Paul Hughes

Final Report: Automating the Identification and Quantification of Adverse Weather Events for Crop-Insurance Claims

Problem Statement

Crop insurance provides financial protection to farmers against losses caused by adverse weather events (e.g., droughts, excess precipitation, heat, damaging freezes, wind, hail, etc.), disease, and price fluctuations. Weather-related payouts in crop insurance are triggered based on predefined weather indices, thresholds, or deviations from climatological normals. The payout amount is determined by the severity and duration of the weather event, the insured crop's susceptibility to the weather event, and the coverage level selected by the farmer.

In the event of a weather-related loss to an insured crop, the farmer has the great responsibility to: 1) file the claim within a specified time window following the occurrence of the weather event, 2) provide supporting documentation that details the cause and amount of crop loss, and 3) corporate in the investigation of the claim. If the farmer fails to initiate the claim within the designated time window, the claim may be denied. In addition, problems with the supporting documentation can significantly hinder the rate at which the claim is processed and resolved.

A component of the supporting documentation is the weather data associated with the adverse farming conditions. Therefore, can the identification and quantification (severity and duration) of adverse weather events be automated using anomaly detection algorithms to help farmers successfully submit insurance claims and receive payouts on their insured crops?

Data Wrangling

For this Proof-of-Concept study, two separate datasets were acquired from the National Oceanic and Atmospheric Administration (NOAA) National Centres for Environmental Information. The first dataset contained 30-year averages (1981–2010) and associated standard deviations of daily maximum and minimum air temperatures for site-specific weather stations. This study focussed on three specific weather stations: 1) Cedar Rapids, IA, 2) Kansas City, MO, and 3) Springfield, IL. The second dataset contained the daily maximum and minimum air temperatures for the year 2012 for the same three weather stations.

The climatological data was provided in four separate (.csv) files: 1) mean daily maximum temperatures, 2) standard deviation of mean daily maximum temperatures, 3) mean daily minimum temperatures, and 4) standard deviation of mean daily minimum temperatures. The following steps were taken to get the data in a useable format:

1. The mean and standard deviation values were provided as strings. However, before the values could be converted to floating points, the character 'C' needed to be removed.
2. Missing values were initially represented by '-8888.' This value was replaced by 'NaN.'

3. A provided scale factor needed to be applied to convert the data into real values.

The daily data was provided in .csv files. Each file contained 365 rows and 12 columns. The rows represent the days in a year. The columns contained information about the weather station location along with the temperature observations. Of the 12 columns, nine were eliminated, keeping only the timestamp, maximum air temperature, and minimum air temperature. The timestamp was then set as the index and the temperature data was converted from string values to floating points. Lastly, after viewing the air temperature time series, missing values were replaced using a linear interpolation method. Lastly, after reconfiguring the climatological data, it was combined with the daily data.

Modeling Results

The identification and quantification of daily maximum and minimum air temperatures was explored using three different anomaly detection techniques: 1) z-score, 2) Isolation Forest, and 3) One-class Support Vector Machine (SVM). The results of the first two techniques will be presented below. The one-class SVM model did not work for this application and the results for this model will not be discussed in this report.

Z-score

Figure 1 shows the results of using z-scores to identify anomalies in daily maximum air temperatures over the time period extending from April 1, 2012 to November 31, 2012. This time period represents the general growing season of corn from planting to harvest. The red dots indicate when the maximum air temperature for that particular day exceeded two standard deviations from the corresponding 30-year climatological mean value (which is shown in orange). The clustering of red dots around day 200 (which is July 18th) corresponds to the timing of the June/July record heat wave that impacted a large portion of the United States. Figure 2 shows similar results for the daily minimum air temperatures. Again, the majority of the anomalies identified by z-scores are clustered around day 200, aligning with the timing of the heat wave. Outside of the June/July time period, there are relatively few maximum and minimum air temperatures that are identified as anomalies, and the duration of these events is typically only a single day.

Isolation Forest

Figues 3 and 4 show the results of using the Isolation Forest technique to identify anomalies in the same daily maximum and minimum temperature time series as shown in Figures 1 and 2. For the daily maximum temperatures (Figure 3), the Isolation Forest technique identified a clustering of anomalies around day 200 at all three locations, aligning again with the June/July 2012 heat wave as seen before. Figure 4 shows similar results for the daily minimum air temperatures. The Isolation Forest technique again identified a clustering of minimum-air-temperature anomalies centered around day 200, aligning with the timing of the heat wave. However, unlike the z-score technique, the Isolation Forest method identified a second clustering of anomalies. Figures 3 and 4 show this second cluster of anomalies occurring after day 300, which corresponds to the month of November.

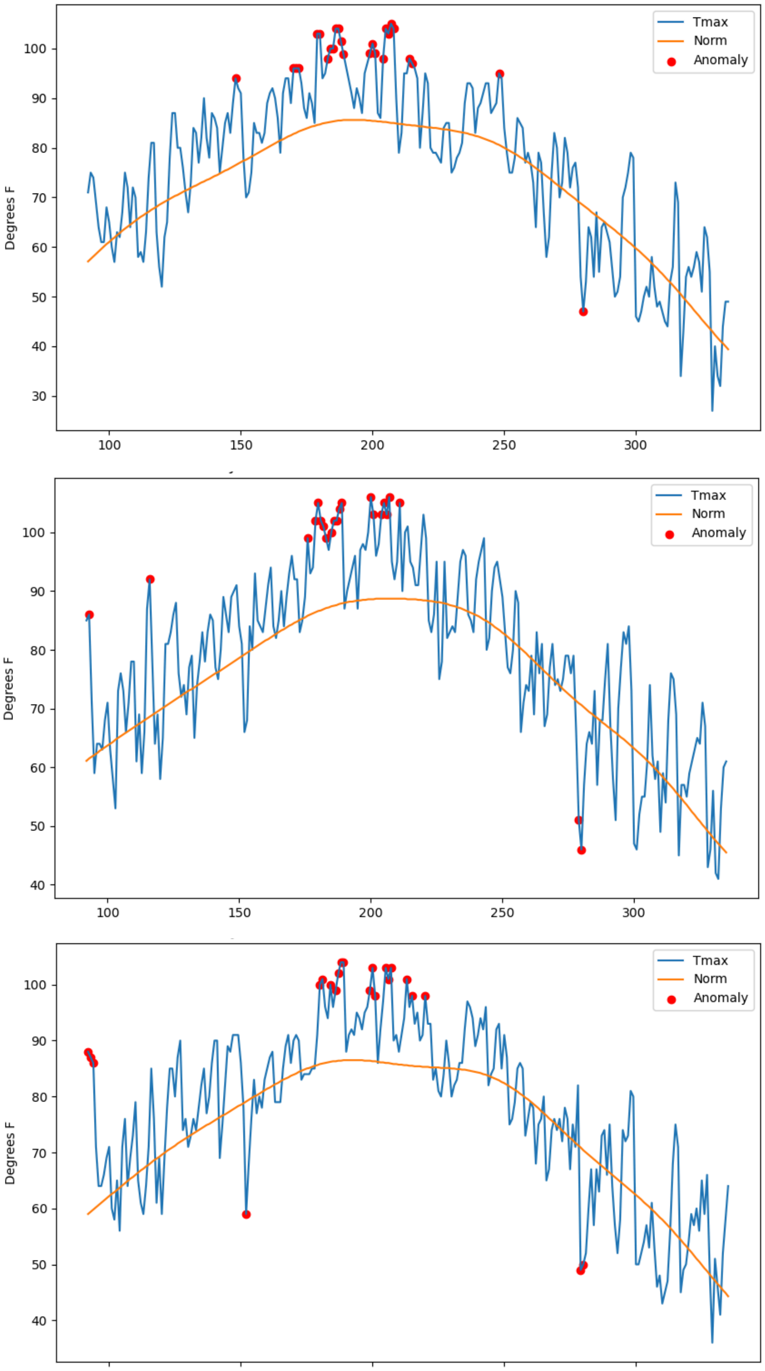

Figure 1. Time series of daily maximum air temperature for the period extending from April 1, 2012 to November 30, 2012 for (top) Cedar Rapids, IA, (middle) Kansas City, MO, and (bottom) Springfield, IL. The blue line represents the daily maximum air temperature, the orange line represents the 30-year mean of daily air temperature, and the red circles represent the daily maximum air temperatures that were identified as anomalies using z-scores (greater than two standard deviations away from the 30-year mean).

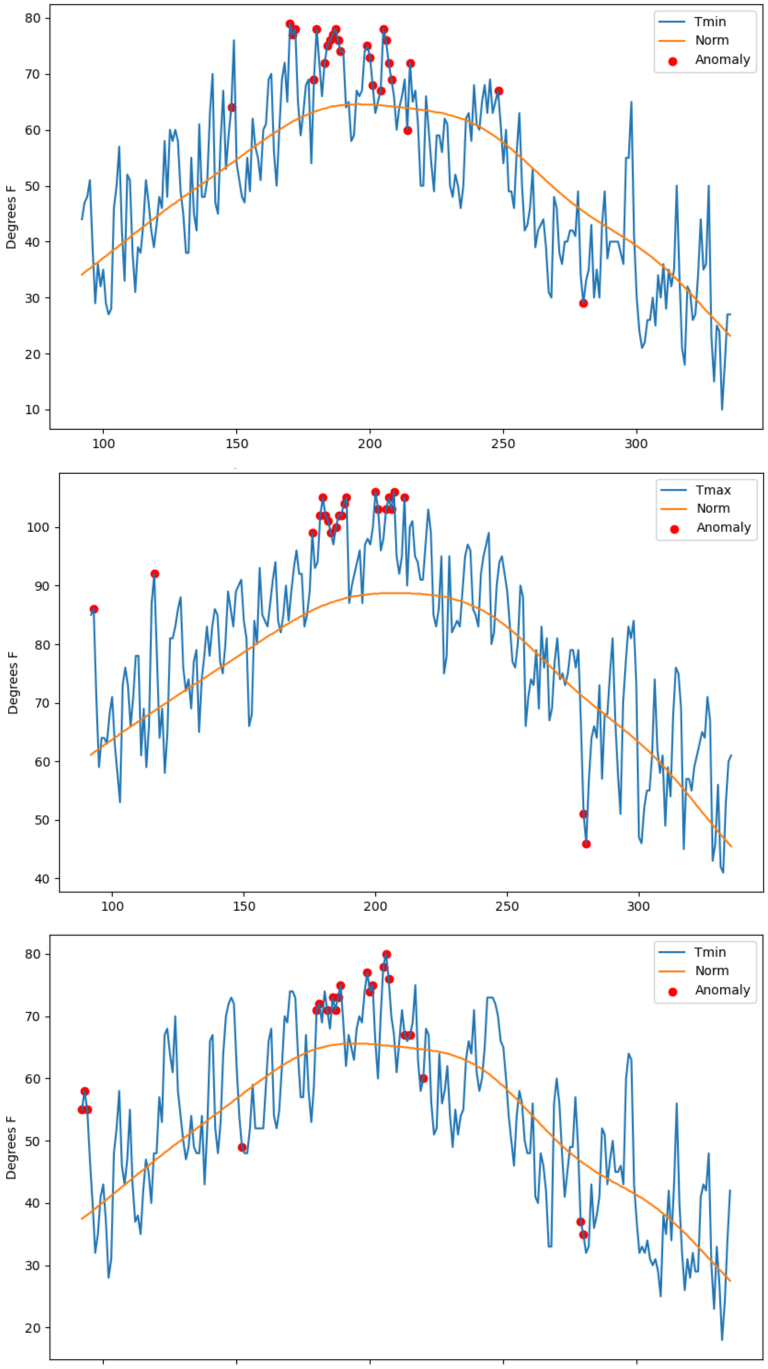

Figure 2. As in Figure 1 expect for daily minimum temperature.

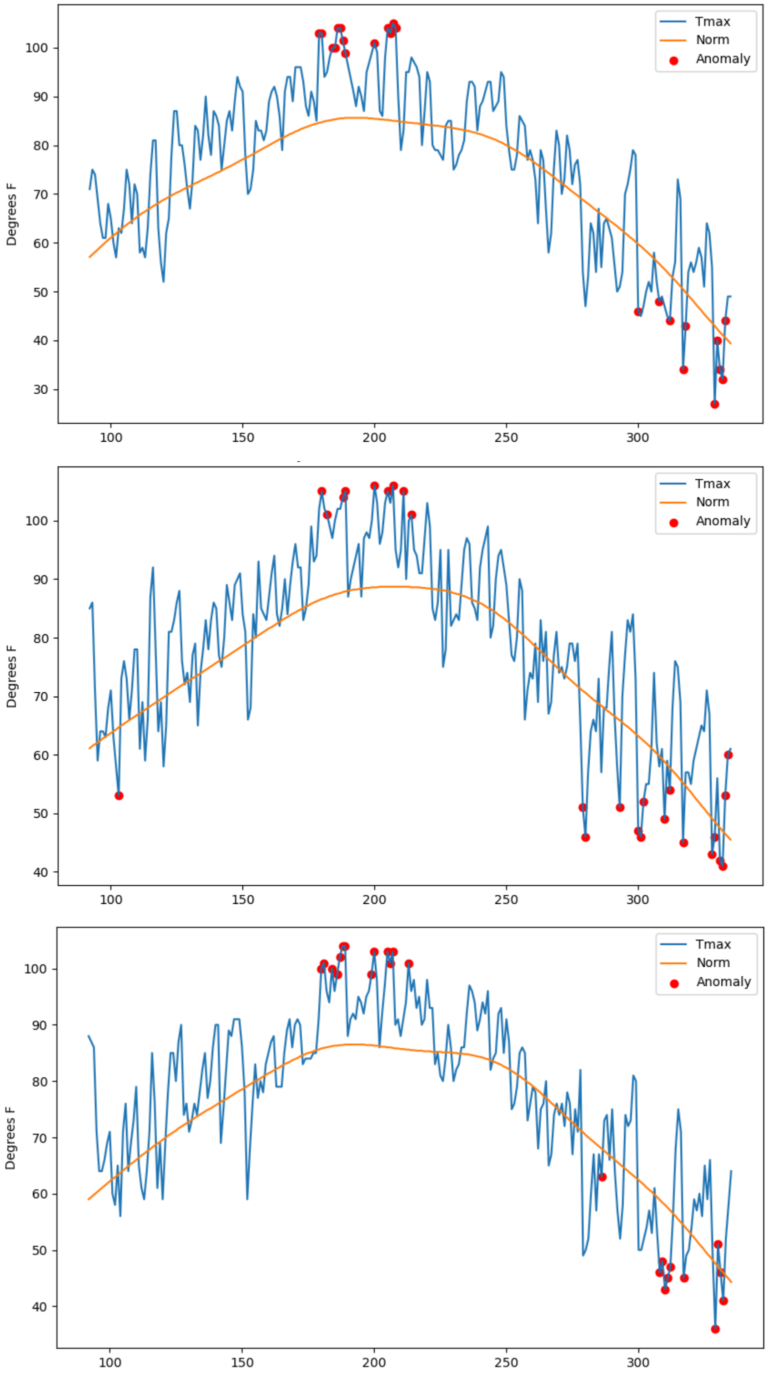

Figure 3. As in Figure 1 except the red circles represent the daily maximum air temperatures that were identified as anomalies using the Isolation Forest algorithm.

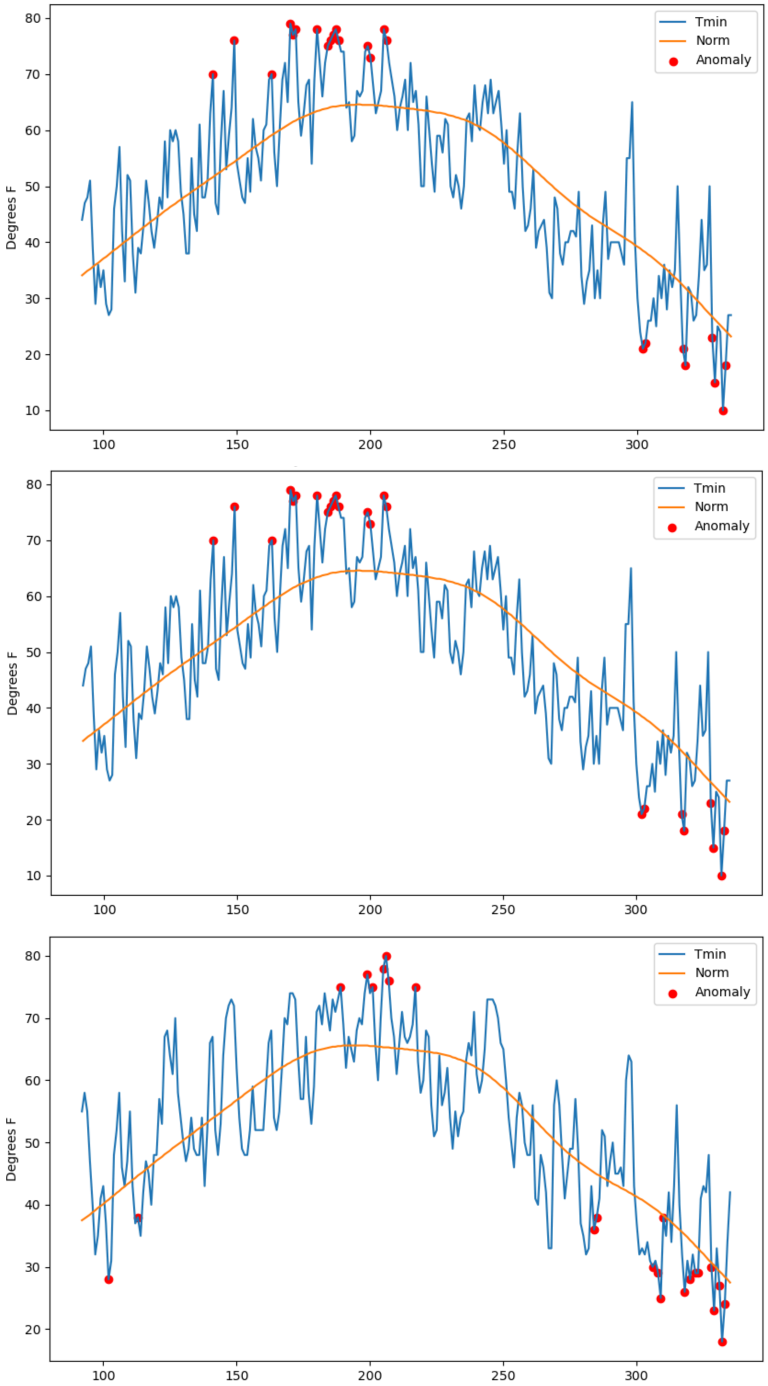

Figure 4. As in Figure 3 except for daily minimum air temperature.

Conclusions

Overall, this study validated the Proof of Concept, showing that the identification and quantification (of severity and duration) of adverse weather can be automated using anomaly detection algorithms to help farmers successfully submit insurance claims and receive payouts on their insured crops. At this point, more work is needed to understand why the Isolation

Forest technique identified a second cluster of anomalies and the z-score method did not. Additionally, the physical significance of this second cluster with respect to crop health needs to be better understood. Only after exploring the second cluster further can a model be given preference.

Next Steps

Once a model is given preference, the immediate next steps of this project will be as follows:

1. Incorporate other meteorological variables (e.g., precipitation, wind speed, growing degree days, humidity, etc.)
2. Incorporate thresholds for identifying anomalies that physically connect to the needed growing conditions for particular crops.