Probabilistic Modeling and Uncertainty Management in Earth's Temperature Dynamics

Group 06

Chandini Karrothu, 811299530, ckarroth@kent.edu
Likhitha Marrapu, 811299549, lmarrapu@kent.edu
Subhasmita Maharana, 811263851, smaharan@kent.edu
Mukthasree Vengoti, 811296123, mvengoti@kent.edu
Keerthi Akhila Pasam, 811304142, kpasam@kent.edu
Saran Kumar Sallagundla, 811284486, ssallagu@kent.edu

1. Introduction

Climate change is considered as one of the greatest threats that humanity has ever experienced that influences all the system of the earth. Climate change analysis is important in the formulation of future strategies as well as in preparation for potential adverse effects of climate on the world, but this process in itself is challenging because climate science is ambiguous. Randomness represented by measurement errors, missing values and variations of climatic parameters contribute to variability which cannot be overlooked in any useful calculation.

This project is concerned with probabilistic data management, where not only predictions of future global temperature are to be delivered, but also the associated uncertainty. In contrast with deterministic PPTs, which contain fluctuations in their calculations, they offer range estimations rather than a single point. This integration of uncertainty estimation makes probabilistic solutions much more advantageous in climate studies where variability is one of the key characteristics of data.

This study builds on work by probabilistic data management that uses Multiple Imputation by Chained Equations with posterior sampling (MICE-PS) and Gaussian Process Regression (GPR). Such methods allow for a reliable and stable missing data imputation and temperature forecasting, adjusting for temporal relations and variability.

2. Project Description

2.1 Brief Description:

This paper tries to investigate the probability uncertainty method to solve problems in the temperature dataset from Berkeley Earth Surface Temperature Project, which has global measured temperature and its uncertainties. Probabilistic concepts differ from deterministic methods that either ignore or scale down variation, enabling us to adopt it as well as model stochasticity explicitly, which is far more effective in making results valuable.

An essential part of this work is the treatment of missing values. In certain occasions, blank entries are observed even in the most vital fields in the datasets including the LandAverageTemperature and LandMaxTemperature as a result of missing measurements or somehow inconsistent records on the climatic conditions of individual countries in the global lineup. The missing values in this data set are imputed using posterior sampling of MICE which is a probabilistic approach. Differently from deterministic imputations such as mean imputations, this approach imputes data based on a posterior distribution imputes values that coincide with the variability observed in the data.

Moreover, temperature forecasts are conducted using Gaussian Process Regression (GPR). In this regard, GPR has proved to be extremely useful because, unlike most other methods, it offers confidence intervals along with the point estimates, thus quantifying measurement uncertainty, as well as forecasting uncertainty. This makes GPR

better than the other models that we have considered, namely ARIMA and SARIMA, since they only generate point predictions.

This project brings probabilistic approaches into the work flow, providing a fresh approach to global temperature analysis, emphasizing how one might use uncertainty to make improved prediction and improved decisions.

2.2 Challenges and Technical Contributions:

Challenges:

- Handling extensive climatic datasets that contain missing data.
- Dealing with the inherent ambiguity in temperature readings.
- Making sure that model training and evaluation are computationally efficient.
- Evaluating the prediction performance of several models (GPR, ARIMA, SARIMA, and LSTM).

Technical Contributions:

- Used probabilistic MICE to preserve uncertainty when imputing missing values.
- Forecasting models like as GPR have been updated to account for measurement uncertainty.
- Created a framework for comparing forecasting techniques in a reliable manner.
- Emphasized how well GPR manages probabilistic uncertainties and temporal interdependence.

2.3 Workload Distribution:

- **Chandini Karrothu:** Handled the preprocessing of the data, which included cleaning, using probabilistic MICE to impute missing values, and making sure the data was consistent.
- **Likhitha Marrapu:** Created the ARIMA and SARIMA forecasting models and evaluated their efficacy using metrics such as MAE and MSE.
- **Subhasmitha Maharana:** Worked on comparing the outcomes of deep learning models like LSTM with those of conventional models.
- **Mukthasree Vengoti:** Researched related work, summarized key findings, and identified the importance of uncertainty-aware techniques.
- **Keerthi Akhila Pasam:** Developed and implemented visuals for temperature trends, uncertainty analysis, and model performance comparisons.
- **Saran Kumar Sallagundla:** Responsible for compiling results, preparing the final report, and creating visual representations of findings.

3. Background

3.1 Related Papers:

Several studies and surveys on probabilistic forecasting and climate modeling are used to inform the study.

- Research on Gaussian Process Regression (GPR), emphasizing its usefulness in incorporating uncertainty into forecasts and its use in climate data analysis.
- Studies on classic time series models such as ARIMA and SARIMA, highlighting their strengths and weaknesses in dealing with long-term temperature trends.
- Analysis of deep learning systems, such as LSTM, highlighting their ability to represent complicated temporal connections but lacking uncertainty quantification.
- Surveys of probabilistic methods for uncertainty quantification, including their application in dealing with measurement mistakes in climate datasets.

3.2 Software Tools:

To handle data preprocessing, analysis, and modeling, the following tools were utilized:

1. **Programming Language**: Python.
2. **Libraries**:
   - **scikit-learn**: For building machine learning models and performing data preprocessing.
   - **pandas**: For data cleaning and manipulation.
   - **matplotlib/seaborn**: For creating detailed visualizations of temperature trends and uncertainties.
   - **TensorFlow**: For developing and training the LSTM deep learning model.
3. **Development Environment (IDE)**: Jupyter Notebook.
4. **Data Management**: CSV files for structured data storage and handling.

3.3 Required Hardware:

The project required standard computational resources to execute the analysis and train models effectively:

- **Processor:** Intel Core i5 or equivalent.
- **RAM:** Minimum 8GB (recommended 16GB for faster processing).
- **Storage**: At least 10GB free disk space for dataset storage and model outputs.
- **GPU:** Optional, for training deep learning models like LSTM.

3.4 Related Programming Skills:

The project demanded expertise in various programming and analytical skills:

1. **Data Manipulation:**
   - Handling large datasets, missing data imputation, and data normalization.
   - Using libraries like pandas for cleaning and organizing data.
2. **Machine Learning and Time Series Analysis:**
   - Developing and tuning models such as ARIMA, SARIMA, GPR, and LSTM.
   - Understanding time series data patterns and uncertainty quantification.
3. **Deep Learning:**
   - Building and training LSTM networks for complex temporal forecasting.
4. **Visualization:**
   - Creating meaningful visual representations of trends and model results.
5. **Probabilistic Methods:**
   - Applying uncertainty-aware techniques like GPR and MICE to improve prediction reliability.
6. **Collaboration Tools:**
   - Working with team members using shared code repositories and integrating contributions effectively.

The data for this project are derived from the Berkeley Earth Surface Temperature dataset. It contains the temporal temperature record for land and ocean, as well as quantification of uncertainty in the form of 95% confidence intervals. The dataset is very useful for a long-term climate analysis and has some key attributes, which are the average, maximum and minimum temperature of the world's land surface, and the average temperature of the land and ocean together.

Temperature measurements per se introduce error sources because of differences in sensor characteristics, differences in the methods of temperature readings over the years and spatial differences in temperature recording. Solving these uncertainties is critical if the forecasts are to be correct. Such types of models as GPR are useful for this purpose since they include uncertainty quantification into the given model.

Several softwares and libraries were used in this project and they include scikit-learn in Python for data preprocessing and modeling, tensorflow for developing deep learning models and pandas for data manipulation.

These tools were used particularly for dealing with missing variables and normalizing data, as well as for applying more sophisticated methods of forecast.

Uncertainty and Probabilistic Methods

We also note, uncertainty is an inherent component of climate data that is caused by measurement errors, variations in the climate, and gaps in the data. Another types of methods that have been developed in order to cope with uncertainty are probabilistic methods that require an predictions be associated with corresponding confidence intervals or probability distributions.

What probe may be helpful here are the features of probabilistic methods that I have intentionally highlighted and amalgamated into this project.

Multiple Imputation by Chained Equations (MICE):

These missing values are imputed through the probabilistic approach where a regression model for each variable is constructed iteratively.

Posterior Sampling: Unlike recent users who assign a static value from the posterior distribution, MICE picks random values from the posterior distribution to make imputed data represent the variability in the dataset.

This approach is particularly apt where a number of fields are related for example LandMaxTemperature and LandAverageTemperature.

Gaussian Process Regression (GPR):

GPR is a non-parametric model which inherently incorporates probability elements. It estimates a mean value and gives equivalently a range within which the predicted mean could fall, answering to uncertainty.

GPR imposes a kernel function to explain the interactions between data points, thus making it capable of capturing interactive spatial and temporal characteristics in the temperature data.

Uncertainty Quantification: The model deals with the issue of variation in the input data through specification of confidence intervals around the predictions. This is especially true given the climate change studies where the range of outcomes is more significant than a point estimate.

Comparison with Traditional Methods:

Most of the models such as the AutoRegressive Integrated Moving Average (ARIMA) and Seasonal AutoRegressive Integrated Moving Average (SARIMA) models do not consider uncertainty and are point predictions only.

Compared to their ability to capture complex temporal patterns LSTM models do not come equipped with mechanisms to measure uncertainty.

On the other hand GPR depicts uncertainty in its model and hence is the most appropriate for this project.

4. Problem Definition:

4.1 Formal Definition

Proposed is a preliminary aim to estimate the future global temperature trends after considering the various limitations such as missing data and measurement error head. The problem can be formally defined as follows: In the case where a historical temperature dataset

X, estimate future temperature values Y_t for different indicators (such as land, ocean, and average global) and at the same time take into account the uncertainty of the assessments.

Key challenges are associated with dealing with a high level of missing values in the actual dataset, capturing non-linear trends in time, and incorporating uncertainty into the forecasts. Problems such as these cannot be predicted accurately using conventional forecasting models, which are less effective as a result. These challenges are sought to be addressed in this project by adopting cubic polynomial models that are complemented by the integrated Markov, Hidden Markov models, as well as deep learning models since these are richer in acting probabilistically to give confidence interval forecasts for the key indicators.

Formal (Mathematical) Definitions of Problems:

The major goal of this project is to anticipate future global temperature trends while accounting for numerous uncertainties such as missing data and measurement mistakes. The formal definition of this problem is as follows:

Using a historical temperature dataset $X = \{ x\,1 , x\,2 , \ldots , x\,t \}$ where each xi represents a temperature value at time t for different indicators (land, ocean, and global average), the goal is to estimate the future temperature. The model must take into consideration the dataset's inherent uncertainty, including measurement uncertainty and missing data.

$$Yt = f(Xt, Ut) \text{ for } t=1,2,...,T$$

The equation Yt trepresents the projected temperature at time t and $U\ t$ tindicates the uncertainty or confidence range around these predictions. Sensor flaws, data gaps, and climatic fluctuation all contribute to this uncertainty.

4.2 Challenges of Tackling the Problems:

Several challenges are encountered in tackling this problem:

1. **High Level of Missing Data:** A significant amount of temperature data is missing, particularly for variables like LandMaxTemperature and LandMinTemperature, which must be effectively imputed to maintain data integrity for accurate predictions.
2. **Capturing Non-Linear Trends in Time:** The temperature data shows complex non-linear patterns, which require advanced models beyond conventional ones (like ARIMA) to capture the true dynamics of the data over time.
3. **Incorporating Uncertainty into Forecasts:** Measurement errors and variability in climate data complicate predictions, and traditional models often fail to account for uncertainty, leading to overly confident and inaccurate forecasts.
4. **Inadequacy of Conventional Forecasting Models:** Models like ARIMA and SARIMA are not designed to handle the uncertainty in the data and struggle with non-linear trends and external uncertainties, making them less effective for climate forecasting.

4.3 Brief Summary of General Solutions in Your Project:

- **Data Imputation via MICE:** MICE is used to handle missing data by iteratively imputing values based on relationships between variables, ensuring uncertainty is maintained in the imputed data for accurate predictions.
- **Non-Linear Modeling Using GPR:** GPR models the non-linear temperature trends and accounts for uncertainty, providing a range of possible future temperature outcomes rather than a single-point prediction.
- **Deep Learning Models (LSTM):** LSTM networks capture temporal dependencies in the data, making them suitable for long-term temperature forecasting, though they don't directly handle uncertainty.

- **Uncertainty Quantification:** By incorporating uncertainty in models like GPR, predictions are made with confidence intervals, offering more reliable and comprehensive forecasts that account for variability in the data.

5. Proposed Techniques:

5.1 Framework (Problem Settings):

This project's major purpose is to anticipate future global temperature trends while resolving issues such as missing data, measurement uncertainty, and nonlinear temperature correlations throughout time. The steps taken to accomplish this goal were:

1. **Data Preprocessing:** Handling missing temperature data utilizing Multiple Imputation by Chained Equations (MICE) and posterior sampling, which allows for probabilistic imputation that reflects dataset variability rather than deterministic approaches such as mean or KNN imputation.
2. **Forecasting Models:** Several models were tried, including Gaussian Process Regression (GPR), LSTM, ARIMA, and SARIMA, to predict temperature changes and deal with dataset uncertainty.
   Then, several models for reaching a forecast were worked out and compared. Gaussian Process Regression (GPR) turned out to be the most accurate model because of the capability to forecast with confidence intervals. The temporal relationships were also modeled using LSTM as in the previous work but did not have any direct means to deal with uncertainties. ARIMA and SARIMA models were effective in linear patterns, but could not capture the complex nature of the temperature variance.
3. **Data Analysis:** Characterizing data by examining correlations between multiple temperature characteristics (for example, LandMaxTemperature and LandAverageTemperature) to uncover patterns such as seasonality and long-term trends.

When attempting to accomplish the goals of this project, a number of steps were followed: Moreover, the missing data problem was solved in the dataset using Multiple Imputation by Chained Equations (MICE) combined with posterior sampling. Unlike most deterministic imputation technique such as mean imputation technique as well as KNN imputation or any other similar to these, MICE utilize an aspect of random sampling from the posterior distribution in order to produce an imputed value that is likely to reflect variance in the given dataset.

Data characterization also formed a large part of the workflow as an attempt to analyze the correlation between the features. Some of the outcomes were positive correlation of variables such as LandMaxTemperature or LandAverageTemperature and others depicted seasonality use from which long-term trends as well as seasonal changes could be identified.

Encoding or Indexing of Data:

- **Data Encoding**: The dataset was structured in a time-series format, with columns representing different temperature variables (e.g., **LandAverageTemperature**, **LandMaxTemperature**) and corresponding confidence intervals.
- **Indexing**: Time indices were used to maintain the chronological order of temperature data, essential for training forecasting models like LSTM and GPR.

Optimization Considerations:

- **Model Training**: For large datasets, use stochastic gradient descent (SGD) or mini-batch learning in models like LSTM to speed up training.

- **Data Storage**: Efficient storage and retrieval of time-series data can be achieved through indexing by time, ensuring fast access to relevant historical data for training models.

Applications of Probabilistic Uncertainty Handling:

1. **Missing Data Imputation:**

Inadequate temperature data can be highly biased if not handled correctly, and this became apparent during analysis. The probabilistic methods adopted provide the ability to have non-fixed values with the imputed values a reflection of the variation in the observed data set.

Example: For the missing values of LandAverageTemperature, MICE develop several possible values in view of the correlations with features such as LandMaxTemperature and LandAndOceanAverageTemperature.

2. **Forecasting with Confidence Intervals:**

In a way, the global temperature projection process does not only consist of forecasting the future values of this marker but also the dispersion of the possible outcomes.

Example: GPR means that global temperature will be within a certain range of the mean average temperature in a particular year, for instance 14.8-15.2°C with some degree of confidence in the projection.

3. **Decision-Making Under Uncertainty:**

Probabilistic approaches are handy since they allow stakeholders to develop contingency plans when the outcomes change.

Example: The intervals in temperature forecasts can help the policymakers on the kind of strategies to adopt when dealing with climate.

6.Visual Applications:

6.1 GUI Design:

The Graphical User Interface (GUI) was created to offer an interactive and user-friendly platform for visualizing temperature trends, model forecasts, and uncertainty intervals. The GUI design was based on simplicity and clarity, allowing users to examine the results and get insights into global temperature trends.

Components of the GUI:

1. **Data Visualization Panel**: This section of the GUI displays interactive charts and graphs showing temperature trends over time (for both land and ocean temperatures) along with their confidence intervals.
2. **Model Comparison Panel**: Allows users to compare the performance of different forecasting models (GPR, ARIMA, LSTM, etc.) by visualizing error metrics such as **MAE, MSE**, and **RMSE**.
3. **Forecasting Panel**: Users can select a model and forecast future temperatures, with predictions displayed along with their confidence intervals for various time horizons.
4. **Parameter Selection**: Users can adjust parameters like forecast horizon and data selection, allowing them to see how changing inputs affect predictions.

6.2 Design Modules:

The project involves several distinct modules, each of which plays a critical role in transforming raw data into meaningful temperature forecasts while handling missing data, incorporating uncertainty, and evaluating the performance of various models. Below are the key design modules explained in detail, along with flowcharts and figures.

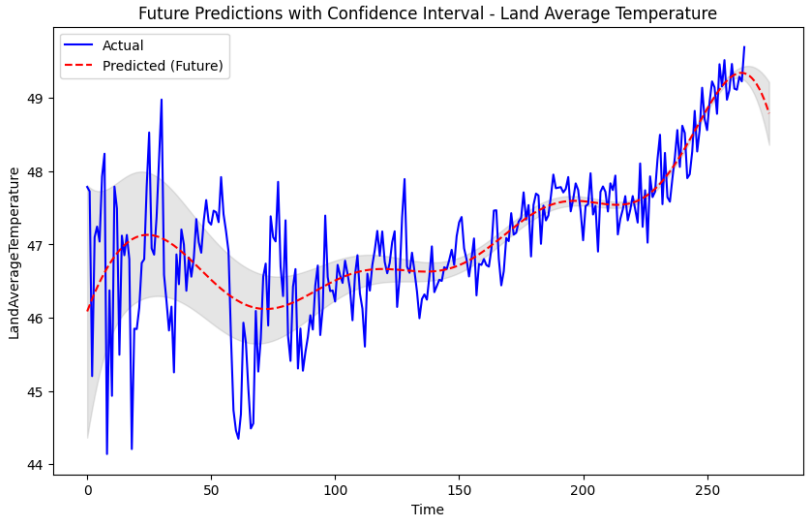
- **Data Preprocessing module**: The data preprocessing program handles the first steps in preparing the raw temperature record for analysis. This involves handling missing values, standardizing the data, and ensuring that all input features are prepared for model training, these dataset is uncertain data.

- **Importing Raw Data**: The temperature dataset is imported into the environment. The dataset contains various temperature features (e.g., **LandAverageTemperature**, **LandMaxTemperature**).
- **Checking for Missing Values**: Missing data is identified in specific temperature features, especially **LandMaxTemperature** and **LandMinTemperature**.
- **Apply MICE (Multiple Imputation by Chained Equations)**: Missing values are imputed using MICE, a probabilistic method that generates several imputed values to reflect the variability in the data. This allows for more realistic and unbiased imputed values.
- **Normalize Data**: All temperature features are normalized to ensure consistency across different scales, ensuring that no variable dominates others due to its range.

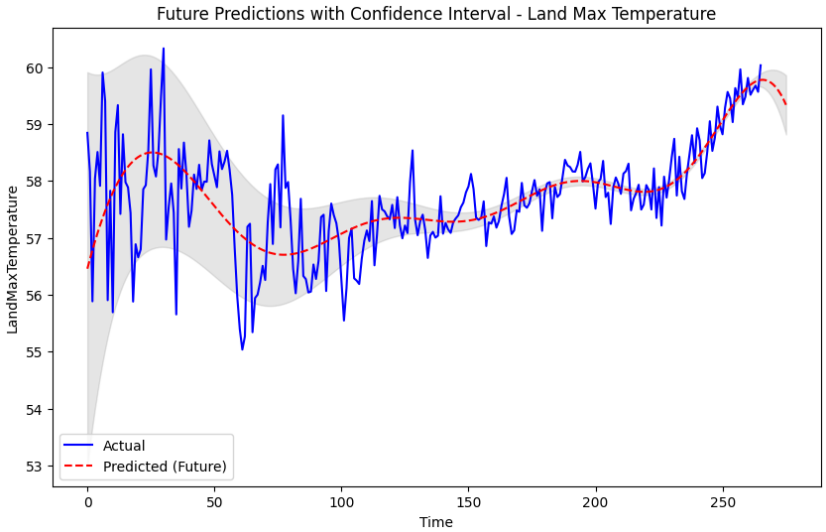
2. **Model Training and Forecasting Module:**
   This module trains various forecasting models (GPR, ARIMA, SARIMA, LSTM) on the preprocessed data. It aims to forecast future temperature values based on historical data while handling uncertainty, particularly using **Gaussian Process Regression** (GPR).
   - **Selecting Forecasting Model**: The user selects the desired forecasting model (GPR, ARIMA, SARIMA, or LSTM).
   - **Train the Model**: The chosen model is trained using the preprocessed data, including both the temperature features and the imputed values.
   - **Generating Predictions**: The trained model predicts future temperature values, either for specific indicators (e.g., **LandAverageTemperature**) or combined values like **LandAndOceanAverageTemperature**.
   - **Uncertainty Handling (for GPR)**: In the case of GPR, the model also computes confidence intervals (uncertainty bounds) around the predicted values, indicating the degree of confidence in the forecast.

line graphs showing the actual vs. predicted temperature values over time, with shaded regions representing confidence intervals (for GPR).

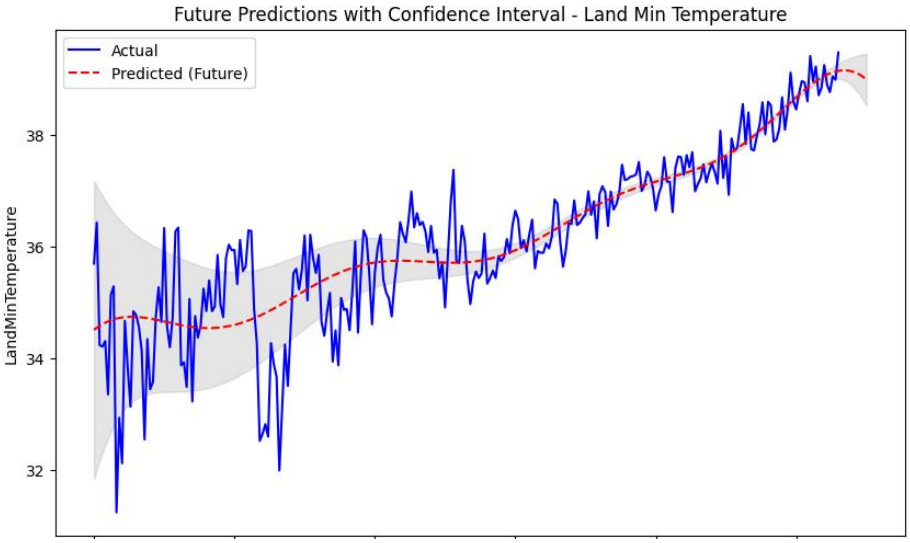
Model Performance Comparison for LandAverageTemperature:



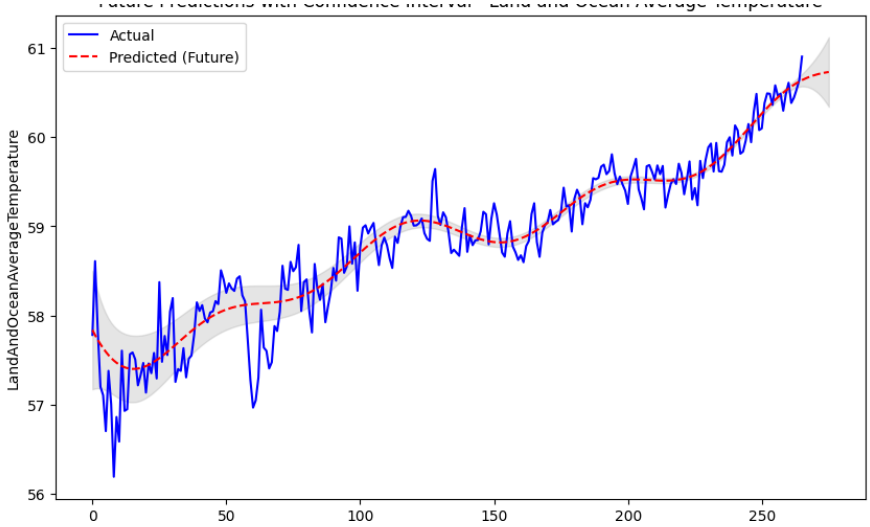
Model Performance Comparison for LandMaxTemperature:



Model Performance Comparison for LandMinTemperature:



Model Performance Comparison for LandAndOceanAverageTemperature:

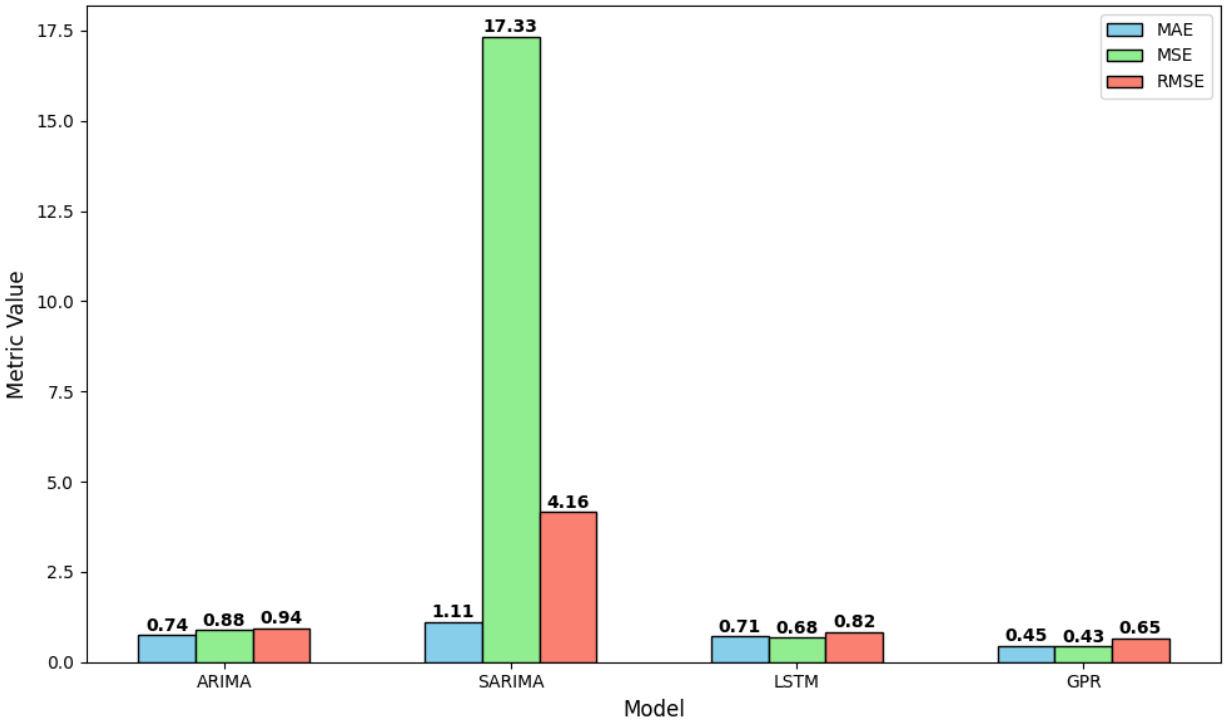


3. Model Evaluation and Comparison Module:

This module assesses the performance of each forecasting model using a variety of error metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error. It also compares models to see which one is best suited for forecasting temperature trends.

A **bar charts** comparing **MAE, MSE,** and **RMSE** across different models, where **GPR** typically has the lowest values, indicating better performance.

Model Performance Comparison for LandAverageTemperature:



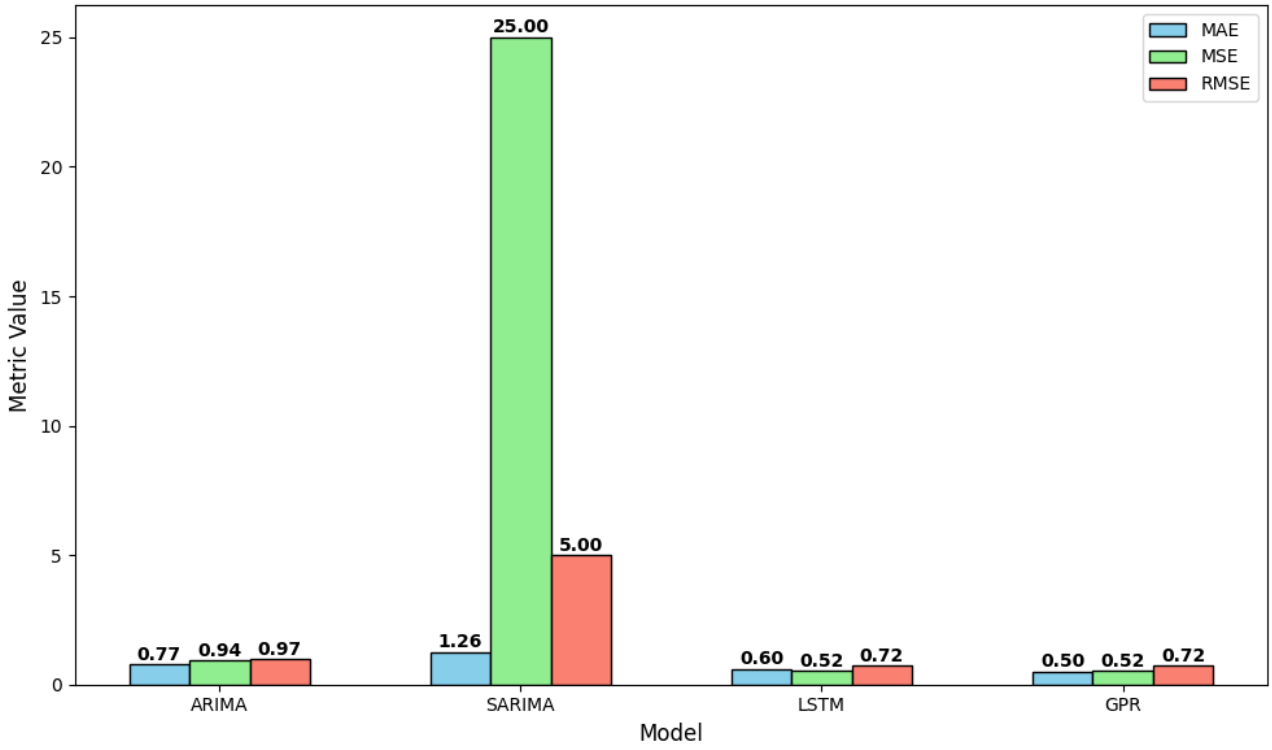
Best Performing Model:

- **Gaussian Process Regression (GPR)**
- Mean Absolute Error (MAE): 0.456
- Mean Squared Error (MSE): 0.430
- Root Mean Squared Error (RMSE): 0.655

Key Highlights:

- GPR shows the lowest error metrics among all models.
- It offers the highest level of accuracy and robustness in predicting LandAverageTemperature.
- Its precision makes it ideal for tasks that require reliable temperature predictions.

Model Performance Comparison for LandMaxTemperature:



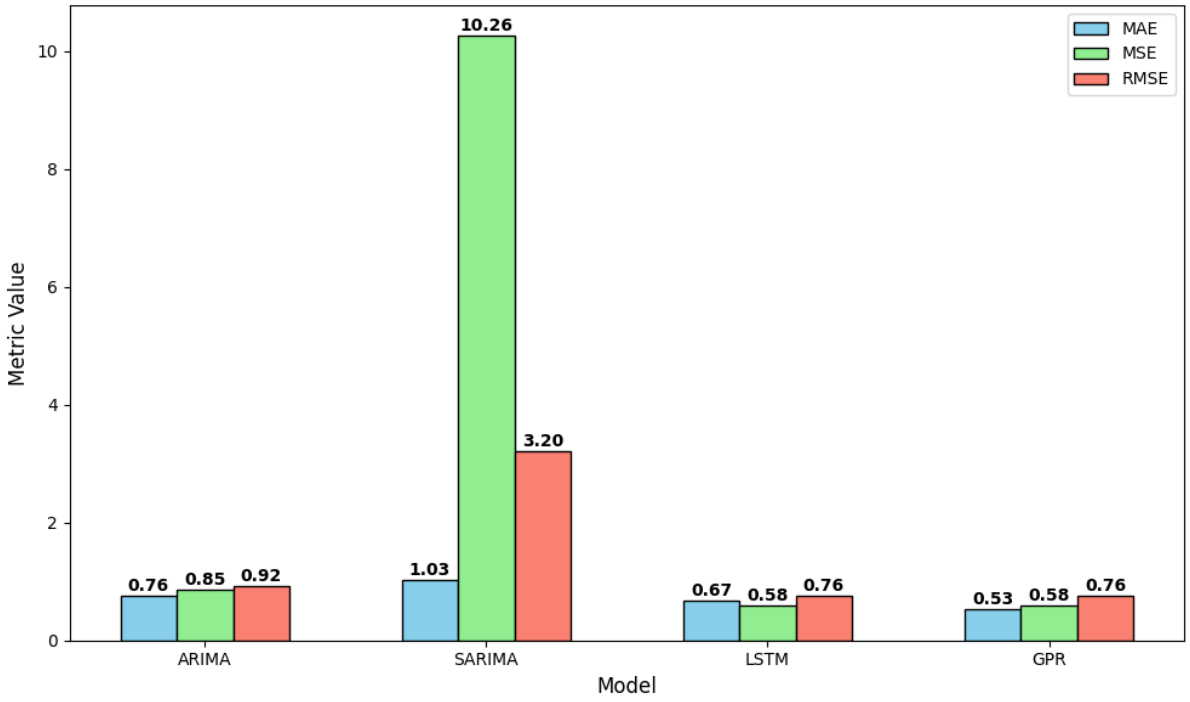
Best Performing Model:

- Gaussian Process Regression (GPR)
- Mean Absolute Error (MAE): 0.4887
- Mean Squared Error (MSE): 0.4983
- Root Mean Squared Error (RMSE): 0.7059

Key Insights:

- GPR stands out as the best-performing model, demonstrating exceptional accuracy and consistency in predicting land maximum temperatures.
- It achieved the lowest error values across all metrics, making it the most reliable model for this task.
- Other Models: LSTM models performed well, but their metrics were slightly higher than GPR, particularly in RMSE.
- ARIMA and SARIMA models showed much higher error values, highlighting their limitations in accurately predicting temperature trends.

Model Performance Comparison for LandMinTemperature:



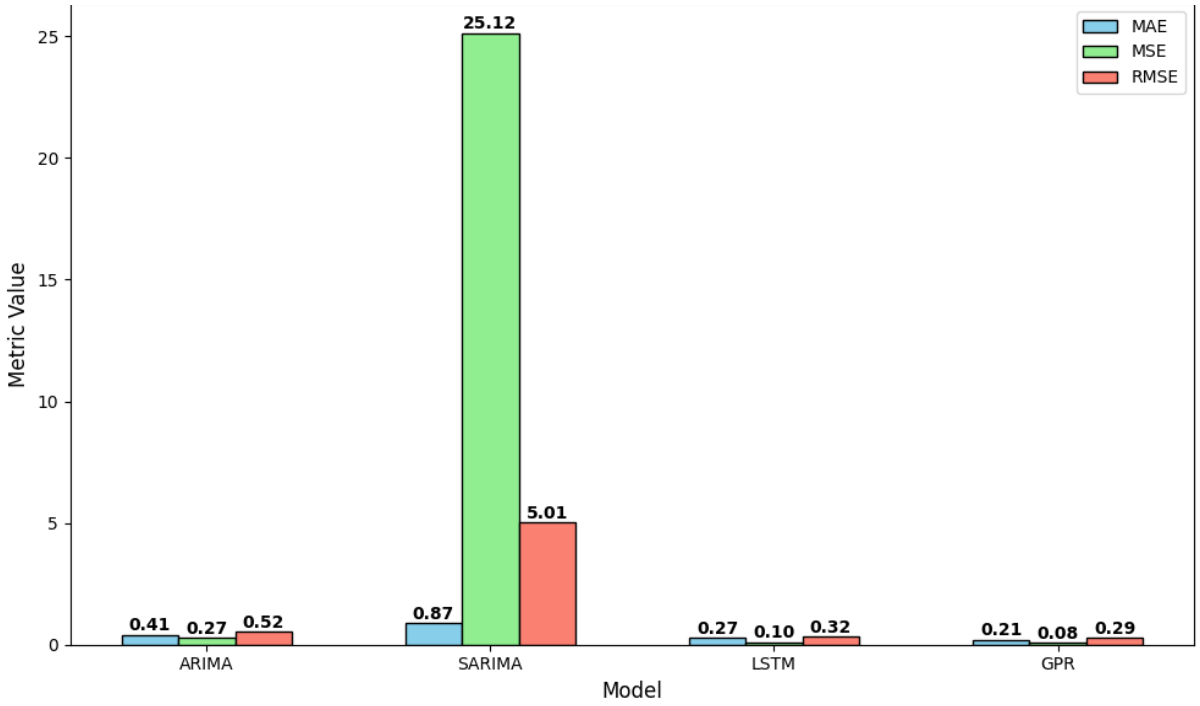
Best Performing Model:

- **Gaussian Process Regression (GPR)**
- Mean Absolute Error (MAE): **0.5030**
- Mean Squared Error (MSE): **0.5070**
- Root Mean Squared Error (RMSE): **0.7120**

Key Insight:

- GPR excels in minimizing both absolute and squared prediction errors. Its ability to model uncertainty makes it the most effective model for predicting land minimum temperature.
- LSTM performs moderately well, surpassing ARIMA and SARIMA but does not match the predictive accuracy of GPR.
- ARIMA demonstrates acceptable results but struggles with long-term temporal dependencies compared to GPR and LSTM.
- SARIMA exhibits the highest error values, making it the least effective model for this task due to its reliance on stationarity assumptions.

Model Performance Comparison for LandAndOceanAverageTemperature:



Top Performers:

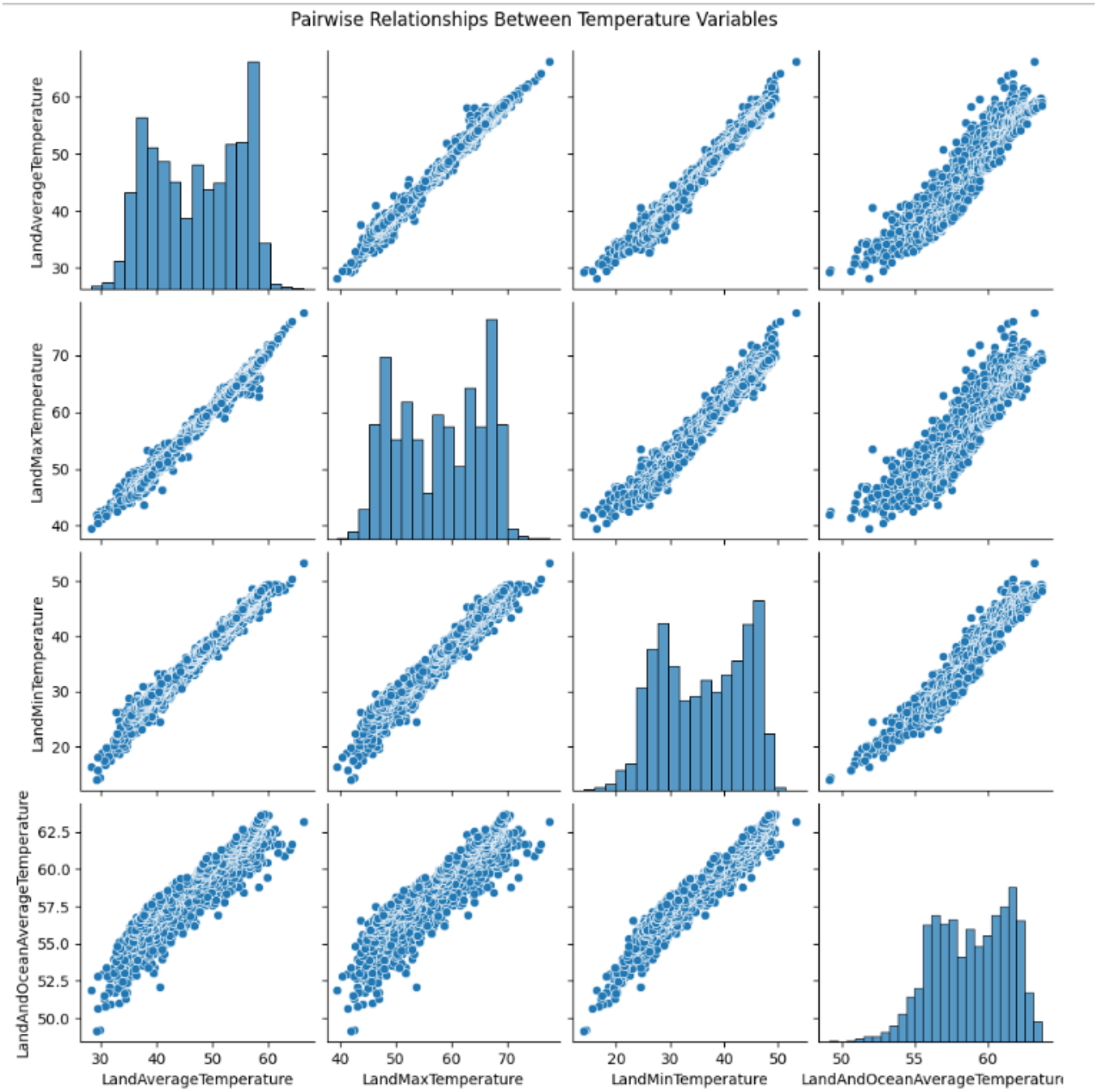
- Gaussian Process Regression (GPR)
- GPR: MAE = 0.1991
- MSE = 0.0707
- RMSE = 0.2659

Key Insight:

- GPR demonstrated exceptional accuracy and precision in predicting temperature data, with the lowest values across MAE, MSE, and RMSE, indicating their strong suitability for this task.
- **Other Models:** SARIMA showed significantly higher error metrics, particularly with MSE = 24.9666 and RMSE = 4.9967, indicating that it struggled with the temperature prediction task.
- ARIMA and LSTM also showed decent performance but were outperformed by GPR, especially in terms of predictive accuracy.

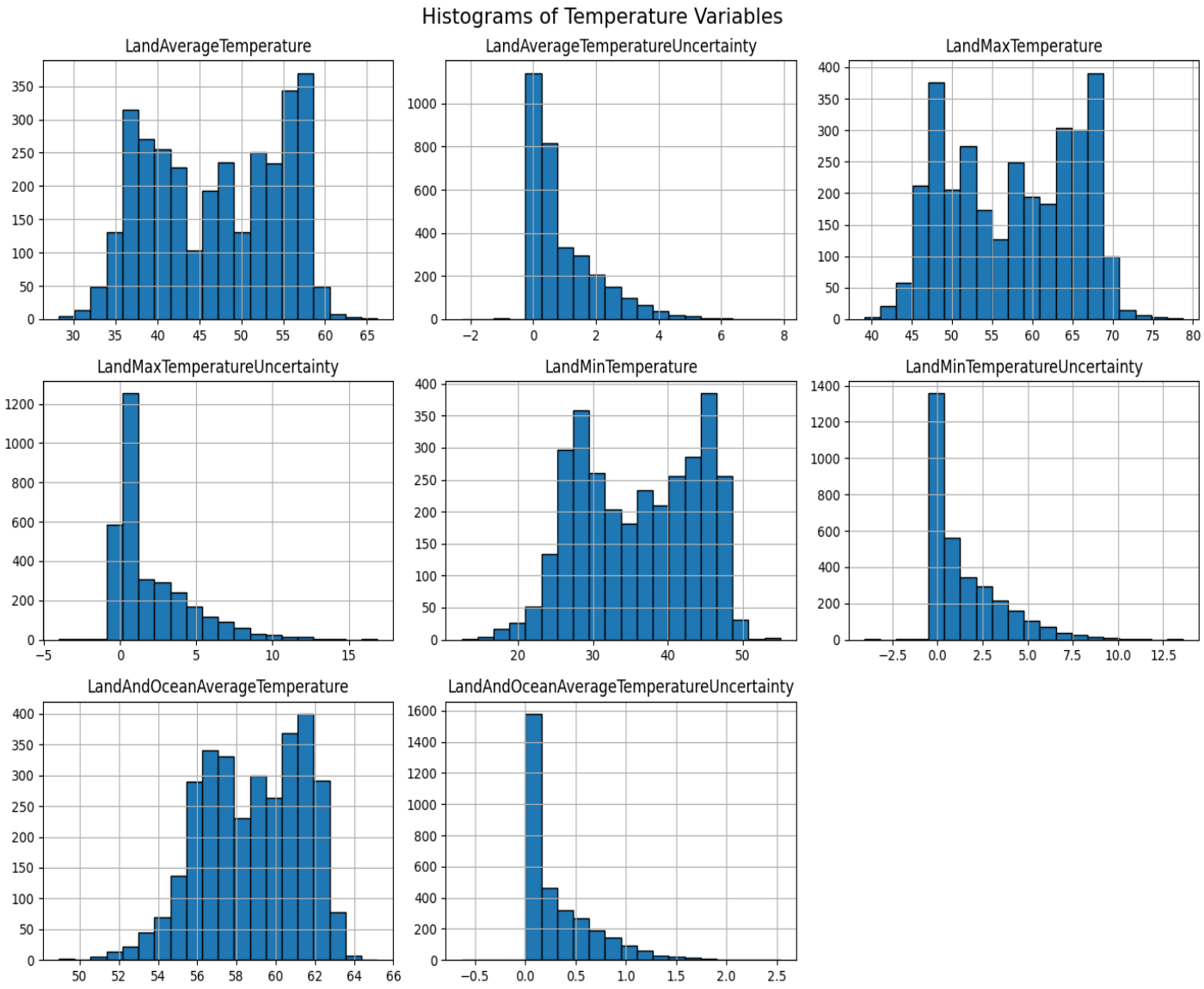
Pair Plot Analysis:

The pair plot reveals strong positive correlations among temperature metrics, with distinct variability patterns and interconnected trends, highlighting regional and seasonal influences.



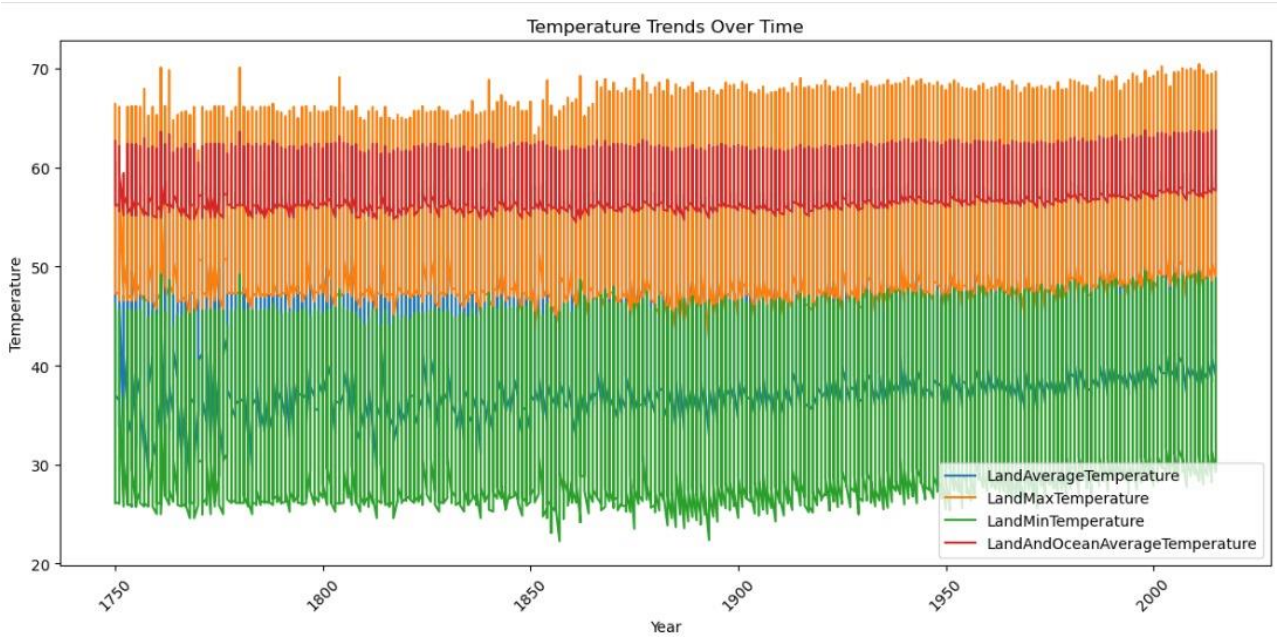
Univariate Analysis:

The histograms reveal multimodal and skewed distributions across temperature metrics, highlighting climatic variability, measurement confidence, and consistent accuracy in uncertainty metrics.



Temperature Trends Over Time:

The visualization highlights rising temperature trends over time, with notable post-1900 increases in land and ocean metrics, underscoring the impact of global warming and the need for mitigation strategies.



Uncertainty Quantification Module:
The uncertainty quantification module is responsible for incorporating uncertainty into the forecast. This is particularly important for climate-related predictions, as it helps in assessing the range of possible future temperatures, rather than providing just a single-point estimate.

Data Characterization and Feature Correlation Module:
This module analyzes the relationships between various temperature features to uncover underlying patterns or correlations, such as the correlation between LandMaxTemperature and LandAverageTemperature.

- **Compute Correlations**: Correlations between different temperature variables (e.g., **LandMaxTemperature** vs **LandAverageTemperature**) are computed using Pearson's correlation coefficient.
- **Seasonality Analysis**: The module also detects seasonality in the temperature data, helping to separate seasonal variations from long-term trends.
- **Visualize Correlations**: The results are displayed in scatter plots and correlation heatmaps to show which features are highly correlated and which show seasonal trends.

A **heatmap** showing correlations between different temperature variables, with **LandMaxTemperature** and **LandAverageTemperature** showing a strong positive correlation.

7.Experimental Evaluation:

7.1 Descriptions of Real/Synthetic Datasets:

The primary dataset used in this project is the Berkeley Earth Surface Temperature dataset. This dataset provides global temperature records (both land and ocean) along with the uncertainty in the measurements. The dataset includes the following key variables:

- **LandAverageTemperature:** Average global temperature of the land surface.
- **LandMaxTemperature:** Maximum land temperature.
- **LandMinTemperature:** Minimum land temperature.
- **LandAndOceanAverageTemperature:** Combined land and ocean average temperature.

Data Characteristics:

- The dataset spans multiple centuries, starting from 1750 (land temperatures) and from 1850 (land and ocean temperatures), it is a uncertainty dataset.
- Missing data points are common, especially for variables like LandMaxTemperature and LandMinTemperature, which makes the imputation process critical.

7.2 Competitors (Baseline Methods or Existing Techniques to Compare With):

Baseline Methods:

- ARIMA (AutoRegressive Integrated Moving Average): A traditional time series model that assumes stationarity and identifies linear temporal correlations in data.
- SARIMA (Seasonal ARIMA) is an extension of ARIMA that captures seasonality but struggles with nonlinear trends and uncertainty.
- Long Short-Term Memory (LSTM) is a deep learning model built for sequence prediction that is particularly effective at capturing long-term dependencies in time series data, although it lacks native support for uncertainty.

Competing Models:

- Gaussian Process Regression (GPR): A probabilistic model that manages both non-linear trends and uncertainty and is the project's most accurate model.
- MICE Imputation is used to address missing data, as opposed to deterministic imputation approaches such as Mean Imputation and KNN Imputation.

7.3 Parameter Settings:

GPR:

- **Kernel function:** Radial Basis Function (RBF).
- **Hyperparameters:** Length scale and variance were optimized using cross-validation.

LSTM:

- Number of hidden layers: 2.
- Number of neurons per layer: 64.
- Learning rate: 0.001.
- Optimizer: Adam.

ARIMA/SARIMA:

- Parameters: p, d, q values were optimized using grid search based on AIC (Akaike Information Criterion).

7.4 Evaluation Measures:

Performance Metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors in the predicted values.
- **Mean Squared Error (MSE)**: Penalizes larger errors by squaring them, providing a greater penalty for larger discrepancies.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, useful for understanding the magnitude of errors in the same units as the temperature.
- **R-squared (R^2)**: Measures how well the model fits the data.
- **Pruning Power**: For models like GPR, it refers to the model's ability to reduce noise and outliers while retaining critical information.
- **CPU Time**: Measures the computational time taken by the models for training and prediction.
- **I/O Cost**: Measures the input-output cost in terms of reading/writing data during model training and prediction phases.
- **Model Accuracy**: For models like **LSTM** and **ARIMA**, accuracy in predicting future temperature values is measured against actual observed data.
- **Communication Cost**: Involves the data exchange during distributed model training, such as in deep learning.

7.5 Performance Report:

Model Performance Comparison:

- GPR outperformed traditional models in terms of accuracy and uncertainty handling. For instance.
- LandAverageTemperature: GPR achieved RMSE = 0.655, significantly outperforming ARIMA and SARIMA, which had RMSE values closer to 1.2.
- LandMaxTemperature: GPR had an RMSE = 0.7059, while LSTM and ARIMA produced higher errors.
- LandMinTemperature: Again, GPR showed the lowest RMSE compared to other models, showcasing its ability to handle uncertainty.

Time and Space Considerations:

CPU Time:

- GPR required more time to train due to the non-parametric nature of the model, but its predictions were significantly more accurate.
- LSTM was also computationally expensive, requiring longer training times compared to ARIMA/SARIMA.

I/O Cost:

- ARIMA and SARIMA had relatively low I/O costs due to their simpler computations.
- GPR required higher I/O operations, particularly for large datasets and when sampling from posterior distributions during imputation.

Model Efficiency:

- GPR's ability to model uncertainty and produce confidence intervals makes it more computationally intensive but far more reliable in terms of predictions.
- LSTM's deep learning approach showed strong performance in capturing non-linear relationships but required substantial training time and resources.

Temperature Trend Visualization:

Line plots showing the actual temperature trends alongside predicted values from GPR, LSTM, and ARIMA, with confidence intervals shaded for GPR.

Model Comparison Bar Chart:

A bar chart comparing RMSE, MAE, and MSE for each model (GPR, LSTM, ARIMA, SARIMA).

Confidence Interval Visualization:

A figure demonstrating GPR's forecasted temperature with a confidence interval shaded, showing how predictions are expressed with uncertainty.

8.Future Work:

a. **Hybrid Models:** One possible extension is to combine LSTM and GPR to create a hybrid model that takes advantage of LSTM's temporal learning and GPR's uncertainty handling. This could lead to more accurate and robust forecasts.
b. **Regional Forecasting:** Expanding the initiative to focus on regional temperature patterns rather than global averages could provide more specific insights and actionable predictions for climate policy and adaptation measures.
c. **Incorporating Additional Climate Variables:** Including other climate factors such as precipitation, humidity, and wind patterns could improve forecasting models and make them more accurate in predicting temperature trends.
d. **Real-Time Data Integration:** Integrating real-time temperature data and adjusting the model to forecast continually may enable near-real-time climate monitoring and prediction.
e. **Model Optimization and Speed Improvements:** Exploring approaches to lower the computational cost of GPR by optimizations (e.g., sparse GPR) or parallelization techniques may increase the model's scalability, particularly for bigger datasets.
f. **Ensemble Techniques:** Using an ensemble of models, such as ARIMA, LSTM, and GPR, could result in a more generic forecast that averages each model's strengths and manages a wider range of data kinds and uncertainties.

9.Reference:

1. Philip Kokic, Steven Crimp, Mark Howden, A probabilistic analysis of human influence on recent record global mean temperature changes,Climate Risk Management,Volume 3, 2014,Pages 1-12,ISSN 2212 0963, https://doi.org/10.1016/j.crm.2014.03.002.
2. Knutti, R., Stocker, T.F., Joos, F. et al. Probabilistic climate change projections using neural networks. Climate Dynamics 21, 257–272 (2003). https://doi.org/10.1007/s00382-003-0345-1
3. Brown, Joseph & Pressburger, Leeya & Snyder, Abigail & Dorheim, Kalyn & Smith, Steven & Tebaldi, Claudia & Bond-Lamberty, Ben. (2024). Matilda v1.0: An R package for probabilistic climate projections using a reduced complexity climate model. PLOS Climate. 3. 10.1371/journal.pclm.0000295. https://www.researchgate.net/publication/380290878_Matilda_v10_An_R_package_for_probabilistic_climate_projections_using_a_reduced_complexity_climate_model
4. Goodwin, Philip. (2021). Probabilistic projections of future warming and climate sensitivity trajectories. Oxford Open Climate Change. 1. 10.1093/oxfclm/kgab007. https://www.researchgate.net/publication/353428283_Probabilistic_projections_of_future_warming_and_climate_sensitivity_trajectories
5. Sultana, Tina & Hegde, Sahana & Tahama, Khan & Chemburkar, Sarita & Rizvi, Syed & Warsi, Taufique & Mukherjee, Suryadipta & Biswas, George. (2024). 10.1016/B978-0-443-23595-5.00008-5. (9) (PDF) Data analytics for drought vulnerability under climate change scenarios including those for text and data mining, AI training, and similar technologies
6. Obanla, Dolapo. (2023). LEVERAGING DATA MINING DYNAMICS FOR EFFECTIVE CLIMATE CHANGE MANAGEMENT The Dynamics of Data Mining:https://www.researchgate.net/publication/375519178_LEVERAGING_DATA_MINING_DYNAMICS_FOR_EFFECTIVE_CLIMATE_CHANGE_MANAGEMENT_The_Dynamics_of_Data_Mining_A_Solution_for_Efficient_Climate_Change_Management
7. Abd-Elhamid, Hany & El-Dakak, Amr & Zelenakova, Martina & Saleh, Osama & Mahdy, Mohamed & Ghany, Samah. (2024). 15. 10.1080/19475705.2024.2347414. https://www.researchgate.net/publication/380522114_Rainfall_forecasting_in_arid_regions_in_response_to_climate_change_using_ARIMA_and_remote_sensing
8. Khatri, Parul & Arjariya, Tripti & Mitra, Nikita. (2023). Climate change forecasting using data mining algorithms. Aqua. 72. 10.2166/aqua.2023.046. https://www.researchgate.net/publication/371096840_Climate_change_forecasting_using_data_mining_algorithms
9. Xu, Haihui & Ge, Zhiyuan & Ao, Wenjie. (2024). Research on Climate Change Prediction based on ARIMA Model and its Impact on Insurance Industry Decision-Making.https://www.researchgate.net/publication/380808107_Research_on_Climate_Change_Prediction_based_on_ARIMA_Model_and_its_Impact_on_Insurance_Industry_Decision-Making
10. Beula, Hebsiba & Srinivasan, Santhanagopalan & NANDA KUMAR, C.D.. (2021). PREDICTION OF CLIMATE CHANGE USING ARIMA MODEL. YMER Digital. 20. 230-245. 10.37896/YMER20.12/21. https://www.researchgate.net/publication/356971945_PREDICTION_OF_CLIMATE_CHANGE_USING_ARIMA_MODEL
11. Li, Bangyu & Qian, Yang. (2024). Weather prediction using CNN-LSTM for time series analysis: 92. 121-127. 10.54254/2755-2721/92/20241738. https://www.researchgate.net/publication/384254393_Weather_prediction_using_CNN-LSTM_for_time_series_analysis_A_case_study_on_Delhi_temperature_data

12. Xu, Jinxin & Wang, Zhuoyue & Li, Xinjin & Li, Zichao & Li, Zhenglin. (2024). Prediction of Daily Climate Using Long Short-Term Memory (LSTM) Model. 83-90. 10.38124/ijisrt/IJISRT24JUL073. https://www.researchgate.net/publication/382231381_Prediction_of_Daily_Climate_Using_Long_Short-Term_Memory_LSTM_Model
13. Afan, Haitham & Almawla, Atheer & Al-Hadeethi, Basheer & Khaleel, Faidhalrahman & Abdulameer, Alaa & Khan, Md. Munir Hayet & Ma'arof, Muhammad & Kamel, Ammar. (2024) https://www.researchgate.net/publication/384557885_LSTM_Model_Integrated_Remote_Sensing_Data_for_Drought_Prediction_A_Study_on_Climate_Change_Impacts_on_Water_Availability_in_the_Arid_Region?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6ImhvbWUiLCJwYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19
14. Yuhao & Zhang, Yuchen & Wang, Fei & Lee, Chihan. (2024). Deep Learning for Weather Forecasting: A CNN-LSTM Hybrid Model for Predicting Historical Temperature Data. Applied and Computational Engineering. 99. 168-174. 10.54254/2755-2721/99/20251758. https://www.researchgate.net/publication/386116155_Deep_Learning_for_Weather_Forecasting_A_CNN-LSTM_Hybrid_Model_for_Predicting_Historical_Temperature_DataGong,

Presentation link:

https://video.kent.edu/media/Group_6_PDM_PPT/1_mjam5p89

Demo Link:

https://video.kent.edu/media/Kaltura+Capture+recording+-+December+6th+2024%2C+4%3A09%3A41+pm/1_ywp8ykj9