Semester Project

Joshua Elms[1]*

Abstract

This data analytics project is focused on predicting the size of largest hailstone that comes from any given storm. To do this, reports of more than 20,000 hail events and their associated meteorological data have been collected and tabularized.

Keywords

Meteorology — Numerical Weather Prediction — Machine Learning

[1] *Computer Science, School of Informatics , Computing and Engineering, Indiana University, Bloomington, IN, USA*

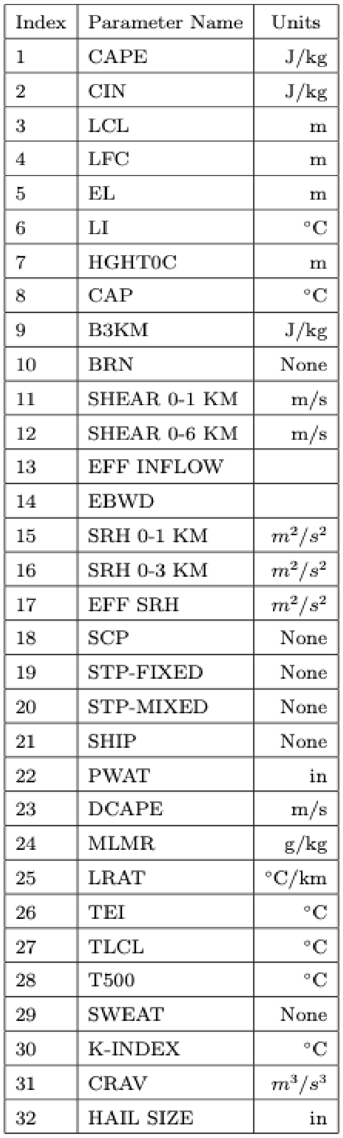
Contents

1. Problem and Data Description

According to the NOAA Annual Severe Weather website, there were 3,762 recorded severe hail events (hail over 0.75”) in the US in 2021 alone. Beyond the noble goal of furthering our understanding of the natural world, the ability to accurately predict which locations will experience hail storms could benefit insurance companies who pay out an average of $12,000 for residential damage claims and $4,000 for automobile claims. Multiplied by thousands of storms per year and hundreds of thousands of possible claims per storm, it is clear that increasing advanced warming times for likely hail conditions could have massive benefits for both the general public and insurance companies alike.

The data for this project consists of approximately 29,000 reports of severe hail events and the meteorological conditions present during the event. The hail event data was collected from the 2012-2016 entries in the NOAA Storm Prediction Center database. After extracting just the hail sizes, coordinates, and dates/times, those parameters were entered into the North American Mesoscale Model to determine vertical profiles of temperature, humidity, and wind within 3 hours and 20 km of the individual hail events. The meteorological profiles were passed to an open-source python program, SHARPpy, to calculate the exclusively continuous, numerical parameters we will be discussing in the following table.

After the data is extracted from the various sources and tabularized, it forms a CSV with 21902 entries and 55 fields, including the target variable of hailstone size.



Index	Parameter Name	Units
1	CAPE	J/kg
2	CIN	J/kg
3	LCL	m
4	LFC	m
5	EL	m
6	LI	°C
7	HGHT0C	m
8	CAP	°C
9	B3KM	J/kg
10	BRN	None
11	SHEAR 0-1 KM	m/s
12	SHEAR 0-6 KM	m/s
13	EFF INFLOW	
14	EBWD	
15	SRH 0-1 KM	m^2/s^2
16	SRH 0-3 KM	m^2/s^2
17	EFF SRH	m^2/s^2
18	SCP	None
19	STP-FIXED	None
20	STP-MIXED	None
21	SHIP	None
22	PWAT	in
23	DCAPE	m/s
24	MLMR	g/kg
25	LRAT	°C/km
26	TEI	°C
27	TLCL	°C
28	T500	°C
29	SWEAT	None
30	K-INDEX	°C
31	CRAV	m^3/s^3
32	HAIL SIZE	in

2. Data Preprocessing & Exploratory Analysis

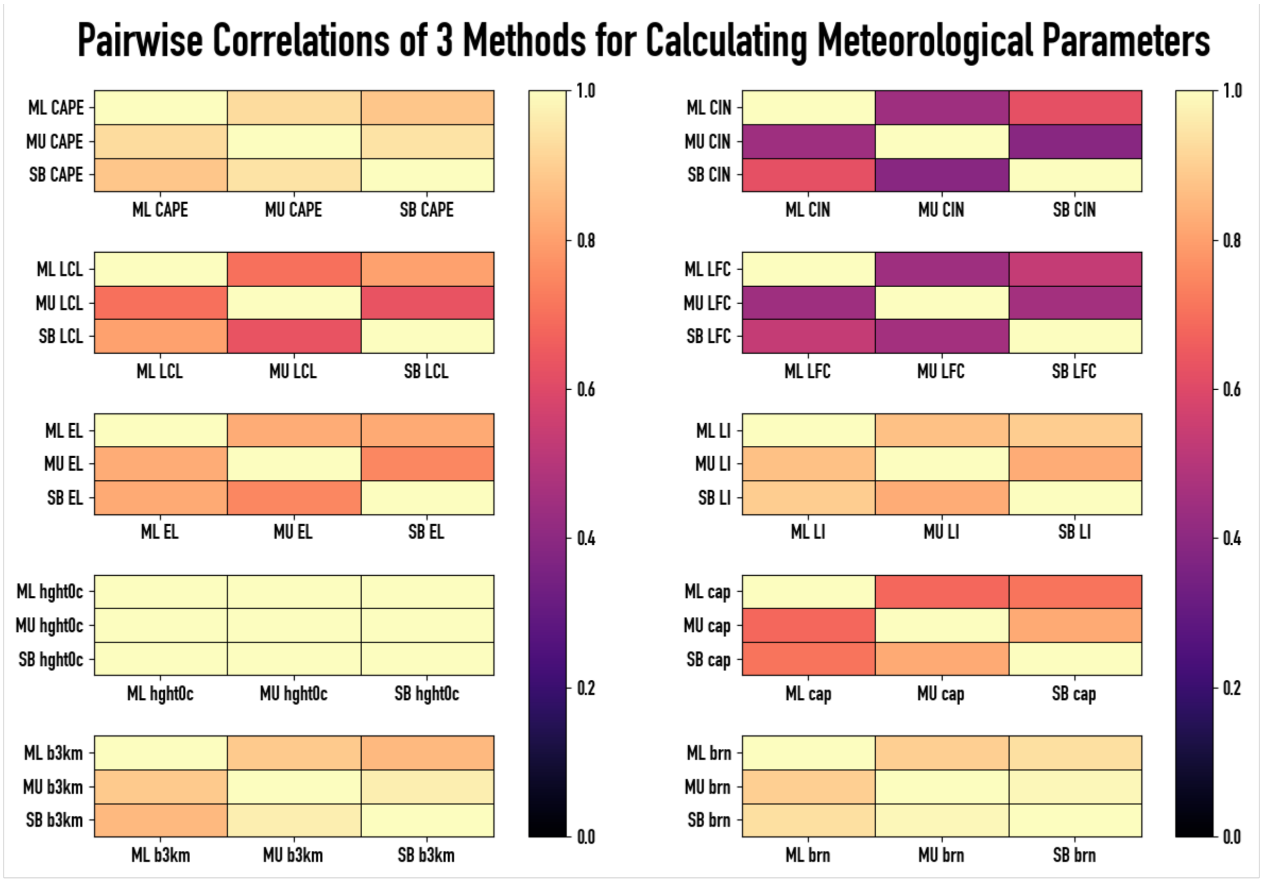
2.1 Parameter Selection

After the pipeline detailed in the Data Description was traversed, there were 53 parameters associated with each severe hail event (not including Hail Size). The 11 of the parameters (index positions 1-10 and 27) were repeated three times throughout the data, with the only difference being slightly different calculation methods for each. The three methods used were SB (Surface Based), ML (Mixed Layer), and MU (Most Unstable), each of which describes a process for calculating thermodynamic and wind related parameters. In short, SB uses the temperature at the surface, ML uses an average of the conditions up to an altitude of 100 mb (millibars), and MU uses the temperature of the most unstable air parcel found in the lowest 300 mb of the atmosphere. For a more in-depth explanation, refer to the NOAA Storm Prediction Center's guide on the subject.

We first attempted to determine how well these these various calculation techniques track each other by generating correlation plots corresponding to each of the first 10 variables. It our hope that they would all perform almost identically and thereby allow only one set to be used for visualization, analysis, and modeling purposes. However, the chart makes it clear that this is not the case.

I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.

I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing. I include more text here for the purpose of... uh... testing.



2.2 Handling Missing Values

2.3 Exploratory Data Analysis

3. Algorithm and Methodology

Briefly explain the algorithms in this section, i.e., linear regressions. You can add more subsections if needed, i.e., regression trees etc.

4. Experiments and Results

5. Summary and Conclusions

Acknowledgments