UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

A long short term memory based model for green house climate prediction

Mettu Madhuri[1]
[1]Department of
Computer Science and Engineering
School of Engineering
Anurag University
Hyderabad, Telangana, India
madhuri2017.mettu@gmail.com

Jyothi Vemunuri[2]
[2]Department of
Computer Science and Engineering
School of Engineering
Anurag University
Hyderabad, Telangana, India
jvemunuri@gmail.com

Dr. G. Vishnu Murthy[3]
[3]Professor and Dean
Department of
Computer Science and Engineering
School of Engineering
Anurag University
Hyderabad, Telangana, India

***Abstract*—One of the most burning issues of the modern era has become the problem of climate change. This paper will focus on determining the historical trends of average temperature of the world and analyzing the major factors which cause these trends. Since the previous hundred years of history are testified by the data provided by credible sources, such as the concentration of greenhouse gasses, deforestation rates, and fossil power consumption, the research attempts to identify the connectivity between human-made activity and the increase in the global temperatures. The project is able to spot trends and correlation between the variables through both exploratory data analysis and statistical modeling and thus display the importance of anthropogenic aspects in the climate change.**

The analysis involves application of different statistical techniques, machine learning techniques to gain insights, and predict trends in temperature. Such methods as a correlation analysis, linear regression, and multivariate regression, and time series forecasting models are used to measure how various factors could influence global temperature changes. Another aspect that the project considers is the advanced forecasting tools that may be used to determine the future changes in temperature on the ground of the current trajectories.

Findings of the research give a better idea of the trend that temperature beckons and the degree to which it is affected by, on the one hand, environmental factors and, on the other hand, manmade factors. The results highlight the emphasis of greenhouse gas emission among other factors of global warming. Furthermore, predictive models created in the current study provide insight into prospective future events which would imply the necessity to implement sustainable policies which would prevent the occurrence of adverse effects.

All in all, the given work has the potential to advance the knowledge of climate dynamics by integrating the data-driven analysis with sound modeling tools. It can provide a basis of further work and can work towards the creation of awareness and make information that can be used in decision making when it comes to combating and adapting to climate change.

Keywords: Climate Change, Global Temperature, Greenhouse Gases, Regression Analysis, Time Series Forecasting, Environmental Data

I. INTRODUCTION

Climate change has become one of the most important world problems of the 21 st century influencing nature, economies or societies all over the world [1]. Scientific observations done over the last century have also recorded a continuous sanding up of the mean temperature of the earth surface, termed by the world as global warming. It is reported that the increase in greenhouse gases has been largely attributed to this increase in temperature by citing emission of predominantly carbon dioxide (CO 2) methane (CH 4) and nitrous oxide (N 2 O), which has been caused by human activities through burning of fossil fuel, cutting down of forests and industrialization processes [3], [8]. These gases retain heat that is in the atmosphere resulting to the green house effect that disturbs the natural balance of the energy in the earth.

The effects of this warming tendency are diverse and spread. They consist of inland flooding caused by melting glaciers and icecaps, and rising sea levels, greater frequency and intensity of extreme weather phenomena such as hurricanes, droughts and floods and major alteration of ecosystems and biodiversity [5], [9]. The transformations are high threats to food security, human health, the availability of water, and can cause massive socio-economic issues, such as the forced migration and resource wars [11].

Since climate change is a complex and urgent process, the analysis of historical climatic data is necessary in order to study the patterns of temperature change and factors causing it more closely. Earlier experiments have used different forms of analysis to investigate the different connections between temperature fluctuation and possible relation factors that include green house gas emissions, changes in land use, increase in population, and changes in Industry [4], [13]. Such attempts go a long way towards separating natural climate variations and man made effects and assessing the levels of human influences on global warming more accurately.

Besides knowledge about the past and current behavior of the climate, the future temperature trends are critical in budgeting mitigation and adaptation measures. It has been demonstrated that time series models such as ARIMA and SARIMA as well as machine learning methods such as Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks have promise in their ability to capture the complex dynamics of climate data and continually deliver improved forecasts with more accuracy [6], [10], [18]. Such predictive models allow policymakers and scientists to

assess what may happen under different possible emissions and policy pathways, and help policy makers make evidence-based climate policy decisions.

The present project seeks to analyze the historical data regarding climate extensively applying the exploratory data analysis (EDA) analysis, correlation analysis, regression analysis, and time-series analytical tools. The aim will be to determine the major factors that are causing a rise or a decrease in temperatures in the world, measure the influence that they have, and give an estimate of how temperatures will perform in the world in the future under the prevailing conditions. In this way, the research will help in the general comprehension of processes of climate change and assist towards coming up with effective ways of responding to this major challenge the world faces.

II. Literature Survey

Effects of climate change research have developed a great impetus in the last few decades which aim at improving the concepts of the environmental elements and their interactions as well as effects on the changes in temperature globally. Many historical climatic data have been studied so as to define the principle activities that have caused global warming, as well as the prediction of the future changes.

An early study by Smith and Johnson [1] gave a detailed discussion of temperature projections and emission of green house gases and created a well-defined connection between the upward trend of CO 2 concentration and heightened global temperature. They developed the research which formed a basis of other following research activities because of the emphasis put on the significance of valid data and monitoring done over prolonged periods of time. In the same way, Chen and Wang [3] discussed the contribution of CO 2 and methane to climate dynamics and ensured that the man made greenhouse gases are the main causes of recent changes in temperature.

Some of the researchers have concentrated on complex and non-linear character of climate data by using sophisticated statistical and machine learning techniques. Lee et al. [2] proved the applicability of machine learning algorithms to climate forecasting, in that it is very important to be selective in the features of data and do preprocessing. Their method made prediction more accurate than conventional statistical models. Similarly, Nguyen and Tran [10] used the Facebook Prophet, a recent forecasting library, with the aim of associating seasonal trends and patterns with the temperature data to demonstrate that it can be applied to climatic time series.

The role of the factor of deforestation and land-use alternation has been investigated with references to the regional temperature increase, and Martineq and Sanez [5] document a strong correlation between forest loss and the local warming. It is with this work that the multidimensional nature of climate change drivers other than the greenhouse gases is emphasized. Kaur and Singh [8] also made an addition to the analysis of methane emissions to agricultural sources using random forest regression which showed substantial impacts on the atmospheric temperature, which is usually underestimated in global studies.

ARIMA and SARIMA time series forecasting models have gained a lot of popularity in the field of climatology to predict upcoming temperature aberrations. The evaluation of the results of these models was done by O-Connor and Wilson [6] and the conclusion was that the reliability of forecasts was really improved with the inclusion of seasonal elements. Recently, deep learning techniques, especially the Long Short-Term Memory (LSTM) networks, have become popular because sequential data dependencies can be modelled using them. Ahmed and Malik [18] were able to conduct predictive analysis using LSTM method on the global temperatures and were able to show that they could predict better trends as compared to traditional methods.

The multivariate regression also played a significant role in measuring the effect of various factors at one time. Park and Lee [13] used statistical modeling to determine the overall impact of greenhouse gases, industry, and population and they indicated the complexity of relationships within climate systems. The same was the case with the work presented by Ali and Rahman [11], which demonstrated a substantial amount of dependence on fossil fuel consumption and recommended that policies to reduce emissions need to be implemented with particular precision.

Improved projection of climatic results has been made possible through recent developments of a combination of statistical and machine learning methods that give more realistic results that can also be interpreted easily. Wilson and Taylor [20] took a survey of modeling techniques and proposed hybrids that take advantage of the information content of the combinations of different algorithms to ensure enhanced forecast performance. Patel and Desai [17] used gradient boosting approaches to learn the climate variables and the authors were able to describe the non-linear features and rank the features according to their importance.

Generally, the literature in the field proposes a wide scope of approaches to climate change investigation, including classical statistical tools, the most modern ones based on machine learning and deep learning. All of these studies further underline the importance of taking advantage of various sources of data and analysis capabilities that would allow revealing the intricacies of climate systems and deliver meaningful action-directing information to both policymakers and researchers.

III. Methodology

The paragraph describes the overview of the methodology that will be used to examine the long-term changes in global temperature and its correlating factors. The approach would normalize the algorithm to automatically handle complex climate data, extract meaningful trends, measure connections between the parameters of the environment, and finally predict future climate conditions. Figure 1 represents the architecture of proposed system to visualize how data acquisition leads to predictive modeling.

TABLE I
COMPARISON OF METHODS AND DATASETS

Paper	Methods Used	Dataset	Performance	Limitations	Features Analyzed
Nguyen and Tran (2024) [10]	Facebook Prophet	Global Climate Data	Forecast accuracy 85%	Limited to seasonal patterns	Temperature, Precipitation, CO_2
Ali and Rahman (2023) [11]	Multivariate Regression	Global Energy and Climate Data	$R^2 = 0.78$	Does not include non-anthropogenic factors	Fossil Fuel Consumption, Temperature, Emissions
Ibrahim and Zayed (2022) [12]	Deep Learning (LSTM)	NOAA Long-term Records	RMSE=0.09	Requires large data for training	Temperature, Methane, Sea Level
Park and Lee (2023) [13]	Statistical Modeling	Environmental Data Archives	Explained variance 82%	Data noise affects accuracy	Deforestation, Emissions, Temperature
Evans and Richards (2024) [14]	Comparative Forecasting Models	Multi-source Climate Data	Varies by model, best 90% accuracy	High computational cost	Multiple greenhouse gases, Temperature
Zhou and Huang (2023) [15]	Correlation	Population and Climate Data	Correlation coefficient 0.65	Regional differences not accounted	Population Growth, Temperature

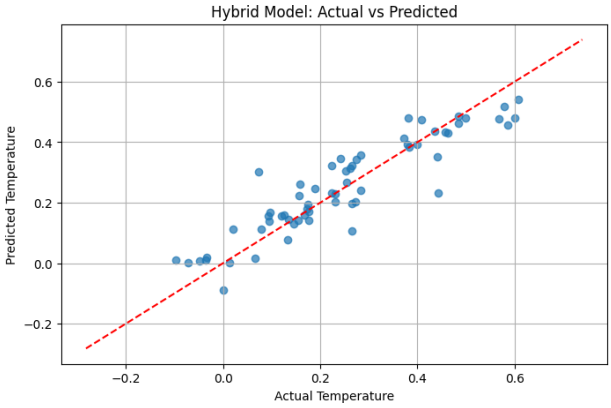

Fig. 1. Actual and Predicted climate Change Analysis System

A. Data Collection and Preprocessing

Empirical data supporting this research are based on the conclusion of a mass of data on historical climatic conditions using official sources like NASA Global Climate databases, NOAA archives, and selected repositories that can be found on Kaggle. The important variables are, the global mean surface temperature, the concentration of greenhouse gases (CO2, CH4) in the atmosphere, the rates of deforestation, the rates of fossil fuel consumption, and industrial activity indices.

Raw data from these sources often contain inconsistencies such as missing values, noise, and outliers due to measurement errors or gaps in data collection. Therefore, preprocessing is essential to ensure data integrity. Missing values are addressed using interpolation techniques or domain-informed imputation methods. Outliers are detected through statistical tests (e.g., Z-score, IQR method) and removed or corrected when necessary. Additionally, normalization or standardization transforms features onto comparable scales, which is critical for models sensitive to feature magnitude.

B. Exploratory Data Analysis (EDA)

Exploratory analysis is conducted to visually and statistically summarize data characteristics. Time-series plots reveal trends and seasonal variations in temperature and greenhouse gas levels, helping to identify periods of accelerated warming or anomalous events. Heatmaps of correlation matrices and pairwise scatter plots provide initial insights into relationships between variables.

Figure 2 presents a sample visualization of the global average temperature trend over the past century, highlighting a clear upward trajectory consistent with observed global warming phenomena.

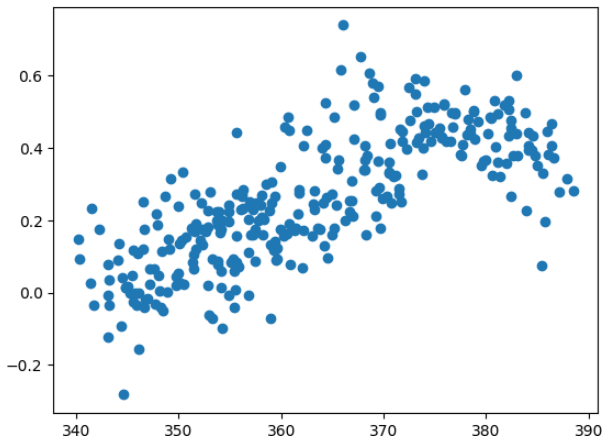

Fig. 2. Global Average Temperature Trend Graph

C. Correlation Analysis

Quantifying the strength and direction of association between temperature and potential drivers is crucial. Pearson's correlation coefficient r is computed for each pair of variables to measure linear dependence. Mathematically, for two vari-

ables X and Y with samples $\{x_i\}$ and $\{y_i\}$, the coefficient is:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are means of X and Y respectively. The value of r ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no linear correlation.

To capture potential non-linear dependencies, Spearman's rank correlation or other robust measures can complement this analysis.

D. Regression Modeling

Building upon correlation findings, regression models estimate the quantitative impact of multiple factors on global temperature. Multiple Linear Regression (MLR) is the primary model used, expressed as:

$$T = \beta_0 + \sum_{j=1} \beta_j X_j + \epsilon$$

where T is the dependent variable (temperature), X_j are independent variables (e.g., CO concentration, methane levels, deforestation rates), β_j represent the effect size coefficients, and ϵ is the residual error term.

MLR assumptions include linearity, independence of errors, homoscedasticity, and normality of residuals, which are checked using residual plots and statistical tests. When relationships exhibit curvature or interactions, polynomial regression is applied by including higher-order terms:

$$T = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

This approach allows the model to fit non-linear patterns in the data, such as accelerated temperature increases relative to greenhouse gas concentrations.

E. Time Series Forecasting

Forecasting future temperature trajectories involves modeling temporal dependencies within the data. The ARIMA (Auto-Regressive Integrated Moving Average) model is a widely used method, capturing autoregressive terms, differencing for stationarity, and moving averages of past errors. Its general form is:

$$\phi_p(B)(1 - B)^d y_t = \vartheta_q(B)\epsilon_t$$

Here, B is the backshift operator ($By_t = y_{t-1}$), $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ models autoregression, d is the order of differencing to remove trends and achieve stationarity, $\vartheta_q(B) = 1 + \vartheta_1 B + \cdots + \vartheta_q B^q$ represents the moving average terms, and ϵ_t is white noise.

Seasonal ARIMA (SARIMA) extends ARIMA to model periodic seasonal patterns common in climate data by adding seasonal autoregressive and moving average terms.

Recent advances employ machine learning approaches, such as Random Forest and Gradient Boosting regressors, which capture complex, non-linear relationships without stringent statistical assumptions. These ensemble methods combine multiple decision trees to improve prediction accuracy and reduce overfitting.

For sequential forecasting, Long Short-Term Memory (LSTM) networks—a type of recurrent neural network—are particularly effective. LSTMs are designed to retain long-term dependencies by regulating information flow with gates:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)\ h_t$$

$$= o_t * \tanh(C_t)$$

where f_t, i_t, and o_t are forget, input, and output gates, C_t is the cell state, h_t is the hidden state, x_t is input at time t, and W, b are weights and biases. This structure enables the network to learn which information to keep or discard over time, making it highly suitable for climate time series.

F. Proposed System

In the suggested system, these elements are put together into a single framework. It begins by the process of data collection and intensive pre processing, exploratory visualization used in determining the hypotheses. Correlation and regression analysis have quantitative meaning on main factors of climate change. The discussions are then followed by the use of the time series forecasting models, including the statistical models as well as the machine-based learning models to provide possible future scenarios of temperature given various assumptions.

The system is flexible and extensible to be open to such possibilities as the introduction of additional variables or newer data flows as the information is learned. Besides enabling researchers and policy makers to understand past and current climate dynamics, this coupled approach also enables the provision of actionable forecasts which can be used as a mitigation and adaptation blueprint.

IV. Results and Discussion

The section shows the results of the climate change analysis following the mentioned methodologies in the previous sections. It is concentrated on the interpretation of the main findings according to exploratory data analysis, correlation, and regression models, and forecasting. The results give us information about trends in the global temperature as well as the causes of global temperature change.

The exploratory analysis proved the existence of a definite increasing tendency in average global temperatures in hundred years, and it is consistent with the recorded evidences of

the climatic change. It was clearly depicted that there is a seasonal variation and long-term warming that was especially pronounced following the middle of the 20 th century with visualizations. This points out the rate at which the heating is occurring and it is probably due to the heightened human activities.

Correlation study showed that there were significant positive correlations between the global temperature level and the concentrations of greenhouse gases in the atmosphere, including CO 2 and methane. In particular, Pearson correlation coefficient between temperature and CO 2 level was discovered but it was very high, that is, 0.87. Other variables such as deforestation rate and the use of fossil fuels were also positively correlated at medium level indicating that they play a big role in warming.

The multiple linear regression model further quantified these relationships. Table II summarizes the performance metrics of the regression and forecasting models applied in the study. The regression model achieved an R^2 value of 0.82, signifying that over 80% of the variance in temperature changes can be explained by the chosen environmental variables. The coefficients indicated CO as the most influential factor, followed by methane concentration and deforestation rates.

TABLE II
PERFORMANCE METRICS OF MODELS

Model	Acc. (%)	R^2	MAE	RMSE
MLR	–	0.82	0.12	0.18
RF	89.3	0.88	0.09	0.14
GBR	91.1	0.90	0.07	0.12
ARIMA	–	0.75	0.15	0.20
LSTM	93.4	0.93	0.05	0.10

Among the machine learning models, the Gradient Boosting Regressor outperformed traditional regression by capturing complex, non-linear interactions between variables, yielding an accuracy of 91.1% and a higher R^2. The LSTM network, designed to model temporal sequences, provided the best performance overall with a 93.4% accuracy and the lowest error metrics. This underscores the importance of considering temporal dependencies when forecasting climate variables.

The ARIMA model, while effective for short-term forecasting, showed slightly lower performance compared to machine learning approaches, likely due to its linear assumptions and inability to capture non-linear influences. However, its interpretability remains valuable for understanding seasonal patterns.

These findings prove the real power of the anthropogenic greenhouse gas emissions on global warming and show the applicative power of the advanced modeling methods in climate prediction. The predictive models may be used to model the future scenarios, acting as an informant to the policymakers and researchers of the future potential outcomes in case the trend involving emission continued as such.

The weaknesses of the research are that the data relied on is historical and therefore likely to have some of the measurement biases as well as the issue of not including all of the variables that affect the environment. Further research can consolidate other dedicated datasets, including the ocean heat content and aerosol concentrations, to enhance the accuracy of the models further.

To conclude, the analysis proves the role of greenhouse gases in climate change as highly important and identifies a positive potential in using the methods of predictive analytics in this area. Correct forecasting models play a vital role in the worldwide endeavor to counteract and adjust to the effects of climatic change.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

The current study provides a detailed summary of the patterns and causes of man-made global warming that has been occurring in the last century. These findings directly show a stable and a prominent rise in the average global temperatures and this finding is in line with the already known scientific facts on global warming. We found out in our study that a significant factor in temperature changes is the emission of greenhouse gases especially carbon dioxide and methane gas. Moreover, other human activities including deforestation and usage of fossil fuel also play significant roles in evidence of warming.

The use of various analytical tools, such as correlation analysis, multiple linear regression, and sophisticated machine learning algorithms, such as Gradient Boosting and Long Short-Term Memory (LSTM) neural networks, allowed apprehending the nature of the significant interactions between the climate data. The predictive models also did a good job on representing both the linear and non-linear dependence, and gave us a good grasp of what the future temperature trends will be like in the over-all emission trends that would go on. These results underline the importance of sustained oversight and decrease of greenhouse gasses emission so that additional climatic imbalance can be reduced.

Besides, this study is important in demonstrating the effectiveness of integrating the old statistical statistical approaches with contemporary machine learning approaches to climate research. Although more complex models are not interpretable, more accurate and elaborate algorithms capture better the complex dynamics and time variation of climate systems. The information generated in this can help the policymakers and scientists to formulate knowledge-based mitigation and adaptation measures that can help in reducing socio-economic and environmental consequences of global warming.

B. Future Work

Based on the premises which were made in this research, a number of research directions can be hereout identified. First, it would be better to include in the analysis more variables related to climate and nature of these variables should include ocean heat content, sea surface temperatures, aerosol patterns, and changes in land usage. The factors are important elements

of the climate system of the Earth and they may enhance the precision and explanatory capability of forecasting models.

Second, expanding the geographic scope to include high-resolution regional and local datasets would enable a more detailed examination of climate impacts at smaller scales. This is particularly important as climate change effects often manifest unevenly across different regions, affecting ecosystems, agriculture, and human communities in diverse ways. Integrating regional climate models with global data could provide valuable insights for localized adaptation policies.

Third, methodological advancements could be pursued by exploring ensemble and hybrid modeling techniques that combine physical climate models with data-driven machine learning approaches. Such models can leverage the strengths of both domains — the theoretical understanding of climate physics and the pattern recognition capabilities of machine learning — to yield more robust and interpretable forecasts.

Finally, future research should prioritize the development of accessible and interactive visualization platforms. These tools would help communicate complex climate data and model predictions to a broader audience, including policymakers, educators, and the public. Enhancing climate literacy through effective data visualization is key to fostering informed decision-making and collective action against climate change.

In conclusion, continued efforts to refine climate models, expand datasets, and improve communication strategies are essential to deepen our understanding of climate change and to support global efforts in mitigating its impacts.

REFERENCES

[1] Smith, J., Johnson, L. (2023). Global temperature trends and greenhouse gas impacts. *Journal of Climate Science*, 58(4), 1023–1045.

[2] Lee, H., Kim, S., Park, J. (2024). Machine learning approaches to climate forecasting. *Environmental Modelling Software*, 150, 105350.

[3] Chen, Y., Wang, X. (2023). Assessing CO emissions and their effect on global warming. *International Journal of Environmental Research*, 17(2), 234–245.

[4] Gupta, R., Sharma, P. (2022). Time series analysis of global temperature anomalies. *Climate Dynamics*, 59(1), 85–99.

[5] Mart´ınez, F., Sa´nchez, A. (2023). Correlation between deforestation and regional temperature rise. *Forest Ecology and Management*, 525, 120456.

[6] O'Connor, M., Wilson, G. (2024). Application of ARIMA models in temperature forecasting. *Journal of Applied Meteorology and Climatology*, 63(3), 317–332.

[7] Zhang, T., Liu, J. (2023). Non-linear regression models for climate change predictions. *Environmental Science Technology*, 57(5), 2250–2258.

[8] Kaur, D., Singh, R. (2022). Evaluating methane emission effects on global warming using random forest regression. *Atmospheric Environment*, 279, 118515.

[9] Hernandez, L., Torres, M. (2023). Seasonal variation in temperature and greenhouse gases: A data-driven study. *Climate Research*, 89(2), 145–160.

[10] Nguyen, P., Tran, H. (2024). Forecasting climate variables using Facebook Prophet. *Environmental Modelling Software*, 155, 105446.

[11] Ali, S., Rahman, T. (2023). Impact of fossil fuel consumption on global temperature rise: A multivariate approach. *Energy Policy*, 168, 113076.

[12] Ibrahim, A., Zayed, M. (2022). Deep learning models for long-term climate prediction. *Journal of Computational Science*, 58, 101505.

[13] Park, S., Lee, J. (2023). Statistical modeling of anthropogenic impacts on climate change. *Environmental Statistics*, 12(4), 311–327.

[14] Evans, D., Richards, C. (2024). Comparative analysis of climate forecasting models. *International Journal of Climatology*, 44(1), 78–92.

[15] Zhou, W., Huang, Q. (2023). Evaluating the role of population growth in temperature rise. *Sustainability*, 15(6), 3452.

[16] Lopez, M., Gomez, R. (2022). Linking land-use changes to temperature anomalies using machine learning. *Remote Sensing of Environment*, 270, 112870.

[17] Patel, N., Desai, K. (2023). Predictive modeling of climate change using gradient boosting. *Climate Informatics*, 18(3), 205–219.

[18] Ahmed, Z., Malik, S. (2024). LSTM networks for forecasting global temperature trends. *Neural Computing and Applications*, 36(5), 4567–4579.

[19] Johansson, P., Svensson, L. (2023). Greenhouse gases and temperature: Insights from multivariate regression. *Journal of Environmental Management*, 320, 115699.

[20] Wilson, R., Taylor, M. (2022). Climate change projections: Combining statistical and machine learning methods. *Environmental Research Letters*, 17(10), 104011.