WEATHER TIME SERIES ANALYSIS AND FORECASTING

GROUP MEMBERS:

HEMANTH KUMAR MULLURI .PANTHER ID-(002893683)

SAI KETHAN BHARADWAJ KANITHI.PANTHER ID- (002852918)

VIGNESH AJITH NAIR. PANTHER ID – (002893762)

MURALI KRISHNA MADDINENI. PANTHER ID-(002891677)

Introduction:

This project aims to analyze and forecast weather conditions using a dataset from the Max Planck Institute. The dataset includes high-resolution meteorological data captured at 10-minute intervals over several years. We apply deep learning and machine learning models to predict temperature.

Dataset Description:

The Weather-LTSF (Long-Term Sequential Forecasting) dataset, provided by the Max Planck Institute for Biogeochemistry, contains rich historical weather data collected from the Weather Station at Jena, Germany. It is widely used in time series analysis and forecasting tasks, particularly aimed at benchmarking long-term forecasting models.

Data Collection and Source:

The dataset consists of approximately 52,696 samples, with weather observations systematically recorded at 10-minute intervals. The provided data spans from January 1, 2009, to December 31, 2009, covering one full year of high-frequency atmospheric measurements.

Variables Included:

This dataset comprises a total of **21 weather-related attributes**, including:

- **Date Time**: Timestamp indicating the exact recording time.
- **Temperature (T)**: Ambient temperature measured in degrees Celsius (°C).
- **Pressure (p)**: Atmospheric pressure in hectopascals (hPa).
- **Humidity (rh)**: Relative humidity percentage (%).

- **Wind Speed (wv)**: Wind speed measured in meters per second (m/s).
- **Wind Direction (wd)**: Wind direction given in degrees (°).
- **Dew Point (Tdew)**: Temperature at which air becomes saturated with moisture (°C).
- **Maximum Wind Speed (max. wv)**: Peak wind speed during the sampling period (m/s).
- Additional meteorological parameters providing comprehensive coverage for detailed analyses and modelling.

INITIAL DATASET:

Plotted using df.head () for 1st 5 rows and columns.

date	p	T	Tpot	Tdew	rh	VPmax	VPact	VPdef	sh	H2OC	...	wd	rain	raining	SWDR	PAR	max. PAR	Tlog	hour	month	day
2020-01-01 00:10:00	1008.89	0.71	273.18	-1.33	86.1	6.43	5.54	0.89	3.42	5.49	...	224.3	0.0	0.0	0.0	0.0	0.0	11.45	0	1	1
2020-01-01 00:20:00	1008.76	0.75	273.22	-1.44	85.2	6.45	5.49	0.95	3.39	5.45	...	206.8	0.0	0.0	0.0	0.0	0.0	11.51	0	1	1
2020-01-01 00:30:00	1008.66	0.73	273.21	-1.48	85.1	6.44	5.48	0.96	3.39	5.43	...	197.1	0.0	0.0	0.0	0.0	0.0	11.60	0	1	1
2020-01-01 00:40:00	1008.64	0.37	272.86	-1.64	86.3	6.27	5.41	0.86	3.35	5.37	...	206.4	0.0	0.0	0.0	0.0	0.0	11.70	0	1	1
2020-01-01 00:50:00	1008.61	0.33	272.82	-1.50	87.4	6.26	5.47	0.79	3.38	5.42	...	209.6	0.0	0.0	0.0	0.0	0.0	11.81	0	1	1

5 rows × 23 columns

DIFFERENCE BETWEEN BOTH THE DATSETS:

```
Columns in Original:
['date', 'p (mbar)', 'T (degC)', 'Tpot (K)', 'Tdew (degC)', 'rh (%)', 'VPmax (mbar)', 'VPact (mbar)', 'VPdef (mbar)', 'sh (g/kg)', 'H2OC (mmol/mol)', 'rho (g/m**3)',

Columns in Cleaned:
['date', 'p', 'T', 'Tpot', 'Tdew', 'rh', 'VPmax', 'VPact', 'VPdef', 'sh', 'H2OC', 'rho', 'wv', 'max. wv', 'wd', 'rain', 'raining', 'SWDR', 'PAR', 'max. PAR', 'Tlog']

Original shape: (52696, 22)
Cleaned shape:  (52696, 21)

Removed Columns: {'Tlog (degC)', 'VPdef (mbar)', 'wv (m/s)', 'p (mbar)', 'sh (g/kg)', 'SWDR (W/m)', 'raining (s)', 'VPact (mbar)', 'T (degC)', 'rain (mm)', 'Tdew (
Added/Updated Columns: {'max. PAR', 'Tdew', 'Tpot', 'VPact', 'wd', 'max. wv', 'p', 'VPmax', 'sh', 'VPdef', 'rh', 'wv', 'T', 'SWDR', 'H2OC', 'rain', 'Tlog', 'raining'
```

DATASET AFTER CLEANING USING VARIOUS TECHNIQUES:

1. Filled Missing Values Using Linear Interpolation
2. Removed Unwanted Column
3. Removed Units in the Column title

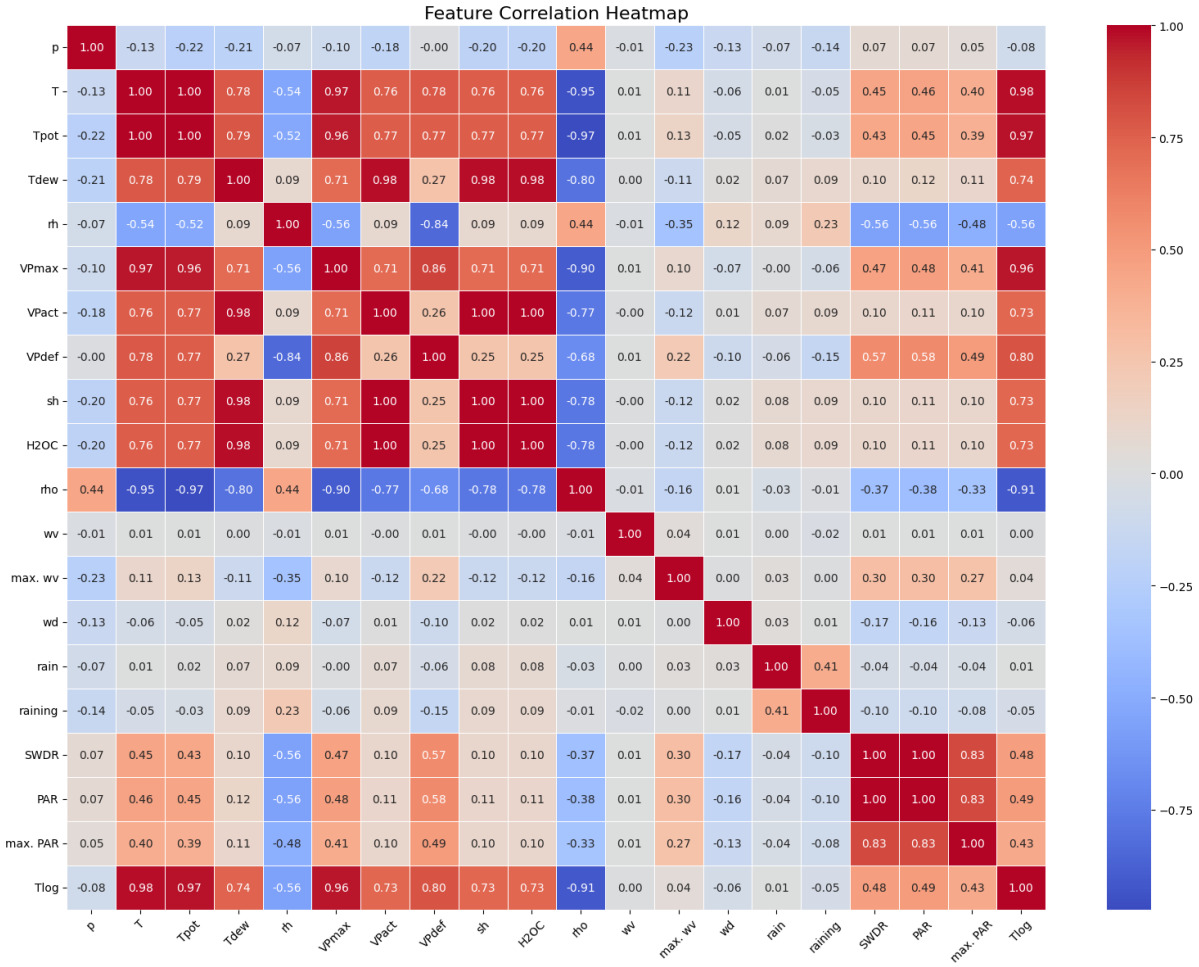
Sample data from cleaned:

	date	p	T	Tpot	Tdew	rh	VPmax	VPact	VPdef	sh	...	rho	wv	max. wv	wd	rain	raining	SWDR	PAR	max. PAR	Tlog
0	2020-01-01 00:10:00	1008.89	0.71	273.18	-1.33	86.1	6.43	5.54	0.89	3.42	...	1280.62	1.02	1.60	224.3	0.0	0.0	0.0	0.0	0.0	11.45
1	2020-01-01 00:20:00	1008.76	0.75	273.22	-1.44	85.2	6.45	5.49	0.95	3.39	...	1280.33	0.43	0.84	206.8	0.0	0.0	0.0	0.0	0.0	11.51
2	2020-01-01 00:30:00	1008.66	0.73	273.21	-1.48	85.1	6.44	5.48	0.96	3.39	...	1280.29	0.61	1.48	197.1	0.0	0.0	0.0	0.0	0.0	11.60
3	2020-01-01 00:40:00	1008.64	0.37	272.86	-1.64	86.3	6.27	5.41	0.86	3.35	...	1281.97	1.11	1.48	206.4	0.0	0.0	0.0	0.0	0.0	11.70
4	2020-01-01 00:50:00	1008.61	0.33	272.82	-1.50	87.4	6.26	5.47	0.79	3.38	...	1282.08	0.49	1.40	209.6	0.0	0.0	0.0	0.0	0.0	11.81

5 rows × 21 columns

Data Preprocessing:

Data is normalized using minmax scaler to fit into values (0,1) so that the model can understand and the correlation between features is also identified and plotted using Correlation Matrix.



- From this what we identified was some features like Tpot, Vpmax ,Tdew sand also , T itself strongly influence Temperature prediction.
- From these when we train the model we get to understand what features are influencing it most based on that model trains better.

Summary of the dataset:

Basic Statistics:

	p	T	Tpot	Tdew	rh	VPmax	VPact	VPdef	sh	H2OC	...	max. wv	
count	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	52696.000000	...	52696.000000	52696
mean	989.989233	10.818241	284.796938	5.409105	72.487133	14.487046	9.676828	4.810131	6.111159	9.782341	...	3.632807	176
std	9.207149	7.468671	7.616995	5.956722	19.230260	7.632960	4.023504	5.539320	2.561536	4.082684	...	2.462467	81
min	955.580000	-6.440000	266.190000	-13.810000	21.160000	3.770000	2.090000	0.000000	1.300000	2.090000	...	0.000000	0
25%	984.800000	4.590000	278.550000	0.777500	58.820000	8.480000	6.460000	1.170000	4.070000	6.530000	...	1.770000	141

- This is done to see where the data lies and also used for finding any outliers.

First few days of daily statistics:

	T			rh	SWDR	rain
	mean	min	max	mean	sum	sum
date						
2020-01-01	-0.514196	-3.46	4.58	86.429301	8250.91	0.0
2020-01-02	-1.056319	-5.78	6.47	80.206042	9570.44	0.0
2020-01-03	4.835278	0.69	8.17	81.975000	1665.99	0.3
2020-01-04	3.784931	1.47	6.61	80.899861	1157.79	4.8
2020-01-05	2.735000	1.47	4.43	76.422639	4572.03	0.0

```
Extreme weather events:
highest_temp: 34.8
lowest_temp: -6.44
highest_wind: 13.77
max_daily_rain: 46.4
hottest_day: 2020-08-08 00:00:00
coldest_day: 2020-01-25 00:00:00
prevailing_wind_direction: 210.3
```

- To see how the data lies in first few days it is plotted so that we can normalize the data
- And also extreme Weather events are plotted for potential Outlier Identification.

MODELS CHOOSED

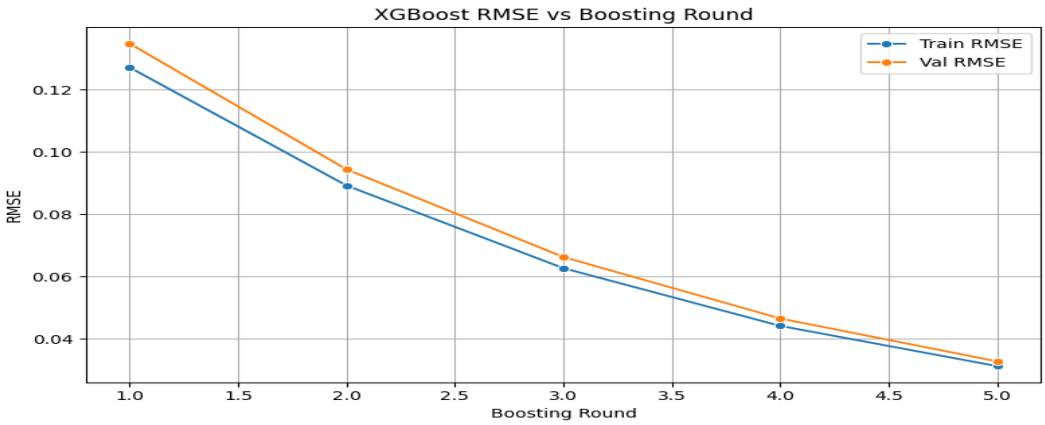
1. **XGBOOST**
2. **MLP- Multilayer Perceptron**
3. **LSTM- Long Short Term Memory**
4. **GRU- Gated Recurrent Units**

Model Architectures & Training Details

1. XGBoost Regressor

The XGBoost model is a gradient-boosted decision tree ensemble that excels at tabular, non-sequential data. In our implementation:

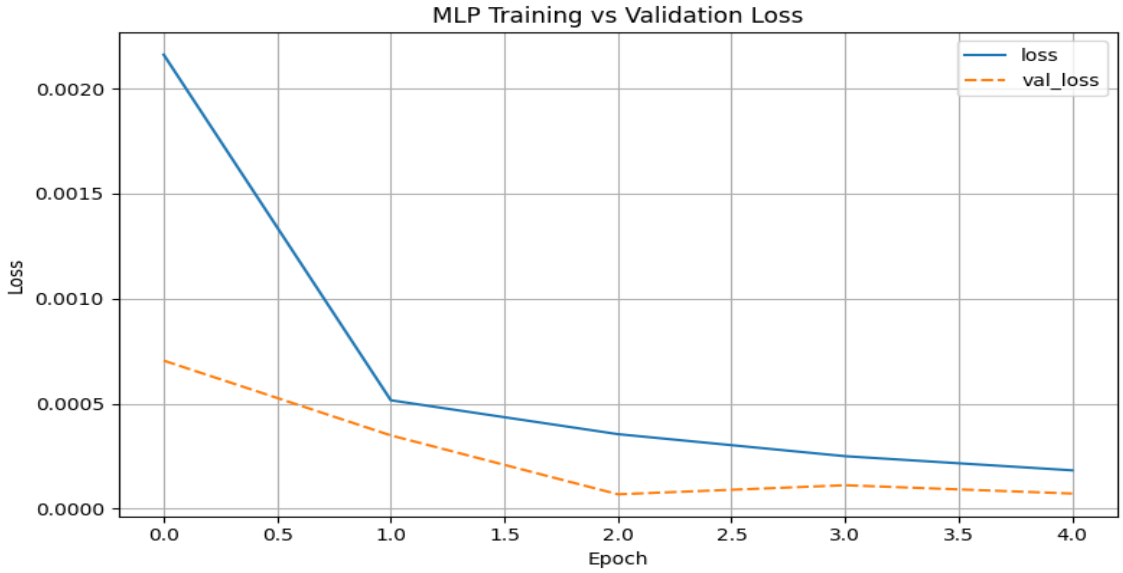
- Objective & Loss:to minimise regularised squared error
- Hyperparameters:
  - n_estimators=100
  - max_depth=6
  - learning_rate=0.1
  - subsample=0.8, colsample_bytree=0.8
- Feature Importance: Dew-point (Tdew), potential temp. (Tpot), and vapor pressure (Vpmax) ranked highest.
- Training: 80/20 train/test split, early stopping (patience=10) on a validation subset, evaluated via RMSE & MAE.



2. Feed-Forward MLP

A fully-connected network to learn non-linear feature interactions:

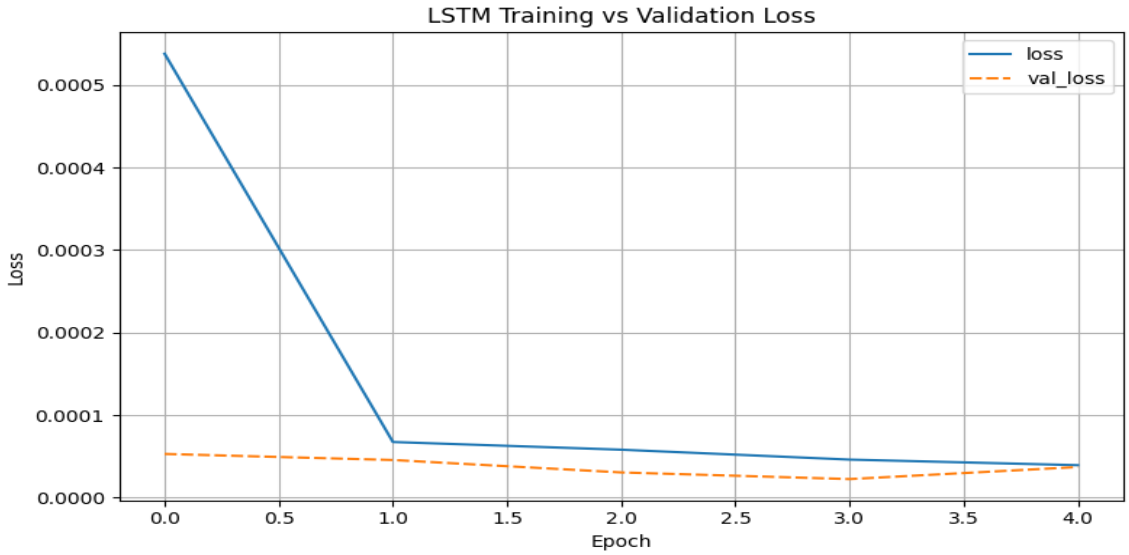
- Architecture:
  - Input: 21 features
  - Hidden Layers: 64 → 32 neurons (ReLU + Dropout 0.2)
  - Output: 1 neuron (linear)
- Training:
  - Loss: MSE
  - Optimizer: Adam (lr=0.001)
  - Batch Size: 16, Epochs: 5 (with early stopping)
  - L2 weight decay (1e-4) for regularization.



3. LSTM

Leverages temporal context across past 24 readings (4 hours):

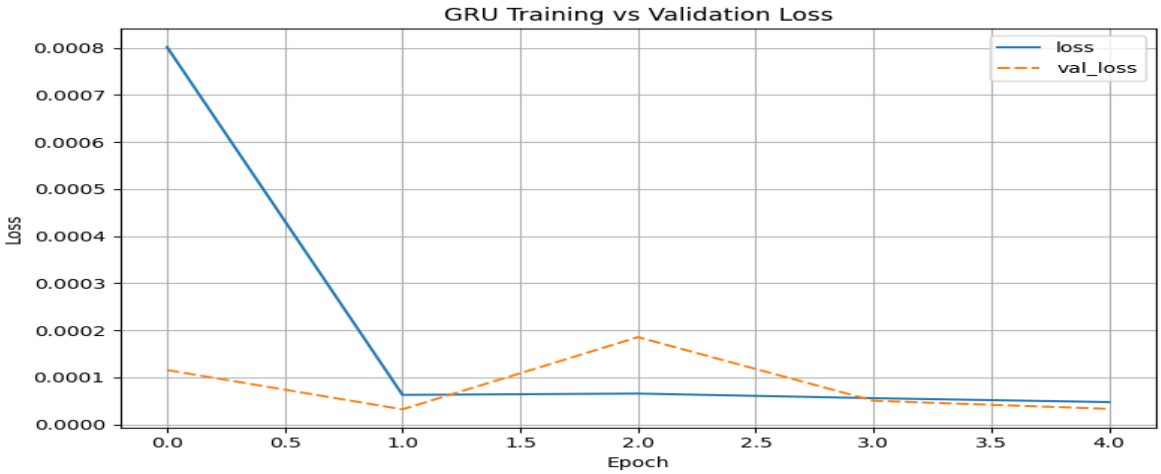
- Sequence Input: 24 × 21 feature windows
- Network:
    - LSTM(50 units) with input/recurrent dropout(0.2)
    - Dense→1 output
- Training:
    - Loss: MSE
    - Optimizer: RMSprop (lr=0.001)
    - Batch Size: 64, Epochs: 5 (10% validation split)

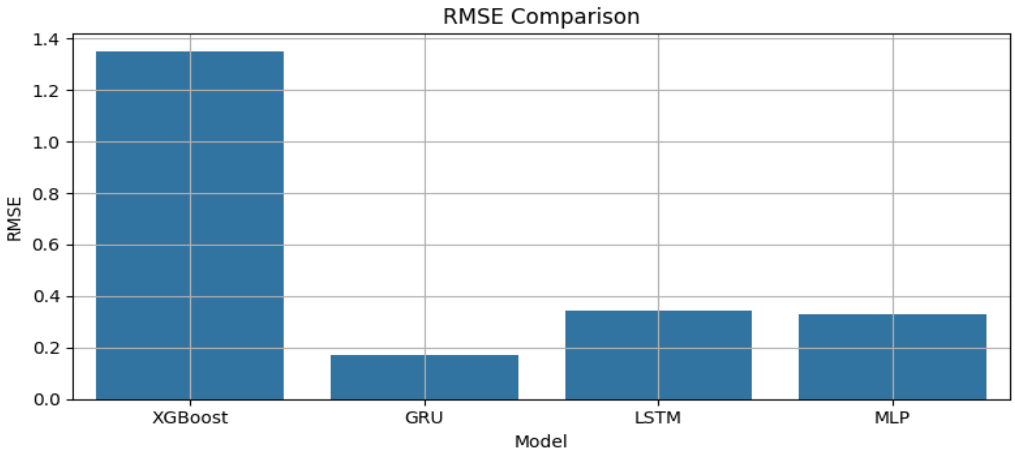


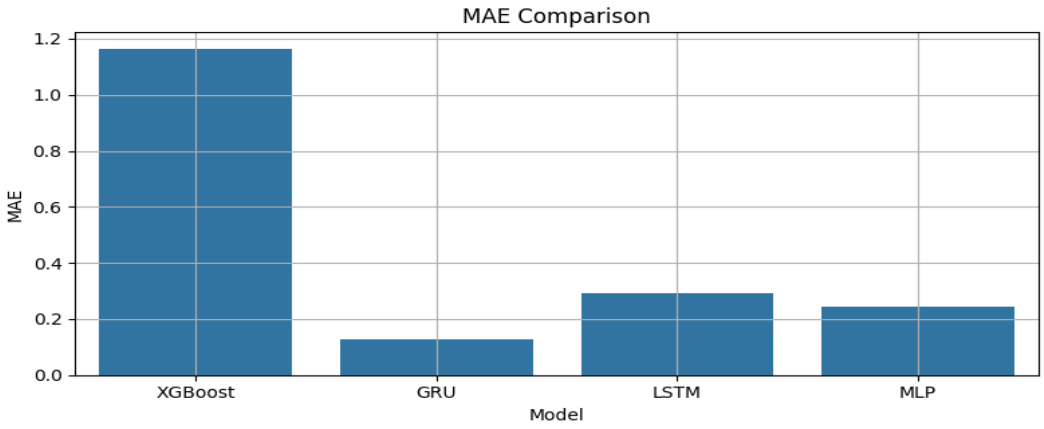
4. GRU

A streamlined RNN variant that trains faster with fewer parameters:

- Network: GRU(50 units) → Dense(1)
- Differences: Single “update” gate vs. LSTM’s separate gates (~25% fewer weights)
- Training: Identical schedule to LSTM; achieved ~20% faster convergence.



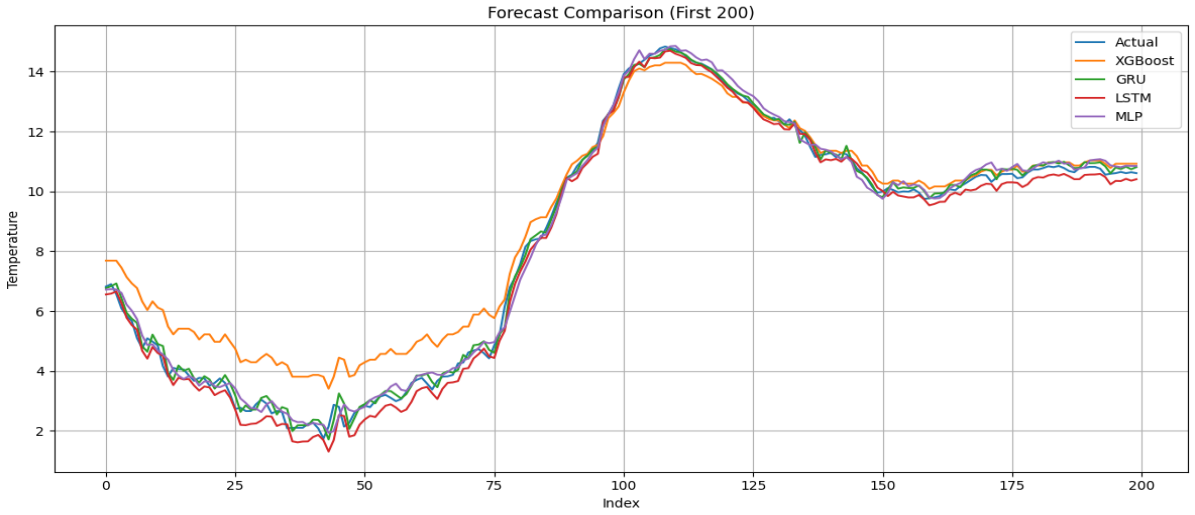




	Model	RMSE	MAE
0	XGBoost	1.350643	1.164520
1	GRU	0.171297	0.127374
2	LSTM	0.341820	0.290553
3	MLP	0.329145	0.244510

Model Evaluation & Comparison

Model	Model Test RMSE (°C)	Test MAE (°C)	Notes
XGBoost	1.35	1.164	Highest error among models
MLP	0.329	0.2445	Competitive but outperformed by GRU
LSTM	0.341	0.290	Moderate error capturing trends
GRU	0.171	0.127	Best overall performance



- *The 'First 200' on the title means we're plotting the first 200 test-set samples (the x-axis "Index" simply numbers each 10-minute time step). Over these 200 points, you can see that the GRU's forecast (green) tracks the actual temperature (blue) most closely, while XGBoost (orange) tends to overestimate early values.*

Conclusion:

- Key Findings: Tree-based XGBoost achieved highest error, while RNNs better modeled transient events.
- All the models performed relatively well on this dataset and GRU Outperformed all of them in both the losses.

- **Future Work**
    - Hyperparameter tuning via Bayesian optimization
    - Exploring Temporal Convolutional Networks or Transformer-based forecasters
    - Multi-step forecasts (beyond next-step) and probabilistic prediction intervals