Accelerating Brazil’s Weather Forecasting with CHTC

1. Introduction

This project aims to predict Brazilian weather using statistical modeling techniques. Hourly Brazilian weather dataset is used and time series and machine learning techniques are applied to build predictive models for key meteorological indicators. CHTC is used to accelerate both training and forecasting.

This report first describes the dataset and preliminary analyses, then introduces the ARIMA-Random Forest model, and concludes with a comparative evaluation demonstrating how CHTC enhances performance.

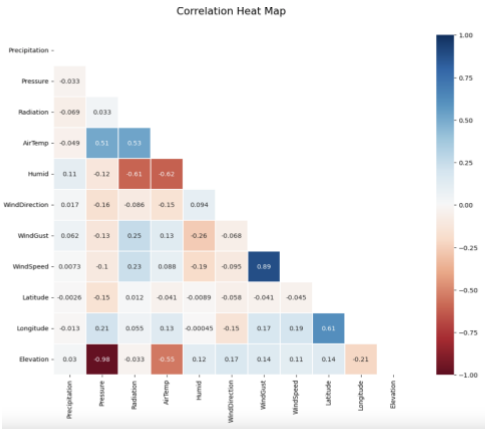
2. Core Analysis and Results

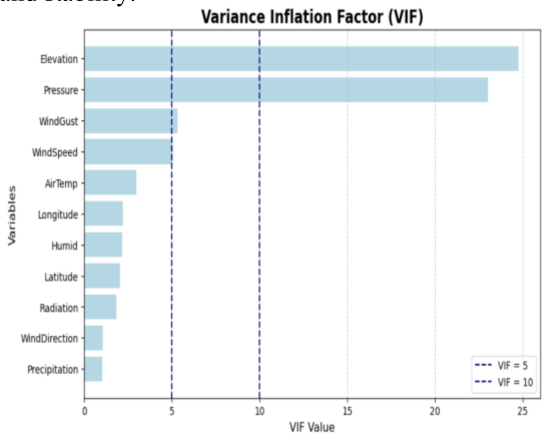
2.1 Data Description

The dataset is derived from hourly observations recorded by weather stations distributed across Brazil between 2000 and 2021. It encompasses approximately 10.11 GB of data, including 27 variables—such as temperature, humidity, precipitation, and location.

2.2 Data Preprocessing

In the data preprocessing stage, missing values are first imputed using the mean of the nearest neighboring values in time. This approach helped ensure the continuity and consistency of the dataset. Preliminary analysis is conducted using a correlation heatmap and a Variance Inflation Factor (VIF) plot. These analyses revealed that certain variables, such as Elevation, Pressure, and WindGust, exhibit high multicollinearity. This finding highlighted the need for careful consideration when selecting variables for the model, as multicollinearity could affect the model’s accuracy and stability.



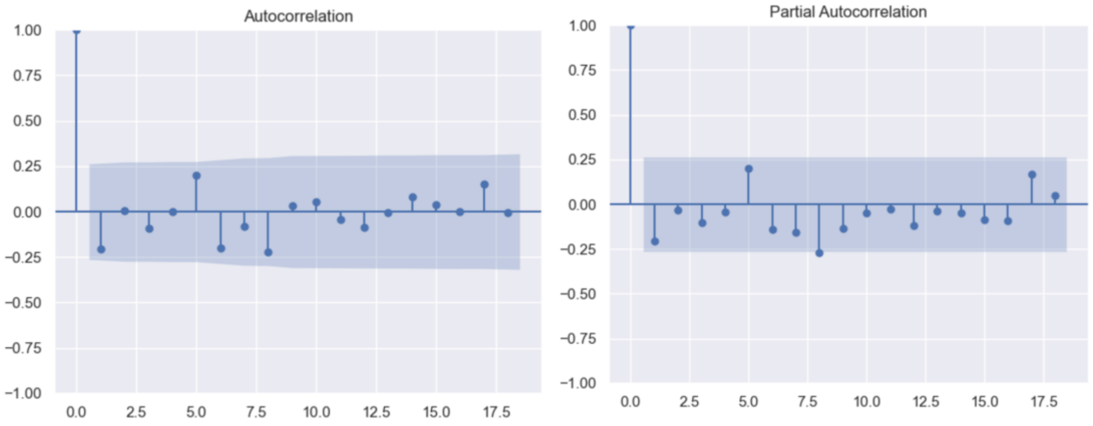


Meteorological data was then aggregated using averages to represent state-level conditions over specific periods, preparing for parallel computing.

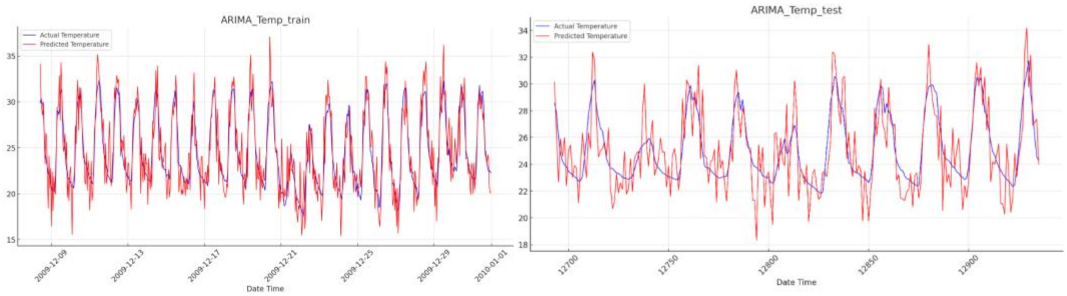
2.3 Statistical Modeling & Results

Key variables, including temperature, pressure, precipitation, and humidity, were selected for training. Auto-ARIMA was applied to those with clear time series patterns using a 7:3 train-test split.

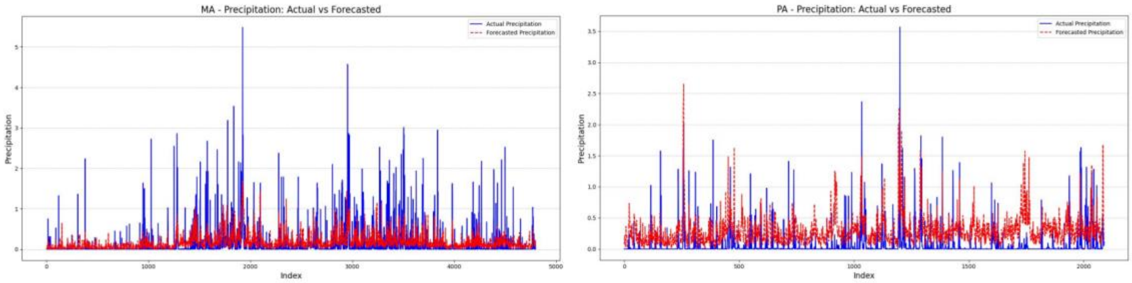
Taking temperature as an example, the ACF and PACF plots show the data has significant autocorrelation at lag 1 and a clear pattern, making it suitable for ARIMA modeling to capture these dependencies and forecast future values.



The model performs well on the training set and the test set, as shown in the figure below.



For variables with weak time-series characteristics, such as precipitation, a random forest model is employed for prediction. Random forests are robust to multicollinearity, thus effectively addressing the issues identified during preliminary data analysis. By incorporating the time-series forecasts of other variables into the trained random forest model, precipitation predictions can be obtained.



The final results can be expressed as follows.

Variable	**Model**	**MSE**	**MAE**
Pressure	ARIMA	0.13373568771117	0.24706566148439
Radiation	ARIMA	77407.9914380233	192.569016059212
AirTemp	ARIMA	0.629315326056324	0.548305677088805
Humid	ARIMA	8.95212648043136	1.95757175699927
WindDirection	ARIMA	5289.97586500201	56.5128957072755
WindGust	ARIMA	1.06925726087981	0.670525675738969
WindSpeed	ARIMA	0.253930594851785	0.355909992507996
Precipitation	Random Forest	0.227573903729218	0.140765488556314

2.4 Parallel Processing

On CHTC, data was divided into five geographical regions for parallel data preprocessing, and parallel processing was also applied during model training and prediction.

Parallel computation was also implemented in R when training ARIMA model with different variables, accelerating training but increasing resources demand.

	Local Machine	Single Node on CHTC	Multiple Nodes on CHTC
Data	RJ (34 MB for test)	ES (lager than RJ, have been held because didn't have enough resources)	26 datasets
Resource	Memory: 224.4MB	CPU:1 Disk:5243904KB Memory: 5120MB	Requested: CPU: 1 Memory: 5GB Disk: 5GB
Run Time	15min	18min	38min

Parallelizing a single task can speed it up, but it increases resource demand, which may lead to longer queue times on CHTC, resulting in no overall reduction in total runtime.

3. Conclusion

This project modeled and predicted weather data in Brazil using a combined ARIMA-Random Forest. The models effectively captured key weather variables with strong performance. CHTC was crucial in speeding up both training and prediction tasks.

Looking ahead, improving model accuracy and balancing the complexity of tasks with CHTC's resources will be key. Finding this balance could reduce queue times and runtime, maximizing efficiency.

Contributions

Member	Proposal	Coding	Presentation	Report
Shumeng Fang	1	1	1	1
Rui Zhang	1	1	1	1
Yifan Chen	1	1	1	1
Shixin Zhang	1	1	1	1