U.S. Storm Event Predictability

By Denise Macias

Table of Contents

Purpose

The purpose of this project will be to determine the predictability of recent storm events in the U.S. through historical data.

Data

- Source: National Climatic Data Center

- Original dataset:
  - 1,740,597 observations
  - 51 features
  - 70 event types
  - 1950-2022

- Cleaned dataset:
  - 1,740,597 observations
  - 17 features
  - 70 event types
  - 1950-2022

Data Exploration

- **Target:** Magnitude
- **Focus:** Identify storm events that hold the most promise for a reliable magnitude forecast using the following criteria:
    - 1) Events for which magnitude is tracked
    - 2) Events with the highest event counts
    - 3) Events with the highest casualties

Data Exploration: Events Identified

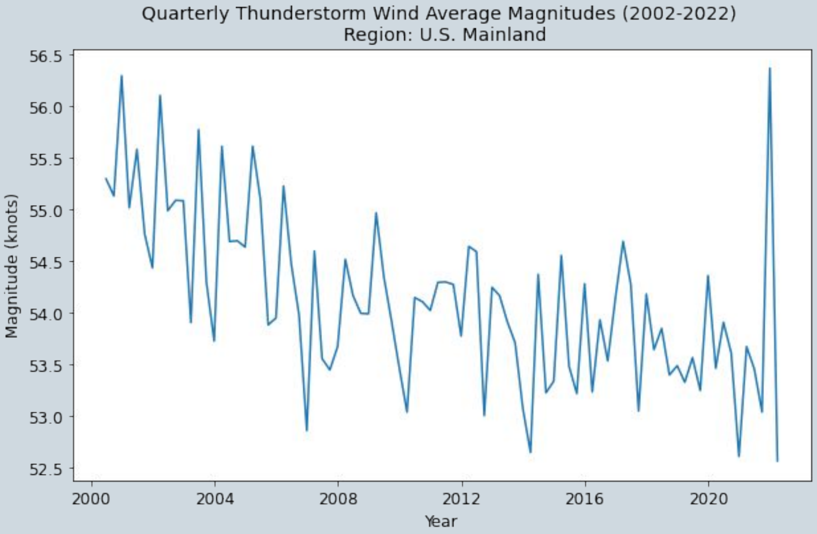
- **Thunderstorm Wind**
  - 484,908 events
  - 11,609 injuries; 2,048 deaths; 15 billion in damages
- **Tornado**
  - 73,987 events
  - 97,472 injuries; 12,230 deaths; 7 billion in damages
- **Hail**
  - 382,234 events
  - 1,563 injuries; 30 deaths; 34 billion in damages
- **High Wind**
  - 79,894 events
  - 1,986 injuries; 706 deaths; 16 billion in damages

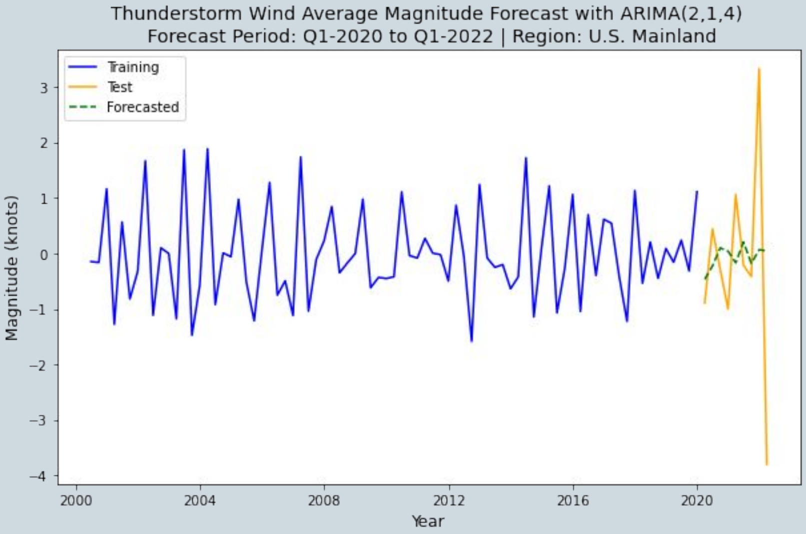
Data Exploration: Event Findings

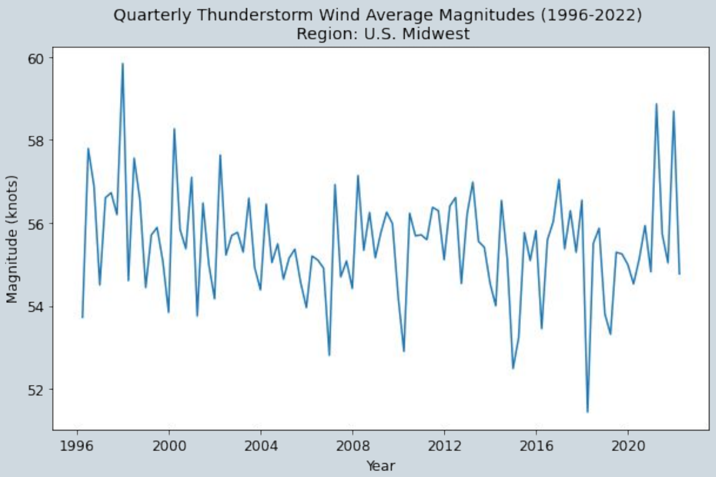
- **Thunderstorm Wind (1955-2022) → Contender**
  - Very chaotic from 1955-1996 and not representative of recent data
  - Better consistency from 1996-2022
  - Regions vary greatly from the national trend
    - % of data per region: South (48%), Midwest (36%), Northeast (11%), West (5%)
    - Midwest and West experience most severe thunderstorm wind events
- **Tornado (1950-2022) → Disqualified**
  - No records after 2006
- **Hail (1955-2022) → Contender**
  - A little chaotic from 1955-2000 and not representative of recent data
  - Better consistency from 2000-2022
  - Regions do not vary as much from national trend
    - % of data per region: Midwest (44%), South (42%), West (9%), Northeast (4%)
    - Midwest and South experience most severe hail events
- **High Wind (1996-2022) → Disqualified**
  - Chaotic all throughout and no consistency

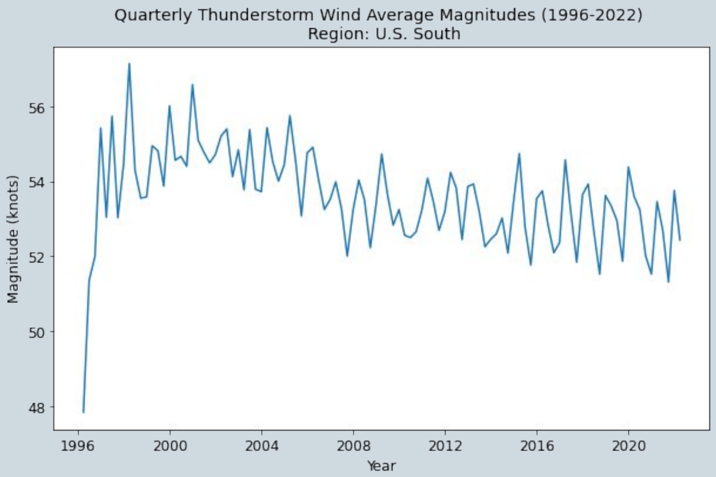
Thunderstorm Wind

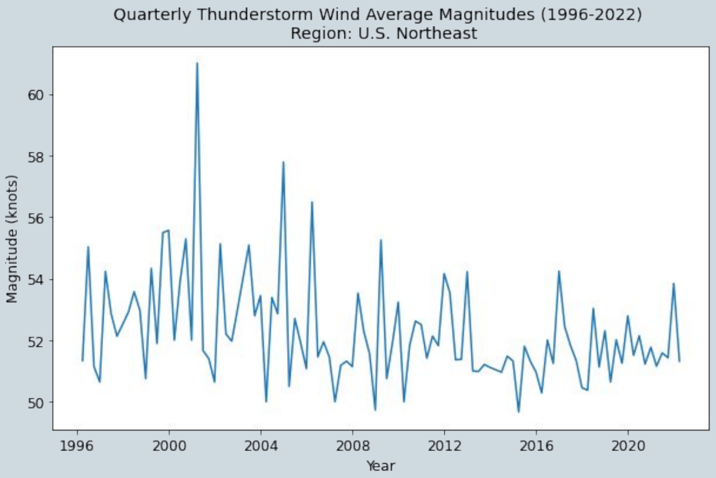
ARIMA Model Forecasts

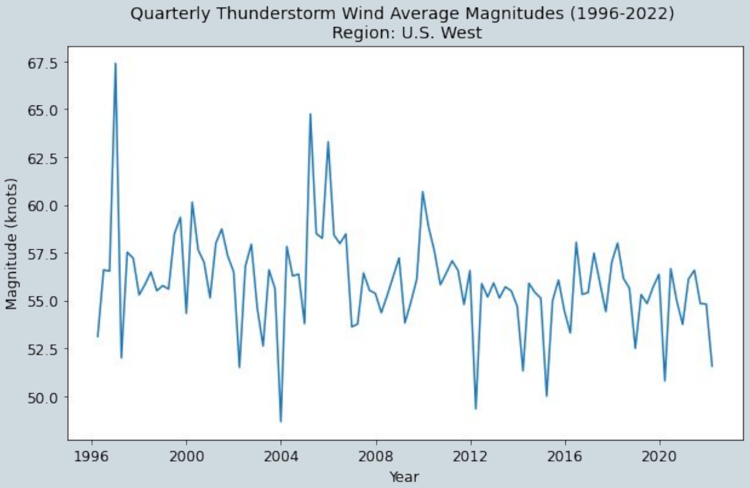


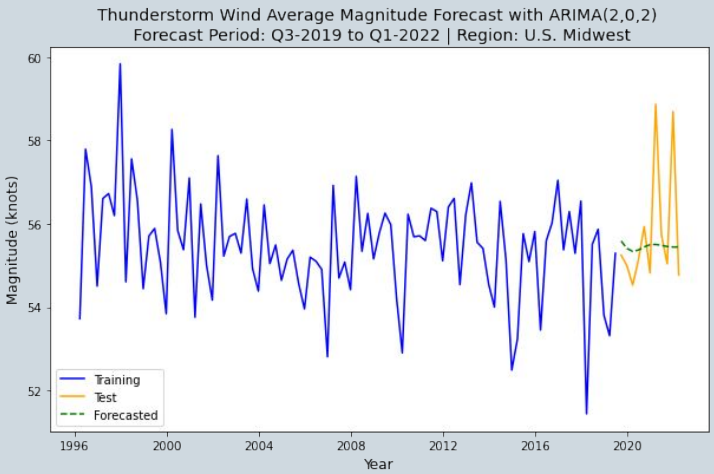


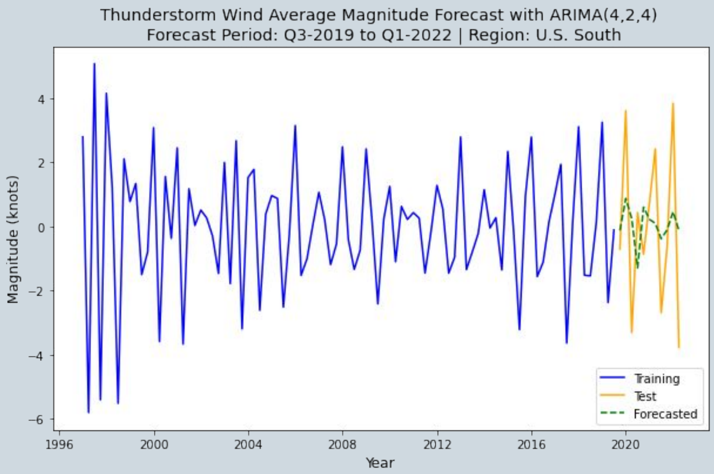

Quarterly Thunderstorm Wind Average Magnitudes (1996-2022)
Region: U.S. Midwest
Magnitude (knots)
Year
60
58
56
54
52
1996
2000
2004
2008
2012
2016
2020

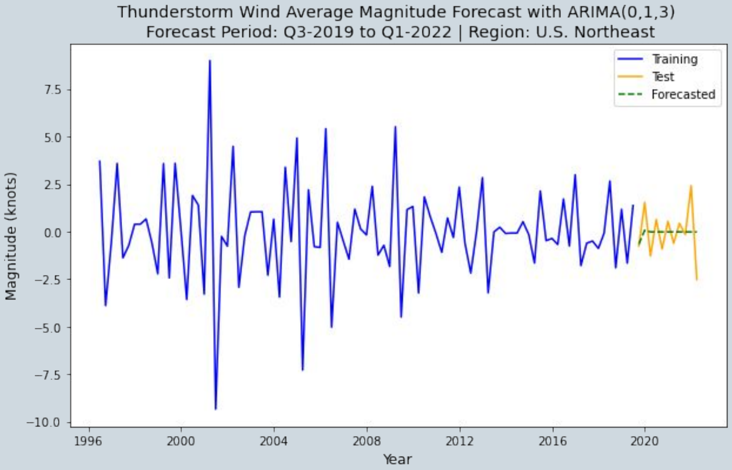

Quarterly Thunderstorm Wind Average Magnitudes (1996-2022)
Region: U.S. South
Magnitude (knots)
Year
56
54
52
50
48
1996
2000
2004
2008
2012
2016
2020

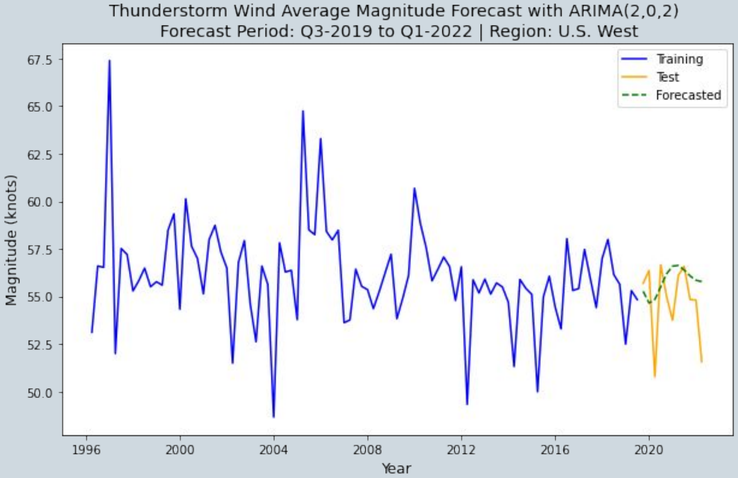

Quarterly Thunderstorm Wind Average Magnitudes (1996-2022)
Region: U.S. Northeast
Magnitude (knots)
Year
60
58
56
54
52
50
1996
2000
2004
2008
2012
2016
2020


Quarterly Thunderstorm Wind Average Magnitudes (1996-2022)
Region: U.S. West
Magnitude (knots)
Year
67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
1996
2000
2004
2008
2012
2016
2020


Thunderstorm Wind Average Magnitude Forecast with ARIMA(2,0,2)
Forecast Period: Q3-2019 to Q1-2022 | Region: U.S. Midwest
Magnitude (knots)
Year
Training
Test
Forecasted


Thunderstorm Wind Average Magnitude Forecast with ARIMA(4,2,4)
Forecast Period: Q3-2019 to Q1-2022 | Region: U.S. South
Magnitude (knots)
Year
Training
Test
Forecasted


Thunderstorm Wind Average Magnitude Forecast with ARIMA(0,1,3)
Forecast Period: Q3-2019 to Q1-2022 | Region: U.S. Northeast
Magnitude (knots)
Year
Training
Test
Forecasted

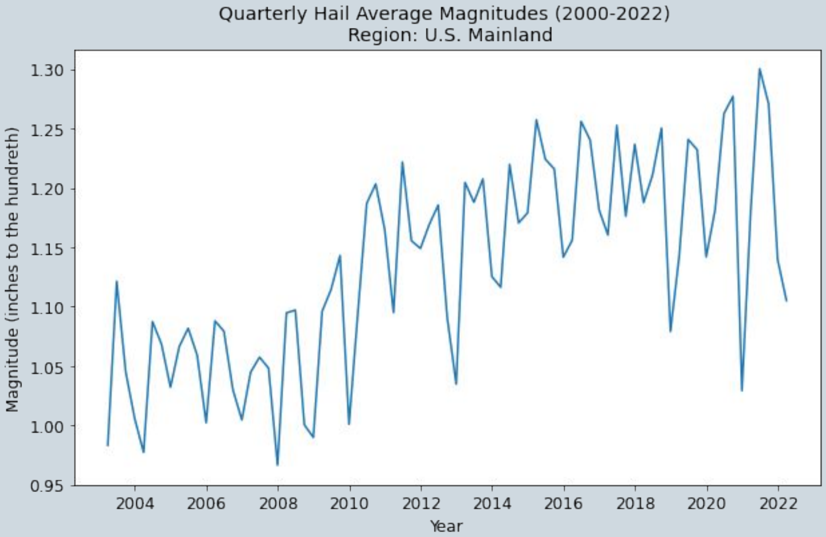

Thunderstorm Wind Average Magnitude Forecast with ARIMA(2,0,2)
Forecast Period: Q3-2019 to Q1-2022 | Region: U.S. West
Magnitude (knots)
Year
Training
Test
Forecasted

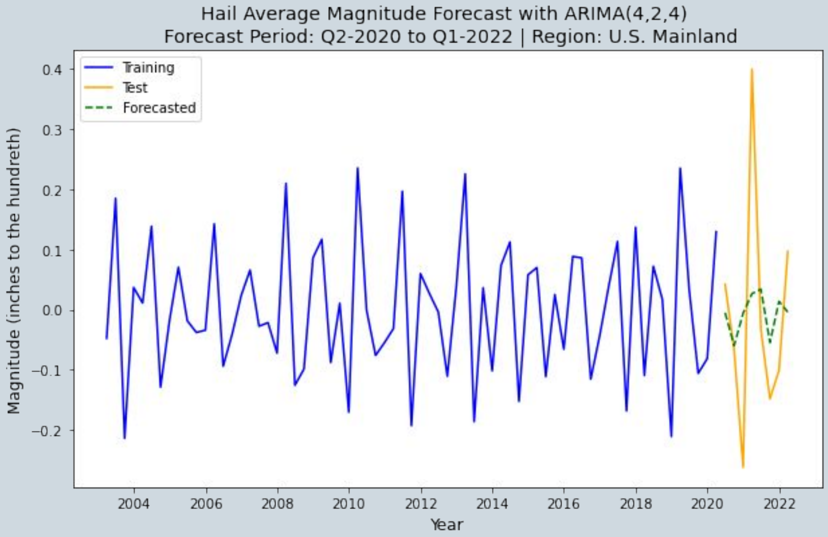
ARIMA Model Findings

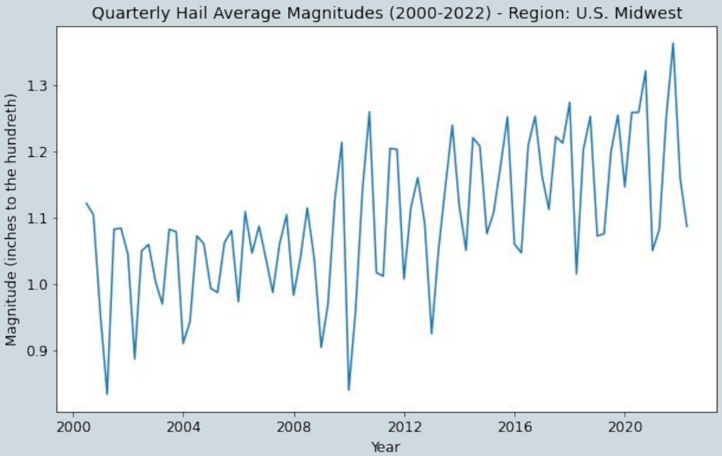
- **U.S. Mainland**
  - The ARIMA (2,1,4) model explained -2% of the variability
  - Did not produce a reliable forecast → RMSE (1.80) > STD (0.98)
- **Midwest**
  - The ARIMA (2,0,2) model explained -4% of the variability
  - Did not produce a reliable forecast → RMSE (1.49) > STD (1.32)
- **South**
  - The ARIMA (4,2,4) model explained 10% of the variability
  - Did not produce a reliable forecast → RMSE (2.37) > STD (2.12)
- **Northeast**
  - The ARIMA (0,1,3) model explained 4% of the variability
  - Produced a reliable forecast→ RMSE (1.28) < STD (1.52)
- **West**
  - The ARIMA (2,0,2) model explained -30% of the variability
  - Produced a reliable forecast → RMSE (2.16) < STD (2.66)

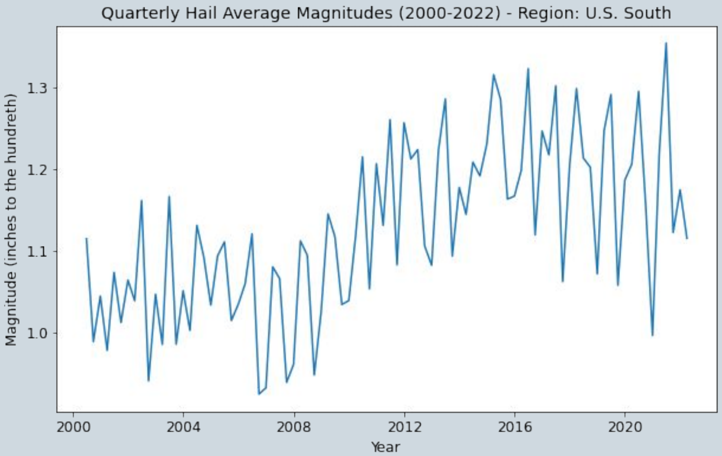
Hail

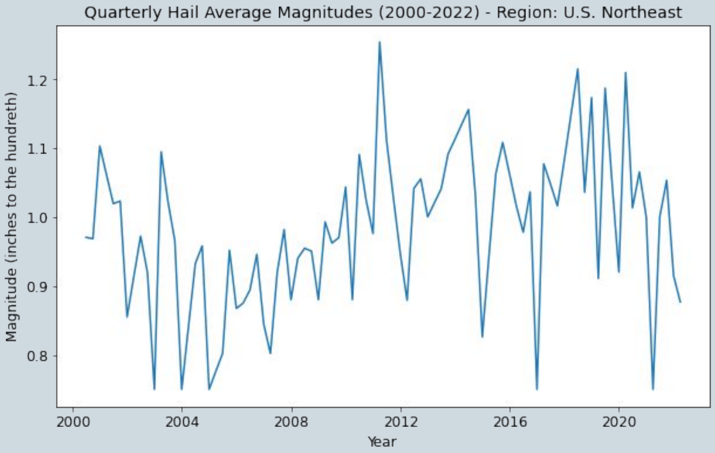
ARIMA Model Forecasts

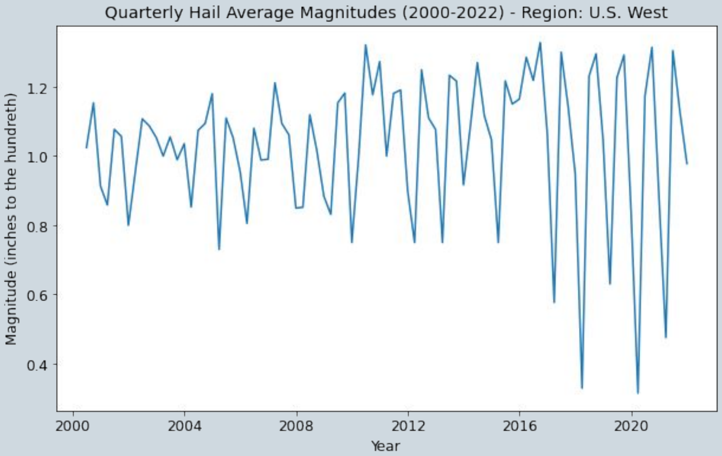


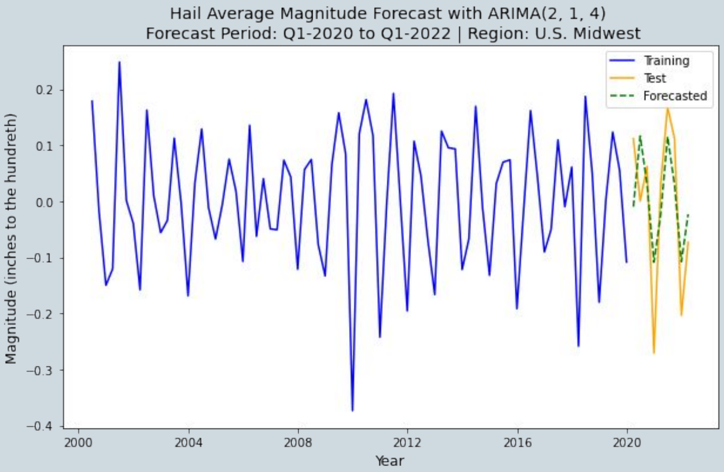


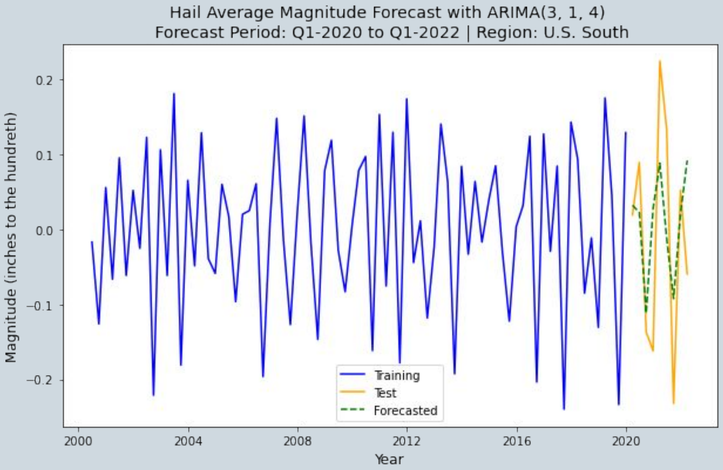

Quarterly Hail Average Magnitudes (2000-2022) - Region: U.S. Midwest
Magnitude (inches to the hundreth)
1.3
1.2
1.1
1.0
0.9
2000
2004
2008
2012
2016
2020
Year

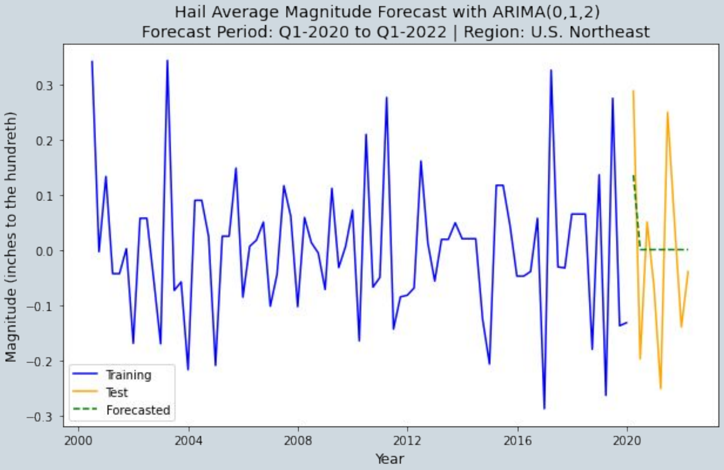

Quarterly Hail Average Magnitudes (2000-2022) - Region: U.S. South
Magnitude (inches to the hundreth)
1.3
1.2
1.1
1.0
2000
2004
2008
2012
2016
2020
Year

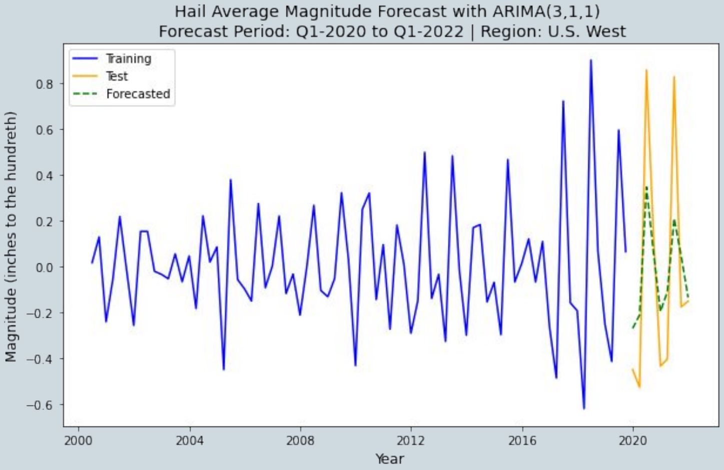

Quarterly Hail Average Magnitudes (2000-2022) - Region: U.S. Northeast
Magnitude (inches to the hundreth)
1.2
1.1
1.0
0.9
0.8
2000
2004
2008
2012
2016
2020
Year


Quarterly Hail Average Magnitudes (2000-2022) - Region: U.S. West
Magnitude (inches to the hundreth)
1.2
1.0
0.8
0.6
0.4
2000
2004
2008
2012
2016
2020
Year


Hail Average Magnitude Forecast with ARIMA(2, 1, 4)
Forecast Period: Q1-2020 to Q1-2022 | Region: U.S. Midwest
Training
Test
Forecasted
Magnitude (inches to the hundreth)
Year


Hail Average Magnitude Forecast with ARIMA(3, 1, 4)
Forecast Period: Q1-2020 to Q1-2022 | Region: U.S. South
Training
Test
Forecasted
Magnitude (inches to the hundreth)
Year


Hail Average Magnitude Forecast with ARIMA(0,1,2)
Forecast Period: Q1-2020 to Q1-2022 | Region: U.S. Northeast
Training
Test
Forecasted
Magnitude (inches to the hundreth)
Year


Hail Average Magnitude Forecast with ARIMA(3,1,1)
Forecast Period: Q1-2020 to Q1-2022 | Region: U.S. West
Training
Test
Forecasted
Magnitude (inches to the hundreth)
Year

ARIMA Model Findings

- **U.S. Mainland**
  - The ARIMA (4,2,4) model was able to explain 12% of the variability
  - Did not produce a reliable forecast → RMSE (0.17) > STD (0.12)
- **Midwest**
  - The ARIMA (2,1,4) model was able to explain 55% of the variability
  - Produced a reliable forecast → RMSE (0.09) < STD (0.12)
- **South**
  - The ARIMA (3,1,4) model was able to explain 31% of the variability
  - Did not produce a reliable forecast → RMSE (0.118) > STD (0.114)
- **Northeast**
  - The ARIMA (0,1,2) model was able to explain 21% of the variability
  - Did not produce a reliable forecast → RMSE (0.16) > STD (0.13)
- **West**
  - The ARIMA (3,1,1) model was able to explain 58% of the variability
  - Did not produce a reliable forecast → RMSE (0.33) > STD (0.30)

Conclusion

- Not very predictable on the regional or national level
- Removing outliers would create biased models that wouldn't necessarily translate into the real world

- **Next Steps:**
    - Changing the timeframe of the averages from quarterly to monthly
    - Introducing exogenous variables
    - Exploring other models such as SARIMA
    - Exploring additional events

Thanks!

Denise A. Macias

Data Science Candidate @ General Assembly

deniseamacias1@gmail.com