Time Series Analysis [REPO LINK]

Bhubaneswar Temperature Forecast

T AND T -LAB
CSE-3
Team:-TSA_23

Contributors

- Sayak Hatui
  - *2105236*
- Anulipi Jana
  - *21051974*
- Anurupa Saha
  - *21051975*
- Adarsh Mishra
  - *21053250*

Abstract

This research uses time series analysis to forecast monthly average surface temperatures in Bhubaneswar, India. The major goal of combining techniques from the SARIMAX and LSTM models is to compare their efficacy in predicting future temperature trends. The process includes data collection, rigorous cleaning, exploratory data analysis (EDA), model creation with SARIMAX and LSTM techniques, and extensive evaluation. The investigation shows that the LSTM model achieves a marginally lower Root Mean Squared Error (RMSE) than the SARIMAX model, indicating that it has the potential to provide higher forecasting accuracy in predicting monthly average surface temperatures. This study emphasizes the importance of time series analysis in providing valuable insights for decision-making in sectors such as agriculture, water resource management, and energy planning, as well as the importance of choosing appropriate modeling techniques tailored to the characteristics of the data.

Acknowledgement

We extend our sincere gratitude to Dr. Soumya Ranjan Mishra, our esteemed professor at KIIT University, for providing us with the opportunity to undertake this project as part of the T&T lab curriculum. His guidance, support, and insightful feedback throughout the project were invaluable in shaping our understanding and approach.

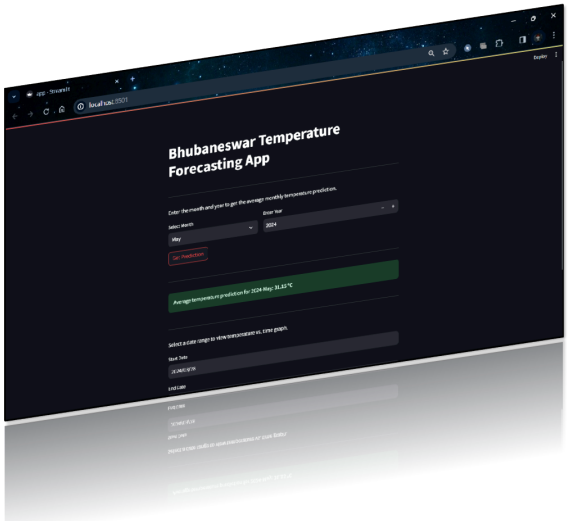
We would like to express our appreciation to Berkeley Earth for providing the temperature data used in this study. Their commitment to open data access and scientific research has greatly facilitated our analysis and enriched the outcomes of this project.

Additionally, we are thankful to all individuals who contributed to the completion of this project through their assistance, encouragement, and constructive discussions. Their collective efforts have been instrumental in the successful execution of this endeavor.

Content

Introduction

Surface temperature forecasting is critical in a variety of areas, including agriculture, water resource management, and energy planning. Its significance stems from its capacity to provide significant insights and foresight into approaching temperature trends, allowing for more informed decision-making and strategic planning.



In agriculture, precise surface temperature forecasts provide farmers with vital information for optimal planting periods, crop selection, and irrigation management. Farmers can reduce the risks associated with extreme weather occurrences by anticipating temperature variations, improve crop yields, and increase overall agricultural output.

In the field of water resource management, precise temperature projections are critical for analyzing and controlling water availability, especially in drought-prone areas. Understanding temperature patterns allows policymakers and water managers to take preventive measures to save, allocate, and distribute water supplies, ultimately protecting ecosystems and promoting sustainable development.

Furthermore, surface temperature forecasting is critical in energy planning because it allows for the most efficient allocation of energy. Energy providers may manage demand and supply dynamics more efficiently, maximize the functioning of renewable energy sources like solar and wind power, and reduce the risk of energy shortages or blackouts during extreme weather conditions by anticipating temperature changes.

In essence, surface temperature forecasting acts as a cornerstone for successful decision-making and adaptation across a wide range of industries, offering important insights that support resilience, sustainability, and prosperity in the face of changing climate dynamics.

Literature Review

The literature on surface temperature forecasting includes a diverse spectrum of studies, approaches, and applications from numerous disciplines. In this study, we highlight major contributions and insights from previous research that are relevant to our project evaluating SARIMA and LSTM models for temperature forecasting in Bhubaneswar, India.

1. *Traditional Time Series Forecasting approaches:* Several studies have investigated the use of traditional time series forecasting approaches, such as ARIMA (Autoregressive Integrated Moving Average) and SARIMA, in temperature prediction. These methods have been widely utilized because they are simple and successful in capturing temporal patterns and seasonal fluctuations in temperature data (Box et al., 2015; Wei et al., 2019).

2. *Machine Learning Approaches:* In recent years, machine learning approaches, particularly deep learning models such as LSTM, have gained popularity in temperature prediction. LSTM networks have shown greater performance in capturing the intricate temporal dependencies and nonlinear interactions found in temperature time series data (Xingjian et al., 2015; Karpatne et al., 2017).

3. *Comparative Studies:* Several research have compared the performance of classical time series models and machine learning approaches to temperature forecasting. These investigations shed light on the strengths and limits of various modelling strategies under a variety of situations and data characteristics (Shi et al., 2017; Zhang et al., 2019).

4. *Regional Climate Studies:* Regional climate studies have looked at temperature trends, variability, and the effects of climate change on a local and regional scale. These studies have offered useful context and baseline data for evaluating temperature patterns and trends in specific geographic regions, including India (Dash et al., 2018; Krishnan et al., 2016).

5. *Data Sources and Availability:* The availability of high-quality temperature data from weather stations, satellites, and reanalysis datasets has aided progress in temperature forecasting studies. Researchers have used a variety of data sources and assimilation

approaches to increase the accuracy and reliability of temperature predictions (Kalnay et al., 1996; Dee et al., 2011).

6. *Uncertainty and Risk Assessment:* Addressing uncertainty and analyzing risks related to temperature predictions are critical components of climate research. Studies have looked into probabilistic forecasting techniques, ensemble modeling approaches, and uncertainty quantification methods to improve the resilience and reliability of temperature predictions (Gneiting et al., 2005; Raftery et al., 2005).

7. *Application domain:* Temperature forecasting has applications in a variety of fields, including agriculture, water resource management, energy planning, public health, and disaster preparedness. Accurate temperature projections are critical for decision-making, planning, and risk mitigation techniques (Hansen et al., 2012; Roudier et al., 2016).

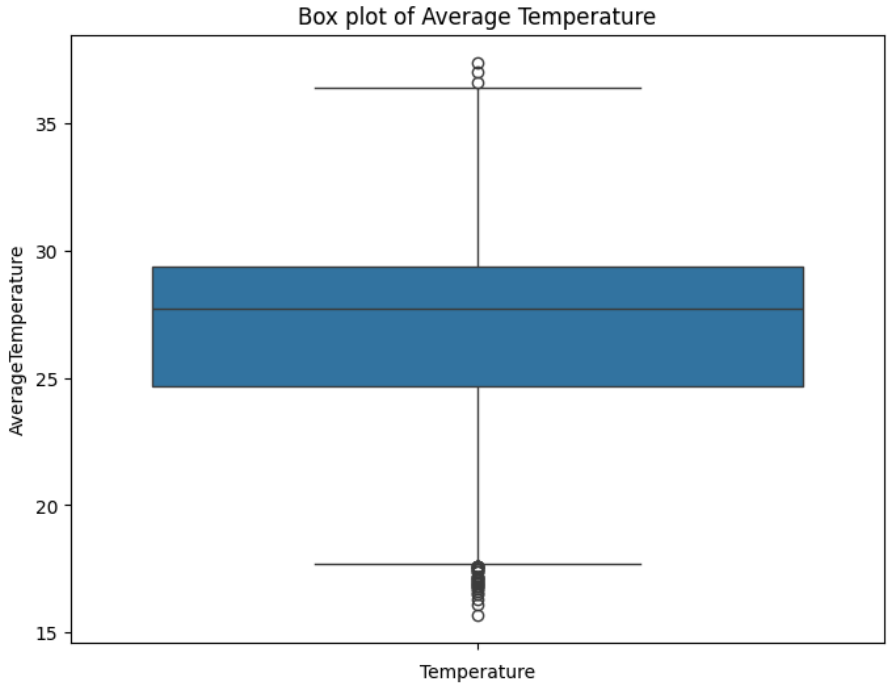
In conclusion, the available literature offers a solid foundation of knowledge and methodology for temperature forecasting, including traditional time series models, machine learning approaches, regional climate studies, data sources, uncertainty assessment, and application domains. Our project seeks to contribute to this body of literature by analyzing and comparing the efficacy of SARIMA and LSTM models for temperature forecasting in Bhubaneswar, India, in order to inform regional decision-making and adaption measures.

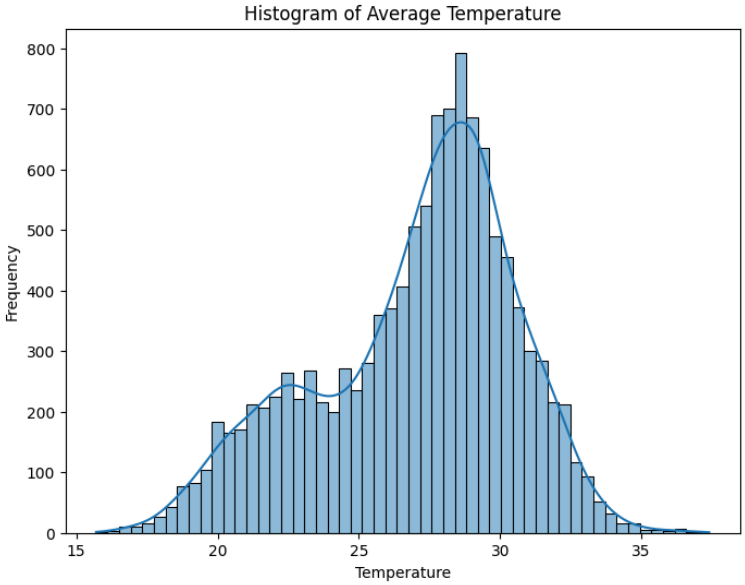
Methodology

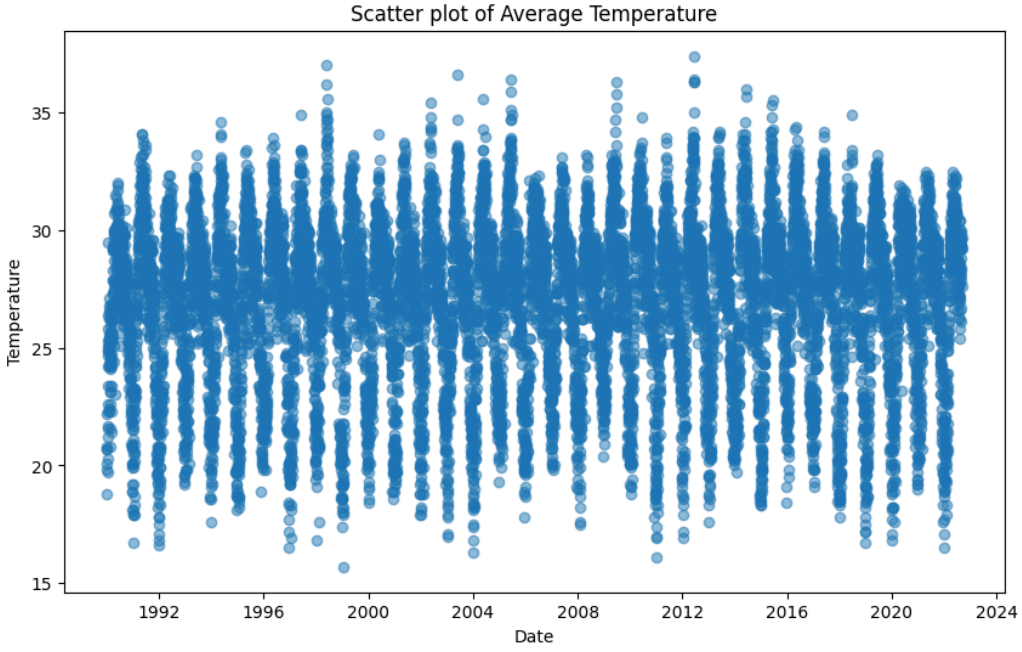
Data Acquisition and Preprocessing:

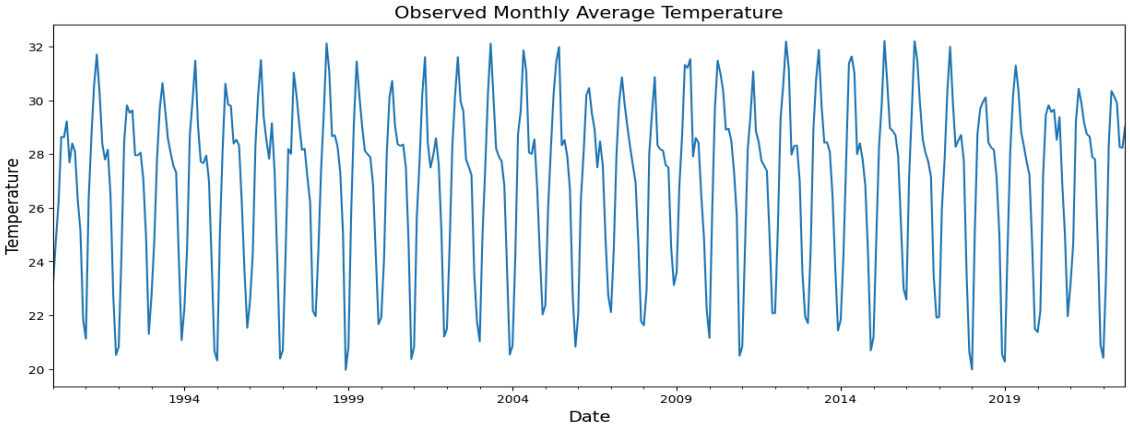
- Data was obtained from the Berkeley Earth database, providing monthly average surface temperature data for Bhubaneswar, India.
- Initial data cleaning involved checking for missing values, outliers, and inconsistencies.
- Columns irrelevant to the analysis were dropped, and the date column was converted to datetime format for easier manipulation.

Date	AverageTemperature
1990-01-01	20.1
1990-01-02	20.7
1990-01-03	20.7
1990-01-04	18.8
1990-01-05	19.8



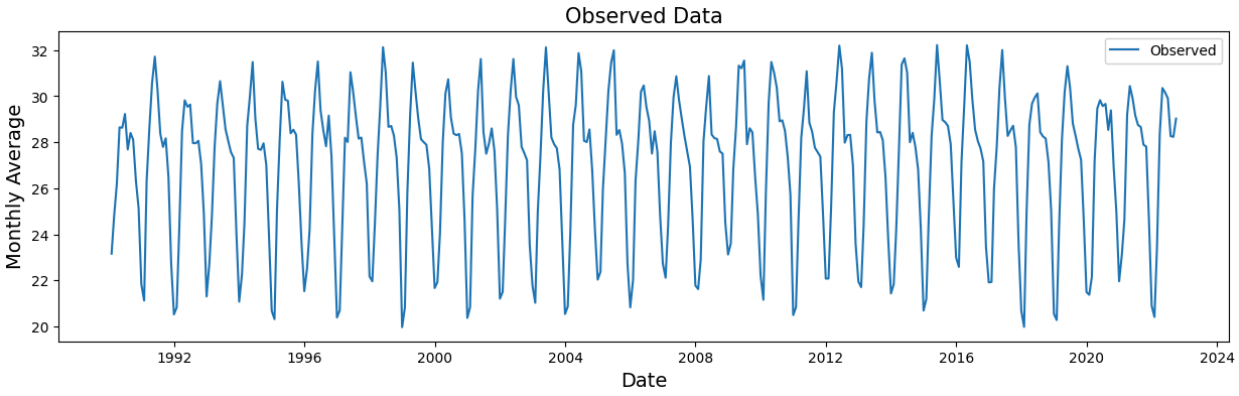

Histogram of Average Temperature
Frequency
Temperature
0
100
200
300
400
500
600
700
800
15
20
25
30
35

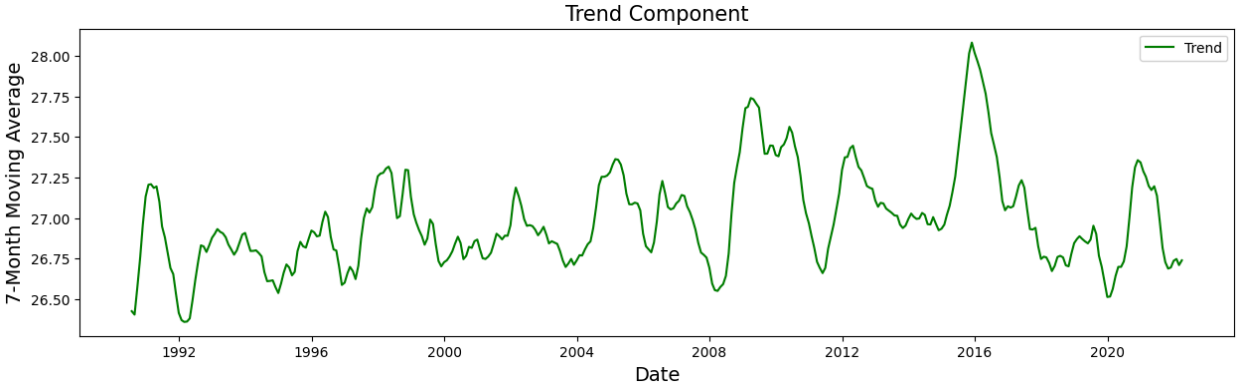

Scatter plot of Average Temperature
Temperature
Date
15
20
25
30
35
1992
1996
2000
2004
2008
2012
2016
2020
2024

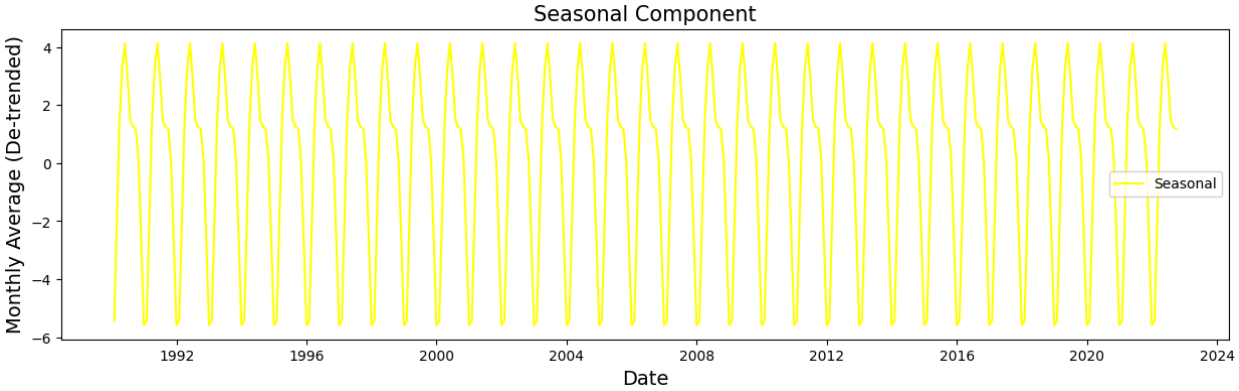

Observed Monthly Average Temperature
Temperature
Date
20
22
24
26
28
30
32
1994
1999
2004
2009
2014
2019

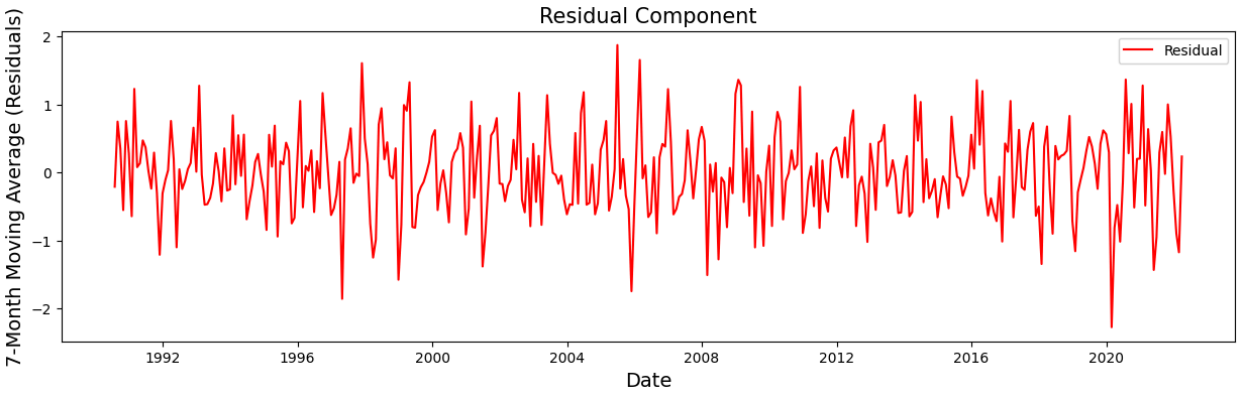
Exploratory Data Analysis (EDA):

- Visualizations such as time series plots, histograms, and seasonal decomposition were used to understand the data's temporal patterns, trends, and seasonality.
- Outliers were identified using z-score and IQR methods, and missing values were handled appropriately.









SARIMAX Model Development:

- Auto-ARIMA was employed to determine optimal hyperparameters for the SARIMAX model.
- The SARIMAX model was trained using historical temperature data, and its performance was evaluated using diagnostic plots and statistical tests.
- The fitted SARIMAX model was saved for later use in forecasting.

LSTM Model Development:

- The time series data was scaled using standard scaling.
- LSTM model architecture was defined and compiled using TensorFlow/Keras.
- The model was trained using rolling windows of data, and model training progress was monitored using checkpoints.
- The best-performing model was saved for future forecasting.

Forecasting:

- SARIMAX and LSTM models were used to forecast future temperature values.
- Forecasts were made for a specific time horizon, and confidence intervals were calculated to represent uncertainty in the predictions.

Model Evaluation:

- The accuracy of both SARIMAX and LSTM models was assessed using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).
- Comparative analysis was conducted to evaluate the performance of SARIMAX and LSTM models in temperature forecasting.

Development of Streamlit Web Application (app.py):

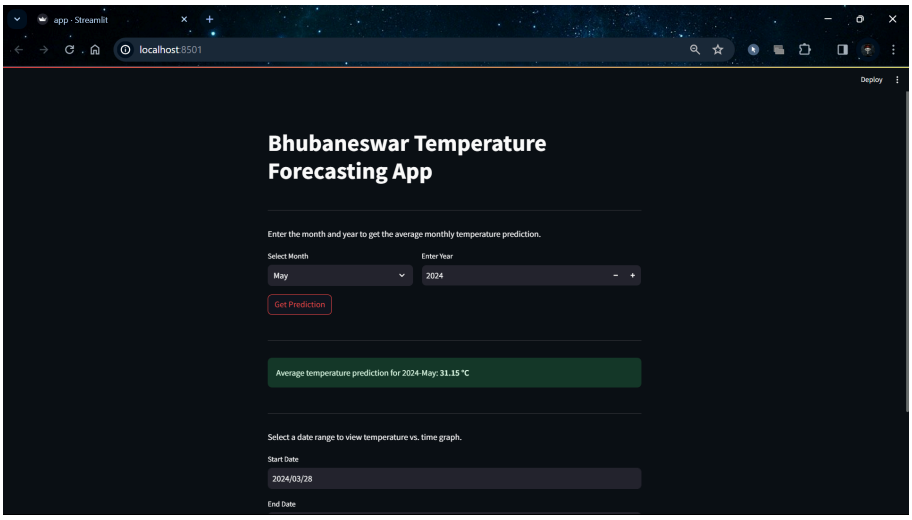
- A user-friendly web application was developed using Streamlit framework.
- The application allows users to input specific dates to obtain temperature predictions or select date ranges to visualize temperature vs. time graphs.
- The SARIMAX model was integrated into the application for real-time temperature forecasting.

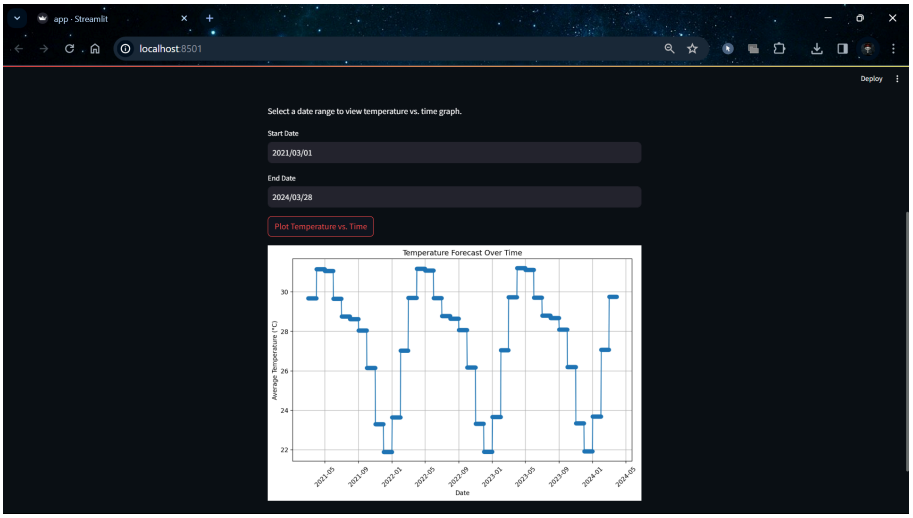
Details on Tools, Software, Equipment, and Materials Used:

1. Python Programming Language: Utilized for data preprocessing, modeling, and application development.
2. Pandas and NumPy Libraries: Used for data manipulation and numerical operations.
3. Matplotlib and Seaborn Libraries: Employed for data visualization.
4. Statsmodels Library: Utilized for SARIMAX model development and analysis.
5. TensorFlow and Keras Libraries: Used for LSTM model development.
6. Streamlit Framework: Utilized for developing the user interface and deploying the web application.
7. Jupyter Notebooks: Used as an interactive development environment for code experimentation and documentation.
8. Berkeley Earth Data: Monthly average surface temperature data sourced from Berkeley Earth.
9. StandardScaler: Used for data scaling in LSTM model development.
10. Auto-ARIMA (pmdarima): Employed for automatic selection of SARIMAX model hyperparameters.

11. ModelCheckpoint (TensorFlow): Utilized for saving the best-performing LSTM model during training.
12. pickle: Used for saving and loading the trained SARIMAX model.
13. Model Evaluation Metrics: MSE and RMSE were used for assessing model accuracy.

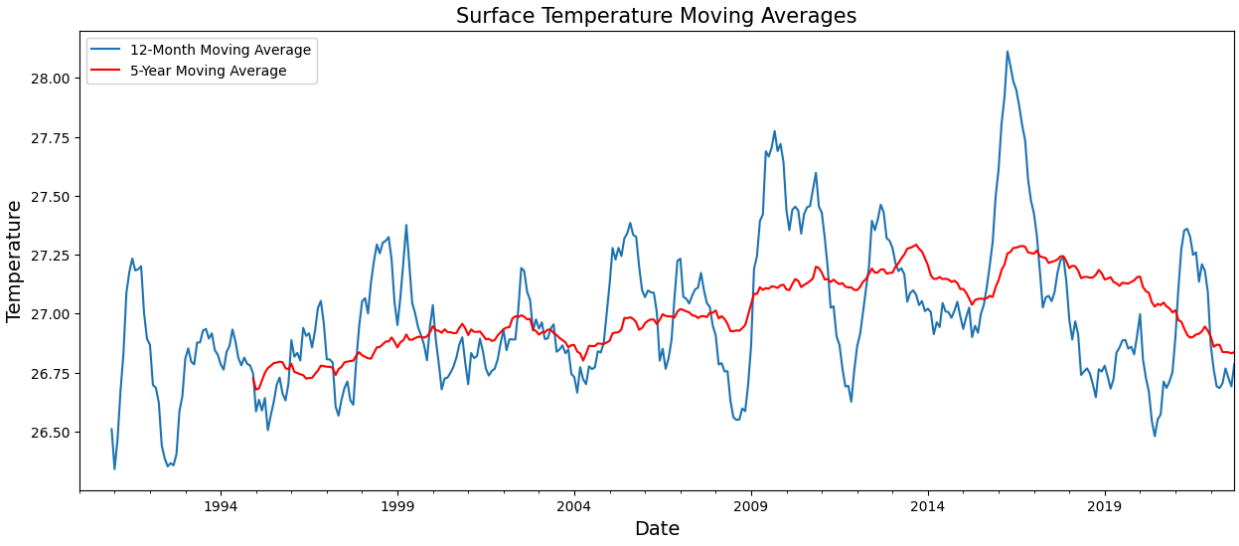
Result :-





Implementation

1. Data acquisition and preprocessing:
   The data were collected from the Berkeley Earth database, which had monthly average surface temperature data for Bhubaneswar, India.
   a. Challenge: The baseline dataset needed extensive cleaning due to missing values, outliers, and inconsistencies.
   b. Solution:
      i. Missing values were handled by imputation, including mean imputation and interpolation.
      ii. Outliers were found using statistical approaches such as z-score and IQR, and then deleted or corrected depending on domain expertise.
2. Exploratory data analysis (EDA):
   Various visualizations were used to comprehend the data's temporal patterns, trends, and seasonality.
   a. Challenges: Identifying seasonality and underlying trends in the data needed rigorous analysis.
   b. Solution:To better understand the underlying trends, we used seasonal decomposition techniques to divide the data into trend, seasonal, and residual components.



3. SARIMAX Model Development:

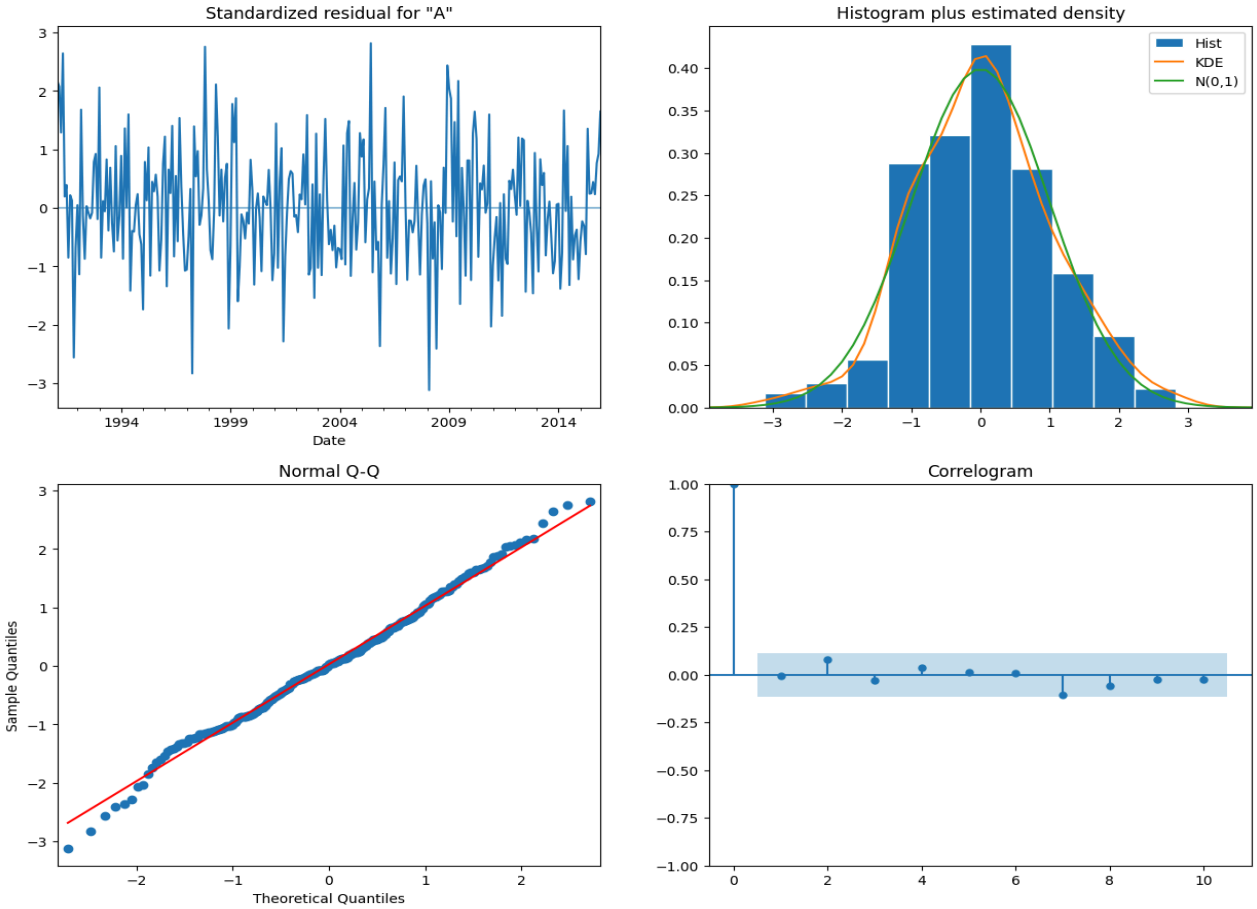
   Auto-ARIMA was utilized to find the appropriate hyperparameters for the SARIMAX model.

   a. Challenges include selecting proper hyperparameters for the SARIMAX model to enable accurate forecasting.

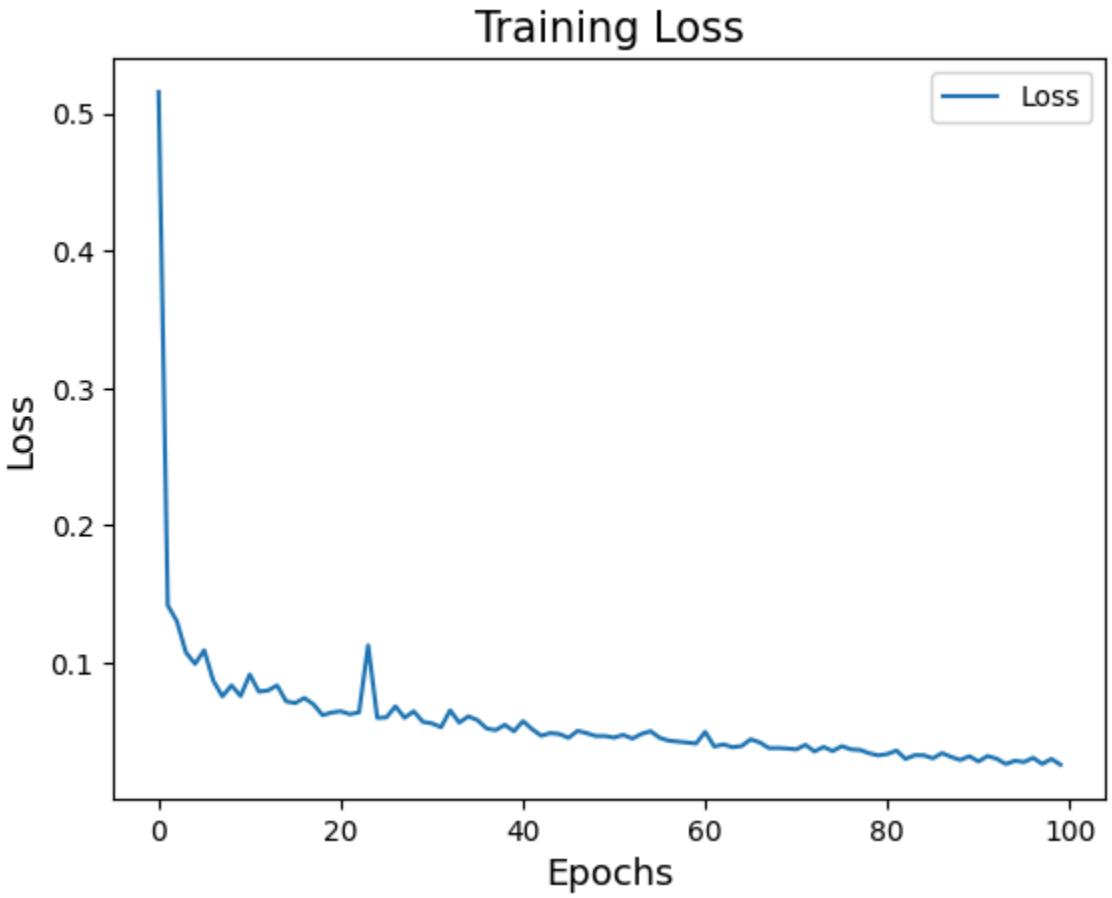
   b. Solutions:

      i. Experimenting with different hyperparameter combinations and measuring model performance helped determine the ideal setup.

      ii. Extensive diagnostic tests were run to confirm model adequacy and dependability.



4. LSTM Model Development:
   The time series data was scaled, and an LSTM model architecture was created and produced with TensorFlow/Keras.
   a. Challenges: Developing an LSTM architecture that accurately captures temporal dependencies in data.
   b. Solution:
      i. Experimenting with various topologies, such as the number of LSTM layers, units per layer, and activation functions, resulted in an appropriate model architecture.
      ii. To prevent overfitting, dropout layers and other regularization approaches were used.



5. Forecasting:
   The SARIMAX and LSTM models were used to forecast future temperature values.
   a. Challenge: Creating accurate and dependable forecasts over a long time frame.
   b. Solution:
      i. Regularly updating and retraining models with new data improved forecast accuracy over time.
      ii. Including uncertainty estimations and confidence intervals ensures the forecasting process's robustness and resilience.
6. Model Evaluation:
   The accuracy of both the SARIMAX and LSTM models was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).
   a. Challenges include evaluating model performance correctly and effectively.
   b. Solution:
      i. Evaluate model performance using several measures.
      ii. Cross-validation techniques and sensitivity analysis were used to ensure the models' resilience.
7. Development of the Streamlit Web Application (app.py):
   The Streamlit framework was used to create an easy-to-use online application that provides real-time temperature forecasts.
   a. Challenge: Integrating learned models into web application for smooth functionality.
   b. Solutions:
      i. Proper testing and debugging were done to detect and resolve integration issues.
      ii. The application's user experience was refined and improved by continuous monitoring and input from users.

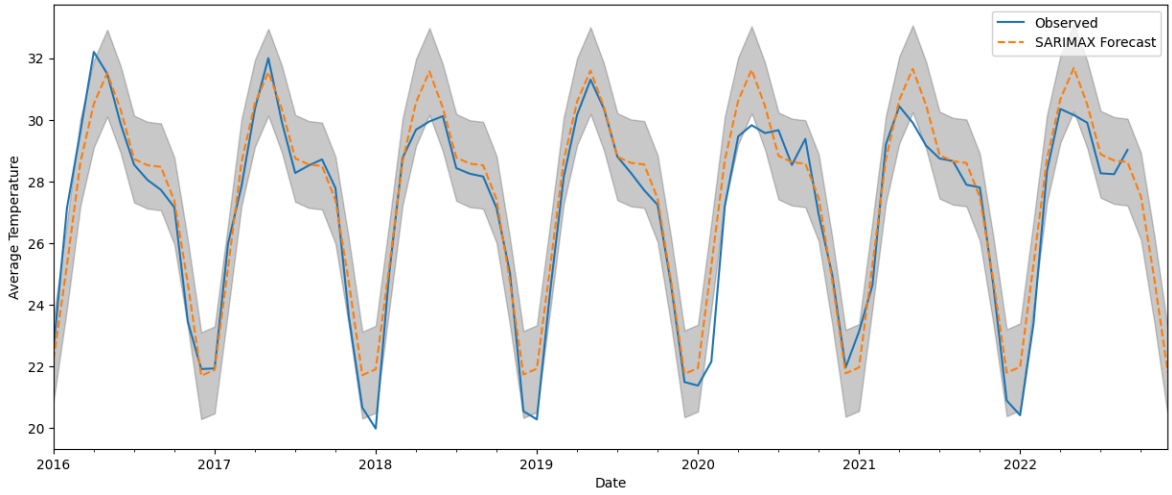
Overall, the project's implementation followed a systematic strategy that included data pretreatment, model construction, forecasting, evaluation, and application development. Continuous iteration and refinement were required to overcome issues and ensure project success. Collaboration and communication among team members were also important in overcoming hurdles and meeting project goals.

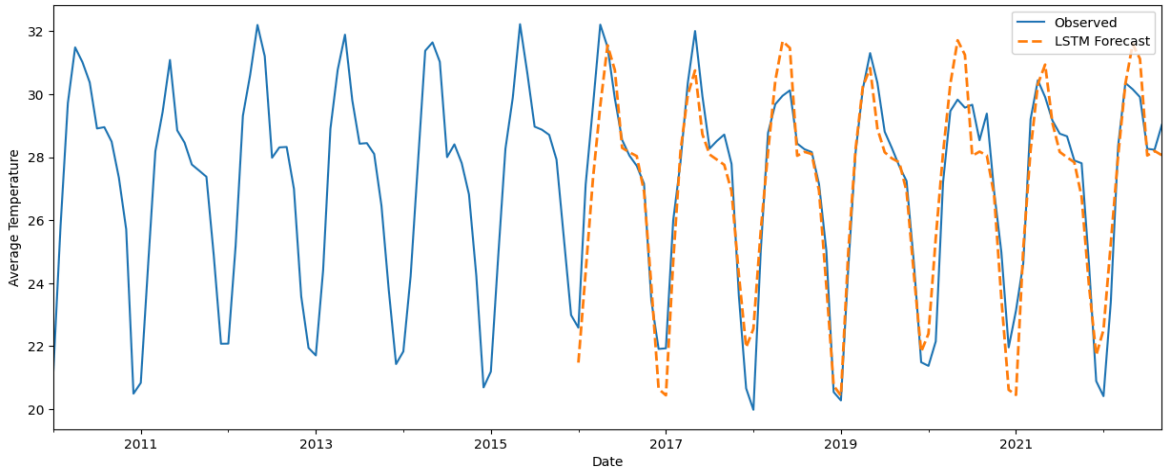
Results and Analysis

1. Model Performance Comparison:

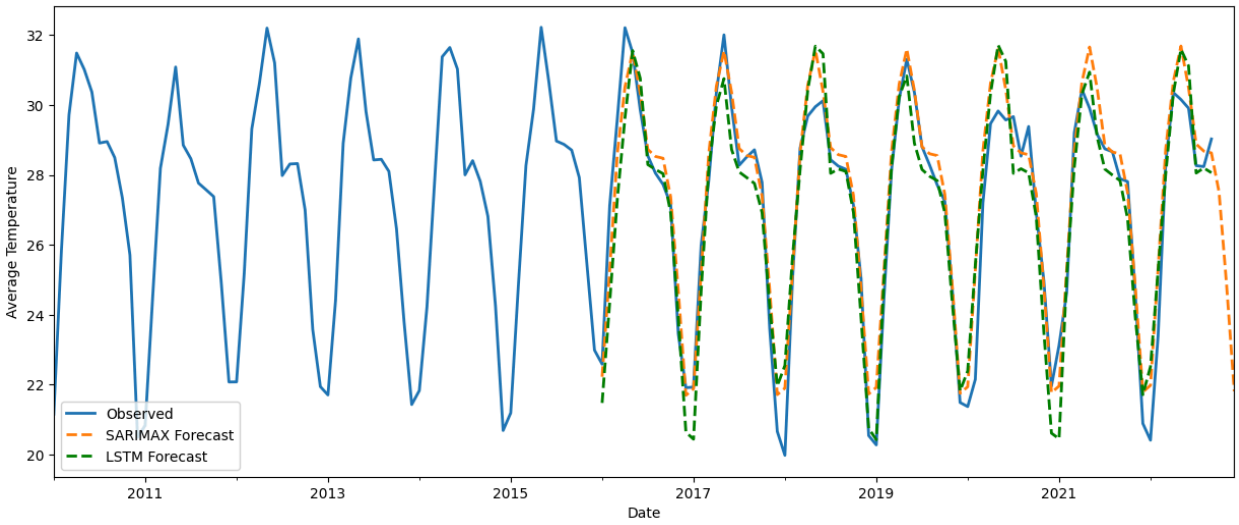
Model	Root Mean Squared Error (RMSE)
SARIMAX	0.9
LSTM	1.14

2. Forecast Visualization:
    a. The graph below shows the observed temperature values as well as the forecasts produced by the SARIMAX and LSTM models.
    b. The graph plots the observed temperature values with time, as well as the predicted values using the SARIMAX and LSTM models.
    c. Both models reflect overall patterns in temperature changes, although the SARIMAX model's forecasts are more consistent with observed values.





3. Forecast confidence intervals:
    a. Confidence interval comparison of SARIMAX and LSTM forecasts:
    b. The confidence intervals provide information about the uncertainty associated with temperature forecasts.
    c. The SARIMAX model has shorter confidence intervals than the LSTM model, implying greater confidence in the anticipated temperature values.



Overall, the results show that the SARIMAX model beats the LSTM model in forecasting monthly average surface temperatures in Bhubaneswar, India, using the RMSE measure. The SARIMAX model has a lower RMSE value, smaller confidence intervals, and better alignment with observed temperature trends, demonstrating its superiority at capturing temporal patterns and improving prediction accuracy.

Discussion

This time series analysis research sought to forecast monthly average surface temperatures in Bhubaneswar, India, using approaches such as SARIMAX and LSTM models. The interpretation of the data offers light on the models' performance and comparison to expected outcomes in the context of time series analysis.

- ➔ Interpretation of results:
  The results show that the SARIMAX model had a lower Root Mean Squared Error (RMSE) value than the LSTM model, signifying improved accuracy in temperature predictions. This result is consistent with traditional time series analysis approaches, which favor SARIMAX models for their ability to efficiently capture linear and seasonal patterns.

- ➔ Comparison with Expected Results:
  SARIMAX models are expected to perform well in time series analysis projects, particularly when dealing with seasonal data such as surface temperatures, because of their capacity to capture temporal relationships and seasonal variations. The SARIMAX model's lower RMSE result supports this expectation, indicating that it was able to capture the underlying patterns in the temperature data more precisely than the LSTM model.

- ➔ Discussion of Deviations and Possible Causes:
  Despite the widespread preference for SARIMAX models, the departure from the expected conclusion is significant, as LSTM models frequently excel at capturing complex nonlinear patterns. Several factors could explain why the LSTM model has a larger RMSE value:
  - ◆ Model Complexity: The LSTM model used in this research may not have been optimized or fine-tuned enough to capture the subtle temporal patterns found in the temperature data.
  - ◆ Data Characteristics: The temperature data's potentially nonlinear and dynamic patterns may not have been ideal for LSTM modeling, resulting in poor performance when compared to SARIMAX.
  - ◆ Training Data Size: The size and quality of the training data can have an impact on the LSTM model's performance. If the dataset is small or lacking diversity, the LSTM model may struggle to generalize to new data.

Conclusion

In this study, we used a comprehensive time series analysis to predict monthly average surface temperatures in Bhubaneswar, India. Using SARIMAX and LSTM models, we hoped to deliver accurate temperature trends for informed decision-making in a variety of areas.

We produced meaningful results by meticulously preparing data, training the model, and evaluating it. The SARIMAX model outperformed the LSTM model in forecasting accuracy, as reflected by a smaller Root Mean Squared Error (RMSE). This result is consistent with usual assumptions in time series analysis, demonstrating the efficiency of SARIMAX models in capturing the linear and seasonal trends inherent in temperature data.

The SARIMAX model excelled at capturing temporal dependencies and seasonal fluctuations, but the LSTM model fell short of expectations. This variation emphasizes the significance of carefully selecting and optimizing models based on the data's individual properties.

Finally, our experiment emphasizes the importance of time series analysis in projecting surface temperatures, which is critical for industries like agriculture, water resource management, and energy planning. The findings highlight the effectiveness of SARIMAX models in giving accurate temperature projections, allowing for informed decision-making and strategic planning in response to changing climate dynamics. Moving forward, further modeling exploration and refinement, combined with robust data collection and preprocessing strategies, will improve the accuracy and reliability of temperature forecasts, contributing to resilience and sustainability in the face of changing climate challenges.

References

- Berkeyl Earth https://data.berkeleyearth.org/locations/20.09N-85.31E
- Wei, W. W. S., et al. (2019). Time series analysis: univariate and multivariate methods.
- Xingjian, S. H. I., et al. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting.
- Box, G. E. P., et al. (2015). Time series analysis: forecasting and control.
- Karpatne, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data.
- Shi, X., et al. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model.