Temperature Forecasting Project

Group ID: 17
202418009, 202418053, 202418043, 202418039

Abstract

This study focuses on forecasting temperature using historical weather data. Weather prediction is a complex and vital challenge in today's world, relying heavily on advancements in science and technology. This research explores various forecasting techniques, including statistical models such as AR, MA, and ARMA, as well as deep learning approaches like RNN, LSTM, and GRU. The objective is to analyse temperature trends across different regions and observe how they evolve over time. Through extensive experimentation, our results demonstrate that deep learning models, particularly those based on recurrent architectures, provide the most accurate and reliable forecasts for temperature prediction tasks. The detailed code can be found here.

Problem Statement

Accurate and timely temperature forecasting is essential for disaster preparedness, agriculture, transportation, and daily life decisions. However, existing weather prediction models often struggle to provide reliable and region-specific forecasts due to the complex, dynamic, and non-linear nature of atmospheric data. This research aims to develop and evaluate advanced forecasting models capable of predicting key weather parameters like temperature across multiple regions with improved accuracy and computational efficiency, leveraging.

Dataset Description

The data used to conduct this research was extracted from **Services — National Centres for Environmental Information (NCEI)**. This is a U.S. government site, so data credibility is high. The dataset consists of 3,597 observations and 29 attributes, each representing a daily weather report from a specific station.

Data Characteristics

- No missing values across any columns.
- Numerical features: temperature, pressure, precipitation, wind speed, visibility.

- Categorical features: metadata attributes (_ATTRIBUTES), special weather event indicators.
- Spans multiple meteorological stations, suitable for regional forecasting.

Literature Survey

1. **Sindhu P. Menon et al.** used time series and multiple linear regression to study the Urban Heat Island effect.
2. **Shen Rong et al.** applied linear regression to examine temperature's effect on iced product sales.
3. **Pinki Sagar et al.** used linear equations to forecast trends in humidity and temperature time series.
4. **Senlin Zhu et al.** used M-GASVR (a GA-tuned SVR model) to forecast reservoir water temperatures.

Methodology

1. Data Collection

Collected from NCEI here, including temperature, humidity, wind speed, atmospheric pressure, timestamps, and geolocation.

2. Data Preprocessing

- No missing values; no imputation needed.
- Engineered features: day, month, season, lags.
- Normalized data.
- Split: 80% train, 20% test.
- Stationarity test using Dickey-Fuller:

```
Test Statistic: -3.497198
p-value: 0.008059
=> Data is stationary (p < 0.05)
```

3. Model Selection

- Statistical: AR, MA, ARMA, ARIMA
- Deep Learning: RNN, LSTM, GRU

4. Model Training

-¿ TimeSeries Split and then training model on train data then evaluating on test data.

5. Model Evaluation

Evaluate on test set using MSE, MAE, RMSE, R^2. Visual comparisons for trends.

Visual Results and Plots

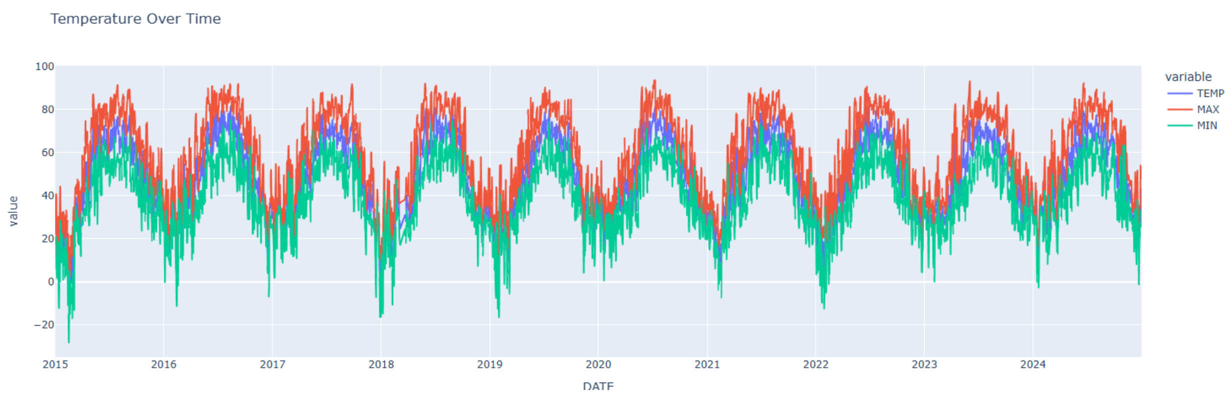

Figure 1: Temperature Over Time

This image shows a variation in temperature over the years.

Decomposition of Time Series Data

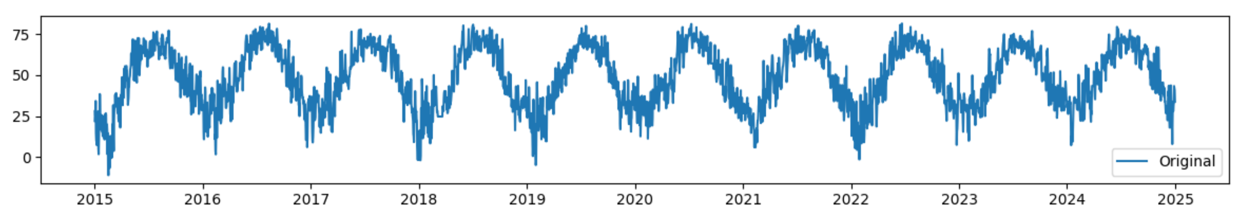

Figure 2: Original Time Series Data

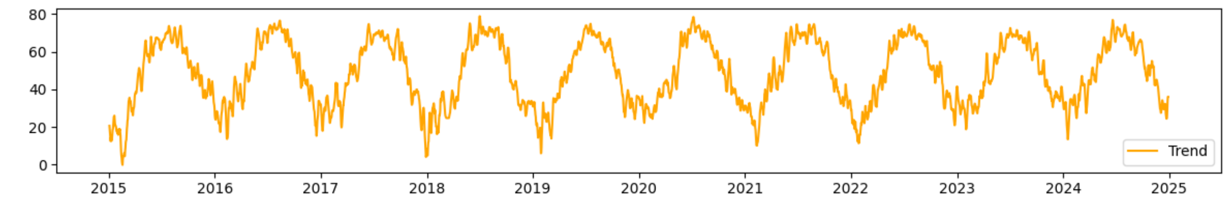

Figure 3: Trend Component

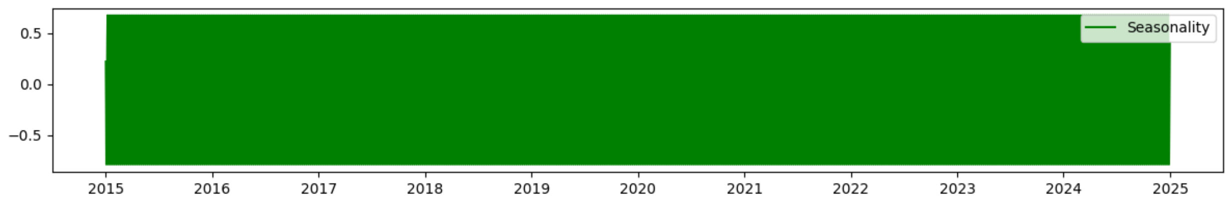

Figure 4: Seasonality Component

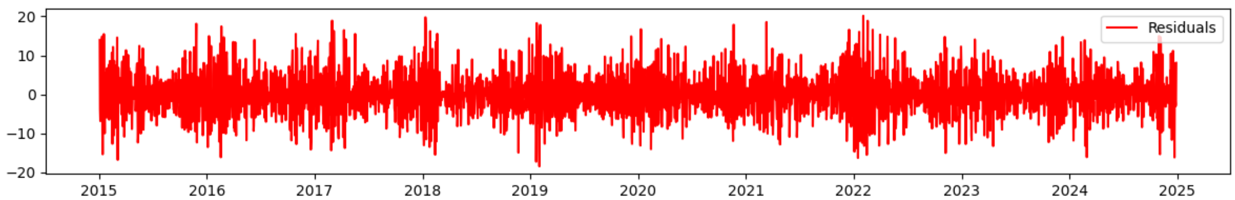

Figure 5: Residual Component

Applying time series model

1. The Nave base model

Navie method is used as a benchmark model, to compare the performance of a more advanced forecasting model.

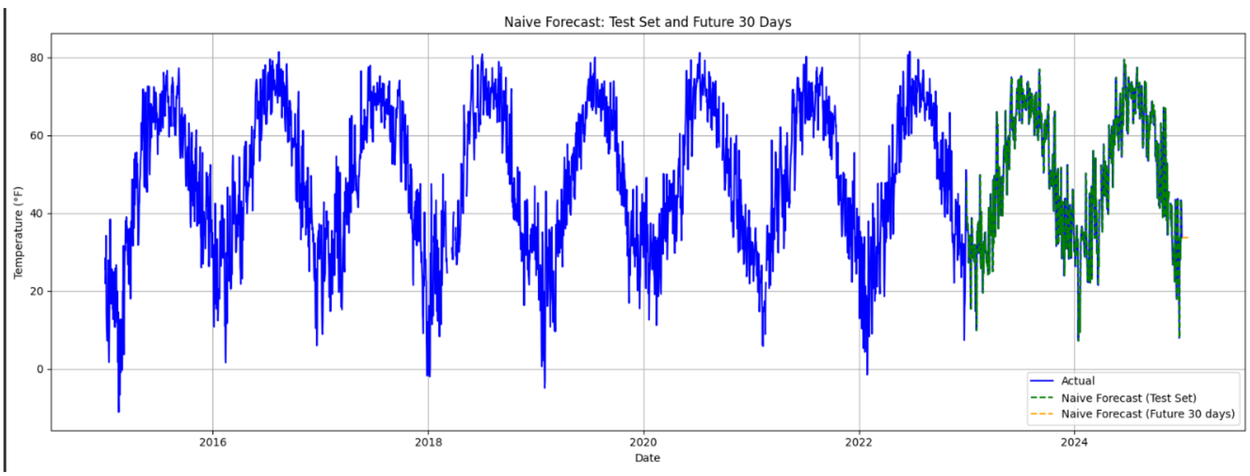

Figure 6: Naive Base model

Naive Forecast Evaluation

Train Set Metrics:
Mean Absolute Error (MAE): 4.87
Mean Squared Error (MSE): 42.06
Root Mean Squared Error (RMSE): 6.49
Coefficient of Determination (R^2 Score): 0.88
Test Set Metrics:

Mean Absolute Error (MAE): 4.30
Mean Squared Error (MSE): 32.81
Root Mean Squared Error (RMSE): 5.73
Coefficient of Determination (R^2 Score): 0.87

ACF and PACF plot

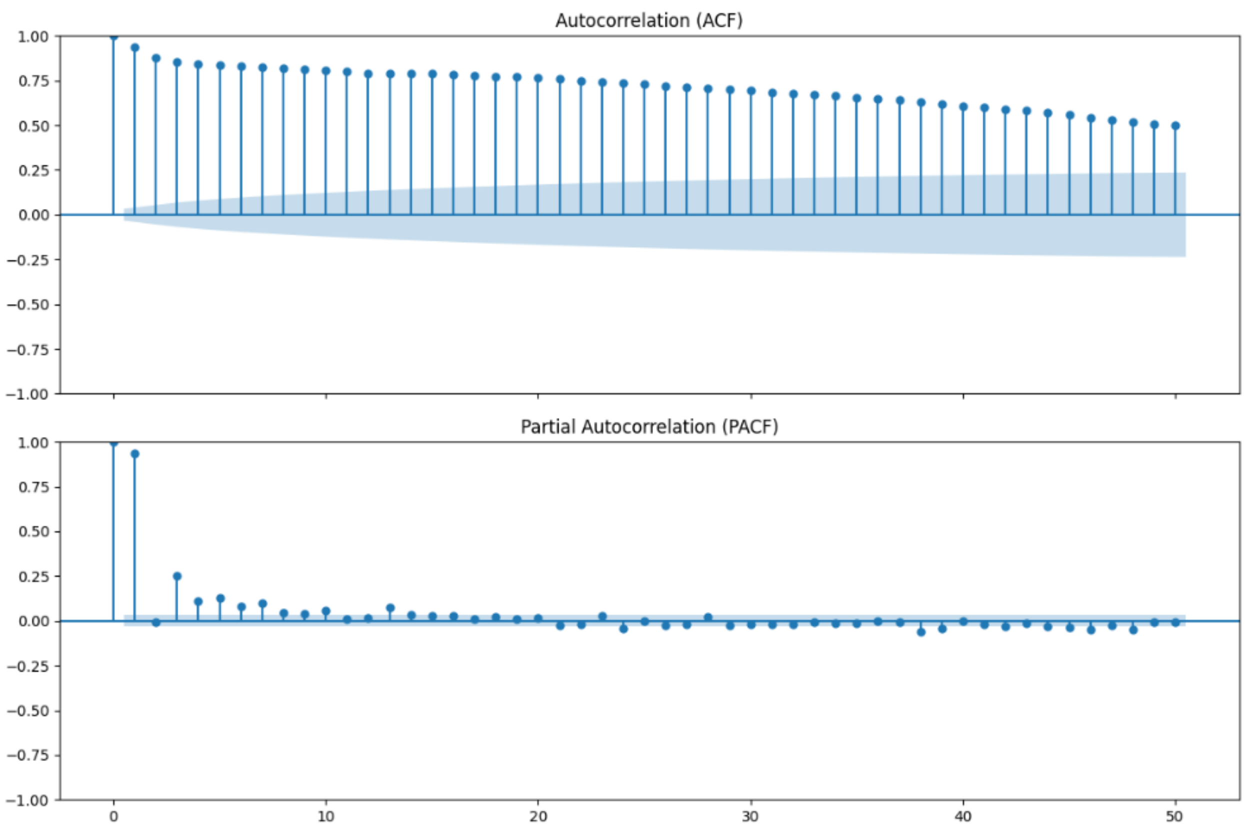

Figure 7: ACF and PACF Plot

The ACF plot shows a gradual decline in autocorrelation with significant lags up to around 10-15, indicating some temporal dependence in the data. The PACF plot exhibits a sharp drop after the first lag, with most subsequent lags within the confidence interval, suggesting an AR(1) process might be appropriate, though further lags could be considered for a more comprehensive model.

2. Moving Average (2) Model

Model Description

Moving Average (MA) is a type of time series forecasting model where the current value of the time series depends linearly on the mean of the error term and past error terms.

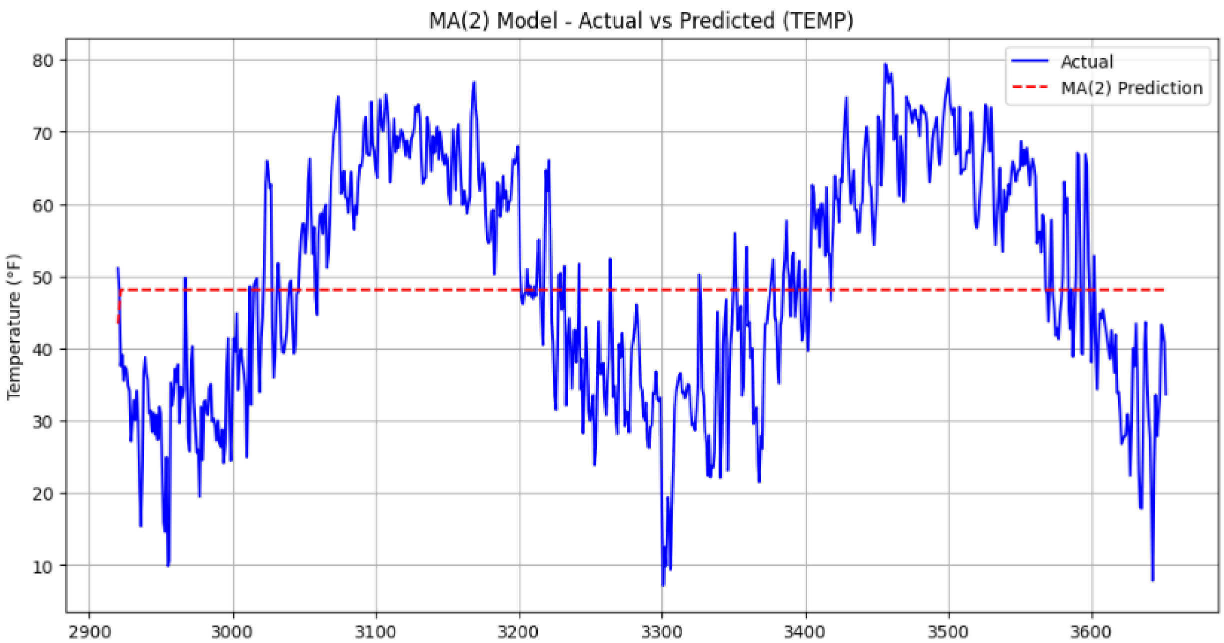

Figure 8: Moving Average Plot

Model Interpretation

The plot titled *"MA(2) Model - Actual vs Predicted (TEMP)"* compares actual temperature data (blue) with MA(2) predictions (red) over a time series indexed by date. The actual temperatures fluctuate significantly between 10°F and 80°F, while the MA(2) prediction remains nearly constant around 50°F, indicating poor tracking of the data's variability. The evaluation metrics show a high Mean Absolute Error (MAE) of 14.2181 and Mean Squared Error (MSE) of 263.1740, with a negative coefficient of determination ($R^2 = -0.0104$). This suggests that the MA(2) model fails to capture the temperature dynamics effectively. A more sophisticated model, such as ARIMA or SARIMA, may be required to account for underlying trends and seasonality.

Model Evaluation

- MA(2) MAE: 14.2181
- MA(2) MSE: 263.1740
- R^2 Score: -0.0104

3. Autoregressive (1) Model

Model Description

In this Autoregressive(AR) model, the current value of a variable is modeled as a linear combination of its past values.

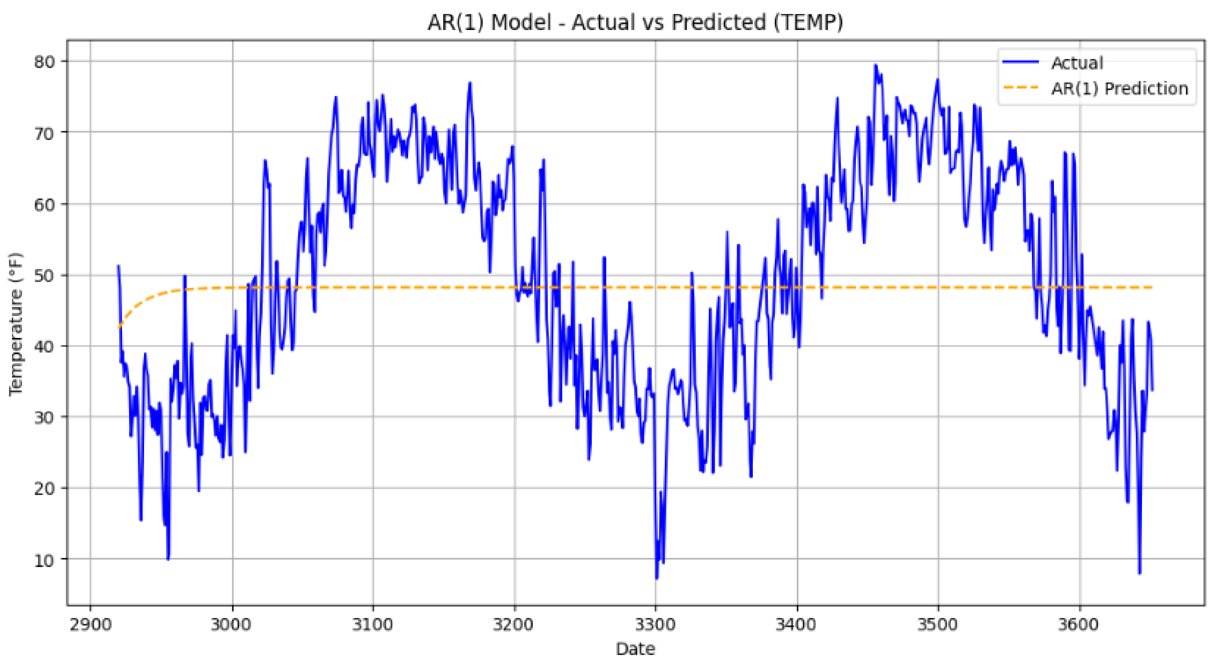

Figure 9: Autoregressive (AR(1)) Model Plot

Model Interpretation

The plot for the AR(1) model shows that the actual temperature (blue) varies widely between 20°F and 80°F, while the AR(1) prediction (orange) remains nearly constant around 50°F. This indicates poor adaptability to the data's fluctuations. The evaluation metrics reveal a high Mean Absolute Error (MAE) of 14.1068 and a Mean Squared Error (MSE) of 259.7374, with a near-zero coefficient of determination ($R^2 = 0.0028$). These results confirm the model's inability to capture the temperature trends effectively. This suggests that the AR(1) model, using only one lag, is too simplistic for this dataset. A more complex model, such as ARIMA or SARIMA, may be more suitable for improving forecasting accuracy.

Model Evaluation

- AR(1) MAE: 14.1068
- AR(1) MSE: 259.7374
- R^2 Score: 0.0028

4. ARMA (AutoRegressive Moving Average)

Model Description

ARMA is a statistical model that combines Autoregressive (AR) and Moving Average (MA) components to predict future values based on past observations and past forecast errors. It is suitable for stationary time series data.

AIC Comparison

All AIC Values:

```
(4, 0, 4)  18724.353215
(2, 0, 2)  18725.038517
(1, 0, 4)  18726.525229
(3, 0, 3)  18726.581575
(3, 0, 2)  18726.946078
(2, 0, 3)  18726.967873
(3, 0, 4)  18728.047363
(4, 0, 3)  18728.074177
```

Optimal Parameters:
Best (p, d, q): (4, 0, 4) — AIC: 18724.35

Model Interpretation

The plot titled *"ARMA(4, 0, 4) Forecast (AIC = 18724.35)"* compares actual temperature data (blue) with ARMA(4, 0, 4) predictions on the test set (red) and a 30-day future forecast (orange), spanning from 2015 to 2025. The actual temperature shows strong seasonal variation between 0°F and 80°F, while the model's predictions remain clustered around the mean. This lack of responsiveness indicates that the model struggles to capture both short-term and long-term dynamics. The evaluation metrics further support this: a high Mean Absolute Error (MAE) of 14.3953, Mean Squared Error (MSE) of 300.8314, and a negative

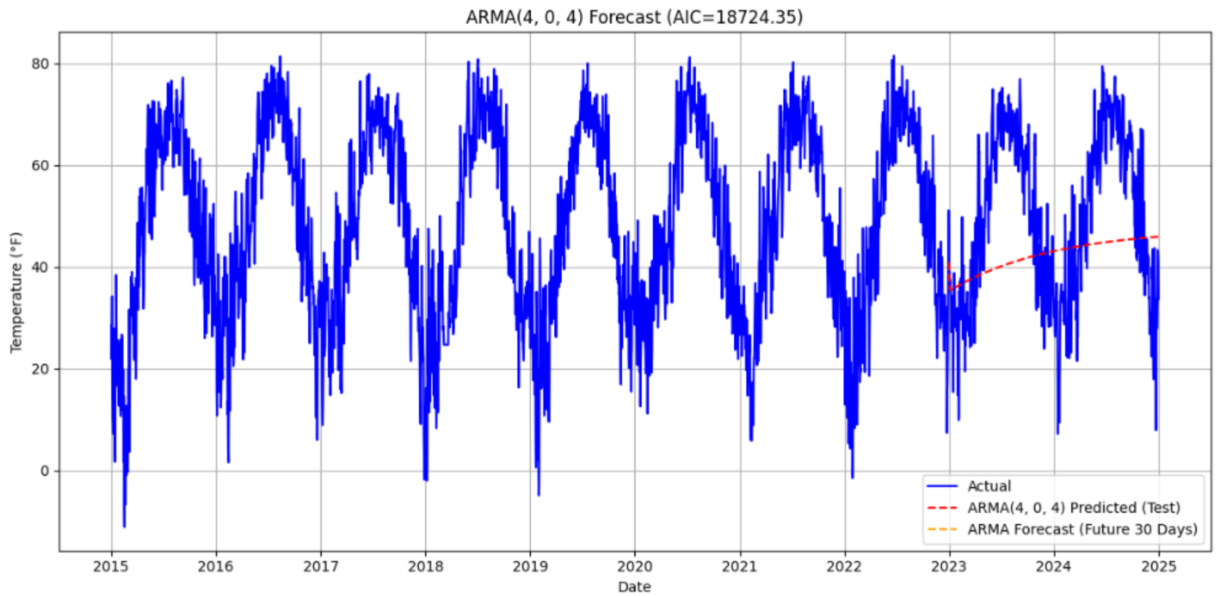

Figure 10: ARMA(4, 0, 4) Model - Actual vs Predicted and 30-Day Forecast

coefficient of determination ($R^2 = -0.1658$) suggest the ARMA(4, 0, 4) model is inadequate for forecasting this time series. A seasonal model such as SARIMA may be better suited to capture the patterns present in the data.

Model Evaluation

- MAE: 14.3953
- MSE: 300.8314
- R^2 Score: -0.1658

Q-Q Plot

The Q-Q plot of residuals shows a noticeable curvature deviating from the reference diagonal line, indicating that the residuals are not normally distributed. This suggests:

- Potential model misspecification or omitted variables.
- Presence of non-linear patterns that the model fails to capture.
- Overall, the model may not fully represent the underlying data distribution.

Residual Plot

In the residual plot, residuals should ideally be randomly scattered around zero with no visible patterns. Any structure in the plot (e.g., curvature, clustering) indicates issues such as non-linearity or heteroscedasticity, implying that model assumptions may be violated.

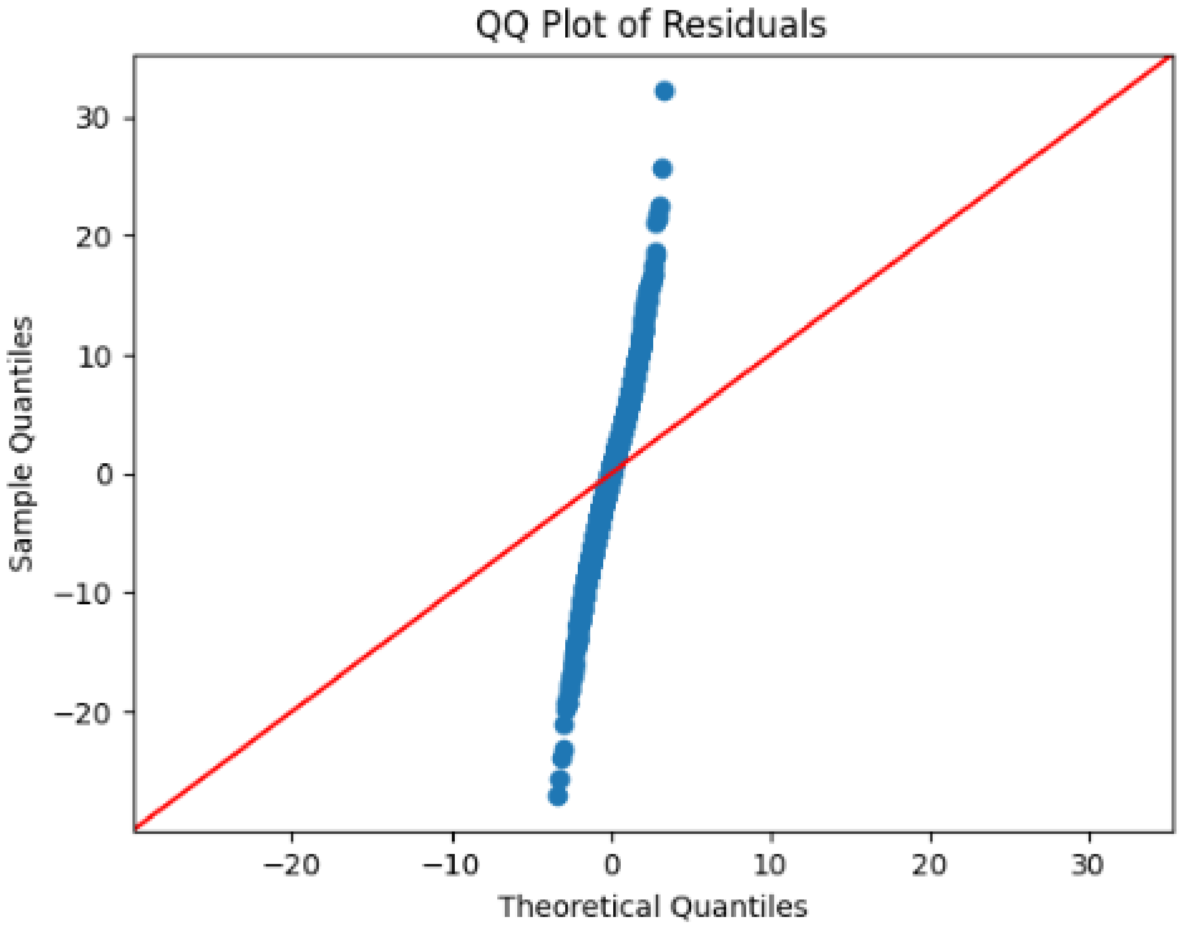

Figure 11: Q-Q Plot of Residuals

Ljung-Box Test for Autocorrelation

Ljung-Box Test Results:

```
     lb_stat  lb_pvalue
1   0.162915   0.686487
2   0.575492   0.749952
3   0.631931   0.889085
4   1.181342   0.881161
5   1.189163   0.945913
6   1.411790   0.965137
7   2.094999   0.954396
8   2.355151   0.968144
9   2.358051   0.984460
10  2.411339   0.992106
```

The Ljung-Box test checks for autocorrelation in the residuals. A **p-value greater than 0.05** at various lags indicates that the residuals resemble white noise, i.e. there is no significant autocorrelation. This supports the assumption that errors are independently dis-

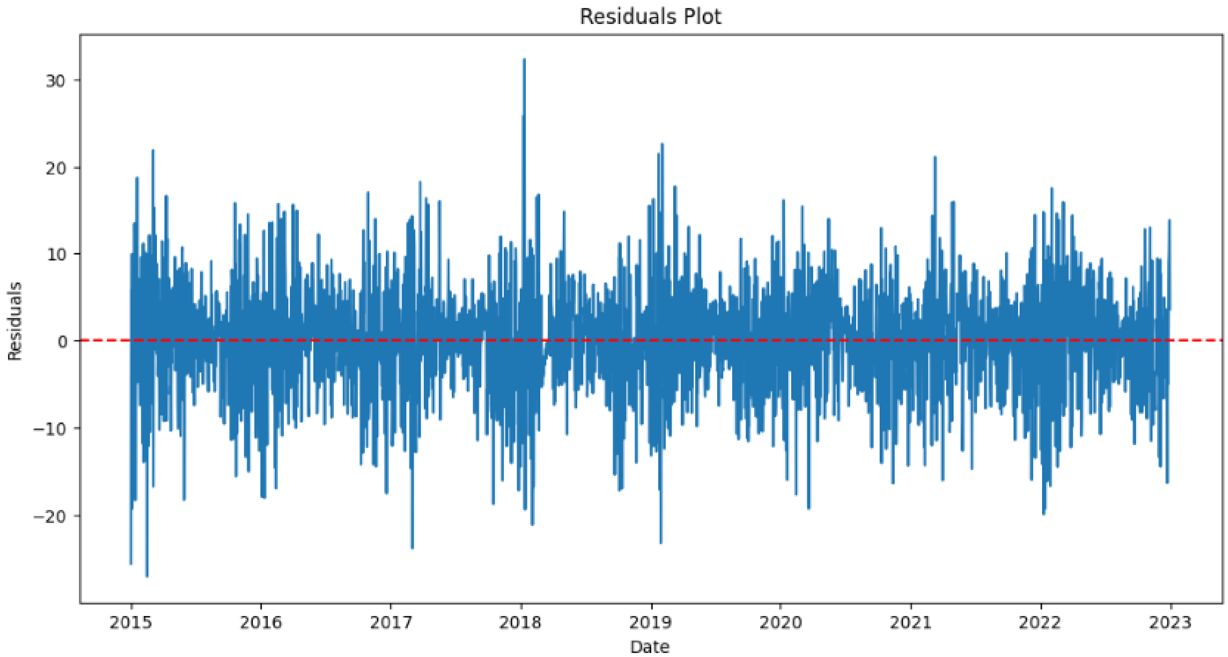

Figure 12: Residual Plot

tributed.

5. RNN Model

Model Description

A Recurrent Neural Network (RNN) is a type of deep learning model designed specifically for sequential data. It processes data in temporal order by feeding outputs from previous time steps as inputs to the current step, allowing the model to learn temporal dependencies and patterns effectively.

Prediction Interpretation

- The plot compares actual (blue) and predicted (red) temperatures over time (index 0 to 700), ranging from 0°F to 80°F.
- The RNN predictions closely follow actual temperature trends, showing alignment in peaks and troughs.
- Minor deviations are present but overall indicate good temporal pattern recognition.
- Compared to simpler models, the RNN provides improved fit and adaptability to variations in data.

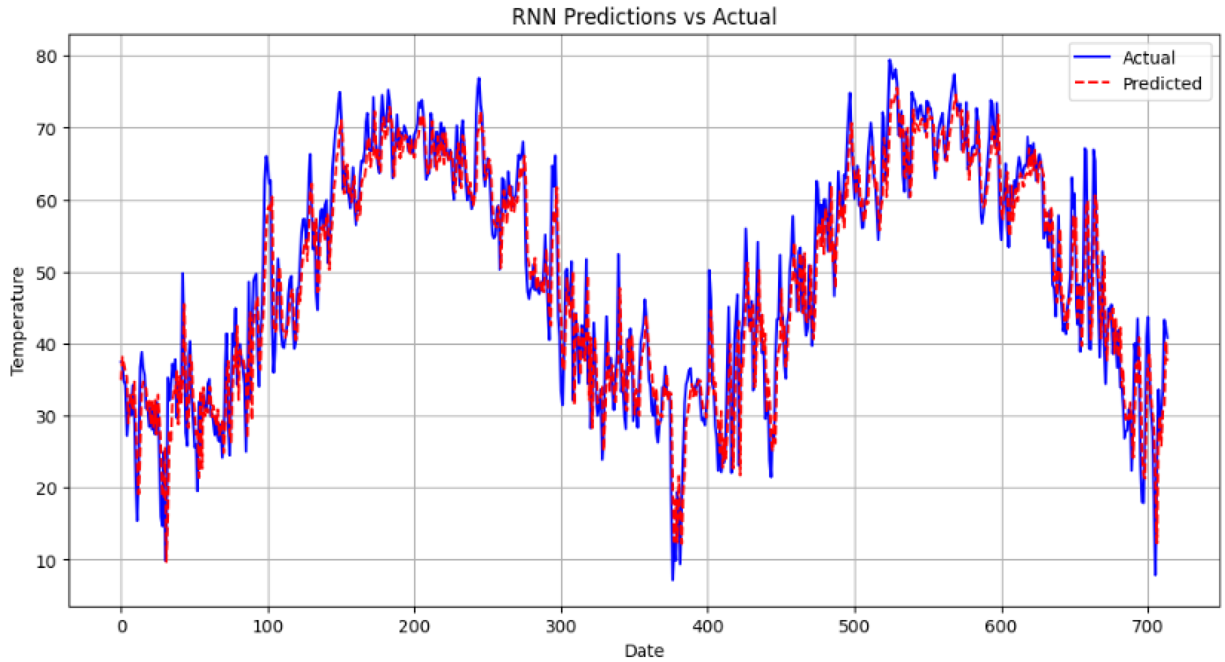

Figure 13: RNN Model - Actual vs Predicted Temperature

Figure 14: RNN Model - Training vs Validation Loss

Training Curve Interpretation

- Training loss (blue) decreases sharply within the first 2 epochs and stabilizes around 0.005, indicating effective learning.
- Validation loss (red) initially follows training loss but increases after epoch 6, peaking

around 0.007.

- This divergence suggests the onset of overfitting, where the model learns training data well but generalizes less effectively to unseen data.

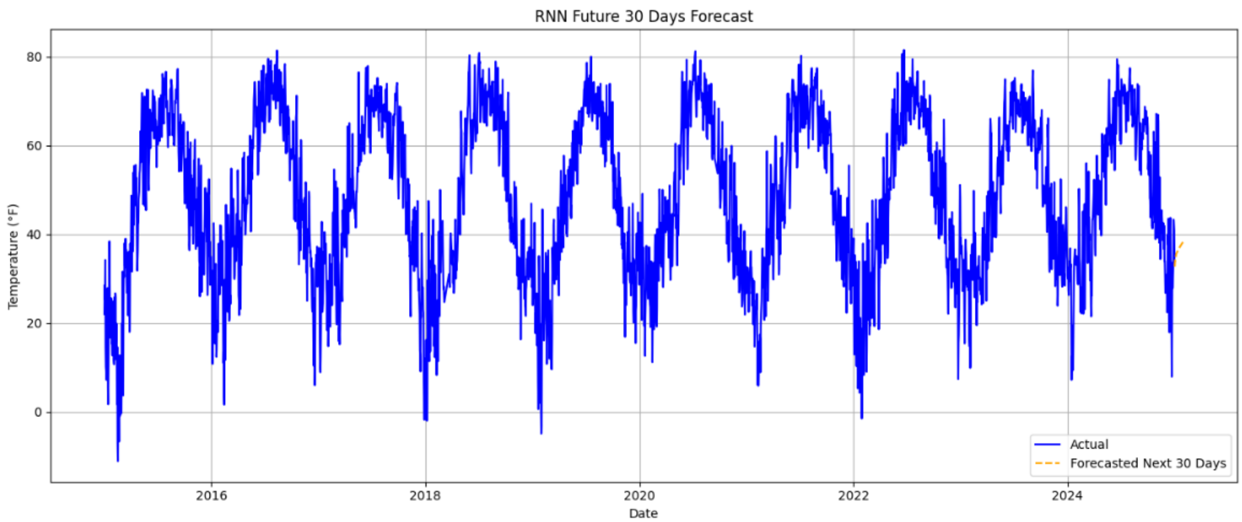

Figure 15: RNN Model - Evaluation Metrics

Evaluation Interpretation

- The RNN achieves a strong R^2 score of 0.90 on training data and 0.89 on test data, indicating high explanatory power.
- The MAE values for train and test (4.62 and 4.21, respectively) show low average error.
- MSE (35.82 train, 29.46 test) and RMSE (5.99 train, 5.43 test) suggest consistent performance and generalization.

Model Evaluation

Train Set Metrics:

- MAE: 4.62
- MSE: 35.82
- RMSE: 5.99
- R^2 Score: 0.90

Test Set Metrics:

- MAE: 4.21

- MSE: 29.46
- RMSE: 5.43
- R^2 Score: 0.89

6. Long Short-Term Memory (LSTM) Model

Model Description

Long Short-Term Memory (LSTM) networks are an advanced type of Recurrent Neural Network (RNN) that utilize memory cells to capture long-term dependencies in sequential data. They are especially effective for time series forecasting with long-range temporal patterns.

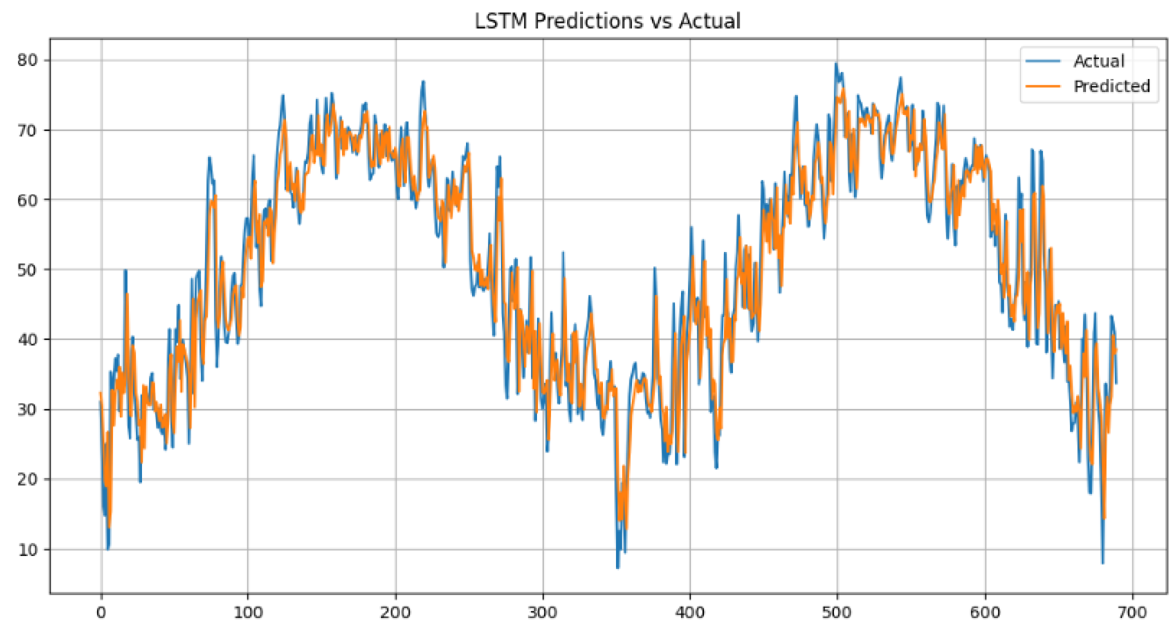

Figure 16: LSTM Model - Actual vs Predicted Temperature

Prediction Interpretation

- The plot compares actual (blue) and predicted (orange) temperatures over a date index from 0 to 700, with values ranging between 0°F and 80°F.
- LSTM predictions closely follow the peaks and troughs of the actual temperature data, indicating strong pattern recognition.
- Minor deviations are present but less pronounced than in simpler models, showing better adaptability.

- This alignment reflects LSTM's ability to model complex, time-dependent trends effectively.
- Overall, the model demonstrates robust predictive performance on the temperature dataset.

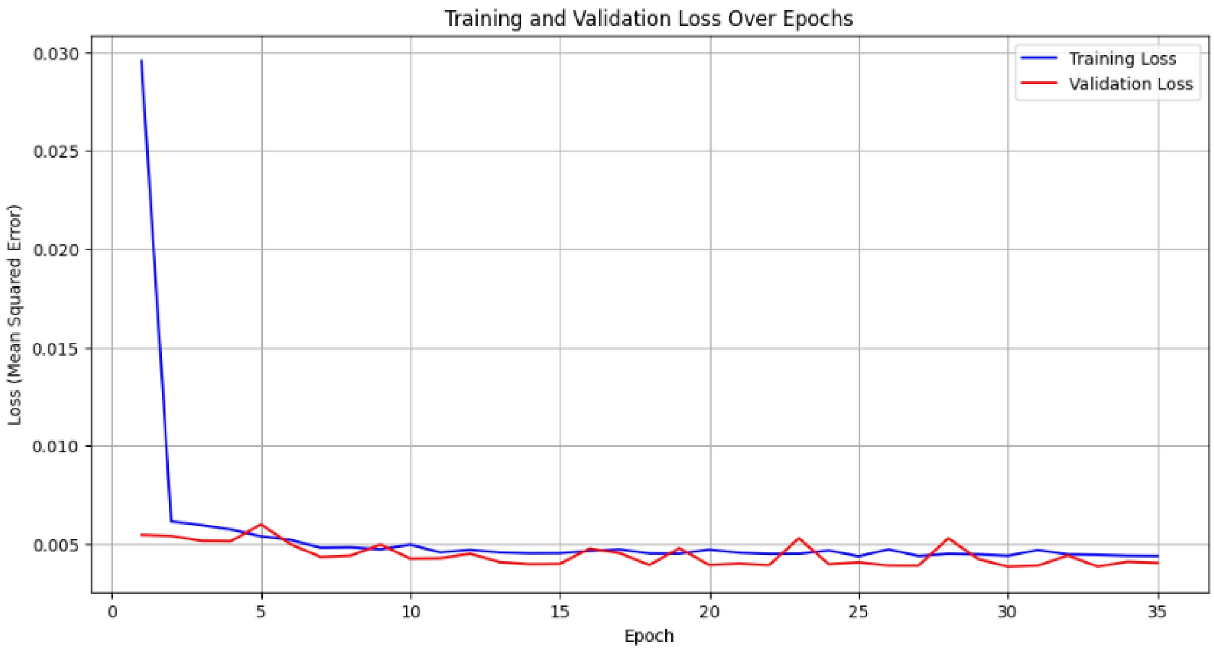

Figure 17: LSTM Model - Training vs Validation Loss

Training Curve Interpretation

- Training loss (blue) drops steeply from 0.03 to around 0.005 by epoch 5 and stabilizes, indicating efficient learning.
- Validation loss (red) initially decreases but then fluctuates slightly around 0.005, suggesting minor instability.
- The small gap between training and validation loss indicates minimal overfitting.

Evaluation Interpretation

- The model achieves a strong R^2 score of 0.897 on training data and 0.887 on the test set, showing high predictive accuracy.
- MAE values (4.541 for training and 4.137 for testing) reflect low average errors.
- MSE values (35.316 train, 29.092 test) and RMSE (5.943 train, 5.394 test) further support the model's consistency and reliability.

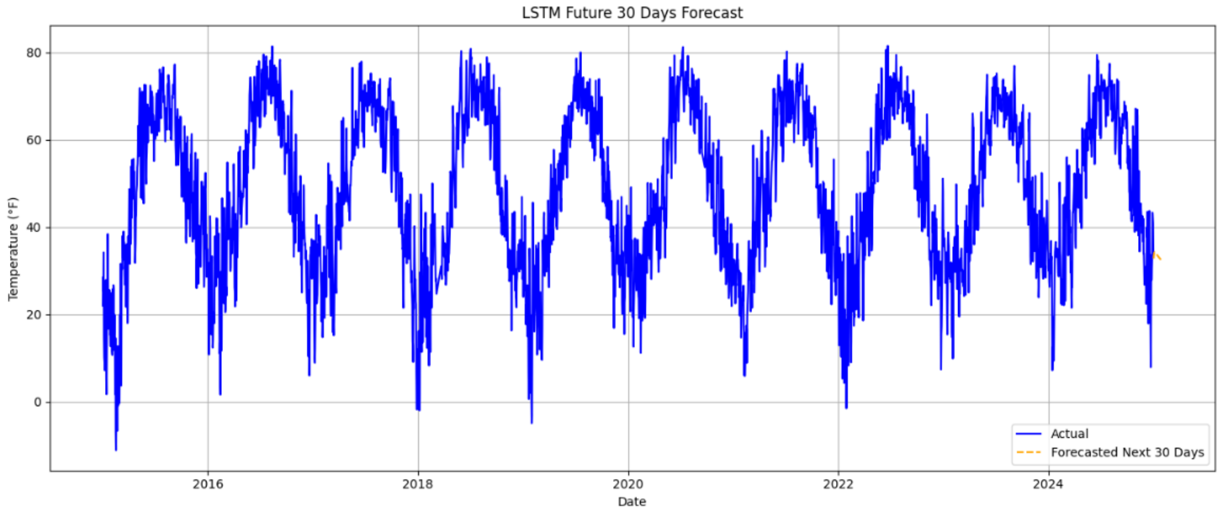

Figure 18: LSTM Model - Evaluation Metrics

Model Evaluation

Train Set Metrics:

- RMSE: 5.943
- MAE: 4.541
- MSE: 35.316
- R^2: 0.897

Test Set Metrics:

- RMSE: 5.394
- MAE: 4.137
- MSE: 29.092
- R^2: 0.887

7. GRU (Gated Recurrent Unit)

Model Description

The Gated Recurrent Unit (GRU) is a simplified variant of the LSTM that also captures long-term dependencies in sequential data. It uses fewer parameters, offering a balance between performance and computational efficiency.

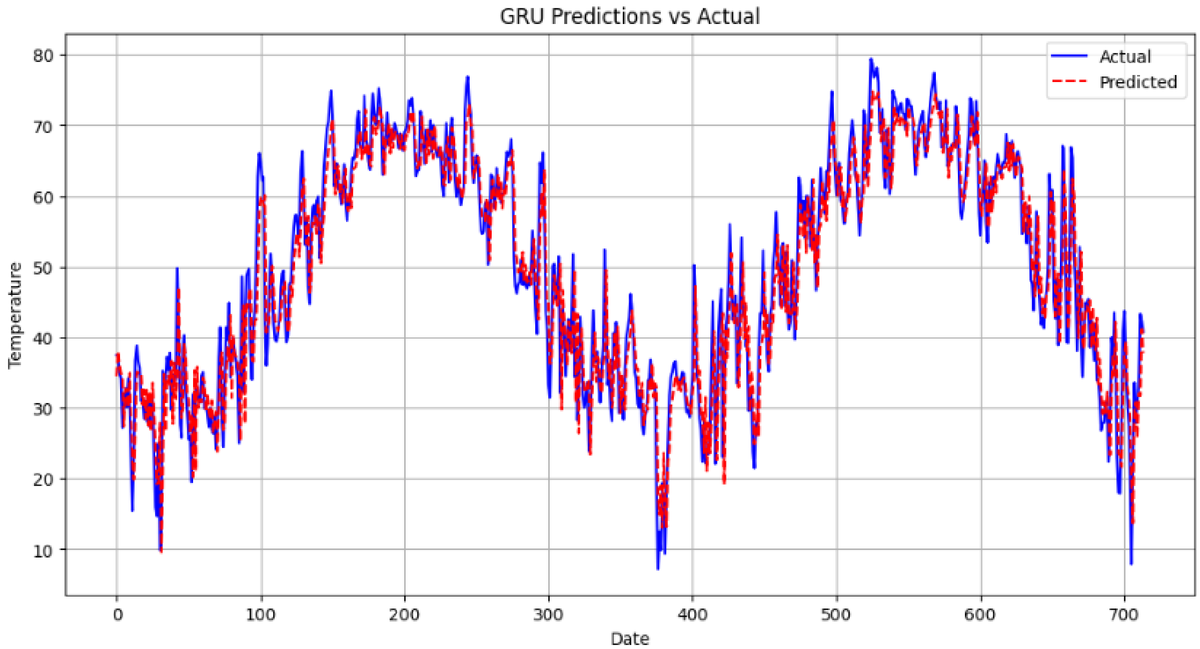

Figure 19: GRU Model - Actual vs Predicted Temperature

Prediction Interpretation

- The plot compares actual (blue) and predicted (red) temperatures over a date index from 0 to 700, with values ranging from 0°F to 80°F.
- GRU predictions align closely with the actual data, successfully capturing the peaks and troughs of the temperature.
- Minor deviations are observed, suggesting that while the model captures overall trends, it may miss some fine details.
- Overall, the GRU model demonstrates strong performance in predicting temperature patterns.
- The model shows robust predictive capability and is competitive with more complex models like LSTM.

Training Curve Interpretation

- Training loss (blue) decreases sharply from 0.016 to around 0.005 by epoch 2 and stabilizes, indicating effective learning.
- Validation loss (red) decreases initially but fluctuates around 0.005 with a slight increase around epoch 6, suggesting minor overfitting.
- The gap between training and validation loss remains small, indicating minimal overfitting and efficient model training.

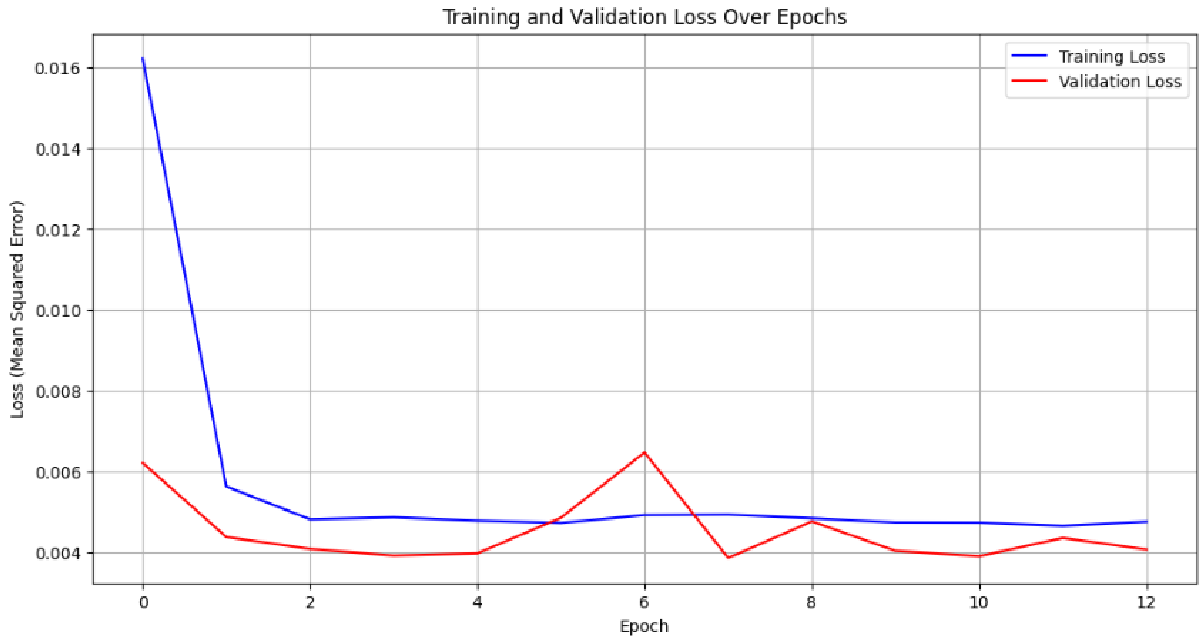

Figure 20: GRU Model - Training vs Validation Loss

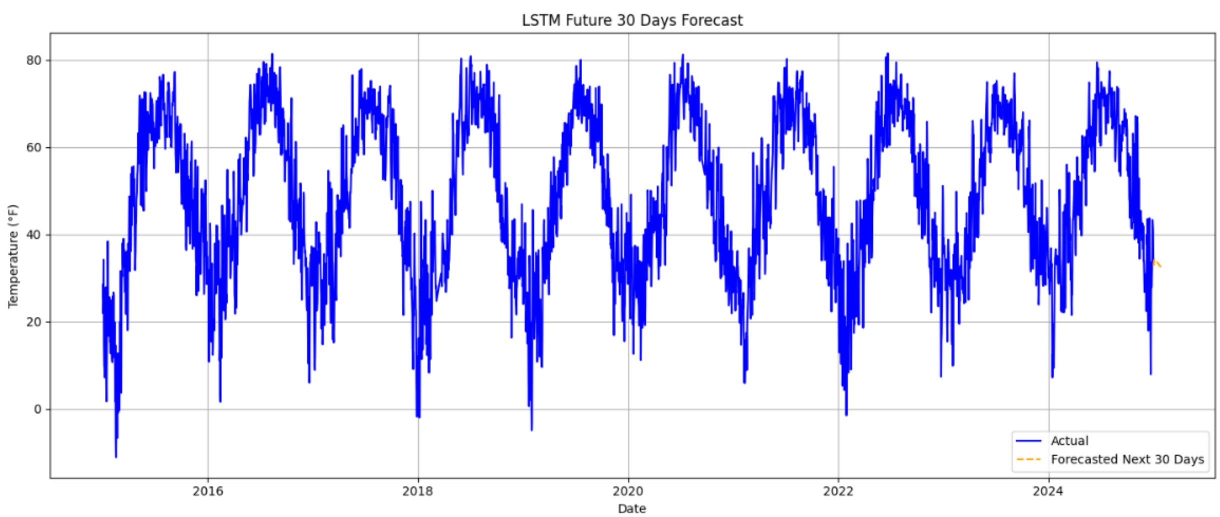

Figure 21: GRU Model - Evaluation Metrics

Evaluation Interpretation

- The GRU model achieves a high R^2 score of 0.90 on the training set and 0.89 on the test set, reflecting strong explanatory power.
- MAE values (4.57 for training and 4.13 for testing) suggest the model's errors are consistently low.
- MSE values (35.50 train, 28.74 test) and RMSE (5.96 train, 5.36 test) further confirm the model's stability and performance across different data sets.

Model Evaluation

Train Set Metrics:

- RMSE: 5.96
- MAE: 4.57
- MSE: 35.50
- R^2: 0.90

Test Set Metrics:

- RMSE: 5.36
- MAE: 4.13
- MSE: 28.74
- R^2: 0.89

Final Interpretation and Forecasting

Model Evaluation Summary

Based on the evaluation matrix, the **GRU Model** emerges as the best model for temperature prediction. It achieves the lowest test MAE (4.13) and MSE (28.74), with a competitive R^2 score of 0.890, indicating the highest accuracy and explanatory power among the models. Compared to LSTM (test MAE: 4.137, MSE: 29.092, R^2: 0.887), RNN (test MAE: 4.21, MSE: 29.46, R^2: 0.89), and Naive Forecast (test MAE: 4.30, MSE: 32.81, R^2: 0.87), GRU consistently outperforms with the smallest error metrics. The slight improvement over train metrics (MAE: 4.57, MSE: 35.50, R^2: 0.90) suggests minor overfitting, but its overall performance justifies its selection as the best model, pending further tuning for optimal generalization.

Model Evaluation Table

Metric	Naive Forecast	RNN Model	LSTM Model	GRU Model
Train MAE	4.870	4.620	4.541	4.570
Train MSE	42.060	35.820	35.316	35.500
Train R^2	0.880	0.900	0.897	0.900
Test MAE	4.300	4.210	4.137	4.130
Test MSE	32.810	29.460	29.092	28.740
Test R^2	0.870	0.890	0.887	0.890

Table 1: Model Evaluation Metrics

Temperature Forecasts (2025-01-01 to 2025-01-30)

Below are the temperature forecasts for January 2025 based on the Naive, RNN, LSTM, and GRU models.

Date	Naive Forecast	RNN Forecast	LSTM Forecast	GRU Forecast
2025-01-01	33.7	32.73	32.78	32.61
2025-01-02	33.7	34.18	33.71	34.06
2025-01-03	33.7	34.16	34.07	34.23
2025-01-04	33.7	34.70	34.23	34.04
2025-01-05	33.7	35.16	34.27	33.93
2025-01-06	33.7	35.60	34.24	33.85
2025-01-07	33.7	35.70	34.18	33.85
2025-01-08	33.7	35.85	34.10	33.69
2025-01-09	33.7	35.99	34.01	33.62
2025-01-10	33.7	36.13	33.92	33.55
2025-01-11	33.7	36.23	33.82	33.48
2025-01-12	33.7	36.34	33.73	33.42
2025-01-13	33.7	36.36	33.64	33.36
2025-01-14	33.7	36.71	33.54	33.30
2025-01-15	33.7	36.84	33.45	33.25
2025-01-16	33.7	36.96	33.36	33.20
2025-01-17	33.7	37.08	33.27	33.15
2025-01-18	33.7	37.19	33.18	33.10
2025-01-19	33.7	37.31	33.09	33.05
2025-01-20	33.7	37.42	33.01	33.01
2025-01-21	33.7	37.53	32.92	32.97
2025-01-22	33.7	37.63	32.84	32.93
2025-01-23	33.7	37.74	32.76	32.90
2025-01-24	33.7	37.85	32.68	32.87
2025-01-25	33.7	37.95	32.61	32.83
2025-01-26	33.7	38.05	32.53	32.80
2025-01-27	33.7	38.15	32.46	32.77
2025-01-28	33.7	38.25	32.38	32.74
2025-01-29	33.7	38.34	32.31	32.71
2025-01-30	33.7	38.44	32.25	32.68

Table 2: Temperature Forecasts from January 1, 2025 to January 30, 2025

Conclusion

The **GRU model** shows superior performance in predicting temperature, achieving the lowest test error metrics among all models evaluated. With a minimal deviation from actual data and a high explanatory power, the GRU model is the most effective model for this time series prediction task. Further tuning and validation could further optimize its generalization capabilities, but, as of now, it is the best model for forecasting future temperatures.