Predicting Lightning Strikes Through Generalized Linear Mixed Model and Binary Hurdle Model

Sarah Yang, sarahxy2@illinois.edu, Mallory Klostermann, mallory6@Illinois.edu

Introduction

The aim of this project is to create two different models for predicting lightning in the northern Midwest region of the United States, with one model predicting whether there will be lightning present in an area and another model predicting the number of strikes.

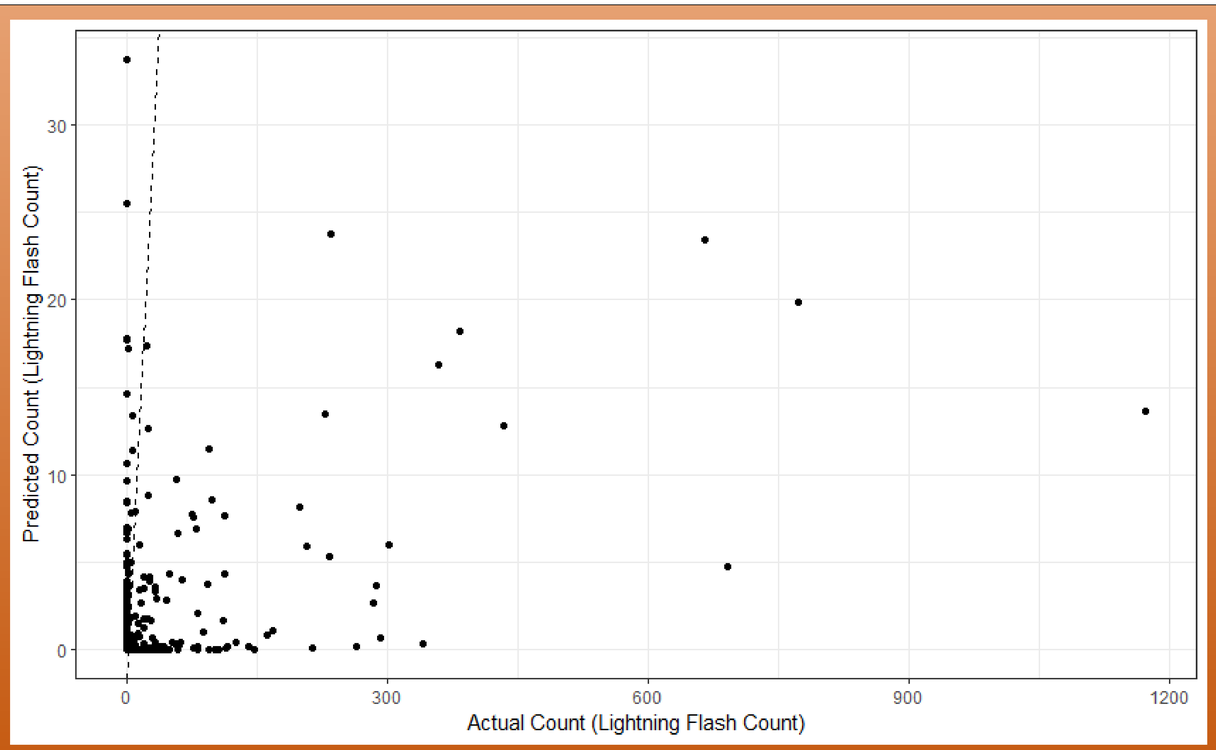
Analyses

Generalized Linear Mixed Logistic Regression Model:

- To incorporate the most recent atmospheric conditions into the predictions, each prediction was made with a model trained on all previous data, up through the hour before the time to be predicted
- With this method, there was a mean sensitivity and specificity of 70.42% and 83.08%, respectively

Hurdle Model:

- The binary hurdle model predicts the probability of observing a non-zero count, while the truncated count model models the distribution of non-zero counts
- Assumes that the probability of observing a non-zero count is influenced by a set of covariates with decreasing correlation as time passes in each grid, we use a zero-inflated negative binomial model to predict the actual lightning count for the next 24 hours



Methods

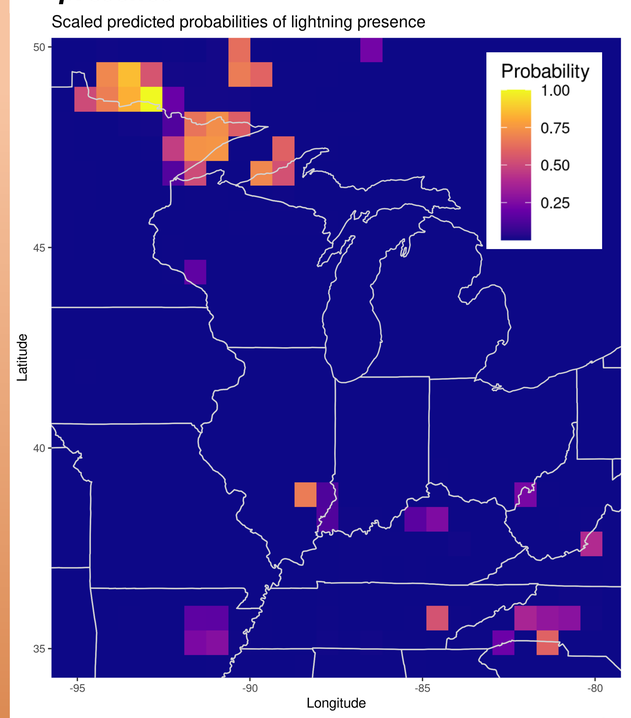
Generalized Linear Mixed Logistic Regression Model:

- Fixed effect for the atmospheric radiance measured at each location over time
- Random effect for geographic location with AR(1) covariance structure

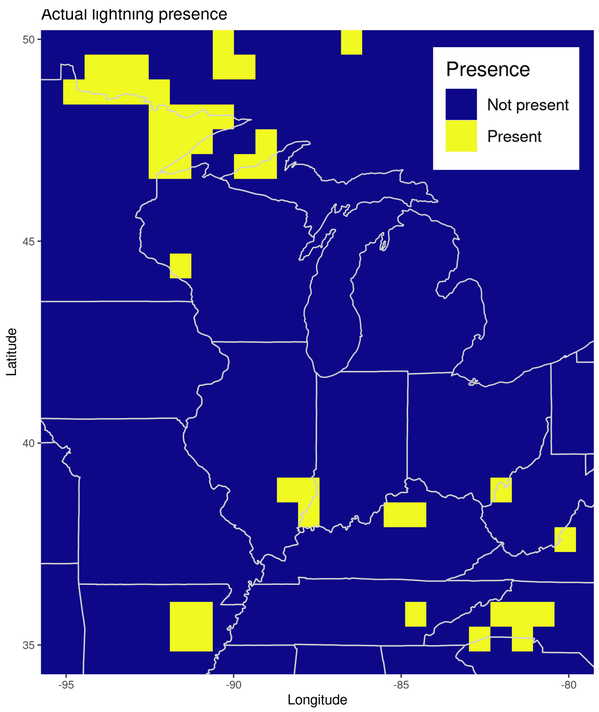
Hurdle Model:

- Binary part modeling the non-zero probability of a lightning observation
- Continuous part modeling distribution of non-zero lightning observations

Scaled predicted probabilities of lightning presence



Actual lightning presence



The above plots are for July 6th, 2020 at 2:00 AM.

Next Steps

Severely unbalanced data led to difficulties predicting lightning presence in the logistic model. Resampling techniques for time series data could be explored to avoid the low probability threshold necessary and improve model performance.

Currently, the models are only trained on a few days of data (and are therefore only able to meaningfully predict within a few days). In the future, adding more training data will allow the models to make better predictions for a much larger range of dates.