Vol. 7, No. 2 | July December 2023

PREDICTIVE ANALYSIS OF CLIMATE DISASTER DATA

Anum Aziz[1], Shaukat Wasi[2], Muhammad Khaliq-ur-Rahman Raazi Syed[3]

Abstract:

In this paper, the Total deaths and Cost per Index (CPI) of worldwide climate disaster dataset has been modelled. The time period of the dataset is from 1900 to 2021. The Autoregressive Integrated Moving Average (ARIMA) has been applied to forecast the Total Deaths and CPI of the study area. The total of 75% of the train data is used for construction of the model and the remaining 25% dataset is used for testing the model. The ARIMA model is general provides more accurate projection especially interval forecast and is more reliable than other common statistical techniques. The best-fitted model is identified as ARIMA*(2,0,1) and (2,1,2) for Cost per Index CPI and Total Deaths* respectively, generated on the basis of minimum values of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) procedures. The accuracy parameter considered as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) both parameters shows the model is accurate respectively. There is a 7% difference between the auto and manual models for the CPI feature, similarly, there is a 4% difference for Total Deaths, indicating that CPI plays a significant impact in climatic disasters. In order to identify best fitted model, we applied the model manually and automatic processing. By means of Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots, the most appropriate order of the ARIMA model are determine and evaluated. Accordingly the created model can help in determining future strategies related to climate disaster dataset of the world. From the forecast result it is found that the results seems to show an increasing trend in CPI values and the minimal decreasing in total death condition and economic activities of the world.

Keywords: *Climate Disaster,Climate Predictions, Climate evolution, Disaster Management*

1. Introduction

The utilization of large-scale climate data is becoming increasingly important in forecasting climate change and understanding its influence on a variety of issues, including environmental illnesses. Climate data can now be collected on a worldwide scale thanks to technological improvements and the availability of sophisticated sensors, reflecting seasonal oscillations and offering vital insights for weather forecasting, climate modeling, and seasonal disease analysis. The availability of comprehensive global climate data has also resulted in breakthroughs in environmental legislation and international accords. Regulators and policymakers rely on current and precise data to develop plans and make educated decisions about reducing greenhouse gas emissions, adapting to climate change, and allocating resources to combat it. Furthermore, combining large-scale climatic data with socioeconomic aspects allows for a better understanding of the social implications

[1] Department of Computer Science, Muhammad Ali Jinnah University, Karachi 75400, Pakistan. Corresponding Author: anumaziz9900@gmail.com

of climate change. Stakeholders and legislators can design effective ways to alleviate the detrimental consequences of climate change on those most vulnerable by examining relationships between climate-related factors and health outcomes, food security, migratory patterns, and economic indicators.

The effects of climate change are becoming more visible, resulting in natural catastrophes and extreme weather conditions that significantly impact people in need, particularly in countries that are developing. Rising temperatures, shifting rainfall patterns, and extreme weather events all have serious repercussions, including higher fatality rates. Effective temperature forecasting and understanding the implications of climate change are critical for future planning and reducing the effects of natural catastrophes. Intelligent infrastructure is critical in disaster management for gathering, integrating, managing, and analyzing disparate dispersed data sources. Climate change-related difficulties are exacerbated by shifting rainfall patterns. Prolonged droughts in certain areas cause water scarcity, agricultural failure, and food instability. Others may experience more heavy rains and a higher danger of floods, resulting in relocation, infrastructure devastation, and the spread of waterborne illnesses. Understanding the effects of shifting rainfall patterns is critical for planning for the future, particularly managing water resources, agricultural practices, and susceptible urban growth. Disaster prevention and response activities rely heavily on intelligent infrastructure. Intelligent infrastructure may acquire, integrate, manage, and analyze various data sources by using contemporary tools such as remote sensing, satellite photography, and real-time data-gathering systems. This allows for precise and timely information on weather patterns, environmental conditions, and possible threats, allowing for early warning systems, evacuation strategies, and resource allocation during natural disasters. Intelligent infrastructure also aids in disaster recovery and resilience-building by promoting data-driven decision-making and effective collaboration among multiple stakeholders. Predictive analytics seeks to estimate future system behaviors based on past data. One of the most significant benefits of machine learning is its capacity to deliver insights via various sorts of analytics. Descriptive analytics is concerned with analyzing historical data in order to comprehend underlying processes, discover patterns, and solve critical concerns. It assists stakeholders in gaining a full knowledge of previous events, assessing their impact, and making well-informed choices on the basis of that information. On the other hand, predictive analysis seeks to forecast future system behaviors by analyzing historical data. Machine learning algorithms can anticipate and predict the future climate by recognizing patterns and connections in previous climate data. This allows stakeholders to predict future hazards, arrange for mitigation measures, and efficiently allocate resources. Another key part of machine learning is predictive analytics. It extends further descriptive and predictive analytics by making suggestions and recommending actions to improve results. Prescriptive analytics can provide practical insights on how to mitigate the effect of climate-related disasters by analyzing large-scale climate data and taking into account numerous influencing factors.

Finally, prescriptive analytics entails making the best future judgments based on the outcomes of analytical methods such as descriptive and predictive. Given these improvements and obstacles, the goal of this thesis is to perform a thorough examination of several data processing prototypes, such as spatial autocorrelation models, binary segmentation models, closest neighbor algorithms, and principal component analysis. Forecasting and future forecasts are critical in climate-related research for understanding the possible implications of climate change and establishing effective strategies for mitigation and adaptation. This method makes use of historical climate data, statistical modeling, and advanced analytical tools.

2. Literature Review

Large-scale climate data have been used to anticipate climate changes and disease from the environment. It consists of two increase scalability and feasibility, various data processing prototypes have been developed, including spatial autocorrelation models, binary segmentation models, nearest neighbor algorithms, principal component analysis as an unsupervised model, and nearest neighbor algorithms. This work conducts a thorough analysis of the aforementioned approaches to develop a fresh paradigm to handle complex climate data. Data analytics is becoming increasingly important in fields such as healthcare, social networking, climate modeling, and so on. Climate data, which reflects seasonal fluctuations, might be acquired with the sophisticated sensor. Meteorological data is used to anticipate the weather, and weather data is also valuable for analyzing seasonal diseases and reflecting seasonal changes [1][12]. The ongoing difficulty in worldwide healthcare research is determining the risk presented by epidemics of infectious diseases as our understanding of them improves and the geographical range that exists naturally increases. As the size of spatial epidemiology data expands, utilizing usable intelligence in these data has become a priority. They share volume, velocity, diversity, value, and authenticity, which are all data analysis properties. Spatial epidemiology data is a critical component of big data and healthcare analytics in digital epidemiology. The purpose of this study is to examine the geographical climate data issues in infectious illness monitoring, with an emphasis on influenza epidemics [2]. Climate change is a crucial role in determining the magnitude of other factors. Handling catastrophes has played an important role in reducing and minimizing loss of life and property damage. Intelligent infrastructure for the gathering, integration, administration, and analysis of disparate distributed data sources is required to facilitate efficient disaster management [3]. Infrastructure disaster including those caused by hydrological, climatic, and climatological consequences, have become more severe and frequent, putting cities around the world to the test. The effects of climate change have been related to higher snow loss, faster sea level rise, more frequent heat waves and droughts, stronger hurricanes, and, most importantly, a continuous and rapid rise in global temperatures. [4][11]. Descriptive analytics is focused with analyzing historical data in order to comprehend the processes under consideration, answering critical questions about these processes, and making meaningful conclusions. Predictive analytics seeks to forecast the future behavior of systems and entities based on such findings. Finally, prescriptive analytics focuses on determining the optimum future decision(s) based on descriptive and predictive analytics results [5]. The current study uses data analytics and machine learning approaches to provide a performance prediction for CPI infrastructure networks [13]. Data analytics, which is separated into descriptive, predictive, and prescriptive analytics, tries to find hidden information that cannot be investigated using traditional statistical and mathematical methods [6]. An interesting application topic for social sensing is emergency management. Despite this, little effort has been made to acquire fast estimates of the effects of disasters on the people and infrastructure. The use of crowd sourced social data, such as eyewitness testimonies, in estimating damage had long been claimed. However, the present methods dependent on citizen reporting may take days to get final conclusions [7][8].

Many studies have been published that investigate shifting patterns in temperature, precipitation, and discharge, as well as their interactions across the world. Trend analysis of historical climatic data is a critical step in determining a region's climate status. It offers an overall assessment of the fluctuations in climatic variables during a certain time period [9].

3. Methdology

This research work comprises of many essential phases to achieve the goals of analyzing complicated climate data, building a new paradigm for dealing with such data, and utilizing data analytics for climate modeling death monitoring, and CPI. The essential components of the technique are outlined below.

3.1 Climate Dataset

Data gathering for climate catastrophe research may come from a variety of sources of information, and Kaggle is one site where you can get datasets connected to climate and catastrophic occurrences. It is a great resource for field researchers. The dataset chosen for study has a total number 16,127 records spanning the years 1900 to 2021. This broad time span enables the study of climate-related disasters over a long period of time. The dataset has 45 Characteristics, which provide a comprehensive collection of variables capturing various elements of the events. The dataset contains metadata about past disasters, such as the type of disaster (e.g., floods, hurricanes, wildfires), the date and time period when the disasters occurred, the geographical area or location where the events occurred, and potentially additional details such as severity or magnitude. These qualities provide useful information for understanding the characteristics, trends, and effects of climate-related disasters throughout history.

However, it has been shown that certain records include a high amount of null or trash data, rendering them unsuitable for analysis. As a result, these faulty entries must be removed from the dataset. Similarly, records containing trash values must be detected and eliminated, as they may reveal data input mistakes or discrepancies. These records can skew the analysis and lead to incorrect conclusions. By removing them, we verify that the dataset contains valid and useful data. After preprocessing, just 18 features of the initial 45 features with 15,071 records regarded compatible and adequate for the specified study purpose.

3.2 Dataset Analysis

This EDA process gives a basic knowledge of the dataset and aids in hypothesis formulation, modelling tool selection, and assumptions. We often investigate many features, such as summary statistics, distributions, correlations, and visualizations.

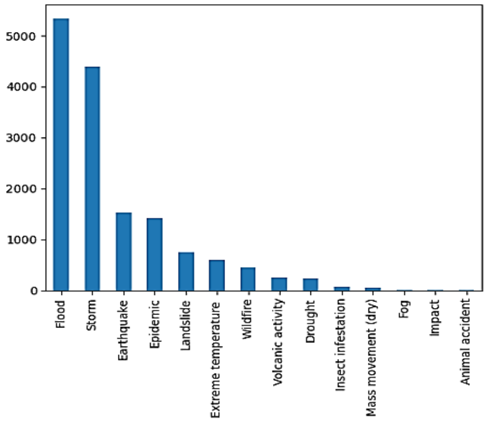

Fig. 1: Trend of Disaster Type

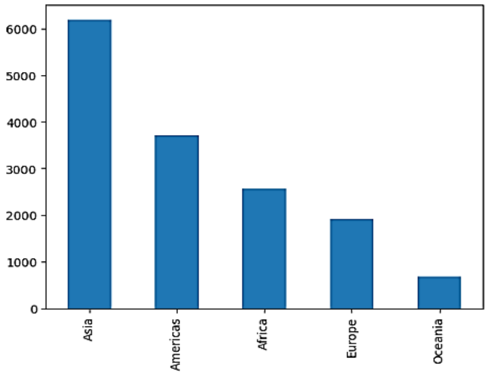

Fig. 2: Number of Disasters Occur in Regions

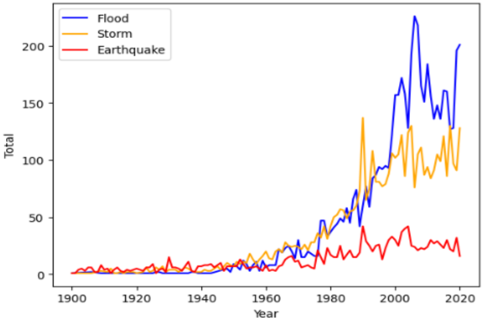

Fig. 3: Trend of Flood, Storm, Earth Quake over the time period

3.2.1 Check Data is Stationary or not?

The Augmented Dickey-Fuller (ADF) test is one of the most frequent and commonly used methodologies for determining stationarity. The ADF test is a statistical test used to detect whether or not a time series is stationary. Many statistical software packages and computer languages support it, notably Python's statsmodels module**.** The null hypothesis in the ADF test asserts that the time series is non-stationary. If the p-value produced from the test is less than or equal to the desired significance threshold (often set to 0.05), the null hypothesis can be rejected, indicating that the data is stationary.
After calculating,
p-value: 0.00010390767452394873 is less than 0.5 that means data is stationary

3.3 Predictive Analysis Modelling

The ARIMA model is used for prediction and forecasting in the dataset trend analysis. Two strategies are used: ARIMA modeling, both automatic and manual. The first step in choosing an appropriate model is ensuring that the time series data is stationary. Stationarity is a time series property in which statistical parameters like mean, variance, and autocorrelation stay constant across time [14].

Basic Components of ARIMA:

1. Autoregressive (AR) model: To predict Y_t by one or multiple lagged value. This is represented by equation mentioned below.

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$$

2. Moving Average (MA) model: To predict Y_t by one or multiple lagged value of the error. This is represented by equation mentioned below.

$$Y_t = c + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q}$$

3. Differencing (Integration): In ARIMA Model, Time Series data must be stationary to obtaining useful information.

$$Y'_t = Y_t - Y_{t-1}$$

A stationary time series is essential for ARIMA modeling since it implies data stationarity. The next step is to choose the AR (Autoregressive) and MA (Moving Average) parameters for the ARIMA model [3][10[15]. The ACF plot, also known as the autocorrelation plot, aids in establishing the order of the MA component. After a given lag, the autocorrelation values in an MA process drop to zero. This suggests that the MA component is capturing the time series' random shocks or mistakes. The autocorrelation plot for an AR process, on the other hand, diminishes gradually or geometrically. This suggests that prior observations in the time series have persisted. We can establish the optimal order of the AR component in the ARIMA model by watching the decline of autocorrelation in the ACF plot. Following are the four main stages the model takes inputs and provides desired outputs:

3.3.1 Stage 1: Model Inputs

Several factors are included in the current study on disaster prediction to forecast the (CPI) and Total Deaths. Among the variables chosen are: *Start_Month_Year*, *End_Month_Year*, *Disaster Type*, *Disaster Subgroup*, *Region*, *Dis Mag value*, *Continent.* We can investigate the correlations, trends, and possible predictive value of each variable alone or in combination by including them in the study. The combining of several characteristics allows for a more complete study and can improve the accuracy and resilience of CPI and Total Deaths prediction models in the context of disaster situations.

3.3.2 Stage 2: Model Selection and Prediction

The ARIMA (Autoregressive Integrated Moving Average) model was chosen as the optimal way for analyzing and predicting the dataset throughout the model selection and prediction procedure. The ARIMA model is frequently used for time series forecasting and is particularly excellent at capturing the data's temporal patterns and statistical behavior. To use the ARIMA model, you must first establish the AR (Autoregressive) and MA (Moving Average) components. The ARIMA model is chosen and applied by analyzing the statistical behavior of the data, choosing the suitable AR and MA values, training the model, and generating future predictions [3][10][15].

3.3.2.1 The Auto ARIMA Model

The Auto ARIMA is a valuable tool since it automates the laborious task of manually picking the best parameters for an ARIMA model, which may be difficult and time-consuming. It is especially useful for those who do not have a comprehensive grasp of time series modeling but wish to make use of its forecasting capabilities. This approach is intended to make it easier to choose the right order of lags and variance variables for an ARIMA model. It uses a stepwise strategy to search through various parameter combinations, assessing the performance of each model using a chosen criterion (such as AIC, BIC, or AICs), and identifying the hypothesis with the best performance. We calculate the Auto ARIMA model for CPI Table1 and Total Deaths Table2 below.

Table 1: Auto AR and MA values for CPI

ARIMA(p,d,q)	*AIC*	*TIME*
ARIMA(0,0,0)	132463.091	0.15 sec
ARIMA(0,0,1)	infinity	1.28 sec
ARIMA(0,0,2)	103230.402	4.77 sec
ARIMA(0,0,3)	infinity	23.71 sec
ARIMA(1,0,0)	infinity	0.69 sec
ARIMA(1,0,1)	33134.731	6.40 sec
ARIMA(1,0,2)	33041.430	1.36 sec
ARIMA(1,0,3)	33036.118	8.94 sec
ARIMA(2,0,0)	infinity	7.25 sec
ARIMA(2,0,1)	33041.399	11.05 sec
ARIMA(2,0,2)	33138.72	8.68 sec
ARIMA(2,0,3)	33038.861	11.42 sec
ARIMA(3,0,0)	infinity	8.67 sec
ARIMA(3,0,1)	33083.273	8.46 sec
ARIMA(3,0,2)	33018.448	12.34 sec

Table 2: Auto AR and MA values for Total Deaths

ARIMA(p,d,q)	*AIC*	*TIME*
ARIMA(0,0,0)	145385.231	0.18 sec
ARIMA(0,0,1)	144316.049	0.85 sec
ARIMA(0,0,2)	143833.524	2.38 sec
ARIMA(0,0,3)	143628.520	4.20 sec
ARIMA(1,0,0)	143832.983	0.97 sec
ARIMA(1,0,1)	141902.743	3.41 sec
ARIMA(1,0,2)	141787.031	5.94 sec
ARIMA(1,0,3)	141765.330	10.00 sec
ARIMA(2,0,0)	143315.159	0.94 sec
ARIMA(2,0,1)	141776.757	6.66 sec
ARIMA(2,0,2)	infinity	9.17 sec
ARIMA(2,0,3)	infinity	19.75 sec
ARIMA(3,0,0)	143071.613	1.41 sec
ARIMA(3,0,1)	141764.384	8.01 sec
ARIMA(3,0,2)	infinity	16.45 sec

The model that has a minimum AIC Value is the best-fitted model. *ARIMA (3,0,2)* has minimum AIC for CPI and *ARIMA(3,0,1)* for Total Deaths . These (p,d,q) values will be used for auto-modeling in a dataset.

3.3.2.2 The Manual ARIMA Model

The manual ARIMA model incorporates the autoregressive (AR), differencing (I), and moving average (MA) components, as does the Auto ARIMA model. The manual technique, on the other hand, gives researchers greater power and flexibility in determining the optimal parameters for the ARIMA model. It is critical to validate the dataset's stationarity before using the manual ARIMA model. The next stage in the manual ARIMA model is to find the appropriate AR and MA parameters after confirming stationarity.

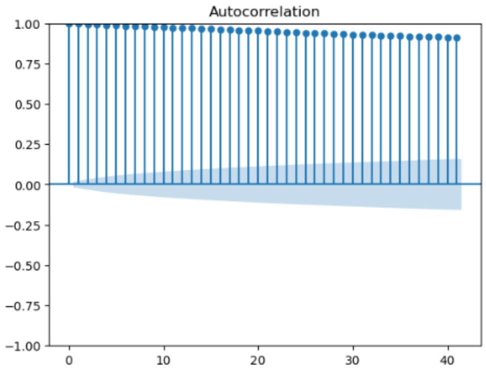

Fig. 4: ACF for CPI

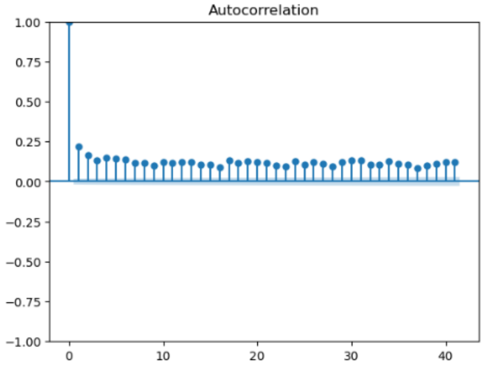

Fig. 5: ACF for Total Deaths

The Autocorrelation Function (ACF) is a useful tool for finding the lag value (p) for strong correlation in a collection of independent features. Figures 4 and 5 are most likely ACF plots demonstrating correlation coefficients at various lag levels. We can uncover significant association patterns and appropriate values for p and q by analyzing the ACF and PACF plots.

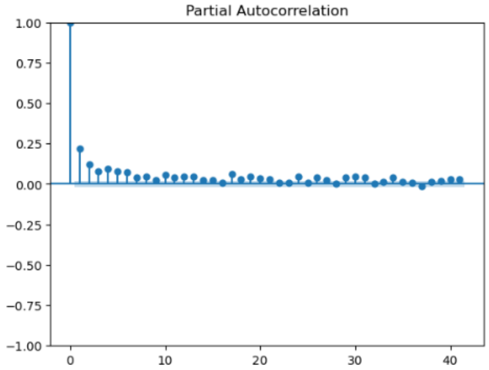

Fig. 6: PACF for CPI

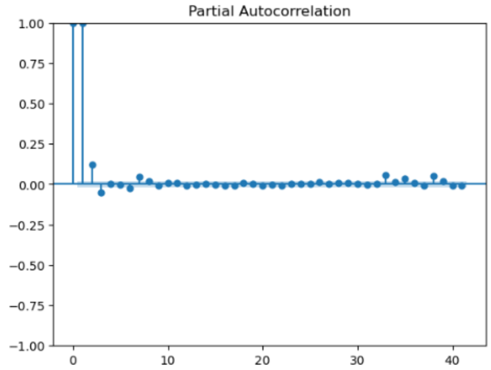

Fig. 7: PACF for Total Death

These parameters are critical in building the manual ARIMA model. According to the text, ARIMA (2,0,1) is the best-fitted model for CPI, with an autoregressive order of 2 (AR = 2), no differencing (d = 0), and a moving average order of 1 (MA = 1). Similarly, the best-fitting model for Total Deaths is ARIMA (2,1,2), which implies an autoregressive order of 2 (AR = 2), differencing order 1 (d = 1), and a moving average order of 2 (MA = 2). These parameter values are calculated by analyzing the ACF and PACF plots, which correspond to the patterns and statistical behavior seen in the corresponding datasets.

We generate predictions and projections for CPI and Total Deaths using the manually specified ARIMA (2,0,1) and ARIMA(2,1,2) models, which incorporate the indicated autoregressive, differencing, and moving average components. These models provide a more tailored approach, allowing researchers to fine-tune the ARIMA model to their individual dataset and analysis needs.

3.3.3 Stage 3: Model Validation

It entails assessing the model's performance, correctness, and generalizability in order to assure its dependability and use in real-world applications. Accurate forecasts and credible models are critical for successful disaster management, risk assessment, and decision-making in the context of climate disaster frameworks. In this research, we will go deeper into the model validation process, its significance, and the numerous strategies used. Model validation evaluates how well the model works on previously unknown or future data, since it is critical to guarantee that the model can generalize beyond the data provided for training.

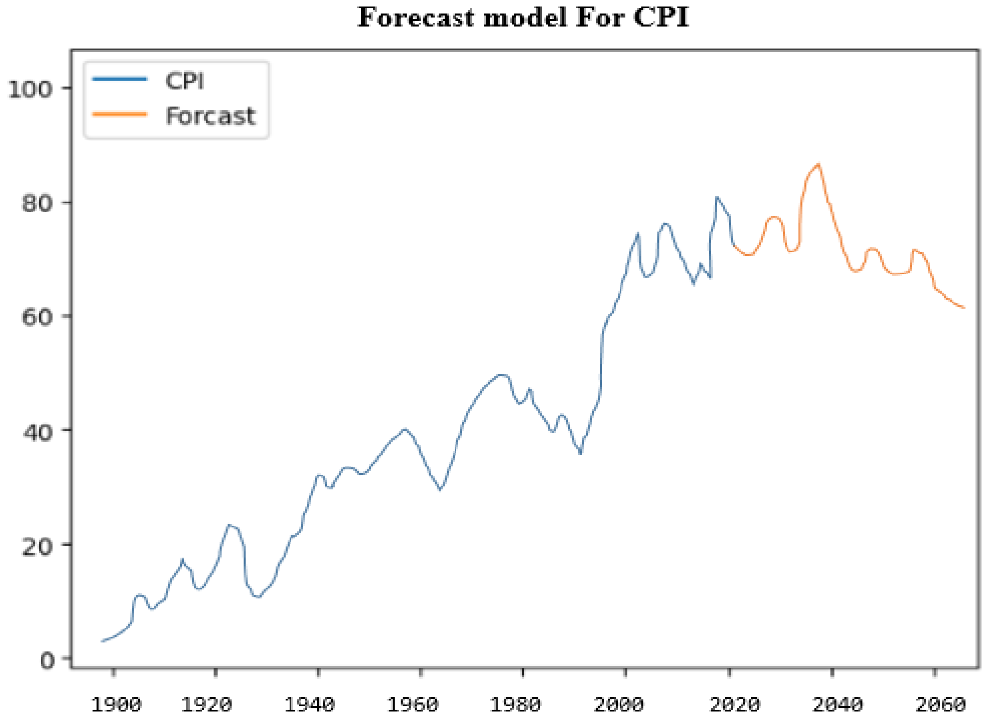

Fig. 8: Actual and Forecast for CPI

3.3.4 Stage 4: Actual and Forecast Prediction

To evaluate the accuracy of the model's forecasts, the forecast and actual lines are frequently shown on the same graph for comparison. This combined graph shows the model's predictions match the actual data. Figure 08 and 09 shows the actual and forecast values of CPI and Total Deaths with respect to the time period of 1900 to 2021. The Forecast were generated from the best fitted ARIMA (2,0,1) and ARIMA(2,1,2) models respectively. This shows the model seems to be accurately. The models' accuracy will be measured by MAE and RMSE tests.

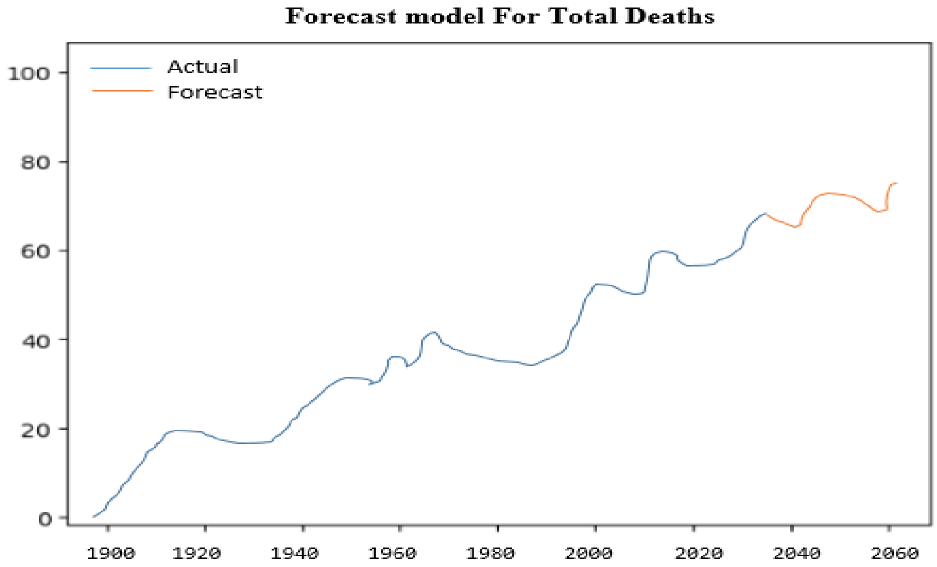

Fig. 9: Actual and Forecast for Total Deaths

4. Results

It is critical to validate the ARIMA model in order to evaluate its performance and correctness. Both the auto and manual ARIMA models are examined in this scenario using two assessment metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics give insight into the models' forecast performance and quantify the Gap between expected and actual values. Tables 3 and 4 are most likely the anticipated performance indicators for CPI and Total Deaths, respectively. These tables offer an overview of the model's predictive accuracy for the various variables.

Table 03: Forecast Performance measures for CPI feature

1900-2021	MAE	RMSE
Auto	0.32193142421649	0.3403954082930152
Manual	0.23143605915805	0.261272408584486

Table 04: Forecast Performance measures for Total Death feature

1900-2021	MAE	RMSE
Auto	0.5240172128499798	0.5990340634941405
Manual	0.4602817487618751	0.5580509383657063

In summary, the MAE and RMSE evaluation metrics give quantitative assessments of the forecast performance of the auto and manual ARIMA models. By comparing these measurements, researchers may identify which model predicts CPI and Total Deaths more accurately, allowing them to make educated judgments and interpretations based on the anticipated numbers.

5. Conclusion

Time series analysis is critical in forecasting and predicting the CPI and Total Deaths during global climatic disasters. The goal of this work is to analyze the Climate Disaster dataset from 1900 to 2021 and construct models to make forecasts. To find the best-fitted model for the dataset, both automated and human testing methods are used. This assumption is critical to the model's dependability. The MAE and RMSE measures are used to assess how well the auto and manual models predict CPI and Total Deaths. These measures assess the difference between anticipated and actual values. The analytical findings show that both traits are important. There is a 7% difference between the auto and manual models for the CPI feature, indicating that CPI plays a significant impact in climatic disasters. Similarly, there is a 4% difference between the two models for Total Deaths, underscoring the importance of this variable in assessing the effect of climatic catastrophes. This gives important insights into the probable effects and implications of climate change on economic parameters like the CPI and human lives as indicated by Total Deaths.

The study's findings and insights might be useful to policymakers, researchers, and practitioners working in the fields of climate science and disaster management. These findings add to a better understanding of the effects of climate change and serve as a foundation for future studies employing more advanced modeling methodologies in risk reduction initiatives. Researchers can use the tools and approaches employed in this work to expand on and construct more complex and robust models. Future research may build on these limits to refine and improve climate disaster prediction models, eventually leading to more effective risk reduction measures and climate change adaption plans.

References

[1] V.Nandhini & Geetha Devasena, 2019, *Predictive Analytics for Climate Change Detection and Disease Diagnosi.* 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

[2] M. Gunasekaran, Senthil Murugan, Harpreet Kaur, Kaja M. Abbas, 2014, *Spatial Big Data Analytics of Influenza Epidemic in Vellore, India,* IEEE International Conference on Big Data.

[3] Muhammad Amjad, 2022, Analysis of Temperature Variability, Trends, and Predictions in the Karachi Region of Pakistan Using ARIMA Model, Pakistan, Academic Editor

[4] Cheng Zhang, Chao Fan, Wenlin Yao, Xia Hu, Ali Mostafavi, 2019, *Social media for intelligent public information and warning in disasters: An interdisciplinary review,* International Journal of Information Management.

[5] Jedsada Phengsuwan, 2021**,** Use of Social Media Data in Disaster Management: A Survey, Future Internet

[6] May Haggag, 2021, Infrastructure performance prediction under Climate-Induced Disasters using data analytics, International Journal of Disaster Risk Reduction.

[7] Marco Avvenuti, 2017, *Nowcasting of Earthquake Consequences using Big Social Data,* IEEE Internet Computing

[8] Gary Feng, 2016, *Trend analysis and forecast of precipitation, reference evapotranspiration and rainfall deficit in the Blackland Prairie of Eastern Mississippi,* Journal of Applied Meteorology and Climatology.

[9] Rashid Mahmood, 2017, *Spatial and temporal hydro-climatic trends in the transboundary Jhelum River basin*, Journal of Water and Climate Change

[10] Andrea L. Schaffer, 2021, *Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions*, BMC Medical Research Methodology.

[11] Benoˆıt Sarr, 2012, *Present and future climate change in the semi-arid region of West Africa: a crucial input for practical adaptation in agriculture,* Royal Meteorological Society.

[12] Brian R. Pickard, 2015, *Translating Big Data into Big Climate Ideas*, Research Gate

[13] N. W. Arnell, 2019, *Global and regional impacts of climate change at different levels of global temperature increase*, Climate Change

[14] Rashid Mahmood, 2019**,** *Global and regional impacts of climate change at different levels of global temperature increase*, Scientific Report.

[15] YUCHUAN LAI, 2019, *Use of the Autoregressive Integrated Moving Average (ARIMA) Model to Forecast Near-Term Regional Temperature and Precipitation*, American Meteorological Society