Using Machine Learning for Weather Prediction at a Local Scale
Kate Anderson and Logan Becker
Professor Etai Roth
2 May 2025

Introduction/Abstract

Weather forecasting is a vital tool used worldwide to support decision-making for individuals, businesses, and governments (Fathi et al., 2022). These decisions range from daily choices, like what to wear, to critical actions such as evacuation planning (Fathi et al., 2022). Forecasting is complex and typically involves recording weather parameters like temperature, humidity, and wind speed at regular intervals, then analyzing the data with models to detect patterns and make predictions (Fathi et al., 2022). Local weather prediction is especially challenging due to high uncertainty and error rates (Fathi et al., 2022). As a result, most studies focus on regional forecasting using large datasets. However, we believe that accurate local forecasting is possible. We hypothesize that meteorological data can predict precipitation amounts in Albany, NY, resulting in an R^2 value greater than 0.50.

To test this, we used daily weather data from the NOAA National Center for Environmental Information, focusing on Albany, NY. We applied linear regression to predict precipitation amounts based on features such as humidity, temperature, and other variables. The initial model showed low error but explained only a small amount of the variance in precipitation. We then tested the same approach on predicting temperature departures from the average. This model had a higher error but captured more variance. These results highlight the difficulty of predicting precipitation amounts at a local level but also suggest that predicting temperature locally may be more feasible.

Methods

Our dataset consists of 2,668 entries from a single weather station (Station ID: 72518014735) in Albany, NY, which records data at 11:59 PM daily from January 1st, 2025, to May 31st, 2022. For the purposes of our model, we utilized all measurable weather variables or events in the dataset, including factors like humidity, temperature, wind speed, sunset, and sunrise. Non-weather variables such as station and date were excluded (to name a few), as were daily snowfall and snowfall depth, since these are already included in the daily precipitation totals. Our goal was to predict precipitation amounts based on other meteorological measures as features.

We chose to use linear regression because it works well for continuous variables like weather data. However, weather patterns can often be nonlinear, especially over extended periods. In this case, we assumed that the relationship between the weather features and the target variable was linear for the model. The models were implemented using Python in VS Code, relying on the sklearn machine learning package. We began by creating a correlation matrix to identify significant relationships between the features and the target variable. Since no highly correlated features were found, we included all measurable weather variables in the final model. To prepare the data, we scaled the features and then began training. To assess the model's performance, we split the data into training (80%) and test (20%) sets. After training the model, we predicted on the test data and evaluated its performance using Mean Squared Error and R^2. These metrics allowed us to evaluate how well the model predicted precipitation amounts based on the weather variables. After observing a low accuracy performance of the initial model, we adjusted our target variable to focus on temperature departures from the average, which we hypothesized might lead to better predictions.

Results

Precipitation amounts in Albany, NY, from 2015 to 2022 averaged 0.11 inches per day with a standard deviation of 0.27 inches. During this time, the daily precipitation ranged from 0 inches to 3.92 inches in one day. Despite the low average, there is significant variation in daily precipitation amounts (Figure 1). The highest spikes generally occur in the summer and early fall months (Figure 1).

Our first linear regression had a mean squared error of 0.05 and an R^2 value of 0.32 when predicting precipitation amounts using other meteorological measurements as features (Figure 2). Our second linear regression had a higher mean squared error (MSE = 21.26) and a higher R^2 value ($R^2 = 0.65$) than the first when predicting departure from the normal average temperature (Figure 3).

Discussion

Meteorological data did not predict precipitation amounts with an R^2 value greater than 0.50, so we reject our hypothesis. Weather forecasts are often ridiculed by the public for inaccuracy, but weather prediction, especially numerical, is challenging (Meng et al., 2024). In our study, we were unable to explain much of the variance in the data using a simple linear regression despite having low error (Figure 2). We also attempted to use Elastic Net and Polynomial regressions, but these models provided worse results than the linear regression. Most likely, accurate numerical precipitation predictions would have required more complicated models beyond the scope of this project. For example, the Weather Research and Forecasting (WRF) model is commonly used to predict precipitation with some accuracy (Meng et al., 2024). Meng et al. (2024) found that adding other methods, such as the adaptive noise-robust empirical mode decomposition (CEEMDAN) method, helped improve the accuracy of the WRF model. More research into predicting precipitation amounts is still required, but evidently, some models can explain more variance in precipitation data than our linear regression was able to.

Predicting the departure from the normal average temperature in Albany, NY, using a linear regression was more successful at predicting the variance in the data, but had a higher error than when predicting precipitation amounts. These data have significant implications for predicting and understanding the local impacts of climate change. However, one limitation to this result is that the model used other temperature records in the dataset to make the prediction. Even still, the fact that these data fit a linear regression implies that temperatures are rising in Albany. The high error in this model indicates significant variation in the departure from the normal average temperature on a day-to-day basis, but there is an upward trend overall. This change will likely have implications for other weather events and temperature-reliant activities.

GenAI Statement and Contributions

The help of Copilot was used in the creation of our code, and Grammarly was used to catch spelling errors. Logan did the introduction and methods, and Kate did the results and the discussion. Logan and Kate both contributed to finding sources and coding figures and results in VS Code.

Citations and Attributions

Link to dataset:
https://www.kaggle.com/datasets/die9origephit/temperature-data-albany-new-york

Climate data—New york state. (2023). Retrieved April 29, 2025, from https://www.kaggle.com/datasets/die9origephit/temperature-data-albany-new-york

Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E. (2022). Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering*, *29*(2), 1247–1275. https://doi.org/10.1007/s11831-021-09616-4

Meng, C., Hu, Z., Wang, Y., Zhang, Y., & Dong, Z. (2024). A forecasting method for corrected numerical weather prediction precipitation based on modal decomposition and coupling of multiple intelligent algorithms. *Meteorology and Atmospheric Physics, 136(5)*, 32-. https://doi.org/10.1007/s00703-024-01030-2

Figures

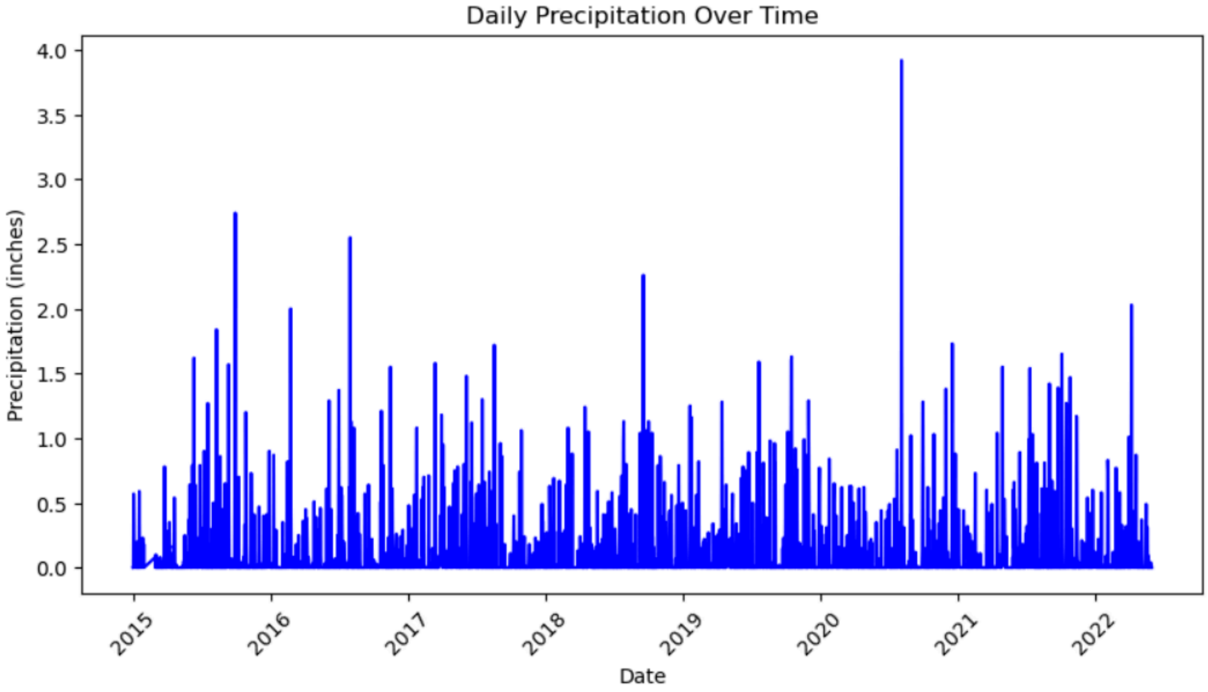

Figure 1: Plot demonstrating the daily precipitation amounts (inches) from 2015 to 2022 in Albany, NY.

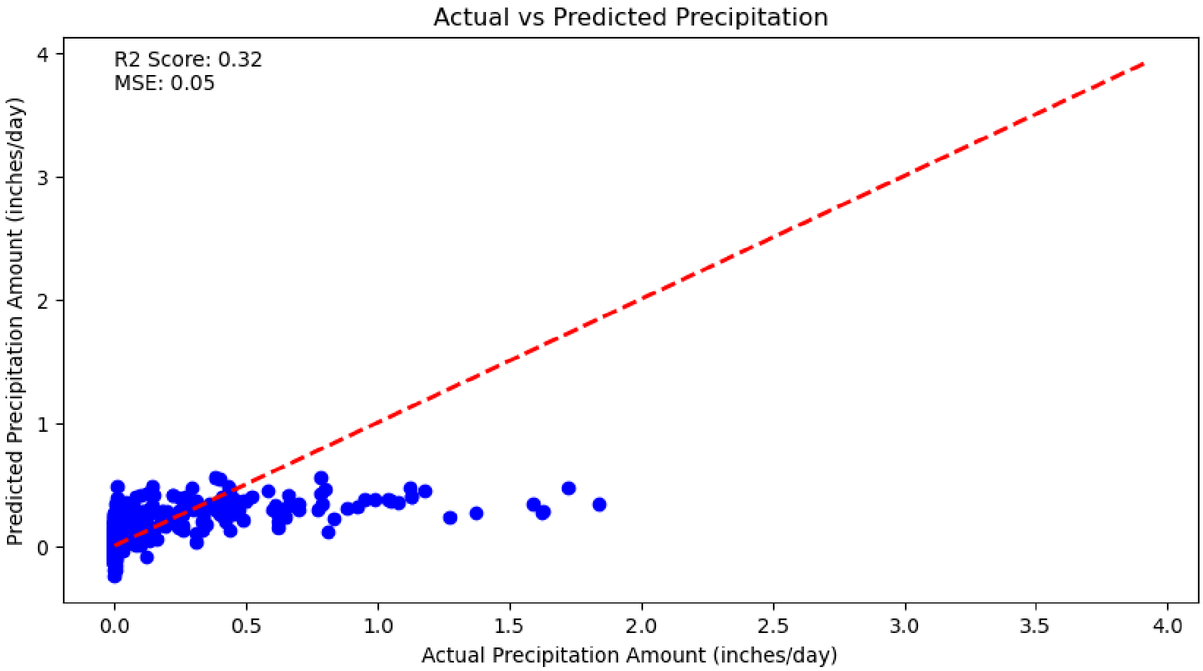

Figure 2: Linear regression predicting the amount of precipitation (inches) in one day using other meteorological measurements, including temperature, dew point, and humidity. Data were collected from 2015 to 2022 in Albany, NY.

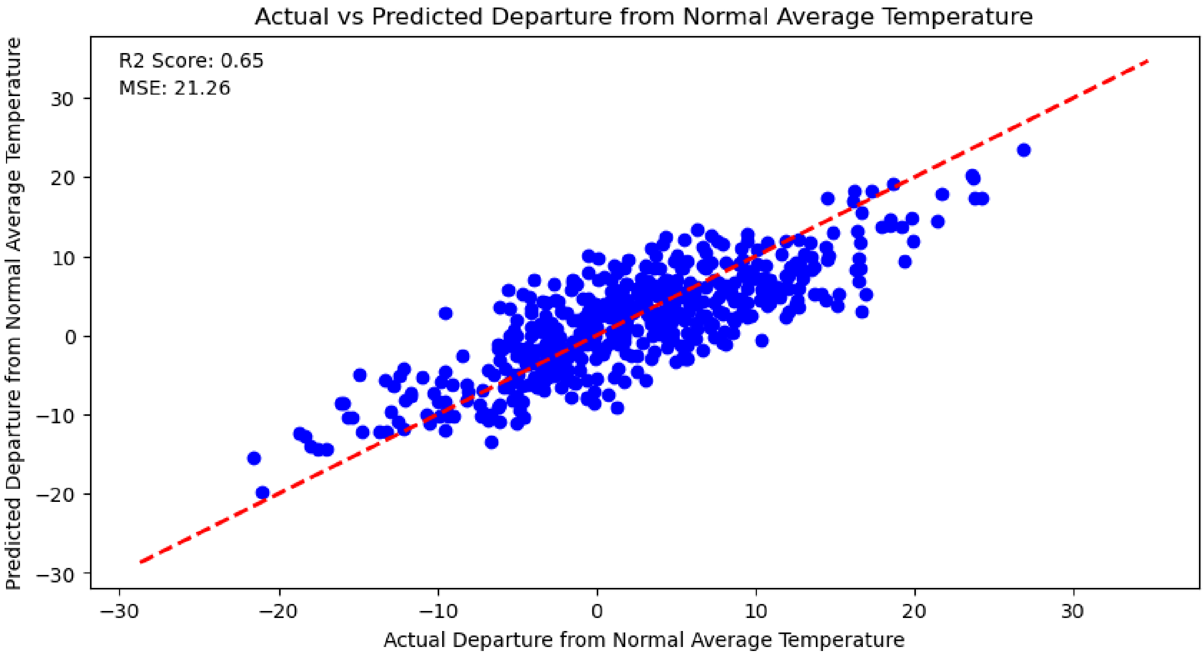

Figure 3: Linear regression using other meteorological measurements such as precipitation, dew point, and humidity to predict the departure from the normal average temperature. Data were collected from 2015 to 2022 in Albany, NY.