CSC 293 - Final Project
Rainfall Anomaly Detection in MacLeish
Elaine Chen, Ashley Qian

Background: Due to the design of the MacLeish station's rainfall gauge, snow can accumulate in the funnel and get temporarily stuck. When the snow later melts, it may be incorrectly recorded as rainfall on days when it neither rained nor snowed, leading to errors in the measured rainfall data. So we aim to build a model that identifies such anomalies.

Data origin: We will use five years of daily data (2019–2023) to train and evaluate the model provided by the CEEDS.The model will also rely on precipitation and snow data from the Leverett Number 2 station from the National Centers for Environmental Information as ground truth, since that station distinguishes between rain and snow.

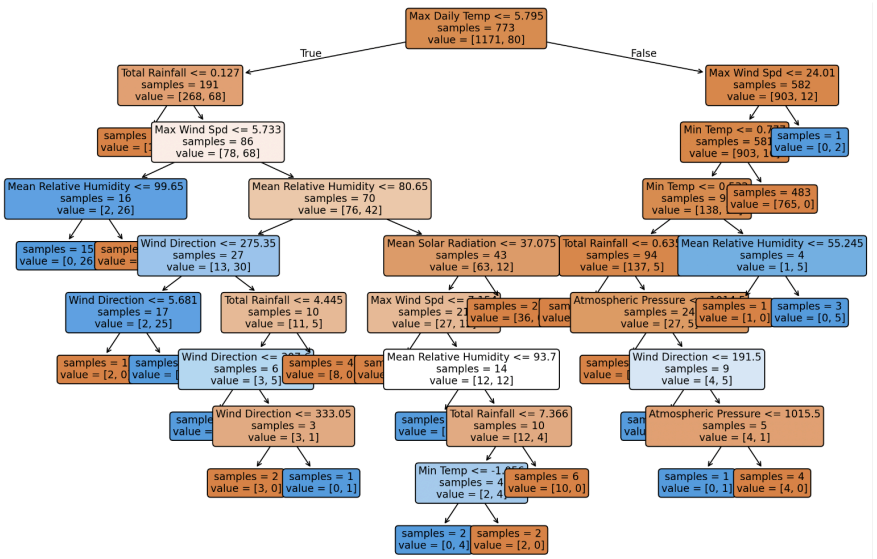
Result:

Data	Model	Lag Features	Classification Error
All-Year	Bagging	No	3.72%
All-Year	Random Forest	No	3.72%
All-Year	Bagging	Yes	5.04%
All-Year	Random Forest	Yes	5.22%
Winter	Bagging	No	8.24%
Winter	Random Forest	No	9.74%
Winter	Bagging	Yes	7.89%
Winter	Random Forest	Yes	9.77%

(Note: winter data spans from November to April; lag features indicate whether there is snow or rain one day ago, two days ago, and three days ago)

Upon reviewing the table above, we notice that bagging models almost always demonstrate a lower classification error when compared to random forest models. Incorporating the lag features generally increases the classification error, except for bagging models trained on winter-only data. The lag features also demonstrate fairly low importance across all models (below 0.01). Additionally, models trained on winter-only data generate higher classification errors than those trained on all-year data.

The lowest classification error among all models is 3.72%, achieved by the bagging and random forest models trained on all-year data without lag features. The most important features of both models are identical: Total Rainfall, Mean Relative Humidity, and Min Temp, though the ranking slightly differs.



	Coefficient	Std. Error	z-statistics	p-value
Intercept	-20.8374	15.444	-1.349	0.177
Min Temp	-0.1412	0.014	-10.186	0.000
Wind Direction	-0.0034	0.001	-3.636	0.000
Mean Relative Humidity	0.0276	0.007	3.899	0.000
Atmospheric Pressure	0.0165	0.015	1.104	0.270
Total Rainfall	0.0103	0.011	0.967	0.333

The table above shows the logistic regression results using the most important features (importance factor ≥ 0.1) as predictors. The result is relatively intuitive. For example, the increase in minimum temperature will decrease the anomaly probability since the melting speed will increase, and there will be less snow stuck in the gauge. Based on these coefficients, our final equation for the model is:

$$\text{logit}(p) = -20.8374 - 0.1412 \times (\text{Min Temp}) - 0.0034 \times (\text{Wind Direction}) + 0.0276 \times (\text{Mean Relative Humidity}) + 0.0165 \times (\text{Atmospheric Pressure}) + 0.0103 \times (\text{Total Rainfall}).$$

Future work: We can improve this model in several possible ways.

1. **Time series models**: Since the data is time-based, applying models designed for time-dependent data, like LSTM or ARIMA would be more logical.
2. **Collinearity**: There might be collinearity within the predictors, for example, minimum/maximum temperature and average temperature.
3. **Anomaly definition**: The current definition of anomalies includes a restriction on minimum temperature to filter out summer rainfall events, considering the ~10-mile distance between MacLeish and Leverett. However, CEEDS could further refine this definition.

Data Appendix

Column Name	Unit	Meaning
TIMESTAMP	YYYY/MM/DD HH:MM	Date and time of the observation
Average Temp	°C	Daily average temperature
Max Daily Temp	°C	Maximum daily temperature
Min Temp	°C	Minimum daily temperature
Wind Speed	m/s	Average wind speed
Wind Direction	Degrees (°)	Average wind direction
Max Wind Speed	m/s	Maximum wind speed
Min Wind Speed	m/s	Minimum wind speed
Mean Relative Humidity	%	Daily average relative humidity
Atmospheric Pressure	mb	Atmospheric pressure
Mean Solar Radiation	W/m²	Average solar radiation
Total Rainfall	mm	Total rainfall recorded by MacLeish station
Precipitation	mm	Precipitation recorded by Leverett Station No. 2
Is_rain	1/0	1 if precipitation > 0, else 0
Snow	mm	Snowfall recorded by Leverett Station No. 2
Is_snow	1/0	1 if snowfall > 0, else 0
Is_anomaly	1/0	1 if Precipitation == 0 and Snow == 0 and Total Rainfall ≠ 0 and Min Temp < 1°C