Predictive Analysis of Global Warming

Sufian Ahmad Reg no 450554

Hassan Saud Reg no 450006

Ommair Ahmed Reg no 450437

Abstract: *Rising in the temperature level is a great concern for humankind because it is not only affecting the biological lives but also the environment. This project implemented different machine learning techniques on global warming datasets and highlighted the key insights of this concern. The main objective of this project is to evaluate the different machine learning models and choose the best one model to predict the rising temperature.*

Introduction:

Global Warming can be defined as the 'average rise in earth temperature and the effects of this rise on its surface and the atmosphere' [1]. This warming trend is not new, though its pace has significantly increased in the recent years mainly due the burning of the fossil fuels [2]. Natural phenomena and several cosmic fluctuations have caused the earth's temperature to rise over a period of time but the current concerns are mainly caused by the human activities [3] such as the CO_2 emission and the uncontrolled release of greenhouse gases. Rise in sea levels, floods, droughts, uncontrolled spread of diseases are some of the effects of the climate change that is linked with the global warming [4].

Machine Learning techniques can be used to analyze and predict the global warming trends that can help in timely management of the effects, these include

- **Climate Modeling:**

  By analyzing the climate data, it can be used to predict temperature changes, climate patterns and sea-level rising.

  hurricanes, heatwaves, and heavy rainfall can be predicted using weather data which helps the governments to prepare in advance in these situations.

- **Renewable Energy Optimization:**

  It can be used to optimize the energy production from wind power and solar panels by predicting the pattern of wind and sunlight.

- **Ecosystem Monitoring:**

  Satellite imagery and sensor data can be used to monitor and predict the ecosystem changes.

Area of interest:

The major goal is to implement the different machine learning techniques on global warming datasets for climate modelling to predict the rise in temperature change and evaluate different machine learning models.

Problem Statement:

The planet is facing an enormous and serious problem from global warming and the rising temperature level is directly affecting the atmosphere in the form of forest fires, sea level rising, unprecedent weather changes, droughts and also economies, biological lives, human lifestyle and habits. One of key challenges that we are dealing with is the melting of glacier which are responsible for 69% of water storage for the earth and they are melting at the alarming rate [5]. Therefore, if they completely melt down, our future generation will face water shortage. This is one example of many challenges that we will face in future if temperature level keeps rising. Therefore, proposed paper will predict the rising temperature globally by using machine learning techniques.

Potential Applications:

- It will help to see the change in temperature over the years and make predictions for upcoming years.
- The prediction model will help the governments to take measures by knowing the rate of increase or decrease of temperature.
- Better policies and decision can be made by knowing the fact that temperature is rising or decreasing. For example, material used in construction of building can be made more robust to face the temperature randomness.
- Overall, it will create the awareness among people to use less carbon emission products or take any other precautions for safety of environment.
- The animal breeders are dependent on the condition of weather. Therefore, it will give them clue for perfect condition for breeding.
- Plants and Seasonal vegetables can be planted by considering the facts that in which region temperature is going to up or down in coming years.

Literature Review:

- One of the direct effects of global warming is the unprecedented rate of precipitation and the in order to predict the rate of precipitation and the temperatures in Pakistan Multi Model Ensembles technique is used [6] in which different models are evaluated and the best ensemble is used to make the prediction model. In the same study optimum number of GCMs are found using the same MME techniques.
- Time series data from the year 1961 to 2020 is used to make a prediction model for predicting the Global Change in Temperature [7]. Performance of Extra Trees and Light Gradient Boosting is also explored along with other algorithms including Bayesian and Random Boosting and KNN and Random Forest. Authors used different evaluation criteria including MAE (C), MSE (C), MAPE(C), RMSLE (C), R2 and RMSE (C).
- Useability of different ML algorithms is explored by the authors [8]. Different ML algorithms are explored for their useability in predicting the Extreme Events like unprecedented rains and droughts and extreme conditions like heatwaves and floods. Two different techniques of Model Ensembling are explored i.e bagging and boosting and different ML algorithms are used in these ensembling techniques.
- Effect of climate change with Pakistan as a case study is discussed [9]. Support Vector Machines based ML model is designed to analyze the previous 29 years climate data based on ocean atmospheric variables. And the model gives predictions for the next 5 years. Authors conducted the study to evaluate the heat wave days in Pakistan in Summer. They used four different

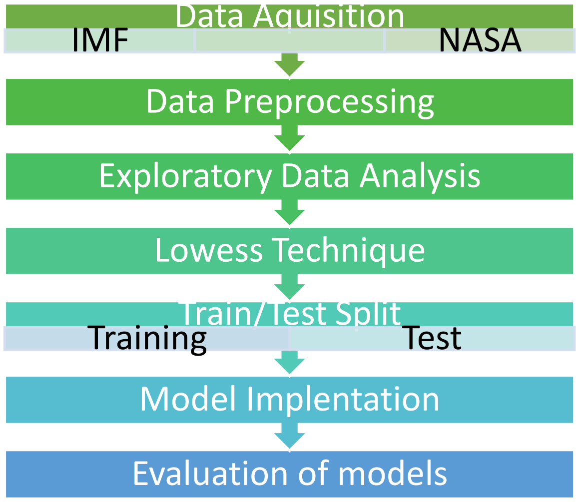
evaluation metrics including R^2, Normalized-RMSE, Percentage of Bias (PBias) and Modified Index of Agreement (MD).

- The potential and applications of ML methods in the Climate Change Risk Assessment are explored [10]. The timeframe for the evaluation is selected to be the year 2000 to the year 2020. The authors also noted the lack of ML applications in the assessment of compound hazards. Different ML algorithms are evaluated and their applications in climate change predictions are presented.
- Droughts are the result of the extreme weather conditions and a committee-based drought prediction model is developed and evaluated with Pakistan as a case study [11]. Different Evaluation metrics are used Root Mean Square Error, Mean Absolute Error, Correlation Coefficient, Willmott's Index.
- Wildfires are a result of extreme weather conditions [12]. Historic data of the last few decades is used to predict the forest fires using the traditional ML techniques ensembled with modern deep learning methods.
- Carbon Dioxide is the major contributing factor in the rise in global temperatures [13]. Historical data is used to make the predictions on which year will be the most damaging due to CO_2 emissions. The year predicted by the authors is 2047.
- The failure in the global yield of the crops is discussed [14] and the Random Forest ML algorithm is used in the prediction.

Proposed Methodology:

The methodology used in this project aims to predict the temperature change in the coming years by using the different machine learning models.

The complete life cycle of the project is given below



The details of each step are given below

Data Acquisition

Different datasets are used to address the global warming, and Worlds Leading Organizations (IMF, NASA) gathered the data from all over the world and it is open source. Two datasets of NASA and one dataset of IMF are used. These datasets have readings of temperature change over the years. The reason to choose three dataset is to perform the EDA analysis on each dataset to extract prominent information from it.

- Dataset 1 contains the temperature change record from 1960 to 2022 for every country. Therefore, this dataset is used to visualize the temperature change in every country[15].

- Dataset 2 contains the temperature change records from 1880 to oct 2023 for every single month. This dataset helps to visualize the temperature change over the months [15]
- Dataset 3 contains the temperature change record from 1880 to 2023 for every year. This dataset is used for predictive analysis and it is normalized by using the statistics technique of Lowess [16].

Data- Preprocessing

In this step, all the datasets go through pre-processing in which data cleaning, data preparation, data normalization and other steps are performed on data to make it clean and ready for EDA analysis.

Exploratory Data Analysis:

In this step, different statistical techniques were used to calculate the key information from data. These techniques included mean, variance, moving average and correlation. Moreover, in this step bar charts, box-plot, line graph and histogram were used to visualize the information from the data.

Lowess Technique:

The dataset was not normalized therefore lowess technique was used to normalize data. In this technique linear regression model is used on specific window of data to make it smooth which helps the machine learning models to learn better and make prediction accurate. The whole process is done by selecting the size of window which it was selected 0.1% and it find the middle point between the window and fit the linear regression curve and make prediction. Based on those prediction new curve is draw which is smooth and easy to interrupt.

Train Test Split:

After the EDA analysis, dataset 3 was used for machine learning modelling. It consists of single column of recorded temperature from 1880 to 2023 and it was transformed into supervised dataset with the lag of 10 values which is common task in time series prediction. After that, datasets were split into two parts train and train split in which last 10 years of data was used for testing and remaining data was reserved for the training. Important note here is that we used the lag of 10 values throughout the work.

Model Implementation

Different models of machine learning and statistics were tried for time series prediction. The list is as fellows

- **Linear Regression**

  Linear regression is widely used in time series analysis specially in stock market prediction.

- **ARIMA**

  ARIMA is statistical model which is integration of the Auto-regressor and moving average and it is extensively used in time series analysis. It has three hyperparameters p,d,q where p is know for auto regressor d for difference and q for moving average. These hyperparameters were set by choosing the different combination of them while taking account the Mean square error.

- **Decision Tree & Random Forest & XGBoost**

  These are common machine learning models that used for regression problem.

- **LSTM**

  It is type of RNN which is used for sequential data input. It has three gates forget, input, output which perform operations to consider the context of the previous data that is why it is widely used in time series analysis. In this model, 10 neurons in input layer with the 20 neurons of LSTM in hidden layer and 1 neuron in output layer were used. The layers and number of neurons decided after experimentation and the default activation functions of LSTM was used.

- **CNN**

  Basically, it is widely used in Image classification but it can be used to analyze the time series data. It works by using the time series data as image and utilize its architect according to data. The architect of CNN is consisted of single input layer of 10 neuron follow by 1D convolution layer of 16 neurons with the kernel size of 3. After that we used the 1D max pooling layer followed by the 1D convolution layer of 32 neurons with kernel size of 3. At last, we applied the global max pooling and dense layer with single neuron. The number of neurons and layers are hyperparameters which was selected after some experimentation to find optimal values. Relu activation function was used due to its general use in regression problems.

Evaluation of models

Mean Squared Error was used as the loss function and the model's performance was evaluated on R2 Score. At last, best models were choose to forecasts the temperature change for next 10 years

Findings and Results:

Three datasets were used to analyze the global warming over the years and each dataset has its own unique features and datapoints. Therefore, results and findings of each dataset is given below

Dataset 1

The dataset contains the global temperature change from 1960 till 2022 for every country in the world.

	Country	ISO3	1961	1962	1963	1964	1965	196(
0	Afghanistan, Islamic Rep. of	AFG	-0.113	-0.164	0.847	-0.764	-0.244	0.22(
1	Albania	ALB	0.627	0.326	0.075	-0.166	-0.388	0.55(
2	Algeria	DZA	0.164	0.114	0.077	0.250	-0.100	0.43(
3	American Samoa	ASM	0.079	-0.042	0.169	-0.140	-0.562	0.18(
4	Andorra, Principality of	AND	0.736	0.112	-0.752	0.308	-0.490	0.41(

Fig 2.1: Dataset 1 overview

Findings from the dataset 1 is given below

- Temperature is significantly changed over the years in every country and most of temperature values are in range of 2 degree to -1.5 degree. The mean of temperature is also changing over years which indicated temperature change is not constant and it keep changing.
- If the means temperature plots with respect to years, it is observed that mean fluctuates over the years as shown in fig 2.2.
- From the dataset it is noticed that Mongolia is the country where highest temperature change is observed followed by the Mauritania, Austria, Western Sahara and Finland. from the year of 1960 to 2022.
- From the year 2000 to 2022 Estonia is the country where highest temperature change is observed followed by the Belarus and Russia.
- Overall, temperature change is observed in all countries from 1960 to 2022 as shown in fig 2.3. From 2000 to 2022, higher temperature difference is observe as compared to past years of 1960 to 1999.

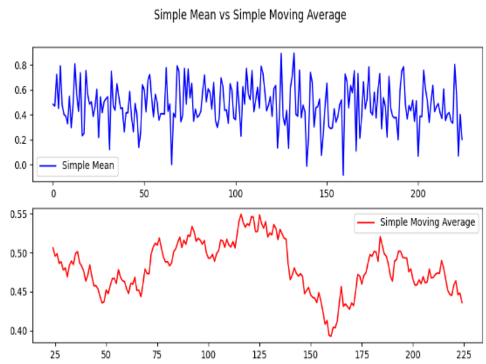


Fig 2.2: Simple mean vs Simple Moving Average

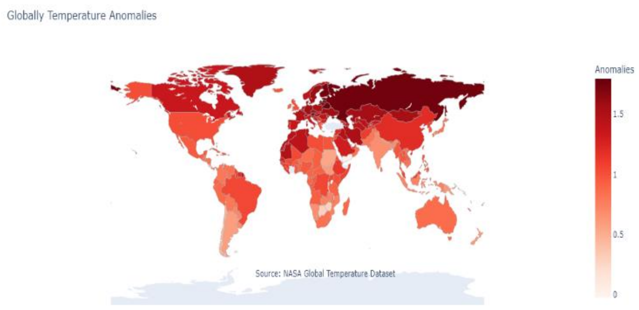


Fig 2.3: Global Temperature Anomalies

For Dataset 2

Dataset 2 contains the temperature anomalies from 1880 till October 2023 for every single month. This dataset helps to analyze the temperature change in months with respect to every year.

	Year	Jan	Feb	Mar	Apr	May
0	1880	-0.19	-0.25	-0.09	-0.16	-0.10
1	1881	-0.20	-0.14	0.03	0.05	0.06
2	1882	0.16	0.14	0.04	-0.17	-0.15
3	1883	-0.29	-0.36	-0.13	-0.18	-0.18
4	1884	-0.13	-0.09	-0.37	-0.40	-0.34

Fig 2.4: Dataset 2 overview

The findings from dataset 2 is given below

- From the correlation matrix, it is concluded that there is strong positive relation between months. It indicates that temperature change in one month strongly effect the temperature change in other months.
- From March to October, there is correlation between 0.91 to 0.96 which indicates the higher relationships of temperature change in these months and same is true for others.
- Mean value of temperature change fluctuates over the months and February is the month where higher temperate change and June and July are months where small amount of temperature change is observed.
- From the year of the 1975 temperature change is constantly increasing in all months and the pattern of abrupt increase in temperature anomalies after the 1950 is observed. From 1975 onwards, it is noticed that there is constant increase in temperature anomalies till 2023 for every single month as shown in fig 2.5

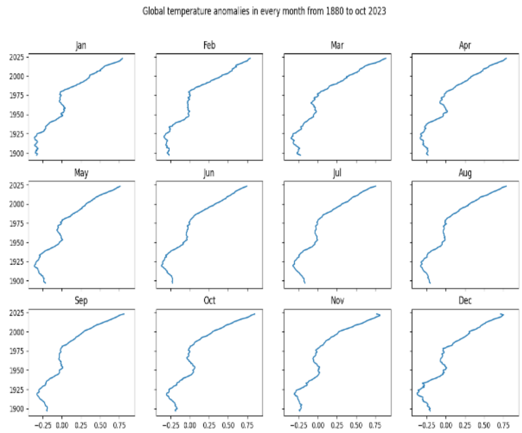


Fig 2.5: Monthly Temperature changed from 1880 till October 2023.

For Dataset 3

The dataset 3 contains the yearly global temperature anomalies from 1880 to October 2023. It is the latest data on temperature anomalies. The datapoints are normalized by using the statistical technique of lowess. This dataset helps to visualize the yearly global temperature anomalies and it is also used for machine learning modelling.

	Year	No_Smoothing	Lowess
0	1880	-0.17	-0.10
1	1881	-0.09	-0.13
2	1882	-0.11	-0.17
3	1883	-0.18	-0.21
4	1884	-0.29	-0.24

Fig 2.6: Dataset 3

The findings from the dataset 3 is given below

- From the correlation matrix of no smoothing and Lowess features it is observed that there is strong positive relation between them.
- From the correlation matrix, value of 0.98 indicated the strong positive relation between them. Therefore, Lowess feature was used for training of the model because on the normalized data the model optimized well and the non-normalized data was used for forecasting of temperature anomalies for next years which is known as out of sample forecasting.
- There are two years 2016 and 2020 that are outliers which indicated higher temperature change in these years from expected value.
- If the comparison is observed between the No smoothing data and normalized data it is concluded that all models perform well on smooth data. Smooth curve showed the constant increasing pattern which is good for learning of machine model as compared to no smooth data shown in fig 2.7
- From the dataset it is observed that 2016 and 2020 was the hottest year from 1880 to 2023 with the temperature change of 1.01 degree in both years.
- It is observed that 1909 and 1904 was the coldest year from 1880 to 2023.

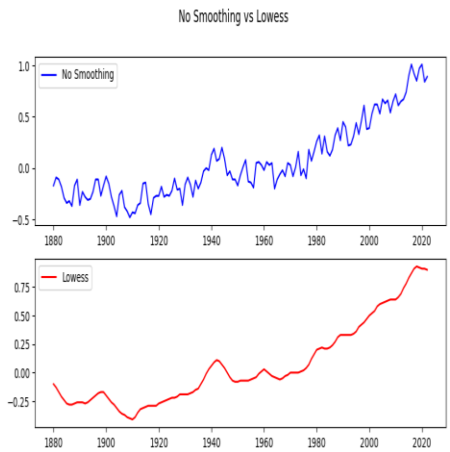


Fig 2.7 : No Smooth vs Lowess data

Machine Learning models Comparison

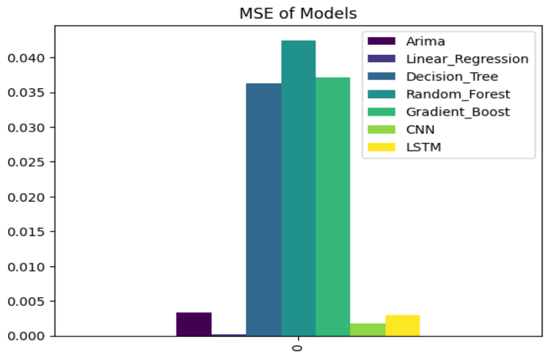
Since it was the regression problem of time series data therefore different machine learning models were used and evaluated based on their Mean Square Error (MSE).

The Chosen machine learning models are given below

- Linear Regression
- Decision Tree Regressor
- Randon Forest Regressor
- XGBoots Regressor
- CNN
- LSTM
- ARIMA

MSE of all tried models is given below

Arima	Linear Regression	Decision Tree	Random Forest	Gradient Boost	CNN	LSTM
0.003297	0.000172	0.03624	0.042492	0.037177	0.001758	0.003015



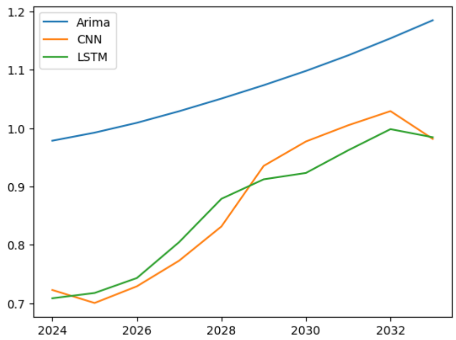
R2 Score of ARIMA, Linear regression, CNN and LSTM are given below

Arima	0.953666
Linear Regression	0.984624
CNN	0.820965
LSTM	0.903409

Tabel 2.2: R2 Score of Model

Lowess technique was applied on the data and data was already train by the linear regression model that's why it performed very well on data. But it does not perform well on unseen data. Therefore, it is discarded for the final prediction

The predicted temperature from 2024 to 2032 from these models is give below



	Arima	CNN	LSTM
2024	0.978479	0.722602	0.708418
2025	0.992308	0.700326	0.717490
2026	1.009176	0.728893	0.743092
2027	1.029030	0.772877	0.804673
2028	1.050689	0.831493	0.878894
2029	1.073624	0.935325	0.912307
2030	1.098127	0.977186	0.923248
2031	1.124802	1.004950	0.962079
2032	1.153979	1.029213	0.998308
2033	1.184934	0.981614	0.984405

Comparison with other research paper

The comparison was made with the other research paper which was published in 2022 [17] .

The below table makes the comparison between the proposed methodology and their [17] methodology

	Their approach	Our approach
Dataset	Global Surface Tempeature Change dataset from NASA	Global Surface Temperature Change dataset from NASA
Models	ARIMA	LR,RF,DT,XGBoot,LSTM,CNN,ARIMA
Best Model	ARIMA	ARIMA
R2 Score	0.92	0.953666

Limitations:

Gathering the dataset is a basic challenge for using the machine learning models. Climate change data changes with reference to time. So this models is only as good as the data. This can be further enhanced using more versatile and real time data.

Conclusion:

Global warming poses the upward linear trend with time and we might face the hottest years in coming decade. Different models were tried for the forecasting of global warming anomalies and it is finds out that Linear regression outperform the other models followed by the ARIMA, CNN and LSTM. These models were evaluated based on their MSE which is usually used for regression problem and at last means prediction was taken to male prediction robust and reliable.

Future Work:

The further enhancement can be made which are given below

- The forecasting can be made more robust by including other factors that stimulates the global warming like CO2 emission, forest fires, chemical exposure into environment etc.
- Real time data can give more accurate forecast since it includes different variable at a time to forecast.
- Since it is only univariate analysis, therefor it can be upscale to multivariate analysis.

References:

[1] "Global warming | Definition, Causes, Effects, Solutions, & Facts | Britannica." Accessed: Dec. 31, 2023. [Online]. Available: https://www.britannica.com/science/global-warming

[2] "Global Warming." Accessed: Dec. 31, 2023. [Online]. Available: https://education.nationalgeographic.org/resource/global-warming/

[3] "Global Warming 101 - Definition, Facts, Causes and Effects of Global Warming." Accessed: Dec. 31, 2023. [Online]. Available: https://www.nrdc.org/stories/global-warming-101#causes

[4] "Global warming and climate change effects: information and facts."

Accessed: Dec. 31, 2023. [Online]. Available: https://www.nationalgeographic.com/environment/article/global-warming-effects

[5] M. Purushotham Reddy, A. Aneesh, K. Praneetha, and S. Vijay, “Global Warming Analysis and Prediction Using Data Science,” in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2021, pp. 1055–1059. doi: 10.1109/I-SMAC52330.2021.9640944.

[6] K. Ahmed, D. A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan, “Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms,” *Atmos Res*, vol. 236, no. November 2019, p. 104806, 2020, doi: 10.1016/j.atmosres.2019.104806.

[7] S. M. Malakouti, “Utilizing time series data from 1961 to 2019 recorded around the world and machine learning to create a Global Temperature Change Prediction Model,” *Case Studies in Chemical and Environmental Engineering*, vol. 7, no. February, p. 100312, 2023, doi: 10.1016/j.cscee.2023.100312.

[8] S. Salcedo-Sanz *et al.*, “Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review,” *Theor Appl Climatol*, 2023, doi: 10.1007/s00704-023-04571-5.

[9] N. Khan, S. Shahid, T. Bin Ismail, and F. Behlil, “Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change,” *Stochastic Environmental Research and Risk Assessment*, vol. 35, no. 7, pp. 1335–1353, 2021, doi: 10.1007/s00477-020-01963-1.

[10] F. Zennaro *et al.*, “Exploring machine learning potential for climate change risk assessment,” *Earth Sci Rev*, vol. 220, no. June, p. 103752, 2021, doi: 10.1016/j.earscirev.2021.103752.

[11] M. Ali, R. C. Deo, N. J. Downs, and T. Maraseni, “Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting,” *Comput Electron Agric*, vol. 152, no. April, pp. 149–165, 2018, doi: 10.1016/j.compag.2018.07.013.

[12] R. S. Priya and K. Vani, “Climate Change Forecast for Forest Fire Risk Prediction using Deep Learning,” *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, vol. 1, pp. 1065–1070, 2023, doi: 10.1109/ICACCS57279.2023.10112983.

[13] H. Bhatt, M. Davawala, T. Joshi, M. Shah, and A. Unnarkat, “Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques,” *Cleaner Chemical Engineering*, vol. 5, no. June 2022, p. 100095, 2023, doi: 10.1016/j.clce.2023.100095.

[14] T. Schillerberg and D. Tian, “Changes in crop failures and their predictions with agroclimatic conditions: Analysis based on earth observations and machine learning over global croplands,” *Agric For Meteorol*, vol. 340, no. April, p. 109620, 2023, doi: 10.1016/j.agrformet.2023.109620.

[15] “Data.GISS: GISS Surface Temperature Analysis (GISTEMP v4).” Accessed: Dec. 31, 2023. [Online]. Available: https://data.giss.nasa.gov/gistemp/

[16] “Annual Surface Temperature Change | Annual Surface Temperature Change | Climate Change Indicators Dashboard.” Accessed: Dec. 31, 2023.

[Online]. Available: https://climatedata.imf.org/datasets/4063314923d74187be9596f10d034914/explore

[17] G. Chen, B. Xiao, L. Wei, Z. Liao, J. Li, and Z. Zhu, “Analysis of global land-based ocean temperature indices based on time series analysis,” in *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2022, pp. 679–683. doi: 10.1109/IAEAC54830.2022.9929784.