

Temperature and humidity prediction for three locations of North America

Mihail Batura, Group 6, F10

Github: <https://github.com/MihBat/F10-Weather-prediction>

Kaggle data: <https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data/data>

Open Meteo data: <https://open-meteo.com/>

Background

Assume that some businesses in southern Canada experience financial losses due to inaccurate winter weather forecasts. Current forecast accuracy is approximately 75% for humidity and 80% for temperature. These errors directly affect energy costs during the heating season, because in coastal areas temperatures can vary from -5°C to 10°C . Warehouses must maintain stable temperature and humidity levels to protect stored goods, making better forecasts essential. The goal is to develop a more accurate model that can reduce winter losses by approximately 10–15%.

Goals

- Select and compare suitable machine learning methods for weather forecasting
- Train the models to capture weather patterns in North America
- Evaluate performance on actual observations and achieve $\text{RMSE} < 1.5^{\circ}\text{C}$ for temperature and $\text{RMSE} < 5\%$ for humidity

Data

Historical hourly weather data for Montreal, Toronto, and Dallas (2013–2016) from Kaggle was used for model training. Hourly weather data from Open-Meteo for 2023 served as the test dataset. Dallas was included as a contrasting location due to its more continental and variable southern climate relative to southern Canada.

Methodology

Two machine learning models were used for weather forecasting in this study: a Random Forest regressor and a Multilayer Perceptron (MLP) neural network. Both models were trained using historical hourly weather data from Montreal, Toronto, and Dallas. The feature set included variables (temperature, humidity, pressure, wind speed/direction) as well as engineered time-related features such as hour, month, cyclical month encoding (sin/cos), and temperature lag values (1-hour, 24-hour, and 24-hour rolling mean). Model performance was evaluated on 2023 test data using two standard metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

Results and discussion

Overall, both machine learning models produced fairly accurate forecasts. For temperature, MAE ranged from 0.8°C to 1.3°C , and for RMSE, from 1.1°C to 1.7°C . For humidity, MAE ranged from 4% to 5%, and for RMSE, from 5.5% to 6.6%. It means that both models usually were off by 1°C in predicting temperature and 5% in predicting humidity. Prediction of both variables was more accurate for Montreal and Toronto than for Dallas.

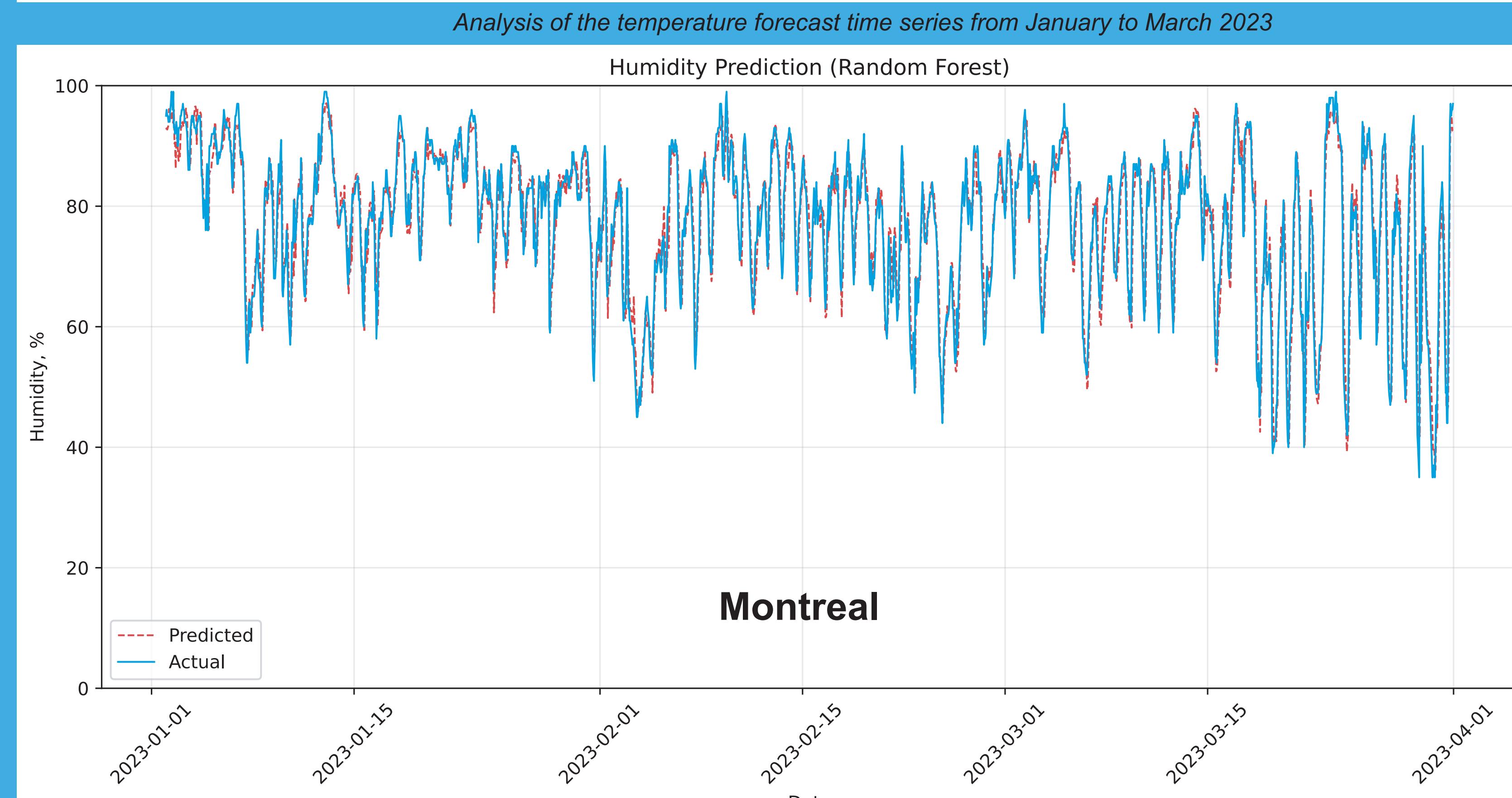
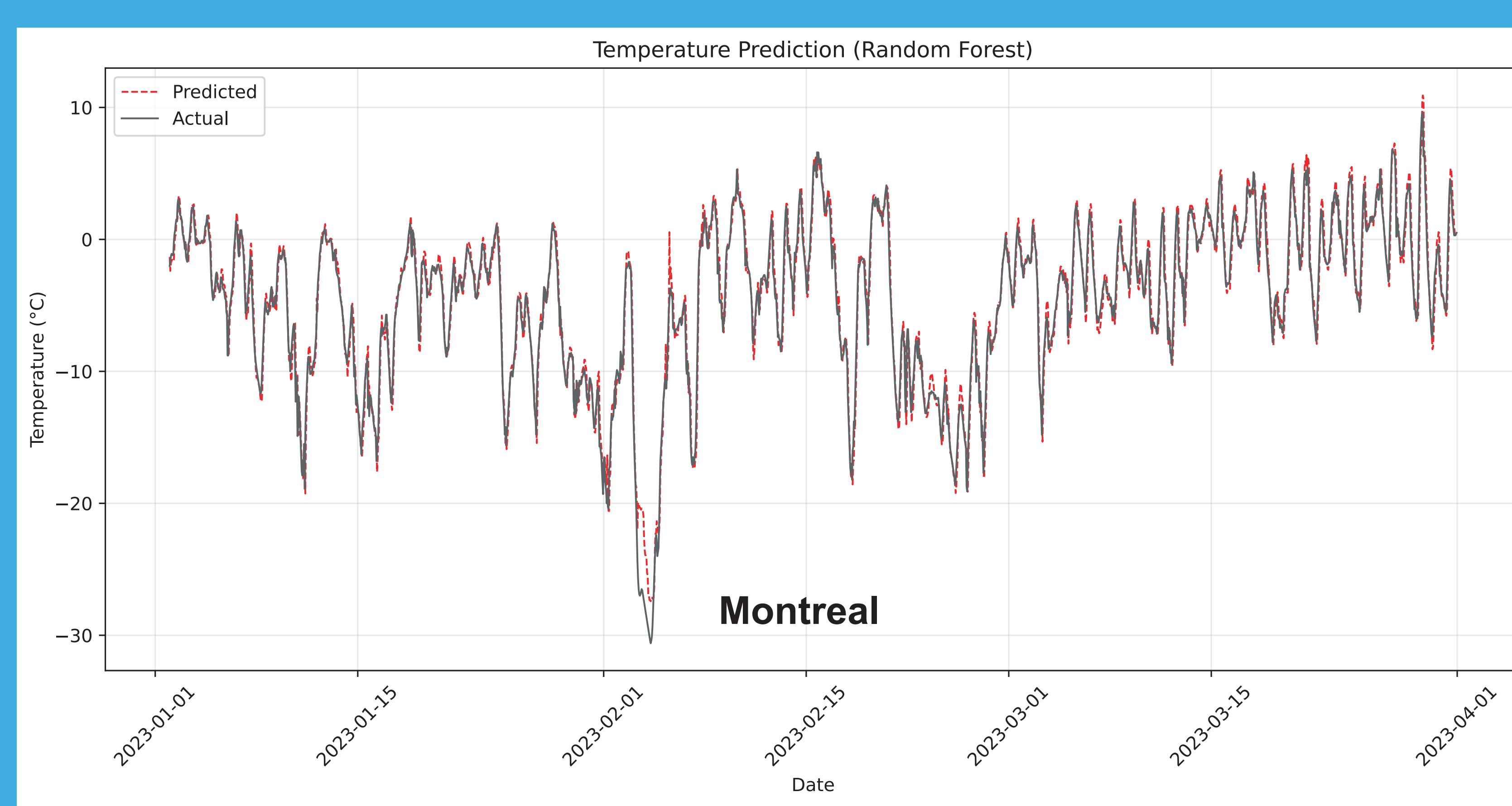
The inclusion of lag features (e.g., 1-hour temperature lag, 24-hour lag, and 24-hour rolling mean) substantially increased forecasting accuracy. These features allowed the models to learn temporal dependencies and the strong daily cycles present in temperature data.

A less accurate forecast for Dallas may be due to more variable weather conditions, as atmospheric flows from the Gulf of Mexico toward the southeastern United States significantly influence the forecast. However, since Dallas is not a target location and was included in the sample for contrast, its forecast accuracy will not be considered in the context of achieving the targets.

Humidity forecasting was generally less accurate than temperature forecasting. This is expected, as humidity is influenced by additional atmospheric processes such as vertical moisture transport, cloud formation, and precipitation. Including precipitation or other moisture-related variables could improve humidity predictions, as these parameters provide direct information about atmospheric water content.

Conclusions

- Random Forest performed slightly better in predicting temperature, while MLP performed better in predicting humidity in Montreal and Dallas
- The RMSE target of $<1.5^{\circ}\text{C}$ was achieved for temperature forecasting in Montreal and Toronto
- More atmospheric variables are likely needed to more accurately predict humidity



model	location	RMSE_temperature_°C
Random Forest	Toronto	1.129
Random Forest	Montreal	1.173
MLP neural network	Montreal	1.207
MLP neural network	Toronto	1.223
Random Forest	Dallas	1.621
MLP neural network	Dallas	1.737

Temperature prediction comparison of two models

model	location	RMSE_humidity_%
MLP neural network	Montreal	5.387
Random Forest	Montreal	5.525
Random Forest	Toronto	5.838
MLP neural network	Toronto	6.190
MLP neural network	Dallas	6.326
Random Forest	Dallas	6.590

Humidity prediction comparison of two models