three lead times (15-min, 30-min, 60-min).

Stage 3: Based on a downsized and balanced set of parameters from two data sources (10 MYRORSS-based hail predictors and 10 ERA5-based hail predictors) with the least relevant hail predictors from each source removed. The multi-class probabilistic RF classifiers predict the maximum hail size with probabilities distributed across several size ranges. Two independent models were developed to forecast four hail diameter categories (< 1 mm, 1–5 mm, 5–20 mm, ≥ 20 mm) at two lead times (15-min, 60-min).

The first two stages in the experiment are identical, except one used only remote sensing predictors, and the other used only reanalysis predictors. Both were trained and tested using data from one randomly selected severe hail day (June 3, 1999). Stage 3 is the most comprehensive exercise with an improved RF model design that takes advantage of lessons learned from the earlier stages. The final RF classifiers were trained with four-year data from 1999 to 2002 and tested with one-year data in 2003, combining hail predictors from two sources. Additional significant upgrades include the shift from deterministic to probabilistic hail prediction and ensuring a balanced number of samples across the four classes of hail size distribution with a modified data sampling technique. All RF model calculations, including the pre-processing and post-analysis operations, were done using the Python programming language and its packages supporting the ML techniques.

2.2.4 Generation of the Input Data Sets

In Section 2.1.5, we briefly touched on the three input data categories (training, validation, and test sets) and their relationships in a supervised learning model. With an understanding of the design principle of our hail nowcasting model, we can now present the methodology behind the generation of the input data sets.

The training set is composed of N training examples of the form $\{(x_1, y_1), ..., (x_N, y_N)\}$ such that x_i is the feature vector of the i-th example and y_i is its label (i.e., class, or the verification). Consider an example from the training set, its feature vector encompasses the most recent estimated values of hail predictors (radar and reanalysis parameters selected for the ML model) for a single grid point at a time instance. Depending on the forecast lead time of the model, the corresponding label is the hail-size class determined from the MESH value at the time of forecast verification. For example, if the elements of the feature vector have a time stamp of 20:00 local time, and the nowcasting model under training has a forecast lead time of 15 minutes, the MESH value at 20:15 for this location will be referenced to label this training example. Say MESH equals 7 mm in this case, it will be labeled as "True" for the "5–20 mm" hail-size class. The RF algorithm will then treat this training example as a "small hail" observation as it "learns" to predict the maximum hail size range to be expected while processing the training set. Ultimately, the learning algorithm seeks a function g: $X \rightarrow Y$, where X is the input space (instances of the current state of the atmosphere represented by the chosen meteorological parameters), and Y is the output space (instances of predicted local maximum hail size at a specified time and location). The generation of the test set follows the same principle as the training set, except that the feature vectors are not pre-labeled. When the training and validation of the RF model are complete, the test set is then used to assess the performance of this fully specified classifier. The fully trained nowcasting model predicts the hail size class for each feature vector example. Those predictions are compared to the examples' "true" classifications to verify the model's accuracy.

The input training set is composed of individual examples pre-labeled with one of the four hail-size classes. To achieve an unbiased hail-size prediction outcome, the training data must possess an equal number of examples drawn from each of the four hail-size classes. For any given time in our domain, the majority (> 99%) of the grid points are usually labeled as no hail (< 1 mm). Even during a massive thunderstorm outbreak on a severe convective day, the proportion of grid points with large hail (≥ 20 mm) rarely exceeds 0.01 % at its peak level. This means that an algorithm that always predicts "no hail" would be right more than 99% of the time, but it would be valueless. By training the algorithm with equal numbers of examples of the four classes, we encourage the random forest to learn what characterizes the four distinct outcomes. A consequence of this approach is that the disproportionally low count of large hail cases at any

given time suggests the sampling size of the other three hail sizes (< 1 mm, 1–5 mm, 5–20 mm) would be limited by the number of large hail training examples.

In this study, we developed the following input data sampling procedure to ensure a 1:1:1:1 ratio of training examples across four hail-size classes while at the same time sampling as many days, times, and areas as possible. We sought to sample every large hail pixels and systematically one pixel every so many of the other hail categories such that we get samples from every input radar map and region (the training for "no hail" followed a somewhat different approach described later). Using a subset composed of 13 days of historical MESH data drawn between May and June 1999, we estimated that the ratio of grid point observations with the three largest MESH categories (≥ 20 mm, 5–20 mm, 1–5 mm) was approximately 1:10.7:21.3, recognizing that there is some uncertainty to that ratio. We then sought to prepare a dataset containing all the large hail cases and subsets of the other two outcomes. To build this dataset in a computationally efficient manner while not knowing a priori how many pixels would end up in all four categories, a two-step process was followed. First, a temporary dataset was built containing all the large hail events, a subset of $1/5^{th}$ of the 5–20 mm category, and a subset of $1/10^{th}$ of the 1–5 mm category. Note that the resulting datasets for these two categories is larger than what would be expected given the previously mentioned estimated ratio of 1:10.7:21.3. This was done so that whatever number of samples we end up with in the large hail category, we have at least as many to choose from in the other categories while saving CPU by not extracting every sample. This was exactly what we implemented while generating the training set. In a second phase, a further round of fine-tuned random sampling was carried out to obtain an absolute equal number of training examples across the three hail-size labels.

The sampling approach for the fourth label class (no-hail or MESH < 1 mm) is different from the rest due to its unique nature. No hail observations exist in both stormy and non-stormy environments such as sunny days, while encounters of MESH > 1 mm are confined to stormy environments. To build an unbiased all-weather-capable hail prediction system, our hail nowcasting model should be trained with "no hail" examples from both stormy and non-stormy situations on the same scale. Therefore, we sought to sample an equal amount of MESH < 1 mm examples from stormy frames and non-stormy frames. To estimate how many examples to sample per frame, we relied on the results obtained from the subset experiment mentioned earlier. On average, 81 out of the 2,003,001 grid points (over a 1,001 × 2,001 gridded data

frame) have a MESH value indicative of large hail (≥ 20 mm). We then developed a custom threshold to differentiate stormy from non-stormy frames. A stormy scenario is defined to be a radar composite with at least 500 grid points with RALA ≥ 45 dBZ and at least 50 grid points with RALA ≥ 55 dBZ. Applying this definition, 66.13% of the frames in the experimental subset are considered stormy. The sampling procedure for MESH < 1 mm examples is thus the following:

1. Determine whether the gridded data frame under review is stormy or not.
2. If it is a stormy frame: using an approximate oversampling factor of 2.5, we randomly sample 2.5*(81/2)/0.6613 ≃ 150 eligible examples per frame.
3. If it is a non-stormy frame: using the same safety factor of 2.5, we should randomly sample 2.5*(81/2)/(1–0.6613) ≃ 300 eligible examples per frame.
4. Repeat the above steps for all frames in the set and accumulate the sampled examples into one list (MESH < 1 mm class label training set).
5. Conduct one last round of random sampling within the list to match its size with the limiting class of MESH ≥ 20 mm.

This sampling procedure was also applied when deriving the test set when the fully trained model was evaluated for prediction performance on unseen data. A summary of statistics on the training and test sets built for the final hail nowcasting model suite is presented in Table 2.4.

Input Data Set	RF Model Forecast Lead Time	# of Frames Processed	Average # of Examples per Class per Frame*	Total Sample Size (from four classes)
Experimental Subset	N/A	1,872	81.26	4 × 152,127
Training Set	15 mins	27,072	90.59	4 × 2,452,483
Test Set	15 mins	6,768	94.82	4 × 641,756
Training Set	60 mins	25,344	88.31	4 × 2,238,194
Test Set	60 mins	6,336	90.67	4 × 594,266

Table 2.4. Summary statistics on the hail nowcasting model input data sets. Note that the average number of examples per class per frame (*) is identical to (or limited by) the average number of MESH ≥ 20 mm data points per frame for the corresponding data set.

From Table 2.4, we can notice that we had underestimated the case limiting factor – the average number of MESH ≥ 20 mm data points per frame when constructing the experimental subset. Nonetheless, our estimated large hail case sample size at 81.26 was no more than 15% below the actual average size obtained from any model input training set or test set. With the chosen safety factor of 2, none of the large hail examples encountered were discarded in the sampling process to obtain a class-balanced input data set. As expected, the training sets are approximately four times larger than their respective test sets because of the 80/20 split applied in this study. The number of data frames processed by the 60-min forecast model was slightly lower than that for the 15-min model since frames spanning the last 55 minutes of each day are obsolete for a 60-min forecast but still relevant for a 15-minute forecast.

2.2.5 Model Verification Measures

For the classification predictor, the prediction results of binary events can be classified into 2 × 2 contingency tables that arrange observed and forecasted cases into 4 categories: correct rejections (true negative, TN), false alarms (false positive, TN), misses (false negative, FN), and hits (true positive, TP) as presented in Figure 2.5. Based on the confusion matrix, seven evaluation indices are calculated to assess the model's performance, as presented in Table 2.4.

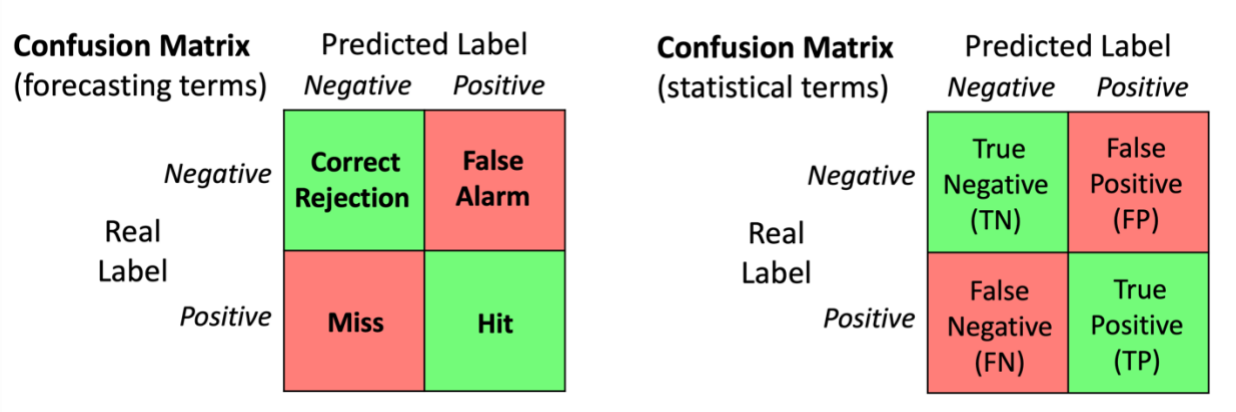


Figure 2.5. Schematic of a sample confusion matrix (2 × 2 contingency table) with interchangeable cell elements in forecasting terms (left) or statistical terms (right).

Besides the fundamentals such as Proportion Correct (PC), Probability of Detection (POD), False Alarm Ratio (FAR), and False Alarm Rate (F), it was decided to follow the study by Czernecki et al. (2019) to calculate measures that are more suitable for rare events. The Critical

Success Index (CSI) assesses hits relative to all positive signals in a forecast of observations. It emphasizes accurate hits while penalizing both misses and false alarms. A minor enhancement to CSI is the Equitable Threat Score (ETS), which considers an adjustment for the number of hits that could be achieved by random chance (Jolliffe and Stephenson, 2012). The final selected metric is the Heidke Skill Score (HSS), which is akin to PC but benchmarks the forecast performance that could be achieved by a random model (Wilks, 2008). A positive HSS may be interpreted as a model having some added skill in comparison to a random model.

Score *(statistical term)*	**Formula**	**Range**	**Optimum**
Proportion Correct = PC *(Accuracy)*	$\frac{TP + TN}{TP + FP + FN + TN}$	[0, 1]	1
Probability of Detection = POD *(Recall)*	$\frac{TP}{TP + FN}$	[0, 1]	1
False Alarm Ratio = FAR *(1 - Precision)*	$\frac{FP}{TP + FP}$	[0, 1]	0
False Alarm Rate = F	$\frac{FP}{FP + TN}$	[0, 1]	0
Critical Success Index = CSI	$\frac{TP}{TP + FP + FN}$	[0, 1]	1
Equitable Threat Score = ETS	$\frac{TP - TP_r}{TP + FP + FN - TP_r}$ where $TP_r = \frac{(TP+FP)(TP+FN)}{TP+FP+FN+TN}$	[-1/3, 1]	1
Heidke Skill Score = HSS	$\frac{TP + TN - TP_r - TN_r}{TP + FP + FN + TN - TP_r - TN_r}$ where $TN_r = \frac{(FP+TN)(FN+TN)}{TP+FP+FN+TN}$	[-1, 1]	1

Table 2.5. Statistical and forecast verification scores computed to evaluate hail nowcasts.

Evaluation and analysis of the verification measures for all RF models developed across the three stages of the experiment are presented in the Results and Discussion sections of the thesis.

3. Results and Discussion

3.1 Stage 1 (Radar-only) Results and Discussion

During the first two stages of the hail nowcasting experiment, we conducted a storm case study exercise based on meteorological data obtained for June 3, 1999. On this day, there were a total of 49 hail reports within the selected domain of study, meeting the minimum (10) hail reports criteria to be considered a severe hail day (see Figure 3.1).

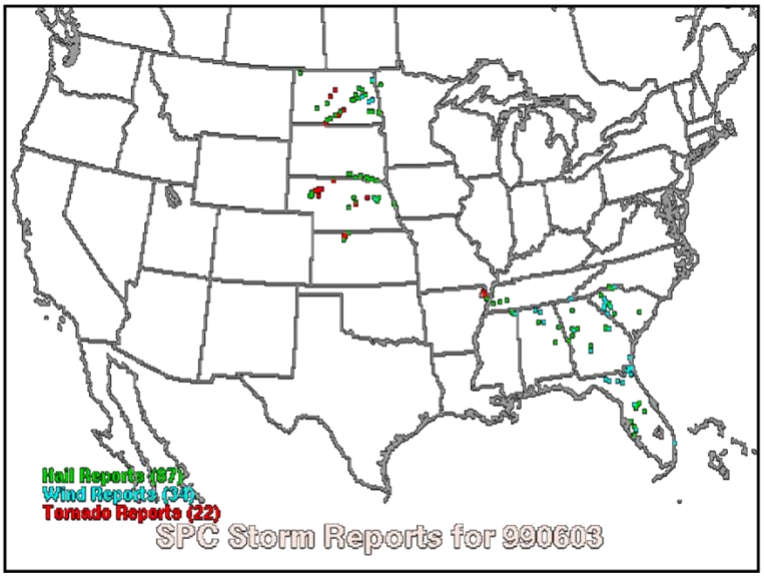


Figure 3.1. Storm Reports Map from the Storm Prediction Center for CONUS on June 3, 1999 (the day chosen for the storm case analysis in Stage 1 and 2). Green-colored dots represent verified hail reports on this day. Image adapted from the U.S. NWS (1999).

During Stage 1, we took 18 parameters derived from the MYRORSS remote sensing dataset as predictors (i.e., features) to build an ensemble of nine binary deterministic random forest classifiers for hail nowcasting: Each RF classifier specializes in predicting hail occurrence over three hailstone size thresholds D (> 1, > 5, > 20 mm) and three lead times T (15, 30, and 60 mins), making up the 3×3 or 9 classifiers. The feature importance scores that indicate how much each feature contributes to the model's predictions were analyzed (see Figure 3.2).

The position of a feature in a tree as a decision node can indicate its importance in predicting the target variable. Features at the top influence a higher fraction of input samples, making their

expected contribution a measure of their importance. In Scikit-Learn (2023), the portion of samples a feature impacts and the impurity reduction from splitting are combined for a normalized prediction power measure known as the mean decrease in impurity (MDI). Averaging this across randomized trees reduces variance and supports feature selection (Louppe, 2014).

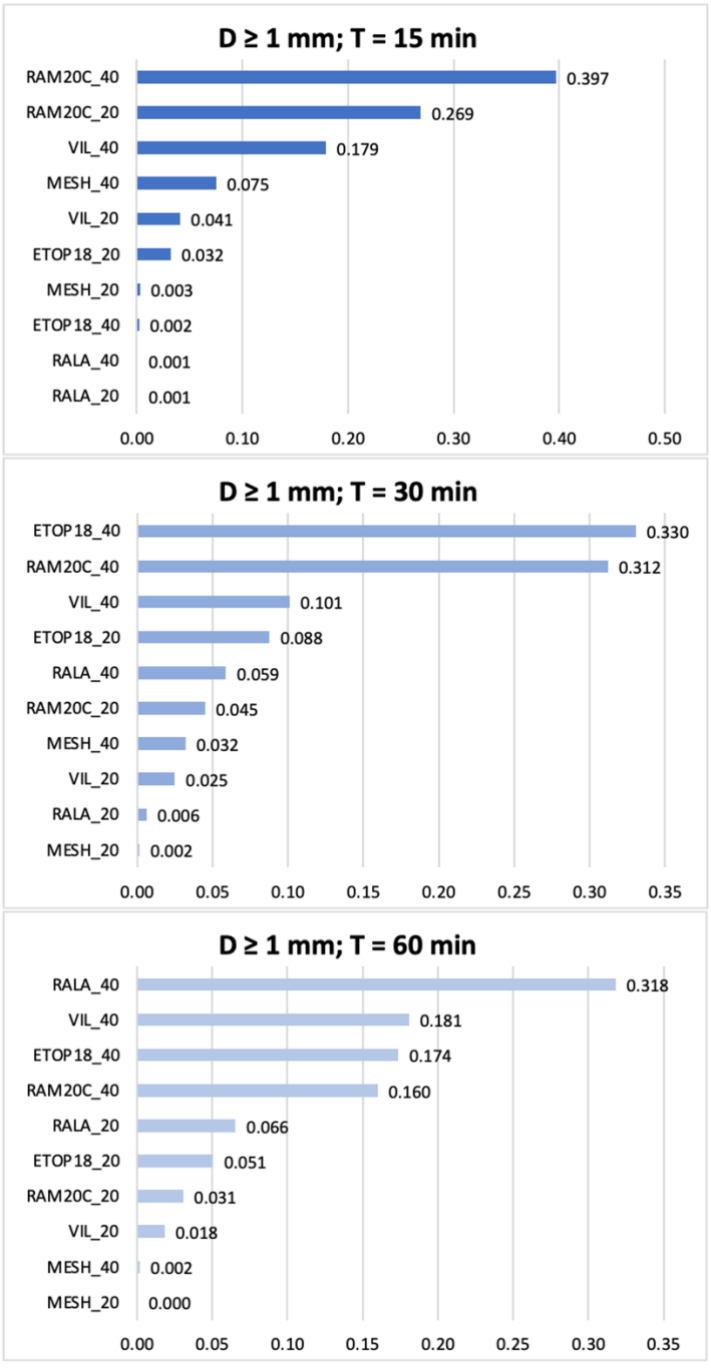


Figure 3.2. (On previous page) Top 10 hail predictors ranked by their feature importance based on the mean decrease in impurity. Computed for a total of nine independent RF models (three hail-size thresholds D and three forecast lead times T) used during Stage 1 (radar data only) of the experiment. Only the three sub-models for the $D \geq 1$ mm hail-size threshold were shown here, with a decrease in color shade for an increase in T.

Hail predictors are crucial to RF model construction as they directly impact prediction accuracy. These predictors can also unveil potential physical laws governing hail event occurrence and development. Note that we have not included results for the $D \geq 5$ mm and $D \geq 20$ mm hail prediction models as the sample size for these hail cases were uneven and insufficient (less than 10,000) to be worthy for analysis.

Consider the three RF sub-models with $D \geq 1$mm in Stage 1 (Figure 3.2), the top 5 most important remote sensing factors overall in hail prediction are Reflectivity at –20°C (RAM20C), Reflectivity at the lowest altitude (RALA), Echo top heights for 18 dBZ (ETOP18), Vertically integrated liquid (VIL), and Maximum estimated size of hail (MESH). RAM20C is the only factor with an average score above 0.4 (at 0.405). The fact that RAM20C significantly outperformed other reflectivity factors confirmed conventional wisdom that sufficient supercooled liquid water content at the –20°C temperature level is the most significant ingredient to hail formation and growth.

With an average score of 0.226, RALA came at the second place. Its magnitude signifies the intensity of precipitation closer to the ground level than that of RAM20C. However, precipitation condition at the lower altitude played a less significant role in predicting hail occurrence. The next two factors (ETOP18 and VIL) have importance scores (0.150, 0.115) in the range of 0.10 to 0.15, suggesting that the vertical amount and extent of precipitation in the storm clouds have relatively equal contributions to hail nowcasting. An a priori surprising result is that MESH, the parameter that is also what the RF models aim to predict, only came at fifth place (0.04) in the ranking. It could be possible that considering MESH at the current time alone is not the best strategy in telling MESH at some later time due to the short duration and rapidly evolving nature of hailfall activities. Knowing MESH at present is still valuable when forecasting MESH, just not with the same degree of relevancy compared to other reflectivity factors.

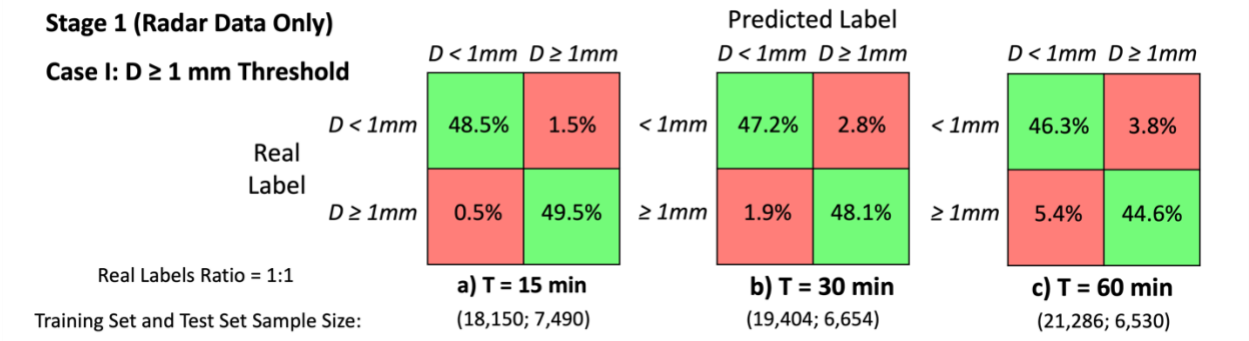
For every RF sub-model built in Stage 1, the azimuthal shear factors and their quality indices obtained the lowest scores (no higher than 0.01). Azimuthal shear was initially selected as a predictor because of its effectiveness in identifying mesocyclones and tornadoes, features associated with severe thunderstorms also capable of producing hail. Here we offer three possible explanations as to why azimuth shear came at the bottom of the feature importance ranking. First, thunderstorms with mesocyclones and tornadoes are often associated with hail, but hail-producing thunderstorms do not guarantee mesocyclones and tornadoes. In other words, azimuthal shear may be a sufficient condition for some hail but not a necessary condition for all hail. Second, it could be that azimuthal shear was exceptionally not relevant for this particular storm event, or that its value cannot be properly evaluated with the limited number of samples. Third, the azimuthal shear used here is an averaged value of all grid cells within a 20 km × 20 km (or 40 km × 40 km) domain. Since mesocyclones and tornadoes have much smaller dimensions compared to the domain size, only a few grid cells in the domain can capture a non-zero azimuthal shear value (usually on the order of 0.0001 s^{-1}). This tiny distinction was further diluted when the average is computed over a large domain, rendering the azimuthal shear factor largely meaningless. This may call for a grid-wise maximum shear to be used instead of an average value. Despite attempts made to capture the scale of the rotation using some quality index, none of the azimuthal shear factors made it into the Top 10 list in feature importance. Subsequently, all eight azimuthal shear parameters were removed from the list of hail predictors designed for Stage 3 of the experiment. Further investigation might be made on how to better express azimuthal shear as a hail predictor candidate after this study.

One advantage in the design of the Stage 1 (and 2) experiment is that one can compare the feature importance tables across RF models with different lead times (T). We focus on the results from three RF sub-models in Figure 3.2. RAM20C outperformed all other parameters for T = 15 mins. But, as the lead time increased, it was overtaken by parameters VIL, ETOP18, and RALA. RALA became the top factor at T = 60 mins. While a very high RAM20C value may suggest active hail production aloft in proximity that may generate hailfall in the next half an hour, it is less clear why the present RALA value upstream can become the most important factor to predict hailfall in an hour. It could be that when signs of ultra-high reflectivity (capable of hail) near the ground are observed upstream, the likelihood of getting hail downstream in the next

hour significantly increases. Another possibility is that other reflectivity factors simply became less relevant over time, leaving RALA as the top hail predictor.

Two additional findings can be stated. First, as the forecast lead time increases, the most important hail predictors gradually shift from a mix of 20 km × 20 km and 40 km × 40 km grid averages to being dominated by 40 km × 40 km grid averages. As the 40 km × 40 km grid covered an area four times larger, it gained importance in the long run as it encompassed more developing storm cells upstream that may impact the location of interest while deemphasizing smaller-scale patterns with shorter times of predictability. Second, as T increases for the RF model, the weight of the most important predictor decreases, and the gap in the scores among the chosen factors narrows. This suggests there is probably less dominance by any remote sensing parameter alone when the RF model forecasts large hail in the long run. It also suggests that it may become increasingly difficult to rely on a limited set of remote sensing signatures to make skillful hail forecasts as lead times increase.

The confusion matrices for the three RF sub-models under the $D \geq 1$mm prediction case were made (Figure 3.3) to visualize the proportional distribution of hail classification outcomes (sum to 1), while the data table below summarized the seven statistical scores for model performance evaluation.



Sub-model Forecast Lead Time	T = 15 min	T = 30 min	T = 60 min
Proportion Correct (PC)	98.0%	95.3%	90.8%
Probability of Detection (POD)	98.9%	96.2%	89.1%
False Alarm Ratio (FAR)	2.9%	5.5%	7.8%
False Alarm Rate (F)	3.0%	5.6%	7.5%

Figure 3.3. Confusion matrices and the model evaluation metrics (performance scores) for the three RF sub-models with $D \geq 1$ mm hail-size threshold at Stage 1 (radar data only) of the hail nowcasting experiment.

For all sub-models in the $D \geq 1$ mm case, the quality in the performance of all scores computed decreases with increasing lead time. This makes sense as it is more challenging to forecast hail risks in the distant future due to the growing uncertainty associated with the meteorological conditions. Although predictions of hail vs. no hail for the next 0 to 1 hr have high PC scores of above 90% and low FAR scores below 8%, predictions of large hail (results not shown) have an average accuracy of 80% and an average false alarm ratio of 24%. We suspect that the relatively poor performance and high discrepancies in scores of the RF models for large hail prediction were due to the lack of large hail observations resulting from a much smaller sample size used. Having insufficient data for the training set would lead to poor generalization of unseen data, which results in inferior performance by the RF model. To tackle this problem, the hail data sampling technique was modified in Stage 3 to ensure a more balanced distribution of hail observations across different sizes.

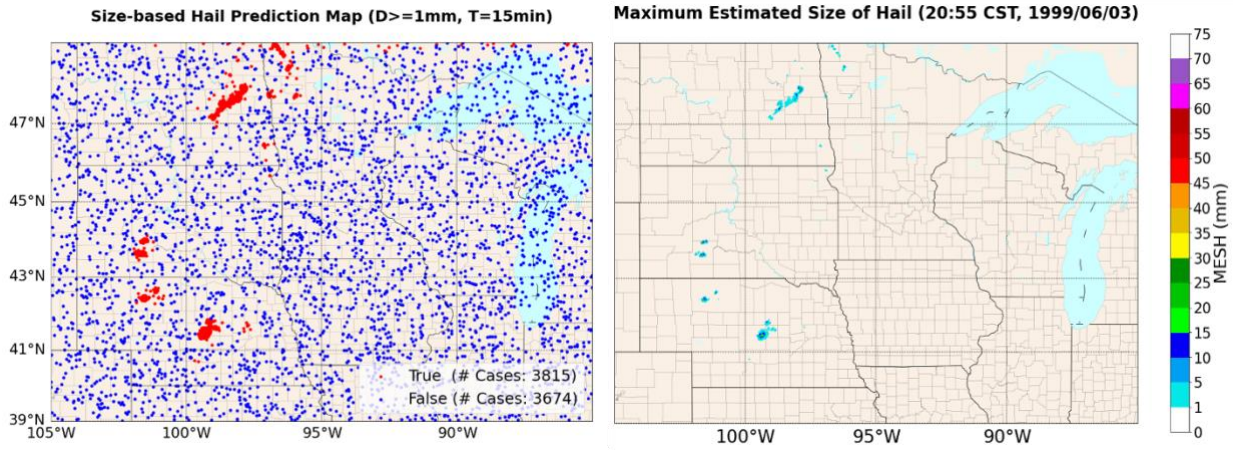


Figure 3.4. Left subplot: Deterministic hail-size prediction map generated by the RF sub-model with a hail-size threshold of 1 mm and forecast lead time of 15 minutes in the storm case analysis exercise (Stage 1, radar-only predictors). Right subplot: MESH color map at the time of verification for the RF sub-models with a forecast lead time of 15 minutes in the storm case analysis exercise (Stage 1 and 2).

To visualize the achievement of the hail nowcasting exercise in Stage 1, we plotted the deterministic hail occurrence predictions made by one of the sub-models ($D \geq 1$ mm, $T = 15$ mins) on a map (see Figure 3.4, left). It can be compared with the map of MESH at the time of forecast verification (see Figure 3.4, right) to examine the performance of the hail prediction

model. The positive hail-size claims (red dots in Figure 3.4, left) generally matched regions with MESH ≥ 1 mm (colored pixels in Figure 3.4, right), indicating this model can classify hail locations from no-hail locations in this case study with good accuracy. Similar prediction maps can be made for probabilistic forecasts with hail-size categories represented in color contours.

3.2 Stage 2 (Reanalysis-only) Results and Discussion

Stage 2 of the experiment resembles the design scheme of Stage 1, except for replacing the 18 radar-derived parameters with 12 reanalysis parameters as hail predictors. The feature importance, confusion matrix, and performance scores were all computed again for each RF sub-model. The identical experiment setup means we can cross-examine the results from Stages 1 and 2. Findings from these two parts were then used to improve the design of the Stage 3 experiment, with new RF models trained and tested on a dataset of a much larger scale. Again, only results for the three RF sub-models with $D \geq 1$ mm hail-size prediction threshold were shown due to the unsatisfactory data sampling outcome in other cases.

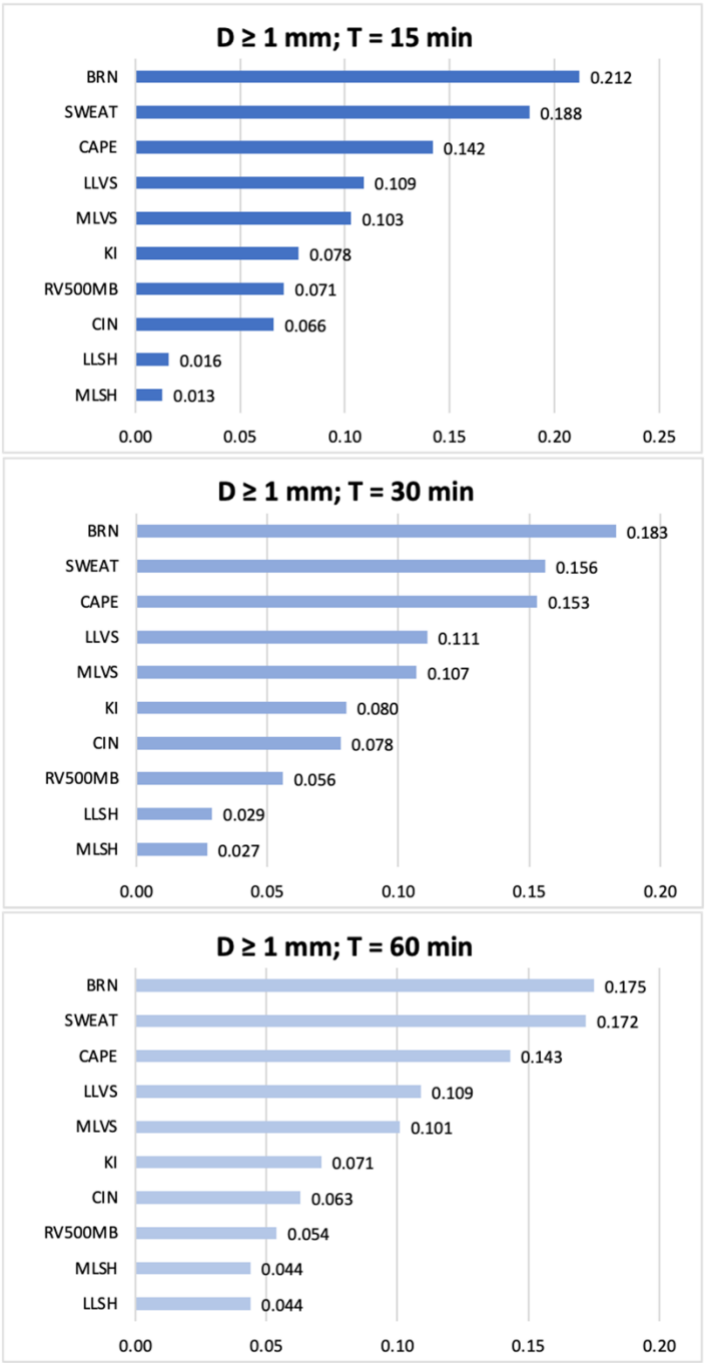


Figure 3.5. Top 10 hail predictors ranked by their feature importance based on the mean decrease in impurity. Computed for a total of nine independent RF models (three hail-size thresholds D and three forecast lead times T) used during Stage 2 (reanalysis data only) of the experiment. Only the three sub-models for the $D \geq 1$ mm hail-size threshold were shown here, with a decrease in color shade for an increase in T.

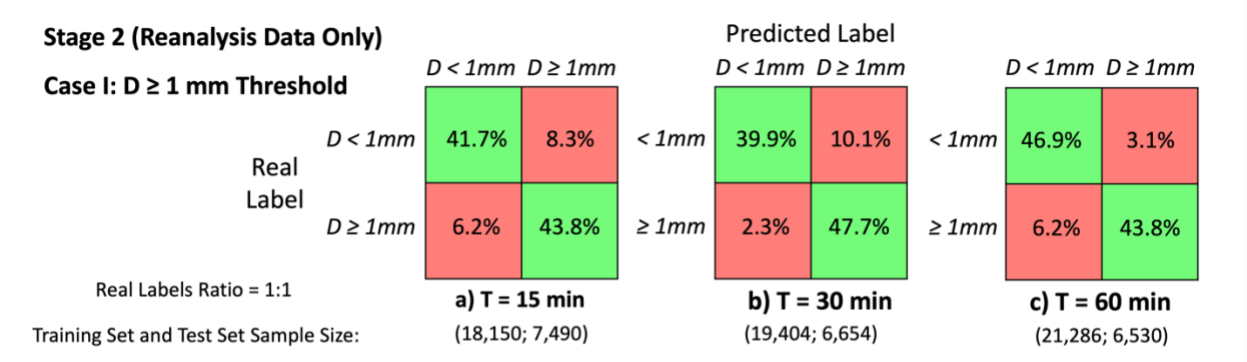
The first observation when examining the feature importance tables of reanalysis parameters (Figure 3.5) is that composite indices such as BRN and SWEAT usually rank higher than single-element factors such as vertical shear and specific humidity. This is within expectations as composite indices were designed to combine information from multiple aspects, which can lead to a more comprehensive representation of the underlying physics in the data. As an example, BRN incorporates CAPE with vertical shear while SWEAT considers temperatures, dewpoints, and shear at different pressure levels of the atmosphere. As the ranking suggests, both BRN and SWEAT outperformed their constituent environmental elements. The third composite index, KI, has a lower rank (usually 4th to 6th place) but still outperformed some single-element factors such as specific humidity (LLSH, MLSH) and relative vorticity (RV500MB).

For hail nowcasting models with a size threshold of $D \geq 1$ mm, BRN has a slight lead over SWEAT as the most important predictor. For predictions of large hail ($D \geq 20$ mm, not shown) risks, however, SWEAT surpasses BRN at all lead times. As introduced in Section 2.1.4, SWEAT is an index developed to evaluate the potential of severe thunderstorms (not ordinary thunderstorms), while BRN is usually an indicator of convective storm types. It is not surprising that SWEAT scored higher than BRN in large hail prediction models as most large hail observed were products of severe thunderstorms.

Compared to the feature importance tables in Stage 1, the variation in rankings across lead times and hail sizes was not as significant, with adjustments of no more than two places for any factor. In other words, there exist more commonalities among reanalysis-based RF sub-models. While the top-performing radar reflectivity parameters can earn a score as high as 0.4, the highest score for reanalysis parameters was just above 0.2. Additionally, the mean decrease in impurity for the reanalysis factors has a lower variance. Thus, the reanalysis parameters selected to build the RF models contribute more equally (in weights) to hail prediction.

Surface solar irradiance (SSI) and Zero-degree level (ZDL) were the only two factors that repeatedly scored the lowest among the 12 reanalysis parameters. This does not mean that SSI and ZDL are not relevant to hail potential, they were just much less influential in predicting hail when compared with the remaining factors in the RF model. If we could compute the wet-bulb-zero level (WB0), a proven hail forecasting parameter (Morgan, 1970), it is almost certain to outperform ZDL. As for SSI, the amount of solar radiation earlier in the day may be indicative of

thunderstorm potential. But when considered for hail nowcasting in the next 0 to 1 hr, it is no longer as effective. Due to their poor feature importance scores, SSI and ZDL were later removed from the list of predictors used in the final stage of the study.



Sub-model Forecast Lead Time	T = 15 min	T = 30 min	T = 60 min
Proportion Correct (PC)	85.6%	87.6%	90.7%
Probability of Detection (POD)	87.7%	95.3%	87.5%
False Alarm Ratio (FAR)	15.9%	17.4%	6.5%
False Alarm Rate (F)	16.5%	20.1%	6.1%

Figure 3.6. Confusion matrices and the model evaluation metrics (performance scores) for the three RF sub-models with $D \geq 1$ mm hail-size threshold at Stage 2 (reanalysis data only) of the hail nowcasting experiment.

Compared to the hail nowcasting model using radar-only hail predictors (Stage 1), the performance of the reanalysis-based hail predictions was less satisfactory, with PC scores in the range of 85% to 90%. There were a higher proportion of no-hail situations misclassified as positive hail cases, vice versa. The discrepancies in FAR and F scores across sub-models were also higher in Stage 2, with the least error rates for the longest forecast lead time. The lower scores of the statistical metrics obtained in Stage 2 suggest that designing a hail nowcasting RF model with the selected reanalysis predictors alone is no better than replacing them with the selected remote sensing predictors. But this is not surprising as numerical predictors focus on storm and hail potential while the radar predictors rely on related observations such as strong echoes at the surface or aloft.

3.3 Stage 3 (Combined sources) Results and Discussion

A significant portion of the added value of Stage 3 to the study was its data fusion approach compared to earlier predictions obtained using only one data source. A total of 20 input variables were chosen from the earlier stages to form a comprehensive set of hail predictors consisting of both remote sensing and reanalysis predictors. The latest nowcasting model consists of two multi-class RF classifiers that make probabilistic predictions of hail across four size classes at two specific lead times (15 mins, 60 mins).

We first examine the results of the model hyperparameter tuning. Table 3.1 listed the near-optimal values of the 10 hyperparameters investigated in this study for the two independent hail nowcasting models constructed. These values were considered "near-optimal" since not all possible parameter settings (3^{10}) were tried out, but rather a fixed number (k × n_iterations = 5 × 1000) was sampled from the search space during cross-validation.

Hyperparameters	**Tuning Candidates**	**15-min Model**	**60-min Model**
n_estimators (N_{tree})	100, 300, 400	300	300
max_samples	0.10, 0.15, 0.20	0.20	0.15
max_depth	10, 15, 20	15	15
criterion	"gini", "entropy", "log_loss"	"gini"	"log_loss"
max_features (M_{try})	5, 6, 7	5	6
min_samples_split	0.10, 0.15, 0.20	0.15	0.15
min_samples_leaf	0.05, 0.075, 0.10	0.05	0.075
max_leaf_nodes	20, 25, 30	25	20
min_impurity_decrease	0.01, 0.025, 0.05	0.05	0.05
ccp_alpha	0.01, 0.015, 0.02	0.01	0.02

Table 3.1. Near-optimal values of each hyperparameter, determined by the RF algorithm during model validation via a randomized search in the candidate pool of hyperparameter settings.

In general, several mismatches exist in the hyperparameter configuration between the two models. A given hyperparameter is more likely to be assigned the middle value of its three tuning candidates as part of the overall "near-optimal" setting, except for min_impurity_decrease and ccp_alpha during tuning (while not applicable to criterion). Both models preferred the same tuning candidate for n_estimators, max_depth, min_samples_split, and min_impurity_decrease, although we do not know whether these agreements were by coincidence during sampling. It is also worth mentioning that both models used the same 30 hail predictors (input features) and favoured a random forest with 300 trees, which agreed with the proposal by Boehmke and Greenwell (2019) to start with 10 times the number of features as the number of trees. Despite making attempts, our over-simplified approach to hyperparameter tuning in this study means we are still much less confident about the actual optimal hyperparameter configuration for the RF classifiers. Further emphasis should be placed on hyperparameter optimization to improve the ML model performance.

A quick examination of the new feature importance ranking for both sub-models (Figure 3.7) revealed a balanced mix of parameters from two sources in the Top 10, with small leads by remote sensing parameters. In the T = 15 mins sub-model, 20 km × 20 km grid averages consistently ranked higher than their 40 km × 40 km counterparts, but the opposite is true for the T = 60 mins sub-model. This remarkable contrast confirmed that as the forecast lead time increases, the RF model favours predictors computed using data collected over a larger spatiotemporal domain.

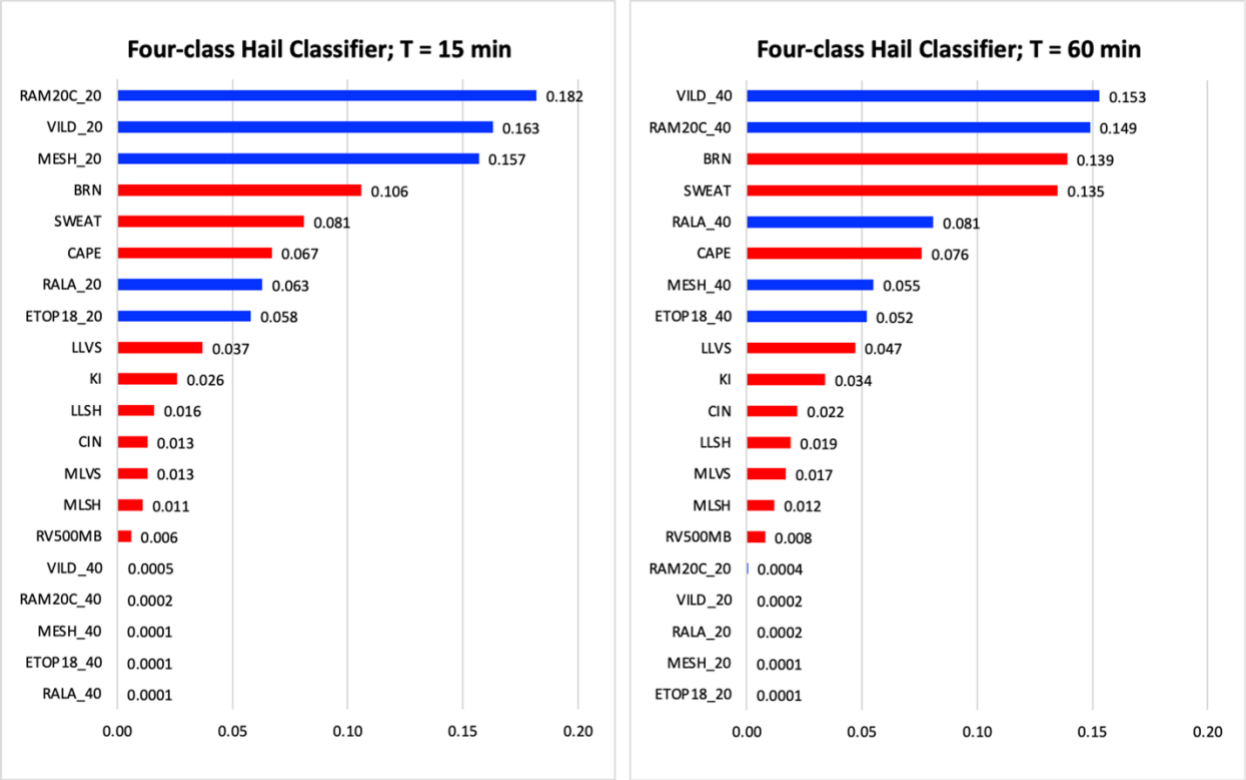
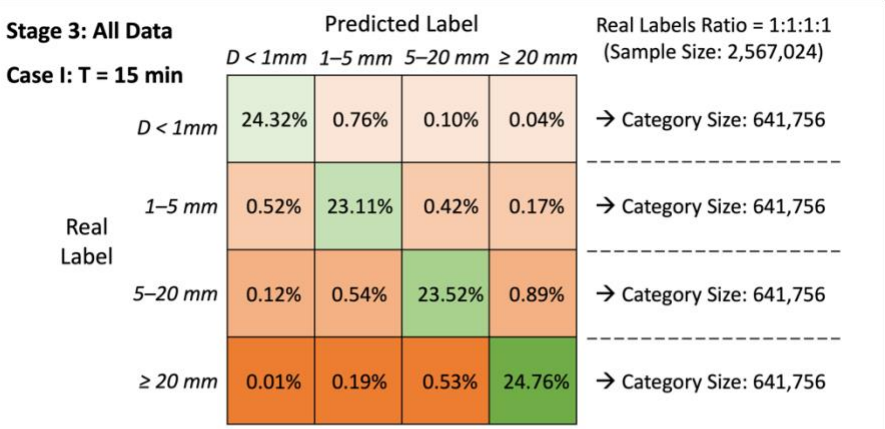


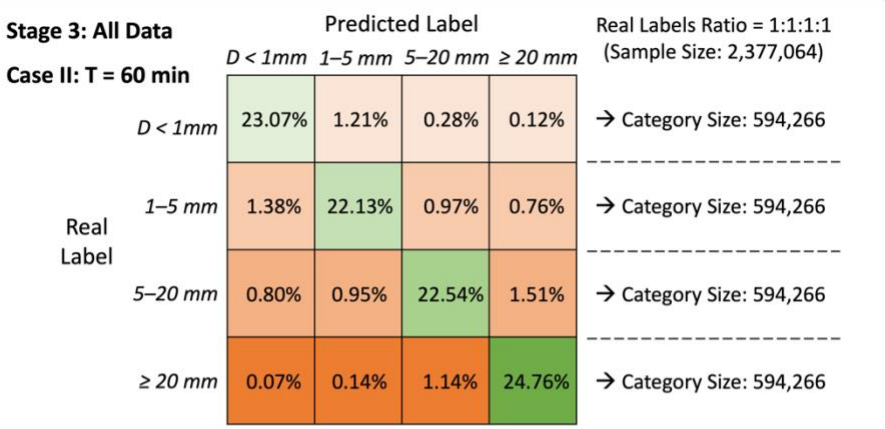
Figure 3.7. All 20 hail predictors ranked by their feature importance based on the mean decrease in impurity. Computed for the two independent four-category hail classifier RF models (left chart for a lead time of 15 mins, right chart for a lead time of 60 mins) used during Stage 3 (remote sensing and reanalysis data combined) of the hail nowcasting experiment. Blue (red) color bars for radar-derived (reanalysis-based) predictors.

There exist little differences in the ranking of reanalysis parameters between the two RF models, with the top three scoring factors standing solid as BRN, SWEAT, and CAPE. The main relative change is that SWEAT, our best indicator of storm severity, gains importance with increasing lead times as remote-sensed predictors of past storm severity lose value. For the radar-based parameters, the most relevant predictors ranked from top to bottom for T = 15 mins were RAM20C, VILD, MESH, RALA, and ETOP18. As for the T = 60 mins model, the order became VILD, RAM20C, RALA, MESH, and ETOP18. An increase in forecast time saw higher rankings in the relative importance of VILD and RALA but a significant decrease in the significance of MESH. These pattern changes also resemble the trends identified in Stage 1.



Label Category	D < 1 mm	1 – 5 mm	5 – 20 mm	≥ 20 mm
Proportion Correct (PC)	98.45%	97.40%	97.40%	96.56%
Probability of Detection (POD)	96.43%	95.42%	93.82%	97.14%
False Alarm Ratio (FAR)	3.57%	4.58%	6.18%	2.86%
False Alarm Rate (F)	0.90%	2.04%	1.45%	1.52%
Critical Success Index (CSI)	94.01%	89.89%	90.05%	93.12%
Equitable Threat Score (ETS)	92.08%	86.84%	86.97%	87.11%
Heidke Skill Score (HSS)	95.88%	92.95%	93.03%	93.11%
Overall Accuracy (PC_{avg})	95.71%			

Figure 3.8. Confusion matrix as 4 × 4 contingency tables and the model evaluation metrics (performance scores) for the multi-class (four hail-size categories) RF model with a forecast lead time of 15 minutes at Stage 3 of the hail nowcasting experiment.



Label Category	D < 1 mm	1 – 5 mm	5 – 20 mm	≥ 20 mm
Proportion Correct (PC)	96.14%	94.59%	94.35%	92.99%
Probability of Detection (POD)	93.48%	87.68%	87.36%	94.44%
False Alarm Ratio (FAR)	6.52%	12.32%	12.64%	5.56%
False Alarm Rate (F)	3.06%	3.12%	3.24%	3.24%
Critical Success Index (CSI)	85.67%	80.36%	79.96%	85.98%
Equitable Threat Score (ETS)	81.34%	74.69%	74.03%	75.30%
Heidke Skill Score (HSS)	89.71%	85.51%	85.08%	85.91%
Overall Accuracy (PC_{avg})	90.67%			

Figure 3.9. Confusion matrix as 4 × 4 contingency tables and the model evaluation metrics (performance scores) for the multi-class (four hail-size categories) RF model with a forecast lead time of 60 minutes at Stage 3 of the hail nowcasting experiment.

Because of the switch of the RF models from a binary classifier to a four-class classifier in Stage 3, the confusion matrix went from a 2 × 2 contingency table to a 4 × 4 layout (see Figure 3.8 and Figure 3.9). The weights of individual cell elements still add up to 1, but there were four versions of statistical scores, one for each category of hail size a sample can be classified into. The good news is that we can still compare the performance scores for predictions across hail sizes and lead times. Both RF models achieved an overall accuracy (PC) score above 90%, with the shorter lead time version having an even better score above 95%. These were significant improvements to the previous generation of RF models.

We also witnessed better performances across statistical scores. However, there exists a noticeable gap across hail-size categories. Hail size predictions with $D < 1$ mm (i.e., no hail) or $D \geq 20$ mm (i.e., large hail) always outperform the predictions with D in the range of 1–5 mm (i.e., too small to be hail) or 5–20 mm (i.e., small hail) by 1–5%. Since we sampled the four

cases of hail observations with an equal distribution over a much larger dataset, the cause of these discrepancies should be rooted elsewhere. We suspect that forecasting hail sizes in the range of 1–5 mm and 5–20 mm is intrinsically more challenging than predicting scenarios with no hail or large hail occurrences, due to the positioning of these two categories between the two extreme cases and their narrow size range. The current RF classifier setup should be retained before a better design becomes available.

Regarding the three forecast verification metrics used, each have its own distinct purpose. The CSI aims to evaluate the number of hits against all positive signals (hits, false alarms, and misses), and therefore reflects the conditional probability of hits but without considering the correct rejections. Since CSI is sensitive to hits, and penalises both misses and false alarms, CSI values are usually low for rare events like hail. Indeed, the CSI scores in both RF sub-models were lower than PC and POD. The ETS made a correct to CSI for the number of hits that would be obtained by chance. As expected, the ETS scores were thus lower than CSI scores for all hail-size categories, with the amount of decrease higher for the 60 minutes forecast model. Lastly, the all-positive HSS values can be interpreted as our RF models have some added skill in comparison to a random model. In general, CSI had the best results while ETS had the worst performance, with no metric receiving a score lower than 74%.

3.4 Limitations and Future Research

There are several limitations with the current approach to hail nowcasting with random forests. Some notable ones include the following:

- The 20-km (and 40-km) grid averaging technique lacks flexibility and are not optimal for a range of forecast lead times. Some past studies used storm tracking algorithms to detect the current location of the storm cells.
- The 20 parameters selected as hail predictors for the RF models were not the best choices. Some better candidate parameters (such as wet-bulb zero height, most unstable CAPE, and storm-relative helicity) were not included due to data unavailability or difficulties in computing them.
- The current RF models do not provide a specific forecast of the MESH, only a probabilistic prediction on the maximum hail size range to be expected. This could,

however, be changed easily by making a regression-based RF instead of a categorical one as in this work.

- The initial screening and further tuning of model hyperparameters procedures were simplified without visiting the majority of possible parameter combinations. This method saves computation time significantly, but at the expense of not obtaining the most optimal set of hyperparameters to construct our RF classifiers.

Compared to previous studies in applying RF to hail forecasting, this study presents several new insights. We have confirmed the finding by Czernecki et al. (2019) that radar reflectivity-based factors tend to rank higher in feature importance than environmental parameters in hail prediction. The 0–6 hr hail nowcasting model developed by Yao et al. (2020) had reanalysis-based predictors CAPE and K Index both in the Top 10 list. We introduced indices such as BRN and SWEAT not used in their research but verified that CAPE and K Index were indeed more important than many other candidates to forecast hail potential. In addition, the RF model suite developed in this study scored better in all performance metrics evaluated except POD.

To improve the hail nowcasting RF model developed in this study, there are three aspects to focus on better data, better algorithm, and better interpretation. First, return to the process of selecting hail predictors. We can further optimize the existing list of parameters by examining data sources and hail-relevant factors not considered before. Second, we can experiment with more sophisticated random forest algorithms such as AdaBoost (Hastie et al., 2009) and Gradient Tree Boosting (Friedman, 2002) or consider other proven supervised learning methods in severe weather prediction. Moreover, additional visualization products such as forecast maps and performance charts can be developed using the output data from this study to facilitate in the understanding of our RF-based hail nowcasting model suite. Lastly, there are several strategies to interpret the results from a machine learning model that we have not implemented. Work can be done to derive more desirable metrics and graphical techniques to investigate the relationship between the selected hail predictors and hail nowcasting outcomes.

4. Summary and Conclusions

In this study, we developed an experimental probabilistic 0–1 hr hail nowcasting regional model through the employment of a multi-class random forest classifier framework. The RF prediction model takes input predictors as a screened list of atmospheric physical quantities and indices sourced from the MYRORSS remote sensing data and the ERA5 reanalysis environmental data. In the early stages of the experiment, two preliminary RF exercises were implemented separately using inputs from either the remote sensing or reanalysis source alone, with corresponding models trained and tested on a single-day storm event split in a 3:1 ratio into two time segments respectively. Based on their limited results obtained with certain assumptions, we refined and integrated hail predictors from both data sources, improved the hail case sampling technique, and modified the output hail predictand format. With these considerations, the ultimate RF model suite was trained on 192 days (from 1999 to 2002) and tested on 48 days (in 2003) out of the 240 days of severe convective weather with hail reports. The hail nowcasting model performed well on the test data, with the predicted potential maximum hail size in good agreement with the MESH product at the time of forecast verification.

The RF model suite was able to predict the maximum hail size range anticipated with a high overall accuracy of 95.7% for a 15-min lead time forecast and 90.7% for a 60-min lead time forecast. In general, the hail-size classifiers had better performance by a few percentage points in making correct classifications (PC) than detecting all positive cases (POD) for each hail-size-class bin. The number of false alarms relative to all positive claims made (FAR) was often less than half of those relative to all negative cases (F) for each bin. The RF model suite was most skillful at identifying large hail (with a diameter $\geq$ 20 mm) and situations without hail, but less satisfactory for graupel (with a diameter < 5 mm) and small hail (with a diameter between 5 mm and 20 mm) classifications. When comparing the three forecast verification scores (CSI, ETS, and HSS), both RF models performed the best in HSS and the worst in ETS. The positive HSS values suggested that our RF models have some added skill in comparison to a random model. As expected, ETS was lower than CSI for both classifiers since ETS accounts for a correction to CSI for the number of hits that would be obtained by chance.

There exist clear differences in the ranking of feature importance among the 20 selected hail predictors between the two RF models, despite a mere difference of 45 minutes in the relatively short nowcasting time horizon of 0–1 hr. The five most significant predictors of a 15-min lead

time forecast were Radar Reflectivity at –20°C, Vertically Integrated Liquid Density, Maximum Estimated Size of Hail, Bulk Richardson Number, and Severe Weather Threat Index. For a 60-min lead time forecast, they were Vertically Integrated Liquid Density, Reflectivity at –20°C, Bulk Richardson Number, Severe Weather Threat Index, and Reflectivity at the Lowest Altitude. Reflectivity-derived products and storm-relevant composite indices dominated the top of this list. It was a surprise that MESH, used as the hail size label for the training samples and as the truth for test samples, did not rank with the highest score. This finding counters the intuition that the current MESH holds the most value to estimate MESH later.

This study also demonstrated that for an RF model suite that combined the radar-derived remote sensing data and environmental reanalysis data to predict hail, radar reflectivity products have a lead in 0–1 hr hail nowcasting. Nonetheless, as the forecast lead time extended from 15-min to 60-min, the total feature importance score associated with radar parameters decreased from 0.62 to 0.49. In addition, hail predictors measured over a larger domain situated farther upstream in the storm track became more prioritized in hail nowcasting with a longer lead time.

The RF model suite has several shortcomings. First, by choosing only readily available or simple-to-compute parameters as model inputs, we ignored many factors potentially more effective in hail prediction. Second, the domain-averaging method to compute two variants of each remote sensing predictor is questionable. Introducing a storm-tracking algorithm to locate the relevant grid cells would be a better alternative. Third, the computationally intensive RF hyperparameters tuning process was simplified using a randomized search approach, which means that the combination of hyperparameters may still be optimized further for a better model performance. Lastly, the model was only trained and tested on severe convective weather days with hail reports, so its generalization performance on days without hail activities will be limited.

The RF model at the current stage of development is not suitable yet for operation forecasting since it still relies on the ERA5 reanalysis database as its environmental data source, which is not timely for nowcasting purposes. To adapt the existing system for operational use, the hail predictors must be derived from a near-real-time data source as the observed and forecasted parameters within an operational NWP ensemble model and a live remote sensing network.

In conclusion, the stated objectives of the thesis research have been accomplished with conditions and assumptions. The application of a random forest classifier to hail nowcasting in

the next 0 to 1 hr has shown to be satisfactory when evaluated with four basic statistical classification scores and three forecast verification scores. There is great potential for research into supervised learning techniques to forecasting severe convective weather hazards such as hail, using predictors sourced from remotely sensed data and NWP model products. The development of the next-generation severe thunderstorm warning system can benefit from lessons learned in this study.

Bibliography

Adams-Selin, R. D., & Ziegler, C. L. (2016). Forecasting hail using a one dimensional hail growth model within WRF. *Monthly Weather Review*, *144*(12), 4919–4939. https://doi.org/10.1175/MWR-D-16-0027.1

Allen, J. T., Giammanco, I. M., Kumjian, M. R., Punge, H. J., Zhang, Q., Groenemeijer, P., Kunz, M., & Ortega, K. L. (2020). Understanding Hail in the Earth System. *Reviews of Geophysics*, *58*(1). https://doi.org/10.1029/2019rg000665

Amburn, S. A., & Wolf, P. W. (1997). VIL density as a hail indicator, *Weather Forecast*, *12*(3), 473–478. https://doi.org/10.1175/1520-0434 (1997)012<0473:VDAAHI>2.0.CO;2

American Meteorological Society. (2012). *Hail - Glossary of Meteorology.* AMS. Retrieved from https://glossary.ametsoc.org/wiki/Hail

Battan, L. J., & Bohren, C. F. (1986). Attenuation of microwaves by spherical hail. *Journal of Applied Meteorology and Climatology*, *25*(8), 1155-1159.

Biryukov, S., Hewitt, E., Krzak, J., Marro, A., & Kubicek, A. (2021). A Deep Learning Based Hail Climatology for the Contiguous United States 1979-2018. *Understory Weather, Preprint.* https://understoryweather.com/papers/CONUS_Hail_Climate_1979-2018_PREPRINT.pdf

Bluestein, H. B., & Woodall, G. R. (1990). Doppler-Radar analysis of a Low-Precipitation severe storm. *Monthly Weather Review*, *118*(8), 1640–1665. https://doi.org/10.1175/1520-0493(1990)118

Boehmke, B., & Greenwell, B. (2019). *Hands-On Machine Learning with R.* Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9780367816377

Borland, S. W., Browning, K. A., Changnon, S. A., Cooper, W. A., Danielsen, E. F., Dennis, A. S., & Danielsen, E. F. (1977). Inherent difficulties in hail probability prediction. *Hail: a review of hail science and hail suppression*, 135-143.

Breiman, L. (1984). Classification and regression trees. *Biometrics*, *40*(3), 874. https://doi.org/10.2307/2530946

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification And Regression Trees. In *Routledge eBooks*. https://doi.org/10.1201/9781315139470

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

Browning, K. A. (1962). Cellular Structure of Convective Storms, *The Meteorological Magazine*, Vol. 91, No. 1085, 1962, pp. 341-350.

Browning, K. A. (1977). The structure and mechanisms of hailstorms. *Hail: A Review of Hail Science and Hail Suppression*, 1–47. https://doi.org/10.1007/978-1-935704-30-0_1

Byers, H. R. (1949). Structure and dynamics of the thunderstorm. *Weather*, *4*(8), 244–250. https://doi.org/10.1002/j.1477-8696.1949.tb01056.x

Changnon, S. A. (1977). The scales of hail. *Journal of Applied Meteorology and Climatology*, *16*(6), 626-648. https://doi.org/10.1175/1520-0450(1977)016<0626:TSOH>2.0.CO;2

Changnon, S. A. (1999). Data and approaches for determining hail risk in the contiguous United States. *Journal of Applied Meteorology and Climatology*, *38*(12), 1730-1739. https://doi.org/10.1175/1520-0450(1999)038<1730:DAAFDH>2.0.CO;2

Changnon, S. A., & Changnon, D. (2000). Long-term fluctuations in hail incidences in the United States. *Journal of Climate*, *13*(3), 658-664. https://doi.org/10.1175/1520-0442(2000)013%3C0658:LTFIHI%3E2.0.CO;2

Cintineo, J. L., Smith, T. A., Lakshmanan, V., Brooks, H. E., & Ortega, K. L. (2012). An objective High-Resolution Hail climatology of the contiguous United States. *Weather and Forecasting*, *27*(5), 1235–1248. https://doi.org/10.1175/waf-d-11-00151.1

Crum, T. D., & Alberty, R. L. (1993). The WSR-88D and the WSR-88D Operational Support Facility. *Bulletin of the American Meteorological Society*, *74*(9), 1669–1687. https://doi.org/10.1175/1520-0477(1993)074

Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A. A., & Szturc, J. (2019). Application of machine learning to large hail prediction - The importance of radar reflectivity, lightning occurrence and convective parameters derived

from ERA5. *Atmospheric Research*, *227*, 249–262. https://doi.org/10.1016/j.atmosres.2019.05.010

Doesken, N. J. (1994). Hail, hail, hail! The summertime hazard of eastern Colorado. *Colorado Climate Publication*, *17*.

Dudhia, J. (1997). Back to basics: Thunderstorms: Part 2 - storm types and associated weather. *Weather*, *52*(1), 2–7. https://doi.org/10.1002/j.1477-8696.1997.tb06241.x

Edwards, R., & Thompson, R. L. (1998). Nationwide comparisons of hail size with WSR-88D vertically integrated liquid water and derived thermodynamics sounding data. *Weather and Forecasting*, *13*(2), 277–285. https://doi.org/10.1175/1520-0434(1998)013<0277:NCOHSW>2.0.CO;2

Fabry, F. (2015). *Radar Meteorology Principles and Practice* (1st ed., p. 256). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107707405

Federer, B., & Waldvogel, A. (1978). Time-resolved hailstone analyses and radar structure of Swiss storms. *Quart. 1. Roy. Meteor. Soc.*, 104, 69-90.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.

Gagne, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and forecasting*, *32*(5), 1819-1840.

Gagne, D. J., McGovern, A., Brotzge, J. A., Coniglio, M. C., Correia, J., & Xue, M. (2015). Day-Ahead Hail Prediction Integrating Machine Learning with Storm-Scale Numerical Weather Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(2), 3954–3960. https://doi.org/10.1609/aaai.v29i2.19053

Gavrilov, M. B., Marković, S. B., Zorn, M., Komac, B., Lukić, T., Milošević, M. V., & Janićević, S. (2013). Is hail suppression useful in Serbia? – General review and new results. *Acta Geographica Slovenica*, *53*(1), 165–179. https://doi.org/10.3986/ags53302

Gensini, V. A., Converse, C., Ashley, W. S., & Taszarek, M. (2021). Machine Learning Classification of significant tornadoes and hail in the U.S. using era5 proximity soundings. *Weather and Forecasting*. https://doi.org/10.1175/waf-d-21-0056.1

George, J. J. (2014). *Weather forecasting for Aeronautics*. Academic Press.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *The University of Texas at El Paso Departmental Technical Reports (Computer Science)*, UTEP-CS-18-09. https://digitalcommons.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techre

Gunturi, P., & Tippett, M. (2017). Impact of ENSO on U.S. tornado and Hail Frequencies - Managing Severe Thunderstorm Risk. *Columbia University.* http://www.columbia.edu/~mkt14/files/WillisRe_Impact_of_ENSO_on_US_Tornado_and_Hail_frequencies_Final.pdf

Hart, J. A. (1998). The occurrence and non-occurrence of large hail with strong and violent tornado episodes: Frequency distributions. Preprints, *19th Conference on Severe Local Storms*, Minneapolis, MN, American Meteorological Society, 285-286.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), 349-360.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Hayduk, R. J. (1973). Hail damage to typical aircraft surfaces. *Journal of Aircraft*, *10*(1), 52-55.

Hill, A. J., Herman, G. R., & Schumacher, R. S. (2020). Forecasting Severe Weather with Random Forests. *Monthly Weather Review*, *148*(5), 2135–2161. https://doi.org/10.1175/mwr-d-19-0344.1

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Huang, W., Jiang, Y., Li, X., Pan, Y., Li, X., Guo, R., Huang, Y., & Duan, B. (2019). Classified early-warning and nowcasting of hail weather based on radar products and random forest algorithm. *2019 International Conference on Meteorology Observations (ICMO)*. https://doi.org/10.1109/icmo49322.2019.9026039

IBM. (2023). *What is Supervised Learning?* IBM. Retrieved from https://www.ibm.com/topics/supervised-learning

Insurance Bureau of Canada. (2023). *Hail protection*. IBC. Retrieved from https://www.ibc.ca/stay-protected/severe-weather-safety/hail

Jewell, R., & Brimelow, J. (2009). Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Weather and Forecasting*, *24*(6), 1592–1609. https://doi.org/10.1175/2009WAF2222230.1

Johns, R. H., & C. A. Doswell III (1992). Severe local storms forecasting. *Weather Forecasting*, 7, 588-612.

Jolliffe, I. T., & Stephenson, D. B. (2011). Forecast verification. In *Wiley eBooks*. https://doi.org/10.1002/9781119960003

Knight, C. A., Ashworth, T., & Knight, N. C. (1978). Cylindrical ice accretions as simulations of hail growth: II. the structure of fresh and annealed accretions. *Journal of the Atmospheric Sciences*, *35*(10), 1997–2009. https://doi.org/10.1175/1520-0469(1978)035<1997:ciaaso>2.0.co;2

Knight, C. A., & Knight, N. C. (2001). Chapter 6. Hailstorms. In *Severe Convective Storms* (pp. 223–254). essay, American Meteorological Society.

Krauss, T. W., & Santos, J. E. (2004). Exploratory analysis of the effect of hail suppression operations on precipitation in Alberta. *Atmospheric Research*, *71*(1–2), 35–50. https://doi.org/10.1016/j.atmosres.2004.03.004

Kumjian, M. R., Khain, A. P., Ben Moshe, N., Ilotoviz, E., Ryzhkov, A.V., & Phillips, V. T. J. (2014). The anatomy and physics of ZDR columns: Investigating a polarimetric radar signature with a spectral bin microphysical model. *Journal of Applied Meteorology and Climatology*, *53*(7), 1820–1843. https://doi.org/10.1175/JAMC-D-13-0354.1

Kumjian, M. R., Lombardo, K., & Loeffler, S. (2021). The evolution of hail production in simulated supercell storms. *Journal of the Atmospheric Sciences*. https://doi.org/10.1175/jas-d-21-0034.1

Labriola, J., Snook, N., Jung, Y., & Xue, M. (2019). Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with

several multi-moment microphysics schemes. *Monthly Weather Review*, *147*(4), 1193–1213. https://doi.org/10.1175/MWR-D-18-0266.1

Lemon L.R. (1980). *Severe thunderstorms radar identification techniques and warning criteria: A preliminary report*. NOAA Technical Memorandum.

Lin, Y., & Kumjian, M. R. (2022). Influences of CAPE on hail production in simulated supercell storms. *Journal of the Atmospheric Sciences*, *79*(1), 179–204. https://doi.org/10.1175/jas-d-21-0054.1

Loken, E. D., Clark, A. J., & McGovern, A. (2022). Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Weather and Forecasting*, *37*(6), 871–899. https://doi.org/10.1175/waf-d-21-0138.1

Louppe, G. (2014). Understanding random forests: From Theory to practice. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/1407.7502

Markowski, P., & Richardson, Y. (2010). *Mesoscale meteorology in Midlatitudes.* Wiley-Blackwell.

Marshall, J. (2000). *Fact Sheets Hail*. UCAR. Retrieved from https://web.archive.org/web/20100327214028/http://www.ucar.edu/communications/factsheets/Hail.html

Marwitz, J. D. (1972). The structure and motion of severe hailstorms. Part I: Supercell storms. *Journal of Meteorology*, *11*(1), 166–179. https://doi.org/10.1175/1520-0450(1972)011<0166:TSAMOS>2.0.CO;2

McCloskey, S. O., Loken, E. D., Jahn, D. E., Karstens, C. D., & Smith, B. T. (2021). Determining when and how a random forest adds value to day 1 spc hail forecasts. *National Weather Center Research Experience for Undergraduates*. https://caps.ou.edu/reu/reu21/finalpapers/McCloskey_FinalPaper.pdf

McClure, M. (2009). New One Inch Hail Criteria for Severe Thunderstorm Warnings. *National Weather Service Quad Cities IA/IL Weather Home Companion*, *6*(1), 1–1.

McGovern, A., Supinie, T., Gagne, D. J., Troutman, N., Collier, M., Brown, R., Basara, J., & Williams, J. (2010). Understanding severe weather processes through spatiotemporal

relational random forests. *2010 NASA Conference on intelligent Data Understanding*. https://c3.ndc.nasa.gov/dashlink/static/media/publication/Paper_17_.pdf

Mecikalski, J. R., Sandmæl, T. N., Murillo, E. M., Homeyer, C. R., Bedka, K. M., Apke, J. M., & Jewett, C. P. (2021). A Random-Forest Model to Assess Predictor Importance and Nowcast Severe Storms Using High-Resolution Radar–GOES Satellite–Lightning Observations. *Monthly Weather Review, 149*(6), 1725–1746. https://doi.org/10.1175/mwr-d-19-0274.1

Miller, L. J., Tuttle, J. D., & Knight, C. (1988). Airflow and hail growth in a severe Northern High Plains supercell. *Journal of the Atmospheric Sciences*, *45*(4), 736–762. https://doi.org/10.1175/1520-0469(1988)045

Morgan, G. M. (1970). An examination of the Wet-Bulb zero as a hail forecasting parameter in the Po Valley, Italy. *Journal of Applied Meteorology*, *9*(3), 537–540. https://doi.org/10.1175/1520-0450(1970)009

Murillo, E. M., & Homeyer, C. R. (2019). Severe hail fall and hailstorm detection using remote sensing observations. *Journal of Applied Meteorology and Climatology*, *58*(5), 947–970. https://doi.org/10.1175/JAMC-D-18-0247.1

NCEI. (2013). *Storm Events Database | National Centers for Environmental Information*. Retrieved from https://www.ncdc.noaa.gov/stormevents/

NOAA's National Weather Service, & Buchanan, S. (2010). *Record Setting Hail Event in Vivian, South Dakota on July 23, 2010*. Retrieved from https://www.weather.gov/abr/vivianhailstone

NSSL. (2023). *Severe Weather 101 - Hail basics.* NOAA National Severe Storms Laboratory. Retrieved from https://www.nssl.noaa.gov/education/svrwx101/hail/

North, G. R., Zhang, F., & Pyle, J. (2015). *Encyclopedia of Atmospheric Sciences (second edition)*. Academic Press.

Ortega, K. L. (2018). Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *Electronic Journal of Severe Storms Meteorology*, *13*(1), 1–36. https://doi.org/10.55599/ejssm.v13i1.69

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://hal.inria.fr/hal-00650905

Porter, K. A. (2022). *A benefit-cost analysis of impact-resistant asphalt shingle roofing.* Institute for Catastrophic Loss Reduction. Retrieved from https://www.iclr.org/wp-content/uploads/2022/04/Benefit-cost-analysis-of-Impact-resistant-asphalt-shingle-roofing2.pdf

Pulukool, F., Li, L., & Liu, C. (2020). Using Deep Learning and Machine Learning Methods to Diagnose Hailstorms in Large-Scale Thermodynamic Environments. *Sustainability*, *12*(24), 10499. https://doi.org/10.3390/su122410499

Punge, H. J., Bedka, K. M., Kunz, M., & Reinbold, A. (2017). Hail frequency estimation across Europe based on a combination of overshooting top detections and the era-interim reanalysis. *Atmospheric Research*, *198*, 34–43. https://doi.org/10.1016/j.atmosres.2017.07.025

Renaud, J. (2022). *Northern Hail Project recovers record-breaking hailstone.* Western News. Retrieved from https://news.westernu.ca/2022/08/record-breaking-hailstone/

Rogers, R. R., & Yau, M. K. (1996). *A short course in Cloud Physics*. Butterworth-Heinemum, 290 pp.

Rudden, J. (2022). *Crop-hail damage U.S. 2014-2018*. Statista. Retrieved from https://www.statista.com/statistics/1015612/crop-hail-damage-usa/

Samuel, A. L. (1959). Machine learning. *The Technology Review*, *62*(1), 42-45.

Sánchez, J. L., Fraile, R., De La Madrid, J. L., De La Fuente, M. T., Rodríguez, P., & Castro, A. (1996). Crop damage: The hail size factor. *Journal of Applied Meteorology and Climatology*, *35*(9), 1535-1541. https://doi.org/10.1175/1520-0450(1996)035%3C1535:CDTHSF%3E2.0.CO;2

Schmid, F., Wang, Y., & Harou, A. (2019). *Nowcasting Guidelines – A Summary*. World Meteorological Organization. Retrieved from

https://public.wmo.int/en/resources/bulletin/nowcasting-guidelines-%E2%80%93-summary

Scikit-learn developers. (2023). *User guide: contents*. Scikit-learn. Retrieved from https://scikit-learn.org/stable/user_guide.html

Souter, R. K., & Emerson, J. B. (1952). *Summary of available hail literature and the effect of hail on aircraft in flight*. National Advisory Committee for Aeronautics.

Stull, R. B. (2017). *Practical Meteorology: An algebra-based survey of Atmospheric Science*. University of British Columbia.

Tang, B. H., Gensini, V. A., & Homeyer, C. R. (2019). Trends in United States large hail environments and observations. *Npj Climate and Atmospheric Science*, *2*(1). https://doi.org/10.1038/s41612-019-0103-7

Waldvogel, A., Federer, B., & Grimm, P. (1979). Criteria for the detection of hail cells. *Journal of Applied Meteorology*, *18*(12), 1521–1525. https://doi.org/10.1175/1520-0450(1979)018

Wang, J., Fan, J., & Feng, Z. (2023). *Climatological Occurrences of Hail and Tornado Associated with Mesoscale Convective Systems in the United States*. https://doi.org/10.5194/nhess-2023-16

Watson, B. (2008). *Fire And Ice: Suppressing hailstorms in Alberta.* Wings Magazine. Retrieved from https://www.wingsmagazine.com/fire-and-ice-suppressing-hailstorms-in-alberta-1555/

Weisman, M. L., & Klemp, J. B. (1986). Characteristics of isolated convective storms. In *American Meteorological Society eBooks* (pp. 331–358). https://doi.org/10.1007/978-1-935704-20-1_15

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Williams, S. S., Ortega, K. L., Smith, T., & Reinhart, A. E. (2022). Comprehensive Radar Data for the contiguous United States: Multi-Year Reanalysis of Remotely Sensed Storms. *Bulletin of the American Meteorological Society*, *103*(3), E838–E854. https://doi.org/10.1175/bams-d-20-0316.1

Witt, A., Burgess, D. W., Seimon, A., Allen, J. T., Snyder, J. C., & Bluestein, H. B. (2018). Rapid-scan radar observations of an Oklahoma tornadic hailstorm producing giant hail. *Weather and Forecasting*, *33*(5), 1263–1282. https://doi.org/10.1175/WAF-D-18-0003.1

Witt, A., Eilts, M. D., Stumpf, G. J., Johnson, J. T., Mitchell, E., & Thomas, K. W. (1998). An enhanced hail detection algorithm for the WSR-88D. *Weather and Forecasting*, *13*(2), 286–303. https://doi.org/10.1175/1520-0434(1998)013

Witt, A., & Nelson, S. P. (1991). The use of single-Doppler radar for estimating maximum hailstone size. *Journal of Applied Meteorology*, *30*(4), 425–431. https://doi.org/10.1175/1520-0450(1991)030<0425:TUOSDR>2.0.CO;2

Wong, R. K., Chidambaram, N., Cheng, L., & English, M. (1988). The sampling variations of hailstone size distributions. *Journal of Applied Meteorology*, *27*(3), 254–260. https://doi.org/10.1175/1520-0450(1988)027<0254:tsvohs>2.0.co;2

Yao, H., Li, X., Pang, H., Sheng, L., & Wang, W. (2020). Application of random forest algorithm in hail forecasting over Shandong Peninsula. *Atmospheric Research*, *244*, 105093. https://doi.org/10.1016/j.atmosres.2020.105093

Zhou, Z. H. (2021). *Machine learning.* Springer Nature.