**Hail Weather Event Detection and Risk Analysis using Deep Learning**

**Group Members:** Bobbie Cavna, Alden Jettpace, Satya Uma Praneetha Parupudi

**Introduction**

Geographic Information Systems (GIS) emerged with the first computers in the 1960s in order to model damages from natural disasters such as earthquakes and hurricanes. Over the years, new technology and theories have been introduced in order to improve the capability of GIS to predict and model such disasters, with models for hurricanes becoming increasingly complex. However, hail modeling approaches have been long neglected. This is in spite of the fact U.S hail losses now on average exceed $1-billion per incident, and account for 45.5% of all homeowner insurance claims [1]. Current models for hail tend to be based on frequency based risk models, which requires assuming no records are missed. Despite this, many have observed that these models, even those deployed by FEMA, tend to have an urban bias in their mapping [2, pp. 164].

**Problem Statement**

A wider detection model, trained to detect possible hail events based on recorded atmospheric conditions and utilizing deep-neural-network technology has the potential to improve current hail modeling algorithms. By creating these models, more accurate risk assessments for hail frequency and damages can be generated, and the biased models currently employed by traditional studies may be enhanced with studies that leverage deep-learning technology.

**Data Description**

For our project, we used two datasets, one which formed our feature data and one which formed our label data. The links to both these datasets can be found in Appendix B, while the storage location of their raw and processed forms can be found in Appendix D.

The first dataset, “ERA5 hourly data on pressure levels”, was extracted manually from the dataset download page of the Copernicus data store’s dataset download page [3]. The data itself consists of four features: V-component of Wind, U-component of Wind, Temperature, and Humidity/Dew-Point. These features were extracted at 37 different atmospheric levels, from the years 2011-2022, across the latitude-longitude range of (37N=>42N, 89W=>84W). The features that we extracted were selected based on information from an existing study [4, pp. 3].

The second dataset “NOAA Hail Event Data” was extracted from the National Oceanic and Atmospheric Administration’s website, and consists of the recorded weather hail events of several states, specifically Indiana, Illinois, Kentucky, Michigan, and Ohio, from the years 2011-2022. These records were used to generate the positive instances within our data, with negative instances being considered a default value at all other times and places. Because of the rarity of hail, this meant that our methodologies saw a severe label imbalance.

**Methodology and Approach**

The first step was to clean the dataset, by ensuring that there were no duplicate hail events within the positive-labels datasets and ensuring that no features were missing in the feature dataset. As the goal was to demonstrate the capabilities of using deep-learning processes and technology in developing Hail Risk Analysis Charts, we approached the process of hail-event detection from multiple methodologies which required different formats for the input and output data. Thus, after cleaning the datasets we processed the data into two new datasets, to match two separate methodologies. Within both methodologies, the years 2011-2020 were used as training data while the years 2021-2022 were testing data. Model performance was measured using each year separately as opposed to in bulk.

**Method 1**

Method 1 was intended to limit the dimensionality of the data, and to make it easier to understand and process. In order to determine which dimensions to limit, background meteorology information was obtained from online presentations by David Stang [5]. Incorporating this information resulted in focusing on the relationships between altitude, temperature, dew-point temperature, wind speed, and vertical velocity of the wind. Altitudes were also limited to 23 values, rather than using all 37.

Transformations were applied to the ERA-5 source data, in order to obtain dew point from temperature and humidity features and to obtain wind speed from the separate U/V components of the wind. Vertical velocity of the wind was not considered in our other models, but was extracted separately and added only for method 1. Once transformed, the resulting 4 weather features were related to the 23 selected altitude levels to create 23 x 4 matrices for the primary model input.

Correlations between latitude and longitude, as well as correlations across different time periods, were not considered in Method 1. This was because hail was reported to be localized over time and space, in an existing work that we reviewed [4, pp. 1]. Additionally, the long-term trend in weather patterns was ignored in favor of focusing on producing output that could be used to derive current-state risk relativities. This current-state focus was also intended to help differentiate this study from the existing methods used for long-term global warming projections and near-term weather forecasting.

As a result of the above design choices, latitude longitude locations were treated as independent observations. The time-series approach, used in our other models, was also converted to a 24 step daily LSTM model. This was done primarily to help with class imbalance and to make it easier to match features with the time of known hail events. As such, the LSTM model input was reshaped from 24 hours x 23 altitudes x 4 features for each day and location to the following dimensions: (704,450 x 24 x 94). Note, that locations were trimmed to include only 193 points falling within or near the borders of Indiana.

The model architecture included 3 LSTM layers with 276 units, 25% Dropout, RELU activations, and L1 regularization followed by a fully connected dense layer with 276 units and RELU activation, and a final fully connected output layer with sigmoid activation and 1 unit. A Keras Balanced Batch Generator was also used to help with the class imbalance.

**Method 2**

In the second methodology, the feature data was processed into the form of a sequential-convolution, similar to that of film data, with each 20\*20 matrix holding 148 features representing the four atmospheric features at each atmospheric level. Using a 5-length sequence our feature data dimensions where (None, 5, 20, 20, 148). Specifically, these represent the conditions in that geographic region in the times 2-hours before the hail event to be detected to 2-hours after that event, or in effect (t-2, t-1, t, t+1, t+2). The label-data was formed by creating 20\*20 0-matrices for each hour in the time-period, and using the NOAA records of hail-events to create positive labels. This methodology creates a large imbalance between labels, with non-hail-events outnumbering hail-events by a factor of 10,000.

Thus, potential tells by the relationship of these features over time, such as a sudden drop in humidity at certain atmospheric levels before or after a hail event, or hail-drift over time thanks to U/V-wind components could give a more true-to-life sense of how atmospheric conditions interact in hail events. Thus, the methodology is based around next-frame prediction, but functions instead as ‘frame-creation’ by using future frames, with each frame being a gray-scale representation of the probability of a hail-event at that location at that time.

From this methodology, two models were made. The first ignored the sequence-form of the data entirely, instead being a simple CNN ending with a reshaped FNN, totaling just 6-layers, and using 3-relu-activation layers and a dropout rate of 0.5, and then trained for 6 epochs. The second methodology took the sequence form into account by using 4-layers of ConvLSTM2D layers which perform convolution within a LSTM cell. Normalizing after each layer, the final layer was a simple sigmoid-activated convolution layer without sequence, and trained for 8 epochs. Both of these models were then optimized with an adjusted loss-valuation with false-negatives having 5000x more weight, or effectively class\_weight = {0:1, 1:5000}.

**Results**

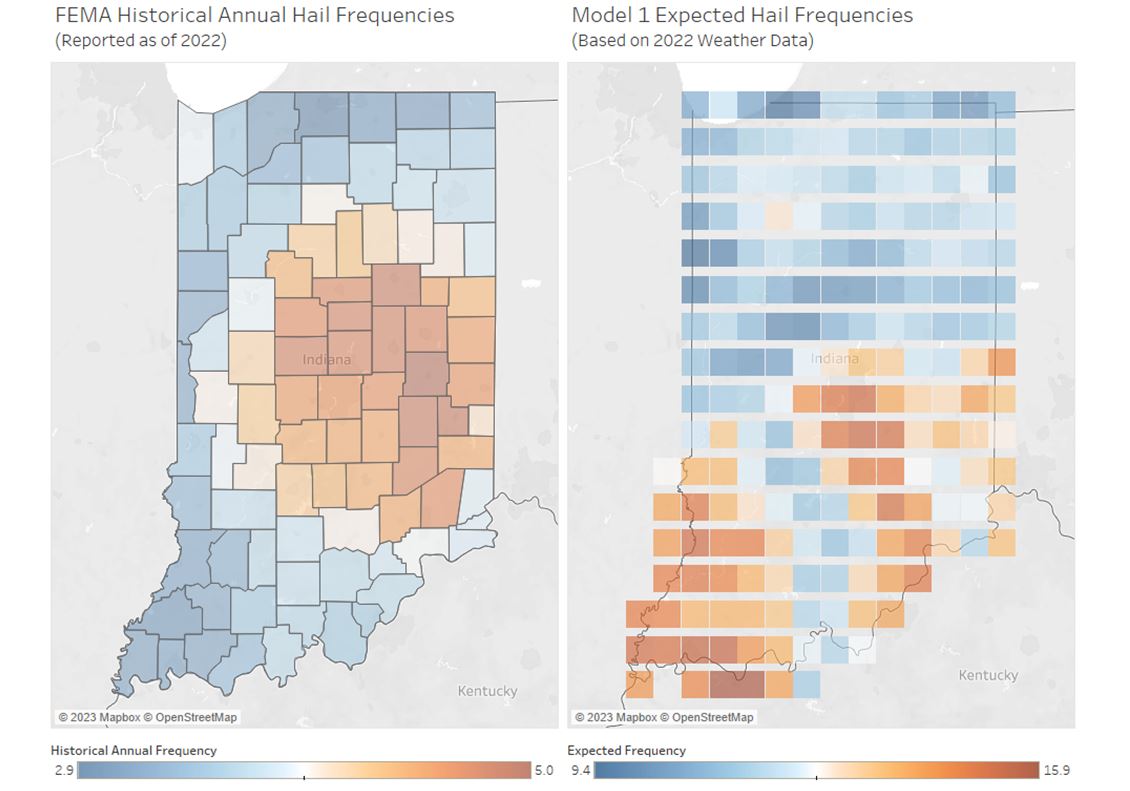
Because of the severe label-imbalance, the success of the models was measured primarily by the models positive precision and recall, or how well it detected hail events in the years 2021 and 2022 separately. Accuracy was considered as mainly determined by the precision of hail detection. We also generated confusion matrices to better visualize the precision and recall of the model. The numerical measures of each model’s performance for each test year can be found in Table1. Overall, we found that precision in such detection models was poor, with false-positives outnumbering true-positives by a factor of nearly 100,000. However, recall tended to be fair, being near to above 50% year by year, except for method 1 in 2022.

| Model | Year | Precision (1) | Recall (1) | Precision (0) | Recall (0) | Accuracy |
| --- | --- | --- | --- | --- | --- | --- |
| Model 1:  method 1 - LSTM | 2021 | 0.0039 | 0.4444 | 0.9996 | 0.9290 | 0.9287 |
| 2022 | 0.0032 | 0.0952 | 0.9993 | 0.9781 | 0.9775 |
| Model 2:  method 2 - Conv&FNN | 2021 | 9.589e-5 | 0.5229 | 0.9999 | 0.8303 | 0.8303 |
| 2022 | 1.583e-4 | 0.4916 | 0.9999 | 0.8413 | 0.8413 |
| Model 3:  method 2 - ConvLSTM | 2021 | 1.385e-4 | 0.4587 | 0.9999 | 0.8969 | 0.8969 |
| 2022 | 2.739e-4 | 0.6759 | 0.9999 | 0.8739 | 0.8739 |

*Table 1: Evaluation metrics for all three models across both test years.*

Each methodology had a different focus on visualizing the performance of its detections. The prediction results of method 1 were transformed into a risk map for hail in Indiana during the year 2022. The results were then compared to existing FEMA risk maps to show how a deep learning model could be used to generate risk relativity factors to improve more traditional types of risk analysis.

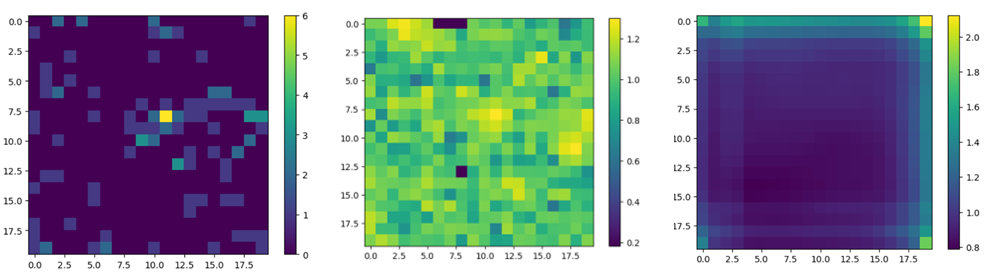
This was done by summing the model’s predicted probabilities, for each date and location. This approach was chosen as a proxy for deriving the expected values that might result from multiple independent Bernoulli random variables with varying parameters. Note that, in practice, this type of estimate would normally be based on a larger data set than what was used with Model 1. However, the estimate shown was based on a single year, for illustrative purposes only, due to the limited scope of our analysis. The comparison can be found in Figure 1 below:

****

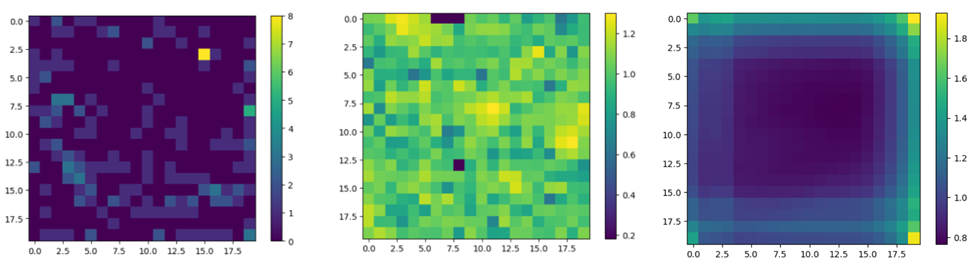
*Figure 1: FEMA 2022 Hail Frequency Map and Method 1 2022 Expected Frequencies Map.*

Despite the tendency towards false-positives by the poor precision, the relative frequency of the cells shows a lack of the urban bias found within the FEMA historical risk map, with higher frequencies in the south rather than around Indianapolis. While the recall performance does raise the possibility of missed hail events in central Indiana, nothing in the model is designed to deliberately snub urban areas so the likelihood of the undetected hail events all occurring within that area is highly unlikely.

Method 2 instead focused on the performance of the model relative to the labels rather than to existing FEMA risk maps. It factors in the tendency to over guess by correcting the summed probabilities for percentage compared to predicted mean for each year, then comparing to the summed labels to see if more frequent predictions were made in areas where hail was more frequent within that year, as found in Figures 2 and 3.



*Figure 2: Normalized hail event frequency for 2021 in labels, FNN-output, and ConvLSTM output*

**

*Figure 3: Normalized hail event frequency for 2022 in labels, FNN-output, and ConvLSTM output*

While at first difficult to see, the FNN output had higher frequency near at least a single region in each year which had a higher frequency of hail events, though outside these singular regions the outputs had little recall. The ConvLSTM model by contrast tended to have higher frequency of positive guesses at boundaries and corners of the image, most likely as a result of padding each layer. While the lack of precision confuses the results, a slight increase in shade corresponds to the areas of higher frequency within the label data.

**Limitations**

While we were able to train and implement three models through two different methodologies, we encountered several issues which limited our project.

The first limitation was the size of the data meant several days of downloading where required in order to have all we needed available, and required specialized storage on Google-Pro, as our personal computers either did not have enough space or a GPU not supported by keras nor pytorch. The time then required to clean, copy, and transform the data for the methodologies required more time, giving us less time to experiment with model architectures.

The greatest limitation was running the models themselves. Because of the size of the data, particularly for method 2, training time for a single epoch ranged between 40-70 minutes, which further limited our ability to experiment and tune hyperparameters. The data also had to be run on Google Collab in order to access its GPU, as without it a single epoch would have taken 11 hours. While Collab's GPU did save time in that aspect, it had a tendency to crash in the middle of training, meaning that many training epochs were lost, wasting time in the long term. So much time was lost thanks to this that an entire model architecture had to be abandoned as there was no time left to tune and train it.

**Future Works**

Two areas for improvement in the future would be model performance and an expansion of use of this model by generating more complex risk-maps as well as comparing to existing GIS analysis tools. Model performance could be improved not just through more time to fine-tune the hyperparameters, but also through combining our approaches, taking into account both potential interaction between geographic locations and time periods, but pre-refining the features based upon a greater understanding of atmospheric phenomena. The features used could also be expanded, and satellite imaging might also be included in the dataset. Expanding the area analysis to include the entire continental United States, while further straining computation resources, would expand the usability of the model, allowing for nation-wide detection. Expanding the region used in training would also be likely to help model precision.

Besides improving our model’s hail detection capabilities, we could then expand on the use-cases for such a model by not only comparing results to existing FEMA risk maps, but also in existing GIS hail modeling technologies which rely solely on historical data and toolset algorithms.

**Appendix A: Contributions**

**Bobbie Cavna:** ERA Data Extraction, Method 1 Data Processing, Method 1 Model, Method 1 Testing

**Alden Jettpace:** NOAA Data Extraction, Method 2 Data Processing, Method 2 ConvLSTM model testing

**Satya Uma Praneetha Parupudi:** Method 2 Hyper Parameter Tuning, Presentation Slides

**Appendix B: Data Sources**

ERA Hourly Data on Pressure Levels: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form>

NOAA Storm Event Database: <https://www.ncdc.noaa.gov/stormevents/choosedates.jsp?statefips=18%2CINDIANA>

**Appendix C: Link to Project Folder**

<https://drive.google.com/drive/u/0/folders/1DL5QaBuUUKXzdFvZyHGum9yxdUfsShCE>

**Appendix D: Folder-Guide**

**RAW\_ERA5:** Raw ERA5 feature data. 144 files, each representing a month of measurements **Bulk\_Weather\_Events:** NOAA Hail-Event Data. 5 files, representing the hail events from 2011-2022 in Indiana, Ohio, Illinois, Kentucky, and Michigan

**ERA\_Numpy\_Files:** Transformed ERA5 feature data for Method 2. 144 files, each representing a month. Matches files-for-file, instance-for-instance with files in **NOAA\_Event\_Labels**

**NOAA\_Event\_Labels:** Transformed NOAA label data for Method 2. 144 files, each representing a month. Matches files-for-file, instance-for-instance with files in **ERA\_Numpy\_Files**

**Method\_1:** Holds files for Method 1, including data generation, model training, and analysis.

**Method\_2:** Hold files for generating files in **Bulk\_Weather\_Events** and **ERA\_Numpy\_Files**, as well as Method 2 models and analysis.

**Models\_and\_Histories:** Holds intermediate and final models for Method 2, as well as recorded training histories.

**Scrap\_Notes:** Contains scrap note-files for model, loader, and data transformation experiments

**References:**

[1] Cape Analytics. “Hail Risk: The Growing Threat for Property Insurers”. Capeanalytics.com. <https://capeanalytics.com/blog/hail-risk-for-property-insurers/#:~:text=In%20the%201990s%2C%20annual%20losses,than%20%243.3%20billion%20in%20damage> (accessed December 9, 2023)

[2] C. Zuzak *et. al.* “National Risk Index Technical Documentation” Federal Emergency Management Agency, Washington, DC, US, March 2023. Accessed: Dec. 9, 2023. [Online]. Available at: <https://www.fema.gov/sites/default/files/documents/fema_national-risk-index_technical-documentation.pdf>

[3] Climate Data Store. [Dataset Index]. <https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset> (accessed December 9, 2023)

[4] I. Lukyanenko, M. Mozikov, Y. Maximov, & I. Makarov. “Long-term hail risk assessment with deep neural networks.” *Advances in Computational Intelligence,* pp 288-301, Sept. 2023. Accessed: Dec. 9, 2023. Doi: arXiv preprint arXiv:2209.01191.

[5] D. Stang. *Youtube.* Available: <https://www.youtube.com/@davidstang2188/featured> (accessed December 9, 2023)